STEPHEN SHENNAN

ARQUEOLOGÍA CUANTITATIVA

Traducción castellana de
JUAN ANTONIO BARCELÓ

EDITORIAL CRÍTICA
BARCELONA

PRÓLOGO A LA EDICIÓN ESPAÑOLA

Ni qué decir tiene que estoy encantado de que mi libro haya sido traducido al castellano, y espero que el éxito justifique el esfuerzo. La publicación de esta traducción me ofrece la oportunidad de volver al tema y considerar los desarrollos que han tenido lugar desde la publicación de la versión inglesa.

Mi intención al escribir el libro fue la de proporcionar al lector una introducción al pensamiento cuantitativo en el análisis de datos arqueológicos, presentando algunos de los métodos más útiles en el análisis de datos y que más han sido aplicados por los arqueólogos, desde los más sencillos hasta las relativamente complejas técnicas multivariantes. Tal introducción no sólo debía utilizar ejemplos arqueológicos para ilustrar los métodos, sino que la descripción de los mismos había de hacerse de forma amena para los estudiantes no acostumbrados a pensar matemáticamente, que aún son mayoría en nuestra disciplina. En otras palabras, las descripciones tenían que basarse en palabras y gráficos más que en ecuaciones. Escribiendo este libro me di cuenta de que las dificultades que yo había experimentado al estudiar esos métodos eran una ventaja y no un inconveniente, porque podía entender los problemas a los que debía enfrentarse un no matemático. La acogida positiva que ha tenido el libro desde su aparición puede haber justificado mis esperanzas e intenciones.

Las razones para escribirlo siguen siendo válidas: el pensamiento cuantitativo es esencial para la definición de asociaciones significativas en los datos arqueológicos. Los arqueólogos debieran ser capaces de realizar por sí mismos al menos las técnicas más simples, así como evaluar aquellas publicaciones que utilicen análisis cuantitativos. Aún lo creo así, a pesar de la tendencia hacia enfoques humanísticos o posprocesuales, de creciente importancia en los últimos años, sobre todo en la arqueología europea, ya que los métodos cuantitativos no tienen por qué estar estrechamente vinculados a la arqueología procesual, de la cual emergieron.

El libro no pretendía describir las técnicas más recientes y complejas que se usan en la arqueología cuantitativa; no obstante, creo que conviene hacer algunos comentarios acerca de los últimos avances, especialmente en las técnicas abordadas en el libro. Quizás lo más significativo haya sido la indudable expansión del análisis de correspondencias, infravalorado por la arqueología

angloamericana cuando escribí este libro. Aunque se aborda su estudio en él, la descripción es muy breve y no se presenta con el mismo detalle que la del análisis de componentes principales. Recientemente se han publicado numerosos ejemplos arqueológicos del uso del análisis de correspondencias (por ejemplo, Ringrose, 1988; Gob, 1988; Bertelsen, 1988; los trabajos más generales de Djindjian, 1989; Madsen, 1989, y los artículos que aparecen en Madsen, 1988), por lo que quizás sea la técnica del análisis multivariante más empleada hoy en día, debido, sobre todo, a las posibilidades de aplicación que ofrece; sin embargo, las descripciones de ese método suelen ser muy técnicas o demasiado superficiales. Uno de los factores que detuvieron inicialmente su difusión en la arqueología fue la falta de programas de ordenador accesibles, deficiencia que empieza a resolverse.*

Una consecuencia indirecta del auge del análisis de correspondencias ha sido el declive del análisis de conglomerados (o cluster analysis*), así como del análisis de componentes principales y de las escalas multidimensionales. La actitud hacia el empleo del análisis de conglomerados en las clasificaciones arqueológicas está cambiando, apareciendo, por un lado, ciertos intentos de reafirmar el valor de los enfoques clasificatorios tradicionales (por ejemplo, Adams, 1988), y por el otro una crítica a los enfoques politéticos en los que se basa el análisis de conglomerados, que emerge de los trabajos en psicología cognitiva. Así, hay quien afirma que las clases no han de caracterizarse como conjuntos politéticos, sino basándose en ciertos individuos «prototípicos», a partir de los cuales se define un «anillo» de densidad decreciente a medida que los objetos se convierten en miembros excéntricos de su clase (Rosch y Lloyd, 1978; en arqueología, Cowgill, 1990). Otros enfoques relativos a la clasificación arqueológica se han desarrollado a partir del análisis de las propiedades de simetría de los artefactos (Washburn y Crowe, 1988), y del uso de «gramáticas de motivos» (por ejemplo, Sharp, 1988; Chippindale, 1986, citados en Cowgill, 1990), cuyo objetivo es definir las reglas de combinación de los atributos decorativos, que a su vez generarán los esquemas observados en conjuntos específicos. Con todo, parece poco probable que alguna de estas técnicas pueda superar el uso del análisis de conglomerados, al menos para ciertos propósitos.*

*Otra área de desarrollo directamente relacionada con el contenido del libro es la disponibilidad creciente de programas de ordenador, especialmente para realizar análisis multivariantes. Hodson y Tyers han creado un paquete de programas informáticos para el análisis de datos procedentes de necrópolis (IAP);** Scollar propone varios programas para diferentes actividades arqueológicas, incluyendo la seriación; Wrigth ha escrito MV-ARCH, un conjunto de programas de análisis multivariante muy útil, especialmente ideado para arqueólogos, y que cuenta con un manual lleno de consejos de gran interés. Todos esos programas incluyen versiones del análisis de correspondencias, que es el único propósito del paquete CANOCO, orientado especialmente para la ecología. Aunque los formatos de entrada de datos son muy poco flexibles, al menos en la versión inicial del programa, contienen un amplio abanico de opciones que incluyen la aleatorización de las pruebas de significación de los resultados.*

* Véanse las Adiciones a la Bibliografía. (*N. del ed.*)

** Véase anexo 2, pp. 337-340. (*N. del ed.*)

*Para aquellos que deseen estar al día en los avances en arqueología cuantitativa existen diversas fuentes: las Actas del Congreso Anual sobre Aplicaciones Informáticas en Arqueología (*Computer Applications in Archaeology *o CAA, como la conocen los asiduos a ella), publicadas por British Archaeological Reports (BAR, Oxford), los volúmenes resultantes de las reuniones de la Comisión 4 de la UISPP (el más reciente es Voorrips, 1990), y revistas como* Quantitative Anthropology, Archeologia e Calcolatori.

Si nos referimos ahora a los avances que se esperan en un próximo futuro, es preciso señalar que los más importantes se asociarán a las posibilidades de los sistemas de información geográficos (o GIS, Allen et al., *1990) en análisis espacial, un área que merece un libro por sí sola. Merecen también atención los sistemas expertos asociados a programas de estadística, que guiarán al usuario en la selección de la técnica más apropiada. En cuanto a lo que es estrictamente análisis estadístico de datos arqueológicos, la tendencia continuará hacia el desarrollo de métodos estadísticos y modelos que no sólo pongan de relieve los rasgos característicos de los datos arqueológicos, sino que intenten representar la relación entre los datos recuperados por el arqueólogo y la realidad del pasado. En el ámbito del análisis de conjuntos cerámicos, un avance importante es el método desarrollado por Orton y denominado «trozo de pastel» (*pie slice*) (Orton y Tyers, 1989, 1991; Tyers y Orton, 1991), mientras que Litton y otros han empezado a explorar las posibilidades de la modelización bayesiana de los problemas arqueológicos, incluyendo los estudios de procedencia y la interpretación de datos de radiocarbono (Buck y Litton, 1991; Litton y Leese, 1991). Sin embargo, el estudio de esos modelos no ha hecho más que empezar.*

Espero que esta traducción castellana aumente el número de arqueólogos hispanohablantes que valoren las posibilidades de los enfoques cuantitativos en esta disciplina, que los usen en su trabajo, evalúen los resultados de otros arqueólogos y que vayan mucho más allá de los límites de este libro y de su autor en el desarrollo de nuevos métodos y modelos.

STEPHEN SHENNAN

PREFACIO

Tal y como se explica en el primer capítulo de este libro, en los últimos años, los métodos cuantitativos han alcanzado una gran importancia en arqueología. Se ha logrado un considerable progreso en el grado de complejidad de los análisis y en la armonización entre los datos arqueológicos y los métodos matemáticos. Ahora bien, los especialistas de un dominio científico que exige el uso de las matemáticas suelen trabajar a un nivel que los no especialistas son incapaces de seguir. Es imprescindible para el avance real de la disciplina que los que la ejercen tengan algún conocimiento acerca de sus implicaciones. Actualmente, en arqueología hay un pequeño grupo de especialistas muy competentes en matemáticas, pero la mayoría de la gente que practica la arqueología ni siquiera entiende de lo que los otros están hablando, no digamos ya de sus métodos y técnicas. Esta es una situación peligrosa, porque conduce, por un lado, al desprecio de unos métodos que pueden ser útiles y, por el otro, a una excesiva credulidad con respecto a las afirmaciones expresadas por los análisis matemáticos. Si no se pueden evaluar los argumentos, ¿cómo será posible pronunciar un juicio correcto? Esta situación contrasta con la que se da en geografía, donde los especialistas matemáticos y estadísticos configuran el vértice de una pirámide de amplia base de practicantes, con un conocimiento básico en matemáticas y estadística.

La diferencia radica en el proceso educativo de las dos disciplinas. Los métodos cuantitativos se integraron pronto en la educación de los geógrafos, como resultado de la revolución de la geografía cuantitativa de los años cincuenta y sesenta. Los desarrollos comparables en arqueología nunca se incorporaron en la misma medida a la educación de los arqueólogos, por razones que, sospecho, tienen más que ver con la sociología de la disciplina, especialmente en las universidades, que con cualquier otra cosa. Así, mientras que los manuales de estadística y matemáticas han proliferado en geografía, hay que reconocer que no ha sucedido nada parecido en arqueología.

Este libro pretende llenar un hueco entre *Matemáticas para arqueólogos* (Orton, 1980), de un lado, y *Mathematics and Computers in Archaeology* (Doran y Hodson, 1975), de otro. El libro de Orton es muy claro, una excelente exposición de la manera que los métodos cuantitativos pueden llegar a ser útiles para

los arqueólogos; sin embargo, no se trata de un manual. El libro de Doran y Hodson es mucho más avanzado que el de Orton. Su tratamiento altamente esquemático de los métodos elementales, aunque modelo de elegante concisión, lo hace demasiado denso para lectores que aún no comprenden la materia.

Este libro procede de un curso sobre métodos cuantitativos de análisis de datos que he impartido durante algunos años en el Departamento de Arqueología de la Universidad de Southampton. Con el paso del tiempo, el curso fue cambiando con respecto a su primera versión, en respuesta, sobre todo, a los intereses de los alumnos y a la evolución de la disciplina (la estadística es una disciplina cambiante, un hecho que los que no trabajan en ella no suelen apreciar). Llegados a este punto, se hacen imprescindibles algunos comentarios acerca del contenido del libro.

En primer lugar, trata sobre el análisis de datos en *arqueología* y por eso está orientado al estudio de objetos, excavaciones y datos arqueológicos procedentes de los informes y memorias, antes que a los estudios basados en el trabajo de laboratorio, tales como propiedades del suelo, análisis químicos, etc.; en general, esos estudios tienen su propia tradición matemática y cuantitativa, derivada de la disciplina en la que esas técnicas se originaron.

Esta orientación ha influido indudablemente en la selección del material presentado aquí. Por eso, el siguiente punto acerca del contenido del libro es precisamente que no se trata de una obra completa, en el sentido de cubrir todas las técnicas cuantitativas que han sido usadas o pueden llegar a ser útiles a los arqueólogos. Para ello, el libro tendría que haber sido mucho más largo. He pretendido trabajar a distintos niveles: en primer lugar, cómo se traduce un problema arqueológico en términos estadísticos y las cuestiones asociadas a esa traducción; en segundo lugar, proporcionar una base técnica a los métodos más importantes. En un manual como este, es natural que la segunda cuestión ocupe la mayoría del espacio disponible.

El propósito de la primera parte es mostrar a los estudiantes cómo llevar a cabo por sí mismos algunas de las técnicas más básicas; el objetivo de la segunda es proporcionar una comprensión intuitiva de algunos de los métodos más complejos, basados en un enfoque geométrico, para poder entender la bibliografía. La familiaridad con el material arqueológico presentado aquí como ejemplo hará mucho más fácil entender la bibliografía estadística escrita para otros públicos, como, por ejemplo, geógrafos y sociólogos.

En el nivel estadístico básico, la omisión más evidente es la prueba de t, así como la falta de un estudio de la teoría distribucional, a excepción del muy esquemático de los capítulos 8 y 14. Esto no se ha hecho sin una considerable reflexión previa. De hecho, ese material estaba cubierto ampliamente en las primeras versiones del curso que originó este libro. Si se eliminó fue porque la cantidad de detalles técnicos complejos que habían de tratarse demostró ser un obstáculo mayor para la comprensión de los métodos cuantitativos en sí,

y sin relevancia directa para la arqueología. El no incluirlos fue mucho más satisfactorio: la ganancia superó la pérdida.

Otra omisión obvia es el análisis espacial. El libro de Hodder y Orton (1976) aún sirve como una buena introducción para los arqueólogos, apoyada por muchos textos geográficos. El uso de un material mucho más elemental que el que aparece en el trabajo de estos autores sobre análisis espacial sólo conseguiría un aumento considerable del tamaño del libro, sin que su utilidad se viese incrementada.

El carácter introductorio del libro excluye un examen de las técnicas avanzadas usadas actualmente en la investigación arqueológica, como aquellas basadas en la simulación por ordenador, si bien hay un breve comentario sobre ellas en el último capítulo. Una o dos secciones del libro, sin embargo, son bastante más difíciles que el resto. El lector quizás prefiera dejarlas de lado en una primera lectura. Especialmente en el caso de la última sección del capítulo 7 y, en menor grado, partes de los capítulos 10 y 11. Han sido incluidas para mostrar los métodos del análisis de datos y cómo pueden realizarse. Son ejemplos, como hay otros en el libro, de un enfoque escéptico de la subdisciplina: si algo me ha guiado, no es que los lectores memoricen los detalles de las técnicas estadísticas, sino que adquieran una actitud informada, escéptica e interrogante acerca de los análisis cuantitativos que ellos mismos u otros investigadores hayan emprendido.

Durante los años en que he enseñado métodos cuantitativos y preparado este libro, he adquirido numerosas deudas que deben ser reconocidas. En primer lugar, y lo más importante, a aquellos alumnos que siguieron mis clases, especialmente Todd Whitelaw, Hans-Peter Wotzka y Nick Winder. Sus preguntas y críticas no me permitieron nunca una salida por la tangente, ¡y yo me he beneficiado mucho de ello! Agradezco también al profesor Colin Renfrew por sus ánimos iniciales y apoyo a la enseñanza de los métodos cuantitativos en Southampton, y a Archie Turnbull, de Edinburgh University Press, por animarme a escribir este libro y ofrecerme muchos comentarios críticos y constructivos respecto a un borrador inicial. Finalmente, tengo una deuda especial de gratitud con el doctor Nick Fieller, del Departamento de Probabilidad y Estadística de la Universidad de Sheffield, por su valiosa y experta ayuda en la lectura del manuscrito, y con el profesor R. Barry Lewis, del Departamento de Antropología de la Universidad de Illinois en Urbana-Champaign, quien hizo numerosos comentarios y sugerencias, útiles y clarividentes. Ni ellos ni nadie más, excepto yo mismo, son responsables de los errores que puedan quedar.

1. INTRODUCCIÓN

El propósito de este texto es familiarizar a los estudiantes con algunos de los métodos cuantitativos básicos usados habitualmente en arqueología. Naturalmente, esas técnicas no son exclusivas de la arqueología, pues se usan en otros muchos campos, pero la experiencia enseña que los estudiantes de arqueología no sacan mucho provecho siguiendo las clases de estadística para sociólogos y biólogos, porque, aunque la teoría estadística sea la misma, los ejemplos usados les son extraños. A un estudiante de arqueología, esos ejemplos le resultan aburridos y a menudo incomprensibles. Para la mayoría de la gente, los métodos cuantitativos ya son lo suficientemente prohibitivos como para necesitar de dificultades de ese tipo. Enseñar en un entorno ajeno es particularmente desafortunado, porque a menudo los que no sienten una inclinación especial por las matemáticas encuentran que es mejor seguir un caso práctico para obtener una visión inicial de una materia, que aprender la teoría que la fundamenta. Por esas razones, un texto introductorio específicamente arqueológico parecía apropiado.

Espero que, al finalizar la lectura del libro, los estudiantes sean capaces de usar por sí mismos las técnicas sencillas que se describen aquí, que reflexionen acerca de las cuestiones que pueden ser traducidas en términos cuantitativos, y que cuenten con una base para poder hablar con los estadísticos profesionales en sus propias palabras, si los problemas son más complejos. Este último punto es muy importante. Si un arqueólogo pide ayuda a un estadístico profesional y ninguno de los dos es capaz de entender lo que el otro está diciendo, se obtendrá la solución equivocada al problema equivocado.

El texto exige muy poco conocimiento previo. Sólo son necesarias las operaciones básicas de suma, resta, multiplicación y división, junto con las raíces, las potencias y los logaritmos. No se precisa ni el cálculo integral o diferencial, ni el álgebra matricial.

¿Por qué necesitamos los métodos cuantitativos?

Hay que dar una respuesta a esta pregunta antes de seguir adelante. De hecho, es posible dividirla en dos cuestiones bastante distintas: ¿por qué los estu-

diantes de arqueología han de aprender los métodos cuantitativos? y ¿por qué la arqueología, como disciplina, ha de tratar con métodos cuantitativos?

Una respuesta a la primera de esas cuestiones es que la bibliografía arqueológica está produciendo cada día más artículos cuya argumentación depende de la aplicación de tales métodos. Un conocimiento de los mismos es, por tanto, esencial si se quiere entender y evaluar sus argumentos. Esto es cierto, pero no responde a la segunda, y más amplia, cuestión. La versión más cínica de la respuesta sería que la arqueología se ha visto implicada en los métodos cuantitativos exclusivamente como resultado de una moda en la disciplina. Los últimos treinta años han presenciado la cuantitativización de las ciencias biológicas, la geografía y muchas de las ciencias sociales; es una cuestión de prestigio para una disciplina que quiere ser «científica» el que los métodos cuantitativos interpreten un papel clave. La arqueología se habría limitado a seguir esa tendencia, adoptando la imagen del «arqueólogo en bata blanca», con lo que el proceso se habría ido imponiendo, aumentando su influencia, el número de personas trabajando en ello y absorbiendo recursos. A su debido término, afirma el argumento, esos enfoques pasan de moda —es posible que ya se haya entrado en esa fase—, y se extinguen progresivamente. Creo que no tendría sentido negar ese aspecto de la «revolución cuantitativa», tal y como ha sido denominada en geografía, pero esos argumentos de la sociología de la ciencia sólo constituyen una parte de la historia.

Un factor clave ha sido la proliferación de los ordenadores. Como todos saben, los ordenadores tienen hoy una gran variedad de aplicaciones en la arqueología. En la última década se han usado cada vez más como útiles de tratamiento de datos para tareas como el registro de los datos de excavación y la construcción de bases de datos regionales de información arqueológica; el primer desarrollo, en particular, ayudado sobre todo por la irrupción de los microordenadores (Richards y Ryan, 1985; Gaines, 1981). El uso de los ordenadores en la configuración de modelos arqueológicos tambien ha sido importante: se han escrito numerosos programas para simular procesos tan diversos como el colapso de la civilización maya (Hosler *et al.*, 1977), o la manufactura y desecho de útiles en los sistemas de subsistencia-asentamiento de los aborígenes australianos (Aldenderfer, 1981); pueden hallarse numerosos ejemplos más en los libros compilados por Hodder (1978), Renfrew y Cooke (1979) y Sabloff (1981).

Aquí sólo quisiera considerar su uso como herramientas para llevar a cabo el análisis de datos, que es el propósito para el cual los ordenadores fueron introducidos en la arqueología y otras disciplinas. Antes del desarrollo y la primera aplicación de los ordenadores, en los años cincuenta e inicios de los sesenta, el uso de las matemáticas y la estadística estaba restringido a las «ciencias duras», porque en ellas la solución a muchos problemas de interés podía obtenerse por medio de métodos elegantes de análisis matemático que no exigiesen una cantidad enorme de cálculos; igualmente, las técnicas estadísticas que por

la misma razón eran posibles sin ordenador demostraron su utilidad en muchas aplicaciones científicas, tecnológicas e industriales. Esto no sucedió con los datos, más intratables, de la geografía, la arqueología y las ciencias sociales; sólo por medio de un instrumento capaz de realizar un enorme número de cálculos a elevadas velocidades se hizo posible la aplicación de los métodos más apropiados al tipo de problemas que presentan los datos característicos de esas disciplinas.

Ahora bien, hay que reconocer que los arqueólogos empezaron a usar el ordenador como si de un juguete nuevo y excitante se tratara. La relación de la arqueología con los métodos cuantitativos procede de los ensayos realizados por parte de aquellos arqueólogos a los que les gustaron los nuevos juguetes y encontraron un uso para ellos. No deberíamos olvidarlo.

Sin embargo, aún no hemos entrado en el corazón del asunto, que no radica en la moda, ni en la disponibilidad de ordenadores, sino en el hecho de que el razonamiento cuantitativo es fundamental en la arqueología, y que si reconociéramos este hecho mejoraría nuestro trabajo como arqueólogos. El libro de Clive Orton *Matemáticas para arqueólogos* (1980) proporciona una excelente demostración de por qué ese es el caso, proponiendo algunas de las preguntas-tipo que los arqueólogos se plantean, como ¿qué es?, ¿qué antigüedad tiene?, ¿de dónde procede? y ¿para qué servía?, mostrando cómo un enfoque cuantitativo puede ayudar a proporcionar las respuestas. De ello se desprende que los métodos cuantitativos han de ser considerados no como una especialidad científica diferenciada dentro de la arqueología, como el análisis de polen, por ejemplo, o las variadas técnicas de caracterización de artefactos, sino formando parte del arsenal de útiles del arqueólogo. Los especialistas en estadística, matemáticas e informática pueden ser necesarios para tratar con problemas particulares, siempre y cuando el arqueólogo tenga el suficiente conocimiento para reconocer cuándo los problemas pueden ser tratados satisfactoriamente de forma cuantitativa. Nadie más hará esto por él.

Por lo tanto, es preciso especificar claramente en qué punto convergen la arqueología y las matemáticas. Parte de la respuesta está en la descripción simple del registro arqueológico: la *cantidad* de fragmentos de distintos tipos, el *tamaño* de las fosas, etc. Esa información cuantitativa es una parte esencial de los trabajos arqueológicos modernos, por lo que la descripción cuantitativa simple es lo primero que trataremos.

Mucho más importante, sin embargo, es la conexión descrita por Orton (1980). El arqueólogo hace inferencias acerca del pasado, basándose en la estructuración y las relaciones en el registro arqueológico. Las matemáticas constituyen un sistema abstracto de relaciones; existe, pues, la posibilidad de que las matemáticas nos ayuden en la tarea de reconocer un esquema en el registro arqueológico y de especificar su naturaleza. La estadística es, precisamente, el área de las matemáticas en la que estas intentan poner orden en la confusión aparente del mundo real. Eso es lo que hace de la estadística una materia difí-

cil, pues implica consideraciones tanto matemáticas como factuales, y porque las relaciones que observamos casi nunca son perfectas.

Orton muestra claramente que toda interpretación del registro arqueológico precisa de la identificación de una regularidad, y que por tanto es capaz de beneficiarse de un enfoque cuantitativo. No obstante, es un hecho histórico el que el principal esfuerzo para la introducción de los métodos cuantitativos en el análisis de datos arqueológicos procedió de la tradición norteamericana de la «Nueva Arqueología» de la década de los sesenta, lo que trajo como resultado el que esa Nueva Arqueología y la utilización de los métodos cuantitativos estuviesen estrechamente asociadas en la consciencia arqueológica general, ambas etiquetadas como «antihumanistas» (Hawkes, 1968). Doran y Hodson (1975) se esmeraron en señalar, correctamente, que no había una conexión necesaria entre ambas, y que los enfoques cuantitativos podían ser empleados para resolver problemas arqueológicos tradicionales. No obstante, es todavía la tradición de la Nueva Arqueología, hoy día conocida como «escuela procesualista», la que ha hecho un mayor uso de tales técnicas; por lo que sería mejor preguntarse por qué los análisis cuantitativos han sido, y son, uno de sus rasgos distintivos, a pesar del hecho de que algunas aplicaciones están consideradas hoy en día ejemplos clásicos de errores y malas interpretaciones (Thomas, 1978).

Creo que hay varias razones para ello. En primer lugar, lo menos encomiable y quizás también lo menos importante: los métodos cuantitativos son considerados «científicos» y la Nueva Arqueología, específicamente, pretende adoptar un enfoque científico, convirtiendo el uso de los métodos cuantitativos en ideológicamente necesarios. En segundo lugar, la Nueva Arqueología enfatizó la objetividad y la claridad, las cuales están eficazmente apoyadas por el rigor de los análisis cuantitativos, y su función vital de eliminar algunas de las fuentes del autoengaño. En tercer lugar, defendió un enfoque hipotético-deductivo para el estudio del pasado, en el que las hipótesis se generaban a partir de una base teórica, de la cual se deducían las implicaciones arqueológicas, y éstas se comparaban al registro arqueológico por la bondad de su ajuste. Fuesen cuales fuesen los valores o defectos del enfoque —y ello ha sido sujeto de un debate considerable—, queda el hecho de que la bondad del ajuste entre hipótesis y datos es una de las principales tareas que conciernen a la estadística.

Finalmente, y como punto fundamental, la Nueva Arqueología adoptó una visión sistémica del pasado. Rechazó la visión según la cual mientras que las diferencias espaciales en el registro arqueológico procedían de aquellas normas de la población que variaban espacialmente, los cambios a lo largo del tiempo eran el resultado del cambio de las normas por difusión o sustitución de un pueblo por otro. La Nueva Arqueología afirmaba que lo que imperaba era el contexto adaptativo —cómo se relacionaba la gente con el entorno y con otras poblaciones—. Dentro de este esquema, la investigación de las relaciones entre las variables que habían sido medidas en el registro arqueológico se hizo importantísima; la única forma de estudiarlas era cuantitativamente.

En resumen, dado que la escuela procesualista se ha visto envuelta, mucho más que otras escuelas, en la formulación de hipótesis acerca de las relaciones entre los rasgos del registro arqueológico, fue inevitable que dependieran cada vez con mayor frecuencia de los análisis cuantitativos. La demostración de los inconvenientes del enfoque normativo y la importancia del estudio de las relaciones sistémicas han supuesto un avance mayor, con unas consecuencias cuantitativas que han de ser tenidas en cuenta.

Naturalmente, en los últimos veinticinco años la escuela procesualista ha cambiado considerablemente. Incluso ha sido duramente atacada, particularmente en Europa, por críticos que argumentan que muchos de sus presupuestos básicos no son válidos (véase Hodder, 1982). Es posible que estemos en una época en la que el uso de los métodos cuantitativos esté pasando de moda y perdiendo importancia. Ciertamente, han habido cambios en el campo cuantitativo. Hoy hay un menor énfasis de lo que solía ser habitual en la contrastación de las hipótesis estadísticas en arqueología, una situación que se refleja claramente en este libro. Además, se ha desarrollado una mayor conciencia de los problemas interpretativos del registro arqueológico, como evidencia de la conducta en el pasado, y con ella un rechazo de la idea optimista según la cual el análisis de datos cuantitativos podría proporcionar de algún modo una comprensión directa del pasado, lo cual no sería posible con enfoques más tradicionales. Sin embargo, no creo que los cambios teóricos que han tenido lugar en la disciplina como un todo produzcan la decadencia del uso de las técnicas cuantitativas, porque el estudio de las relaciones entre fenómenos sigue siendo de importancia fundamental, sea cual sea la orientación teórica adoptada; en muchos casos, incluso, la única forma de investigar la estructura relacional del registro arqueológico es cuantitativamente. Así, por ejemplo, Tilley (1984; y Shanks y Tilley, 1982) hace un amplio uso del análisis multivariante de datos (véase más adelante, capítulos 11 y 12), pero rechaza la base teórica de la arqueología procesualista. La necesidad de los métodos aumenta porque el registro arqueológico se muestra como una masa de material aparentemente muy desorganizada, elocuente en su silencio.

EL LUGAR DE LOS MÉTODOS CUANTITATIVOS EN LA INVESTIGACIÓN ARQUEOLÓGICA

Antes de volver a las técnicas en sí mismas, es mejor decir algo acerca de la situación exacta de los métodos cuantitativos en el procedimiento de la investigación arqueológica. Este análisis generalmente se produce en un momento tardío de ese procedimiento, lo cual puede resultar engañoso. En la fase del diseño de la investigación, el investigador debe decidir no sólo qué hacer, sino cómo hacerlo, incluyendo las formas apropiadas de análisis. Una vez que esas decisiones se han tomado, ellas mismas definen la conducta a seguir durante

el resto del proceso investigador: en ningún momento es esto más importante que a la hora de asegurarse que los datos recogidos y los métodos para su obtención corresponden a las exigencias de las técnicas que uno se propone utilizar, incluyendo los presupuestos teóricos de las mismas. Descubrir los problemas en el momento del análisis es demasiado tarde.

Finalmente, y en lo que lo anterior es obvio, las técnicas usadas tienen un efecto sobre los resultados obtenidos y las conclusiones arqueológicas extraídas de ellos. De hecho, como veremos, la relación entre el método usado y los modelos «descubiertos» puede ser bastante complicada.

La investigación no es un proceso lineal, naturalmente; adopta la forma de un bucle, porque las conclusiones inevitablemente nos devuelven otra vez a la primera fase para diseñar una nueva investigación.

Los ejercicios: un comentario

La forma en la que las técnicas usadas se relacionan, tanto con el diseño inicial de la investigación, como con las conclusiones arqueológicas, es algo de gran importancia —incluso, en términos conceptuales, es más importante que los detalles de las técnicas mismas—. Esas cuestiones irán apareciendo inevitablemente en el texto, pero su objetivo principal es el de familiarizar al lector con las técnicas, de modo que habrá que dedicar más atención a los detalles de cómo se efectúan. La importancia de hacer los ejercicios y problemas no debe ser subestimada. El lector puede pensar que ya ha comprendido todo lo que ha leído, pero descubrirá que si lo ha conseguido es únicamente porque ha intentado resolver los problemas. Sólo por este medio logrará el lector un mayor grado de capacidad numérica. Tal y como ha dicho Colin Renfrew: «los días de lo incontable están contados».

2. LA CUANTIFICACIÓN DE LAS DESCRIPCIONES

Las colecciones de materiales arqueológicos no hablan por sí mismas; es necesario que el arqueólogo especifique los aspectos en los que está interesado, y éstos a su vez estarán determinados por sus objetivos. El proceso de ir de los objetivos a los aspectos relevantes del material a nuestra disposición no es nada fácil. Algunos arqueólogos, Lewis Binford en particular, dirían que en muy pocas ocasiones se ha realizado satisfactoriamente, por lo que muchas reconstrucciones arqueológicas del pasado no son más que ficciones (Binford, 1981).

Consideremos un ejemplo. Supóngase que nos interesa estudiar la estratificación social en un área dada a lo largo del tiempo. Empezaremos observando el registro arqueológico de esa área y decidiremos cuál es el aspecto que mejor muestra los cambios en la estratificación social: creemos que se trata de la variación a lo largo del tiempo en la cantidad del ajuar metálico depositado en las tumbas más ricas del área. Si esas cantidades cambiantes de metal depositadas no estuviesen relacionadas con la riqueza de los individuos, sino con los cambios en la tecnología minera o con los contactos comerciales de esa área, entonces lo que se refleja en el análisis no es la evolución de la estratificación social. Si después de haber argumentado erróneamente que la deposición de metal estaba relacionada con la estratificación social, intentamos explicar las razones del incremento de la estratificación social, complicaríamos aún más las cosas, ¡porque estaríamos explicando algo que nunca ocurrió! Vistos de esta manera, los errores parecen bastante obvios, pero es muy fácil incurrir en ellos en la práctica, por lo que muchas investigaciones actuales han intentado mejorar nuestra comprensión de los procesos que producen el registro arqueológico.

Para el propósito de este libro, este problema será dejado de lado en la mayoría de las ocasiones, asumiendo que se ha seleccionado para la investigación un aspecto del material que es apropiado a nuestros intereses. En la práctica, particularmente en el nivel de la descripción del material recuperado de una excavación, hay un amplio consenso acerca de qué categorías de información han de registrarse. Sin embargo, la dificultad subrayada antes es real, básica para

la arqueología, por lo que es imprescindible que volvamos a ella más adelante.

Una vez definidos los aspectos del material en el que estamos interesados, es preciso hacer un registro del mismo, listo para el análisis. El proceso de asignar un valor o puntuación al material que nos interesa constituye el proceso de medida. Se trata de algo mucho más general que el peso de los objetos en distintas balanzas, o la simple medida de las cosas por medio de un calibrador o pie de rey —las medidas pueden ser de muchas clases—. En un conjunto de cerámicas, por ejemplo, hay muchos aspectos que nos pueden interesar: la altura o el volumen de las vasijas, los motivos decorativos usados en ellas, la pasta, o sus formas. De cada vasija de nuestra colección hemos de registrar la información que nos interesa. El resultado de esta tarea es una gran tabla con las puntuaciones y valores de cada aspecto significativo (tabla 2.1). Los aspectos del material en cuyo estudio estamos interesados suelen denominarse *variables* de interés.

TABLA 2.1. Ejemplo de la información registrada para un grupo de vasijas de cerámica.

	Altura (mm)	Diámetro del borde (mm)	Tipo de la pasta	Tipo del borde	Motivo en 1.ª posición	Motivo en 2.ª posición	...
Vasija 1	139	114	1	1	16	11	...
Vasija 2	143	125	2	1	12	9	...
⋮							
Vasija *n*	154	121	4	3	21	15	...

El proceso de medida, especialmente la codificación de cosas tales como las descripciones de la cerámica, no es tan sencillo como parece, y requiere unas buenas dosis de reflexión (para un mayor análisis sobre el tema véanse Richards y Ryan, 1985; Gardin, 1980). A menudo, esa codificación se lleva a cabo como un preliminar a la introducción de los datos en un archivo del ordenador, antes de emprender análisis más detallados. Es muy importante que los datos estén codificados de forma relevante al análisis que se pretende, de otro modo perderemos mucho tiempo manipulando los datos en los archivos del ordenador para poder otorgarles una forma correcta.

La forma precisa en la que se plantea el problema de la codificación empieza a cambiar hoy, en la medida en que los arqueólogos usan cada vez más programas informáticos de gestión de base de datos para la introducción y tratamiento de sus datos, lo cual se opone a la introducción de los mismos directamente en los programas específicos de análisis de datos, con unas condiciones de formato muy estrictas. Este nuevo desarrollo proporciona un mayor grado de flexibilidad, pero no elimina el problema de la descripción de los datos, tal y como enfatiza Gardin (1980).

Se han planteado otras cuestiones acerca del proceso de codificación en sí

mismo: es importante evitar ambigüedades e inconsistencias lógicas. Codificar la decoración de la cerámica puede ser especialmente difícil, dado que implica la toma de decisiones acerca de las unidades básicas del esquema decorativo, cuáles de esas unidades son simples variaciones dentro de una estructura fundamental, etc. El tema ha sido bien abordado por Plog (1980).

Una cuestión general que a menudo aparece es qué incluir y qué omitir en una descripción, incluso cuando se conocen los objetivos del estudio. Por ejemplo, si estudiamos un cementerio con enterramientos de inhumación que contienen ajuar, e intentamos hacer inferencias acerca de la organización social de la comunidad que usó esas tumbas, ¿se incluirá información acerca de la naturaleza y posición de cada objeto del ajuar en la tumba? ¿O quizás sea significativa la ubicación exacta de las extremidades del esqueleto? La respuesta habitual es pecar por exceso en la inclusión, antes que en la omisión; en un estudio muy amplio, eso puede implicar una enorme cantidad de trabajo no particularmente relevante, que puede llegar a costar mucho dinero, especialmente si se trata de trabajo de campo. Una solución sería realizar un estudio piloto: un análisis preliminar de una pequeña parte de los datos, usando la descripción completa, así como cualquier variable que nos parezca que no ha sido incluida en el registro general de los datos. No resulta exagerado afirmar que las decisiones tomadas en la fase de la codificación pueden tener un efecto muy importante en el resultado de los análisis subsiguientes.

Una vez que hemos construido la tabla de datos, toda la información está allí incluida, pero no es fácilmente accesible. No solemos estar interesados en las características de cada individuo en particular, sino en el conjunto del material como un todo, de forma que cuando preguntamos cuán parecidas son las distintas pastas de esa cerámica o bien si tienen las vasijas un tamaño estandarizado, las respuestas no se pueden extraer directamente de la tabla. Necesitamos resumir nuestros datos (los valores de las variables) de alguna forma. Los gráficos constituyen una de las mejores maneras, pero para que sean apropiados hemos de considerar primero las características métricas de nuestras variables, lo que se conoce como *niveles de medida*. ¿Cuáles son esos niveles o escalas? En orden de potencia matemática: *nominales, ordinales, de intervalo* y *proporcionales*. Para empezar con el menor, la escala nominal se llama así porque sólo incluye los nombres de las distintas categorías. Uno puede pensar que eso no es ninguna medida, sino un proceso de *clasificación*: situar cosas en grupos o categorías, el primer paso en cualquier investigación. Supongamos que estamos estudiando la cerámica funeraria de la edad del bronce británica, y que dividimos los recipientes en urnas con cuello, urnas globulares, urnas en forma de barril y urnas en forma de cubo. Esto representaría una escala nominal, apropiada para ese particular conjunto de cerámicas, en el que hay cuatro categorías. En ese caso, el proceso de medida consistirá en la asignación de una de esas categorías o valores a cada uno de nuestros recipientes. No hay ninguna

ordenación inherente entre los recipientes implicados en esa categorización. Podríamos asignar números a las categorías, así:

1. urna con cuello;
2. urna globular;
3. urna en forma de barril;
4. urna en forma de cubo.

Si lo hacemos, estaremos usando los números simplemente como símbolos que nos resultan convenientes por alguna razón —quizás como una notación simplificada—. No tendría sentido sumar o multiplicar esos números.

Si es posible dar un orden a todas las categorías de acuerdo con algún criterio, entonces se habrá utilizado el nivel ordinal de medida. Así, si dividimos una colección de cerámicas en cerámica fina, cotidiana y grosera, podríamos decir que se trata de una escala ordinal con respecto a cierta noción de calidad. Podríamos disponer la cerámica fina en el lugar 1, la doméstica en el 2, y la grosera en el 3. Igualmente, la conocida clasificación de las sociedades en bandas, tribus, jefaturas y estados (Service, 1962) es una ordenación de las sociedades con respecto a una idea de complejidad de la organización. Cada categoría tiene una posición única en relación con las otras. Si sabemos que la jefatura es mayor que la tribu y que el estado es mayor que la jefatura, automáticamente se desprende que el estado es mayor que la tribu. Por otro lado, no sabemos *en qué grado es* menor la jefatura con respecto al estado, o la tribu con respecto a la jefatura, sólo conocemos el orden —es menor—. Esa propiedad de ordenación es la única propiedad matemática de las escalas ordinales.

En contraste con las escalas ordinales, en donde sólo se define la ordenación de las categorías, en las escalas de intervalo y proporcional las distancias entre las categorías están definidas en unidades fijas e iguales. La diferencia entre las dos, sin embargo, es menos obvia que en las otras que hemos ido viendo, por lo que será mejor ilustrarla con un ejemplo. ¿Es la medida del tiempo en años a.C. o d.C. una escala de intervalo o proporcional? Ciertamente, se trata de algo más que una simple escala ordinal, porque el tiempo está dividido en unidades fijas e iguales —los años—. La distinción entre las dos depende de la definición del punto cero —sea o no arbitrario—. El hecho de definir una cronología en años a.C. o d.C. es una convención arbitraria. Existen otros sistemas cronológicos perfectamente válidos, con distintos puntos de partida, por ejemplo, los sistemas hebraico o islámico. Si, por otro lado, consideramos las dimensiones físicas —distancias, volúmenes o pesos—, entonces el punto cero no es arbitrario; por ejemplo, si medimos la distancia, sean cuales sean las medidas que usemos, la distancia cero estará definida con propiedad: se trata de la ausencia de distancia entre dos puntos; la proporción entre 100 mm y 200 mm es la misma que entre 3,94 pulgadas y 7,88 pulgadas, esto es 1:2. Todo esto no resulta válido para nuestros sistemas cronológicos: la proporción entre el

año 1000 d.C. y el 2000 d.C. es 1:2, pero si tomamos los años correspondientes a la cronología islámica, 378 y 1378, la proporción es 1:3,65. La cronología, por tanto, es un ejemplo de escala de intervalo, mientras que las dimensiones físicas son ejemplos de escalas proporcionales. En la práctica arqueológica, una vez que hemos superado el nivel de las escalas ordinales, usualmente se emplean variables de escala proporcional, como por ejemplo algunas de las dimensiones físicas antes mencionadas, así como las frecuencias de aparición de los elementos.

Si existen estas distinciones es porque afectan a las técnicas estadísticas, tanto en el caso de los complejos análisis multivariantes, como si nos limitamos a dibujar gráficos. En los capítulos siguientes, a medida que se vayan presentando las técnicas, una de las principales consideraciones será siempre el nivel de medida entre los datos más apropiado a los métodos. Es muy fácil aplicar métodos inapropiados cuando se usa un ordenador, ya que éste utilizará los números por su simple valor y no preguntará qué es lo que se pretende con ellos.

Pasando de un nivel de medida a otro

Hasta aquí se ha insistido en las distinciones entre los varios niveles de medida; puede que sea conveniente, sin embargo, acabar mencionando las posibilidades de cambiar un nivel por otro, ya que la escala de medida de una propiedad particular en un conjunto de datos no es inmutable necesariamente.

Volvamos a nuestro ejemplo de dividir un conjunto de cerámicas en fina, de uso cotidiano y grosera, una escala ordinal basada en la idea de finura o calidad. En principio, no hay razón por la que no podamos cuantificar la finura de la pasta, por ejemplo, en términos de la media del tamaño de los granos del desgrasante, o la proporción de inclusiones en la arcilla. Tendríamos entonces una escala proporcional de medida de la finura y podríamos situar cada fragmento o vasija en un punto específico de la línea entre la cerámica fina y la grosera, medida en unidades fijas e iguales.

Algunos quieren ver en el nivel de medida predominante en una disciplina un criterio para su sofisticación científica. Así, una disciplina en la que la mayoría de las variables suelen medirse en la escala proporcional estaría más avanzada que otra en la que la mayoría de las variables estuviese expresada en una escala nominal. Tanto si lo aceptamos como si no, es cierto que las variables de escala proporcional, como en nuestro ejemplo de la pasta de la cerámica, contienen más información acerca de la propiedad en cuestión que una escala ordinal.

No hay razón alguna, en principio, que nos impida dar la vuelta al proceso. Empezando con medidas del tamaño de los granos en la pasta de la cerámica, por ejemplo, es posible categorizarla como fina, cotidiana y grosera. Si hacemos esto, sin embargo, estaremos perdiendo información, lo cual no siempre

es recomendable. El argumento no es tan simple como parece, y la controversia se ha visto reflejada en la bibliografía arqueológica con discusiones acerca de si es posible categorizar las variables de escala proporcional (véanse las contribuciones en Whallon y Brown [1982], particularmente la de Hodson y la de Spaulding).

La mejor opción será siempre hacer uso del nivel de medida que pueda proporcionar, con menos esfuerzo, una respuesta concreta a la cuestión que investigamos. Por volver a referirnos a la cerámica, si nuestras investigaciones requieren simplemente una distinción entre cerámica fina, cotidiana y grosera, será una pérdida de tiempo y dinero producir una descripción cuantitativa detallada de la pasta de cada una de las vasijas. Sin embargo, es posible que pretendamos analizar algunas muestras de cada tipo de pasta para demostrar a otros que nuestra distinción no era puramente subjetiva.

Ejercicios

2.1. Observa la serie de vasijas decoradas del neolítico alemán en la figura 2.1, p. 30 (según Schoknecht, 1980), y sugiere un sistema de codificación que, a tu parecer, proporcione la base para una adecuada descripción de las mismas. Codifica cada vasija usando tu sistema. Escala 3:16. ¿Qué problemas han aparecido en la codificación, si es que los hay?

2.2. Intenta el mismo ejercicio con el conjunto de ilustraciones de los planos de tumbas y sus contenidos respectivos de un cementerio del neolítico tardío en Checoslovaquia, que aparecen en las figuras 2.2 a 2.7, pp. 31-36 (según Buchvaldek y Koutecky, 1970). El contenido de las tumbas está también enumerado, dado que la naturaleza de los objetos no siempre es clara a partir de los dibujos. Escala = planos, 1:27; cerámica y piedra de afilar 1:4; otros elementos 1:2.

Tumba 1	1. Urna
	2. Vaso decorado
	3. Hacha plana
	4. Lámina de sílex
	5. Piedra de afilar
Tumba 2	1. Fragmentos de un vaso
	2. Vaso decorado
Tumba 3	1. Vaso decorado, con asa
	2. Urna decorada
	3. Lámina de sílex
	4. Fragmento de una espiral de cobre

Tumba 4	1. Fragmento de una lámina de sílex
	2. Fragmentos posiblemente de dos recipientes
Tumba 5	1. Urna
	2. Urna decorada
	3. Cabeza de maza
Tumba 6	1. Raspador de cuarcita
Tumba 7	1. Urna
	2. Vaso decorado, con asa
	3. Jarra decorada
	4. Vaso cilíndrico, con asa
Tumba 8	1. Urna
	2. Urna decorada
	3. Vaso decorado, con asa
	4. Hacha-martillo
	5. Lámina de sílex
Tumba 9	1, 2. Vasos decorados
	3. Jarra
	4. Vaso decorado
	5. Jarra
	6. Urna decorada
	7. Urna
	8. Lámina de sílex

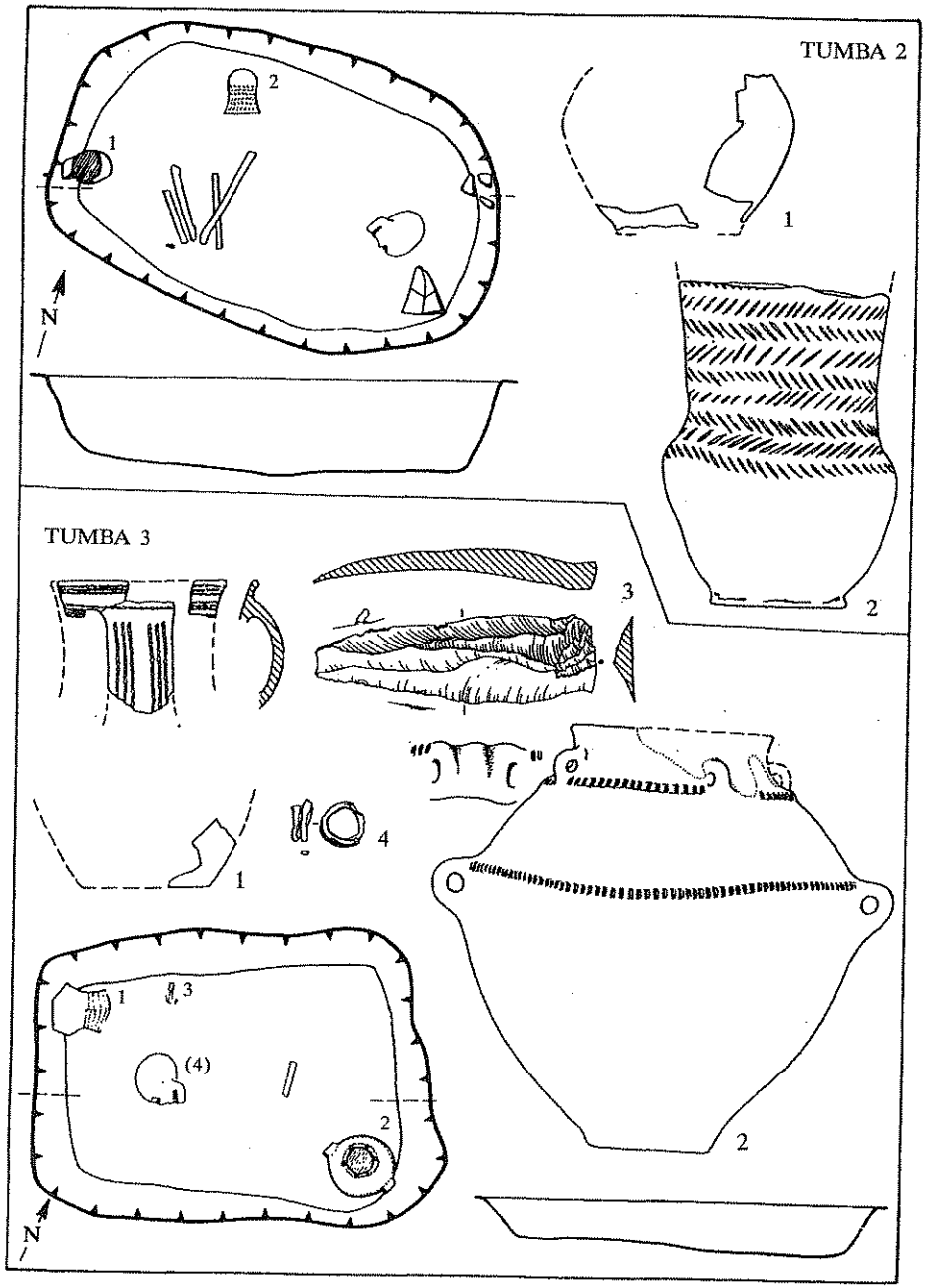


FIGURA 2.3.

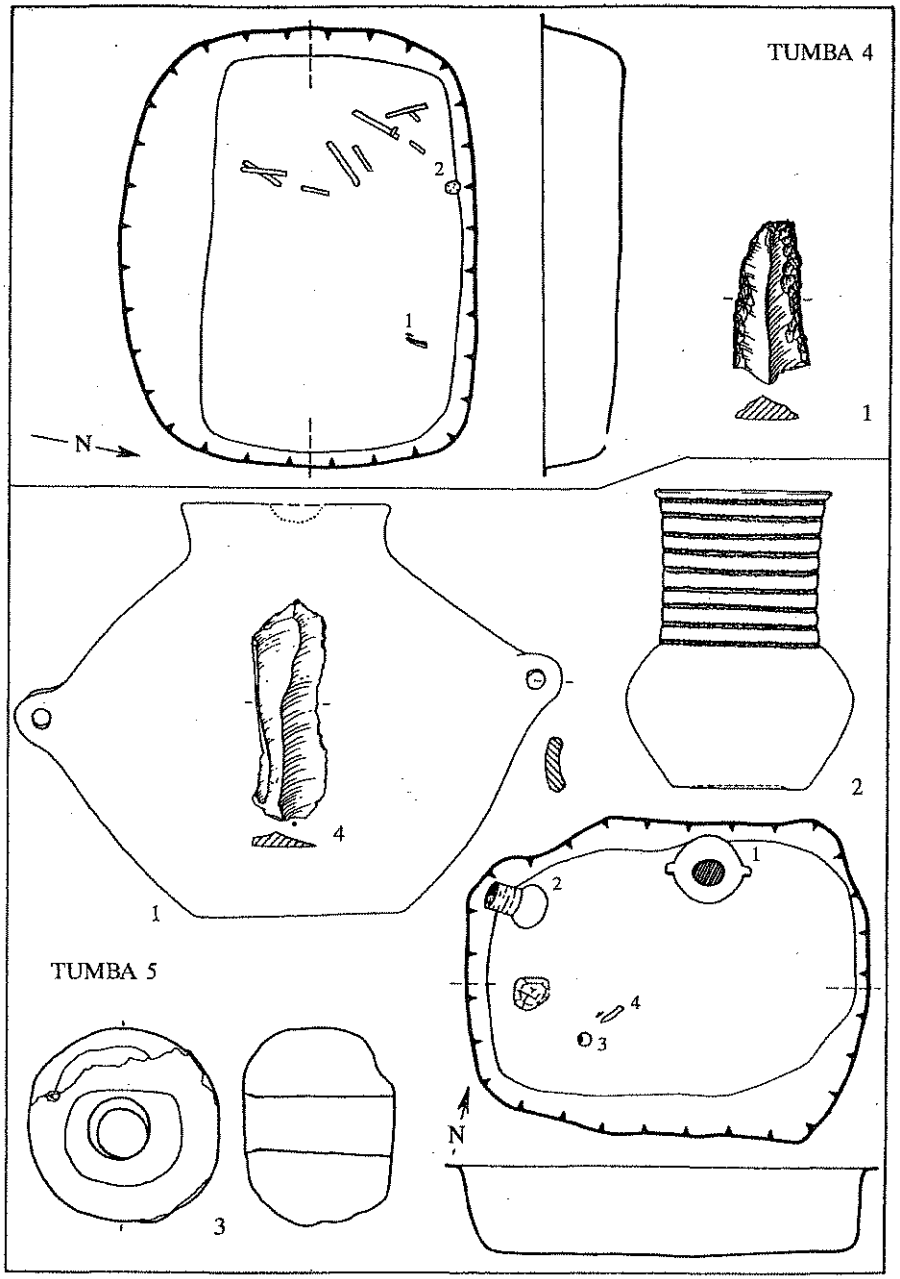


FIGURA 2.4.

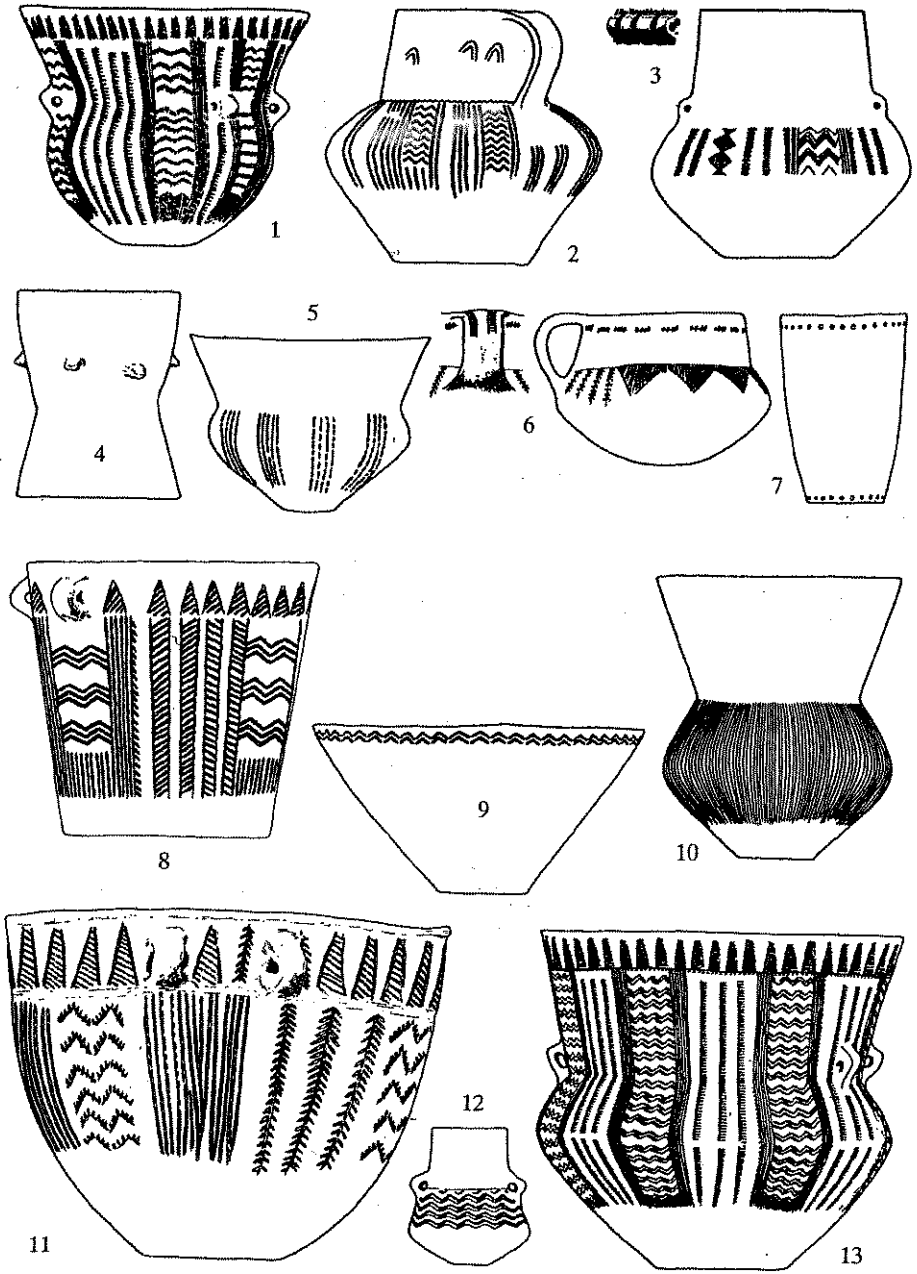


FIGURA 2.1

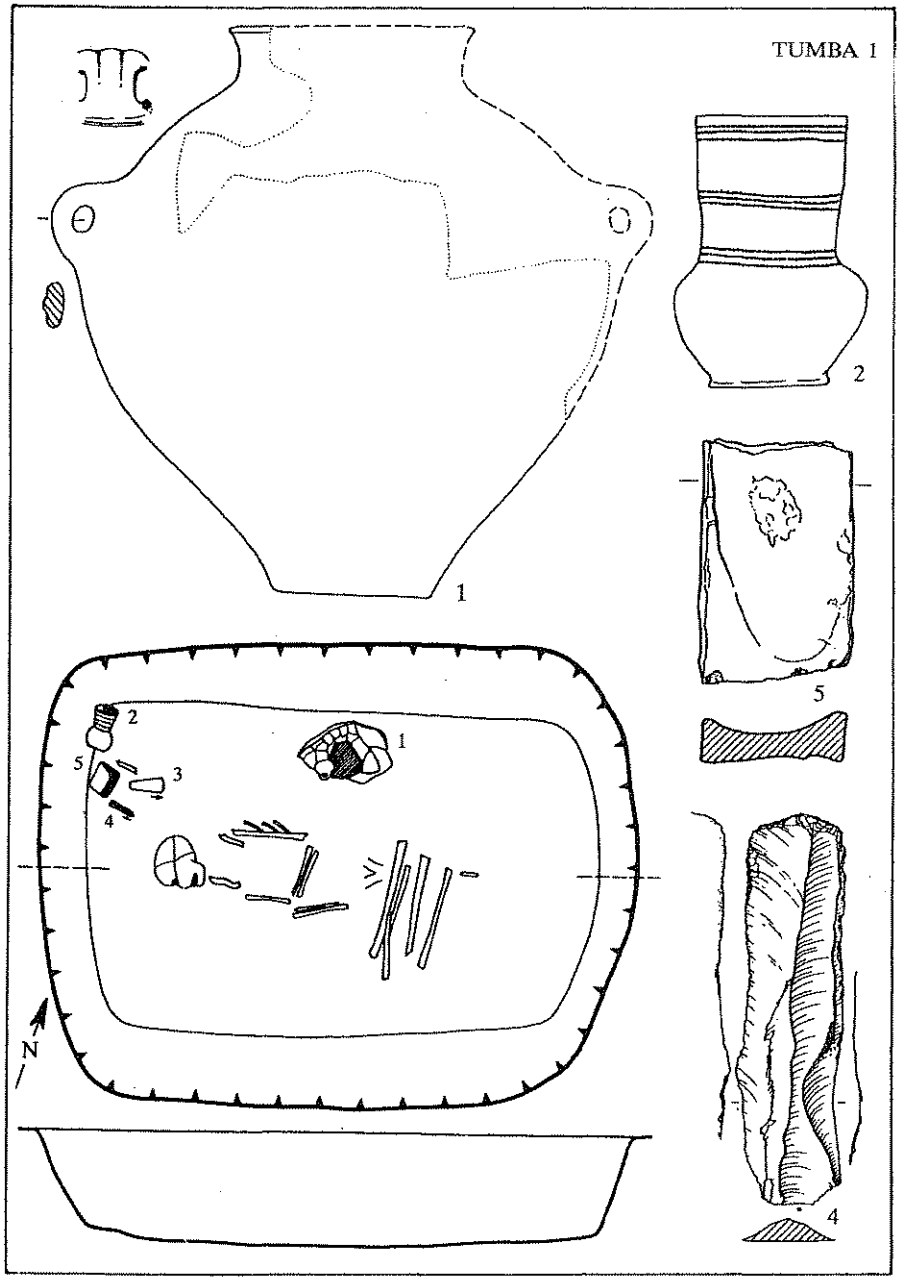


FIGURA 2.2.

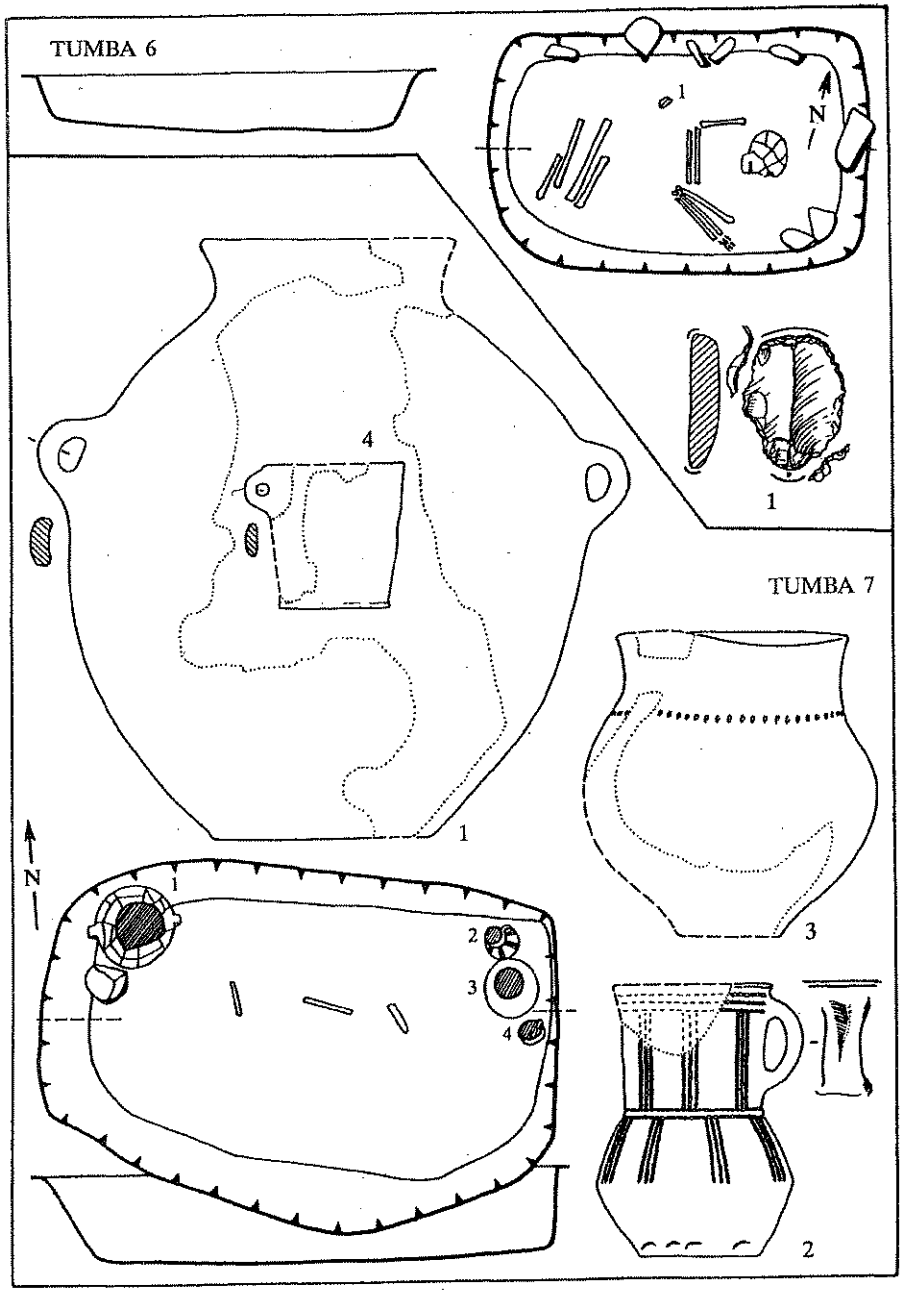


FIGURA 2.5.

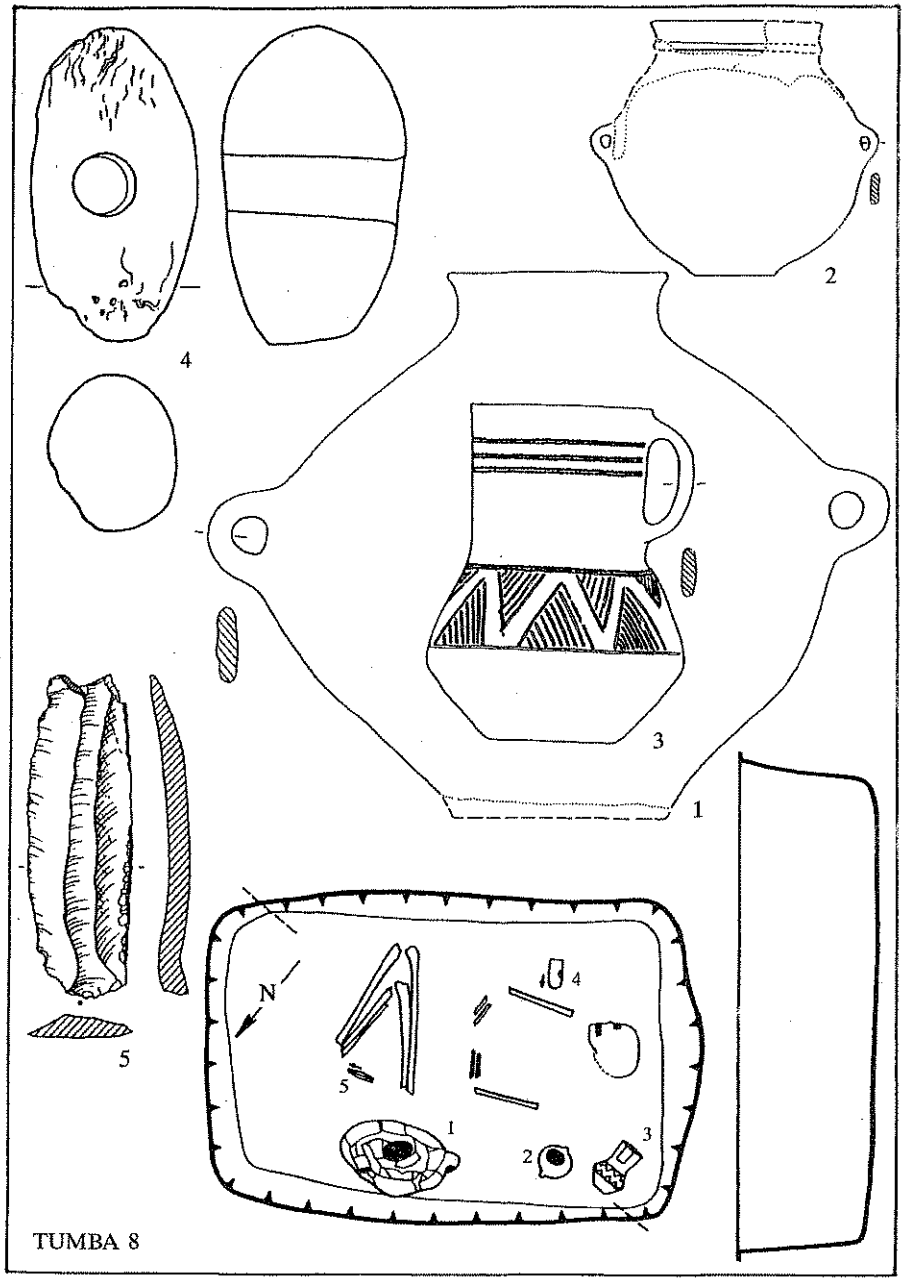


FIGURA 2.6.

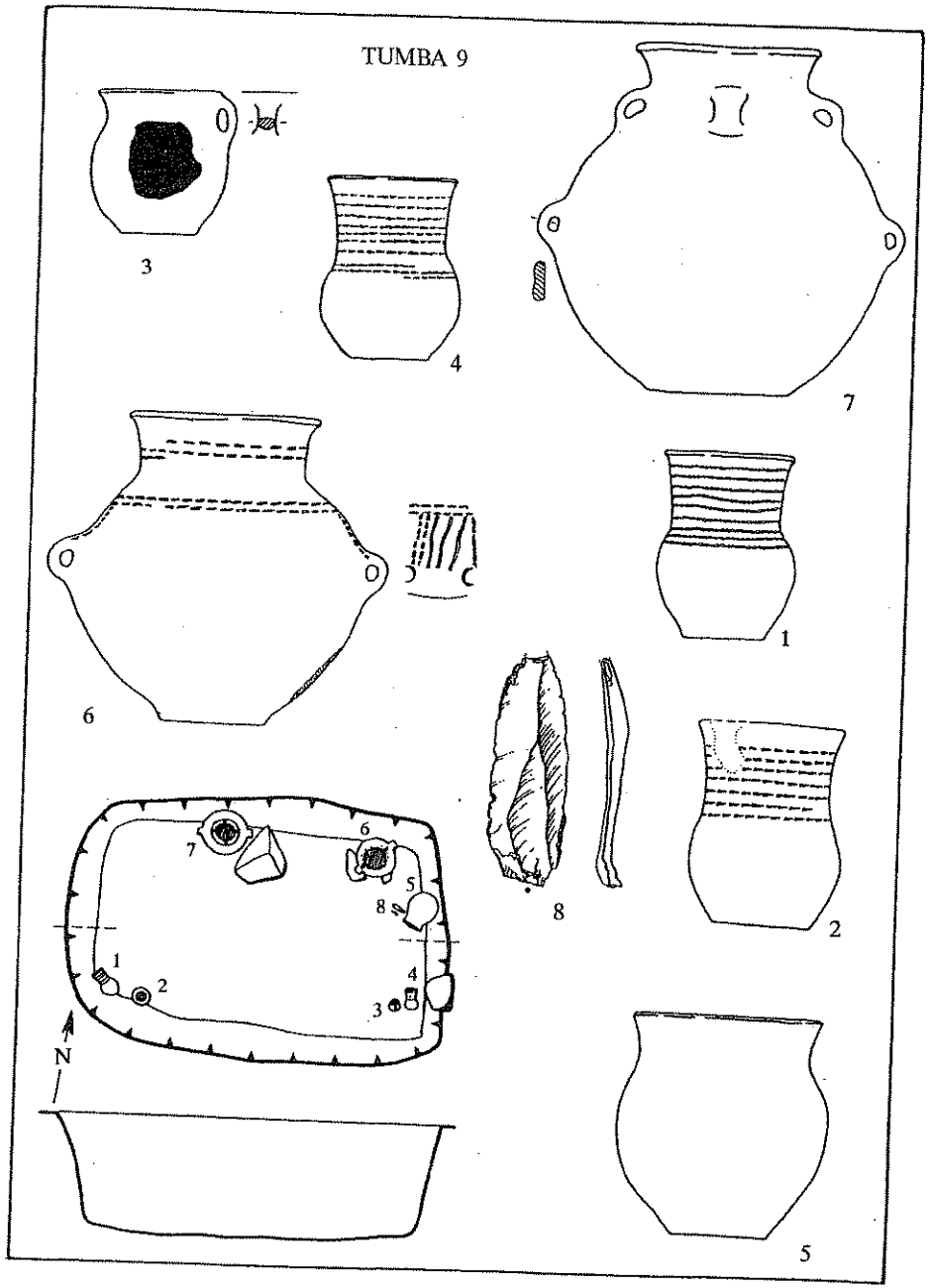


FIGURA 2.7.

3. RESÚMENES GRÁFICOS DE UNA VARIABLE ÚNICA

La simple representación gráfica de los datos en arqueología supone ya una aplicación sustancial de los métodos cuantitativos. De este modo es posible emplear la capacidad del cerebro y del ojo humanos para detectar y afirmar la existencia de estructuras. Algunos métodos muy complejos pueden ser considerados como una manera de obtener la mejor imagen posible de un complicado conjunto de datos. Sin embargo, el ojo humano puede ver esquemas allí donde no hay nada; volveremos más tarde sobre este punto.

El uso de gráficos y cuadros para mostrar la información ha desempeñado siempre un papel importante en la estadística, esencialmente como una fase preliminar al uso de resúmenes numéricos de los datos, los cuales, a su vez, se completan con las inferencias estadísticas (véase más adelante). Recientemente se ha desarrollado un enfoque que presta menos atención a la estadística inferencial; es conocido como *análisis de datos exploratorio*, y se caracteriza por un mayor énfasis en la representación visual de los datos y no en las estadísticas que se derivan de los mismos, así como por un interés muy reducido en las pruebas estadísticas de relevancia (véanse Hartwig y Dearing, 1979; Mosteller y Tukey, 1977; para una discusión arqueológica, Clark, 1982; Lewis, 1986). Su objetivo es explorar el conjunto de los datos a nuestra disposición, definido como relevante para cierto problema, con el fin de buscar algún esquema significativo. La idea fundamental, usando el vocabulario típico del análisis de datos exploratorio (o EDA —*exploratory data analysis*—, tal y como es conocido habitualmente), es la siguiente:

$$\text{Datos} = \text{suave} + \text{basto}$$

En otras palabras, un conjunto de observaciones puede dividirse en dos componentes, un esquema general, lo suave, y las variaciones a partir de ese esquema, lo basto. La tarea del analista de datos es distinguir lo suave de lo basto, poniendo a prueba constantemente lo que está haciendo.

Como Tukey (1980) ha explicado, no se trata de rechazar los métodos tradicionales, como las pruebas de significación (véase el capítulo 5), sino de poner-

las en su sitio, formando parte de aquel bucle que, como veíamos, era el proceso de investigación: la interacción constante entre las ideas y los datos. Tan importante como el papel tradicional de la estadística en la comprobación de las ideas (lo que Tukey denomina *análisis de datos confirmatorio*) es desarrollarlas, ya que a menudo aparecen tras una exploración previa de los datos y no «caídas del cielo». La representación visual de los datos es una buena forma de alcanzar ese objetivo, y no un fin en sí mismo.

Este capítulo va a tratar de los diversos medios visuales de representar las distribuciones de variables únicas, incluyendo los métodos cuyo uso está bien establecido en arqueología, y otro que ha sido muy poco utilizado en nuestra disciplina, el diagrama de tallo y hoja. Sea cual sea la técnica, sin embargo, lo que se pretende es reducir los datos a un orden de algún tipo, de forma que sea posible ver qué es lo que parecen, es decir, obtener una impresión inicial de lo «suave» y de lo «basto». En general, esto implica la presentación de *distribuciones de frecuencia*, en las que las observaciones están agrupadas en un número limitado de categorías.

Probablemente, el más conocido de esos métodos sea el *diagrama de barras*, familiar en nuestra vida cotidiana y cuyo uso en arqueología está establecido desde hace tiempo. Permite distinguir si las categorías están expresadas en una escala nominal, o si hay un orden inherente en las barras. Un ejemplo de lo anterior es la figura 3.1, que resume la cantidad de fragmentos de hueso de diferentes clases, procedentes de un yacimiento hipotético de la edad del hierro británica. No hay ninguna significación particular en el orden de las especies sobre el eje horizontal; podría haberse cambiado por otro cualquiera de las diferentes formas posibles, sin que modificase la información que contiene. El riesgo de caer en el error de interpretar un orden específico leyendo automáticamente el gráfico de izquierda a derecha puede ser reducido usando un *gráfico de sectores*. Requiere la conversión de los números absolutos en proporciones relativas, lo cual representa una ventaja en un sentido, y una pérdida en otro: la idea de la cantidad total se pierde, si bien las proporciones relativas emergen con mayor claridad.

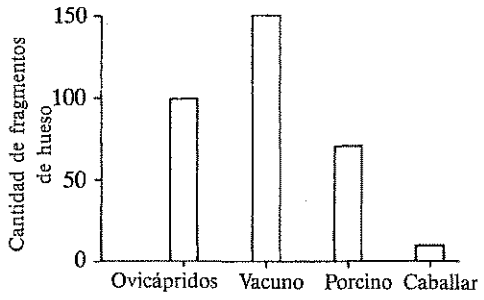


FIGURA 3.1. Gráfico de barras de la cantidad de fragmentos de hueso de distintas especies de animales domésticos, procedentes de un hipotético yacimiento británico de la edad del hierro.

Se considera importante mostrar una indicación del número de casos con los que se trabaja, y no sólo las proporciones; es una buena idea indicarlo al pie del gráfico. Si se representan juntos varios gráficos de sectores, puede darse una impresión de los tamaños relativos de las distintas muestras variando el área de los círculos proporcionalmente al tamaño de la muestra. En el caso del ejemplo de los fragmentos de huesos animales, el gráfico de sectores adoptaría la forma del de la figura 3.2, donde el ángulo del sector apropiado en el centro del círculo es el porcentaje correspondiente, multiplicado por 360/100. Así, si el porcentaje de ganado bovino es del 46 %, se obtendrá $46 \times 360/100 = 166°$.

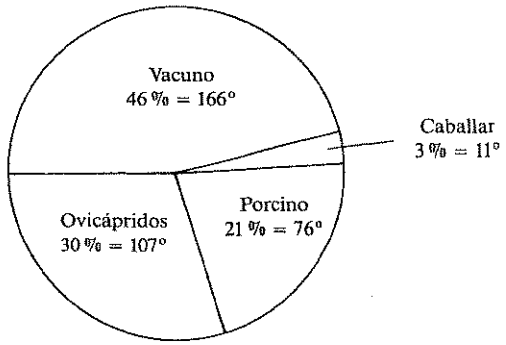


FIGURA 3.2. Gráfico de sectores de las proporciones relativas de fragmentos de hueso de distintas especies de animales domésticos, en el que se usan los datos de la figura 3.1. Cantidad de fragmentos = 330.

El gráfico de sectores es un modo de presentación de los datos muy útil cuando el objetivo es ilustrar las proporciones relativas de las categorías no ordenadas; pero puede confundir si aparecen muchas categorías, hay categorías vacías o con muy pocos elementos y éstas han de ser agrupadas. Algunos autores, sin embargo, presentan objeciones al gráfico de sectores (por ejemplo, Tufte, 1983), diciendo que, para las cantidades de datos relativamente pequeñas que pueden contener, unas tablas que muestren los porcentajes son mucho más útiles.

En una escala ordinal, el orden de las categorías está fijado con respecto a algún criterio, de forma que el orden horizontal de las barras en el gráfico de barras tiene un significado. En un nivel de medida más alto, no sólo es significativo el orden, sino también el intervalo entre las barras; un ejemplo aparece en la figura 3.3, en donde cada barra representa la frecuencia de cada categoría comparada con las categorías adyacentes.

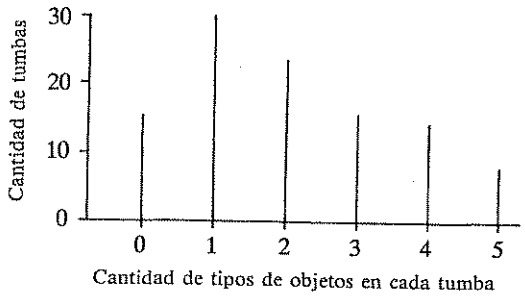


FIGURA 3.3. Gráfico de barras de la cantidad de tumbas que contienen diferentes cantidades de tipos de objetos de ajuar, en un cementerio hipotético de la edad del bronce en Europa central.

Aquí el gráfico de barras resume la cantidad de tumbas en una necrópolis de la edad del bronce que tienen una cierta cantidad de tipos de objetos de ajuar funerario. Tratamos con una escala proporcional —cero significa ausencia de ajuar—, pero la escala tiene una característica peculiar a la que hay que prestar atención: sólo toma en consideración números enteros. Para una tumba, es simplemente imposible contener 3,326 tipos de objetos de ajuar.

Otras escalas de intervalo o proporcionales pueden adoptar cualquier valor, siendo denominadas entonces *escalas numéricas continuas* (también llamadas valores *reales*). Supongamos que se mide, por ejemplo, la altura de unos recipientes, o la longitud de unos huesos; obtendremos resultados como 182,5 mm, 170,1 mm y 153,6 mm. Aunque el conjunto específico de recipientes o de huesos que medimos adopte un conjunto particular de valores, no hay razón teórica que les impida tener cualquier cifra decimal, cuyo número estará determinado exclusivamente por la precisión con que hayamos tomado las medidas.

Si queremos representar la frecuencia de diferentes valores de una misma variable numérica continua, como la altura, la longitud o el peso, habrá que adoptar una estrategia diferente. No podemos tener una categoría separada para 182,5 mm, otra para 170,1 mm y aún otra para 153,6 mm; posiblemente, como mínimo uno de los objetos que nos interesan tendrá exactamente esos valores. Lo que hemos de hacer es dividir nuestra variable en una cierta cantidad de intervalos, cuya anchura ha sido elegida por nosotros mismos, en cada uno de los cuales contaremos la cantidad de observaciones incluidas. Por ejemplo, la figura 3.4 muestra la distribución de frecuencia de las capacidades de una serie de vasos campaniformes. En cada uno de los intervalos se han situado las observaciones que le corresponden. La decisión acerca de cuántos intervalos hay que emplear es arbitraria, pero no ha de ser tomada a la ligera. No nos sirve de nada una distribución con tan pocos intervalos que desaparezca cualquier estructura que pudiera haber. Por otro lado, si tenemos unos intervalos muy

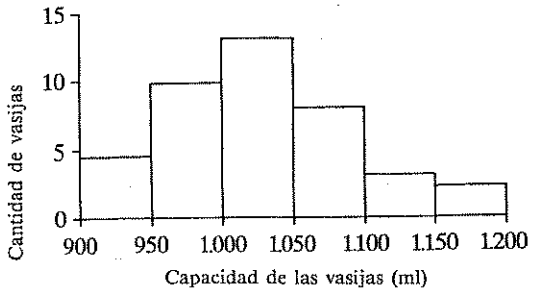


FIGURA 3.4. Gráfico de barras de la distribución de capacidades en un conjunto de 40 vasos campaniformes.

estrechos habrá demasiados huecos y desfases en la figura, lo cual haría difícil distinguir cualquier tipo de tendencia en la distribución, cuando lo que pretendíamos era, precisamente, recuperar tales estructuras relacionales. En general, nunca es bueno tener más de veinte intervalos, porque la imagen que se obtiene es demasiado confusa. Un truco útil que suele producir una representación razonable es calcular un número de intervalos aproximadamente igual a la raíz cuadrada del número de observaciones; así, por ejemplo, si nuestros datos son los volúmenes de 40 recipientes, tendremos que dividir la variable volumen en seis intervalos.

Dado que operamos con variables continuas, es importante especificar claramente los intervalos en el diagrama. En primer lugar, han de ser exhaustivos, es decir, han de incluir todas las observaciones; eso es bastante simple. En segundo lugar, han de ser mutuamente excluyentes. Si la capacidad en uno de los intervalos fuese 900-950 ml y la siguiente 950-1000 ml, sería ambiguo, ya que el valor 950 ml estaría incluido en ambas clases. Hemos de especificar que el rango del primer intervalo es 900-949,9 ml y que el siguiente es 950-999,9 ml, y así sucesivamente.

Otra forma de expresar la información en un diagrama de barras ordenado es la que se denomina *polígono de frecuencias*. La figura 3.5 muestra el mismo ejemplo del ajuar antes empleado (fig. 3.3) en forma de polígono de frecuencias. Esta forma de presentación suele usarse para documentar las transformaciones a lo largo del tiempo, con una escala temporal como eje horizontal y una cantidad en el vertical.

Los métodos presentados constituyen la forma tradicional de representar la distribución de una sola variable en forma de diagrama. El problema con ellas es que la única «verdad» real en un conjunto de observaciones la constituyen las puntuaciones de las observaciones mismas. En cuanto intentamos resumirlas, incluso por medio de una representación como las anteriores, empezamos a perder información (lo cual no es necesariamente perjudicial). Como

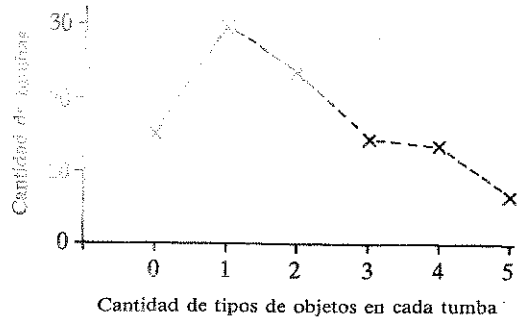


FIGURA 3.5. Polígono de frecuencias de los datos que aparecen en la figura 3.3.

ya hemos visto, a menudo hemos de perder en el detalle para ganar en conocimiento, para ver el bosque y no los árboles. Sin embargo, no hay por qué perder información si uno no quiere hacerlo. Otra dificultad es que no hay una representación «correcta». La forma de un histograma puede variar considerablemente, depende de la anchura de los intervalos y del punto de partida exacto elegido. Por otro lado, una simple lista de los valores de los datos no suele conducir a la detección de las estructuras.

La integración del valor exacto de los datos en un histograma puede conseguirse por medio de un método gráfico conocido como *representación de tallo y hoja*. Podemos ilustrarlo con datos acerca del diámetro de una muestra de 35 hoyos para poste del cercado neolítico de Mount Pleasant, Dorset, Inglaterra (según Wainwright, 1979); están agrupados en la tabla 3.1. Para producir el diagrama, las primeras cifras de cada valor (aquí los diámetros de los hoyos) se separan del resto. Esas primeras cifras se enumeran verticalmente en la parte izquierda del diagrama, formando el tallo (fig. 3.6).

TABLA 3.1. Diámetros (en cm) de 35 hoyos para poste del cercado del neolítico final de Mount Pleasant, Dorset, Inglaterra.

48	57	66	48	50	58	47
48	49	48	47	57	40	50
43	40	44	40	34	42	47
48	53	43	43	25	45	39
38	35	30	38	38	28	27

Las cifras restantes para cada puntuación se sitúan en la fila que corresponde a su primera cifra, en orden creciente, para formar la hoja (fig. 3.7). Esto nos da una imagen que no pierde nada de la información inicial.

Si después de estudiarla consideramos que sería necesario hacer los intervalos más estrechos, podemos reducirlos en cinco unidades, pasando su anchura


```
2 |
3 |
4 |
5 |
6 |
```

FIGURA 3.6. El «tallo» de un diagrama de tallo y hoja (diagrama de los diámetros de los hoyos para poste en Mont Pleasant, enumerados en la tabla 3.1).

```
2 | 5 7 8
3 | 0 4 5 8 8 8 9
4 | 0 0 0 2 3 3 3 4 5 7 7 7 8 8 8 8 8 9
5 | 0 0 3 7 7 8
6 | 6
```

FIGURA 3.7. Diagrama de tallo y hoja de los diámetros de los hoyos para poste en Mount Pleasant, enumerados en la tabla 3.1.

```
2 |
2 | 5 7 8
3 | 0 4
3 | 5 8 8 8 9
4 | 0 0 0 2 3 3 3 4
4 | 5 7 7 7 8 8 8 8 8 9
5 | 0 0 3
5 | 7 7 8
6 |
6 | 6
```

FIGURA 3.8. Diagrama de tallo y hoja de los diámetros de los hoyos para poste en Mount Pleasant, enumerados en la tabla 3.1. Los intervalos en el tallo son de 5 unidades y no de 10.

de diez a cinco; eso es fácil; tenemos simplemente dos filas para cada primera cifra, una para las segundas cifras 0-4, la otra para las segundas cifras 5-9 (fig. 3.8).

Otra utilidad de este tipo de representaciones es que permiten ver fácilmente y con mucha exactitud los valores aberrantes en una distribución y estudiarlos si fuese necesario; de hecho, cualquier peculiaridad de la distribución es fácilmente perceptible.

Muy distinta es la *curva acumulativa*, que no proporciona el mismo tipo de representación, si bien puede darnos, en muchas ocasiones, una imagen de los datos más clara que por medio de otros métodos gráficos, especialmente cuando lo que queremos es comparar un conjunto de datos con otro. En general,

las curvas acumulativas no se basan en las cantidades que hay en las categorías o intervalos, sino en la expresión de esas cantidades como proporción o porcentaje del número total de observaciones. Su funcionamiento puede ilustrarse por medio del ejemplo de los objetos de ajuar funerario. Lo presentaremos primero como una tabla.

TABLA 3.2. Número de tumbas que contienen cantidades diferentes de tipos de objetos de ajuar funerario, procedentes de una hipotética necrópolis de la edad del bronce en la Europa central.

N.° de tipos de objetos	N.° de tumbas	Porcentaje de tumbas
0	17	15,6
1	30	27,5
2	26	23,9
3	17	15,6
4	13	11,9
5	6	5,5
	109	100,0

A continuación se puede trazar un nuevo gráfico, con el eje horizontal indicando, al igual que antes, la cantidad de tipos distintos de objetos en las tumbas, y en el vertical los porcentajes de 0 a 100. Observamos, en primer lugar, que el 15,6 % de las tumbas están agrupadas en la categoría 0-tipos de objetos, lo cual marcamos en el gráfico. A continuación llegamos al 27,5 % de las tumbas en la categoría 1-tipos de objeto, que también sumamos o acumulamos al 15,6 % de la categoría 0; así 15,6 + 27,5 = 43,1 %, que es el valor pará la categoría 1-tipos de objeto, el cual se marca en el gráfico. Este valor nos explica que el 43,1 % de las tumbas tienen un solo tipo de objetos de uso funerario, o menos. Hacemos lo mismo para todas las categorías, hasta que el 100 % de las tumbas haya sido acumulado:

$$\begin{aligned} 43,1 + 23,9 &= 67,0 \\ 67,0 + 15,6 &= 82,6 \\ 82,6 + 11,9 &= 94,5 \\ 94,5 + 5,5 &= 100,0 \end{aligned}$$

Cuando todos los puntos han sido señalados en el gráfico, los unimos con una línea, que es la curva acumulativa representada en la figura 3.9. Muestra la forma de la distribución acumulativa.

Simplemente como medio de representar la forma de la distribución de una única variable, este método puede parecer bastante complicado e innecesario. ¿Por qué no hemos usado el gráfico de barras? La respuesta es que el método

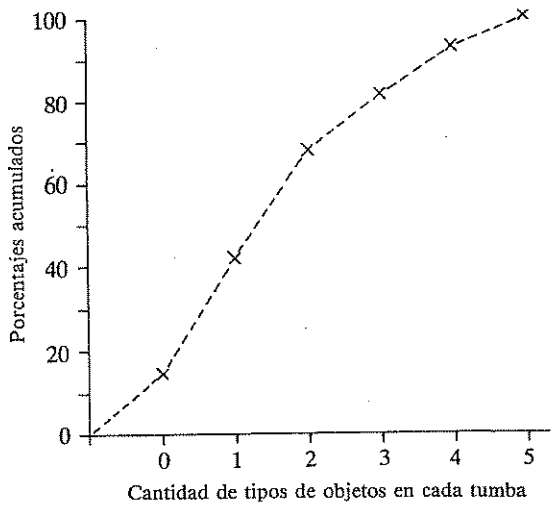


FIGURA 3.9. Curva acumulativa de las cantidades de tipos de objetos de ajuar, presentados en la tabla 3.2.

es efectivo sobre todo a la hora de comparar distribuciones. Los gráficos de barras son bastante difíciles de comparar visualmente, a causa de las diferencias de altura entre las barras. La curva acumulativa, gracias a su progresión continua, permite descubrir con mayor facilidad las semejanzas y las diferencias entre distribuciones.

Esta forma de presentación es significativa únicamente en el caso de que exista un orden real en el eje horizontal, es decir, si tratamos con datos medidos en una escala ordinal, cuando menos. Si el nivel de medida es nominal, cualquier orden será arbitrario, tal y como ya hemos dicho, con lo que la forma de la curva acumulativa será también arbitraria. Siempre sería posible manipular los datos para producir la curva que se pretendiera. Si se adoptase un orden fijo de presentación de las categorías cuando se hacen comparaciones, las curvas acumulativas tendrían utilidad en la presentación de datos nominales, tales como conjuntos paleolíticos descritos en términos de cantidades o porcentajes de tipos particulares de artefactos: así y todo, este enfoque debe ser usado siempre con precaución.

Las técnicas descritas en este capítulo han proporcionado al lector las herramientas básicas para describir las distribuciones de datos por medios visuales. Como tales, pueden usarse siempre que queramos presentar una enumeración resumida de ciertos resultados, o bien para conseguir una primera impresión de cualquier estructura subyacente, presente en los datos. El estudio de tales distribuciones y sus implicaciones es un paso previo imprescindible antes del uso de los métodos estadísticos que se explicarán a lo largo de este libro.

20

EJERCICIOS

3.1. Considera las siguientes cantidades de fragmentos de distintos tipos de cerámica del yacimiento de Mount Pleasant, Dorset, Inglaterra (datos según Wainwright, 1979). Represéntalos por medio de un diagrama de barras y un gráfico de sectores, y di cuál de ellos prefieres.

cuenco neolítico liso	391
cerámica acanalada	657
campaniformes	1.695
cerámica de Petersborough	6
edad del bronce	591

3.2. Los tamaños (en hectáreas) de unos asentamientos del Uruk tardío, en Mesopotamia, (según Johnson, 1973) son:

45,0	37,0	34,8	52,0	75,0	86,0	59,7	74,0	32,0
57,7	65,0	86,0	37,0	38,4	90,5	45,0	67,0	50,0
33,0	30,0	43,2	32,0	35,2	54,5	43,1		

Usa un método gráfico apropiado para representar estos datos. ¿Es posible encontrar alguna estructura relacional en la distribución del tamaño de los asentamientos? ¿Cambia algo si alteras el intervalo?

3.3 Traza la distribución acumulativa del porcentaje de las frecuencias de los siguientes datos, referidos a los individuos enterrados en una necrópolis prehistórica.

Categoría de edad	Cantidad de enterramientos
Infantiles I	10
Infantiles II	16
Juveniles	10
Adultos	32
Maduros	34
Seniles	4

4. RESÚMENES NUMÉRICOS DE UNA VARIABLE ÚNICA

En el último capítulo examinamos varios métodos de representación gráfica para las distribuciones de observaciones medidas a distintos niveles. El presente capítulo, por su parte, trata acerca de los resúmenes *numéricos* de información. Yo sería el primero en aceptar que no es, precisamente, uno de los temas más interesantes, pero creo que hay dos razones que no nos permiten dejarlo de lado. En primer lugar, que esos resúmenes se están convirtiendo en un elemento importante de las descripciones publicadas en los trabajos arqueológicos. Las excavaciones modernas, por ejemplo, producen a menudo tantos hallazgos de ciertas categorías, que la única forma de presentar la información de manera suficientemente compacta como para publicarla es por medio de algún gráfico y un resumen numérico asociado. Presentar la información de este modo no es deshumanizar la arqueología, sino ofrecer de una forma publicable y comprensible la información sobre la cual se basarán las inferencias, de forma tal que los lectores tengan una oportunidad de evaluarla. Lo cual presume, evidentemente, un lector educado para ello.

La segunda razón por la que hay que considerar las descripciones numéricas es que los métodos que serán descritos posteriormente en el libro, y que tratan cuestiones mucho más interesantes acerca de la identificación de las relaciones entre variables, dependen del uso de las medidas de descripción que van a ser presentadas.

Es importante que recordemos lo que hemos hecho con los métodos gráficos de resúmenes de datos; nos hemos olvidado de los individuos, fragmentos individuales, restos de piedra tallada o cualquier otro, y hemos intentado obtener algún tipo de imagen globalizadora de las tendencias generales en la distribución de los datos. Aunque una imagen o diagrama de algún tipo pueden ser útiles a menudo para resumir la información que nos interesa, a veces también es necesario reducir aún más el conjunto de los datos, hasta dejarlo tan sólo en una o dos cifras, las *estadísticas descriptivas*. Es conveniente, sobre todo cuando queremos hacer comparaciones, por ejemplo, entre conjuntos de datos procedentes de yacimientos diferentes. Por otro lado, la reducción a uno o dos nú-

meros sencillos puede ser potencialmente peligrosa. Cuando se reduce una gran cantidad de información a un par de números, existe un mayor riesgo de equivocarse que con una imagen gráfica. La conclusión que hay que extraer de este hecho es que, incluso si reducimos numéricamente los datos, hay que seguir estudiándolos gráficamente.

Para una escala nominal, la cuestión es bastante trivial. Por ejemplo, tenemos varias categorías de huesos animales clasificados por especies, o bien recipientes divididos en tipos; expresamos las relaciones entre categorías por medio de los porcentajes de las distintas categorías en el conjunto; podemos referirnos a la categoría más común, o modal. Una vez hecho esto, poco más puede añadirse a la descripción.

Cuando tomamos en consideración variables medidas en una escala de intervalo o proporcional, y necesitamos el mejor resumen numérico de la información que disponemos, podemos plantearnos varias cuestiones. Para resumir completamente la información, de hecho, hemos de medir cuatro aspectos distintos de los diagramas de barras o histogramas que hemos visto, los cuales son:

1) *Tendencia central*, o ¿cuál es el individuo más típico?

2) *Dispersión*, o ¿cuánta variación hay? En una representación como la de la figura 4.1(a), un individuo típico es mucho más representativo que en una distribución como la de la figura 4.1(b).

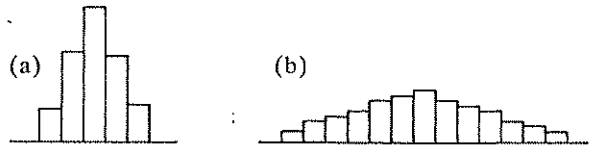


FIGURA 4.1. Dos distribuciones en las que hay: (a) muy poca, (b) mucha dispersión alrededor del valor central.

3) *Forma*, que tiene dos aspectos:

3a) ¿es o no simétrica la distribución? Las figuras 4.2(a), (b), (c) indican algunas de las posibilidades. En los dos últimos casos, la distribución está *inclinada*, es decir, la cola se dirige hacia la derecha (b) o hacia la izquierda (c).

3b) El segundo aspecto de la forma consiste en la longitud de las colas de la distribución, ilustrada en la figura 4.3.(a), (b). El grado de dispersión de esas dos distribuciones es bastante parecido, pero una tiene colas más largas

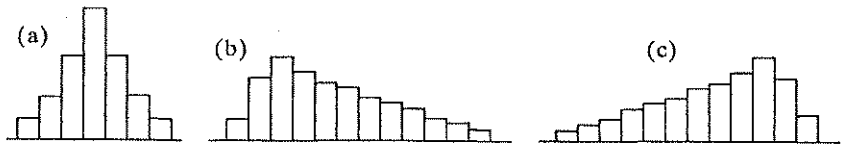


FIGURA 4.2. Ejemplos de distribuciones de formas diferentes.

22

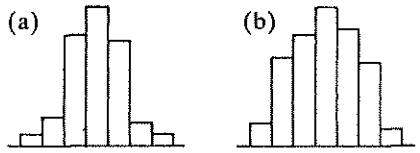


FIGURA 4.3. Ejemplos de distribuciones con colas de distintas longitudes.

que la otra. La longitud de las colas de una distribución es considerada como su grado de *curtosis*. Distribuciones con colas largas son *leptocúrticas* y las distribuciones con colas cortas son *platicúrticas*.

De hecho, las medidas de la forma suelen emplearse menos que las medidas de dispersión y de tendencia central en la mayoría de aplicaciones estadísticas, si bien la cuestión de la oblicuidad y curtosis es de gran importancia, normalmente como problema a resolver (véase el capítulo 8). No obstante, un área donde las medidas de oblicuidad y curtosis tienen un interés propio es en el campo del análisis de partículas. Un ejemplo arqueológico en ese campo podría ser la descripción y comparación de la pasta de la cerámica en términos de los distintos tamaños de las inclusiones (Peacock, 1971).

MEDIDAS DE LA TENDENCIA CENTRAL

Una vez enumeradas esas características descriptivas, podemos acercarnos a la cuestión de la tendencia central con más detalle. Hay varias formas de medirla, y lo mismo es cierto para la dispersión. La más conocida y usual de las medidas de tendencia central es la *media aritmética*, definida como la suma de las puntuaciones dividida por el número total de casos.

Tomemos como ejemplo los diámetros de siete de los hoyos para poste de Mount Pleasant, enumerados en el capítulo anterior:

$$48 + 57 + 66 + 48 + 50 + 58 + 47 = 374$$

Hay siete hoyos para poste, por lo que dividiremos 374 entre 7 y obtendremos 53,4 cm; ese el diámetro medio del conjunto de hoyos. Lo que estamos afirmando es que un hoyo para poste típico mide 53,4 cm de diámetro. No hay ninguno que mida realmente 53,4 cm, si bien este es un valor situado en algún sitio en medio de todos ellos. De hecho, la media representa el centro de gravedad de la distribución, con la propiedad específica de que la suma de desviaciones con respecto a la media de las puntuaciones individuales es siempre igual a 0. Es decir, si tomamos cada una de nuestras observaciones, restamos la media a cada una de ellas, y sumamos todas las diferencias, el resultado será cero. Así:

$$(48 - 53{,}4) + (57 - 53{,}4) + (66 - 53{,}4) + (48 - 53{,}4) + (50 - 53{,}4) + \\ + (58 - 53{,}4) + (47 - 53{,}4) = (-5{,}4) + 3{,}6 + 12{,}6 + (-5{,}4) + (-3{,}4) + \\ + 4{,}6 + (-6{,}4) = 0{,}2$$

(Este resultado no es exactamente igual a cero a causa de los errores de redondeo en los cálculos.)

Llegados a este punto, hemos de hacer una pequeña digresión. Se ha proporcionado una descripción verbal de cómo obtener una media aritmética, y de la propiedad que la caracteriza; esa descripción ha sido apoyada por medio de un ejemplo numérico. Sin embargo, si queremos especificar reglas generales para hacer operaciones con números, resulta mucho más conveniente usar el simbolismo matemático. Los símbolos constituyen una parte esencial de las matemáticas, aunque parezcan ser también su aspecto más complicado y difícil, para aquellos a los que la materia no les atrae directamente. Lo que hay que recordar acerca de los símbolos es que se trata, simplemente, de una forma de notación resumida, fácilmente manipulable.

Consideremos ahora el simbolismo que relaciona la media aritmética convencionalmente escrita como $\bar{x}$. Podemos decir, en general, que:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

donde x_1 es nuestra primera observación, x_2 la segunda, y así sucesivamente. En el caso de los hoyos para poste:

$$x_1 = 48 \qquad x_2 = 57$$
$$x_3 = 66 \qquad x_4 = 48$$
$$x_5 = 50 \qquad x_6 = 58$$
$$x_7 = 47$$

Hay siete observaciones, $n = 7$; así, en nuestro caso: $x_n = x_7 = 47$

Podemos resumir aún más y escribir:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Aquí, x_i está en el lugar de los valores x. $\sum$ es la letra griegra sigma mayúscula, y está en el lugar de la suma. Lo que se pretende decir con la fórmula es que hay que sumar ciertas x. ¿Cuáles? El índice y el exponente de $\sum$ nos lo dicen: de la primera x a la n-ava x, o $i = 1$ hasta n. En el ejemplo hay siete valores x que queremos sumar, por lo tanto:

$$\sum_{i=1}^{7} x_i$$

23

Tras hacer la suma, dividimos por el número de observaciones, otra vez 7 en el ejemplo, para llegar a nuestro valor para $\bar{x}$. Si los números que pretendemos sumar —el rango de la suma— son obvios, prescindiremos del índice y del exponente y escribiremos, simplemente:

$$\sum x_i$$

En estadística, nos encontraremos constantemente con esta expresión, ya que es muy importante; no debe intimidarnos, pues. La notación y el simbolismo son sólo una conveniencia para hacer las cosas más fáciles. Ahora ya podemos volver al tema principal, y usar la notación que acabamos de aprender.

El cálculo de la media en la forma que se ha descrito es sencillo, si es que tenemos sólo una pequeña cantidad de observaciones. Si el número de estas es muy elevado, el cálculo se vuelve tedioso, incluso utilizando una calculadora, pues hay que introducir muchas cifras, y siempre existe la posibilidad de error. Cuando se quiere calcular una media y se tienen muchos datos, suele ser mejor agrupar las observaciones en distribuciones de frecuencias, tal y como ya debe de haber hecho el lector para confeccionar un histograma o diagrama de barras. La fórmula para la media es, en ese caso:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

donde f_i es el número de casos de la i-ava categoría, $n = \sum f_i$, x_i es el valor de la i-ava categoría, y k es la cantidad de categorías.

Como ejemplo, supongamos otra vez que estudiamos unas tumbas y sus ajuares, y que queremos averiguar la cantidad media de los tipos distintos de objetos de ajuar funerario en un grupo de tumbas. Los datos aparecen en la tabla 4.1, de donde $\bar{x} = 215/67 = 3{,}2$.

TABLA 4.1.

N.º de tipo de objetos en las tumbas (x_i)	N.º de tumbas en cada categoría (f_i)	$f_i x_i$
1	1	1
2	22	44
3	15	45
4	20	80
5	9	45
$k = 5$	$\sum_{i=1}^{5} f_i = 67$	$\sum_{i=1}^{5} f_i x_i = 215$

Cuando se trabaja con valores numéricos continuos, la situación se complica, pues las observaciones diferentes habrán sido agrupadas para formar las barras o categorías de la distribución de frecuencias. En este caso, el valor x para cada categoría está dado por el punto central de esas categorías. Las figuras que ilustraban los ejemplos de la capacidad de unas vasijas, en el capítulo anterior, aparecen en la tabla 4.2, de donde $\bar{x} = 41.100/40 = 1.027{,}5$. La capacidad media para ese grupo de vasijas es de 1.027,5 ml.

TABLA 4.2.

Clases de vasijas según su capacidad (ml)	Punto medio de la capacidad en cada clase (ml) (x_i)	N.º de vasijas en cada clase (f_i)	$f_i x_i$
900- 949,99	925	4	3.700
950- 999,99	975	10	9.750
1.000-1.049,99	1.025	13	13.325
1.050-1.099,99	1.075	8	8.600
1.100-1.149,99	1.125	3	3.375
1.150-1.199,99	1.175	2	2.350
k (n.º de categorías) = 6		$\sum_{i=1}^{6} f_i = 40$	$\sum_{i=1}^{6} f_i x_i = 41.100$

Con esto finalizaremos la explicación de la media aritmética. Hay otras medidas de la tendencia central de una distribución, que podemos mencionar. Una importante es la *mediana*, que desempeña un papel crucial en el análisis de datos exploratorio, tal y como veremos más adelante. La mediana es aquel valor a partir del cual la mitad de las observaciones están por encima de él y la mitad por debajo. Obviamente, si queremos encontrar ese valor, habremos de disponer nuestras observaciones en orden ascendente o descendente de tamaño, es decir, un orden de rango. Por esa razón, sólo podremos calcular la mediana de datos ordinales, de escala interválica o proporcional.

Consideremos otra vez los diámetros de los hoyos para poste, que eran 48, 57, 66, 48, 50, 58, 47 cm. El primer paso será ordenarlos, del menor al mayor (o a la inversa): 47, 48, 48, 50, 57, 58, 66 cm. Si buscamos el valor para el cual la mitad de las observaciones están por encima de él, y la mitad por debajo, obviamente buscaremos el valor central. Tenemos siete observaciones. Si contamos hasta la cuarta, desde cualquiera de los dos extremos, nos detendremos en la cifra 50, que es la mediana: hay tres observaciones por debajo de ella y tres por encima. Si el número de casos es impar, la mediana será la puntuación del caso central. Si el número de casos es par, no aparecerá un único caso central, por lo que la mediana será la media de los dos casos centrales.

Supongamos que tuviésemos sólo seis diámetros de hoyos para poste: 48, 48, 50, 57, 58, 66 cm. En ese caso, la mediana se situaría entre 50 y 57, cuya

media es $107/2 = 53{,}5$. Para escalas ordinales, también puede calcularse el rango de la mediana.

Finalmente, hemos de mencionar la *moda*. Se trata, simplemente, del valor más común o más frecuente; obviamente, también existe en escalas nominales. En el ejemplo anterior de los ajuares funerarios (cuadro 4.1), el valor modal es 2; en el ejemplo de la capacidad de los recipientes es la clase 1.000-1.049,99 ml. Evidentemente, no existe moda sin que se disponga de una distribución de frecuencias de algún tipo, una cuestión que es particularmente relevante en el caso de datos numéricos continuos, en donde dos observaciones difícilmente tendrán el mismo valor.

Es posible que una distribución tenga más de una moda (fig. 4.4); o bien, una principal y otra subsidiaria (fig. 4.5). Probablemente sea mejor insistir en el hecho que si una distribución es bimodal, o tiene una moda principal y otra subsidiaria, proporcionar simplemente la medida de tendencia central o de dispersión es totalmente irrelevante. Deberíamos limitar el análisis a presentar las dos modas, o, preferiblemente, dividir la distribución en sus partes constituyentes y calcular las medidas relevantes para cada parte por separado (véase Mellars y Wilkinson, 1980, para un ejemplo de análisis de datos bimodales en arqueología).

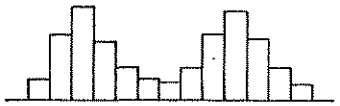


FIGURA 4.4. Un ejemplo de distribución bimodal.

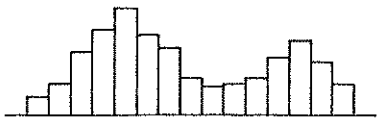


FIGURA 4.5. Una distribución con una moda principal y otra subsidiaria.

En el caso de los diámetros de los hoyos para poste, podemos estar interesados en estudiar específicamente si hay más de una moda en los datos, lo cual sugeriría que los diferentes grosores de los postes desempeñaban distintas funciones. La detección de cierta variación alrededor de una moda única, puede indicar meramente el grado de éxito de los constructores en encontrar troncos del tamaño justo para un único propósito; o bien, la flexibilidad de sus especificaciones.

¿Cómo se comparan entre sí las distintas medidas de tendencia central? En cierto sentido, la que usemos dependerá de lo que intentemos hacer; si bien, en general, si tenemos datos nominales u ordinales no habrá mucho donde ele-

gir, mientras que para datos en escala interválica o proporcional la medida de tendencia central más usada suele ser la media.

La moda nos da simplemente el valor más común, pero no nos explica su relación con los otros. La media emplea más información que la mediana, en el sentido de que se necesitan todas las puntuaciones exactas para calcularla, mientras que la mediana emplea, exclusivamente, las posiciones relativas de las puntuaciones. Muy a menudo es deseable hacer uso de toda la información disponible, por lo que se preferirá la media. De hecho, si la distribución es simétrica, la media, la mediana y la moda coincidirán. Si la distribución es asimétrica, sin embargo, la situación cambia significativamente. Supongamos, por ejemplo, que en el caso de las capacidades de los recipientes ilustrado en la figura 3.4, uno de ellos tiene un volumen de 2.500 ml. Tal distribución será asimétrica, con una observación muy desplazada hacia la derecha. En ese caso, la capacidad media de los recipientes variará considerablemente, desplazándose también hacia la derecha y apartándose del núcleo de las observaciones; por consiguiente, no sería significativa. Por otro lado, la mediana no quedaría muy afectada en ese caso, siendo mucho más representativa de la masa principal de observaciones.

La enseñanza que hay que extraer es que no nos hemos de plantear el problema de la tendencia central —cuál es el individuo más típico— de forma aislada, sino que simultáneamente hemos de reflexionar acerca de la forma de la distribución. Siempre es necesario saber a qué se parece la distribución de frecuencias de los datos, sobre todo si se pretende usar posteriormente métodos estadísticos más complicados; si la distribución presenta alguna peculiaridad, es imprescindible que la conozcamos de buen principio.

Medidas de la dispersión

Si los datos están muy dispersos, una única medida de la tendencia central no será muy representativa del valor típico. Este punto ha sido ilustrado en la figura 4.1.

Existen varias formas de cuantificar la dispersión. La más simple es el *rango*, la diferencia entre la puntuación más alta y la más baja. Su desventaja es que se basa exclusivamente en dos casos y además en los más extremos. Dado que los extremos son, casi por definición, los casos más extraños e inusuales, sería pura suerte si tuviésemos dos observaciones muy extremas en la muestra. Por esa razón, el rango no es una medida de la dispersión particularmente satisfactoria.

Más útil es una cantidad conocida como *rango intercuartil*. En el caso en el que podamos especificar la mediana de una distribución como el valor que tiene el 50 % de las observaciones por debajo de él y el 50 % por encima, podremos definir el primer y el tercer cuartiles de una distribución de valores de datos, medidos, cuando menos, en una escala ordinal. El primer cuartil es el valor que tiene el 25 % de las observaciones por debajo y el 75 % por encima de él, mientras que el tercer cuartil es el valor con el 75 % de las observaciones por debajo y el 25 % por encima de él. La diferencia entre el valor del primer y el tercer cuartiles, el 50 % central de la distribución, es denominado rango intercuartil. Su cálculo es análogo al de la mediana, y sus propiedades parecidas, ya que sólo el orden de las observaciones se tiene en cuenta, sin que influyan la existencia de valores muy grandes o muy pequeños en cualquiera de los extremos de la distribución. Volveremos sobre el rango intercuartil más adelante, en la página 58; por el momento nos limitaremos a decir que su uso en arqueología es meramente ocasional (por ejemplo, Ottaway, 1973).

Dado que el rango intercuartil sólo hace uso del orden de las observaciones, si los datos están medidos en una escala de intervalo o proporcional, se pierde información —no se usan las puntuaciones exactas de las observaciones—. Habitualmente se considera que es preferible utilizar toda la información disponible cuando se calcula una medida de la dispersión, al igual que con la media y las medidas de la tendencia central; pero al igual que ocurría con la media, hay ocasiones en las que el uso de toda la información puede inducir a engaño, es decir, puede producir resultados que dependan de la forma de la distribución (véase p. 58). Sin embargo, si usamos la media como medida de la tendencia central, lo más obvio será utilizar la suma de las desviaciones de las observaciones con respecto a la media, como base de una medida de la dispersión. Desafortunadamente, tal y como hemos visto, esta es siempre 0, ya que las diferencias positivas y negativas se anulan mutuamente. Hay dos formas de solventar el problema: podemos ignorar el signo y tomar en consideración sólo el valor absoluto de las diferencias; o podemos elevar al cuadrado las diferencias, recordando que menos por menos equivale a más, con lo que todas las cantidades se harán positivas.

De hecho, la segunda solución es mucho más usual: la medida de la dispersión se basa en el cuadrado de las diferencias entre la media y los valores de las observaciones individuales, lo cual se conoce como *desviación típica* o *estándar, s*:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

En palabras, tomamos la desviación de cada una de las puntuaciones con respecto a la media, elevamos al cuadrado dicha diferencia, sumamos los resultados, dividimos por el número de casos menos uno y extraemos la raíz cuadrada. El resultado es el siguiente: cuanto mayor sea la dispersión en la distribución, mayor será la desviación típica. Si detenemos el cálculo antes de extraer la raíz cuadrada, obtendremos la *varianza*, s^2, la media de las diferencias cuadradas

entre la media y los valores de los datos. Como la varianza es una cantidad elevada al cuadrado, hay que expresarla en unidades que sean el cuadrado de las unidades de medida originales. Muy a menudo, al describir una distribución, es deseable haber medido la dispersión en las mismas unidades que las medidas originales, por lo que la desviación típica tiende a ser, intuitivamente, más significativa que la varianza. Por ejemplo, si queremos saber el valor del grado de dispersión alrededor de la media en la distribución de las longitudes de un conjunto de láminas de sílex, ya que las hemos medido en milímetros, necesitaremos que la dispersión se mida también en milímetros y no en milímetros cuadrados.

La varianza y la desviación típica desempeñan un papel muy importante en muchas pruebas estadísticas, y por esa razón son las medidas de la dispersión fundamentales en los conjuntos de datos en los que pueden aplicarse, esto es, distribuciones simétricas y unimodales. Algunos de los problemas que plantean serán tratados más adelante; antes es necesario que nos fijemos en el método de cálculo en sí, porque la fórmula anterior es difícil de usar cuando se tienen muchas observaciones. Hoy en día, estas dificultades son menos problemáticas, a causa de la sofisticación de muchas de las modernas calculadoras, aunque no estará de más ofrecer una de las versiones de la misma fórmula mucho más fácil de calcular:

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \left[\left(\sum_{i=1}^{n} x_i\right)^2 / n\right]}{n - 1}}$$

Calcularemos a continuación la desviación típica de los diámetros de los siete hoyos para poste, por medio de ambas fórmulas.

Ya hemos visto que $\bar{x} = 374/7 = 53{,}4$; usando la primera fórmula, la suma de las desviaciones al cuadrado es:

$$(48 - 53{,}4)^2 + (57 - 53{,}4)^2 + (66 - 53{,}4)^2 + (48 - 53{,}4)^2 + \\ + (50 - 53{,}4)^2 + (58 - 53{,}4)^2 + (47 - 53{,}4)^2 = 303{,}71$$

$$s = \sqrt{\frac{303{,}71}{7 - 1}} = 7{,}1$$

Antes de ilustrar el segundo método, y estar seguros de obtener el mismo resultado, es importante tener bien claras las diferencias entre los términos $\sum x_i^2$ y $(\sum x_i)^2$. En el primer caso, tomaremos cada valor x, lo elevaremos al cuadrado y sumaremos todos los cuadrados. En el segundo caso, sumaremos los valores x y elevaremos al cuadrado el total. ¡Lo cual da resultados muy distintos! Así:

$$s = \sqrt{\frac{20.286 - [(374)^2/7]}{7 - 1}} = \sqrt{\frac{303{,}71}{6}} = 7{,}1$$

Si tratamos con datos agrupados, la fórmula de la desviación típica es:

$$s = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}{n - 1}}$$

donde x_i es el valor de la i-ava categoría, f_i es la cantidad de observaciones en la i-ava categoría, k es la cantidad de categorías y n es el número total de observaciones.

La significación precisa de la desviación típica como una medida de la dispersión no será estudiada hasta el capítulo 8, cuando presentemos la distribución «normal», una distribución especial en forma de campana que los estadísticos consideran muy útil; por el momento, sin embargo, esta medida proporciona una «típica» desviación de los datos con respecto a la media. En el caso de los hoyos para poste, 7,1 cm es una típica desviación con respecto a la media.

A veces nos interesa hacer comparaciones entre conjuntos de datos, basándonos en su dispersión. Por ejemplo, si lo que estudiamos es la estandarización de la producción de núcleos líticos en canteras prehistóricas, nos interesará saber si los tamaños de los núcleos de una cantera son más variables que los de otra, lo cual permitirá, probablemente, hacer inferencias sobre distintos grados de especialización en la producción artesanal. Muy a menudo, cuanto mayor es la medida, mayor es la desviación típica, de forma que si en una cantera los núcleos son mayores que en otra, la distribución de tamaños de los núcleos de la primera tendrá una desviación típica mayor por esa razón, y no porque esté menos estandarizada. Podemos eliminar este efecto usando el *coeficiente de variación*, es decir, la desviación típica dividida por la media; a veces el resultado se multiplica por 100, para convertirlo en un porcentaje. El resultado es una medida de la dispersión estandarizada.

El análisis de datos exploratorio y los resúmenes numéricos: descripciones robustas de la tendencia central y de la dispersión

Los resúmenes numéricos tradicionales, basados en la media, no están muy bien considerados en el enfoque del análisis de datos exploratorio, que pone

un mayor énfasis en la importancia de las buenas representaciones gráficas, como ya hemos visto. Dado que esos resúmenes suelen ser convenientes y necesarios, se insiste en que sean lo más robustos posible. En otras palabras, han de hacer lo que se tiene previsto que hagan, proporcionar resúmenes adecuados en una gran variedad de situaciones distintas, y no bajo una condiciones restrictivas. En particular, han de ser resistentes a los cambios en uno o dos valores de la distribución. Si lo que buscamos son resúmenes robustos de la tendencia central y de la dispersión en una distribución, la media y la desviación típica no son muy satisfactorias, ya que su utilidad se restringe a las distribuciones unimodales y simétricas. En otras condiciones, la mediana y el rango intercuartil proporcionan una mejor indicación del valor de una observación típica, y del grado de dispersión alrededor de ese valor. Esa es la postura adoptada por el análisis de datos exploratorio, aunque muchos de sus defensores prefieren usar el término «dispersión central» [*midspread*, en la bibliografía anglosajona] antes que el de rango intercuartil, refiriéndose al valor del cuartil más bajo como «umbral inferior» y al cuartil superior como «umbral superior» de la distribución.

El uso de la mediana y del rango intercuartil puede extenderse a la producción de resúmenes numéricos de una distribución. De esta forma, la distribución se describirá por medio de los valores de la mediana, umbrales superior e inferior, así como por sus valores máximo y mínimo, junto a una indicación de los intervalos entre ellos. Para los 35 diámetros de hoyos para poste, obtendríamos:

	Mín.	Umbral inferior	Mediana	Umbral superior	Máx.
	25	38,5	44	48	66
Intervalos	13,5	5,5	4	8	

A lo cual podemos añadir las distancias entre el mínimo y la mediana («dispersión inferior»), entre los umbrales inferior y superior («rango intercuartil o dispersión central») y entre la mediana y el valor máximo («dispersión superior»). Así:

25	38,5	44	48	66
13,5	5,5	4	8	
	19	9,5	22	

Ahora podemos ver rápidamente que la diferencia de tamaño entre el menor de los hoyos para poste (25 cm) y el umbral inferior (38,5 cm) es de 13,5 cm, mientras que del umbral inferior a la mediana (44 cm) es sólo de 5,5 cm; la diferencia entre el mínimo y la mediana alcanza un total de 19 cm.

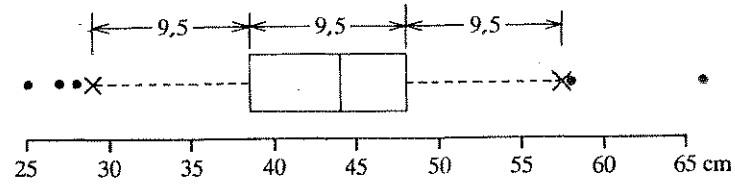


FIGURA 4.6. Gráfico de caja y arbotante de los diámetros de los 35 hoyos para poste en Mount Pleasant.

Para descubrir si una distribución adopta o no una forma parecida a la de una campana —lo que significaría que es apropiada para el uso de las estadísticas basadas en la teoría normal, explicada en el capítulo 8 y en los últimos capítulos de este libro—, podemos hacer unas cuantas comprobaciones con estos números (Hartwig y Dearing, 1979, p. 23). Si la distribución es simétrica y en forma de campana:

1) la dispersión inferior y la superior serán iguales;

2) las distancias del umbral inferior a la mediana y del umbral superior a la mediana serán iguales;

3) las distancias entre el umbral inferior y el valor mínimo y entre el umbral superior y el valor máximo serán iguales;

4) las distancias desde la mediana a los umbrales serán menores que las distancias desde los umbrales a los extremos, a causa de la concentración de casos en la parte central de la distribución.

A partir de estos criterios, parecería que la distribución de hoyos para poste está bastante cerca de la «normalidad», tal y como se deduce también de la forma del histograma de tallo y hoja de esos datos (véase fig. 3.7), pero, como veremos, esta forma de resumen *numérico* no da tanta información como sería necesaria acerca de las colas de la distribución, que también han de ser estudiadas.

Los resúmenes numéricos del tipo de los que acabamos de tratar pueden contrastarse con ayuda de los resúmenes *gráficos* de una distribución, que son distintos de las puras representaciones gráficas, como los diagramas de tallo y hoja. La inusualidad de las distribuciones suele ser particularmente evidente en las colas de la misma, lo cual es fácilmente representable mediante un método gráfico llamado *gráfico de caja y arbotante*.

Un gráfico de este tipo para los datos de Mount Pleasant está ilustrado en la figura 4.6. La pared izquierda de la caja está a la altura del umbral inferior de la distribución, la pared derecha a la altura del umbral superior, y la línea vertical señala la mediana. De este modo, la caja contiene la mitad de los casos de la distribución. Las cruces señalan los casos más alejados de la caja a cada uno de los lados pero todavía dentro de una[1] dispersión central del umbral más próximo. Más allá de ese punto, los casos están señalados individualmente.

1. En MINITAB, los extremos de los arbotantes (*whiskers*) se sitúan a 1,5 dispersiones centrales del umbral más próximo.

En una distribución normal, el 95 % de todos los casos están dentro del rango determinado por los puntos extremos (las cruces) de los arbotantes, por lo que una distribución con más del 5 % de los valores fuera de esos límites empieza a apartarse de la normalidad (véase Hartwig y Dearing, 1979, p. 24). En el caso de los hoyos para poste en Mount Pleasant (fig. 4.6) apreciamos que 5 de las 35 observaciones (14 %) están fuera de los límites, lo que sugiere, después de todo, que la imagen no es muy fácil de interpretar. En este caso, tres de los hoyos son muy pequeños y dos enormemente grandes, en relación con el resto de la distribución. En un estudio real, el paso siguiente sería identificar esos casos extremos en la planta del yacimiento y estudiar si pueden tener o no una función particular.

Los diagramas de tallo y hoja y los de caja y arbotante están disponibles en el paquete de programas estadísticos MINITAB (véase anexo 2), que proporciona una de las mejores maneras de obtenerlos. Tal y como ya hemos visto, las dos técnicas se diferencian en que el diagrama de tallo y hoja proporciona una representación visual de los datos completa, mientras que los diagramas de caja y arbotante sacrifican la exhaustividad en beneficio de una mejor forma de ver si las colas de la distribución se apartan de la normalidad. Este último tipo de gráficos puede ser particularmente útil cuando se hacen comparaciones entre los diferentes conjuntos de datos, porque permite ver de inmediato si existe asimetría en el núcleo principal de las observaciones, y si algunos conjuntos tienen más observaciones extremas que otros. La razón de esas diferencias puede estudiarse posteriormente con más detalle.

Una combinación de diagramas de tallo y hoja, resúmenes numéricos y gráficos de caja y arbotante es capaz de revelar cualquier peculiaridad en la forma de la distribución. Cuando aparecen peculiaridades como la oblicuidad, la multimodalidad o la presencia de valores extremos (*outliers*), la característica más importante de la distribución será su forma y no la tendencia central o la dispersión; confiar exclusivamente en la tendencia central y la dispersión como medidas para resumir una distribución puede hacernos perder la característica más importante del conjunto de observaciones. Además, si no se conoce la forma de la distribución será imposible explicar si las medidas de la tendencia central o de la dispersión son engañosas o de qué forma lo son; la forma es relevante en la selección de las medidas apropiadas para averiguarlo. Finalmente, hemos de recordar que como mejor se reconoce la forma es percibiéndola visualmente.

Debe insistirse en la importancia de considerar la forma de la distribución. Es sorprendente cuánta gente se pierde por ignorar las reglas básicas del análisis de los resúmenes numéricos y visuales de las variables individuales, aun cuando las conozcan bien.

EJERCICIOS

4.1. Refiriéndote a los datos de los asentamientos de Uruk en el ejercicio 3.2: (a) ¿Cómo describirías la forma de la distribución de frecuencias en esos datos? (b) ¿Cuáles son las medidas más apropiadas de la tendencia central y de la dispersión? Calcúlalas.

4.2. Dada la siguiente lista de longitudes (en metros) de unos túmulos funerarios del neolítico en la Inglaterra meridional, analiza su distribución usando las técnicas descritas en este capítulo y en el precedente. Razona tus conclusiones.

33	30	36	60	70	95	75	63	60
34	58	72	70	44	35	71	51	56
60	98	49	70	61	81	74	64	51
95	69	56	37	31	58	51	51	52

28

SHENNAN, S.
"ARQUEOLOGÍA CUANTITATIVA"
ED. CRÍTICA

5. INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

Una vez tratadas algunas de las estadísticas descriptivas básicas podemos considerar el tema de la inferencia estadística y el uso de los métodos estadísticos comparativos en arqueología. Hay que tener en cuenta, de buen principio, que la inferencia estadística no es en absoluto algo simple, desde un punto de vista conceptual, y que muchas de las cuestiones que ha planteado en arqueología son muy controvertidas (véase, por ejemplo, Cowgill, 1977). Hay dos contextos bastante diferentes entre sí en los que estos temas afectan a la arqueología, por lo que será conveniente separarlos claramente desde ahora, incluso aunque, en el fondo, estén relacionados.

En primer lugar, consideraremos la situación en la que el arqueólogo efectúa una selección durante la realización de un proyecto arqueológico. En muy pocas ocasiones los arqueólogos cuentan con recursos suficientes para investigar todo aquello en lo que están interesados, ya sean regiones, yacimientos o conjuntos de artefactos. Normalmente pueden investigar tan sólo una parte de la población que les interesa, y necesitarían que esa parte fuese representativa del total, si es posible. Siempre y cuando los objetivos del estudio estén bien formulados, los conceptos de la inferencia estadística serán útiles en la selección de una muestra que proporcione resultados en cuya fiabilidad y precisión podamos confiar. Hay toda una serie de problemas, relativamente bien definidos, que están implicados en esta situación. Normalmente aparecen incluidos en arqueología bajo el título de *muestreo*, y los discutiremos, por tanto, en el capítulo que trata específicamente ese tema.

El segundo contexto o situación en que la inferencia estadística puede ser relevante está enmarcado por el proceso de comparación. ¿Es la densidad de yacimientos en el área A la misma que en el área B? ¿Es idéntica la proporción de cerámica del tipo X en el yacimiento Y que en el yacimiento Z? ¿Aparecen, en la necrópolis T, los anillos de tipo S con mayor frecuencia en las tumbas femeninas que en las masculinas? A menudo, las cuestiones que se plantea el arqueólogo son de este tipo, ya se trate de comparaciones entre diferentes conjuntos de datos, como en los ejemplos que acabamos de ver, o comparaciones

entre un conjunto de datos observados y las suposiciones derivadas de un modelo teórico: por ejemplo, si la distribución de yacimientos difiere de la aleatoriedad o no.

En todos estos casos, es improbable que los yacimientos, regiones o sexos tengan exactamente los mismos valores en cada caso, para la variable en cuestión. En cualquier comparación entre dos casos siempre aparecen diferencias entre ellos, aunque estas sean pequeñas. El problema es, por tanto, qué magnitud ha de tener la diferencia para que la tomemos en serio y afirmemos que es «real». Esta es una cuestión perfectamente válida, que la estadística puede resolver por medio de las *pruebas de significación*. Para ver cómo funcionan, es preciso estudiar la teoría que las fundamenta, en un nivel abstracto, antes de pasar a su uso arqueológico. El hecho de que consideremos que las pruebas de significación desempeñan un papel importante no obliga a suponer que esta sea la principal justificación de la aplicación de los métodos cuantitativos en arqueología.

Muestras y poblaciones

La inferencia estadística trata de los problemas de la toma de decisiones e incertidumbre. Esta última se cuantifica por medio de la teoría de las probabilidades. Las inferencias se establecen con referencia a «poblaciones», lo que provoca la incertidumbre porque esas inferencias están planteadas a partir de muestras de tales poblaciones. Precisamente, qué es lo que son esas poblaciones está sujeto a discusión en arqueología (veáse más adelante). Hay dos aspectos principales en la inferencia estadística: contrastación de hipótesis y estimación (véase Cowgill, 1977). Ambas están conectadas, si bien en términos generales la primera implica la contrastación de una idea acerca de la población y la segunda implica la estimación del valor de algunas de las características de la población, partiendo de una muestra de los datos, o bien proporcionando los límites superior e inferior en los que se espera encontrar el valor estimado. La manera de presentar las dataciones de radiocarbono es un ejemplo clásico de este último procedimiento. En ambos casos, lo que se pretende es decir algo acerca de algún aspecto de una población, partiendo de una muestra de ella.

Las características de una población reciben el nombre de *parámetros*; las características de una muestra son los *estadígrafos*; la distinción entre ambos es importante. Las características de una población suelen representarse por medio de letras griegas, los estadígrafos de una muestra por medio de letras minúsculas normales. Así, la media de una población se designa mediante la μ griega (my) y la media de la muestra por $\bar{x}$; la desviación típica de la población por σ (sigma) y la desviación típica de la muestra por s.

Los parámetros son valores fijos que se refieren a la población y, por lo general, son *desconocidos*; por ejemplo, la media del diámetro de la boca de las

vasijas tipo Y del yacimiento X. Los estadígrafos, por el contrario, varían de una población a otra, si bien pueden ser calculados; por ejemplo, la media del diámetro de la boca de las vasijas del corte A o del corte B en el yacimiento X. Por otro lado, *no sabemos* si la muestra es representativa de la población o si el estadígrafo obtenido corresponde aproximadamente al parámetro desconocido. Generalmente, nuestro objetivo es hacer inferencias acerca de varios parámetros poblacionales, partiendo de unos estadígrafos muestrales conocidos.

En las pruebas de hipótesis se presumen los parámetros desconocidos y se estudia cómo serían los estadígrafos muestrales si esos supuestos fuesen verdaderos. Pretendemos decidir si los supuestos sobre los parámetros son válidos o no, vista la evidencia a nuestra disposición; de este modo, la contrastación de una hipótesis podrá ser considerada como una manera de tomar decisiones. Los dos tipos de cuestiones que aparecen normalmente en las pruebas de hipótesis son:

1. ¿Cuál es la probabilidad de que dos (o más) muestras hayan sido extraídas de la misma población?
2. ¿Cuál es la probabilidad de que una muestra haya sido extraída de una población con ciertas características definidas?

Esta es una presentación bastante esquemática de la noción general de contrastación de hipótesis. En particular, tal y como indican estas dos cuestiones, las hipótesis son afirmaciones definidas a partir de suposiciones acerca de los datos y, por tanto, potencialmente rechazables. Si somos o no capaces de refutar una hipótesis, es porque nuestros juicios se basan en muestras, con lo que habremos de admitir la posibilidad de error debido a la falta de representatividad de la muestra. La teoría de las probabilidades nos permite evaluar los riesgos de error y tomar en consideración esos riesgos.

En general, se suele empezar comprobando lo que se denomina *hipótesis nula*: la hipótesis de la no diferencia. Refiriéndonos de nuevo a las dos cuestiones anteriores, partimos del supuesto de que dos o más muestras han sido extraídas efectivamente de la misma población; o bien, que la muestra procede realmente de una población con unas características bien especificadas. Clive Orton (1980) presenta este tema en forma de la pregunta: «¿Hay lugar para la respuesta?». Lo veremos mejor mediante un ejemplo. Supongamos que hemos de comparar la densidad de yacimientos arqueológicos en dos áreas distintas: ¿son o no diferentes? El procedimiento habitual comienza planteando una hipótesis nula que afirma la no existencia de diferencias entre las dos áreas, en lo que respecta a la media de la densidad de yacimientos; a continuación se examina la evidencia contra esta hipótesis de no diferencia.

Si imaginamos nuestras dos áreas divididas en cuadros de un kilómetro de lado, será poco probable que cada cuadro en cada una de las dos áreas tenga el mismo número de yacimientos. Habrán diferencias considerables entre ellos, de forma tal que, si elegimos al azar diez cuadros de un área, calculamos la media de la densidad de yacimientos en esos diez cuadros y hacemos lo mismo

con diez cuadros de la otra área, las medias de ambas muestras serán distintas, incluso aunque las medias de las densidades poblacionales (las medias de las densidades para todos los cuadros en cada una de las dos áreas) sean idénticas. Esto sucedería si, al azar, tomásemos diez cuadros de una parte débilmente ocupada de una de las áreas y diez cuadros de una parte densamente poblada de la otra.

Estos efectos aleatorios o *estocásticos* del muestreo están expresamente incluidos en las pruebas estadísticas, y por eso se toman en consideración cuando, por medio de la prueba, decidimos si las medias de las densidades de las poblaciones (como opuestas a las muestras) realmente difieren una de la otra o no.

Normalmente, la hipótesis nula, H_0, tal y como suele ser designada simbólicamente, es comparada a la hipótesis alternativa, H_1. Por el momento, nos limitaremos a decir que esta hipótesis alternativa es, simplemente, la que afirma que sí que existe *una diferencia*; no dice nada acerca del tipo de la diferencia existente. Es habitual, pero no invariable, que en el análisis estadístico aceptemos la hipótesis alternativa cuando se rechaza la hipótesis nula, si bien, ocasionalmente, se propone la hipótesis nula con la esperanza de que sea válida.

Si estamos a punto de tomar la decisión acerca de rechazar o no la hipótesis nula, ¿qué criterio usaremos como fundamento, dados los efectos aleatorios que pueden surgir en la extracción de las muestras? Esencialmente, observaremos los valores de nuestras dos muestras, anotaremos las diferencias entre ellas y nos preguntaremos la probabilidad de que una diferencia como esta pueda ocurrir si las dos muestras procediesen realmente de la misma población. Si la probabilidad de una diferencia tan grande es reducida (partiendo del supuesto de la no diferencia), rechazaremos el supuesto y concluiremos que *hay* una diferencia. Esta probabilidad es conocida como *nivel de significación*, y se expresa simbólicamente por medio de la letra griega α (alfa).

Es una decisión del investigador elegir un nivel de significación aceptable. Esto significa decidir si un resultado es improbable o no bajo el supuesto de la hipótesis nula (la hipótesis de la no diferencia), antes que esa hipótesis sea rechazada. Normalmente, antes de llegar tan lejos, nos interesará que la probabilidad de validez de la hipótesis nula sea muy reducida, dados los resultados, con lo que se dispondría de una cierta confianza a la hora de rechazarla.

Por convención, los dos niveles de significación más empleados son: $\alpha = 0{,}05$ y $\alpha = 0{,}01$. Elegir un nivel de significación de 0,05 significa que hemos decidido aceptar la hipótesis nula como verdadera, a no ser que nuestros datos sean tan atípicos que tan sólo aparezcan así 5 veces de cada 100, o menos, si la hipótesis nula (hipótesis de la no diferencia) fuese cierta, en cuyo caso la rechazaríamos. En otras palabras, si extraemos 100 pares de muestras de dos poblaciones idénticas y anotamos la diferencia entre sus valores, sólo cinco de las diferencias, como media, serán tan grandes como las observadas. En esas circunstancias decidiríamos que los resultados hacen improbable que la hipótesis

nula sea cierta. Igualmente para el nivel de significación 0,01, si bien en este caso nos planteamos la cuestión de si los resultados son tan atípicos que sólo ocurrirían en el 1 % de los casos, o menos, bajo la hipótesis nula, antes de rechazarla; es decir, si la hipótesis de no diferencia es correcta, entonces esperaremos ese resultado sólo una de cada 100 veces, o menos.

Naturalmente, es perfectamente razonable usar otros niveles de significación. Si la decisión es de importancia crítica, sólo querríamos equivocarnos una de cada 1.000 veces.

Podemos pensar que lo mejor sería trabajar siempre con niveles de significación muy conservadores —sólo rechazar la hipótesis nula si la probabilidad de que sea válida es del 1 ‰, o menos; pero hay una trampa en eso, porque si se rechaza la hipótesis nula sólo en circunstancias extremas, se incurrirá en el riesgo de *aceptar* la hipótesis nula cuando, posiblemente, *es falsa*: el error inverso al anterior.

Rechazar la hipótesis nula cuando es verdadera es considerado como un «error de tipo I». En términos estadísticos se trata de un «pecado de comisión», porque significa que se ha afirmado la existencia de una relación o una diferencia relevantes, cuando no hay nada de eso. Aceptar la hipótesis nula cuando es falsa, se conoce como «error de tipo II» y significa el fracaso en la identificación de una relación o diferencia significativas, cuando realmente existen. La mayoría de las veces es más grave cometer un error de tipo I —afirmar la relación cuando no existe— que fallar en la identificación de una relación significativa.

Estableciendo un nivel de significación, decidiremos acerca de la probabilidad de incurrir en un error de tipo I. Es precisamente la gravedad de este tipo de errores —afirmar algo con respecto a los datos cuando no es cierto— lo que ha conducido a los estadísticos profesionales a fijar una serie de condiciones antes de rechazar la hipótesis nula. Esto es así siempre y cuando nos hayamos propuesto estudiar la relevancia de un rechazo de la hipótesis nula, que es lo más habitual; si, por el contrario, esperamos averiguar que la hipótesis nula es cierta, habrá que esforzarse en minimizar la probabilidad de cometer errores de tipo II, la probabilidad de aceptar H_0, cuando es falsa.

Pruebas de significación en arqueología

La discusión anterior sobre muestras, poblaciones, hipótesis nulas y niveles de significación nos da algunas indicaciones acerca de lo que hay que hacer para realizar una prueba de significación, aunque no se han considerado los supuestos requeridos para que esa prueba sea llevada a cabo satisfactoriamente, ni tampoco la manera en que los datos arqueológicos se relacionan con esos supuestos; el ejemplo arqueológico específico que hemos usado como ilustración estaba definido para que se ajustase a todos los supuestos necesarios. Ahora

tendremos que plantear algunas cuestiones implicadas en el uso de las pruebas de significación en contextos arqueológicos, y el método más claro para hacerlo es siguiendo un ejemplo hipotético y tomando en consideración sus implicaciones.

Por ejemplo, estamos estudiando una necrópolis hipotética en Checoslovaquia, y hemos apreciado que las tumbas femeninas pueden dividirse en dos grupos, basándonos en los objetos del ajuar; esas dos clases han sido llamadas «ricas» y «pobres». El problema que se plantea es si las distribuciones de edad en el momento de la muerte de los individuos son distintas en una y en otra clase; la respuesta a esta pregunta es muy importante para la interpretación final de la necrópolis. La información necesaria aparece en la tabla 5.1.

La pregunta es del tipo de las examinadas al principio del capítulo: ¿son iguales las dos distribuciones o no? Ya que no son exactamente iguales, ¿son las diferencias lo suficientemente grandes como para poder afirmar que son realmente distintas? Se trata, una vez más, de la misma pregunta planteada por Clive Orton. ¿Es posible una respuesta? Ciertamente se trata de algo muy importante, ya que si concluimos que la diferencia es real, nos encontraremos ante un fenómeno que hay que explicar. Si, por el contrario, concluimos que las diferencias no son suficientes como para poder ser tomadas en consideración, el análisis se detendrá, necesariamente, en ese punto.

Si queremos usar una prueba de significación como fundamento a la decisión de si tomamos en consideración o no las diferencias, lo primero que hay que hacer es establecer una hipótesis nula y su alternativa:

H_0: no hay diferencia entre tumbas femeninas «ricas» y «pobres», basándonos en la distribución de edades;
H_1: sí hay diferencia entre tumbas femeninas «ricas» y «pobres», basándonos en la distribución de edades.

Tabla 5.1. Distribución de edades en un grupo de tumbas femeninas, procedente de una necrópolis hipotética de la edad del bronce en Checoslovaquia. Las tumbas han sido divididas en «ricas» y «pobres», según los objetos de ajuar asociados a ellas.

Categoría de edad	Categoría de «riqueza» «Ricas»	«Pobres»
Infantil I	6	23
Infantil II	8	21
Juvenil	11	25
Adulta	29	36
Madura	19	27
Senil	3	4
	76	136

Supongamos que en este caso seguimos la convención al uso y elegimos un nivel de significación de 0,05; es decir, rechazaremos H_0 si los resultados observados sólo ocurren 5 de cada 100 veces.

Para llevar a cabo cualquier prueba es necesario plantearse unos supuestos previos acerca de la población que nos interesa estudiar, y acerca de los procedimientos de muestreo que se van a usar. Estos supuestos pueden dividirse en dos categorías: aquellos que podemos aceptar, y aquellos que son dudosos; nos interesan, especialmente, los segundos. La hipótesis nula es el supuesto dudoso más interesante; desde el punto de vista de la prueba estadística, desgraciadamente, todos los supuestos tienen el mismo estatus lógico: si los resultados de la prueba sugieren el rechazo de los supuestos, todo lo más que podremos afirmar es que, como mínimo, uno de los supuestos es falso. La prueba no nos dirá cuál de ellos, por lo que, si queremos que los resultados sean significativos, tendremos que poner en duda únicamente uno de los dos. Por esta razón, en el momento de elegir una prueba es importante seleccionar aquella que implique un único supuesto dudoso, esto es, la hipótesis nula.

Una de las primeras cosas que hemos de tener en cuenta cuando seleccionamos una prueba es el nivel de medida de los datos. Las pruebas para datos medidos en una escala de intervalo o proporcional no son apropiadas para datos medidos en un nivel inferior. Por otro lado, si usamos una prueba apropiada para datos medidos en un nivel bajo, en un conjunto de datos medidos en otro nivel más alto, estaremos desperdiciando información al no aplicar pruebas más poderosas. En el presente caso, imaginaremos que las categorías de edad representan una escala nominal.

Tal y como veremos, muchas pruebas estadísticas requieren supuestos específicos acerca de la forma de la distribución que se está estudiando. Una prueba apropiada para comparar dos escalas ordinales y que especifique muy pocos de esos supuestos es la *prueba de Kolmogorov-Smirnov*. Requiere que las observaciones se dividan en dos categorías mutuamente exclusivas, como mínimo, y que estén medidas a nivel ordinal o superior. La prueba se basa en las diferencias entre las dos distribuciones cumulativas comparadas; para la versión descrita aquí, ambas muestras han de tener más de cuarenta individuos.

El primer paso es convertir las frecuencias originales en proporciones de la categoría total. Así, por ejemplo, si hay 76 tumbas en la categoría «ricas», 6 de las 76 pertenecen a la categoría de edad infantil I, lo cual, expresado proporcionalmente, equivale a $6/76 = 0{,}079$, si la escala va de 0 a 1, o bien, 7,9 %, en una escala de 0 a 100. Esta operación se lleva a cabo con cada una de las categorías de edad de las tumbas, divididas según su «riqueza». Los resultados aparecen en la tabla 5.2.

Las proporciones se acumulan para cada categoría de edad y para cada clase de «riqueza», tal y como ya hemos visto cuando estudiábamos las curvas cumulativas. Así, la proporción de tumbas «ricas» en la categoría infantil II o

TABLA 5.2. Cantidades y proporciones de los enterramientos por riqueza y categorías de edad.

Categoría de edad	Categoría de «riqueza» «Ricas»		«Pobres»	
Infantil I	6	0,079	23	0,169
Infantil II	8	0,105	21	0,154
Juvenil	11	0,145	25	0,184
Adulta	29	0,382	36	0,265
Madura	19	0,250	27	0,199
Senil	3	0,039	4	0,029
	76	1,000	136	1,000

TABLA 5.3. Proporciones acumuladas de los enterramientos por riqueza y categorías de edad.

Categoría de edad	Categoría de «riqueza» «Ricas»	«Pobres»
Infantil I	0,079	0,169
Infantil II	0,184	0,323
Juvenil	0,329	0,507
Adulta	0,711	0,772
Madura	0,961	0,971
Senil	1,000	1,000

menor es $0{,}079 + 0{,}105 = 0{,}184$; en la categoría juvenil, o menor, es $0{,}184 + 0{,}145 = 0{,}329$, y así sucesivamente. El resultado se muestra en la tabla 5.3.

La prueba se basa en el cálculo de la mayor de todas las diferencias entre las dos distribuciones de proporciones acumuladas, de forma que el paso siguiente será calcular las diferencias entre ellas para cada categoría de edad, y anotar cuál de ellas es la mayor (sin fijarse en el signo de las diferencias). Veamos la tabla 5.4.

TABLA 5.4. Proporciones acumuladas de enterramientos por riqueza y categorías de edad y diferencias entre ellas.

Categoría de edad	Categoría de «riqueza» «Ricas»	«Pobres»	Diferencia
Infantil I	0,079	0,169	0,090
Infantil II	0,184	0,323	0,139
Juvenil	0,329	0,507	0,178
Adulta	0,711	0,772	0,061
Madura	0,961	0,971	0,010
Senil	1,000	1,000	0,000

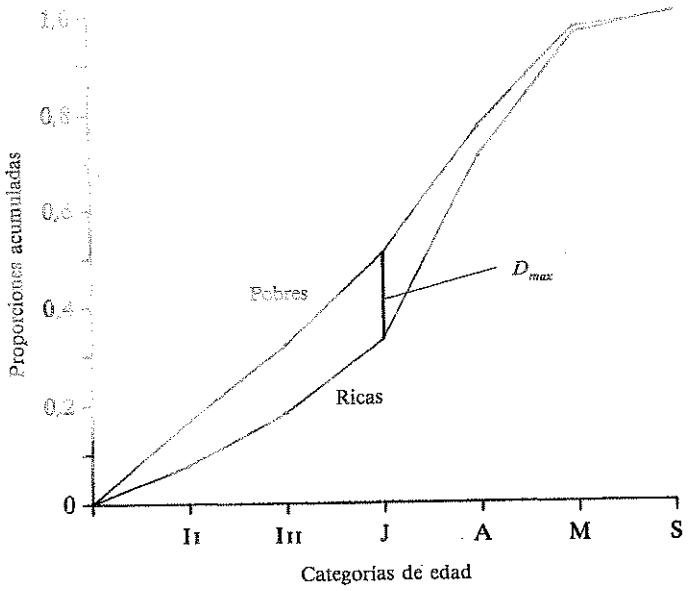


FIGURA 5.1. Gráfico de las distribuciones acumulativas de edad en los enterramientos «ricos» y «pobres» realizado con los datos de la tabla 5.1.

Basándonos en ella podemos apreciar que la mayor de las diferencias aparece en la categoría juvenil, y que es de 0,178. Antes de considerar qué hacer con este número, y cómo se relaciona con el nivel de significación que hemos especificado, será mejor que presentemos esas distribuciones gráficamente, con el fin de obtener una apreciación intuitiva de su apariencia y de lo que representa la diferencia entre ellas (fig. 5.1); la mayor de las diferencias entre ambas curvas está indicada en la figura.

¿Qué significa efectuar una prueba de significación en un problema de esta clase?

Dado que existen tales diferencias entre ambas distribuciones, debiéramos preguntarnos si son o no reales. El empleo de una prueba de significación para resolver la cuestión *¿son diferentes entre sí las dos distribuciones?* presupone reescribir la misma cuestión del siguiente modo: ¿proceden los dos conjuntos de datos muestrales de poblaciones idénticas? Tal y como ya hemos visto, la inferencia estadística pretende hacer inferencias acerca de las poblaciones basándose en muestras. Pero ¿de qué población son muestras nuestros datos? Esta cuestión nos lleva al supuesto clave de la prueba de Kolmogorov-Smirnov, y de todas las pruebas de significación que aún no hemos mencionado. Presupo-

nen que tenemos a nuestra disposición una muestra de una población y que hay independencia en la selección de esa muestra; en otras palabras, que la selección de un individuo no afecta a la selección de otro. El método usualmente empleado para cumplir estas especificaciones es el *muestreo aleatorio*. Una muestra aleatoria tiene la propiedad no sólo de dar a cada individuo una oportunidad igual de ser seleccionado, sino también de dar a cada combinación de individuos la misma oportunidad de selección.

Es obvio que ninguna muestra arqueológica puede ser considerada una muestra aleatoria de lo que existió en el pasado. Es cierto, sin embargo, que a veces los arqueólogos pueden elegir muestras aleatorias del registro arqueológico, cuyos problemas serán tratados en el capítulo correspondiente. Por otro lado, en la mayoría de los casos, y ciertamente en nuestro ejemplo, nos vemos obligados a operar con datos que no han sido recogidos de esa forma. Si el cementerio hipotético del cual deriva nuestro ejemplo hubiera sido totalmente excavado, sí que estaríamos trabajando, en cierto sentido al menos, con una población y no con una muestra. Ahora bien, ¿de dónde procedería entonces la variación muestral, que se supone está provocada por las diferencias entre las muestras?

Quizás fuese conveniente señalar que este no es un problema exclusivo de la arqueología, sino que ocurre en la mayoría de aquellas ciencias sociales en las que se emplea la estadística, incluyendo la geografía y la sociología; en ninguno de esos casos parece estar bien resuelto. Para algunos, la línea de razonamiento que acabamos de indicar conduce a la conclusión de que los procedimientos clásicos de inferencia estadística, por ejemplo, las pruebas de significación, son pura y simplemente irrelevantes, exceptuando ciertas situaciones muy limitadas. Otros, cuyo punto de vista comparto, sugieren que, en muchas circunstancias, es posible postular una población hipotética o ideal de la cual nuestros datos constituyen una muestra. Esto puede parecer un argumento bastante dudoso, pues invoca una población «que difícilmente puede definirse como la población que necesitamos para que la muestra sea considerada aleatoria», como Orton (1982) ha argumentado en un contexto algo distinto. No obstante, la cuestión no es, necesariamente, tan negativa como parece.

Hay dos formas ligeramente distintas de considerar la creación de esas poblaciones. Una de ellas procede del concepto de *aleatorización*. Puede ilustrarse por medio del empleo de la distribución de edad en los grupos de tumbas. Damos por hecho que hay 76 tumbas «ricas» y 136 «pobres». También damos por supuesto que hay un total de 6 + 23 = 29 tumbas en la categoría infantil I; 8 + 21 = 29 tumbas en la categoría infantil II, y así sucesivamente. Esta información es definitiva para el conjunto de datos a nuestra disposición; lo que se pretende averiguar es la forma en que las tumbas en cada una de las categorías de edad se distribuyen en las categorías de riqueza. ¿Acaso la distribución de las edades difiere en ambas categorías de riqueza? Si las distribuciones no son distintas, entonces la proporción de individuos «ricos» e individuos «pobres» en cada una de las categorías de edad es la misma, como si la población

constituyese un todo homogéneo. De este modo, en la categoría infantil I, basándonos en la proporción 76:136, habría 10 tumbas en la categoría rica y 19 en la pobre. Imaginemos que hemos efectuado numerosos experimentos en los que hemos asignado aleatoriamente los diversos enterramientos a las categorías «rico» y «pobre», de acuerdo con la proporción 76:136; si cada vez que lo hacemos trazamos la curva acumulativa de las dos distribuciones y anotamos la mayor de las diferencias entre ellas, después de muchos de esos experimentos dispondremos de una buena cantidad de diferencias máximas, experimentalmente producidas, con las que comparar las diferencias en nuestras dos distribuciones auténticas. Con ese fundamento, decidiremos si son o no inhabituales. Si lo son, decidiremos rechazar la hipótesis nula y afirmaremos la existencia de una diferencia «real» entre las distribuciones.

Esta es una manera de generar una población relacionada con nuestra muestra. En algunas circunstancias, sin embargo, es necesario producir la población aleatoriamente, usando métodos informáticos; en otras, como en el ejemplo en el que hemos estado trabajando, el uso de la prueba puede ser considerado como equivalente al proceso de aleatorización.

Posiblemente sea la aleatorización la forma más directa de conceptualizar, e incluso de efectuar, la idea de una población hipotética. Otra manera de hacerlo consiste en concebir la evidencia arqueológica como un resultado empírico específico de un sistema de conducta, basado en reglas sociales. Cualquier acción o ejemplo de una conducta que produzca evidencias arqueológicas estará basado en reglas, si bien se verá afectado por circunstancias contingentes de todo tipo, por lo que la variación introducida será, de hecho, aleatoria o inducida por el azar, no relacionada sistémicamente con las reglas que han dado lugar a esa conducta. Este tipo de razonamiento ha sido desarrollado no en relación con las pruebas de significación como tales, sino en la identificación y definición de tipos arqueológicos, por parte de una escuela que cree que definiendo los tipos descubrirá sus equivalentes en la mente de aquellos que los fabricaron, estableciendo así las reglas según las cuales fueron producidos (véase la discusión en Whallon y Brown, 1982).

Toda esta argumentación es mucho más apropiada en unas circunstancias arqueológicas que en otras; en particular, en los casos en los que sabemos que estudiamos los resultados de una conducta intencional, y que nuestras observaciones no están desviadas por los factores de recuperación del material. Para aplicar este enfoque en el ejemplo de las tumbas hemos de tener presentes dos limitaciones. Primero, los resultados sólo son aplicables a las tumbas de esta necrópolis en particular; no toman en consideración el hecho de que individuos «ricos» y «pobres», en cualquiera o en todas las categorías de edad, hayan sido enterrados en otro sitio, o bien que no hayan sido enterrados de una manera reconocible arqueológicamente. Segundo, habría que mostrar que cualquier relación entre categorías de riqueza y distribución de edades no es el resultado de otro factor, por ejemplo la variación en las condiciones de conservación de los restos. Dadas estas estipulaciones se puede afirmar que cualquier estructura relacional subyacente a la relación entre la edad del individuo en el momento de su muerte y la categoría de riqueza de su tumba fue el resultado de una conducta socialmente regulada, pero que por muchas razones habrá habido variación en el grado con el que esas reglas se siguieron, produciendo una distribución de la conducta con un valor medio y variación a su alrededor, tal y como Barth, por ejemplo, ha afirmado (1967). Allí donde esa conducta tenga una consecuencia arqueológica, el conjunto particular de conductas cuya evidencia recogemos será, precisamente, uno de entre un abanico de posibilidades.

Los argumentos que acabamos de presentar son bastante complejos y, en muchos sentidos, más profundos de lo que su esquemática presentación en este libro permite suponer. No obstante, es algo muy importante, pues se sitúa en el centro de las discusiones acerca de lo apropiado de las inferencias estadísticas en arqueología, al igual que en otras disciplinas, y no puede ser, por tanto, olvidado.

Completando la prueba de Kolmogorov-Smirnov

Tras esta larga, pero importante digresión, vamos a finalizar el ejemplo de prueba de significación que hemos estado usando. La mayor de las diferencias entre las curvas acumulativas de edad y «riqueza» era de 0,178, para la categoría juvenil. La pregunta, ahora, es: basándonos en la hipótesis nula, según la cual ambas curvas representarían muestras procedentes de idénticas poblaciones y cuyas diferencias de deberían a variaciones al azar, ¿es la diferencia observada inhabitualmente grande? Para saberlo, comparamos la diferencia observada con la distribución esperada de las diferencias derivadas teóricamente. Estas últimas distribuciones suelen presentarse bajo el aspecto de tablas estadísticas; veremos un ejemplo en el próximo capítulo. La prueba de Kolmogorov-Smirnov es ligeramente distinta, ya que la diferencia mínima entre dos distribuciones acumuladas, significativa a un nivel específico, es obtenida por medio de la evaluación de una fórmula. Si la diferencia observada es igual o mayor que ésta, entonces será significativa estadísticamente en el nivel especificado. En este caso hemos fijado el nivel de significación en 0,05 y la fórmula es:

$$1{,}36\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

donde n_1 = cantidad de individuos en la muestra 1, y n_2 = cantidad de individuos en la muestra 2. Aquí:

$$1{,}36\sqrt{\frac{76 + 136}{76 \times 136}} = 0{,}195$$

1,36 es un factor de multiplicación derivado teóricamente, y apropiado para el nivel 0,05. Si se requiere un nivel de significación de 0,01, el coeficiente será 1,63; si se requiere el 0,001, será 1,95.

Habiendo obtenido la diferencia mínima requerida para que H_0 sea rechazada en el nivel de significación especificado, apreciamos que la diferencia máxima observada ($Dmax_{obs}$) en 0,178 no es tan grande como la diferencia mínima requerida para el nivel 0,05 ($Dmax_{0,05}$), en 0,195. Como la diferencia observada es menor que el mínimo requerido para rechazar H_0 en el nivel 0,05, no podemos refutar la hipótesis nula. No hay una diferencia significativa en la distribución de edad entre las categorías «ricas» y «pobres».

Es importante apreciar que esto no significa *que las dos distribuciones sean la misma*. Simplemente significa que la evidencia es insuficiente para sugerir que son distintas; parece que no hay «lugar para la respuesta». No podemos estar seguros de que no haya algo aquí digno de ser estudiado.

Otras pruebas de significación para las diferencias entre dos escalas ordinales

La principal intención al describir la prueba de Kolmogorov-Smirnov en este capítulo fue proporcionar una idea de lo que implica la inferencia estadística. Pero esta no es la única de las técnicas que se pueden emplear para comprobar si dos variables expresadas en una escala ordinal son distintas una de la otra. La *prueba de Mann-Whitney* y la *prueba de las series* [*runs test*, en la bibliografía anglosajona] también pueden usarse en estas circunstancias. No se van a describir aquí sus detalles, que pueden encontrarase en alguno de los numerosos manuales clásicos (por ejemplo, Blalock, 1972). Así y todo, los presentaremos esquemáticamente, para dar una idea de cuándo son apropiados.

Imaginemos que tenemos datos acerca del tamaño de 20 asentamientos del neolítico inicial en dos tipos distintos de suelos en el sur de Italia. La estimación del tamaño del asentamiento a través de una prospección suele ser bastante problemática y, muchas veces, imprecisa. En este caso no consideraremos justificado precisar el área de cada yacimiento, sino que nos limitaremos a ordenarlos por tamaños. ¿Tienden a asociarse los yacimientos menores o los mayores con algún tipo de suelo en particular?

Podemos ordenar los yacimientos por tamaños (de mayor a menor, izquierda a derecha), indicando, para cada uno, el tipo de suelo (A o B) en el que está ubicado. La ordenación resultante es:

AABABBBABAAABBABBAAB

o bien:

BBBBBABBBABBAAAAAAAA

En el primer caso hay series muy cortas de yacimientos en un mismo tipo de suelo, mezcladas con series cortas de yacimientos en el otro tipo; los asentamientos en los dos tipos de suelo están mezclados debido a su tamaño. En el segundo caso, el número de series es mucho menor: hay un predominio de yacimientos en el suelo B en un extremo y en el suelo A en el otro extremo. La prueba de las series nos explica si la cantidad de series o secuencias que tenemos en un caso particular es mayor o menor de lo que esperaríamos si las dos distribuciones estuviesen mezcladas al azar.

La prueba de Mann-Whitney es muy similar, y para ilustrarla usaremos el mismo ejemplo. Volvemos a ordenar los yacimientos por tamaños. Ahora, podemos disponer todos los asentamientos en uno de los tipos de suelo; importa poco cuál, ya que hay diez en cada uno, pero si la cantidad de yacimientos en los dos tipos de suelo no estuviese equilibrada, habría que escoger el tipo con una menor cantidad de frecuencias. Así, anotamos para cada uno de nuestros yacimientos en el suelo tipo B, por ejemplo, cuántos de los yacimientos en el tipo de suelo A tienen un tamaño inferior a él. Para la primera de las dos secuencias anteriores, se observa que el más extenso de los yacimientos en el suelo tipo B se sitúa en el tercer puesto de la serie, con 8 yacimientos en el suelo del tipo A, por debajo. El siguiente en tamaño está en quinta posición, con siete yacimientos en el suelo tipo A, por debajo, y así sucesivamente. En la segunda de las series, cada uno de los cinco primeros yacimientos en B, tienen diez yacimientos en el suelo A, por debajo. Si, en general, los yacimientos en B son menores que los de A, habría muy pocos por debajo; si fuesen mayores, muchos; mientras que si los tamaños estuviesen mezclados al azar, la cantidad de yacimientos en A por debajo de los situados en B se situaría en algún lugar del centro de la secuencia. Usaremos la prueba de Mann-Whitney para descubrir si las cantidades en nuestro caso particular se diferencian significativamente de una mezcla aleatoria.

Conclusión

Este capítulo tan sólo ha intentado abordar el problema de las pruebas de significación en la inferencia estadística; el tema de la estimación será tratado en el capítulo dedicado al muestreo. El propósito ha sido discutir algunas de las cuestiones que implican esas pruebas, en el contexto de un ejemplo específico. Tal y como ya he insistido, las pruebas de significación no son ni la única ni la más importante razón para usar métodos cuantitativos en arqueología. De hecho, como veremos en los capítulos siguientes, la significación estadística y la significación sustantiva en términos arqueológicos no son necesariamente lo mismo: las preguntas acerca de la intensidad y la forma de las relaciones entre variables suelen ser mucho más interesantes e importantes que los problemas sobre su significación estadística.

EJERCICIOS

5.1. A un lado de un asentamiento prehistórico hay una necrópolis de tumbas megalíticas. Entre otros aspectos, nos interesa la significación de la distribución espacial de las tumbas. Se supone que la proximidad de las tumbas al asentamiento es relevante en algún sentido. Las tumbas y sus contenidos varían en distintos aspectos; en particular, se ha visto que es posible dividirlas según su morfología en «complejas» y «simples». La necrópolis ha sido dividida, analíticamente, en varias bandas de unos 200 m de ancho, con alguna variación como resultado de la topografía local. La banda A es la más cercana al asentamiento, aumentando la distancia al mismo hasta la banda F, que es la más alejada. Dada la información siguiente:

Banda	Cantidad de tumbas complejas	simples
A	12	6
B	8	6
C	17	10
D	7	16
E	13	19
F	14	18

¿Hay indicios que permitan concluir que la distancia al asentamiento y el grado de complejidad morfológica de la tumba estén relacionados? Razona los supuestos en los que se basa tu análisis.

5.2. En un estudio de la organización social de un hipotético yacimiento del período formativo en México se lleva a cabo una investigación acerca de los enterramientos. Algunos de ellos son fosas ordinarias, mientras que otros aparecen en tumbas construidas. La pregunta es si algunas características biológicas de los individuos están relacionadas con las diferencias en el modo de enterramiento.

Se adjunta información acerca de la cantidad de individuos en cada una de las series de categorías de edad, divididas según si han sido enterradas en fosas ordinarias o en tumbas construidas. ¿Es distinta la distribución de edad de las poblaciones enterradas en los dos tipos de enterramiento?

	Categoría de edad 1	2	3	4	5	6
Fosas	25	18	29	14	24	9
Tumbas	8	4	6	18	40	5

5.3. Se adjunta información acerca de la longitud de los fragmentos de hueso procedentes de dos cuevas del pleistoceno en la Inglaterra meridional (según Boyle, 1983). ¿Crees que la distribución de la longitud de los fragmentos en los dos yacimientos es distinta?

Categoría de longitud (mm)	Cantidad de fragmentos cueva 1	cueva 2
0-9	1	0
10-19	21	6
20-29	15	11
30-39	5	11
40-49	7	6
50-59	1	6
60-69	2	6
70-79	3	4
80-89	3	2
90-99	0	5
100-109	2	9

6. LA PRUEBA DE χ^2

En el capítulo anterior usamos la prueba de Kolmogorov-Smirnov para mostrar las implicaciones de una prueba de significación y sus supuestos fundamentales. Aunque es una prueba muy útil, hay una restricción para su empleo: el nivel de medida ha de ser ordinal, o más alto. La prueba o test de χ^2 (ji-cuadrado) no tiene esta restricción. Puede utilizarse con datos medidos en una escala nominal, es decir, simplemente clasificados en categorías; es fácil de calcular, si bien hoy en día este punto es menos importante, gracias a la amplia disponibilidad de calculadoras y ordenadores. A causa de esta falta de restricciones, la prueba de χ^2 puede usarse para afirmar la correspondencia entre distribuciones en una gran variedad de situaciones distintas, y como resultado se aplica muy a menudo. Presentaremos aquí esta prueba por varias razones; primero, porque está muy difundida y ha demostrado su utilidad; segundo, porque proporciona un nuevo ejemplo de cómo utilizar las pruebas de significación en contextos arqueológicos, y tercero, porque proporciona un puente muy conveniente entre los conceptos de significación estadística y los de *intensidad* de las relaciones entre variables.

Hay dos versiones ligeramente distintas, si bien el principio es el mismo. La primera, quizás menos familiar a los arqueólogos, es la prueba unimuestral, en la que se compara una muestra a una población especificada teóricamente; la prueba establece el grado de «ajuste» o correspondencia entre esas dos distribuciones. Esta idea es bastante importante a la hora de contrastar modelos teóricos.

Para describir la prueba, lo mejor será empezar con un ejemplo. Una cuestión que suele ser interesante es la distribución del asentamiento según las diferencias de suelo; ¿acaso unas áreas eran más aptas que otras para el asentamiento inicial? Supongamos un área de Francia oriental con tres tipos de suelo: rendzina, aluvión y tierra marrón. Hay 53 asentamientos del neolítico final en el área, y un vistazo al mapa sugiere la posibilidad de una preferencia por las rendzinas. La pregunta es si es posible o no que la distribución de asentamientos según el suelo sea aleatoria. Si los tres tipos de suelo fuesen igualmente atractivos para el asentamiento, sería razonable asumir que encontraríamos aproximadamente la misma densidad de yacimientos en cada uno. Dicho de otro modo, la distribución de asentamientos estaría más o menos equilibrada en la región,

y las variaciones no se relacionarían con el suelo, sino con factores tales como pequeñas diferencias topográficas locales, o los caprichos de las comunidades humanas que allí se asentaron por vez primera. En este contexto, podemos usar la prueba de χ^2.

TABLA 6.1. Cantidad de asentamientos del neolítico final en distintos tipos de suelos de Francia oriental.

Tipo de suelo	Cantidad de asentamientos
Rendzina	26
Aluvión	9
Tierra marrón	18
	53

Lo primero que hay que hacer es anotar la cantidad de yacimientos en cada uno de los tipos de suelos (tabla 6.1). ¿Cómo se calculan las frecuencias esperadas, derivadas teóricamente, a las que hemos de comparar las frecuencias observadas? Ya hemos visto que, si postulamos que las tres zonas eran igualmente atractivas para el asentamiento, tendríamos que esperar la misma densidad de asentamiento en cada una de ellas. Esto equivale a nuestra hipótesis nula derivada teóricamente, para calcular las frecuencias esperadas. Así, es razonable asumir que si las rendzinas constituyen el 32 % del área, tal y como suponíamos, el 32 % de los asentamientos se encontrarán en ese tipo de suelo; los mismo sucede si suponemos que el 43 % del área es de tierra marrón y el 25 % de aluvión. En otras palabras, calculamos la cantidad de asentamientos esperado para cada tipo de suelo, asignando la misma proporción de la cantidad total de asentamientos que el porcentaje del área total que ocupa ese tipo de suelo (tabla 6.2).

Si comparamos los valores observados y esperados en este cuadro para la cantidad de asentamientos en cada tipo de suelo, hay algunas diferencias obvias entre la distribución anticipada si todas las áreas fuesen idénticas en cuanto a densidad del asentamiento, y lo que nosotros podemos observar. La cuestión es: ¿son esas diferencias tan grandes que la probabilidad de que sean un resultado de la variación al azar es lo suficientemente baja? Es aquí donde la prueba de χ^2 demuestra su utilidad.

TABLA 6.2. Cantidades observadas y esperadas de asentamientos del neolítico final en Francia oriental.

Tipo de suelo	Cantidad observada de asentamientos	% del área	Cantidad calculada de asentamientos
Rendzina	26	32	17,0
Aluvión	9	25	13,2
Tierra marrón	18	43	22,8
	53	100	53,0

La prueba de χ^2 unimuestral presupone un conjunto de observaciones dividido en varias categorías mutuamente exclusivas. Se compara la distribución de observaciones en las categorías con una distribución que proceda de ciertas esperanzas teóricamente derivadas, especificadas por la hipótesis nula. Se registran las diferencias entre las dos distribuciones para cada categoría, y se calcula el valor χ^2, basado en la suma de las diferencias. Este último valor es comparado a su vez con el valor mínimo requerido para rechazar la hipótesis nula en el nivel de significación especificado. En efecto, por medio de una prueba de significación nos preguntamos si las observaciones pueden ser una muestra aleatoria de una población que tuviese las características especificadas en la hipótesis nula.

Realizar la prueba exige algunos supuestos previos. Como siempre, es necesario especificar una hipótesis nula y establecer un nivel de significación, y ser capaces de especificar una población de la cual nuestras observaciones sean una muestra, del modo descrito en el capítulo anterior. Tal y como hemos indicado, el nivel de medida requerido no es muy exigente, una simple escala nominal con dos categorías mutuamente excluyentes como mínimo; las observaciones han de ser *cantidades brutas*, y no porcentajes u otras formas de proporción.

La fórmula para χ^2 es:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

donde k es la cantidad de categorías; O_i es la cantidad de casos observados en cada categoría i; E_i es la cantidad de casos esperados en la categoría i; y χ^2 es el símbolo que representa a la prueba, usando la letra griega «ji».

Esta fórmula puede ser interpretada en palabras de la forma siguiente: para cada categoría se restan los valores esperados de los observados, se eleva al cuadrado la diferencia y se divide el resultado por el valor esperado; una vez hecho esto para cada categoría, se suman los resultados de todas las categorías. El resultado será el valor calculado del χ^2.

Una vez calculado éste, habremos de comprobar su significación estadística. En el caso de la prueba de Kolmogorov-Smirnov esto se hacía mediante la comparación de la mayor de las diferencias observadas, con un valor obtenido por la sustitución de las cantidades del tamaño de la muestra en una fórmula apropiada, dado el tamaño de la diferencia requerido para que el resultado fuese significativo. En el caso de χ^2, se han confeccionado unas tablas que proporcionan los valores a los que comparar los calculados por medio de la fórmula (véase anexo I, tabla A). Para encontrar el valor relevante en la tabla y hacer efectiva la comparación, es necesario saber dos cosas: el nivel de significación decidido —lo suficientemente claro— y la cantidad de *grados de libertad* asociados con la muestra.

El concepto de grados de libertad no es nada fácil. Esencialmente, la forma

de la distribución teórica del χ^2, tabulada en la tabla a la que acabamos de referirnos, varía de acuerdo con el número de categorías que dividen las observaciones. Cuanto mayor es la cantidad de categorías, mayor habrá de ser el valor del estadígrafo χ^2, obtenido para las observaciones. Esto tiene sentido, ya que el hecho de que el número de cantidades que se suman depende del número de categorías se desprende del propio enunciado de la fórmula, así que, cuantas más categorías haya, mayor será la suma, es decir, el valor calculado de χ^2. En el caso de la prueba unimuestral, sin embargo, la cantidad de grados de libertad no es igual al número de categorías, sino al número de categorías menos uno; en símbolos:

$$\nu = k - 1$$

donde ν (letra griega «ny») es el número de grados de libertad y k el número de categorías.

¿Por qué? Este comportamiento de la prueba se puede ilustrar bien por medio de nuestro ejemplo, en donde hay 53 observaciones (asentamientos), divididos en tres categorías (tipos de suelos). Ya que hay un total de 53 observaciones, y que $26 + 9 = 35$ están en las dos primeras categorías, el valor de la tercera *ha de ser* $53 - 35 = 18$. En otras palabras, los valores de las dos primeras categorías pueden variar libremente, pero no el valor de la tercera categoría, que está fijado por el requisito de que la suma en las tres categorías ha de ser igual al número total de observaciones con el que hemos empezado.

Cuando se conoce el número relevante de grados de libertad y el nivel de significación, es posible encontrar el valor apropiado en la tabla con el que comparar el valor calculado. En una tabla de χ^2, el número de grados de libertad se encuentra en el lado izquierdo y el nivel de significación en la parte superior. Si, por ejemplo, tenemos dos grados de libertad y usamos un nivel de significación de 0,05, encontraremos la fila para $\nu = 2$, y la seguiremos hasta cruzar con la columna para el nivel de significación 0,05; el número en la intersección es 5,99. Este es el valor χ^2 tabulado que compararemos con el calculado:

$$\text{si } \chi^2_{\text{calc}} \geqslant \chi^2_{\alpha}, \text{ rechazamos } H_0;$$
$$\text{si } \chi^2_{\text{calc}} \leqslant \chi^2_{\alpha}, \text{ aceptamos } H_0.$$

Antes de volver a nuestro ejemplo, sin embargo, hay que indicar ciertos puntos. Los valores tabulados χ^2 sólo proporcionan el nivel de significación correcta para muestras a partir de cierto tamaño mínimo, si bien esta restricción no es muy severa. Si la prueba sólo tiene un grado de libertad, ninguna categoría debiera tener un valor *esperado* inferior a 5; con una mayor cantidad de categorías, esta restricción puede ser atenuada considerablemente. En los casos en los que aparece este problema hay varias formas de soslayarlo, por ejemplo por medio de una prueba de aleatorización.

Una vez que hemos descrito el procedimiento general para efectuar una prueba de χ^2, nos será posible mostrar su empleo en el ejemplo anteriormente esbozado. En primer lugar, hay que presentarlo en la forma apropiada para una prueba de significación:

H_0 = los asentamientos se distribuyen por igual en los tres tipos de suelo.
H_1 = los asentamientos no se distribuyen por igual en los tres tipos de suelo.
Nivel de significación: $\alpha = 0{,}05$.

No hay necesidad de ser muy conservador en la selección del nivel de significación. Nos interesa saber si hay o no huellas de divergencia en la igualdad de la distribución.

Los datos están medidos sólo en una escala nominal; son frecuencias divididas en categorías mutuamente excluyentes. Ninguno de los distintos valores esperados es inferior a 5. El uso de una prueba de χ^2 unimuestral es, por tanto, apropiado en este caso.

Los valores esperados bajo H_0 han sido calculados (tabla 6.2), por lo que es posible efectuar ahora los siguientes cálculos:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(26 - 17{,}0)^2}{17{,}0} + \frac{(9 - 13{,}2)^2}{13{,}2} + \frac{(18 - 22{,}8)^2}{22{,}8} =$$

$$= 4{,}76 + 1{,}34 + 1{,}01 = 7{,}11$$

El resultado se compara con el valor tabulado apropiado. Los grados de libertad son $k - 1$, donde k es el número de categorías: aquí $3 - 1 = 2$. En la tabla, el valor crítico para dos grados de libertad y un nivel de significación de 0,05 es 5,99. Como $\chi^2_{calc} \geqslant \chi^2_{\alpha}$, rechazamos H_0, y aquí $7{,}11 > 5{,}99$, se rechaza la hipótesis nula en este caso.

Pero es importante no detenerse en este punto. Es preciso relacionar este resultado con el problema arqueológico. En este caso, hemos de aceptar la hipótesis alternativa, según la cual los asentamientos no están distribuidos por igual. Para poner en relación nuestros datos con una población en la forma discutida en el capítulo anterior, podemos decir que, si efectuamos numerosos experimentos asignando aleatoriamente 53 asentamientos a esos tres tipos de suelos, bajo el supuesto de una distribución idéntica, la distribución que observemos será bastante inhabitual, y nos veremos en la obligación de rechazarla en un nivel de significación de 0,05. Puede haber motivos para hacer esto, por lo que consideraremos más adelante los problemas que supuso el ir desde la apreciación de unas asociaciones y correlaciones estadísticamente significativas hasta las inferencias sobre la causalidad.

LA PRUEBA DE χ^2 PARA DATOS EN CLASIFICACIONES CRUZADAS

Una vez visto el caso en el que se compara una muestra a una población teórica específica, podemos volver a fijarnos en el uso de χ^2 para probar la independencia de una clasificación en casos en los que los datos han sido clasificados según dos criterios distintos. Como siempre, empezaremos con un ejemplo.

Supongamos que estamos estudiando una necrópolis de inhumación de la edad del hierro en el norte de Alemania, en la que sospechamos que hay una relación entre el sexo de un individuo y el lado de la tumba sobre el cual yace. Disponemos de la información que aparece en la tabla 6.3, denominada a menudo *tabla de contingencia*. Tablas como esta son 2 × 2 (dos por dos), ya que sólo hay dos filas —lado derecho y lado izquierdo— y dos columnas —masculino y femenino—. Las entradas individuales en la tabla —por ejemplo, las tumbas masculinas con el cadáver sobre el lado derecho— se denominan *celdas*. Las cantidades en el extremo de cada fila son los totales de fila, y en el pie de las columnas, los totales de las columnas. En la parte inferior derecha aparece el número total de observaciones, aquí 87.

TABLA 6.3. Lado sobre el cual están colocados los cadáveres en la tumba, tabulados por sexo. Necrópolis de inhumación de la edad del hierro en el norte de Alemania.

	M	F	Total
Der.	29	14	43
Izq.	11	33	44
Total	40	47	87

Básicamente, la prueba en este tipo de tablas es muy similar a la que acabamos de ver, en la que los datos eran frecuencias, divididas en categorías mutuamente excluyentes. Esta vez, sin embargo, en lugar de comparar la distribución de una muestra observada con una población especificada teóricamente, nos preguntaremos si dos clasificaciones distintas de los datos son independientes una de la otra, es decir, si la pertenencia a una categoría particular de una clasificación no está relacionada con la pertenencia a una categoría particular de la otra. Con todo, en ambos casos estaremos comprobando lo que los estadísticos denominan «bondad del ajuste».

El supuesto previo exigido en este caso es, nuevamente, muy similar al de la prueba unimuestral: escalas nominales, o bien un nivel de medida más alto; y sin ninguna frecuencia esperada menor que 5, en el caso de un grado de libertad (véase p. 85) acerca de los grados de libertad en tablas de contingencia). En este caso, no obstante, contamos con dos criterios de clasificación distintos, divididos en dos categorías mutuamente excluyentes, cuando menos. Refiriéndonos a nuestros ejemplos, mientras que en la prueba unimuestral los asentamientos estaban divididos de acuerdo a una sola variable, el tipo de suelo, en

la tabla de contingencia las tumbas están divididas o clasificadas según dos variables: el sexo y el lado sobre el cual yace el cadáver.

El cálculo del χ^2, como antes, se basa en la diferencia entre los valores observados y esperados para cada categoría. El número de categorías es el número de celdas en la tabla: en nuestro ejemplo hay dos categorías según el sexo, y dos categorías según el lado, por lo que el número de celdas en la tabla 6.3 será $2 \times 2 = 4$.

En la prueba unimuestral, los valores esperados eran generados por la población teórica postulada por la hipótesis nula. Aquí la idea es muy parecida, ya que nos preguntamos si las tumbas masculinas y femeninas tienen la misma división proporcional entre enterramientos sobre el lado izquierdo y enterramientos sobre el lado derecho. Así, si hay 43 enterramientos sobre el lado derecho y 44 sobre el izquierdo, esperaremos que las 47 tumbas femeninas y las 40 tumbas masculinas se dividan entre las categorías lado derecho y lado izquierdo, de acuerdo con la proporción 43 : 44. De hecho, mejor que efectuando esta operación, los valores esperados apropiados para una celda dada en la tabla pueden obtenerse multiplicando la suma de la columna correspondiente a la celda por la suma de la fila correspondiente a la celda, y dividiendo ese resultado por el número total de observaciones. De este modo, para la celda situada en la esquina superior izquierda, el valor esperado es $(40 \times 43)/87 = 19{,}8$.

Es posible obtener los restantes valores esperados de la misma manera. Sin embargo, dado que conocemos los totales marginales de la tabla y el valor esperado para la celda superior izquierda, también podemos conseguir los valores esperados de las demás por sustracción:

$$43 - 19{,}8 = 23{,}2$$
$$40 - 19{,}8 = 20{,}2$$
$$44 - 20{,}2 = 23{,}8$$

A continuación construimos una tabla incluyendo los valores esperados entre paréntesis (tabla 6.4), tras lo cual podremos especificar la prueba de significación para los datos de los enterramientos.

TABLA 6.4. Lado sobre el cual están colocados los cadáveres en la tumba, tabulados por sexo, con las frecuencias esperadas en cada categoría entre paréntesis.

	M	F	Total
Der.	29 (19,8)	14 (23,2)	43
Izq.	11 (20,2)	33 (23,8)	44
Total	40	47	87

H_0 = la distribución de tumbas masculinas y femeninas, según las categorías de postura del cadáver «sobre la izquierda» y «sobre la derecha», es la misma.

H_1 = la distribución de las tumbas masculinas y femeninas en las dos categorías es distinta.

Nivel de significación: $\alpha = 0{,}05$

Los datos cumplen con los requisitos de toda prueba de χ^2 para datos en clasificaciones cruzadas, por lo que el paso siguiente será calcular el valor χ^2 para los datos, usando la fórmula anterior (tabla 6.5).

TABLA 6.5. Tabla de cálculos para obtener el valor χ^2 a partir de los datos en la tabla 6.4.

Categoría	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\dfrac{(O_i - E_i)^2}{E_i}$
1	29	19,8	9,2	84,64	4,27
2	14	23,2	−9,2	84,64	3,65
3	11	20,2	−9,2	84,64	4,19
4	33	23,8	9,2	84,64	3,56
					$\chi^2 = 15{,}67$

El proceso de comprobar la significación de esta prueba es igual al de antes, pues se compara el valor calculado con el que aparece en la tabla χ^2 correspondiente al nivel de significación requerido y al número apropiado de grados de libertad. Para la prueba con datos en clasificaciones cruzadas, sin embargo, el número de grados de libertad se obtiene de manera distinta, definido por ν = (el número de filas en la tabla -1) (el número de columnas en la tabla -1). En nuestro caso tenemos $(2 - 1)(2 - 1) = 1$.

Este hecho puede relacionarse con la manera de calcular las frecuencias esperadas (por sustracción) una vez conocida la de la celda superior izquierda. Si miramos ahora los valores tabulados del χ^2 para un grado de libertad y el nivel de significación 0,05, encontramos que es 3,84. Como $\chi^2_{calc} = 15{,}67$ y $15{,}67 > 3{,}84$, rechazaremos H_0. Podemos señalar, incidentalmente, que un valor de 15,67 sería significativo incluso en el nivel 0,01. Por lo tanto, las tumbas masculinas y femeninas no están distribuidas de la misma manera en las dos categorías de postura del cadáver.

Una nota final acerca del cálculo. El método que acabamos de describir es el procedimiento general para calcular el χ^2, sea el que sea el número final de columnas en la tabla. De hecho, en el caso de una tabla 2×2, una tabla con 2 filas y 2 columnas, hay una fórmula alternativa más conveniente:

$$\chi^2 = \frac{n\,(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

en donde n es el tamaño de la muestra, y a, b, c, d se refieren a las celdas de una tabla, etiquetadas del siguiente modo:

a	b
c	d

¿CUÁL ES LA UTILIDAD DEL χ^2?

De todo lo dicho acerca de la prueba de χ^2, debiera quedar claro que se trata de una técnica muy útil e informativa, si bien sería conveniente acabar este capítulo con una lista de sus limitaciones:

1) El χ^2 no nos explica nada acerca de la intensidad de una relación; simplemente, explica la probabilidad de que una tal relación exista o no. Este punto será ampliado en el próximo capítulo. Por el momento, podemos indicar que, incluso si la conexión entre dos variables es débil, obtendremos un resultado estadísticamente significativo. Siguiendo el ejemplo anterior, obtendremos un resultado significativo incluso si hubiese tan sólo una ligera tendencia a que los hombres fuesen enterrados sobre el lado derecho y las mujeres sobre el izquierdo. Aunque ligera, esa tendencia sería bien real (véase el tercer punto).

2) No nos dice nada acerca de la forma en que las variables están relacionadas; simplemente mide la distancia entre los valores esperados y los observados.

3) Al igual que en otras pruebas estadísticas, el tamaño de la muestra afecta a la magnitud del χ^2. Para una diferencia dada con respecto a la independencia, su tamaño es proporcional al tamaño de la muestra; esto significa que siempre podemos obtener en la práctica una relación significativa, haciendo que la muestra sea lo suficientemente grande. La dificultad, entonces, radica en la distinción entre *significación estadística* y *significación sustantiva*.

Lo que hemos de hacer cuando disponemos de datos presentados en forma de una clasificación cruzada es examinarlos en detalle; el simple cálculo de una prueba de χ^2 no es suficiente; muchas veces, incluso, se sabe antes de tiempo que el resultado de χ^2 va a ser significativo, mas confiando sólo en esa impresión ganaríamos muy poca información.

Va a ser preciso que desarrollemos ampliamente estos puntos.

EJERCICIOS

6.1. En la excavación de una necrópolis de inhumación, el 35 % de los esqueletos tienen anillos de bronce como ajuar. Se ha encontrado un grupo de 15 tumbas separadas, pero adyacentes al resto, de las cuales el 10 % contienen anillos de bronce. ¿Difiere este grupo de tumbas del resto, a causa de la deposición de anillos?

6.2. La investigación y análisis del arte rupestre prehistórico necesita a menudo de la identificación de esquemas recurrentes de asociación entre motivos. En el caso que consideramos aquí hay 9 motivos diferentes. El 21,2 % de las frecuencias de aparición individuales de los motivos presentan superposiciones con otros motivos. 15 de 24 apariciones del motivo «oveja» y 13 de 127 apariciones del motivo «humano» muestran superposiciones. ¿Son diferentes esos esquemas de asociación de los de la población en conjunto?

6.3. En un análisis de la estructura espacial del intercambio local en Mesopotamia, se examina un tipo particular de cerámica producido en el centro (datos según Johnson, 1973). El examen sugiere que la anchura de las líneas pintadas usadas en la decoración es distinta según los dos centros de manufactura; el estudio del histograma de las anchuras sugiere que las líneas pueden dividirse en dos categorías: gruesas y finas. Los asentamientos en el área estudiada se dividen en dos grupos, oriental y occidental. ¿Está distribuida la cerámica con los dos tipos de líneas de forma diferente en la división este-oeste? Disponemos de la siguiente información:

	Área oriental	Área occidental	Total
Línea gruesa	42	10	52
Línea fina	17	21	38
Total	59	31	90

6.4. Estamos estudiando un cementerio en el que hay tres tipos distintos de tumbas: fosas simples, tumbas con cámara de madera y tumbas con cámara de piedra. Suponemos que puede haber una relación entre el tipo de tumba en el que los individuos fueron enterrados y la edad del individuo en el momento de la muerte; podemos definir tres categorías de edad por medio del examen antropológico de los esqueletos: menos de 21, 21-40, más de 40. Los individuos se distribuyen como indica el cuadro. ¿Hay alguna relación significativa entre la edad y el tipo de tumba?

	<21	21-40	>40
Fosa simple	23	19	11
Cámara de madera	12	17	13
Cámara de piedra	10	16	15

7. MÁS ALLÁ DEL χ^2: DESCRIPCIÓN DE LA ASOCIACIÓN ENTRE DOS VARIABLES EN LA ESCALA NOMINAL

Las limitaciones del χ^2, enumeradas al final del capítulo anterior, pueden ser consideradas, en realidad, limitaciones generales de las pruebas de significación: no nos conducen muy lejos en la comprensión del objetivo de nuestra investigación. Incluso cuando se usan sin precauciones, pueden inducir a error, ya que es posible que la significación estadística haya sido considerada más importante de lo que es. Este capítulo expondrá algunos métodos utilizables para extraer más información de la escala nominal, que no limitándonos al χ^2. Si bien las técnicas descritas se refieren específicamente a las tablas de contingencia, el enfoque subyacente puede ser generalizado a cualquier tipo de análisis y no es muy distinto del enfoque EDA discutido en un capítulo anterior; de hecho, el papel que desempeñan las pruebas de significación en el resto del libro es relativamente pequeño. La primera fase en la exposición del enfoque es ampliar los comentarios hechos al final del capítulo anterior.

Se sugirió que, para que una relación fuese significativa estadísticamente, no era necesario que su significación estuviese basada en su intensidad; es posible que una relación estadísticamente significativa sea muy débil. Esto se debe a que la significación estadística procede del efecto combinado de dos factores distintos: la intensidad de la relación y el tamaño de la muestra. Por consiguiente, *no podemos* usar el valor del χ^2 o su nivel de probabilidad asociado como una medida de la fuerza de una relación, y decir, por ejemplo, que un resultado significativo al nivel 0,001 indica una relación más fuerte que una significativa al nivel 0,05.

El efecto del tamaño de la muestra en el valor χ^2 y el nivel de significación pueden ilustrarse de nuevo con el mismo ejemplo funerario del capítulo precedente, ligeramente cambiado, de forma que todos los números sean pares (tabla 7.1.) Aquí, el $\chi^2 = 18{,}33$ con un grado de libertad es significativo mucho más allá del nivel 0,001. Si dividimos los números por la mitad, manteniendo la misma distribución proporcional en las categorías, obtendremos la tabla 7.2. Aquí el $\chi^2 = 9{,}16$, con un grado de libertad, es significativo al nivel 0,01.

TABLA 7.1. Lado sobre el cual yace el individuo en la tumba, clasificado según el sexo.

	M	F	Total
Der.	30	14	44
Izq.	10	34	44
Total	40	48	88

TABLA 7.2. Lado sobre el cual yace el individuo en la tumba, clasificado según el sexo. Las cantidades son la mitad de las que aparecen en la tabla 7.1.

	M	F	Total
Der.	15	7	22
Izq.	5	17	22
Total	20	24	44

Igualmente, si multiplicásemos por dos las cifras originales, se obtendría un $\chi^2 = 36{,}66$. Así pues, en general, si mantenemos constantes las proporciones entre las celdas y nos limitamos a multiplicar los números por un factor k, multiplicaremos los χ^2 resultantes por k.

Todo esto tiene sentido. Si nos planteamos la pregunta clásica de una prueba de significación —¿existe o no una relación?—, tendremos más confianza en nuestra respuesta si está basada en un gran número de observaciones. Si el número de observaciones es muy grande, entonces, incluso si entre las variables sólo existe una débil relación, o una pequeña diferencia entre las muestras, llegaremos a la conclusión de que ésta es «real». Si el número de observaciones es muy pequeño, entonces, para cualquier diferencia o relación considerada «real», ésta habrá de estar muy marcada. Tales diferencias bien marcadas o relaciones fuertes son casi siempre de gran interés para nosotros, si bien lo mismo no es necesariamente cierto para las débiles: una relación o diferencia muy débil puede llegar a ser «real», pero ¿tiene algún significado?

La discusión anterior pone de manifiesto la necesidad de medir la intensidad de una relación separadamente de su significación estadística, ya que la prueba del χ^2 no es una forma apropiada de hacerlo, a no ser en aquellos casos raros en los que nuestro objetivo implique, simplemente, el hacer comparaciones entre muestras de idéntico tamaño.

Esta cuestión acerca de la comparación es importante. Generalmente, no estamos interesados en medir la intensidad de la relación en un único caso particular. Lo más frecuente es hacer comparaciones, por ejemplo, con la misma medida en otros conjuntos y en otros datos. Por esta razón, tales medidas han de estar estandarizadas. Es conveniente también tener unos límites superior e inferior bien definidos, convencionalmente 1,0 para el superior y 0 o bien $-1{,}0$ para

el inferior. La mayoría de las medidas adoptan un valor de 1,0 o bien −1,0 cuando la relación es perfecta, y un valor de 0 cuando no hay relación entre las variables.

Dado que χ^2 depende del tamaño de la muestra, resulta obvio dividir el valor de χ^2 por n, el número de observaciones en la muestra; esto significa que obtendremos los mismos resultados siempre que las proporciones entre las celdas sean las mismas, y sea cuál sea la cantidad de observaciones. El coeficiente obtenido dividiendo χ^2 por n es conocido como φ^2 (fi al cuadrado); su valor es 0 cuando no hay relación entre las dos variables. En tablas 2 × 2 (o 2 × k), φ^2 tiene un límite superior de 1,0, que se alcanza cuando la relación entre ambas variables es perfecta, tal y como muestra la tabla 7.3. En este caso, $\chi^2 = 100$ y $\varphi^2 = 100/100 = 1{,}0$.

TABLA 7.3. Un ejemplo de relación o asociación perfecta en una tabla 2 × 2.

	M	F	Total
Der.	50	0	50
Izq.	0	50	50
Total	50	50	100

En una tabla 2 × 2, siempre que dos celdas opuestas diagonalmente están vacías, el χ^2 de la tabla será igual al número de observaciones, y el φ^2 será por tanto 1,0; este comportamiento es denominado a veces *asociación absoluta*. Refiriéndonos a este caso peculiar, podemos decir que la variación en lo que respecta al lado sobre el que yacen los cadáveres en la tumba está explicada completamente por el sexo del cadáver, o bien que está asociada a su sexo.

Tal y como quedó dicho con anterioridad, el φ^2 alcanza un límite superior de 1,0 sólo cuando la tabla tiene dos filas y/o dos columnas. Esto se mantiene para una tabla de 2 filas y 20 columnas, o 2 columnas y 20 filas, pero no para una tabla de 3 × 20 o incluso 3 × 3. Para tablas en las que tanto el número de filas como el de columnas es mayor que 2, φ^2 tendrá un límite superior a 1,0. Para poder fijar ese límite para tablas grandes, φ^2 debe estar, a su vez, estandarizado. La más conocida de tales estandarizaciones es la V^2 de Cramer:

$$V^2 = \frac{\varphi^2}{\min\,(r-1,\ c-1)}$$

donde min ($r - 1$, $c - 1$) se refiere al (número de filas [*rows*] − 1) o (número de columnas − 1) que sea el menor. Esta ecuación adopta un valor máximo de 1,0 incluso cuando el número de filas y columnas no es igual, y para las tablas mayores que 2 × 2 o 2 × k; en estos dos últimos casos, V^2 se reduce obviamente a φ^2.

La Q de Yule es otra medida de asociación o relación usada con bastante frecuencia, aunque sólo es aplicable a tablas 2 × 2:

$$Q = \frac{ad - bc}{ad + bc}$$

donde *a, b, c, d* se refieren a las frecuencias en las celdas en una tabla etiquetada como sigue:

a	b
c	d

Imaginemos una tabla 2 × 2 en la que representamos la presencia/ausencia de algo, por ejemplo un tipo particular de objeto de ajuar en una tumba, frente a la presencia/ausencia de otra cosa, otro tipo de objeto de ajuar. La tabla adopta esta forma:

	+	−
+	+ + (*a*)	+ − (*b*)
−	− + (*c*)	− − (*d*)

La celda superior izquierda indica la presencia conjunta, la celda inferior derecha, la ausencia conjunta, y las otras dos los casos en los que uno está presente y el otro ausente. Las celdas *a* y *d* son los casos en los que los dos atributos *covarían* positivamente: cuando uno está presente, también lo está el otro, y viceversa. La multiplicación del número de casos de presencia conjunta (*a*) y ausencia conjunta (*d*) nos proporciona una medida de la *covariación positiva** entre los dos atributos. Por otro lado, multiplicando la cantidad de casos en que uno está presente y el otro ausente (*b*) y viceversa (*c*), nos proporciona una medida de la *covariación negativa* entre los dos atributos: el grado en el que la presencia de uno implica la ausencia del otro. Si cuando uno está presente, el otro puede estar presente o ausente, entonces no existe ninguna relación sistemática entre ambas. El ejemplo definitivo de la no relación es cuando *ad* (la covariación positiva) es igual a *bc* (la covariación negativa), por lo que $Q = 0$. De otro lado, Q tendrá un límite de $Q = +1{,}0$ para la covariación positiva perfecta o asociación y $Q = -1{,}0$ para la asociación negativa perfecta. Así pues, mientras que φ^2 sólo puede ser positivo, Q adopta también valores negativos. No obstante, la mayor diferencia entre las dos medidas radica en la forma en que tratan la asociación, lo cual quedará mejor explicado por medio de un ejemplo.

* No confundir con la covarianza, que es la covariación media. (*N. del t.*)

En las tablas 7.4 y 7.5, el valor de una de las celdas es 0. Es una consecuencia de la fórmula de Q el que adopte un valor de 1,0 en ambas, e incluso en cualquier tabla 2 × 2 con una entrada = 0. En este caso, podemos ver que refleja la asociación perfecta entre la categoría masculina y exactamente una de las dos categorías de posición —el lado derecho—. Por contraste, en la primera tabla, las tumbas femeninas están distribuidas equitativamente entre ambas posiciones, mientras que en la segunda tienden hacia la izquierda, el esquema opuesto al de los varones. En ningún caso las mujeres están asociadas exclusivamente con el lado izquierdo, que sería necesario para que φ^2 adoptase el valor 1,0, sino, naturalmente, φ^2 aumenta de la primera a la segunda tabla, a medida que la distribución de tumbas femeninas se hace más asimétrica.

TABLA 7.4. Comparación entre Q y φ^2, ejemplo 1. Aquí $Q = 1{,}0$ y $\varphi^2 = 0{,}375$.

	M	F	Total
Der.	60	20	80
Izq.	0	20	20
Total	60	40	100

TABLA 7.5. Comparación entre Q y φ^2, ejemplo 2. Aquí $Q = 1{,}0$ y $\varphi^2 = 0{,}643$

	M	F	Total
Der.	60	10	70
Izq.	0	30	30
Total	60	40	100

Q es un buen coeficiente para reconocer asociaciones bastante débiles, si bien una vez alcanza su límite superior o inferior no puede ir, evidentemente, más allá. Ha sido criticado, precisamente, porque no puede distinguir entre lo que se llama a veces «asociación completa», cuando una celda tiene un valor 0, y la «asociación absoluta», cuando dos celdas diagonalmente opuestas tienen el valor 0 y φ^2 alcanza su límite superior. Así y todo, Q puede ser muy útil si tenemos en cuenta esta circunstancia.

OTRAS MEDIDAS DE ASOCIACIÓN

Fi-cuadrado, V de Cramer y Q de Yule no son, en absoluto, las únicas medidas de asociación para variables medidas en una escala nominal. Hay otras que no serán descritas aquí en detalle. El propósito no es abarcarlo todo, sino presentar algunos de los coeficientes útiles en sí mismos y, lo más importante,

dar una idea de lo que implican las medidas de asociación. Una vez comprendida esa idea general, se pueden consultar otros coeficientes en manuales generales, como el de Blalock (1972).

El lector encontrará que muchos paquetes informáticos de estadística incluyen la tau (τ) y la lambda (λ) de Goodman y Kruskal. Ambos coeficientes utilizan la asociación entre variables para reducir el número de errores que cometemos al averiguar el valor de una variable a partir del valor de otra. Así, tomando en consideración los datos de la tabla 7.5, sabemos que hay 100 tumbas, 70 con enterramientos sobre el lado derecho y 30 sobre el lado izquierdo. Supongamos que hemos de predecir para cada enterramiento si está sobre el lado derecho o sobre el izquierdo. Si cogemos 70 tumbas y suponemos que están sobre el lado derecho y las 30 restantes sobre el izquierdo, cometeremos muchos errores. Si, por otro lado, conocemos el sexo del individuo enterrado la predicción mejorará considerablemente, ya que el sexo del individuo y el lado sobre el cual yace están relacionados. Sabiendo que una tumba en particular contiene un varón, podremos predecir que el cadáver está tendido sobre el lado derecho, pues no hay ningún enterramiento masculino con el cadáver tendido sobre el izquierdo. Si sabemos que la tumba contiene una mujer, lo mejor que podremos hacer es suponer que el cadáver está tendido sobre el lado izquierdo, aunque no siempre acertaremos. Cuanto más intensa sea la relación entre ambas variables, más éxito tendremos al utilizar el valor de una variable para predecir un valor de otra. Si no hay relación entre ellas, la predicción de una a partir de la otra no servirá de nada.

La tau y la lambda de Goodman y Kruskal emplean esta idea general de forma ligeramente distinta, si bien ambas son asimétricas. Refiriéndonos otra vez al ejemplo del cuadro 7.5: si sabemos que una tumba contiene un varón, podremos predecir con un 100 % de seguridad que estará tendido sobre su lado derecho; sin embargo, sabiendo que un individuo está tendido sobre su derecha, no tendremos un 100 % de seguridad de que sea una tumba masculina, porque 10 de 70 tumbas con el cadáver en esa postura son femeninas.

ASOCIACIÓN E INFERENCIA CAUSAL

A menudo, cuando observamos una asociación de la forma antes indicada, pensamos en términos de variables *dependientes* e *independientes*. Así, en el caso del sexo de un individuo y el lado sobre el que yace en una tumba, resulta posible visualizar el lado sobre el que está tendido el individuo como *dependiente* de su sexo, y no que el sexo dependa del lado sobre el que yace. Esto será así mientras la hipótesis funcione. Sin embargo, aunque hemos hablado en un sentido estadístico acerca de una variable que explica otra o que está asociada a otra, no podemos inferir necesariamente una relación causal entre ambas. Todos los libros de estadística llaman la atención acerca de los peligros

de inferir causalidad a partir de la asociación, debido a la posibilidad de correlación espúrea.

Naturalmente, las relaciones causales no pueden descubrirse por medio de simples análisis estadísticos; en el proceso de descubrimiento los métodos estadísticos pueden ser muy útiles o bien inducir a error. Si nos limitamos a aceptar la primera estadística obtenida por su simple valor aparente, fácilmente caeremos en el error. Es importante asegurarse de que cualquier conexión que infiramos entre objetos, hechos o procesos es real y verdadera; y generalmente se sugiere que la prueba para una relación real es que no cambie sean cuales sean las condiciones de observación; en otras palabras, ¿persiste o desaparece una relación entre dos variables cuando introducimos una tercera?

El proceso de investigar relaciones entre variables bajo una gran variedad de condiciones distintas es de mucha importancia, si es que van a hacerse inferencias válidas acerca de esas relaciones. Iremos viendo esto a lo largo de todo el libro; aquí nos limitaremos a introducir la idea general por medio de la Q de Yule. Hemos visto cómo funciona la Q en tablas simples de 2 × 2. La pregunta se planeta ahora acerca de lo que pasa cuando introducimos una tercera variable, de forma que la tabla resultante sea de 2 × 2 × 2. Las distintas posibilidades se muestran por medio de ejemplos.

TABLA 7.6. Volumen de una tumba, tabulado en referencia el sexo del individuo enterrado.

	Volumen de la tumba	
	$\leqslant$1,5 m^3	>1,5 m^3
M	22	47
F	33	26

El coeficiente Q simple entre dos variables, como las de la tabla 7.6, donde

$$Q = \frac{572 - 1.551}{572 + 1.551} = -0{,}461$$

es denominado coeficiente *de orden cero*; no toma en consideración los efectos de ninguna otra variable. En términos de algunos de los argumentos clásicos acerca de las prácticas funerarias y estatus social, podemos concluir que aquí hay evidencia para afirmar un nivel menor para las mujeres, pues se ha gastado menos energía para excavar sus tumbas. ¿Qué sucedería si tomamos en consideración la estatura de los individuos, utilizando estimaciones derivadas de las medidas de los huesos largos? Si queremos introducir una variable adicional como esta, habremos de partir la tabla original en dos (tabla 7.7).

TABLA 7.7. Volumen de la tumba, tabulado por el sexo del individuo y su estatura.

		Volumen de la tumba	
		$\leqslant$1,5 m^3	>1,5 m^3
Estimación estatura $\leqslant$155 cm	M	18	4
	F	30	6
Estimación estatura >155 cm	M	4	43
	F	3	20

¿Cómo se analiza esta nueva tabla? Lo que interesa es lo que sucede cuando «controlamos» la tercera variable. Este es un concepto importante, al que recurriremos con frecuencia (véase especialmente el capítulo 11), mas ¿qué significa? La idea subyacente es que observamos la relación entre nuestras dos variables originales teniendo en cuenta, o manteniendo constante, el efecto de la nueva. Lo hacemos fijándonos, por el momento, en la relación sexo/volumen de la tumba en una de las categorías de estatura, a continuación para la otra categoría; finalmente, reunimos las dos. Este procedimiento contrasta con el coeficiente original de orden cero, que no distinguiría entre las estaturas de los individuos. El nuevo coeficiente ya no es de orden cero, sino un coeficiente *parcial*; es parcial de *primer orden* porque sólo «controlamos» una única variable, la estatura en este caso.

TABLA 7.8. Una tabla de contingencia general 2 × 2 × 2.

		y	no y
t	x	a	b
	no x	c	d
no t	x	e	f
	no x	g	h

Para entender el efecto de la tercera variable en una relación, comparamos los valores del coeficiente de orden cero y del coeficiente parcial, porque esto proporciona la respuesta a la cuestión: qué sucede a la relación cuando esa variable está controlada. Puede que no haya cambio alguno, o puede que la relación se refuerce o que se debilite. Lo primero que hemos de hacer es calcular el coeficiente parcial para la tabla en cuestión. Supongamos una tabla general (tabla 7.8), en la que el coeficiente de orden cero entre x e y se obtiene uniendo simplemente las tablas.

$$Q_{xy} = \frac{[(a + e)(d + h)] - [(b + f)(c + g)]}{[(a + e)(d + h)] + [(b + f)(c + g)]}$$

El coeficiente parcial, controlada t, está dado por

$$Q_{xy/t} = \frac{(ad + eh) - (bc + fg)}{(ad + eh) + (bc + fg)}$$

Si hacemos estos cálculos en la tabla numérica que ejemplificaba el tamaño de las tumbas, obtendremos:

$$Q_{xy/t} = \frac{188 - 249}{188 + 249} = -0{,}139$$

Ya indicamos que una de tres opciones es posible cuando comparamos los coeficientes parcial y de orden cero. No se registran cambios cuando el coeficiente parcial es igual al de orden cero. Significa que no hay diferencias en las relaciones entre x e y, esté controlada la tercera variable o no; en otras palabras, la tercera variable no tiene efecto en la relación original bivariable. En nuestro ejemplo, no obstante, se observa que el coeficiente parcial es menor que el de orden cero, es decir, que la relación entre x e y se debilita cuando controlamos la tercera variable. La conclusión que se deriva de este caso es que *es la variación de la tercera variable la que explica la existencia de la relación* xy. Por eso, refiriéndonos al ejemplo, la relación entre sexo y volumen de la tumba desaparece en gran medida cuando se controla la estatura de los individuos. La relación de orden cero entre el sexo y el volumen de la tumba es una relación entre la estatura de los individuos y el volumen de la fosa, ya que las mujeres tienden a ser más pequeñas que los hombres.

Consideremos otro ejemplo, también basado en el análisis de enterramientos. La tabla 7.9, donde $Q = -0{,}34$, indica que las mujeres tienden a llevar anillos más frecuentemente que los hombres. Pero supongamos que disponemos de información que nos permite dividir la necrópolis en dos fases: la tabla 7.10, donde $Q_{xy/t} = -0{,}524$. Dicho de otro modo, tenemos un caso en el que el coeficiente parcial es mayor que el de orden cero. La intensidad de la asociación entre x e y desaparece si no se toma en consideración la tercera variable; cuando ésta se controla, la asociación mejora.

TABLA 7.9. Presencia/ausencia de anillos, tabulados por sexo del individuo enterrado. Necrópolis hipotética del norte de Alemania en la edad del hierro.

	Anillos en la tumba	
	Presencia	Ausencia
M	42	66
F	53	41

TABLA 7.10. Presencia/ausencia de anillos, tabulados por sexo. Subdivisión en fases.

		Anillos en la tumba	
		Presencia	Ausencia
Inicial	M	31	27
	F	25	5
Tardía	M	11	39
	F	28	36

Podemos salir ganando si calculamos Q separadamente para cada una de las dos subtablas. Esos dos coeficientes se denominarán *coeficientes condicionales*. Uno de ellos es el coeficiente de orden cero para aquellos individuos caracterizados por uno de los estados de la tercera variable, el otro es el coeficiente de orden cero para los individuos caracterizados por el otro estado de la tercera variable.

$$Q_{xy/t} = \frac{ad - bc}{ad + bc}$$

$$Q_{xy(\text{no})t} = \frac{eh - fg}{eh + fg}$$

En nuestro caso, el coeficiente condicional de la primera fase sería:

$$Q = -0{,}62$$

y el de la segunda fase:

$$Q = -0{,}47$$

El coeficiente parcial es la media de estos dos, tomando en consideración los distintos tamaños de las muestras en las que cada uno se basa.

Volviendo a la tabla 7.10, se observa que en la fase inicial la mayoría de las mujeres tienen anillos, mientras que los hombres están divididos en aquellos que tienen anillos y aquellos que no. En la fase tardía, las mujeres aparecen divididas con mayor igualdad entre las que tienen y las que no, mientras que hay menos hombres con anillos que hombres con ese objeto. Introducir la noción de tiempo, pues, provoca grandes diferencias en la interpretación de los dos coeficientes condicionales, y, por consiguiente, el coeficiente parcial adopta los valores que adopta porque en la primera fase inicial hay muy pocas mujeres sin anillos, mientras que en la fase tardía hay muy pocos hombres con anillos. Ambas situaciones tienen el mismo efecto en el coeficiente Q, por supuesto, ya que los valores de las celdas opuestas diagonalmente se multiplican.

TABLA 7.11. Presencia/ausencia de cuentas, tabuladas por el rango social del individuo

		Cuentas en las tumbas	
		Presencia	Ausencia
Rango	Alto	53	39
	Bajo	44	55

Falta por mostrar otra posibilidad. Permaneciendo en el mismo dominio (análisis de una necrópolis), imaginemos que según un criterio razonable hemos dividido las tumbas en un grupo de «alto rango» y otro de «bajo rango». Este criterio no contempla la presencia o ausencia de cuentas de collar en la tumba, y nos interesa saber si ese objeto del ajuar funerario está relacionado con el rango social del individuo enterrado en la tumba. De la tabla 7.11, donde $Q = 0{,}25$, se sugiere que la presencia de cuentas está ligeramente asociada con el alto rango. Intentamos ahora controlar el sexo: la tabla 7.12, donde $Q_{xy/t} = 0{,}35$.

TABLA 7.12. Presencia/ausencia de cuentas, tabuladas por rango social. Subdivisión por sexo.

		Cuentas en las tumbas	
		Presencia	Ausencia
M	Alto	9	33
	Bajo	15	28
F	Alto	42	6
	Bajo	29	27

La comparación entre los coeficientes Q parcial y de orden cero sugiere al principio que nos hallamos ante un caso de supresión débil; cuando controlamos la variable sexo, la intensidad de la relación entre cuentas y rango aumenta. Pero calculemos ahora los coeficientes condicionales para cada una de las categorías de la variable de control:

$$Q_{xy(\text{varones})} = -0{,}32$$
$$Q_{xy(\text{mujeres})} = -0{,}73$$

Las diferencias son muy grandes, de forma que el coeficiente parcial de 0,35 obtenido de la media de los dos condicionales es totalmente erróneo. Obviamente hay una relación muy diferente entre el rango y la presencia de cuentas en las tumbas masculinas que la que existe en las femeninas; pocos varones de alto rango tienen cuentas, mientras que la mayoría de las mujeres de alto rango las ostentan. Este resultado recibe el nombre de *especificación* o *interacción*; el efecto de la tercera variable es especificar cuál de las dos relaciones se mantiene entre las variables x e y. Avances subsiguientes sólo son posibles dividiendo la muestra en dos grupos según la variable específica (aquí, el sexo) y continuando la investigación de las conexiones causales dentro de cada grupo por separado.

Hasta aquí hemos desarrollado unas técnicas específicas para examinar relaciones y hemos observado el efecto que una tercera variable, t, puede tener entre x e y; de este modo, podremos comprobar, por ejemplo, dentro de los límites de un sistema de tres variables, si una relación es o no espuria.

Para completar nuestra investigación de las conexiones entre las tres variables, podemos estar también interesados en el efecto de y sobre la relación xt y el de x sobre la relación yt. Esto supondría ser capaz de definir las conexiones causales existentes en el sistema de tres variables, algo que se opone a la simple asociación por pares. Una pequeña reflexión mostrará al lector que se trata de algo bastante complicado. Para empezar, nos enfrentamos a 12 coeficientes Q, y si no se toma en consideración el hecho de que las relaciones pueden ser positivas o negativas, la dependencia puede tener un sentido u otro, o los dos, con lo que resultará una cantidad enorme de relaciones diferentes.

Sería mejor estudiarlas utilizando el conocimiento *a priori* para predecir ciertas relaciones —hipotetizar una serie de afirmaciones causales— y proceder a contrastarlas. Si nos equivocamos, habrá que reconsiderar las afirmaciones y contrastar un nuevo modelo. Obviamente, para hacerlo hemos de ser capaces de predecir la conducta de los coeficientes de asociación bajo los supuestos *a priori*. Si los coeficientes se comportan como esperamos, consideraremos las afirmaciones válidas provisionalmente. Resumiendo:

1. Se intenta empezar con una hipótesis bien meditada; la aplicación simple de Q (o cualquier otro coeficiente) para ver simplemente qué sucede, suele confundirse a menudo con un sustituto del análisis de relaciones de causa y efecto. Se debe empezar definiendo la variable que más necesita una explicación, seleccionando la explicación más probable; las variables se desarrollan y expanden a medida que continúa el análisis.

2. Lo más importante es controlar las otras variables, incluso si la asociación entre las dos variables es muy intensa: puede ser espuria. Por otro lado, una relación que inicialmente parece débil, puede esconder algo más, por lo que el control de las variables aparece como algo esencial.

3. Una vez que aparece la interacción o la especificación, se ha de emprender una acción apropiada. En general, esto significa efectuar por separado los análisis de la relación entre las dos primeras variables para cada una de las dos categorías de la tercera variable, y asegurar que se ha fijado claramente a qué categoría se refiere cualquier conclusión, ya que pueden ser diferentes.

UN ENFOQUE MODERNO PARA INVESTIGAR LAS RELACIONES ENTRE VARIABLES DE ESCALA NOMINAL: INTRODUCCIÓN A LOS MODELOS LOGARÍTMICOS

De hecho, el análisis de las relaciones entre variables a escala nominal no suele hacerse por medio del coeficiente Q u otro cualquiera de los coeficientes de asociación que hemos visto, sino por medio de *modelos logarítmicos* y el método, estrechamente relacionado, de la *regresión logarítmica* (logit) (y, potencialmente, por medio de la técnica conocida como análisis de correspondencias, presentada en el capítulo 13). La aplicación de estos modelos es un desarrollo relativamente reciente, que depende, como muchos otros en estadística, de la disponibilidad de ordenadores. A causa de su utilidad y empleo cada vez más amplio en arqueología (véase, por ejemplo, Clark, 1976; Spaulding, 1977; MacIntosh y MacIntosh, 1980), es importante saber para qué sirven y cómo funcionan. Lewis (1986) proporciona una explicación más amplia de la que se presenta aquí.

Conceptualmente, sus implicaciones son bastante claras, aunque la notación pueda ser algo difícil y los cálculos tan tediosos que en la práctica requieren el uso de un programa informático apropiado.

Los problemas que resuelven son aquellos que ya hemos visto en la sección anterior: cómo investigar la relación entre más de dos variables de forma coherente y en un entorno único, de forma que podamos afirmar cuáles son las relaciones importantes, y teniendo en cuenta las diversas posibilidades y evitar las situaciones, ya contempladas en este capítulo, donde la introducción de una tercera variable variaba y en ocasiones alteraba los efectos sobre las relaciones entre las otras dos variables.

El enfoque actual nos vuelve a conducir al χ^2. Hasta ahora, sólo lo hemos usado para contrastar la hipótesis nula de no asociación entre variables, partiendo de la observación de las discrepancias entre los valores de los datos observados y los esperados bajo la hipótesis nula. En el modelo logarítmico no estamos limitados a una hipótesis nula. El fundamento de la técnica es construir modelos de las relaciones posibles entre variables en un conjunto de datos, para derivar valores esperados para los diferentes modelos, y decidir cuál de ellos se ajusta mejor a los datos, comparando los valores esperados producidos por los modelos, con los valores de los datos observados, estipulando que el modelo elegido sea el más simple de los que muestran un ajuste razonable.

La parte logarítmica radica en la construcción de los modelos. La exposición que sigue está muy influida por Lewis (1986). El modelo más simple posible es el que ya hemos visto: la hipótesis de independencia nula. En el caso de dos variables vimos que, para encontrar el valor esperado para una celda en particular de la tabla, se multiplicaba la suma de fila de esa celda por la suma de la columna respectiva, y dividíamos por el número total de observaciones:

$$\text{Valor esperado} = \frac{\text{(suma fila) (suma columna)}}{\text{número total de observaciones}}$$

Obteníamos el valor esperado por un procedimiento de multiplicación y división. Si, no obstante, nos interesase el logaritmo del valor esperado, la forma se alteraría ligeramente:

$$\log\,(\text{valor esperado}) = \log\,(\text{suma fila}) + \log\,(\text{suma columna}) - \log\,(\text{n.° total de observaciones})$$

ya que la suma de logaritmos corresponde a la multiplicación normal, y la resta a la división; así pues, sumaremos y restaremos en vez de multiplicar y dividir. Por esa razón, la fórmula del valor esperado es considerada *aditiva* o *lineal*, expresada en logaritmos de los valores originales.

¿Para qué sirve todo esto? Tal y como veremos al tratar el análisis de regresión, los modelos lineales suelen ser mucho más fáciles de tratar, y usándolos podremos proseguir el análisis de las relaciones entre variables, y no limitarnos a la simple contrastación de su independencia. Recuérdese que el propósito de los modelos logarítmicos es construir modelos que se ajusten a los datos, sujeto a la condición de parsimonia, explicada anteriormente.

Supongamos que tras obtener el logaritmo de los valores esperados de la manera que se acaba de explicar, observamos que éste se diferencia considerablemente del logaritmo del valor observado. Esto significa que la modelización (o descripción) de los valores esperados, expresados en términos de la suma de la fila, la suma de la columna y el número total de observaciones, es insuficiente. Algo falta. Si postulamos ahora que nuestras dos variables están relacionadas y *añaden* un término extra a la relación, entonces, si queremos tener en cuenta esa relación y sólo tratamos con dos variables, apreciaremos enseguida que el nuevo modelo ajusta de inmediato; si tratamos con más de dos variables, habrá que añadir sucesivamente términos extra a la ecuación, intentando mejorar el ajuste entre lo esperado y lo observado. Nuestra selección entre los diversos modelos posibles estará determinada por la bondad del ajuste y por el criterio de simplicidad.

TABLA 7.13. Los datos de la tabla 7.7; $n = 128$.

		Volumen de la tumba $\leqslant 1{,}5$ m³	Volumen de la tumba $> 1{,}5$ m³
Estimación estatura $\leqslant 155$ cm	M	18	4
	F	30	6
Estimación estatura > 155 cm	M	4	43
	F	3	20

Como es habitual, sus implicaciones quedarán más claras por medio de un ejemplo. Consideremos, de nuevo, uno de los utilizados para mostrar el uso del coeficiente Q con tres variables: el ejemplo (tabla 7.7) en el que tratábamos de entender las relaciones entre el volumen de las fosas y la estatura y el sexo de los individuos enterrados en ellas, en el análisis de una necrópolis hipotética de enterramientos de inhumación simple. La forma más sencilla de comprobar un modelo para ese cuadro es que las tres variables no estén relacionadas las unas con las otras; la hipótesis nula, en otras palabras. Sin embargo, ahora usamos tres variables en lugar de dos. Para la celda superior izquierda de la tabla 7.13, el modelo es el siguiente, usando el enfoque antes descrito:

log (cantidad esperada de varones bajos en fosas pequeñas) =
log (n.º total de varones) +
log (n.º total de fosas pequeñas) +
log (n.º total de individuos bajos) −
log (n.º total de observaciones)

Lo que dice es que el número total de varones bajos en fosas pequeñas es, simplemente, una función de la cantidad total de varones, fosas pequeñas e individuos bajos, teniendo en cuenta el número total de observaciones.

TABLA 7.14. Valores esperados añadidos a los datos de la tabla 7.13.

		Volumen de la tumba	
		$\leqslant$1,5 m^3	>1,5 m^3
Estimación estatura	M	18 (13,4)	4 (17,8)
$\leqslant$155 cm	F	30 (11,5)	6 (15,3)
Estimación estatura	M	4 (16,2)	43 (21,5)
>155 cm	F	3 (13,9)	20 (18,4)

Si este fuese un buen modelo, el número esperado de varones bajos en fosas pequeñas producido por él tendría que ser muy próximo al del número observado; si no, habría una discrepancia. Podemos modelizar los valores esperados de todas las celdas de la tabla de esta manera y, naturalmente, si existen muchas discrepancias entre los valores observados y los valores esperados, concluiríamos con un valor χ^2 significativo. Los valores esperados y observados aparecen en la tabla 7.14; podemos efectuar ahora una prueba para las diferencias entre ellos. De hecho, y por razones que serán evidentes más adelante, la prueba de χ^2 no se usa en estos casos, sino una prueba equivalente, llamada G^2.

$$G^2 = 2 \sum \left[(\text{valor observado}) \log_e \left(\frac{\text{valor observado}}{\text{valor esperado}} \right) \right]$$

cuya suma afecta a todas las celdas de la tabla. En el ejemplo que estamos siguiendo:

$$G^2 = [18 \log_e (18/13,4)] + [4 \log_e (4/17,8)] + [30 \log_e (30/11,5)] + [6 \log_e (6/15,3)] + [4 \log_e (4/16,2)] + [43 \log_e (43/21,5)] + [3 \log_e (3/13,9)] + [20 \log_e (20/18,4)] = 87,54$$

La significación estadística del valor G^2 puede obtenerse en la tabla χ^2, si bien requiere que conozcamos el número correcto de grados de libertad (= cantidad de celdas − cantidades estimadas de los datos). Hay ocho celdas, y hemos precisado calcular cuatro estimaciones (número de varones, número de fosas pequeñas, número de individuos bajos y número total de observaciones), de aquí que el número de grados de libertad sea cuatro. Con cuatro grados de libertad, el valor G^2 es muy significativo. La hipótesis nula según la cual las tres variables son independientes unas de otras ha de rechazarse.

El uso tradicional del χ^2 llegaría hasta aquí, tal y como vimos en el capítulo 6. Lo que hemos de hacer ahora, sin embargo, es mejorar el modelo. Ya que la hipótesis de independencia no se mantiene, ha de existir alguna relación entre el sexo, el tamaño de la fosa y la estatura del individuo que aún no hemos descubierto, y que habría que incluir en el modelo. Las posibilidades son bastante considerables. En primer lugar, todas esas variables han de relacionarse entre sí formando pares:

Sexo/tamaño de la fosa
Sexo/estatura de los individuos
Estatura de los individuos/tamaño de la fosa

En un nivel más complejo, todos los pares deben relacionarse también entre sí:

Sexo/tamaño de la fosa y sexo/estatura de los individuos
Tamaño de la fosa/sexo y tamaño de la fosa/estatura de los individuos
Estatura de los individuos/sexo y estatura de los individuos/tamaño de la fosa

Aún más complicado: también esas agrupaciones han de ponerse en relación:

Sexo/tamaño de la fosa, sexo/estatura de los individuos y estatura de los individuos/tamaño de la fosa

Finalmente, las tres variables se relacionarán simultáneamente a otra: el sexo se relacionará con el tamaño de la fosa no directamente, sino indirectamente, por intermedio de la estatura; eso será válido para las tres variables.

Podemos ver que, a medida que avanzamos a través de los distintos niveles de complejidad, cada uno incluye al situado por debajo, de forma que si dos

pares de variables están relacionados, eso implicará que uno de los pares estaba ya relacionado en el nivel precedente; igualmente, en el caso de relaciones en un nivel superior, dos de los pares han de estar relacionados en el nivel inmediatamente inferior.

La idea es empezar en el nivel más bajo e ir avanzando en la jerarquía de la complejidad, deteniéndonos en el modelo más simple que se ajuste a los datos. En cada nivel se pierde un grado de libertad, ya que los datos se están usando para estimar la asociación entre cada par de variables. En el modelo de independencia o no asociación teníamos cuatro grados de libertad al principio. Cuando se alcanza el nivel de complejidad más alto, donde todas las variables están relacionadas entre sí simultáneamente, ya no queda ningún grado de libertad. Las cantidades esperadas corresponderían a las cantidades observadas y reproduciríamos los datos con los que empezamos; ¡de ahí que ese modelo final, que suele denominarse *saturado*, no tenga mucho interés!

La discusión anterior está resumida en la tabla 7.15 (basada en Fienberg, 1980), que enumera los diversos modelos, la cantidad de grados de libertad (g.l.) asociados a ellos y su descripción en símbolos matemáticos, usando la notación de Fienberg. El lector apreciará que para los modelos 2 y 3 sólo se enumera una de las tres posibles opciones.

TABLA 7.15. Modelos logarítmicos posibles para las relaciones entre tres variables.

Modelo	g.l.	Abreviación	Descripción simbólica
1. No asociación	4	[1][2][3]	$\log E_{ijk} = u + u_1 + u_2 + u_3$
2. Asociación de 1 par de variables	3	[12][3]	$\log E_{ijk} = u + u_1 + u_2 + u_3 + u_{12}$
3. Asociación de 2 pares de variables	2	[12][23]	$\log E_{ijk} = u + u_1 + u_2 + u_3 + u_{12} + u_{23}$
4. Asociación de 3 pares de variables	1	[12][23][13]	$\log E_{ijk} = u + u_1 + u_2 + u_3 + u_{12} + u_{23} + u_{13}$
5. Interacción entre las 3	0	[123]	$\log E_{ijk} = u + u_1 + u_2 + u_3 + u_{12} + u_{23} + u_{13} + u_{123}$

Para ilustrar cómo han de leerse esas ecuaciones, tomemos el modelo 3, relacionado con nuestro ejemplo, y supongamos, al igual que cuando comprobábamos la independencia, que queremos averiguar el valor esperado de la cantidad de varones bajos en fosas pequeñas, si bien ahora afirmamos que hay una relación entre sexo y tamaño de la fosa, y entre tamaño de la fosa y estatura.

log (cantidad esperada de varones bajos en fosas pequeñas) [log E_{ijk}] =
log (n.º total de varones) [u_1] +
log (n.º total de fosas pequeñas) [u_2] +
log (n.º total de individuos bajos) [u_3] +

log (interacción entre sexo y tamaño de la fosa) [u_{12}] +
log (interacción entre tamaño de la fosa y estatura) [u_{23}]

Habiendo rechazado el modelo de no asociación, podemos ahora contrastar los diferentes modelos posibles del nivel 2, afirmando una asociación entre cualquiera de los pares:

2a) Sexo y tamaño de la fosa
2b) Estatura individual y tamaño de la fosa
2c) Sexo y estatura individual

Las tablas 7.16-7.18 muestran los valores observados en los datos, junto con los valores esperados bajo cada uno de esos modelos. Los valores G^2 (equivalentes a χ^2) para esos modelos, y para la prueba de independencia inicial, aparecen en la tabla 7.19, donde: variable 1 = sexo, 2 = volumen, 3 = estatura.

TABLA 7.16. Valores esperados para el modelo 2a añadidos a los datos de la tabla 7.13.

		Volumen de la tumba	
		⩽1,5 m³	>1,5 m³
Estimación estatura ⩽155 cm	M	18 (10,0)	4 (21,3)
	F	30 (15,0)	6 (11,8)
Estimación estatura >155 cm	M	4 (12,0)	43 (25,7)
	F	3 (18,0)	20 (14,2)

TABLA 7.17. Valores esperados para el modelo 2b, añadidos a los datos de la tabla 7.13.

		Volumen de la tumba	
		⩽1,5 m³	>1,5 m³
M	⩽155 cm	18 (25,9)	4 (5,4)
	>155	4 (3,8)	43 (34,0)
F	⩽155	30 (22,1)	6 (4,6)
	>155	3 (3,2)	20 (29,0)

TABLA 7.18. Valores esperados para el modelo 2c, añadidos a los datos de la tabla 7.13.

		Sexo	
		M	F
⩽1,5 m³	⩽155 cm	18 (9,5)	30 (15,5)
	>155	4 (20,2)	3 (9,9)
>1,5 m³	⩽155	4 (12,6)	6 (20,5)
	>155	43 (26,8)	20 (13,1)

TABLA 7.19. Resumen del ajuste de los modelos logarítmicos de las relaciones entre sexo, estatura y volumen de la fosa.

Modelo	Abreviación	G^2	g.l.
1. No asociación	[1][2][3]	87,54	4
2. Asociación de 1 par de variables:			
a)	[12][3]	79,70	3
b)	[1][23]	11,38	3
c)	[13][2]	76,17	3

Todos esos valores G^2 son significativos, como mínimo en el nivel 0,01, lo que significa que es altamente improbable que cualquiera de esos modelos se ajuste a los datos: las diferencias entre los valores esperados bajo el modelo y los observados es demasiado grande. No obstante, está claro que el modelo 2b, afirmando una relación entre la estatura y el volumen produce una marcada caída del valor G^2 y proporciona el mejor ajuste de cualquiera de los modelos probados.

Trasladémonos al nivel siguiente, en el que los modelos presuponen relaciones entre dos pares de variables:

3a) El sexo y la estatura están relacionados, así como la estatura y el tamaño de la fosa; el sexo y el tamaño de la fosa no están relacionados.
3b) El sexo y la estatura están relacionados, así como el sexo y el volumen de la fosa; la estatura y el tamaño de la fosa no están relacionados.
3c) El sexo y el volumen de la fosa están relacionados, la estatura y el tamaño de la fosa también, no así el sexo y la estatura.

TABLA 7.20. Valores esperados para el modelo 3a, añadidos a los datos de la tabla 7.13.

		Volumen de la tumba $\leqslant 1{,}5$ m³	Volumen de la tumba $>1{,}5$ m³
Estimación estatura $\leqslant 155$ cm	M	18 (18,2)	4 (3,8)
	F	30 (29,8)	6 (6,2)
Estimación estatura >155 cm	M	4 (4,7)	43 (42,3)
	F	3 (2,3)	20 (20,7)

TABLA 7.21. Valores esperados para el modelo 3b, añadidos a los datos de la tabla 7.13.

		Volumen de la tumba $\leqslant 1{,}5$ m³	Volumen de la tumba $>1{,}5$ m³
M	$\leqslant 155$ cm	18 (7,0)	4 (15,0)
	>155 cm	4 (15,0)	43 (32,0)
F	$\leqslant 155$ cm	30 (20,1)	6 (15,9)
	>155 cm	3 (12,9)	20 (10,1)

TABLA 7.22. Valores esperados para el modelo 3c, añadidos a los datos de la tabla 7.13.

		Sexo M	Sexo F
$\leqslant 1{,}5$ m³	$\leqslant 155$ cm	18 (19,2)	30 (28,8)
	>155 cm	4 (2,8)	3 (4,2)
$>1{,}5$ m³	$\leqslant 155$ cm	4 (6,5)	6 (3,5)
	>155 cm	43 (40,5)	20 (22,5)

TABLA 7.23. Resumen del ajuste de los modelos logarítmicos de las relaciones entre sexo, estatura y volumen de la fosa.

Modelo	Abreviación	G^2	g.l.
1. No asociación	[1][2][3]	87,54*	4
2. Asociación de 1 par de variables:			
a)	[12][3]	79,70*	3
b)	[1][23]	11,38*	3
c)	[13][2]	76,17*	3
3. Asociación de 2 pares de variables:			
a)	[13][23]	0,36	2
b)	[12][13]	69,17*	2
c)	[12][23]	3,98	2

Las tablas 7.20-7.22 muestran de nuevo los valores observados, pero ahora con los valores esperados para cada uno de los tres modelos 3. Un vistazo a esos cuadros permite apreciar que el ajuste del modelo 3a es excelente, el del modelo 3b es muy pobre, mientras que el del modelo 3c es también muy bueno. Los valores G^2 de bondad del ajuste para esos tres modelos, junto al resto de los que hemos considerado hasta aquí, aparecen en la tabla 7.23, señalándose los que son estadísticamente significativos al nivel 0,05.

Este es el momento de destacar que el uso del G^2 y no del χ^2 es mucho más apropiado. Si consideramos, por ejemplo, el valor G^2 de un modelo en el nivel 2, digamos el 2b, y los restamos del valor G^2 del modelo de independencia en el nivel 1, obtenemos una medida de la mejora de la bondad del ajuste. En este caso, $87{,}54 - 11{,}38 = 76{,}16$. Estas comparaciones pueden efectuarse entre modelos de dos niveles distintos cualesquiera. Podríamos obtener, por ejemplo, la diferencia entre el G^2 del nivel 1 y el del modelo 3a: $87{,}54 - 0{,}36 = 87{,}18$, la mejora obtenida por la predicción de las variables de la celda no se basa en el supuesto de la independencia, sino en el supuesto de una relación entre las variables 1 y 3 (sexo y estatura) y 2 y 3 (estatura y volumen de la tumba).

Además, la significabilidad de las diferencias puede ser contrastada estadísticamente, usando el número de grados de libertad obtenido al restar el número

de grados de libertad del modelo de nivel más alto de aquel a un nivel inferior. Así, en nuestro segundo ejemplo, el número de grados de libertad para el modelo del nivel 1 de no asociación es 4, y el número para el modelo 3a es 2, con lo que tenemos una diferencia de G^2 de 87,18 con dos grados de libertad, que puede ser comparada con la tabla de χ^2 para establecer su significabilidad. Si el modelo más complejo produce una caída estadísticamente significativa en él, podremos adoptarlo.

Contemplando los resultados para los modelos del nivel 3, apreciamos que dos de los tres valores G^2 representan una mejora considerable con respecto a los del nivel 2. El único que no mejora, el modelo 3b, tiene un ajuste peor que el modelo 2b, lo que no sorprende, pues omite la relación entre las variables 2 y 3, tamaño de la fosa y estatura del individuo, que el modelo 2b establece como muy importante.

Claramente, el más ajustado de todos los modelos es el 3a, con un valor G^2 cercano a cero, que indica un ajuste casi perfecto entre los valores esperados por el modelo y los valores de los datos. ¿Es significativamente mejor que el 2b? Comparemos los valores G^2: 11,38 − 0,36 = 11,02 con un grado de libertad, que es altamente significativo. El modelo apropiado, siguiendo la formulación de la tabla 7.15, es:

$$u + u_1 + u_2 + u_3 + u_{23} + u_{13}$$

De nuevo, si lo queremos poner en relación con el valor predicho de una celda en particular, la que representa los varones bajos en fosas pequeñas, tenemos:

log (cantidad esperada de varones bajos en fosas pequeñas) =
log (n.º total de observaciones) +
log (n.º total de varones) +
log (n.º total de fosas pequeñas) +
log (n.º total de individuos bajos) +
log (interacción entre tamaño de la fosa y estatura) +
log (interacción entre sexo y estatura)

En otras palabras, el tamaño de la fosa está relacionado con la estatura de los individuos y la estatura de los individuos está relacionada con su sexo, si bien el sexo y el tamaño de la fosa no están directamente relacionados. Esas relaciones, que tienen sentido intuitivamente, explican lo que sucede en los datos, por lo que no hay necesidad de utilizar un nivel de complejidad más alto.

Tal y como el lector ya se habrá dado cuenta, la única parte realmente complicada de los modelos logarítmicos es el cálculo de los valores esperados. Eso importa poco, porque no se efectúan a mano, sino por medio de un ordenador. Numerosos paquetes informáticos de programas estadísticos incluyen este tipo de modelos (véase anexo 2).

La explicación que se ha presentado aquí puede ser ampliada recurriendo a los trabajos de Fienberg (1980) y Lewis (1986), cuyo contenido es específicamente arqueológico. Incluye también una descripción del caso más restringido en el que podemos afirmar que una de las variables es dependiente, y es su variación lo que intentamos comprender en términos de los efectos de las variables independientes: el modelo logit. De hecho, los modelos logarítmicos, en particular el modelo logit, junto con el análisis de regresión (capítulos 9-11) y el análisis de varianza (no tratado aquí; véase Blalock, 1972), son todos instancias específicas de lo que se conoce como *modelo lineal generalizado* (Baker y Nelder, 1978; Everitt y Dunn, 1983), en el que los valores de los datos están explicados en términos de los efectos *aditivos* de diversas variables.

Ejercicios

7.1. Se ha registrado el tipo de borde y la forma del cuello de unos fragmentos cerámicos de un asentamiento de Mesopotamia. Los resultados aparecen abajo, divididos en fragmentos decorados y sin decorar:

Cantidad de fragmentos	Tipo de borde	Forma del cuello	Decoración
16	1	1	con
9	1	2	con
14	1	1	sin
32	1	2	sin
7	2	1	con
14	2	2	con
30	2	1	sin
18	2	2	sin

a) ¿Cuán intensa y qué forma tiene la relación global entre la forma del cuello y el tipo del borde? ¿Es significativa esa relación?

b) ¿En qué forma la introducción de una tercera variable, la decoración, afecta a la relación entre el tipo de borde y la forma del cuello, tanto global como condicionalmente?

7.2. Se ha llevado a cabo un análisis de la asociación entre dos motivos diferentes que aparecen en un conjunto peculiar de recipientes cerámicos en una necrópolis centroeuropea de la edad del hierro. Basándose en los siguientes datos:

		Motivo 1	
		Presente	Ausente
Motivo 2	Presente	29	17
	Ausente	23	32

¿Se alteran las conclusiones extraídas de las asociaciones entre ambos motivos cuando se dividen las tumbas en dos fases cronológicas? Véanse los datos siguientes:

		Motivo 1	
		Presente	Ausente
Fase inicial			
Motivo 2	Presente	15	7
	Ausente	9	11
Fase tardía			
Motivo 2	Presente	14	10
	Ausente	14	21

8. VARIABLES NUMÉRICAS: LA DISTRIBUCIÓN NORMAL

Los capítulos inmediatamente precedentes han estudiado las variables nominales, en particular las formas de analizar las relaciones entre ellas. Los próximos capítulos tratarán acerca de las relaciones entre variables numéricas, medidas en una escala interválica o proporcional. No obstante, antes de describir los métodos, será preciso que volvamos a los temas introducidos en los capítulos 3 y 4, que versaban sobre la descripción de variables únicas, para considerar un tipo de distribución en particular, la distribución normal o de Gauss, ya mencionada en el capítulo 4.

Incluso si el lector no sabe lo que es una distribución normal, lo más probable es que haya oído decir que se trata de algo muy importante. Eso es cierto, pues una gran cantidad de las distribuciones observadas son aproximadamente normales, y también a causa de la significación teórica de esta distribución en la estadística inductiva, como fundamento de muchos de los métodos estadísticos.

Por esa razón, muchos manuales de estadística otorgan a la teoría de la distribución normal un papel fundamental por su uso como base de la inferencia estadística, incluyendo las pruebas de significación. En este libro se ha dejado un tanto de lado, sobre todo lo que que se refiere a su papel en la inferencia estadística. Hay varias razones: las pruebas que se basan en ella son conceptualmente difíciles y no tienen gran importancia. Tal y como ya hemos visto, se puede hacer mucho sin emplear la distribución normal; en estadística, su empleo es cada vez menos importante. Cuando la estadística estaba en su fase de desarrollo, la teoría normal proporcionó un punto de referencia para el desarrollo de los métodos estadísticos. Gracias a la disponibilidad y la capacidad cada vez mayores de los ordenadores, es posible, hoy en día, simular las distribuciones estadísticas directamente, usando técnicas numéricas. Además, los defensores del análisis de datos exploratorio han señalado que los métodos basados en la teoría normal suelen ser muy sensibles a las irregularidades en los datos; este punto volverá a abordarse en este capítulo.

Así y todo, la distribución normal no puede ignorarse por completo, y el

29

motivo por el que debemos prestarle ahora nuestra atención es que los métodos usados actualmente para investigar las relaciones entre las variables medidas en escalas de intervalo o proporcionales se basan en ella. Además, como veremos más adelante, pueden aproximarse razonablemente a la normalidad ciertas distribuciones de datos que no eran normales al principio.

En las páginas siguientes, el objetivo es contemplar la distribución normal desde un punto de vista puramente descriptivo, y en particular considerar cómo está relacionada con la desviación típica (véase cap. 4).

La distribución normal

Cuando se expuso el uso de los gráficos de barras para representar la distribución de frecuencias de variables numéricas continuas (cap. 3), se insistió en que la anchura de los intervalos era una cuestión importante. En particular, si los intervalos de la distribución se estrechaban paulatinamente en una muestra de un cierto tamaño, la distribución empezaba a adoptar una apariencia irregular, con huecos y desfases en ella. Aumentando el número de casos, sin embargo, y manteniendo el incremento al estrechar los intervalos, la distribución afina sus divisiones, conservando la misma forma. Tal es así que, en la figura 8.1, sería posible ir de (a) a (b).

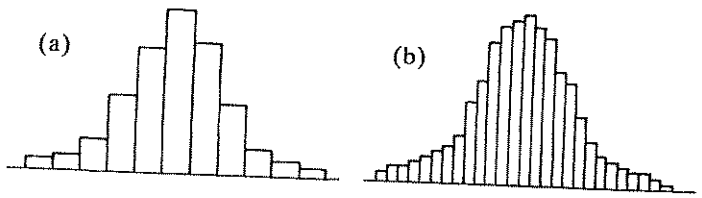


FIGURA 8.1. (a) Histograma con intervalos anchos; (b) histograma con intervalos muy estrechos, basado en una gran cantidad de observaciones.

Asumiendo que la distribución tiene la forma ilustrada en la figura 8.1, si imaginamos que los intervalos se estrechan infinitamente y el número de observaciones aumenta al mismo ritmo, obtendremos al final una curva atenuada en forma de campana (fig. 8.2).

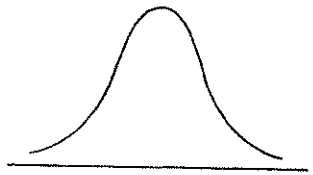


FIGURA 8.2. Una distribución normal.

Del mismo modo que el área dentro de un gráfico de barras puede ser calculada sumando las áreas de los rectángulos individuales, el área bajo la curva atenuada puede calcularse sumando la cantidad infinita de rectángulos bajo la curva; es la operación de cálculo denominada integración.

La curva normal es una curva simétrica, atenuada y en forma de campana, definida por una ecuación particular; una de sus características es que las dos colas extendidas al infinito en cualquiera de las direcciones nunca alcanzarán el eje horizontal. Al nivel que nos movemos en este libro, la ecuación no tiene gran interés. Lo que importa es que sea cual sea la media y la desviación típica particulares que tenga una curva normal, hay siempre una proporción constante del área bajo la curva, o bien una proporción constante de los casos en una distribución de este tipo, entre la media y una distancia desde la media expresada en unidades de desviación típica (fig. 8.3).

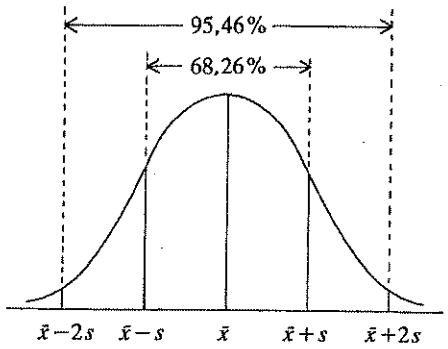


FIGURA 8.3. Porcentaje del área bajo la curva normal, dentro de una y dos desviaciones típicas de la media.

Será mejor ver algunos ejemplos para aclarar la cuestión. Así el área bajo la curva entre la media y un punto que sea una desviación típica mayor o menor que la media será el 34,13 % del área total bajo la curva. Entre una desviación típica menor que la media y una desviación típica mayor que la media, será el doble del 34,13 %, es decir, el 68,26 % del área bajo la curva. Los porcentajes correspondientes para dos desviaciones típicas son: 47,43 % y 95,46 % y para tres desviaciones típicas 49,86 % y 99,73 %.

Aunque esas proporciones se basan en la curva normal teóricamente definida, muchas distribuciones de frecuencia empíricamente obtenidas están lo suficientemente próximas a ella para que las reglas anteriores sean aplicables, por lo que es posible usar esas proporciones constantes.

El hecho de que muchas distribuciones de frecuencia reales estén bastante próximas a la normalidad, de forma que puedan usarse esos resultados teóricos, no es accidental. Si el valor de alguna variable es el resultado del efecto

acumulado de un gran número de otras variables independientes, podrá probarse matemáticamente que la distribución de los valores de esa variable será aproximadamente normal. Un ejemplo de una tal variable en biología, disciplina en la que la distribución normal se aplicó por vez primera, es la estatura del cuerpo, que está determinada por muchos factores genéticos, pero también por factores como la nutrición y el entorno. Todos esos factores tienen tendencia a actuar en distintas direcciones. El resultado es que la distribución de estaturas en una población será normal, como de hecho lo es. Hay muchos ejemplos arqueológicos de variables en escala proporcional, particularmente dimensiones físicas, como longitudes, anchuras, pesos, volúmenes, etc., que están afectadas a su vez por muchos otros factores, con el resultado de que la distribución de los valores de esas variables es normal, o, cuando menos, no muy alejada de la normalidad.

Será conveniente que mostremos ahora cómo se pueden usar e interpretar esas proporciones constantes, características de la distribución normal, en un caso arqueológico específico. Inevitablemente, va a ser un tanto artificial, ya que lo habitual es usarlas como un medio para un fin, y no un fin en sí mismas, que es como las abordaremos. Supongamos que estamos estudiando un gran conjunto de puntas de flecha del suroeste de Estados Unidos. Sus longitudes están distribuidas normalmente, con una media de 110 mm y una desviación típica de 20 mm (véase fig. 8.4). Inicialmente, pretendemos averiguar la proporción de longitudes entre 110 y 140 mm.

En primer lugar, es necesario descubrir cuántas desviaciones típicas está alejado 140 de 110; en milímetros es 30, mientras que la desviación típica es 20. Si dividimos la diferencia entre la media y el valor en el que estamos interesados por la desviación típica, obtendremos la proporción que pretendemos: 30/20 = 1,5. El valor 140 está 1,5 desviaciones típicas alejado de la media. Cuan-

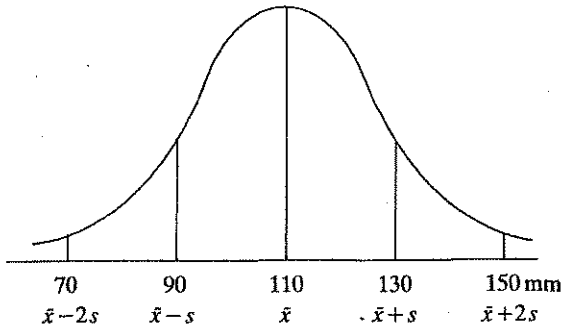


FIGURA 8.4. Distribución de longitudes de una gran cantidad de puntas de flecha del suroeste de Estados Unidos.

do una unidad aparece bajo el aspecto de varias unidades de desviación típica alejada de su media, se la denomina puntuación Z (o puntuación estándar), donde Z representa la desviación respecto a la media en unidades de desviación típica. La expresión general es:

$$Z = \frac{x - \bar{x}}{s}$$

donde $\bar{x}$ es la media, s el valor de la desviación típica y x el valor del límite del intervalo que nos interesa.

¿Cómo pasamos de un valor para Z a un valor para la proporción de casos dentro del intervalo que nos interesa estudiar? Se han construido ciertas tablas con ese fin, lo que se conoce como forma estándar de la curva normal, expresadas en términos de puntuaciones Z (véase el Anexo 1, tabla B). La tabla asume que el área bajo la curva normal suma 1,0, con 0,5 a la izquierda de la media y 0,5 a la derecha. Los valores de Z están dados en los márgenes de la tabla y a lo largo del eje superior. Los primeros dos dígitos de Z se obtienen leyendo hacia abajo y el tercero leyendo a lo largo; la página izquierda de la tabla es para los valores Z negativos, es decir, valores inferiores a la media, y la página derecha para los valores Z positivos, mayores que la media. En este caso, nos interesa un valor Z de +1,50, así que buscaremos en la columna izquierda de la página derecha el valor Z = 1,5, mientras que, a lo largo del eje superior, nos detendremos en la primera columna, que corresponde a Z = 1,50. El resultado indica la proporción del área total bajo la curva entre el valor Z y el extremo derecho de la curva. En este caso, ese valor es 0,06681 o 6,7 %. Pero lo que pretendemos no es encontrar el área entre Z = 1,5 y el extremo derecho, sino entre la media y Z = 1,5. Sabemos que la proporción entre la media y el extremo derecho es 0,5, así la proporción entre la media y Z = 1,5 ha de ser 0,5 − 0,06681 = 0,43319. Redondeando las dos últimas cifras, obtenemos 0,433 o 43,3 % de la curva entre la media y una línea que pase por Z = 1,5. Traduciendo esto a nuestro ejemplo, podemos decir que el 43,3 % de las longitudes de las puntas de proyectil están antre 110 y 140 mm (véase fig. 8.5).

Si nos hubiesen pedido encontrar la proporción de longitudes entre 110 y 80 mm, o 1,5 desviaciones típicas menos que la media, hubiésemos buscado en la tabla el valor correspondiente a Z −1,50, que es 0,93319; es decir, un 93,3 % del área total bajo la curva está entre una línea que pasa por Z = −1,50 y el extremo derecho de la curva. Nos interesa el área entre Z = −1,5 y la media, de forma que restamos 0,5 para obtener 0,43319; no es ninguna sorpresa que esa cifra coincida con la correspondiente al área entre Z = +1,50 y la media. Si nos hubiesen preguntado la proporción de longitudes entre 80 y 140 mm, o dentro de 1,5 desviaciones típicas a cada lado de la media, nos hubiésemos limitado a doblar el valor para cada una de las mitades: 0,433 + 0,433 = 0,866. Obviamente, la proporción o porcentaje puede ser traducida fácilmente en nú-

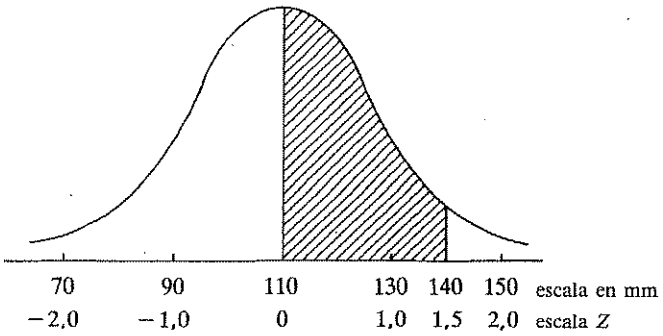


FIGURA 8.5. Distribución de las longitudes de las puntas de flecha con las puntuaciones Z correspondientes a sus valores efectivos en la desviación típica.

meros reales, si fuese necesario, siempre y cuando supiésemos el número total de observaciones de nuestra distribución.

Si la incógnita a resolver fuese la proporción de puntas de flecha con longitudes mayores de 140 mm, el problema habría sido menos complicado. Hubiésemos tenido que saber tan sólo el área entre $Z = +1{,}5$ y el extremo derecho de la curva. Esto lo obtenemos fácilmente por medio del valor $Z = 1{,}50$ en la tabla, tal y como ya habíamos hecho para responder a la primera cuestión: el 6,7 % del área bajo la curva está entre $Z+1{,}50$ y el extremo derecho, de forma que el 6,7 % de las puntas tienen una longitud mayor de 140 mm.

Para puntas menores de 80 mm el procedimiento es similar al primero de los dos casos que hemos visto. El área bajo la curva correspondiente a $Z = -1{,}50$ es 0,93319, tal y como acabamos de ver, por lo que tenemos $1{,}0 - 0{,}93319 = 0{,}06681$, o 6,7 %.

No todas las tablas de la distribución normal estandarizada son iguales a la que figura como la tabla B del anexo 1, si bien son muy similares y no es difícil trabajar con ellas.

El cálculo de las proporciones de la longitud de las puntas de proyectil e intervalos específicos de la distribución global de longitudes tiene interés en sí mismo, siempre y cuando dispongamos de hipótesis específicas acerca del significado cultural o funcional de la longitud de la punta de la flecha. Ahora bien, en este caso en concreto nuestro único propósito era ilustrar la forma en que la distribución normal estándar y los datos reales se relacionan una con otros.

En efecto, lo que hacemos al realizar las operaciones anteriores es efectuar una estandarización de los datos originales. Partimos de una distribución normal en concreto, con una media y una desviación típica expresadas en términos de las unidades en las que se han hecho las observaciones; milímetros en el caso anterior. A continuación, reexpresamos las observaciones en términos de unidades de desviación típica a cualquiera de los lados de la media. La media llega a ser cero, por lo que las observaciones menores que la media serán cantidades negativas, y aquellas mayores que la media serán positivas, con lo que la distribución adquirirá una media de cero y una desviación típica de uno. No importa cuáles sean las unidades originales de medida, podemos convertir cualquier distribución normal en forma de unidades de desviación típica, y tendrá las propiedades que, como hemos visto, caracterizan la distribución normal, en términos de la proporción del área bajo la curva, o si los casos están dentro de la distribución, en el interior de un intervalo dado, según la información de la tabla Z.

La manera más obvia en que la distribución normal es utilizable en arqueología es en la presentación de las fechas de radiocarbono, donde los datos están dados en forma de media y desviación típica (para una discusión más detallada de esta cuestión véanse Thomas, 1976; Orton, 1980). Es muy fácil olvidar que existe sólo una probabilidad del 68,26 % que los datos se encuentren sólo a una desviación típica de la media. La práctica estadística convencional indica que no debiéramos quedar satisfechos con una probabilidad menor del 90 % o 95 %. El problema es que los intervalos de tiempo de ± 2 desviaciones típicas son generalmente tan amplios que, consciente o inconscientemente, los arqueólogos prefieren omitirlos e incurrir en la precisión espuria.

¿QUÉ HEMOS DE HACER SI LOS DATOS NO ESTÁN DISTRIBUIDOS NORMALMENTE?

La cuestión, de hecho, plantea cómo podemos llegar a saber si nuestros datos están o no distribuidos normalmente. Hay varios métodos para averiguarlo. Uno es representando gráficamente la distribución de la frecuencia acumulativa de los datos sobre un papel cuadriculado especial, llamado *papel de probabilidad aritmética* (véase fig. 8.6). Como se puede ver, la escala horizontal está representada en unidades regulares e iguales para el rango de la variable estudiada, pero la escala vertical (dividida en 1.000 partes) registra la distribución acumulativa de las observaciones en una escala variable, de forma que la distancia vertical entre el 50-60 % (500-600 en esa escala) es similar a la distancia vertical entre el 1-2 %. Debe apreciarse que la escala vertical va de 0,1 a 999,9. Esto es así porque la curva normal es *asíntota*: se aproxima a cero en cualquiera de los dos extremos sin que llegue a alcanzarlo, por lo que 0 % y 100 % (0 y 1.000) están distanciados infinitamente. La escala constante horizontal y la escala variable vertical tienen el efecto de convertir la curva acumulada de una distribución normal en una línea recta. Alternativamente, existen programas de ordenador para realizar lo mismo.

Otros dos métodos muy útiles de comprobación de la normalidad ya han sido descritos en la sección del capítulo 4 dedicada al análisis de datos exploratorio. Un estudio de los intervalos entre el valor mínimo, el umbral inferior,

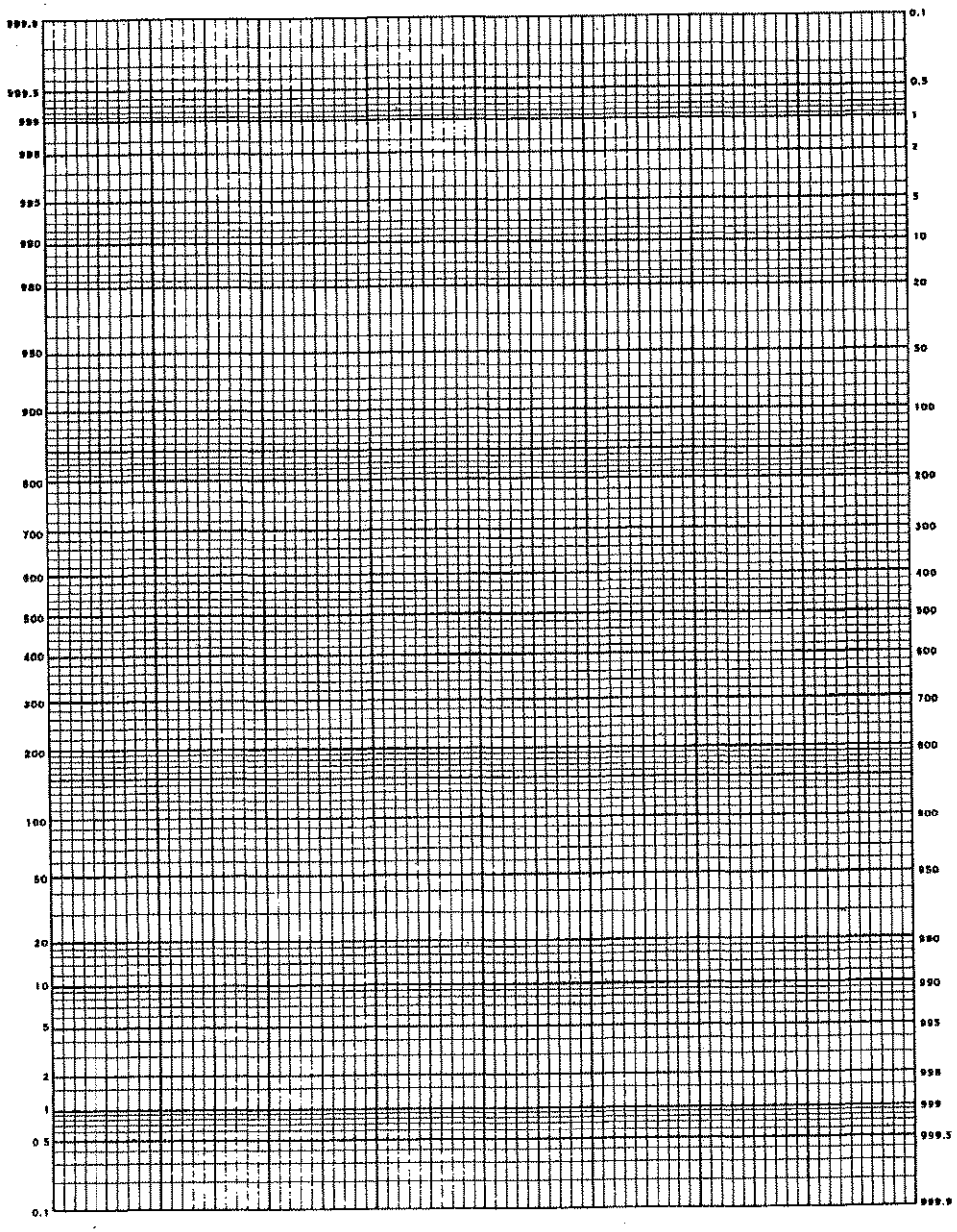


FIGURA 8.6. Un ejemplo de papel de probabilidad aritmética.

la mediana, el umbral superior y el valor máximo dará una buena idea de la simetría global y del grado de concentración de los valores centrales de la distribución. El uso de los gráficos de caja y arbotante pone de manifiesto las peculiaridades de las colas de la distribución. Esto es especialmente importante porque puede que sea sólo ahí donde se observe una desviación con respecto a la normalidad.

¿Qué sucede si los datos no son normales y queremos que lo sean por alguna razón, por ejemplo la aplicación de un método que presuponga distribuciones normales? Sin duda podemos hacerlo por medio de transformaciones. La estandarización Z ya ha sido descrita en este capítulo, si bien se limita a cambiar la escala original en una nueva, sin afectar la forma de la distribución. Otras transformaciones pueden aplicarse a los datos para cambiar la forma de la distribución, cambiando las longitudes relativas de las distintas partes de la escala.

Hace algún tiempo hubo un cierto debate acerca del empleo y validez de la transformación de los datos; algunos pensaban que se trataba de una forma de «camuflarlos». El punto de vista adoptado aquí es que las transformaciones constituyen un útil apropiado para el análisis de datos, como cualquier otro; de hecho, ya hemos visto su uso en el capítulo anterior, donde los modelos logarítmicos se basaban en los logaritmos de las cantidades que aparecían en las tablas, y no en sus valores originales. El uso de una transformación nos permitió ir más adelante en la comprensión de los datos de lo que hubiera sido posible. Sucede muy a menudo que ciertos esquemas emergen más claramente en los datos transformados que en los no transformados; el uso de algunos métodos, por su parte, exige que los datos aparezcan bajo una forma determinada. Si el método que se quiere usar en particular presupone una distribución normal, entonces no hay razón para no transformarla. ¿Por qué habríamos de privilegiar una forma de escala numérica sobre otras? La única condición es que la transformación sea interpretable, pues tendemos a sentirnos más a gusto con escalas de medida próximas a nosotros en la vida real. Sin embargo, no hay razón para imponer esas restricciones en el análisis de datos.

En la práctica arqueológica, una de las situaciones que aparece más frecuentemente son las distribuciones asimétricas positivas, con una cola superior muy larga. En este caso, la transformación en distribuciones normales es bastante simple. Lo que hay que hacer es recortar la cola superior, mientras que el resto de la distribución se deja igual. Resulta fácil hacerlo calculando la raíz cuadrada de cada observación; un efecto más radical se consigue calculando los logaritmos. Lo veremos más claro con un ejemplo.

Supongamos que hemos realizado una prospección y hemos recogido artefactos líticos procedentes de un área bastante extensa, usando un sistema de cuadrículas. Como resultado, tenemos información acerca de la cantidad de artefactos por cuadro, en cada una de las cuadrículas. Queremos efectuar un análisis de correlación de esos datos (véase el capítulo siguiente), y para ello es pre-

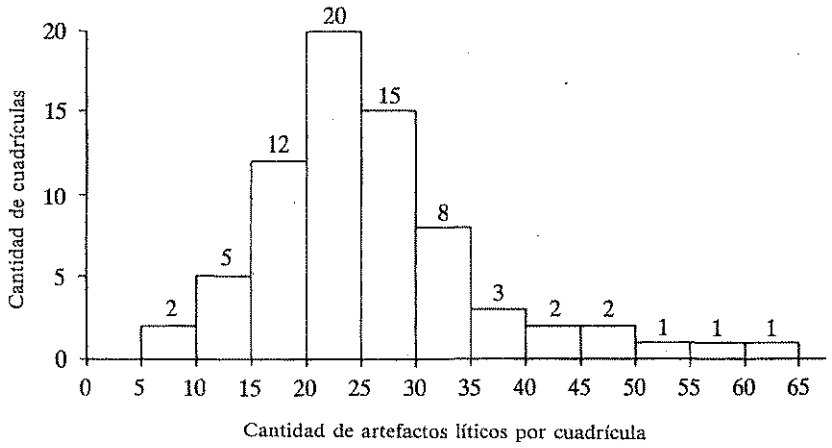


FIGURA 8.7. Distribución de la cantidad de cuadrículas que contienen distintas cantidades de artefactos líticos: datos procedentes de una prospección hipotética.

ferible que los datos estén normalizados. Hemos trazado un histograma con los datos, en el que se aprecia que la distribución es asimétrica positiva, de forma que utilizaremos la técnica de transformación que ya hemos mencionado. Más que transformar cada observación, preferimos transformar el punto medio de cada intervalo; es menos complicado y más sencillo de demostrar. La distribución sin transformar aparece en la figura 8.7.

Si intentamos una transformación por raíces cuadradas, necesitaremos una nueva escala horizontal en unidades de $\sqrt{x}$. Para obtener esto, observamos el valor de los puntos medios de cada intervalo en el histograma original, calculamos su raíz cuadrada y entonces situamos los casos de cada intervalo original en el intervalo transformado que le corresponde. Tal y como se puede ver en la figura 8.8, los datos muestran ahora una mayor aproximación a la normalidad.

Si hubiésemos realizado una transformación logarítmica, habríamos necesitado una nueva escala en unidades de log x (aquí logaritmo en base 10). Por analogía con el caso de la raíz cuadrada, calculamos el logaritmo del punto medio de los intervalos, cambiamos la escala y trazamos el nuevo histograma (fig. 8.9).

De hecho, tal y como puede verse, en este caso el resultado es muy semejante en ambas transformaciones, la raíz cuadrada y el logaritmo. Esto se debe a que la cola positiva no era muy larga. Supongamos que la mayor de las observaciones sea 1.000.000. La raíz cuadrada de ese número es 1.000, pero su logaritmo es 6, con lo que la diferencia entre ambos es considerable. Como principio general, los logaritmos son apropiados para datos positivos en los que los valores están próximos a cero (por ejemplo, densidades), mientras que las raíces cuadradas suelen usarse para transformar datos en forma de frecuencias.

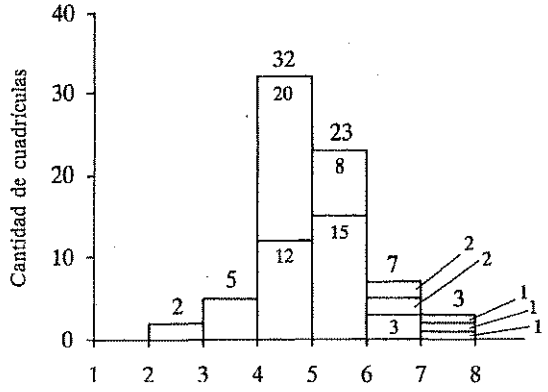
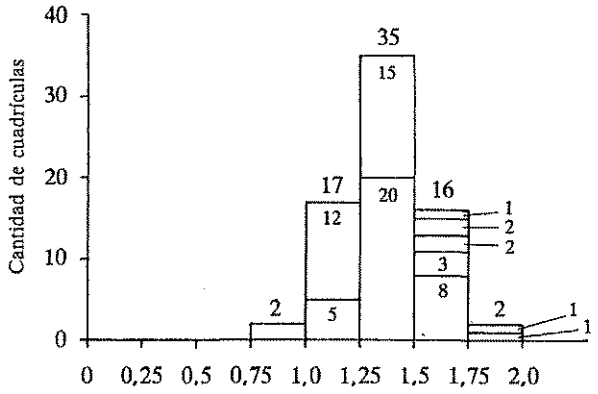


FIGURA 8.8. Distribución de la cantidad de cuadrículas que contienen distintas cantidades de artefactos líticos: cantidad de artefactos por cuadrícula, habiendo sustituido el valor original por su raíz cuadrada.



Logaritmo del punto	0,875	1,096	1,352	1,511	1,759
central de cada clase		1,243	1,439	1,574	1,795
				1,628	
				1,676	
				1,720	

FIGURA 8.9. Distribución de la cantidad de cuadrículas que contienen distintas cantidades de artefactos líticos: cantidad de artefactos por cuadrícula, habiendo sustituido el valor original por su logaritmo común.

34

EJERCICIOS

8.1. Se ha encontrado un conjunto de recipientes cuya capacidad media es 950 ml con una desviación típica de 56 ml. La forma de la distribución de volúmenes es normal. a) ¿Qué proporción de recipientes tiene una capacidad cúbica mayor que 1.050 ml? b) ¿Qué proporción tiene una capacidad menor que 800 ml? c) ¿Qué proporción de capacidades existe entre 900 y 1.000 ml?

8.2. En la investigación de un conjunto de bifaces se ha decidido estudiar la relación entre el peso y otras variables. Los métodos que se precisan exigen que la distribución de pesos esté normalizada. Dada la tabla que muestra la distribución de las frecuencias, comprueba si es normal y, si no lo es, normalízala.

Intervalo (g)	N.º de bifaces	Intervalo (g)	N.º de bifaces
200-249	5	650- 699	3
250-299	10	700- 749	3
300-349	13	750- 799	2
350-399	17	800- 849	2
400-449	13	850- 899	2
450-499	8	900- 949	1
500-549	5	950- 999	1
550-599	4	1.000-1.049	1
600-649	4		

9. RELACIONES ENTRE DOS VARIABLES NUMÉRICAS: CORRELACIÓN Y REGRESIÓN

MÉTODOS GRÁFICOS: DIAGRAMAS DE DISPERSIÓN

El estudio de las relaciones entre dos variables numéricas tiene una gran ventaja sobre el estudio de las relaciones entre variables nominales: las relaciones pueden presentarse en forma gráfica, por medio del *diagrama de dispersión*, en el que se representa una variable en relación a la otra. Este tipo de figuras puede proporcionar mucha información, e incluso prevenir de ciertos errores en los que incurriríamos de fijarnos sólo en los resúmenes numéricos de las relaciones.

TABLA 9.1. Cantidades de cerámica de New Forest encontrada en yacimientos situados a distintas distancias de los hornos de producción.

Yacimiento	Distancia (km)	Cantidad (fragmentos por m^3 de tierra)
1	4	98
2	20	60
3	32	41
4	34	47
5	24	62

Para cada observación tenemos un valor en cada una de las dos variables. Supongamos que nos interesa estudiar la relación entre la cantidad de cerámica procedente de los hornos britanorromanos de New Forest, en el sur de Inglaterra, que alcanzó lugares muy alejados del centro de producción. Disponemos de la información en la tabla 9.1. Podemos trazar un diagrama de dispersión en el que la distancia aparece en el eje horizontal y la cantidad de cerámica

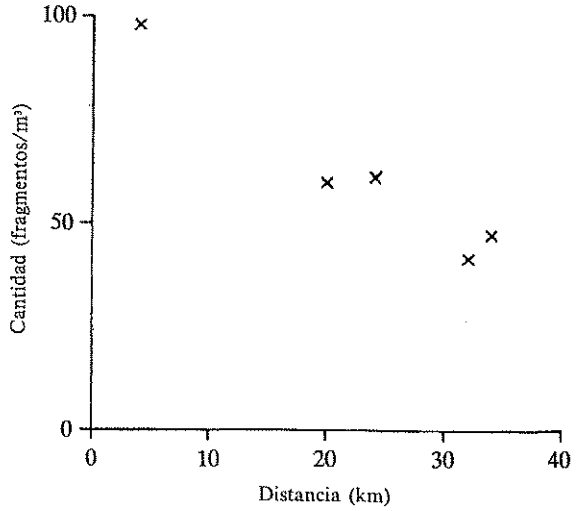


FIGURA 9.1. Gráfico de la cantidad de cerámica britanorromana procedente de los hornos de New Forest, encontrada en distintos yacimientos, en relación con la distancia entre esos yacimientos y los hornos de procedencia.

en el eje vertical. Cada yacimiento se sitúa en el punto que le corresponde del eje horizontal y del vertical, según sus valores, en ambas variables (fig. 9.1).

Este gráfico ya es de por sí extremadamente rico en información. Podemos ver que la cantidad de cerámica disminuye a medida que aumenta la distancia desde el centro de producción. Podemos ver también que la relación adopta la forma de una línea recta: es posible trazar una línea recta que pase muy cerca de todos los puntos. En otras palabras, para un incremento específico de la distancia, se constata una disminución específica en la cantidad de cerámica; todos los yacimientos, en mayor o menor grado registran este comportamiento.

En este caso podemos decir que una de las variables es independiente y la otra dependiente. Imaginemos que la cantidad de cerámica está afectada, de algún modo, por la distancia y, por tanto, depende de ella; la inversa, que la distancia esté afectada por la cantidad de cerámica, no se mantiene. En tales circunstancias, por convención, la variable independiente se representa como eje horizontal (o *x*) y la dependiente en el vertical (o *y*).

No siempre podemos especificar las variables dependiente e independiente. Supongamos que estamos estudiando las dimensiones de unos vasos neolíticos de Hungría y las relaciones entre ellos, para poder caracterizar los principales aspectos de variación morfológica (véase Whallon, 1982, para un estudio de recipientes neolíticos de Suiza); podemos trazar la altura y el diámetro del bor-

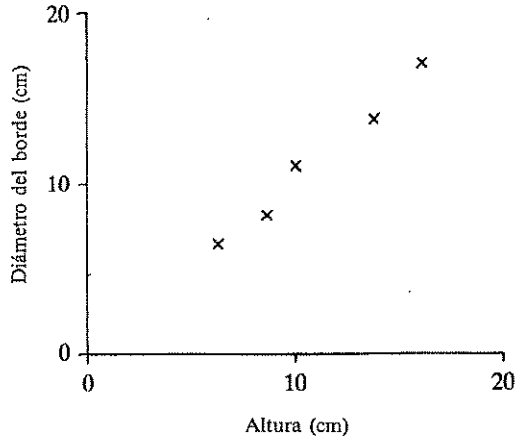


FIGURA 9.2. Diámetro del borde en relación con la altura en un grupo hipotético de vasos campaniformes en Hungría.

de (fig. 9.2). Se observa que ambas variables coinciden, de forma que entre el diámetro del borde y la altura hay una proporción constante. En este ejemplo, el diagrama de dispersión desempeña el mismo papel que en el caso anterior, si bien no hay razón alguna que nos permita afirmar que una de las variables depende de la otra.

Los diagramas de dispersión constituyen el más importante medio de estudio de las relaciones entre pares de variables. Gracias a ellos nos hacemos una idea de la *dirección* de una relación: ¿es positiva o negativa? La relación altura/diámetro del borde es un ejemplo de la primera: a medida que la altura aumenta, también lo hace el diámetro del borde. En el caso de la cantidad de cerámica y la distancia, se trata de de una relación negativa: a medida que aumenta la distancia, decrece la cantidad de cerámica.

El diagrama de dispersión nos explica también la forma de la relación. En los dos casos anteriores hemos obtenido una línea recta o relación *lineal*. Pero no todas las relaciones son de ese tipo. De hecho, los gráficos que representan la relación entre la cantidad de una mercancía cualquiera y la distancia a su centro de producción suelen ser como el que aparece en la figura 9.3. Esta relación es curvilínea, pero aún *monótona*, es decir, a todo lo largo de la relación, a medida que aumenta la distancia, disminuye la cantidad; no es el caso de que, para una parte de la distancia, la cantidad disminuya y para la parte siguiente vuelva a aumentar.

Un ejemplo de relación *no monótona* sería el de la figura 9.4, que interesaría a un arqueólogo que intentase estimar la edad de los animales cuyos huesos estudia. En este caso, la altura del diente (primer molar) va aumentando desde

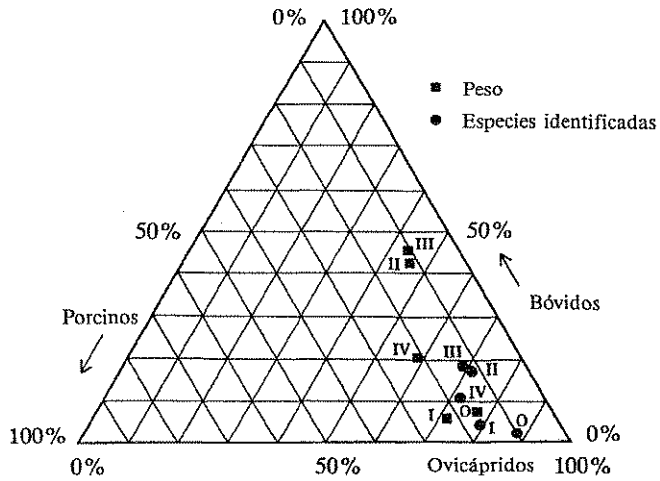


FIGURA 9.6. Gráfico tripolar de los porcentajes faunísticos en cinco de las fases de Filácope: fases 0, I (bronce inicial), fases II, III (bronce medio), fase IV (bronce final) (según Gamble, 1982).

DESCRIBIENDO RELACIONES POR MEDIO DE CIFRAS

La forma de una relación

El procedimiento de describir una relación se denomina *regresión* y el de medida del ajuste a los datos, *correlación*.

La regresión se diferencia de las técnicas que hemos ido viendo hasta ahora (con la excepción de los métodos descritos al final del capítulo 7) en que no descubre si existe o no una relación, o la intensidad de la misma, sino su naturaleza. Por ese motivo es importante, no sólo en la estadística clásica, sino también en la construcción de modelos, pues trata acerca de la predicción. Usamos una variable independiente para estimar los valores de una variable dependiente.

La forma más general de definir matemáticamente una relación hipotética es $y = f(x)$. Esta ecuación no nos dice mucho, tan sólo que el valor de y (la variable dependiente) en un punto en particular es una función del valor de x (la variable independiente) en ese mismo punto. No dice nada acerca de la naturaleza específica de la relación, aunque no sería difícil expresarla por medio de algunas figuras. Por ejemplo,

$$y = x \quad \text{(a)}$$
$$\text{o} \quad y = 2x \quad \text{(b)}$$
$$\text{o} \quad y = x^2 \quad \text{(c)}$$

38

Se verá mucho más claro si lo explicamos con palabras: (a) nos explica que el valor y en un punto dado es idéntico al valor de x en ese punto; (b) nos dice que el valor de y es el doble del valor de x; (c) expresa que el valor de y en un punto dado es el cuadrado del valor de x en ese punto. Tales ecuaciones pueden representarse por medio de líneas en un gráfico; las de (a), (b) y (c) aparecen en la figura 9.7.

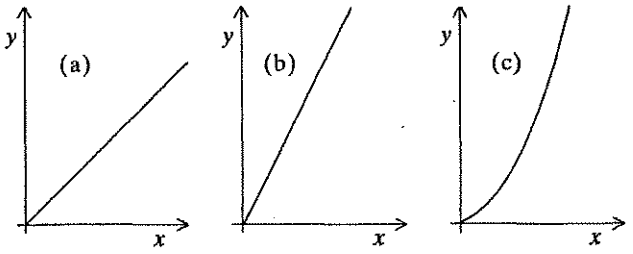


FIGURA 9.7. Gráfico de las ecuaciones: (a) $y = x$; (b) $y = 2x$; (c) $y = x^2$.

Si la especificación que hemos hecho de la relación entre dos variables por medio de una de estas funciones es perfecta, podremos predecir el valor de y en un punto a partir del conocimiento que tengamos de x en el mismo punto. Por ejemplo, si hubiese una relación perfecta en un caso particular entre la densidad de hallazgos de obsidiana en un yacimiento y la distancia entre el yacimiento y la fuente de la obsidiana, entonces, conociendo la distancia, podríamos predecir exactamente la densidad de hallazgos; o si hubiese una relación perfecta entre la altura y el diámetro del borde de un grupo de recipientes, conociendo el diámetro del borde podríamos predecir exactamente la altura (y viceversa, en este caso).

Naturalmente, en la mayoría de los casos, incluso en las ciencias naturales «duras», las cosas nunca son totalmente predecibles. En muchos casos, particularmente en las ciencias experimentales, se debe a la imperfección de nuestras técnicas de medida; a veces, porque los efectos suelen ser el resultado de muchas y diversas causas operando juntas, muchas de las cuales están sujetas a su vez a influencias aleatorias. Lo que hemos de hacer, entonces, es buscar tendencias generales en nuestros datos, estimando la relación entre x e y, y también la exactitud con que los valores de y pueden derivarse de la estimación de esa relación.

La situación en la que existe una relación puede compararse a la situación contraria, cuando x e y son independientes. En este caso no se podrá predecir y a partir de x, o mejor dicho, el conocimiento de x no mejorará la predicción de y; cuanto mayor es la dependencia, mejor será la predicción.

Más adelante veremos cómo es el gráfico el que proporciona un vínculo entre los diagramas de dispersión que ya hemos visto y las ecuaciones matemáticas.

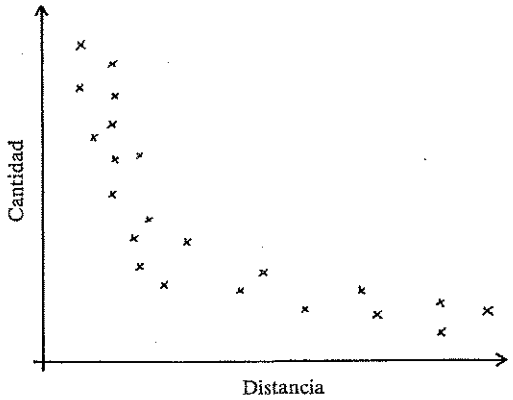


FIGURA 9.3. Gráfico de las cantidades hipotéticas de mercancías que llegan a ciertos yacimientos, en relación con la distancia entre esos yacimientos y la fuente de procedencia de las mercancías. El resultado muestra una relación curvilínea.

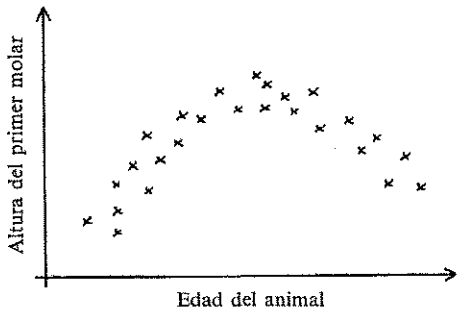


FIGURA 9.4. Gráfico de la altura del primer molar en relación con la edad del animal al morir. Datos: mandíbulas de oveja, cuya edad es conocida.

que sale y a medida que crece el animal, pero a medida que se usa ese diente para masticar, se va erosionando y su altura disminuye.

El diagrama de dispersión también nos proporciona una idea de la intensidad de una relación. Comparemos los dos gráficos de la relación entre peso y cantidad de lascas que se han extraído de dos grupos hipotéticos de bifaces de sílex, procedentes de las terrazas del valle del Támesis, en el sur de Inglaterra (fig. 9.5). En un caso la relación es claramente más intensa que en el otro, porque los puntos están mucho más concentrados que en el otro: están mucho más próximos de cualquier línea recta que tracemos a través de la nube de puntos.

Diagramas de dispersión como estos, con dos ejes perpendiculares entre sí, uno para cada variable, son los más habituales en arqueología y en otras disci-

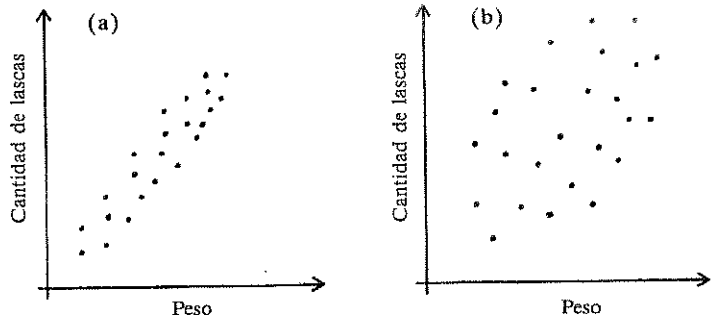


FIGURA 9.5. Diagrama de dispersión del peso en relación con la cantidad de lascas para dos grupos de bifaces.

plinas. Sin embargo, hay otra forma que merece mención, el *gráfico tripolar*, un ejemplo del cual aparece en la figura 9.6. Como puede verse, hay tres ejes unidos por ángulos de 60°, y formando un gráfico de forma triangular. Esos gráficos pueden usarse para trazar no sólo dos, sino hasta tres variables, en los casos en los que las tres variables adopten una escala cerrada. Una escala cerrada es aquella que tiene una suma fija, como, por ejemplo, la escala porcentual. Los conjuntos faunísticos procedentes de yacimientos suelen describirse a menudo como porcentajes de huesos de distintas especies. Muy a menudo, en los contextos agrícolas europeos o del Próximo Oriente sólo interesan tres especies: vacuno, cerdo y ovicápridos. Para un conjunto en particular, si el vacuno alcanza el 20 % y el cerdo el 30 %, los ovicápridos alcanzarán el 50 %, si no los porcentajes no sumarían 100.

La figura 9.6 muestra un gráfico tripolar de los conjuntos de fauna procedentes de las sucesivas fases de ocupación del yacimiento de la edad del bronce de Filácope, en la isla de Melos (Gamble, 1982). Para explicar la forma en que debe leerse, tomemos la fase 0, por el peso. Si observamos la escala de ovicápridos, veremos que su valor es aproximadamente del 76 %; en la escala de porcinos, cerca del 17 % y en la escala de bóvidos, alrededor del 7 %. En la fase III, por el peso, tenemos 44, 11 y 45 % respectivamente. Usando el gráfico podemos trazar fácilmente las tendencias temporales en la composición de los conjuntos faunísticos de Filácope.

En cualquier estudio de las relaciones entre variables de escala interválica, siempre es esencial, como primer paso, trazar el diagrama de dispersión y ver qué forma adopta. Pero no hemos de contentarnos sólo con esto.

Es muy posible que nos interese también describir matemáticamente las relaciones descubiertas en el gráfico, quizás para poder compararlas con otros conjuntos de datos. Igualmente, nos puede interesar definir matemáticamente la intensidad de una relación: ¿qué grado de ajuste a los datos tiene la relación propuesta?

37

Si pensamos en el ejemplo más conocido de análisis de regresión en arqueología: la disminución de la cantidad de un tipo peculiar de material con respecto a la distancia hasta la fuente de dicho material —por ejemplo, la obsidiana de Lípari en el Mediterráneo occidental—, podemos imaginar que para cada valor fijo de la variable independiente, la distancia, habrá una distribución de cantidades de material. No todos los yacimientos situados a la misma distancia tienen la misma cantidad de ese material; pero cada una de esas distribuciones (para cada uno de los valores de la variable distancia) tendrá una media, por lo que podremos trazar su posición en el gráfico. La línea trazada a través de todas las medias de y con valores fijos de x se conoce como la *ecuación de regresión*. (En la práctica es muy raro el caso en el que haya distintos valores de y con un valor dado de x; esto ocurre sólo en el caso de experimentos diseñados precisamente para hacerlo. El método no depende de eso, sin embargo; más adelante se verá el supuesto sobre el que se basa.)

Esa línea puede adoptar cualquier forma, si bien aquí sólo trataremos el caso más simple, cuando la ecuación de regresión es lineal y la relación adopta la forma de una línea *recta*. No se trata de una limitación, como pudiera pensarse, pues muchas relaciones empíricas adoptan esa forma y porque, a menudo, resulta posible transformar las variables de forma que lleguen a ser lineales. Estas relaciones tienen la virtud de ser más fáciles de comprender intuitivamente. Podemos describir la ecuación para una relación lineal, de este modo:

$$y = a + bx$$

donde y es la variable dependiente, x la independiente, y los coeficientes a y b son constantes, es decir, están fijos para un conjunto de datos específico.

Si $x = 0$, la ecuación se reduce a $y = a$, por lo que a representa el punto en que la recta de regresión cruza el eje y (véase fig. 9.8); lo que se conoce habitualmente como punto de corte. La constante b define la pendiente de la recta de regresión, la cantidad de cambio en una dirección vertical (a lo largo del eje y) para una distancia horizontal dada (a lo largo del eje x). Así, en el caso de cantidades de cerámica y la distancia al centro de producción, el valor b representa el incremento de la cantidad de cerámica para un aumento dado en la distancia al centro de producción (está calculado más adelante, p. 134); para el ejemplo de la altura y el diámetro del borde, es el incremento del diámetro del borde asociado a un incremento dado en la altura. La figura 9.8 ilustra esta cuestión.

A medida que la pendiente se acentúa, la cantidad de cambios en y en relación a un cambio dado en x se hace mayor. Cuando esa línea llega a ser horizontal, es decir, $b = 0$, no hay ningún cambio en y en relación a x. Evidentemente, esto significa que no hay relación entre las dos variables, o bien, visto de otra manera, conocer los valores de x en un conjunto de observaciones no ayuda predecir sus valores y. Pero hemos de matizar esta afirmación. Si $b = 0$, no hay relación *lineal* entre las dos variables, aunque ciertas formas de relación

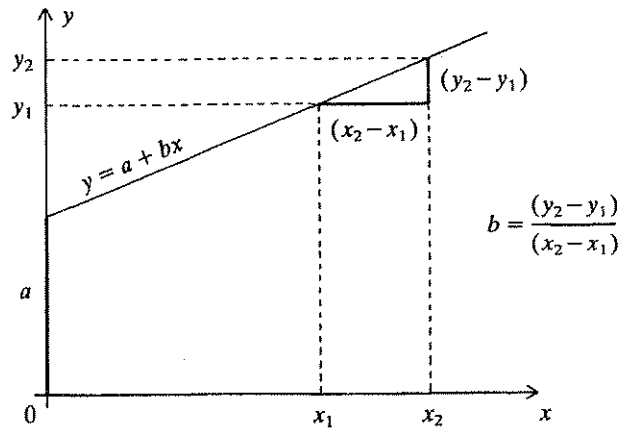


FIGURA 9.8. Pendiente y punto de corte (coeficientes a y b) de una recta de regresión.

no lineal pueden tener ese valor en su coeficiente b. Finalmente, si y decrece a medida que x aumenta, en otras palabras, si la relación es inversa, el signo del coeficiente b será negativo. Veamos cómo funciona todo esto por medio de un ejemplo.

Una vez realizado un diagrama de dispersión de la relación entre la cantidad de cerámica y la distancia al centro de producción (fig. 9.1), basándonos en la información de la tabla 9.1, pretendemos describir matemáticamente esa relación. Esto supone encontrar los valores del punto de corte y la pendiente apropiados (coeficientes a y b) a ese conjunto de datos en particular, e incluirlos en la ecuación $y = a + bx$. Sin embargo, una simple mirada al diagrama de dispersión (fig. 9.1) nos mostrará en seguida que la relación dista de ser perfecta: no hay ninguna línea recta que atraviese exactamente todos los puntos. Lo que intentaremos, entonces, es buscar la recta que proporcione un *mejor ajuste* a los puntos. ¿Cómo hacerlo?

Una forma intuitiva sería trazando el diagrama de dispersión y dibujando a ojo la recta que pase por más puntos. Podemos entonces calcular la pendiente e interceptar valores para la línea. Pero esta forma no es satisfactoria. El método usual para ajustar una recta a la nube de puntos es el llamado de los *mínimos cuadrados*. Para cada punto anotamos su valor y. Es obvio que el valor y predicho por la regresión en ese punto no será exactamente el mismo valor y observado: habrá alguna discrepancia. La figura 9.9 muestra qué significa esto, bajo la forma de un segmento en particular de la recta de regresión.

El método de los mínimos cuadrados define aquella recta que minimice la siguiente expresión:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

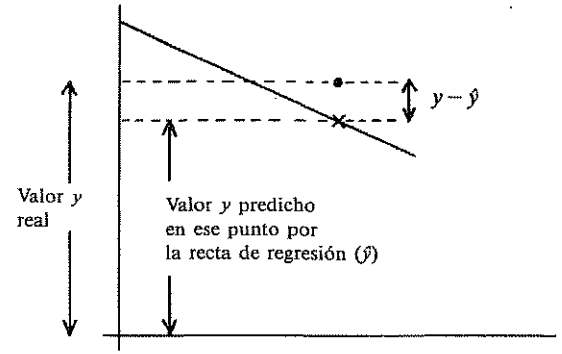


FIGURA 9.9. Diferencia entre el valor y real en un punto y el valor predicho por la regresión.

donde n = cantidad de datos; y_i = el valor y real en un punto i; $\hat{y}_i$ = el valor del punto i predicho por la regresión. Veamos lo que esto significa en palabras más simples.

Para cada punto obtenemos la diferencia entre sus valores y reales y los predichos (en nuestro ejemplo, la diferencia entre la cantidad de cerámica de New Forest en un yacimiento y la cantidad de esa cerámica en ese yacimiento predicha por la recta de regresión); ese es el término $(y_i - \hat{y}_i)$. A continuación debemos elevar al cuadrado esa diferencia, y repetir la operación para todos y cada uno de los puntos, sumando, finalmente, los resultados. Este total ha de ser minimizado; en otras palabras, hemos de ir ensayando posiciones de la recta de regresión hasta que encontremos aquella que produzca la menor suma posible de diferencias al cuadrado entre los valores actuales y los predichos.

Usamos las desviaciones al cuadrado por la misma razón que usamos las desviaciones de la media al cuadrado para definir la dispersión en el cálculo de la varianza y de la desviación típica de una variable: si nos limitamos a calcular las diferencias sin elevarlas al cuadrado, el resultado de su adición sería cero. Inevitablemente, sin embargo, el resultado de este procedimiento es que la pendiente y la posición de la recta de regresión están afectadas por los puntos con las mayores desviaciones con respecto a la media; ésa es una de las razones de la debilidad de la regresión por mínimos cuadrados, pues uno o dos valores extremos puede tener un gran efecto sobre los resultados.

Lo que se minimiza es, en realidad, la suma de las distancias *verticales* al cuadrado, ya que estamos interesados en la regresión de y sobre x, o el efecto de x sobre y. Si quisiéramos hacer la regresión de x sobre y, usaríamos las distancias horizontales. En capítulos posteriores, veremos algunos métodos que implican el uso de distancias perpendiculares a la recta mejor ajustada.

De hecho, no necesitamos hacer probaturas para fijar la posición de la recta mejor ajustada que satisfaga el criterio de los mínimos cuadrados. Se han

obtenido ecuaciones que permiten el cálculo de los coeficientes a y b apropiados para cualquier conjunto de datos:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

En palabras, empezando por el numerador: tomamos el valor de x de un punto en particular y le restamos la media de las x. Tomamos a continuación el valor y de ese punto y le restamos la media de las y. Habiendo hecho esto, multiplicamos las dos cantidades resultantes. Se efectúa esta operación en cada uno de los puntos y se suman los resultados. Esta cantidad, denominada *covariación* entre x e y, es dividida por el denominador. Para este último tomamos el valor x de cada punto, le restamos la media de las x, elevamos al cuadrado la diferencia obtenida, repetimos la operación para todos los puntos y sumamos de nuevo los resultados. Esta suma se utiliza para dividir la suma del numerador y obtener el valor b, la pendiente de la recta de regresión.

Para el coeficiente a, tenemos:

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}$$

Como es fácil de ver, es mucho más sencillo de calcular.

Para b hay otra versión de la fórmula, que, en general, es menos tediosa y más fácil de calcular manualmente, si bien esto último tiene cada vez menos importancia:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

en donde n es la cantidad de puntos; $\sum_{i=1}^{n} x_i y_i$ significa que, para cada punto, multiplicamos el valor x por el valor y y los sumamos todos; $(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)$ significa que sumamos todos los valores x de los puntos, luego todos los valores y, y multiplicamos los dos totales. La misma distinción se produce en el denominador entre $\sum_{i=1}^{n} x_i^2$ y $(\sum_{i=1}^{n} x_i)^2$.

TABLA 9.2. Cantidades de cerámica de New Forest encontrada en yacimientos situados a diferentes distancias de los hornos.

Yacimiento	Distancia (x) (km)	Cantidad (y) (fragmentos por m³ de tierra)
1	4	98
2	20	60
3	32	41
4	34	47
5	24	62

A continuación podemos calcular los valores de a y b para describir la relación entre la cantidad de cerámica y la distancia a los hornos. (Los datos se han reproducido de nuevo en la tabla 9.2 para mayor comodidad.) Usando la fórmula anterior para b, las diversas cantidades relevantes para su cálculo son como sigue: $n = 5$, $\sum y_i = 308$; $\sum x_i = 114$; $\sum x_i y_i = 5.990$; $\sum x_i^2 = 3.172$. Así:

$$b = \frac{(5 \times 5.990) - (114 \times 308)}{(5 \times 3.172) - 12.996} = -\frac{5.162}{2.864} = -1,80$$

Habiendo obtenido ya el coeficiente b, necesitamos el valor del punto de corte:

$$a = \frac{\sum y_i - b\sum x_i}{n} = \frac{308 - (-1,8 \times 114)}{5} = \frac{513,2}{5} = 102,64$$

Basándonos en esta información podemos escribir la ecuación de regresión como:

$$\hat{y} = 102,64 - 1,8x$$

que afirma que en el centro de producción han de haber 102,64 fragmentos de cerámica por m³ de tierra de acuerdo con la recta de regresión, y que esta cantidad declina 1,8 fragmentos/m³ por cada kilómetro que nos alejamos del centro de producción. La recta resultante está trazada en la figura 9.10.

La intensidad de la relación: la correlación

Hasta ahora hemos visto cómo establecer dos parámetros de una ecuación de regresión, a y b, y de este modo indicar la forma de la relación entre x e y. Pero esto no nos explica nada acerca de la precisión de las estimaciones de y proporcionadas por la recta de regresión. Para descubrir la precisión de una recta, habremos de emplear el coeficiente de correlación, que mide la intensidad de la relación entre dos variables. La intensidad de una relación es una cuestión con la que ya estamos familiarizados, tras haberla estudiado en el capítulo 7, en el caso de las variables nominales. En variables de escala interválica, la idea general es la misma, si bien difieren los detalles de la forma de cálculo.

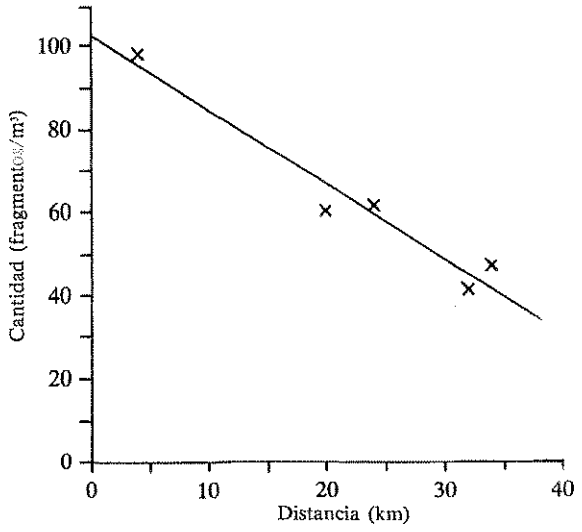


FIGURA 9.10. Gráfico de la ecuación de regresión $\hat{y} = 102,64 - 1,8x$

El coeficiente de correlación ha sido de importancia fundamental para la aplicación de las técnicas cuantitativas en arqueología en los últimos 25 años. Tal y como hemos visto en el capítulo 1, uno de los temas más importantes en la arqueología procesualista fue estudiar la manera en que algo varía en relación con otra cosa. El coeficiente de correlación ha sido, probablemente, la herramienta matemática más importante para el análisis de estructuras de covariación en los datos arqueológicos. Es importante, tanto por sí mismo, como por ser el fundamento de métodos más complejos, tales como el análisis de componentes principales y el análisis factorial (véase el capítulo 12).

Considerado en términos gráficos, el coeficiente de correlación es una medida del grado con que los datos están dispersos alrededor de la recta de regresión. Si están muy próximos a ella, la correlación será intensa y la predicción de los valores de y a partir de x será muy buena. Si los puntos están muy dispersos alrededor de la recta, la correlación será débil y la predicción de y basada en x será pobre. Esta cuestión puede demostrarse claramente refiriéndonos a la figura 9.5. El coeficiente de correlación será más alto en el diagrama de dispersión de la izquierda que en el de la derecha: los puntos en (a) estarían mucho más agrupados alrededor de la recta de regresión. Es obvio que las predicciones de y basadas en x serán mucho mejores en (a). Si nos fijamos en (b)

veremos que para un valor x dado, hay muchos valores posibles de y. Si la nube de puntos es circular, la correlación será cero y el conocimiento que tengamos de x no contribuirá a la predicción de y. Así, el coeficiente de correlación es una medida del grado en que ambas variables *covarían*, aunque es importante recordar que se trata de una medida de correlación lineal, y que ciertos tipos de relaciones curvilíneas pueden producir una correlación de valor cero, aun en el caso de relación perfecta (lo cual demuestra, una vez más, la importancia de un buen estudio previo del diagrama de dispersión; véase fig. 9.4).

Todo esto recuerda mucho la argumentación que hacíamos para el coeficiente b, lo cual no debiera de sorprendernos. Obsérvese, sino, la fórmula para el coeficiente de correlación (r):

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Una versión de la misma, apta para ser calculada a mano:

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}$$

Tal y como puede verse, el numerador de la expresión para r y para b es el mismo, la covariación entre x e y.

La diferencia entre ambos radica en el denominador: para el coeficiente de correlación, la covariación está estandarizada en términos de la variación tanto de x como de y. El máximo valor posible que la covariación puede alcanzar es igual al denominador, la raíz cuadrada del producto de la variación entre x e y. Así, el valor máximo que r puede alcanzar será 1,0: positivo si el término de la covariación es positivo y negativo cuando éste lo sea. El valor máximo es alcanzado cuando todos los puntos se encuentran sobre una línea recta (fig. 9.11). Tal y como ya hemos comentado anteriormente de paso, cuando x e y son independientes, el coeficiente de correlación, al igual que la pendiente de la recta (dado que tienen el mismo numerador), será cero.

Hay, sin embargo, dos importantes diferencias entre r y b, las cuales proceden de las diferencias en el denominador. En primer lugar, el coeficiente de correlación es simétrico, ya que está estandarizado en términos de la variación en ambas variables: no importa cuál de las variables es independiente, ni tampoco si ninguna de ellas lo es; la correlación entre x e y es la misma que la que existe entre y y x. La pendiente de la regresión de y sobre x, sin embargo, no es igual a la pendiente de x sobre y, aunque el ángulo de la recta sea de 45°. La figura 9.10 aclara este problema; muestra la recta para la regresión de la cantidad de cerámica sobre la distancia al centro de producción. Si la volvemos

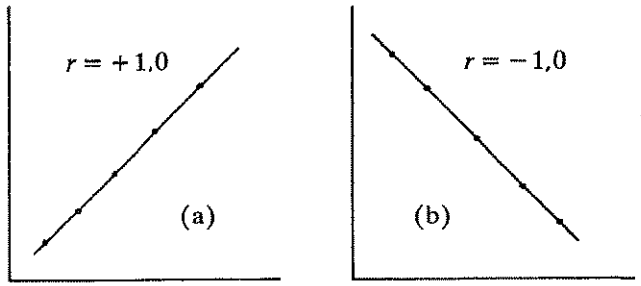


FIGURA 9.11. Diagrama de dispersión y recta de regresión, con: (a) correlación positiva perfecta; (b) correlación negativa perfecta.

del revés por un momento, e imaginamos que la cantidad de cerámica es el eje horizontal, se aprecia que la pendiente con referencia a ese eje es mucho mayor que en relación al eje real x. Es obvio que la proporción de cambios en la cantidad de cerámica a medida que aumenta la distancia al centro de producción es algo muy distinto a la proporción de cambios en la distancia a medida que disminuye la cantidad de cerámica. En segundo lugar, mientras que la pendiente se mide en unidades de las variables originales (por ejemplo, el volumen en que ha disminuido la cerámica por cada incremento dado de la distancia), la correlación es un índice sin unidad a la que referirse, el cual puede usarse como término de comparación en distintas circunstancias.

Antes de que dejemos el coeficiente de correlación, hemos de considerar su valor al cuadrado (r^2). Es conocido como *coeficiente de determinación* y tiene sus propias e interesantes propiedades, que examinaremos a continuación.

Vimos antes que una forma de abordar el análisis de regresión era estudiando si la regresión mejoraba nuestras estimaciones del valor de y en ciertos puntos en particular, usando la información disponible acerca de sus valores x. Si el conocimiento de x mejorase nuestras predicciones de y, significaría que las dos variables están relacionadas de algún modo; así y todo, se deben tener presentes las advertencias expresadas en los capítulos previos: la asociación no necesariamente significa explicación de uno en términos de otro.

Si el conocimiento de x no nos ayuda a predecir y, entonces nuestra mejor estimación de cualquier valor y será la media de y ($\bar{y}$). Tal y como vimos en el capítulo 4, si este es un valor típico o no, depende del grado de dispersión de la distribución alrededor de la media (asumiendo, por el momento, que la distribución sea simétrica y no asimétrica). Por lo tanto, una forma de asegurar la exactitud de una predicción de y basada en $\bar{y}$ es fijándose en la dispersión alrededor de la media, dada por $\sum(y_i - \bar{y})^2$. Esto es, naturalmente, la suma de los cuadrados, o variación en y, la primera fase en el cálculo de la varianza o de la desviación típica.

42

Si a continuación efectuamos la regresión de y sobre x para mejorar nuestra predicción de y, podremos afirmar la precisión de nuestras predicciones fijándonos en la dispersión de las observaciones: no en la media de y, sino alrededor de la recta de regresión, dada por $\sum(y_i - \hat{y})^2$. Esa es la cantidad a la que nos hemos referido antes, y que la regresión por mínimos cuadrados trataba de minimizar. Suele denominarse *variación residual* alrededor de la recta de regresión. $\sum(y_i - \hat{y}_i)^2$ no puede ser en ningún caso mayor que $\sum(y_i - \bar{y})^2$. Precisamente porque es menor, podemos mejorar las predicciones usando la recta de regresión (es decir, el conocimiento de los valores x) como fundamento de la predicción y no la media de los valores y. Por lo tanto, la cantidad de mejora es $\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2$. O bien, dicho en palabras, la variación extraída por la regresión es igual a la variación original menos la variación residual. Este incremento de la «mejora», es decir, la cantidad de variación «extraída» por la regresión, se conoce como variación «explicada», término que, en este contexto, puede inducir a error.

Si dividimos la variación extraída por la regresión por la variación original, obtenemos la proporción de la variación original que puede llegar a ser extraída por la regresión, y que es lo que se conoce como r^2, el coeficiente de determinación. En muchos sentidos, su significación es más intuitiva que el anterior. Su valor suele multiplicarse por 100 para situarlo en una escala porcentual, por lo que recibe el nombre de «porcentaje del nivel de explicación».

Ya va siendo hora de que ilustremos esos coeficientes con ayuda de nuestro ejemplo sobre cerámicas y distancias, para el cual habíamos obtenido una ecuación de regresión. Usando la fórmula de cálculo manual para r:

$$r = \frac{(5 \times 5.990) - (114 \times 308)}{\sqrt{[(5 \times 3.172) - 12.996][(5 \times 20.938) - 94.864]}} =$$

$$= \frac{-5162}{\sqrt{2.864 \times 9.826}} = -0{,}973$$

que nos dice que la relación entre la cantidad de cerámica y la distancia a su centro de producción es, virtualmente, una perfecta asociación lineal negativa, tal y como esperábamos tras haber visto el diagrama de dispersión. Si elevamos al cuadrado ese valor para obtener el coeficiente de determinación tenemos

$$r^2 = -0{,}97^2 = 0{,}94 \quad \text{(o bien 94 \%)}$$

que nos dice que, usando la distancia para estimar la cantidad de cerámica en varios yacimientos, reducimos la variación original en los datos (la variación alrededor del valor medio de las cantidades) aproximadamente en un 95 %, es

decir, que un 95 % aproximadamente de la variación en la cantidad de cerámica está relacionada con la distancia; sólo se deja un 5 % como variación alrededor de la recta de regresión. El que la distancia «explique» la variación en la cantidad de cerámica es otra cuestión, si bien toda explicación debiera tener en cuenta esa marcada relación.

Decir que el 95 % de la variación en la cantidad de cerámica procedente de New Forest en distintos yacimientos está relacionado con la distancia entre ese yacimiento y los hornos en los que se fabricó esa cerámica pudiera considerarse como el empleo de una cifra innecesaria para poner de manifiesto algo que ya era obvio en el diagrama de dispersión. ¡Hasta cierto punto, ese argumento está justificado! Podemos criticarlo, no obstante, en dos contextos específicos. En primer lugar, si estamos comparando, por ejemplo, la relación entre la cantidad de una mercancía en un yacimiento y la distancia al centro de producción de esa mercancía desde el yacimiento en cuestión, el uso de cifras es mucho más satisfactorio que las impresiones visuales a partir de diversos diagramas. En segundo lugar, una vez que emprendemos el estudio de las relaciones entre una gran cantidad de variables (más de dos), volvemos a precisar de las cifras, tanto para hacer las comparaciones como para manipularlas, tal y como veremos más adelante.

Acabaremos esta sección señalando que r^2 generalmente proporciona una afirmación más real que r de la intensidad de una relación, si es que consideramos lo que esos números significan para propósitos interpretativos. Así pues, un valor $r = 0{,}4$ sugiere, como mínimo, una relación moderada entre dos variables. Si lo elevamos al cuadrado veremos que significa que sólo el 0,16 (16 %) de la variación en una variable está relacionado con la otra.

CONCLUSIONES

Hemos visto los fundamentos de la investigación de una relación entre dos variables cuando éstas han sido medidas en una escala interválica o proporcional. El aspecto más importante de todo el procedimiento es la producción y el examen del diagrama de dispersión, si bien podemos profundizar aún más en dos aspectos: *a*) es posible obtener la ecuación de la recta de regresión que mejor se ajuste a la nube de puntos, especificando así la forma en que la variable dependiente cambia en relación con los cambios que suceden en la independiente; *b*) podemos obtener también una medida de la bondad del ajuste de los datos a la recta de regresión por medio del coeficiente de correlación y del coeficiente de determinación.

Existen muchos usos arqueológicos de estas técnicas, ya que los arqueólogos necesitan investigar las relaciones entre pares de variables numéricas. Estos métodos pueden usarse como un fin en sí mismos (véanse los ejemplos en este

capítulo) o como fundamento de las técnicas más avanzadas y complejas, algunas de las cuales se describen en los capítulos 11 y 13.

Antes de desarrollar, en el capítulo siguiente, algunos de los aspectos más complicados de la correlación y de la regresión sobre variables interválicas, no estará de más que nos refiramos brevemente a la *correlación de orden*, para que puedan evaluarse sus posibilidades.

En el capítulo 7 vimos diversas formas de examinar las relaciones entre variables nominales; en este capítulo hemos examinado relaciones entre variables interválicas; es evidente que han de existir también métodos apropiados para escalas ordinales. Tal y como podía esperarse, esos métodos son más potentes que los empleados para las variables nominales, pero no tanto como los empleados para las interválicas. El más conocido de esos métodos es probablemente el coeficiente de rangos de Spearman, si bien la tau b y la tau c de Kendall son más apropiadas en el caso en que haya muchas asociaciones, es decir, si muchas de las observaciones tienen el mismo rango. Una explicación detallada de estas técnicas puede verse en Blalock (1972) y Norusis (1983).

Un ejemplo del uso de los coeficientes de rango aparece en Shennan (1985). Como parte del estudio de una prospección se investigó la habilidad de varias personas para encontrar diferentes tipos de material arqueológico sobre el terreno. Por medio de unos métodos bastante complejos fue posible situar a cada individuo en una escala de orden según su capacidad de recoger cerámica y utensilios líticos. El rango de la habilidad para recoger cerámica se comparó con el rango de la habilidad para recoger artefactos líticos, para ver si, en general, la buena o mala capacidad para una estaba relacionada con la buena o mala capacidad para la otra. En realidad no lo estaban: alguien que sabe reconocer la cerámica sobre el terreno, no tiene por qué saber reconocer utensilios líticos.

En este ejemplo no había rangos iguales o asociados, es decir, no había equivalencias entre unos individuos y, digamos, el tercer puesto en la escala. En otros casos sí que las hay. Supongamos que hemos dividido los yacimientos de un área en particular aproximadamente por tamaños: grande, medio y pequeño; al igual que en el ejemplo de la prueba de Mann-Whitney y la de las series (*runs test*), esto es algo que siempre podemos hacer aunque no hayamos estimado el tamaño exacto de cada yacimiento. Supongamos también que disponemos de una categorización de los suelos de esa región: excelentes, medios o pobres en lo que se refiere a posibilidades de cultivo, y que sabemos cuáles son los yacimientos que se sitúan en cada una de las categorías de suelo. Cada uno de los yacimientos en una categoría de orden en particular puede decirse que está asociado a esa categoría. ¿En qué medida está relacionado el tamaño de los yacimientos con la calidad del suelo?

Podemos construir un cuadro con nuestras observaciones (tabla 9.3). La tau de Kendall (no confundir con la tau de Goodman-Kruskal, mencionada en el capítulo 7) puede calcularse para descubrir si hay correlación entre la categoría del tamaño del yacimiento y el potencial agrícola del suelo en el que está situado.

TABLA 9.3. Categoría del tamaño de los yacimientos tabulada por la categoría de calidad del suelo.

	Calidad del suelo			
	Excelente	Media	Pobre	Total
Grandes	15	7	2	24
Medios	6	11	4	21
Pequeños	7	7	8	22
Total	28	25	14	67

EJERCICIOS

9.1. Como parte de la investigación de la tecnología lítica del paleolítico y su complejidad, se lleva a cabo un estudio de los factores que afectan al número de lascas extraídas de unos bifaces. Se ha sugerido que es simplemente el resultado del tamaño global del bifaz, el cual se mide en términos de peso. Dada la siguiente información, ¿cuál es la relación entre el peso y la cantidad de lascas? ¿Es una relación muy intensa?

Lascas	Peso (g)	Lascas	Peso (g)
18	210	37	620
19	300	72	510
33	195	57	565
28	285	53	650
24	410	46	740
36	375	78	690
45	295	68	710
56	415	63	840
47	500	82	900

9.2. Se ha llevado a cabo una prospección arqueológica en el sur de Inglaterra. No se ha pretendido buscar yacimientos, sino recoger información acerca de la densidad de hallazgos por hectárea. A continuación se indican los datos de la densidad de hallazgos de cerámica en yacimientos de la edad del hierro y britanorromanos. Estudia las relaciones entre ellos.

Edad del hierro	Romana	Edad del hierro	Romana
4	5	12	55
3	20	9	61
7	20	7	62
6	33	13	79
6	46	9	81
9	45	14	98

10. CUANDO LOS DATOS NO SE AJUSTAN A LA REGRESIÓN

En el capítulo anterior vimos los fundamentos de los análisis de regresión y de correlación, evitando en todo momento las posibles complicaciones. Se señaló, sin embargo, que sólo se estudiaba la recta de regresión, y que el examen del diagrama de dispersión era imprescindible para comprobar si la nube de puntos mostraba una tendencia lineal o no lineal. Este punto nos conduce a la cuestión general de los requisitos para llevar a cabo un análisis de regresión válido; obviamente un aspecto muy importante, que aún no habíamos abordado.

Los problemas acerca de las relaciones entre los datos y los requisitos exigidos por el análisis de regresión pueden verse claramente en los *residuales* de la regresión: allí donde la regresión no se ajuste, la diferencia entre los valores y actuales y los predichos por la regresión. Pero los residuales son también importantes desde otro punto de vista. De hecho, pueden ser mucho más interesantes arqueológicamente que la regresión en sí misma, si bien esto requiere un interés previo por el análisis de regresión.

Como ejemplo supongamos que estudiamos el decrecimiento de la cantidad de un material a medida que aumenta la distancia a su fuente. Hay que señalar que, para una distancia dada, la mayoría de los yacimientos tienen sólo una pequeña cantidad de material, y unos pocos, una cantidad mucho mayor. La pregunta es, obviamente, por qué sucede esto. Podemos llegar a descubrir cuáles son los yacimientos y estudiar las características que comparten entre sí algunos de ellos y que son inexistentes en los demás, lo que permitiría explicar el fenómeno. Por ejemplo, puede que los yacimientos en los que aparece una mayor cantidad de material estén cerca de una vía fluvial. Hodder y Orton (1976, pp. 115-117), en un análisis de la distribución de la cerámica britanorromana procedente de los centros de producción de Oxfordshire, fueron capaces de mostrar que los yacimientos con una mayor cantidad de esa cerámica de lo esperado dada su distancia al centro de producción eran aquellos situados más cerca de las rutas fluviales. Shennan (1985) usó métodos de regresión para seleccionar distribuciones de piezas de sílex con cantidades muy grandes o muy pequeñas de piezas retocadas. Trazando su distribución en un mapa, parecía que eran características de ciertos tipos de localización.

45

Tanto por razones sustantivas como metodológicas, es necesario estudiar los residuales de una regresión, aunque ello pueda obligarnos a aumentar el nivel de dificultad de los capítulos anteriores.

RESIDUALES

De la misma forma en que se usa $\sum(y_i - \bar{y})^2$ para el cálculo de la varianza o de la desviación típica de una variable simple, podemos usar $\sum(y_i - \hat{y})^2$ para calcular la varianza o la desviación típica alrededor de la recta de regresión:[1]

$$s^2_{y-\hat{y}} = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

donde $s^2_{y-\hat{y}}$ es la varianza de la distribución alrededor de la recta de regresión, y_i es el valor y actual en el punto i-avo, $\hat{y}_i$ es la estimación de y para el punto i-avo de acuerdo con la regresión y n es el número de observaciones.

La raíz cuadrada de esa expresión constituye la desviación típica de la distribución, llamada también *error típico o estándar.*

TABLA 10.1. Información para calcular el error típico de la regresión para las cantidades de cerámica britanorromana en la tabla 9.1.

y_i	$\hat{y}_i$ *	$(y_i - \hat{y}_i)^2$
98	95,44	6,55
60	66,64	44,09
41	45,04	16,32
47	41,44	30,91
62	59,44	6,55
		104,42

* Calculado a partir de $y = 102{,}64 - 1{,}8x$.

Para el ejemplo de la cantidad de cerámica, disponemos de la información que aparece en la tabla 10.1. Aplicando la fórmula anterior, obtenemos:

$$s^2_{y-\hat{y}} = \frac{104{,}42}{5} = 20{,}88$$

$$s_{y-\hat{y}} = \sqrt{20{,}88} = 4{,}57$$

1. El denominador de esta versión de la fórmula es n, lo que presupone que nos interesa tan sólo la variación alrededor de la regresión para el conjunto de datos en particular analizado. Si queremos estimar la variación alrededor de la recta, para una población de la cual es una muestra, el divisor será $n - 2$, ya que se pierden dos grados de libertad en el cálculo de la regresión. MINITAB usa este divisor en su procedimiento de regresión.

El error típico de la regresión de la cantidad de cerámica y la distancia es 4,57.

Como veremos a continuación, una de las estipulaciones del modelo de regresión es que la distribución de los residuales alrededor de la recta sea normal. Una vez que ese es el caso, podemos apreciar que si disponemos el error típico como líneas paralelas a la recta de regresión, incluiremos, aproximadamente, el 68 % de todas las observaciones; si el espacio entre esas líneas fuese de ± 2 errores típicos, incluiríamos el 95 % de todas las observaciones. Una ilustración de este hecho puede verse en la figura 10.1.

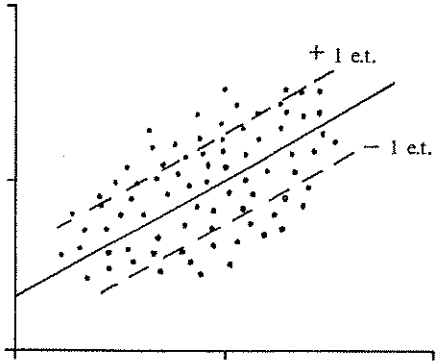


FIGURA 10.1. Región situada a un error típico de la recta de regresión.

De hecho, al igual que con el uso de r o r^2 como indicadores del ajuste de la regresión, podemos usar el error típico de la regresión como un indicador de la precisión de las estimaciones, de la misma forma que usábamos la desviación típica para la dispersión de una variable normal. Podemos añadir un término extra a la ecuación de regresión para reconocer este punto:

$$\hat{y}_i = a + bx_i \pm s_{y-\hat{y}}$$

donde $\hat{y}_i$ es el valor estimado de y_i; a y b el punto de corte y la pendiente, respectivamente; x_i es el valor x del punto relevante; y $s_{y-\hat{y}}$ es el error típico de la regresión. En el ejemplo de la cantidad de cerámica, los valores actuales en la fórmula eran:

$$\hat{y}_i = 102{,}64 - 1{,}8x_i \pm 4{,}57$$

Cuando la distribución de los residuales es normal, podemos decir que alrededor de un 68 % de los residuales entrarán en ese rango. Asumiendo la normalidad del ejemplo anterior, aproximadamente un 68 % de las observaciones estarán entre ± 4,57; o, alternativamente, una estimación de que cualquier valor

46

está en el intervalo ± 4,57 alrededor de la recta de regresión tendrá una posibilidad del 68 % de ser correcta. Una estimación como esta se conoce como *intervalo de estimación*, ya que especificamos el intervalo dentro del cual ha de encontrarse una cantidad determinada, con un grado de probabilidad específico; volveremos a tratar este tema más adelante, cuando abordemos el tema del muestreo (capítulo 14). En cuanto a lo que ahora nos concierne, 3/5 de los datos (60 %) aparecen dentro de un error típico de la recta de regresión, y todos ellos dentro del intervalo definido por dos errores típicos; dado el escaso número de observaciones, la correspondencia con los valores esperados es lo más estrecha que podía ser.

Hay otra propiedad muy útil que también se deriva de la distribución normal de los residuales. Vimos en el capítulo 8 que cualquier observación en una distribución podía transformarse en una puntuación Z que expresa la observación en términos de unidades de desviación típica distantes de la media, donde:

$$Z = \frac{x - \bar{x}}{s}$$

En una distribución normal, la puntuación puede ser trasladada a la tabla para encontrar la proporción de la distribución que se encuentra entre la media y un punto situado a esa distancia de ella. De la misma forma, si tomamos cualquier término residual dado $(y_i - \hat{y}_i)$ de la regresión y lo dividimos entre la desviación típica (error) de la distribución de los residuales alrededor de la regresión, producimos una cantidad análoga a Z, llamada residual estandarizado:

$$\text{Residual estandarizado} = \frac{y_i - \hat{y}_i}{s_{y-\hat{y}}}$$

Tiene las mismas propiedades que Z y puede ser puesto en relación con la distribución normal del mismo modo, usando la tabla normal (asumiendo, claro está, que los residuales estén normalizados).

Para la segunda observación en el ejemplo de la cantidad de cerámica y la distancia, tenemos:

$$\text{Residual estandarizado} = \frac{60 - 66{,}64}{4{,}57} = -1{,}45$$

Este valor y es 1,45 errores típicos menor que el valor estimado por la regresión.

Tal y como veremos en la sección siguiente, el residual estandarizado, por su vinculación con la distribución normal, tiene propiedades que lo hacen muy útil para investigar si los presupuestos de la regresión se han cumplido o no,

e incluso para recuperar esquemas de interés en los resultados de la regresión.[2]

El modelo de regresión

Todo lo que hemos hecho hasta aquí es válido, única y exclusivamente si ciertos presupuestos acerca de los residuales del modelo han sido respetados. Esos presupuestos son de varios tipos, si bien cualquier fallo en su cumplimiento, sea del tipo que sea, se refleja siempre en los residuales. Por ese motivo se deben usar medios gráficos (diagrama de dispersión), no sólo para ver los datos originales, sino para observar la estructura de los residuales. Los análisis basados simplemente en un examen del resumen estadístico de una distribución no son suficientes. La regresión por mínimos cuadrados es bastante robusta, con respecto a violaciones menores de los presupuestos, si bien grandes violaciones de los mismos pueden engendrar conclusiones seriamente distorsionadas. Lo que pretendo hacer a continuación es examinar los presupuestos e indicar cómo detectar una violación de los mismos y qué se puede hacer al respecto.

Presupuestos

1. En la versión de la regresión que estamos comentando se presupone que la variable independiente y la dependiente están medidas en una escala interválica o superior.
2. Ya se ha señalado que tratamos tan sólo con la regresión lineal, en la que la relación entre dos variables adopta la forma de una línea recta. Obviamente, si la tendencia no es lineal, entonces un análisis que asuma esto no será muy satisfactorio; se muestra un ejemplo en la figura 10.2.

En este ejemplo, y aumenta a medida que aumenta x, pero en distintas proporciones en diferentes partes de la escala x. El simple cálculo de la regresión lineal y del coeficiente de correlación asociado sugeriría, de hecho, que hay una fuerte tendencia lineal. Es el examen del gráfico el que muestra que la línea recta representa una descripción insatisfactoria de la relación, ya que subestima al principio y al final de la recta y sobrestima en el centro.

3. La distribución de los residuales alrededor de la recta de regresión ha de ser normal. Esto es particularmente importante si queremos usar la regresión para obtener estimaciones interválicas para y de la manera propuesta en la sección anterior, o bien si queremos llevar a cabo pruebas de significación; véase la figura 10.3, para ejemplos de distribución de residuales normalizados y no normalizados.

2. Recientemente, la tendencia ha sido usar no los residuales *estandarizados*, sino los *studentizados* (distribución de Student). En este caso, el valor de cada residual se tipifica no por medio del error típico de la regresión como un todo, sino por medio del error típico calculado sin incluir el valor de los puntos en concreto; en efecto, esos valores están estandarizados individualmente.

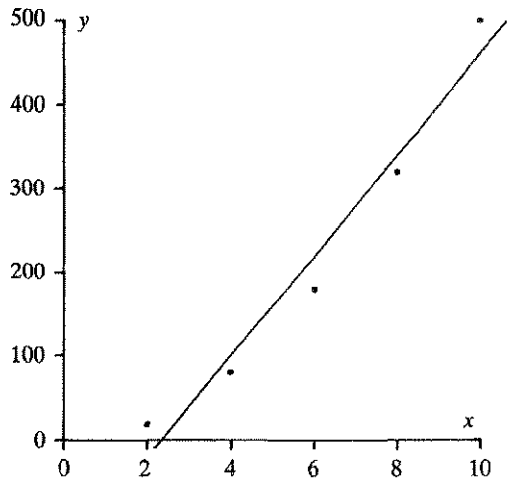


FIGURA 10.2. Diagrama de dispersión de una relación no lineal entre x e y.

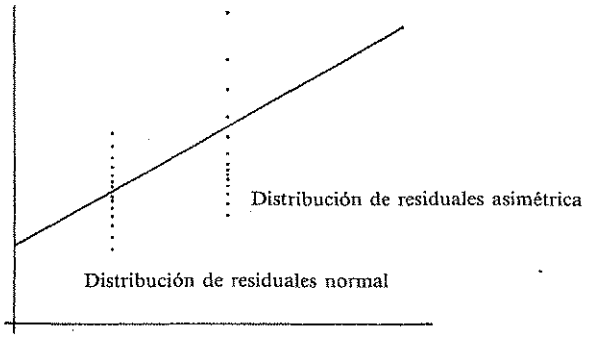


FIGURA 10.3. Distribución de residuales alrededor de la recta de regresión.

4. La media de la distribución de los residuales ha de ser cero para cada valor de x; en otras palabras, la distribución de los residuales ha de estar centrada sobre la recta de regresión. Si no lo están, por lo general es porque se ha producido una violación del presupuesto de linealidad (cf. más arriba), o la presencia de autocorrelación (cf. las páginas siguientes).
5. Uno de los presupuestos más importantes del análisis de regresión es que la variación alrededor de la recta sea *homocedástica*. En otras palabras, que la cantidad de variación alrededor de la recta sea la misma en todos los puntos a lo largo de ella. Si esto no sucede así, la variación es *heterocedástica*. Hay varias maneras en que la heterocedasticidad aparece. Dos de las más corrientes

están ilustradas en la figura 10.4. En 10.4(a) las observaciones con valores pequeños en x e y tienden a estar más cerca de la recta que aquellas con valores mayores, las cuales están más dispersas. En (b) sólo hay una pequeña cantidad de los casos con valores en x e y grandes; la mayoría de las observaciones cuenta con valores pequeños.

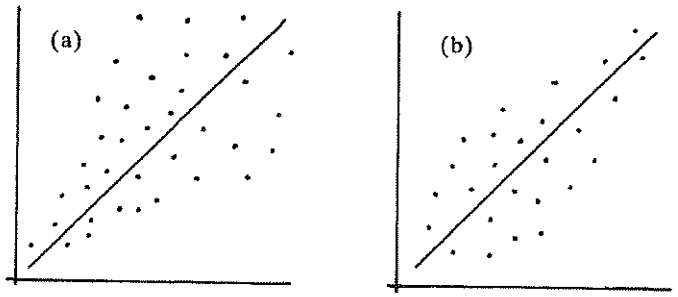


FIGURA 10.4. Distribución de residuales heterocedástica.

6. Autocorrelación. Uno de los principales presupuestos del análisis de regresión es que los términos de error asociados con observaciones particulares no están correlacionados. En otras palabras, que el valor residual en y para un valor en x no ha de estar relacionado con el valor residual de otros valores x. Un ejemplo de distribución autocorrelacionada aparece en la figura 10.5.

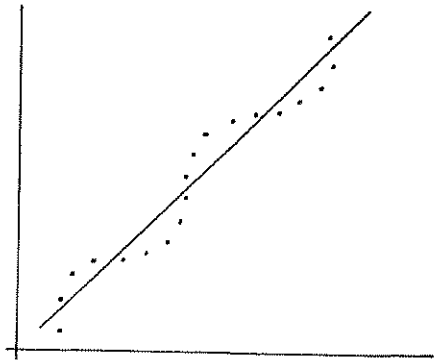


FIGURA 10.5. Ejemplo de autocorrelación de los residuales de una regresión.

En esta figura, los residuales positivos están agrupados y aparecen seguidos por los residuales negativos, también agrupados; se trata de un esquema que se repite a sí mismo a lo largo de la recta y que resulta en una relación no lineal. Junto con los otros presupuestos, cualquier fallo en prevenir la autocorrelación produce resultados engañosos. En el caso ejemplificado anteriormente, el valor del coeficiente de correlación para una relación lineal sería muy alto, lo que implicaría que para cualquier incremento en x habría un incremento equiparable en y; de hecho, dependiendo de la ubicación en el eje x, los aumentos de y variarán considerablemente.

La autocorrelación puede aparecer por muchos motivos, que serán discutidos más adelante.

El modelo de regresión es bastante robusto, si nos referimos tan sólo a las violaciones menores de los presupuestos. Los más importantes son los de linealidad (que subsumen algunos otros), la homocedasticidad y los errores no correlacionados.

Detección y corrección de las violaciones en los presupuestos de la regresión

Como ya se ha sugerido, una de las formas mejores y más simples de detectar discrepancias entre el modelo y los datos es a través del examen de los residuales de la regresión. Ya hemos definido el valor residual de y como la diferencia entre los valores actuales y los estimados de y:

$$\text{res. } y_i = y_i - \hat{y}_i$$

Hemos definido también el residual estandarizado para cada y:

$$\text{res. est. } y = \frac{y_i - \hat{y}_i}{s_{y-\hat{y}}}$$

donde

$$s_{y-\hat{y}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

Tal y como se dijo antes, los residuales estandarizados son como puntuaciones Z, ya que tienen una media igual a cero y una desviación típica igual a uno. En una muestra moderadamente grande, estos residuales han de estar distribuidos aproximadamente con arreglo a la normalidad. La representación gráfica de los residuales revelará si esto es así o no. Si no lo es, es porque hay problemas de algún tipo.

Los gráficos más usados son aquellos en los que los residuales estandarizados están representados como la ordenada (o eje y), en relación (a) a la variable independiente x o (b) al valor estimado de y, por ejemplo $\hat{y}$. Ejemplos de ambos casos aparecen en la figura 10.6. Realmente no se puede decir cuál de ellos es mejor o más adecuado, si bien el primero (residuales por variable inde-

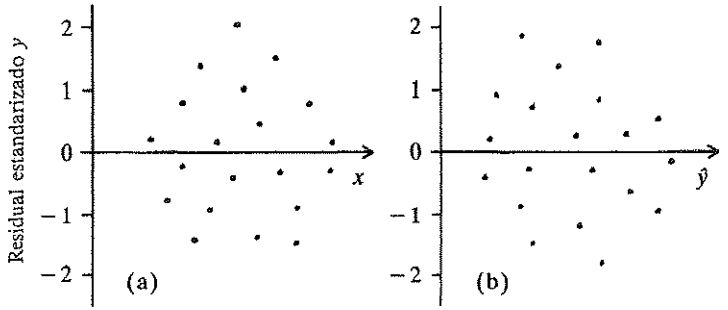


FIGURA 10.6. Ejemplos de los valores residuales estandarizados (eje vertical) con relación a sus: (a) valores x; (b) valores $\hat{y}$ predichos por la regresión.

pendiente) probablemente tenga una interpretación más sencilla. Tal y como veremos en el próximo capítulo, sin embargo, en la regresión múltiple, la única opción es representar los residuales frente al valor estimado de $\hat{y}$.

Si el modelo es correcto, los residuales estandarizados tienden a encontrarse entre los valores $+2$ y -2, distribuidos aleatoriamente (véase fig. 10.6); no muestran ningún esquema de variabilidad distintivo. Si no se mantienen los presupuestos, aparecen ciertos esquemas de regularidad, lo cual nos dice no sólo que hemos de hacer algo para que los datos se ajusten a los requisitos, si es que queremos seguir usando la técnica, sino que también nos proporciona información, revelando estructuras insospechadas entre los datos. Tal y como ya hemos visto, los residuales pueden ser más interesantes que la regresión misma, ya que ésta sólo sistematiza lo que ya sabíamos previamente. Es precisamente cuando esa sistematización revela que la estructuración de los datos es más compleja de lo que creíamos, que podemos ganar en conocimiento. En esos casos, la regresión matemáticamente definida proporciona una base segura para la comparación y detección de irregularidades. Un buen ejemplo de lo que puede llegar a descubrirse ya ha sido citado: la demostración llevada a cabo por Hodder y Orton por medio de un análisis de regresión, según la cual, en el análisis de regresión de la cantidad de cierto tipo de cerámica britanorromana por la distancia a su centro de producción, se obtuvieron residuales positivos muy altos en las áreas accesibles por transporte fluvial (Hodder y Orton, 1976). En la regresión bivariada siempre es posible saber si hay problemas mirando simplemente el diagrama de dispersión de los datos originales; sin embargo, los gráficos de residuales son mucho más efectivos —sirven de lupa para buscar errores—. Cuando lleguemos a la regresión múltiple, en el capítulo siguiente, y tratemos con más de dos variables, la representación de los residuales será la única opción disponible.

Podemos volver ahora a la detección y corrección del fallo para mantener los presupuestos del modelo de regresión lineal.

1. *No linealidad.* Ya ha sido señalado que la no linealidad entre dos variables puede aparecer de muy distintas maneras. Se ha mostrado un ejemplo en la figura 10.2, y aunque la falta de linealidad emerge con bastante claridad del diagrama de dispersión, la figura 10.7 muestra cómo el problema se acentúa en el gráfico de residuales correspondiente, que ciertamente no muestra una nube de puntos distribuida aleatoriamente.

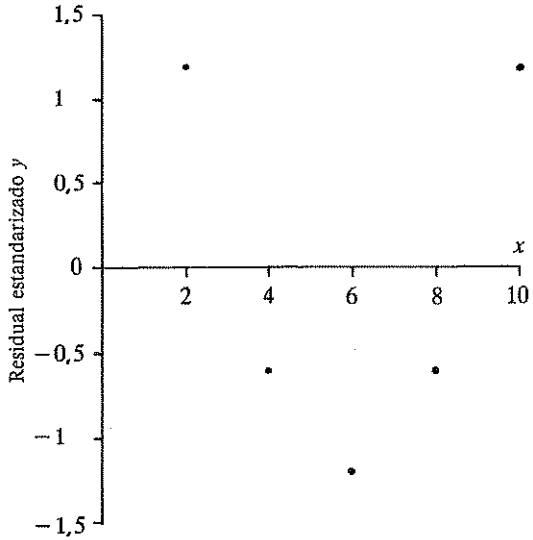


FIGURA 10.7. Gráfico de los residuales estandarizados de la recta de regresión de la figura 10.2, con relación a sus valores x.

En algunos casos, la falta de relación lineal entre x e y puede arreglarse recurriendo a transformaciones. Podemos haber detectado esta no linealidad contemplando los datos, o bien podemos tener razones teóricas para postular una forma particular de curva no lineal, pero «linealizable», y queremos ver si nuestros datos se ajustan a ella; ese ajuste es generalmente más sencillo si la relación aparece en forma lineal. En aplicaciones arqueológicas, el contexto más habitual en el que tendremos razones teóricas para postular una forma particular de relación curvilínea es en los estudios de decrecimiento de la distancia, como en el analizado previamente, en el que se investiga la relación entre el valor cambiante de una variable y la distancia a un punto. Se han llevado a cabo muchos estudios acerca de la forma más probable de esas curvas (véanse Hodder y Orton, 1976; Renfrew, 1977).[3]

3. En casos en los que no tenemos una base teórica para postular curvas en particular, pero tenemos una relación no lineal, es posible permitir que sean los datos los que determinen la selección de la transformación más apropiada. Las técnicas requeridas están mucho más allá del alcance de un libro como este; véase, por ejemplo, McDonald y Snooks (1985).

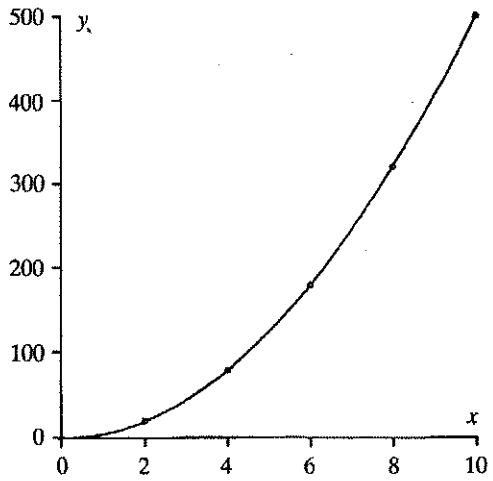


FIGURA 10.8. Gráfico de la relación de Pareto $y = 5x^2$.

Una forma bastante común de relación curvilínea es el doble logaritmo o relación de Pareto, en donde la ecuación de la línea de regresión (fig. 10.8) adopta la forma

$$y = ax^b$$

Para convertirla en lineal, la transformación apropiada es:

$$\log y = \log a + b \log x$$

que significa que hemos calculado los logaritmos de los valores tanto de y como de x y los hemos usado, primero como ejes de un nuevo diagrama de dispersión, y segundo para calcular una ecuación de regresión y un coeficiente de correlación. De hecho, el diagrama de dispersión de la relación no lineal en la figura 10.2 tiene esta forma. Lo que sucede al calcular los logaritmos de los ejes x e y se observa en la figura 10.9.

La otra forma más común de relación no lineal es la curva exponencial (fig. 10.10), con la ecuación

$$y = ab^x$$

y su versión lineal, dada por la fórmula:

$$\log y = \log a + \log bx$$

Estamos obligados a calcular los logaritmos de los valores y y usarlos para producir un nuevo diagrama de dispersión con un nuevo eje vertical en unidades

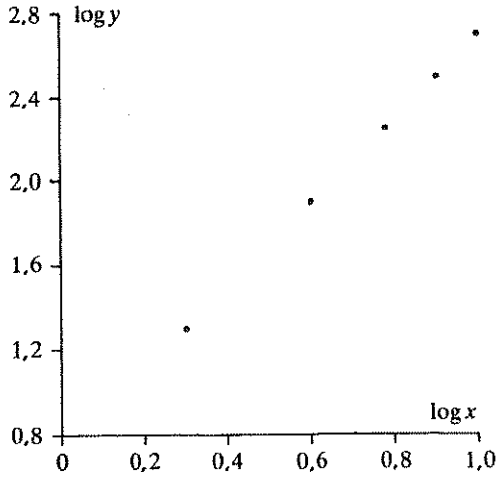


FIGURA 10.9. Datos de la figura 10.2 con sus valores x e y sustituidos por los logaritmos de sus valores originales.

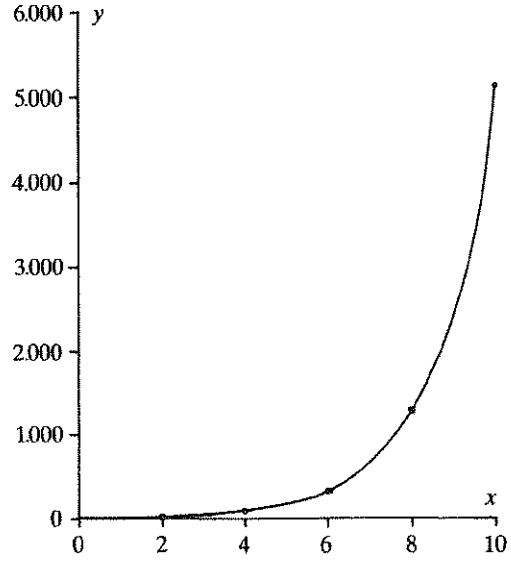


FIGURA 10.10. Gráfico de la relación exponencial $y = 5(2^x)$.

de log y. Precisamente porque el eje y está logaritmizado, se obtienen los exponentes logarítmicos de a y b, pues ambos están expresados en unidades de y: a es el punto en que la recta de regresión corta el eje y, b es el incremento de cambio en y para cada cambio en x. En este caso, x permanece igual.

50

Los casos que hemos visto son algunas de las situaciones no lineales, pero linealizables, más corrientes que pueden surgir en contextos arqueológicos y que son frecuentes en los estudios espaciales de decrecimiento, si bien es importante recordar que son precisamente las versiones negativas de esas curvas las que tienen relevancia, y no las versiones positivas que se han ido citando. Si existe no linealidad, aparecerá en la representación gráfica de los datos y en los residuales estandarizados. Si la representación se corresponde con alguno de los gráficos que hemos visto, entonces será posible llevar a cabo la transformación debida de los datos e intentar de nuevo una regresión lineal, teniendo presente su comprobación antes de representar los residuales.

La discusión precedente sobre funciones no lineales y maneras de convertirlas en lineales ha sido muy abstracta, por lo que el lector necesitará, probablemente, que presentemos un ejemplo concreto; una vez más, consideremos el caso de un estudio hipotético de decrecimiento de cantidades comparadas a la distancia, esta vez la cantidad de un cierto tipo de obsidiana (medida en g/m^3) encontrada en yacimientos mesoamericanos situados a diversas distancias de la fuente. Los datos aparecen en la tabla 10.2.

TABLA 10.2. Densidad de hallazgos de obsidiana de cierto tipo en yacimientos mesoamericanos situados a diferentes distancias de la fuente.

Distancia (km)	Densidad (g/m^3)	Distancia (km)	Densidad (g/m^3)
5	5,01	44	0,447
12	1,91	49	0,347
17	1,91	56	0,239
25	2,24	63	0,186
31	1,20	75	0,126
36	1,10		

La primera fase del análisis consistirá en el cálculo de la regresión y de la correlación:

$$n = 11 \qquad \sum x_i y_i = 284{,}463$$
$$\sum x_i = 413 \qquad \sum y_i = 14{,}715$$
$$\left(\sum x_i\right)^2 = 170.569 \qquad \left(\sum y_i\right)^2 = 216{,}531$$
$$\sum x_i^2 = 20.407 \qquad \sum y_i^2 = 40{,}492$$

$$b = \frac{(11 \times 284{,}463) - (413 \times 14{,}715)}{(11 \times 20.407) - 170.569} = \frac{-2.948{,}202}{53.908} = -0{,}055$$

$$a = \frac{14{,}715 - (-0{,}055 \times 413)}{11} = 3{,}403$$

La ecuación de regresión resulta:

$$\hat{y} = 3{,}403 - 0{,}055x$$

$$r = \frac{(11 \times 284{,}463) - (413 \times 14{,}715)}{\sqrt{[(11 \times 20.407) - 170.569][(11 \times 40{,}492) - 216{,}531]}} =$$

$$= \frac{-2.948{,}202}{\sqrt{53.908 \times 228{,}881}} = -0{,}839$$

$$r^2 = -0{,}839^2 = 0{,}704$$

El coeficiente de correlación tiene el valor de $-0{,}839$ y el valor r^2 de 0,704 indica que algo más del 70 % de la variación en la cantidad de obsidiana está relacionada con la distancia a la fuente. Estos valores indican una fuerte relación lineal entre las dos variables. Ahora bien, prescindiendo de las cifras anteriores, si nos fijamos en el diagrama de dispersión [fig. 10.11(a)], se aprecia que la distribución de puntos no es una línea recta, es decir, que la regresión lineal subestima al principio y al final de la línea y sobrestima en el centro. La representación de los residuales estandarizados [fig. 10.11(b)] pone de manifiesto este hecho con más claridad.

Ya que la regresión no se ajusta, es preciso que hagamos algo al respecto. El examen de los diagramas de dispersión de los datos originales sugiere que una regresión lineal se ajustaría bastante bien a todos los datos excepto al primero. Este primer punto puede ser considerado como *outlier* y excluido del análisis, volviéndose a calcular todo de nuevo sin él. Puede ser correcto rechazar observaciones de esta forma, si bien el procedimiento tiene, obviamente, muchos peligros; se podrían llegar a desechar todos los puntos que no se ajustaran al modelo, pero para eso se necesitarían muy buenas razones; por ejemplo, entre los motivos por los que esa observación en particular no es válida podrían citarse: una mala excavación del yacimiento, o lo reducido del sector excavado.

En este caso supondremos que no hay razones para rechazar el primero de los puntos, por lo que es necesario encontrar un modelo que se ajuste a todos los datos: obviamente, algún tipo de relación curvilínea. El examen del diagrama de dispersión original y el conocimiento de casos similares (véanse Hodder y Orton, 1976; Renfrew, 1977) sugiere que una curva exponencial puede ser la apropiada. Tal y como ya hemos visto, es mucho más sencillo ajustar la regresión a una forma lineal, antes que dejarla en su forma curvilínea original. Por lo tanto, hemos de hacer una transformación. Las referencias antes citadas in-

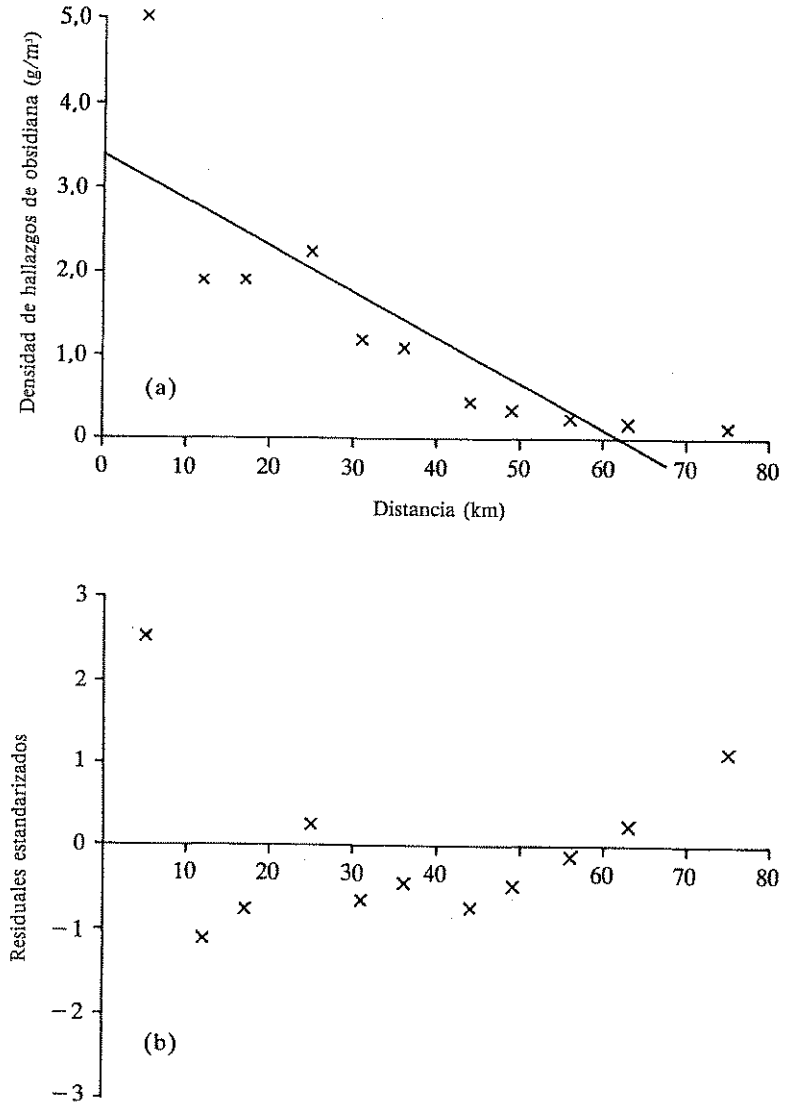


FIGURA 10.11. (a) Gráfico de la densidad de hallazgos de obsidiana con relación a la distancia a la fuente; la recta de regresión $\hat{y} = 3{,}403 - 0{,}055x$ está superpuesta. (b) Gráfico de los residuales estandarizados de las densidades de hallazgos de obsidiana, con relación a la distancia de su fuente.

dican que una curva exponencial puede ser linealizable tomando el logaritmo del eje y; en otras palabras, transformando en logaritmos los valores originales (véase tabla 10.3).

TABLA 10.3. Densidad y logaritmos de la densidad de los hallazgos de obsidiana de cierto tipo en yacimientos mesoamericanos situados a diferentes distancias de la fuente.

Densidad (y)	Logaritmo de la densidad (log y)	Densidad (y)	Logaritmo de la densidad (log y)
5,01	0,6998	0,447	−0,3497
1,91	0,2810	0,347	−0,4597
1,91	0,2810	0,239	−0,6216
2,24	0,3502	0,186	−0,7305
1,20	0,0792	0,126	−0,8996
1,10	0,0414		

Ahora podemos calcular la regresión usando los valores y transformados:

$$n = 11 \qquad \sum x_i y_i = -161{,}865$$
$$\sum x_i = 413 \qquad \sum y_i = -1{,}3285$$
$$\left(\sum x_i\right)^2 = 170.569 \qquad \left(\sum y_i\right)^2 = 1{,}7649$$
$$\sum x_i^2 = 20.407 \qquad \sum y_i^2 = 2{,}8412$$

$$b = \frac{[11 \times (-161{,}865)] - [413 \times (-1{,}3285)]}{(11 \times 20.407) - 170.569} = \frac{-1.231{,}6795}{53.908} = -0{,}0229$$

$$a = \frac{-1{,}3285 - (-0{,}0229 \times 413)}{11} = 0{,}739$$

La ecuación de regresión resulta:

$$\log \hat{y} = 0{,}739 - 0{,}0229x$$

$$r = \frac{[11 \times (-161{,}865)] - [413 \times (-1{,}3285)]}{\sqrt{[(11 \times 20.407) - 170.569][(11 \times 2{,}8412) - 1{,}7649]}} =$$

$$= \frac{-1.231{,}6795}{\sqrt{53.908 \times 29{,}488}} = -0{,}9769$$

$$r^2 = -0{,}9769^2 = 0{,}9543$$

Resulta obvio que una relación exponencial se ajusta a estos datos muchísimo mejor que una simple relación lineal. El valor r^2 indica que, si se postula que el decrecimiento es exponencial, más del 95 % de la variación, expresada en logaritmos de las unidades originales, se relaciona con la distancia de los yacimientos a la fuente de obsidiana; de hecho, ninguna otra función de decrecimiento se ajusta tan bien. Obviamente, debemos volver a estudiar los diagramas de dispersión para la relación transformada y el gráfico asociado de residuales (véanse figs. 10.12 y 10.13).

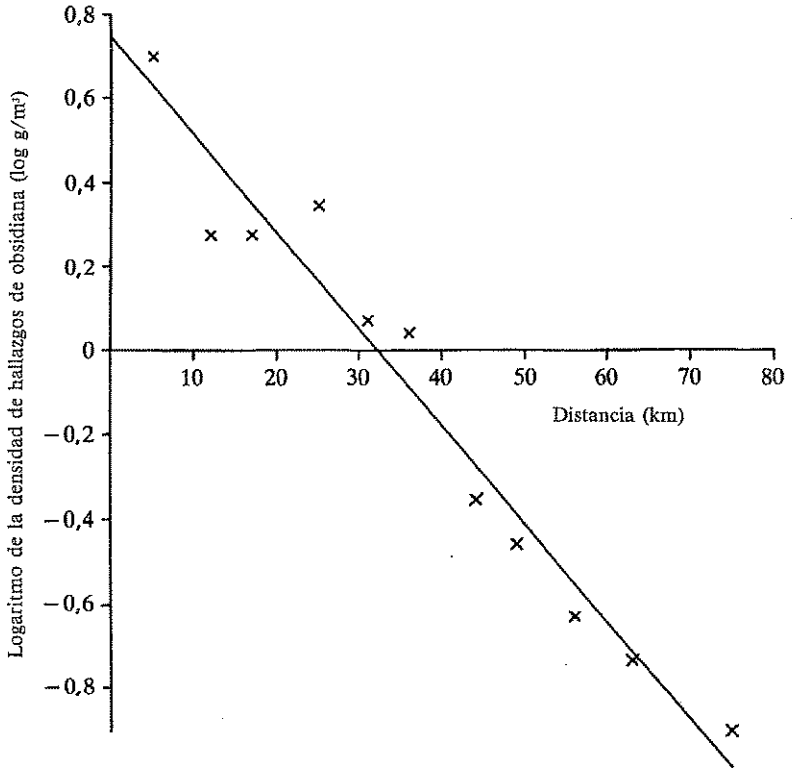


FIGURA 10.12. Gráfico de las densidades de hallazgos de obsidiana (logaritmizados), con relación a la distancia a su fuente; la recta de regresión $\log \hat{y} = 0{,}739 - 0{,}0229x$ está superpuesta.

Se deduce del examen de esas ilustraciones que el ajuste de la línea a sus datos es mucho mejor. La subestimación y sobrestimación anteriores han desaparecido, y la distribución de los residuales se aproxima más a una dispersión amorfa, que es lo esperado si se han respetado los supuestos del análisis de regresión. Restan, sin embargo, unos débiles indicios de heterocedasticidad y auto-

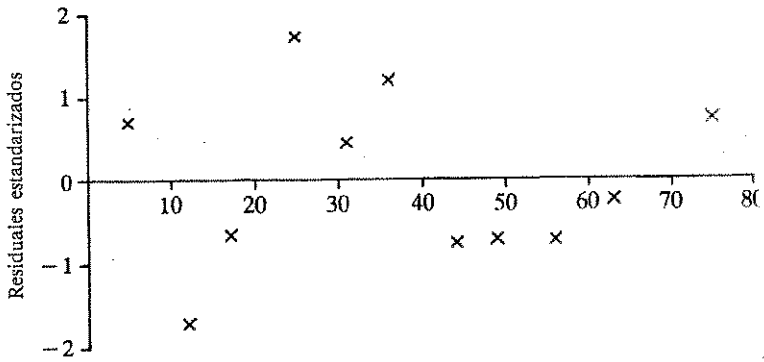


FIGURA 10.13. Residuales estandarizados logaritmizados de las densidades de hallazgos de obsidiana, con relación a la distancia a su fuente.

correlación en los residuales, lo cual merecería un estudio particular. Hay que señalar, además, que esos indicios estaban totalmente ocultos por la no linealidad.

El principal problema potencial con esta o cualquier otra transformación es su interpretación. La ecuación de regresión $\log \hat{y} = 0{,}739 - 0{,}0229x$ significa que hay un decrecimiento de 0,0229 en el logaritmo de y para cada unidad de incremento en x. Suele ser muy fácil olvidar que se ha realizado una transformación, y discutir los resultados como si se hubiesen obtenido a partir de los datos originales sin transformar. Calculando los antilogaritmos es posible volver a expresar la línea de regresión como una curva exponencial (véase fig. 10.14). Pero aunque esto pueda ser intuitivamente útil —nos permite trabajar de nuevo con los datos reales—, tiende a enfatizar el hecho de que las relaciones lineales son más fáciles de entender y que es más sencillo reconocer sus desviaciones.

2. *Heterocedasticidad.* Es preciso ahora que volvamos a tratar la cuestión de la heterocedasticidad y los métodos para estabilizar la varianza del error y volverla homocedástica. Al igual que con otras violaciones de los presupuestos del análisis de regresión, la heterocedasticidad se pone de relieve durante el examen de los diagramas de dispersión de los datos originales y de los gráficos de residuales. Puede aparecer de dos maneras distintas, las ilustradas en la figura 10.4. En 10.4(a) la dispersión de las observaciones alrededor de la recta aumenta a medida que aumenta el valor de la variable independiente. Es la situación que suele aparecer en los estudios de la relación entre el tamaño del asentamiento y la cantidad de población; en asentamientos con mucha población hay una mayor variación en su área de superficie que en aquellos con poca población (véase, por ejemplo, Carothers y McDonald, 1979). En 10.4(b) la

53

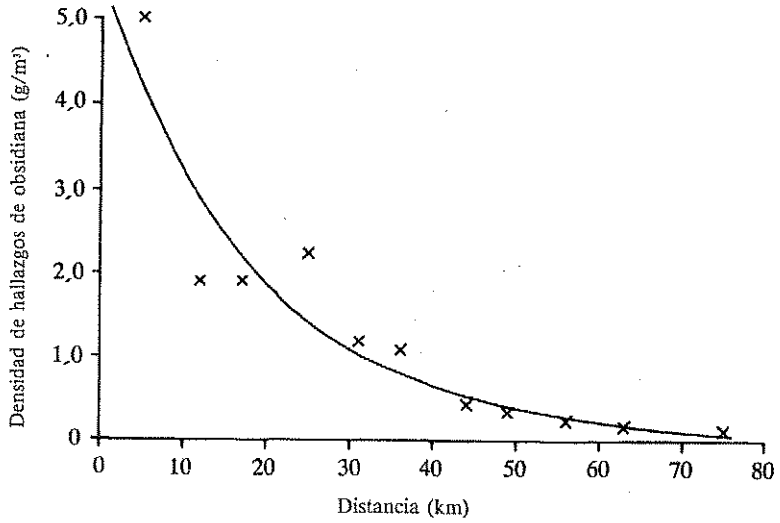


FIGURA 10.14. Gráfico de la densidad de hallazgos de obsidiana con relación a la distancia a su fuente; la versión antilogarítmica de la recta de regresión $\log \hat{y} = 0{,}739 - 0{,}0229x$ está superpuesta.

dispersión alrededor de la línea decrece para las observaciones de valor más alto, básicamente porque hay muy pocos datos con valores altos. Esto puede ocurrir, por ejemplo, en estudios acerca del tamaño de los asentamientos en el vértice de una jerarquía de asentamientos, donde el menor número de yacimientos se registra entre aquellos situados más arriba, que pueden tener tamaños mucho mayores que el resto.

En este segundo caso, podemos considerar la cuestión siguiente: ¿son «outliers» las observaciones mayores en la distribución? ¿Podemos prescindir de ellas en el análisis? En el caso de tener que incluirlas imprescindiblemente, logaritmizando una o ambas variables obtendremos el efecto de «estirar» los valores extremos para que se aproximen al resto y, como resultado, lo más probable es que las varianzas a lo largo de la línea se equilibren.

Más en general, si la dispersión a lo largo de la línea es proporcional a x podremos usar técnicas de regresión por mínimos cuadrados (*ponderadas*), alterando el peso o la influencia de ciertos puntos en los resultados de la regresión. Esta cuestión está fuera del alcance de este libro; esas técnicas alternativas de regresión aparecen en el programa MINITAB, y están descritas en el *Manual MINITAB del estudiante* (Ryan *et al.*, 1985).

3. *Autocorrelación.* La presencia de correlación entre los residuales de una regresión aparece muy claramente durante el estudio del diagrama de dispersión y en el gráfico de residuales, si bien puede efectuarse, también, una prueba estadística, la de Durbin-Watson (véase Chatterjee y Price, 1977). Esta prueba sólo funciona si los datos han sido recogidos en un orden numérico secuencial de las x, por ejemplo, en una secuencia temporal.

A menudo, la presencia de errores correlacionados sugiere que hay alguna otra variable que afecta a la dependiente, y, así como a la x en el modelo. Si ese es el caso, habrá que añadir al modelo las variables que nuestro conocimiento de la situación nos sugiera como relevantes para explicar la aparente autocorrelación. La regresión resultante será múltiple (véase el próximo capítulo).

Muy a menudo la autocorrelación está relacionada con la distribución en el tiempo y en el espacio: observaciones adyacentes en el tiempo o en el espacio tienden a mostrar idénticos residuales. En esas circunstancias, la inclusión de una nueva variable independiente que parezca sustantiva al problema y que varíe en relación con el tiempo o el espacio puede retirar el efecto de la autocorrelación.

Otras veces, sin embargo, la autocorrelación es intrínseca a las tendencias temporales o espaciales, por lo que habrá de intentarse una transformación apropiada. Consideremos un ejemplo. La investigación intensiva de un área del suroeste de los Estados Unidos ha proporcionado información acerca de la densidad de asentamientos por km² para una sucesión de fases cronológicas (tabla 10.4). La pregunta es cómo cambia la densidad de yacimientos a medida que transcurre el tiempo y cuál es la precisión del ajuste de los datos a la relación propuesta.

TABLA 10.4. Cantidad de asentamientos por kilómetro cuadrado para una sucesión de fases cronológicas en un área del suroeste de los Estados Unidos (datos según Plog, 1974).

Período (años)	Yacimientos por km²	Período (años)	Yacimientos por km²
0- 50	0,25	300-350	1,05
50-100	0,25	350-400	1,00
100-150	0,55	400-450	1,15
150-200	0,60	450-500	1,30
200-250	0,95	500-550	1,65
250-300	1,00		

Antes de empezar el análisis hay que señalar que la variable cronología está expresada en intervalos y no en puntos fijos, por lo que puede haber cierto error de medida en esa variable. Para los propósitos del ejemplo, sugiero asumir que todos los yacimientos de una misma fase estaban ocupados en el momento central de esa fase, por lo que consideraremos el punto medio del intervalo como valor fijo.

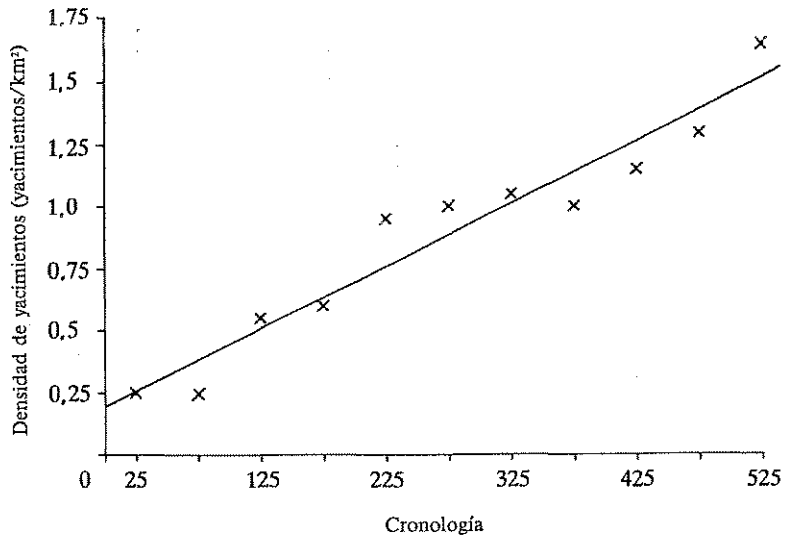


FIGURA 10.15. Gráfico de densidades de yacimientos con relación a su cronología.

El diagrama de dispersión de estos datos, con la recta de regresión superpuesta, aparece en la figura 10.15. La ecuación para esa recta es:

$$\hat{y} = 0{,}191 + 0{,}0025x$$

Nos explica que la densidad de yacimientos aumenta en 0,25 yacimientos/km² cada 100 años. El coeficiente de determinación, o r^2, es 0,932 indicando que el 93,2 % de la variación en la densidad de yacimientos está asociada con la evolución del tiempo, por lo que el ajuste de la regresión a los datos parece, por tanto, muy bueno.

Sin embargo, ¿podemos aceptar este valor simplemente por sí mismo, o bien puede inducir a error? La distribución de los puntos alrededor de la línea de regresión sugiere la posibilidad de autocorrelación en los residuales, si bien esta no es tan grande como para resultar significativa en la prueba de Durbin-Watson. Señala, no obstante, la existencia de un problema sustantivo en el análisis, ya que una mirada a la relación entre los puntos y la recta sugiere que hay cierta variación en la diferencia entre puntos adyacentes, a pesar del altísimo valor r^2, que insiste en mostrar la existencia de un buen ajuste entre la regresión y los datos. ¿Cómo ha podido suceder esto?

En casos como este es muy probable que muchos de los yacimientos ocupados en un período estuvieran también ocupados en el período siguiente, y muy posiblemente en el que vendría después también. Por eso las observaciones no

son independientes una de la otra, ya que un período dado está relacionado con el anterior. El resultado de este proceso de acumulación es que cuando la regresión nos dice que hay una proporción constante de aumento de 0,25 yacimientos/km² cada 100 años, con un r^2 del 93,2 %, puede proporcionar una impresión errónea de la proporción de cambio a través del tiempo, de su carácter constante y de la bondad del ajuste con los datos.

Para eliminar ese efecto acumulativo, en lugar de calcular la regresión de las densidades originales por la secuencia temporal, podemos calcular la diferencia entre la densidad de una fase con la de la precedente, repitiéndolo en todas las fases, y trazarla por la secuencia temporal; en otras palabras, la nueva definición de cambio de la densidad de yacimientos no está dada en los valores y_i originales de las observaciones, sino en términos de los valores $y_i - y_{i-1}$ de ellos (lo cual conlleva la pérdida de la primera de las observaciones). El diagrama de dispersión resultante aparece en la figura 10.16. No se trata, lógicamente, de una regresión como las demás que hemos visto, ya que el eje vertical está definido en términos de los aumentos de cambio entre una fase y la siguiente, y esperamos que la recta sea horizontal. Es decir, si realmente hubiera una proporción constante de cambio a lo largo del tiempo con un buen ajuste a los datos, tal y como indican la regresión original y r^2, las diferencias entre cada fase y la precedente habrían de ser virtualmente constantes a través de todo el período: estaría representada por la media de las diferencias $y_i - y_{i-1}$ y se registraría una variación negligible alrededor de ese valor. De hecho, como muestra el diagrama de dispersión, hay una gran cantidad de variación, lo que demuestra cuánto varía la proporción de cambio de la densidad de yacimientos de fase a fase, durante todo el período. Así, se constata que la regresión original proporciona una imagen distorsionada de la forma en que la densidad de yacimientos cambia a lo largo del tiempo.

Contemplando de esta manera los cambios entre los puntos adyacentes, se observa claramente la variación en las proporciones de cambio, pero no nos proporcionan una nueva imagen global de la relación entre la densidad de yacimientos y la cronología, o bien hasta qué punto esa relación es casi constante, como sugería la regresión original. Para ello no usaremos las medidas originales de densidad y cronología, ni siquiera las diferencias entre puntos adyacentes, sino que nos fijaremos en la diferencia entre todos los puntos, es decir, mediremos la diferencia cronológica y la diferencia de densidad de yacimientos entre todos los pares posibles de puntos y trazaremos un nuevo gráfico. El diagrama de dispersión aparece en la figura 10.17, con la recta de regresión superpuesta.

La pendiente es casi idéntica a la de la regresión original, tal y como debía de ser, pero ahora tenemos una imagen mucho menos confusa de la bondad del ajuste de los datos a la relación. La variación en la densidad entre fases adyacentes, ilustrada en el diagrama de dispersión anterior, se muestra en el rango de los valores y para $x = 50$; el rango de variación en las diferencias de

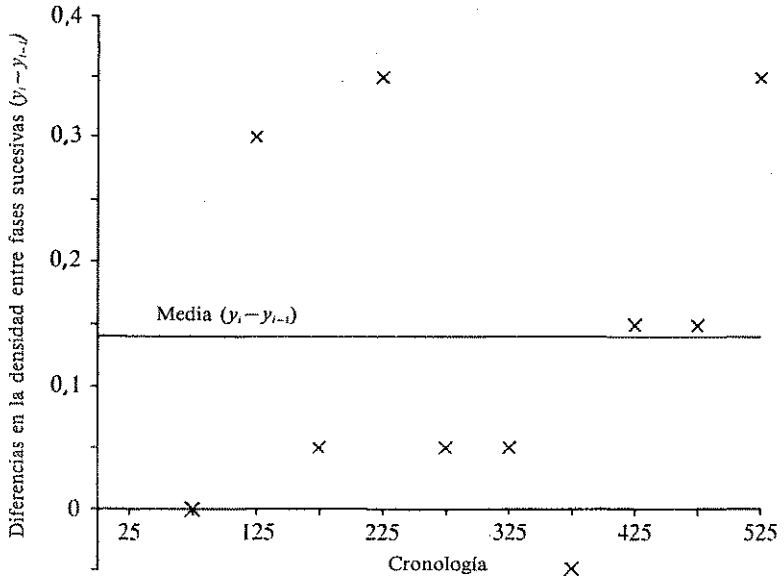


FIGURA 10.16. Gráfico de las diferencias en la densidad entre fases sucesivas, con relación a la secuencia cronológica.

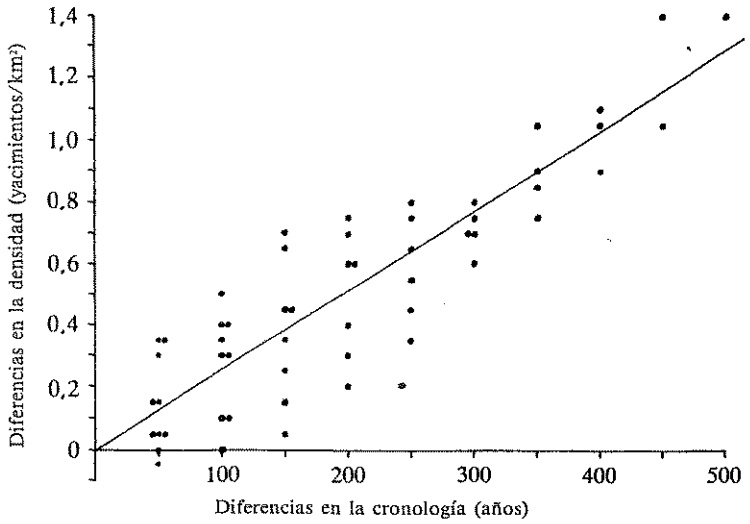


FIGURA 10.17. Diferencias en la densidad de asentamientos entre todos los pares de fases en relación a las diferencias cronológicas entre las dos fases.

densidad para fases distanciadas en 150 años es aún mayor. Todo esto aparece reflejado en una r^2 del 78,9 %, comparada con el 93,2 % de la regresión original.

De esta forma se concluye que la proporción de cambio en la densidad de asentamientos en esta área varió considerablemente durante el período en cuestión, y no fue tan constante como parecía al principio. Este resultado obligaría a plantearse nuevas cuestiones arqueológicas acerca de los motivos de esas proporciones de crecimiento tan cambiantes. Incluso aquí es preciso ser precavido, sin embargo, pues hay rastros de heterocedasticidad en las varianzas, y los residuales adyacentes de la diferencia de densidades tienden a estar correlacionados, pues la prueba de Durbin-Watson es significativa al 5 %.

Finalmente debe mencionarse que el procedimiento de calcular las diferencias de esta forma es conveniente sólo en este caso en particular, y no debe ser considerado como una receta universalmente apropiada para solucionar problemas de esta clase.

La discusión precedente de algunos de los problemas que pueden aparecer con un análisis de regresión por mínimos cuadrados, y los métodos utilizables para solucionarlos, ha introducido una cierta complejidad en la explicación, aunque incluso así se puede comprender el argumento. El propósito seguido es el de mostrar que el objetivo de esta técnica no es calcular mecánicamente dos o tres coeficientes que se limitan a fijar una cifra que ya conocíamos, sino obtener información acerca de las estructuras aparentes en nuestros datos. En esto los modernos enfoques en la teoría normal basados en la regresión son semejantes al análisis de datos exploratorio, con su característico énfasis en distinguir entre lo «áspero» y lo «suave» en una relación. Antes de volver al enfoque EDA aplicado a la regresión, sin embargo, es necesario que comentemos un aspecto de la regresión por mínimos cuadrados que aún no hemos mencionado.

Inferencia estadística

En nuestra explicación de la regresión y la correlación sólo hemos analizado los datos disponibles, describiendo la forma y la intensidad de la relación entre dos variables que nos interesan especialmente, dado un conjunto de datos específico. Es posible, naturalmente, usar la inferencia estadística en un contexto de regresión, cuando nuestros datos son una genuina muestra aleatoria de cierta población hipotética. Puede ser importante, por ejemplo, considerar el grado en que un valor determinado del coeficiente de correlación es estadísticamente significativo a cierto nivel, es decir, si difiere significativamente de una correlación con valor cero.

Más a menudo en arqueología no interesan cuestiones como esta, sino acerca

de los datos disponibles, comparándolos, todo lo más, con otros conjuntos de datos. Además, es discutible cuántas veces la regresión de datos arqueológicos satisface los requisitos de la inferencia estadística. Por esa razón, las pruebas de significación de la regresión y de la correlación no se incluyen aquí; pueden consultarse en cualquier manual (por ejemplo, Blalock, 1972).

Regresión robusta: el enfoque del análisis de datos exploratorio

El enfoque del análisis de datos exploratorio o EDA rechaza la forma clásica de la regresión de la misma manera que rechaza el uso de la media y de la desviación típica para describir la distribución de variables simples: están excesivamente afectadas por aquellos casos con valores extremos en el conjunto de los datos. El argumento es que la descripción de la relación entre dos variables ha de ser *robusta* y no estar influida por los extremos que, tal y como se ha visto, son casi siempre atípicos. Es un buen argumento, en tanto en cuanto funcione en la práctica, pero también hay que señalar que algunas de las transformaciones discutidas anteriormente pueden reducir la influencia de las observaciones extremas en la regresión por mínimos cuadrados, y hacer que los datos se conformen a las especificaciones del modelo de regresión. Usar los métodos EDA al primer indicio de no cumplimiento de los presupuestos del análisis de regresión clásico puede llevar a la pérdida de información, o bien a que no ponga de manifiesto toda la información contenida en los datos. En el estudio del caso de la autocorrelación, antes discutido, hubiese sido equivocado no considerar las implicaciones sustantivas de los indicios de autocorrelación que aparecían tras un primer análisis.

Sin embargo, allí donde se precise de la descripción robusta de una relación, como suele ser el caso, podremos emplear la alternativa EDA a la regresión por mínimos cuadrados, llamada *recta de Tukey*.

Al igual que en el enfoque EDA para la descripción de variables simples, está basada en la mediana y no en la media, pues la mediana es una medida robusta; comparada con el método de los mínimos cuadrados, tiene la ventaja adicional de la simplicidad.

El primer paso es dividir las observaciones en tres grupos más o menos iguales, según los valores del eje x, es decir, aquellas con valores x pequeños, medianos y grandes. Una vez hecho esto, se obtiene la mediana de los valores x y de los valores y en el primer y último grupos. A partir de aquí, hay dos formas de llegar a la recta de Tukey.

El primer método es gráfico y supone establecer la posición de la mediana de x y de y del primero y el último de los grupos de observaciones en el diagrama de dispersión, unirlas mediante una línea recta y mover esa línea, paralelamente, hacia abajo o hacia arriba, hasta que la mitad de los datos estén por encima y la mitad por debajo de la línea (Hartwig y Dearing, 1979, p. 35).

La alternativa es usar métodos aritméticos para calcular la pendiente y el punto de corte de la recta, donde su ecuación es la misma que la regresión por mínimos cuadrados $\hat{y} = a + bx$, si bien las fórmulas para calcular los coeficientes son distintas:

$$b = \frac{(\text{mediana } y_3 - \text{mediana } y_1)}{(\text{mediana } x_3 - \text{mediana } x_1)}$$

en donde la mediana de y_3 significa la mediana del valor y en el tercer grupo de observaciones —aquel con los mayores valores de x—; la mediana de y_1 significa la mediana de los valores del primer grupo de observaciones (aquel con los menores valores de x); la mediana x_3 significa la mediana de los valores de x en el tercer grupo de observaciones, y la mediana de x_1 significa la mediana de los valores de x en el primer grupo.

$$a = \text{mediana de los valores } d_i,$$
$$\text{donde } d_i = y_i - bx_i$$

Una vez calculados los coeficientes, podemos escribir la ecuación de la recta y trazarla de la forma habitual.

Ilustraremos este procedimiento con ayuda de un ejemplo (véase fig. 10.18), un estudio hipotético de la relación entre el tamaño de unos asentamientos mesoamericanos y la cantidad de obsidiana importada encontrada en ellos. Es un caso típico de las situaciones en las que puede emplearse la recta de Tukey, datos en los que parece que no se cumplen los requisitos de la regresión por mínimos cuadrados; en particular, allí donde hay observaciones extremas diferenciadas del resto, que influyen excesivamente en los coeficientes de una regresión ordinaria; en resumidas cuentas, la relación de regresión así definida no sería relevante en la mayoría de los casos.

La aplicación del método gráfico para obtener la recta de Tukey aparece en el diagrama de dispersión (fig. 10.18). Hay cinco puntos en cada uno de los tres grupos; en el tercero el mismo punto representa la mediana de x y la de y; en el primer grupo no es así. En este ejemplo en particular, la recta que une las dos medianas divide la nube de puntos exactamente por su mitad, por lo que no debe moverse ese eje a otro posición; la línea que une las dos medianas es la línea de Tukey que necesitábamos.

En el caso del método aritmético, tenemos:

$$b = \frac{(73{,}0 - 32{,}0)}{(42{,}5 - 7{,}0)} = \frac{41}{35{,}5} = 1{,}15$$

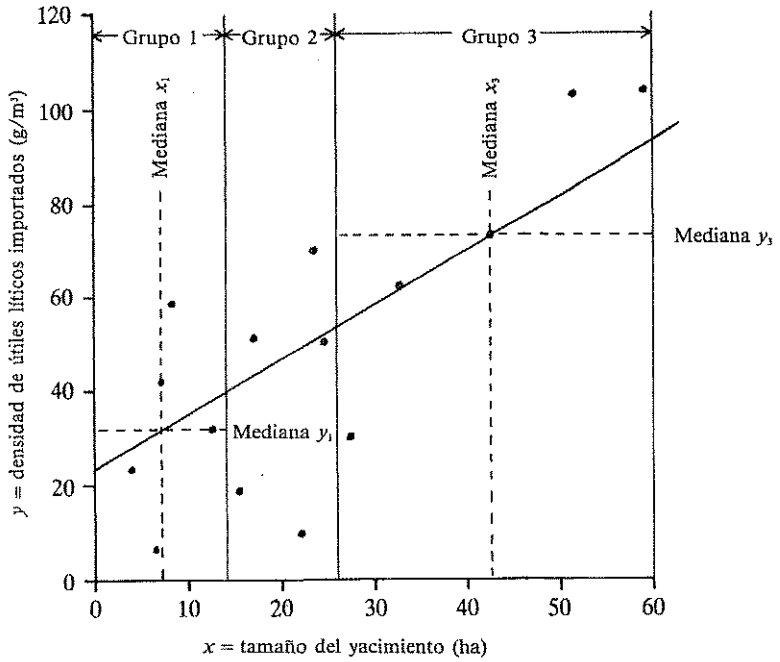


FIGURA 10.18. Cálculo de la línea de Tukey en un diagrama de las densidades de artefactos líticos importados en una serie de yacimientos en Mesoamérica, con relación al tamaño de los yacimientos en que aparecen.

Calcular el punto de corte es algo más laborioso, pues hemos de usar la fórmula:

$$d_i = y_i - 1{,}15\ x_i$$

con la cual obtener los valores d de todos los puntos, calculando a continuación la mediana de todos los valores d. Por ejemplo, en el caso del punto con el menor valor de y:

$$d = 7 - (1{,}15 \times 6{,}5) = -0{,}48$$

Para aquel con el mayor valor y:

$$d = 104 - (1{,}15 \times 59) = 36{,}15$$

La mediana de los valores d es 23,9; una simple mirada al diagrama de dispersión confirma su validez como punto de corte. Así, la ecuación de la recta de Tukey es:

$$\hat{y}_i = 23{,}9 + 1{,}15\ x_i$$

que nos dice que partiendo de un punto de 23,9 se produce un incremento en la densidad de hallazgos líticos de 1,15 g/m³ de tierra, por cada hectárea de aumento en el tamaño del yacimiento.

Esta recta es mucho más robusta que la correspondiente a la regresión por mínimos cuadrados, si eliminamos los dos o tres casos superiores, los cuales tienen un efecto mucho menor en la forma de la relación que el conjunto de los datos como un todo. Es decir, que en el caso que consideramos aquí, la proporción de aumento en la densidad de hallazgos líticos indicada por la ecuación de la recta de Tukey se aplica a la mayoría de los casos agrupados, y no resulta de la diferencia entre los yacimientos mayores y el resto. Incluso si eliminamos esos casos extremos, la proporción entre el aumento en la densidad de hallazgos líticos y el aumento del tamaño de los yacimientos varía muy poco, mientras que, en la relación por mínimos cuadrados, la ecuación varía considerablemente si eliminamos esos casos (fig. 10.19).

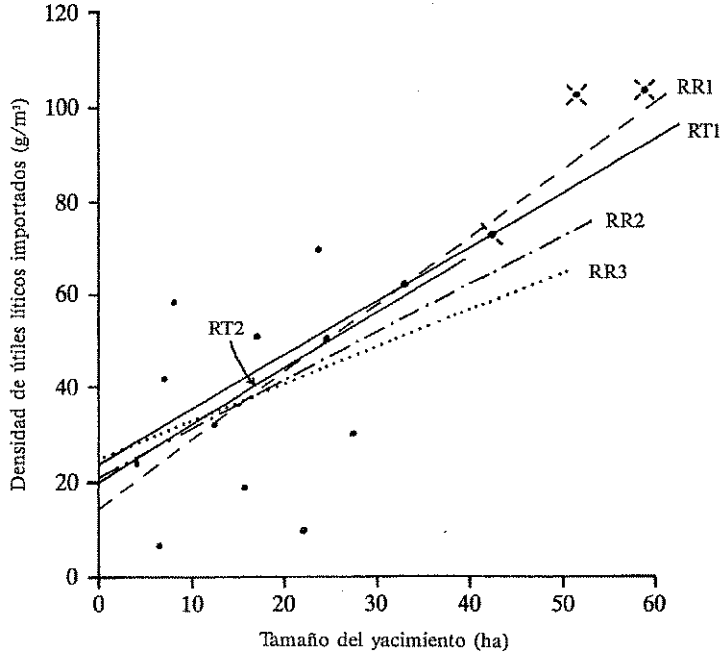


FIGURA 10.19. Comparación de las rectas de Tukey y de regresión por mínimos cuadrados para todos los datos y con las observaciones mayores eliminadas: RT1 es la recta de Tukey para todos los datos; RT2 es la recta de Tukey sin las tres observaciones mayores: RR1 es la recta de regresión por mínimos cuadrados para todos los datos; RR2 es la recta de regresión por mínimos cuadrados, sin las dos observaciones mayores; RR3 es la recta de regresión por mínimos cuadrados, sin las tres observaciones mayores.

58

Tal y como Hartwig y Dearing (1979) señalan, en muchos casos, la idea fundamental que subyace al intento de ajustar una línea a los datos es definir la estructura general (lo suave) y distinguirla de las desviaciones de ese esquema (lo áspero); una caracterización resistente de lo suave es, probablemente, mantener la distinción entre lo suave y lo áspero lo más claramente posible.

EJERCICIOS

10.1. Estudia la relación entre el porcentaje de obsidiana y la distancia a la fuente de los siguientes conjuntos líticos, usando técnicas de regresión para especificar su forma e intensidad.

Distancia	% Obsidiana	Distancia	% Obsidiana	Distancia	% Obsidiana
12	98	85	21	210	8
25	92	82	44	233	16
67	77	112	56	300	10
30	67	150	33	329	5
42	39	154	15	381	8

10.2. Muchos estudios procesuales recientes han considerado la población como una variable fundamental. Habitualmente se infiere el tamaño de la población del tamaño del asentamiento. En este caso, se ha estudiado esa relación en los pueblos actuales del área en estudio, como base para hacer estimaciones acerca de la población en el pasado. Los datos (Carothers y McDonald, 1979) son los siguientes:

Tamaño del asentamiento (ha)	Población	Tamaño del asentamiento (ha)	Población
0,6	20	3,7	300
1,0	70	4,0	250
1,1	100	4,5	500
1,2	130	5,4	270
1,6	120	5,9	190
1,9	170	6,1	630
2,3	195	6,4	650
3,0	190	8,9	310
3,1	210	10,0	730
3,3	360	12,0	850

¿Qué relación se establece entre el tamaño de los asentamientos y la población? ¿Cuán estrecha es? ¿Presentan los datos algún problema para este tipo de análisis? Si es así, ¿qué se puede inferir de ello?

10.3. Se ha medido la longitud de varios huesos y el número de marcas producidas por dientes en diez huesos de animales procedentes de una cueva paleolítica:

Longitud hueso (cm)	4	4	5	6	7	8	9	11	13	14
N.º marcas dientes	0	0	1	2	0	5	0	2	7	0

(a) Representa gráficamente los datos y traza la recta de Tukey. (b) Calcula la línea de regresión para esos datos. Explica cuál es la variable dependiente y cuál la independiente y por qué. ¿Se cumplen los requisitos de la regresión? (c) Comenta brevemente la interpretación de las líneas obtenidas en (a) y (b).

59

11. ENFRENTÁNDOSE A LA COMPLEJIDAD: CORRELACIÓN Y REGRESIÓN MÚLTIPLES

A lo largo del capítulo 10 muchas ideas difíciles y complejas se han ido sumando a los conceptos básicos de regresión y correlación con los que empezamos. Sin embargo, el análisis ha estado limitado en todo momento a la relación entre dos variables, una dependiente y la otra independiente.

Tal y como vimos en el capítulo 7, si pretendemos obtener una comprensión real de una situación en particular, es necesario tratar con más de dos variables. Así, por citar uno de los temas del capítulo anterior, si nos interesa saber por qué las cantidades de un material importado varían en distintos yacimientos, la mayoría de las veces consideraremos que la distancia a la fuente de ese material es el motivo de la variabilidad. El último ejemplo que citábamos, no obstante, era el de una investigación hipotética en la que las cantidades de material importado se relacionaban con el tamaño del yacimiento. Cualquier estudio auténtico, naturalmente, habría de considerar tanto el efecto de la distancia como el del tamaño del asentamiento en la densidad de hallazgos de materiales importados (véase Sidrys, 1977). Como veremos, esto no puede hacerse calculando simplemente dos análisis de regresión bivariantes por separado; las tres variables han de incluirse en un *análisis de regresión múltiple*, en el que haya una sola variable dependiente —cantidad de material— y dos independientes. En general, en los análisis de regresión múltiple siempre habrá una sola variable dependiente y varias independientes, las cuales consideraremos que afectan a la variabilidad de la dependiente, según ciertas hipótesis que habremos desarrollado.

La creciente complejidad del análisis a medida que se añaden variables a las dos originales afecta la forma de llevar a cabo los cálculos. Mientras que virtualmente todas las técnicas que hemos visto hasta ahora pueden realizarse con una simple calculadora, la regresión múltiple y las técnicas que describiremos en los capítulos siguientes necesitan de un ordenador, excepto en los casos más simples y triviales, y ello es debido a la complejidad de los cálculos. Esta

complejidad está asociada, a su vez, con un mayor nivel de dificultad matemática, debido al uso del álgebra matricial.

Parecía poco apropiado en un libro como este presentar una introducción al álgebra matricial y detallar los fundamentos matemáticos de las técnicas. Nos hubiera ocupado mucho espacio y hubiese exigido una sofisticación matemática tal que lo haría poco útil para el público al que este libro está dirigido. Sin embargo, hay que pagar un precio por esta simplificación. Si hasta ahora hemos podido detallar todos los procedimientos de cálculo, a partir de este momento las técnicas de análisis se van a convertir en una «caja negra». Esto comporta sus riesgos, que han conducido al error a muchos arqueólogos en el pasado (véase Thomas, 1978); para aquellos que pretendan ser practicantes serios de estas técnicas no hay otra alternativa que adquirir el conocimiento preciso y (antes que «o») pedir consejo a los estadísticos profesionales. Sin embargo, creo que es posible obtener una comprensión de la estructura teórica de las técnicas sin un conocimiento específico de las matemáticas necesarias, con lo que se gana una visión intuitiva válida acerca de su función y del papel que desempeñan.

Este capítulo empieza con una breve introducción sobre el fundamento del modelo de regresión múltiple. Sigue un examen más detallado de varios aspectos de la regresión y la correlación múltiples, recurriendo a un ejemplo arqueológico, de forma que la argumentación no sea demasiado teórica y se pongan de manifiesto las implicaciones de esas técnicas en el análisis de datos arqueológicos.

El modelo de regresión múltiple

Los principios de la regresión múltiple son los mismos que los de la regresión simple. En general, pretendemos estimar una ecuación de regresión ajustándola a ciertos datos empíricos. Se asume que la relación es lineal y que usamos el criterio de los mínimos cuadrados para obtener el mejor ajuste de la regresión a los datos. Mientras que para la regresión simple la ecuación era $y = a + bx$, ahora es:

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

En el caso de la regresión simple, ajustábamos una línea a una nube de puntos bidimensional (fig. 11.1). En el caso de dos variables independientes, lo que intentaríamos ajustar sería un plano y no una línea (fig. 11.2). Una vez que pasamos el límite de tres variables, la situación se hace más difícil de visualizar, si bien los principios son los mismos: intentamos ajustar un plano, no de dos dimensiones, sino de muchas, tantas como variables independientes haya.

Volviendo al caso de las tres variables. Donde $x_1 = x_2 = 0$, tenemos $y = a$, que es la altura en la que el plano de la regresión cruza el eje y. El coeficiente

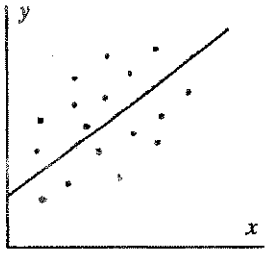


FIGURA 11.1. Diagrama de dispersión de la relación entre una variable dependiente (y) y una independiente (x), con la recta de regresión dibujada.

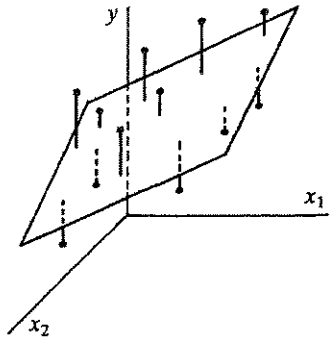


FIGURA 11.2. Diagrama de dispersión de la relación entre una variable dependiente (y) y dos independientes (x_1 y x_2): la *recta* de regresión se ha convertido en un *plano* de regresión, que aparece dibujado en la figura (según Blalock, 1972).

b funciona del siguiente modo. Imaginemos un plano vertical perpendicular al eje x_2, proyectado de forma que cruce el plano de regresión (fig. 11.3). En el punto de intersección con el plano de regresión, este plano vertical es, simplemente, una línea recta sobre la superficie de regresión. Debido a que el plano vertical —la línea de intersección— es perpendicular al eje x_2, todos los puntos en él tendrán el mismo valor para la variable x_2. La pendiente de esa línea es b_1 en la ecuación de regresión múltiple; es decir, es la pendiente de la regresión de y sobre x_1, ya que para esa línea en particular los valores de x_2 son constantes. De la misma manera, si construimos un plano vertical perpendicular al eje x_1, la línea a lo largo de la cual cruza el plano de regresión tendrá como pendiente b_2 y representará la regresión de y sobre x_2, con x_1 constante. El propósito de la regresión múltiple es encontrar los coeficientes a, b_1 y b_2 que produzcan el plano de regresión que mejor se ajuste a los datos, según el criterio de los mínimos cuadrados. Veremos más adelante cómo la bondad del ajuste de ese plano a los datos puede medirse usando el coeficiente de correlación múltiple.

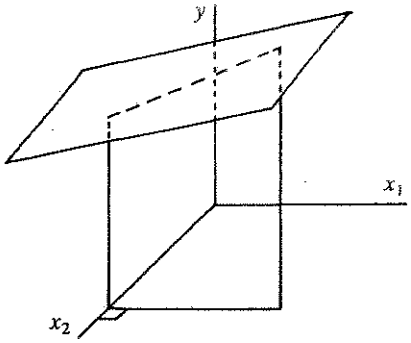


FIGURA 11.3. Un plano vertical, perpendicular a x_2 proyectado hacia arriba hasta que intersecta con el plano de regresión: la recta de intersección representa la regresión de y sobre x_1.

Sin embargo, la regresión múltiple y la correlación no tratan de encontrar el efecto global de un conjunto de variables sobre la dependiente. Tal y como se ha aducido, nos interesa estudiar el efecto de cada variable independiente por separado, manteniendo constantes las demás. El experimentador de laboratorio puede hacerlo en la realidad, manipulando las condiciones de su experimento. Los arqueólogos no pueden hacerlo, obviamente; hemos de controlar nuestros experimentos en la medida de lo posible durante el análisis. Para ese control, hay que usar los coeficientes parciales, tal y como hicimos en el caso de las variables dicotómicas, pero aquí usando los coeficientes de correlación y regresión parciales.

CORRELACIÓN PARCIAL

Empezaremos con la correlación parcial, que es el más importante de las dos, y la ilustraremos recurriendo a un ejemplo. Supongamos que un programa de prospecciones y excavaciones nos proporciona información sobre el tamaño (en términos de área) de unos asentamientos mexicanos. Nos interesa saber los motivos de la variabilidad, y sospechamos que tiene algo que ver con los recursos agrícolas disponibles en la región (véase Brumfield, 1976). Se recoge la información acerca de la extensión de tierra cultivable alrededor de cada yacimiento y de la productividad de esa tierra (tabla 11.1).

TABLA 11.1. Información acerca del tamaño de los yacimientos, extensión de la tierra cultivable y productividad de la misma (en unidades arbitrarias) en 28 yacimientos hipotéticos del período formativo en México.

Tamaño del yacimiento (ha)	Tierra cultivable disponible (km^2)	Índice de productividad relativa
30,0	17,9	0,75
33,0	12,7	0,87
37,0	17,6	0,71
42,0	6,0	0,85
42,0	21,6	0,83
44,9	29,4	0,73
47,0	19,6	0,89
53,2	29,0	0,87
55,0	21,4	0,72
55,0	50,8	0,89
55,2	31,8	0,90
60,0	24,8	0,81
62,0	26,4	0,92
63,1	34,0	0,94
64,5	39,1	0,99
65,0	35,4	0,82
67,7	34,8	0,96
69,7	53,0	0,91
74,0	54,2	0,94
75,0	73,3	1,01
76,0	95,9	1,09
77,0	66,8	1,05
80,5	51,0	1,23
86,0	61,2	1,06
88,0	72,5	1,29
90,0	54,7	1,22
95,3	89,9	1,00
99,0	89,9	1,26

En términos de análisis de regresión:

Variable dependiente (y) = tamaño del asentamiento
Primera variable independiente (x_1) = extensión de tierra cultivable alrededor de cada yacimiento
Segunda variable independiente (x_2) = productividad relativa de la tierra cultivable

Podemos empezar calculando una regresión simple del tamaño del yacimiento sobre la tierra cultivable disponible. Obtenemos los siguientes resultados:

$$\hat{y} = 35{,}4 + 0{,}656x_1$$
$$r_{yx_1} = 0{,}864$$
$$r^2_{yx_1} = 0{,}746$$

El diagrama de dispersión correspondiente aparece en la figura 11.4. En palabras, este resultado nos explica que allí donde no hay tierra cultivable, el tamaño estimado del yacimiento es de 35,4 ha, y por cada incremento de 1 km^2 en tierra cultivable el tamaño del asentamiento aumenta en 0,656 ha. La correlación entre ambas variables es de 0,864. Dado que la disponibilidad de tierra cultivable es la variable independiente y que el tamaño del yacimiento es la dependiente, podemos decir que la variabilidad en la tierra cultivable explica el 74,6 % de la variabilidad del tamaño de los yacimientos.

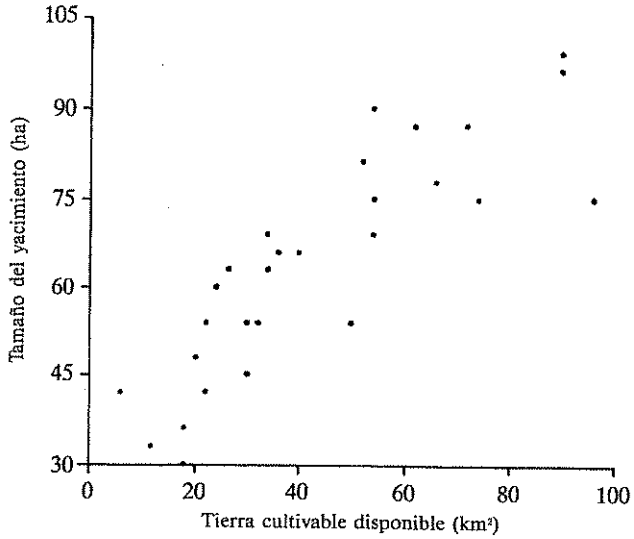


FIGURA 11.4. Diagrama de dispersión del tamaño del yacimiento con relación a la superficie de tierra disponible, según los datos de la tabla 11.1.

Igualmente, si calculamos la regresión del tamaño del asentamiento y la productividad de la tierra, se obtiene:

$$\hat{y} = -28{,}9 + 97{,}9\ x_2$$
$$r_{yx_2} = 0{,}832$$
$$r^2_{yx_2} = 0{,}693$$

El diagrama de dispersión aparece en la figura 11.5; el resultado sugiere que para cada incremento de 1,0 en el índice de productividad, se produce un aumen-

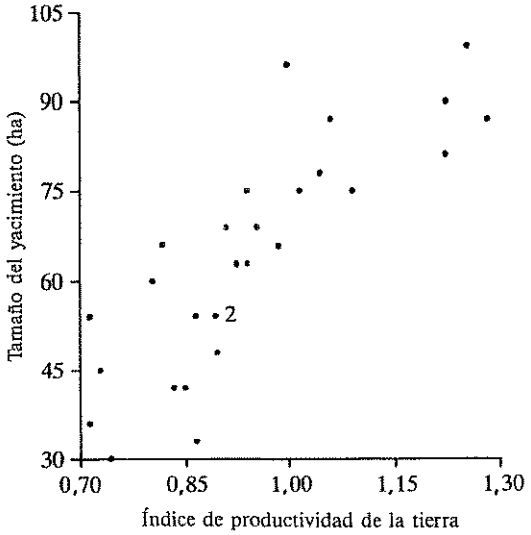


FIGURA 11.5. Diagrama de dispersión del tamaño del yacimiento con relación al índice de productividad de la tierra disponible, según los datos de la tabla 11.1.

to de 97,9 ha en el tamaño del asentamiento. La correlación es 0,832 y la variación en la productividad de la tierra explica el 69,3 % de la variabilidad en el tamaño del yacimiento.

Juntos, los dos valores r^2 parecen indicar que las dos variables independientes explican el 74,6 + 69,3 = 143,9 % de la variabilidad en el tamaño del yacimiento. Claramente, esto debiera hacernos sospechar que algo anda mal. ¿Qué es lo equivocado en el procedimiento?

Supongamos que nos preguntamos si la productividad de la tierra y el área disponible de tierra cultivable están relacionadas. Esto sería así, por ejemplo, si las condiciones geomorfológicas en las que se forman las áreas de suelo más extensas fueran distintas de aquellas en las que se originan áreas de menor extensión; el contraste entre las grandes llanuras aluviales y las cuencas reducidas rellenas de coluvión puede ser un ejemplo. Esta cuestión puede analizarse calculando la regresión de x_2 (productividad) sobre x_1 (extensión de la tierra cultivable):

$$\hat{x}_2 = 0{,}74 + 0{,}0048x_1$$
$$r_{x_2x_1} = 0{,}738$$
$$r^2_{x_2x_1} = 0{,}545$$

Un diagrama de dispersión (con la productividad como eje vertical) aparece en la figura 11.6. La ecuación de la recta afirma que el 54,5 % de la variabilidad de la productividad de la tierra está explicado por la variabilidad en la extensión de la tierra.

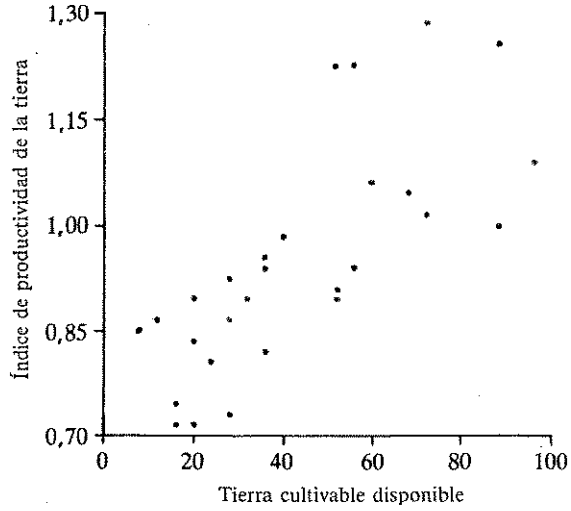


FIGURA 11.6. Diagrama de dispersión de la superficie disponible de tierra con relación al índice de productividad, según los datos de la tabla 11.1.

Esto plantea problemas a las dos regresiones iniciales, pues significa que no trabajamos con dos variables independientes. La segunda regresión, y sobre x_2, era en parte, también, una regresión de y sobre x_1, pues x_1 y x_2 están relacionadas. Esta situación provoca dos dificultades. En primer lugar, no podemos decir cuánta variación del tamaño del asentamiento está relacionada con el área disponible de tierra cultivable, cuánta con la productividad de la tierra y cuánta con la unión de ambas. En segundo lugar, significa que, como sospechábamos, es incorrecto pensar que las variables independientes explican el 144 % de la dependiente: se ha sumado dos veces un mismo componente. No podemos sumar los dos valores r^2 pues los dos están parcialmente superpuestos como resultado de la relación entre x_1 (extensión de la tierra cultivable) y x_2 (productividad). La correlación parcial y la correlación múltiple desempeñan un importante papel en la solución de estos problemas.

Es conveniente cambiar la notación seguida hasta ahora, y designar la variable dependiente como x_0. Nuestro coeficiente de correlación parcial puede simbolizarse entonces como $r_{01.23}$, que se puede leer como «correlación entre las variables 0 y 1, controlados los efectos de las variables 2 y 3». Resulta así mucho más claro que si se incluyen los índices x e y. De las dos variables antes del punto, la primera suele ser la dependiente, y la segunda, la independiente que se está estudiando. Puede controlarse cualquier cantidad de variables inde-

pendientes; el número de ellas bajo control se conoce como orden de la correlación, de forma que si controlamos dos, trabajaremos con una correlación de segundo orden, y si no controlamos ninguna, será una correlación de orden cero.

Es importante señalar que con el coeficiente de correlación parcial, $r_{01.2}$, eliminamos el efecto no sólo de la relación entre x_0 y x_2, sino también el de la relación entre x_1 y x_2.

Los coeficiente parciales de primer orden, aquellos que sólo controlan una variable, pueden calcularse por medio de la siguiente fórmula:

$$r_{ij.k} = \frac{r_{ij} - (r_{ik})(r_{jk})}{\sqrt{1 - r_{ik}^2}\sqrt{1 - r_{jk}^2}}$$

Por ejemplo:

$$r_{01.2} = \frac{r_{01} - (r_{02})(r_{12})}{\sqrt{1 - r_{02}^2}\sqrt{1 - r_{12}^2}}$$

Si queremos obtener los coeficientes proporcionales de segundo orden, el procedimiento es, esencialmente, el mismo:

$$r_{ij.kl} = \frac{r_{ij.k} - (r_{il.k})(r_{jl.k})}{\sqrt{1 - r_{il.k}^2}\sqrt{1 - r_{jl.k}^2}}$$

Es fácil de ver que el cálculo manual de estos coeficientes se hace muy laborioso, a causa de la gran cantidad de números. De hecho, son fácilmente accesibles en paquetes informáticos como SPSS-X (véase anexo 2).

Llegados a este punto, una vez visto en abstracto lo que es un coeficiente parcial y cómo se calcula, es importante volver al ejemplo que introducíamos antes para ver cómo pueden obtenerse en la práctica los coeficientes parciales de correlación y cómo difieren de los coeficientes de orden cero. Hemos de estudiar la relación entre el tamaño del yacimiento y la extensión de tierra cultivable, manteniendo la productividad de la tierra constante; y también, la relación entre el tamaño del yacimiento y la productividad de la tierra, manteniendo constante la extensión de tierra cultivable. Para hacer esto, sustituimos los términos en la fórmula anterior:

$$r_{01.2} = \frac{r_{01} - (r_{02})(r_{12})}{\sqrt{1 - r_{02}^2}\sqrt{1 - r_{12}^2}} = \frac{0{,}864 - (0{,}832)(0{,}738)}{\sqrt{1 - 0{,}832^2}\sqrt{1 - 0{,}738^2}} = 0{,}6678$$

$$r_{01.2}^2 = 0{,}6678^2 = 0{,}446$$

Es decir, que sólo el 44 % de la variabilidad en el tamaño de los asentamientos está explicada por la variabilidad en la extensión de la tierra cultivable, una vez controlada la variación en la productividad. Es evidente la diferencia con el 75 % de la variación explicada por la misma variable cuando la productividad de la tierra no estaba controlada.

A continuación se calcula el mismo coeficiente para la relación entre el tamaño del yacimiento y la productividad, controlando ahora la extensión de tierra cultivable. Usando la fórmula habitual para los coeficientes de primer orden:

$$r_{02.1} = \frac{r_{02} - (r_{01})(r_{21})}{\sqrt{1 - r_{01}^2}\sqrt{1 - r_{21}^2}} = \frac{0{,}832 - (0{,}864)(0{,}738)}{\sqrt{1 - 0{,}864^2}\sqrt{1 - 0{,}738^2}} = 0{,}572$$

$$r_{02.1}^2 = 0{,}572^2 = 0{,}327$$

Así, puede verse que sólo el 32 % de la variabilidad en el tamaño de los yacimientos está explicada por la productividad de la tierra, una vez que la extensión de tierra cultivable está controlada, es decir, cuando se tiene en cuenta el efecto de esa última variable sobre las dos primeras. El resultado difiere claramente del 69 % obtenido en el coeficiente de correlación de orden cero, al no controlar la extensión de tierra cultivable.

De este ejemplo, debiera deducirse que cualquier investigación seria de las relaciones entre variables no puede dejarse en el nivel de coeficientes de orden cero. Sin embargo, el análisis y la comprensión de las relaciones entre un gran número de coeficientes de orden cero y parciales es un asunto complicado; los procedimientos son los mismos que usábamos para el coeficiente *Q*. Es preciso examinar las diferencias en el signo y en la magnitud entre coeficientes parciales y coeficientes de orden cero, para ver si la variable o variables controladas pueden suprimirse al no explicar o no tener efecto relevante alguno sobre las variables que nos interesa estudiar. Como antes, es necesario tener una hipótesis bien definida sobre las relaciones que hay que investigar.

CORRELACIÓN MÚLTIPLE

La sección previa de este capítulo trataba acerca de cómo aislar el efecto de las variables individuales en el contexto de un análisis de regresión con más de una variable independiente. Lo que aún no hemos considerado, sin embargo, es cómo se estudia el efecto global del conjunto de las variables independientes sobre la dependiente. El coeficiente de correlación múltiple (R) mide la bondad del ajuste de la superficie de regresión (criterio de los mínimos cuadrados), considerada como un todo, con respecto a los valores de la variable dependiente. El cuadrado del coeficiente de correlación múltiple (R^2) indica el

porcentaje de variación en la variable dependiente explicado por la superficie de regresión.

Como definición general, la anterior puede pasar; ahora bien, ¿cómo se calcula ese índice, o su cuadrado, ante los problemas vistos al principio en torno a la suma de coeficientes de orden cero y a la necesidad de calcular los coeficientes parciales? Tal y como se ponía de manifiesto en nuestro ejemplo, el porcentaje de variación de la variable dependiente explicado por la regresión global no podía ser la simple suma de los valores r^2 de orden cero, y ello por la misma razón que antes aducíamos.

Podría parecer que la respuesta obvia sería la suma de los coeficientes parciales. Pero esto también es incorrecto. Mientras que la suma de los valores de orden cero implica el contar dos veces la misma cantidad, la suma de los parciales plantea el problema inverso de que no es suficiente. Esto sucede porque cada coeficiente parcial proporciona sólo el efecto de una variable en sí misma, sin considerar la influencia de las otras variables independientes sobre ella; por lo que al sumarlas sólo se consigue el efecto de cada variable independiente en la parte de la variación de la variable dependiente que no se relaciona con las otras variables. Lo que falta es que cuando las variables independientes aparecen correlacionadas entre sí, no producen un efecto por separado sobre la dependiente, sino que existirá un efecto conjunto de todas las variables independientes.

Si consideramos el caso de dos variables independientes, $r_{01.2}$ explica la relación en x_0 no asociada a x_2, y $r_{02.1}$ explica la variación en x_0 no asociada a x_1; sin embargo, como las dos variables están correlacionadas, algo de la variación de x_0 está explicado por la variación conjunta de x_1 y x_2, la cual no está incluida en los coeficientes parciales (véase Johnston, 1978, capítulo 3). Por ese motivo, necesitamos algo diferente a todos los coeficientes que hemos visto hasta ahora.

La fórmula de R^2 múltiple en el caso de tres variables es:

$$R^2_{01.2} = r^2_{01} + r^2_{02.1}\,(1 - r^2_{01})$$

donde $R^2_{01.2}$ es el coeficiente de determinación múltiple entre x_0 de un lado y x_1 y x_2 simultáneamente, del otro. Es la proporción de variación en x_0 explicada por las dos variables independientes, tanto por separado como juntas.

Para calcularlo, hemos de permitir, en primer lugar, que una de las variables independientes «explique» todo lo que pueda. Esto es lo que significa el término r^2_{01}: la proporción de la variación en la dependiente explicada por la primera variable independiente. Si la variación total en la dependiente está definida como 1,0 una vez que la primera variable explica su parte de la variación, la proporción restante es $(1 - r^2_{01})$. De este modo podemos ver que gran parte de la variación restante en la dependiente está explicada por la segunda independiente y se añade a la variación explicada por la primera para obtener

el efecto global de las dos juntas. ¿Por qué entonces la segunda parte de la fórmula es $r^2_{02.1}\,(1 - r^2_{01})$ y no simplemente $r^2_{02}\,(1 - r^2_{01})$?

Precisamente, porque al incluir el término r^2_{01} ya tenemos en la ecuación todo el efecto de la primera variable. Si x_1 y x_2, las variables independientes, están correlacionadas, entonces algún efecto de x_1 estará expresado en r^2_{02}, con lo que si lo usamos en la ecuación volveremos a cometer el error de contarlo dos veces. Hemos de eliminar cualquier efecto de x_1 sobre r^2_{02}, y lo haremos, naturalmente, controlándolo y obteniendo el coeficiente parcial $r^2_{02.1}$.

Volviendo al ejemplo anterior, obtener R^2 significa establecer la proporción de variación en el tamaño de los asentamientos, explicada por el efecto global de la extensión de la tierra cultivable y de la productividad de la tierra, actuando tanto conjuntamente como por separado. Conocemos las cifras necesarias para sustituir las expresiones necesarias en la fórmula:

$$r^2_{01} = 0{,}746$$
$$1 - r^2_{01} = 0{,}254$$
$$r^2_{02.1} = 0{,}327$$
$$R^2_{01.2} = 0{,}746 + (0{,}327)(0{,}254) = 0{,}746 + 0{,}083 = 0{,}829$$

Es decir, que la extensión de la tierra cultivable y la productividad de la tierra explican el 82,9 % de la variación en el tamaño de los asentamientos para ese conjunto de datos en particular, dejando que la extensión de la tierra cultivable explique todo lo que pueda de la variabilidad (74,6 %) y dejando a la productividad de la tierra «explicar» lo que pueda del resto (8,3 %).

Podemos desarrollar aún más la argumentación: no sabemos todavía cuánto de ese 74,6 % es un efecto de la extensión de la tierra cultivable, y cuánto es el efecto conjunto, de esa variable y la productividad de la tierra. El coeficiente parcial $r^2_{01.2} = 0{,}446$ nos explica que el 44,6 % de la variabilidad del tamaño de los yacimientos (x_0) no explicada por la productividad de la tierra (x_2) está explicada por la extensión de tierra cultivable por sí sola, es decir, el 44,6 % del valor $(1 - r^2_{02})$ del 30,7 %. Para descubrir qué proporción de la variación total en el tamaño de los yacimientos está explicada por la variación en la disponibilidad de tierra cultivable por sí sola, calculamos:

$$r^2_{01.2}\,(1 - r^2_{02}) = (0{,}446)\,(1 - 0{,}693) = 0{,}137$$

Es decir, el 13,7 % de la variación en el tamaño del asentamiento está explicada exclusivamente por la variación en la extensión de tierra cultivable.

La cantidad correspondiente para la productividad de la tierra ha sido calculada ya al obtener el valor R^2 múltiple:

$$r^2_{02.1}\,(1 - r^2_{01}) = (0{,}327)\,(1 - 0{,}746) = 0{,}083$$

Así pues, el 8,3 % de la variación en el tamaño de los yacimientos está explicado sólo por la productividad.

Dado que el R^2 múltiple global era el 82,9 % y que solamente 8,3 + + 13,7 = 22 % es atribuible a los efectos de ambas variables independientes por separado, se aprecia que el 60,9 % de la variación en la variable dependiente está explicado por el efecto conjunto de las dos independientes, es decir, las grandes extensiones de tierra cultivable suelen tener una mayor fertilidad que las pequeñas extensiones.

En este caso, por lo tanto, como en muchas otras situaciones empíricas, no es despreciable el grado de superposición en los efectos de las dos variables, de forma tal que R^2 múltiple no es mucho mayor que los valores r^2 simples. El extremo opuesto, claro está, sucede cuando las variables independientes no están correlacionadas y la fórmula de R^2 se reduce a:

$$R^2_{01.2} = r^2_{01} + r^2_{02}$$

es decir, la varianza explicada es simplemente la suma de los dos valores r^2 de orden cero. R^2 obviamente alcanza su valor máximo en una situación como ésta, que es claramente preferible, pues las variables independientes explican *partes distintas* de la variación de la dependiente. Si estuviesen correlacionadas, por otro lado, explicarían la misma variación, introduciendo ambigüedad en nuestras interpretaciones. Este problema, que tiene ramificaciones técnicas muy complejas en el análisis de la regresión, es conocido como el de la *colinealidad* o *multicolinealidad* (véase Chatterjee y Price, 1977).

Una manera obvia de identificar la colinealidad es comparando los coeficientes simples y múltiples. Supongamos dos variables independientes, x_1 y x_2, y una dependiente, x_0. Si calculamos primero el coeficiente de orden cero r^2_{01} y a continuación el múltiple $R^2_{01.2}$, la diferencia entre ambos puede interpretarse como la mejora en la explicación estadística conseguida al añadir el efecto de la segunda variable al de la primera. Si esa diferencia es muy pequeña, nos sugiere que la segunda variable está fuertemente correlacionada a la primera, o bien que no afecta a la dependiente. ¿Cuál de las dos posibilidades es la correcta? Lo sabremos observando el valor r^2_{12}, el coeficiente de determinación entre las dos variables independientes. Si es muy alto, confirmará la existencia de colinealidad.

Una forma de solucionar el problema es, simplemente, eliminar una de las variables independientes del análisis; otra forma sería deshacer la correlación entre las variables, por ejemplo, por medio del análisis de componentes principales (véase el capítulo 13). Por otro lado, no necesariamente hemos de desprendernos o redefinir las relaciones, pues muchas de ellas pueden proporcionar gran cantidad de información, como en el ejemplo del tamaño de los asentamientos.

El coeficiente de regresión múltiple

Hasta ahora hemos estado dejando de lado la ecuación de regresión múltiple, a causa de la gran importancia del concepto de correlación múltiple y parcial.

Se afirmó al principio del capítulo, cuando se introdujo el modelo de regresión múltiple, que los coeficientes de pendiente (las b) se referían a la cantidad de cambio en la dependiente para un cambio dado en una independiente específica, mientras que la otra independiente permanecía constante. A la luz de la discusión sobre la correlación parcial, es evidente que esos coeficientes de pendiente representan algo muy similar. De hecho, son conocidos como *coeficientes de regresión parciales*. Su notación en la ecuación de regresión es también muy parecida:

$$\hat{x}_{0.1\ldots k} = a_{0.1\ldots k} + b_{01.2\ldots k}x_1 + b_{02.1,3\ldots k}x_2 + \ldots + b_{0k.1\ldots k-1}x_k$$

o bien, en el caso de dos variables:

$$\hat{x}_{0.12} + a_{0.12} + b_{01.2}x_1 + b_{02.1}x_2$$

a representa el valor de la recta de corte, cuando los valores de todas las variables independientes son igual a cero.

Las fórmulas para los coeficientes a y b son:

$$a_{0.12} = \bar{x}_0 - b_{01.2}\bar{x}_1 - b_{02.1}\bar{x}_2$$

$$b_{01.2} = \frac{b_{01} - (b_{02})(b_{21})}{1 - b_{12}b_{21}}$$

Es fácil de ver que la fórmula para b es muy parecida a la de r parcial.

Disponemos ahora de un coeficiente de regresión parcial que indica el incremento absoluto en nuestra variable dependiente, siendo constantes las demás; naturalmente, podemos hacer lo mismo para las demás.

En general, no hemos de hacer los cálculos para obtener los índices necesarios, pues la regresión múltiple suele ser un procedimiento informático. La ecuación obtenida por MINITAB (veáse más adelante), para el ejemplo del tamaño del yacimiento, la extensión de la tierra cultivable a su alrededor y la productividad de la misma, es:

$$\hat{x}_0 = -1{,}87 + 0{,}416x_1 + 50{,}3x_2$$

Si hubiésemos calculado x_1 usando la fórmula anterior, así como los resultados de las relaciones bivariadas que hemos ido obteniendo, los cálculos serían:

$$b_{01.2} = \frac{0{,}656 - (97{,}9)(0{,}0048)}{1 - (114{,}0)(0{,}0048)} = \frac{0{,}1861}{0{,}4528} = 0{,}411$$

El error de redondeo es responsable de las diferencias entre esa cifra y el coeficiente 0,416 obtenido por el ordenador.

La ecuación nos dice que el valor del tamaño del yacimiento se podrá predecir mejor si se asume que cuando la extensión de tierra cultivable disponible y su productividad son iguales a cero, el tamaño del yacimiento es −1,87 ha, y que éste aumenta 0,416 ha por cada km² de aumento en la extensión de tierra cultivable y en 50,3 ha por cada unidad de incremento en el índice de productividad de las mismas.

Supongamos, sin embargo, que pretendemos comparar los coeficientes de pendiente para conseguir una medida de la cantidad de aumento/disminución en la dependiente asociada con el aumento de una unidad en cada variable independiente, en términos que sean comparables de una variable a la siguiente. Esto causará, probablemente, algunos problemas, porque las variables independientes estarán definidas, casi con toda probabilidad, en escalas diferentes. En nuestro ejemplo, la primera variable independiente está medida en kilómetros cuadrados, y la segunda en una escala arbitraria, con un rango mucho menor. En estas circunstancias no tiene sentido comparar el cambio de una unidad con el de la siguiente.

Si nos interesaran las proporciones relativas de cambio, lo que se podría hacer es transformar los coeficientes *b* en *coeficientes β* o *pesos β*, coeficientes de regresión parcial estandarizados. Para hacerlo, estandarizamos cada variable, dividiendo sus valores por la desviación típica, en otras palabras, obtenemos su puntuación *Z* y a partir de esas puntuaciones obtenemos pendientes ajustadas comparables de una variable a la siguiente. En símbolos matemáticos, calculamos:

$$Z_{x_i} = \frac{(x_i - \bar{x}_i)}{s_{x_i}}$$

Naturalmente, esto nos va a proporcionar una variable con la media igual a cero y la desviación típica igual a 1. Transformamos a continuación la ecuación de regresión, que para el caso de dos variables independientes será:

$$\hat{Z}_{x_{0.12}} = \beta_{01.2} Z_{x_1} = \beta_{02.1} Z_{x_2}$$

Como tratamos con puntuaciones *Z*, la media de cada variable es cero, y por tanto, el coeficiente *a* será también cero y puede eliminarse de la ecuación.

Nuestros coeficientes de regresión parciales estandarizados, o coeficientes β, indican de este modo los cambios relativos en las variables estandarizadas.

Suelen obtenerse estandarizando los coeficientes *b* por la proporción de desviación típica de las dos variables:

$$\beta_{01.2} = b_{01.2} \frac{s_{x_1}}{s_{x_0}}$$

o, en general,

$$\beta_{ij.k} = b_{ij.k} \frac{s_j}{s_i}$$

Esta fórmula nos explica la cantidad de cambio en la variable dependiente producida por un cambio (en este caso *estandarizado*) en una de las variables independientes, mientras que las restantes están controladas. Usar los coeficientes *b* ordinarios o los pesos β depende de nuestro interés por cambios absolutos o relativos.

Finalizaremos esta sección calculando la ecuación de regresión múltiple usando los coeficientes β (no proporcionados por MINITAB, pero sí por muchos otros programas).

$$\beta_{01.2} = \frac{0{,}864 - (0{,}832)(0{,}738)}{1 - 0{,}545} = 0{,}549$$

$$\beta_{02.1} = \frac{0{,}832 - (0{,}864)(0{,}738)}{1 - 0{,}545} = 0{,}427$$

Con lo que obtenemos la ecuación:

$$\hat{Z}_{x_{0.12}} = 0{,}549 Z_{x_1} + 0{,}427 Z_{x_2}$$

Es decir, de acuerdo con la ecuación, para cada incremento en una desviación típica de la extensión de tierra cultivada se produce un aumento de 0,549 desviaciones típicas en el tamaño del yacimiento, y para cada aumento de una desviación típica en la productividad se registra un aumento de 0,427 desviaciones típicas en el tamaño del yacimiento. La suma de esos dos efectos nos da la mejor predicción posible del tamaño del yacimiento, medido en términos de desviación típica con respecto a su media, y basándose en el criterio de mínimos cuadrados. La comparación de los coeficientes de cada una de las dos variables independientes indica que un incremento dado en la extensión de tierra cultivable tiene un efecto mayor sobre el tamaño del yacimiento que un incremento dado de la productividad de la tierra, comparación ahora válida porque las es-

calas de ambas variables han sido convertidas en las mismas unidades —unidades de desviación típica con respecto a la media de su distribución.

Un interesante ejemplo de análisis que utilice estas técnicas de regresión múltiple es la modelización de los ingresos en las haciendas inglesas medievales registradas en el *Domesday Book*, en términos de los recursos disponibles en esas haciendas, para los cuales también se poseen datos (McDonald y Snooks, 1985).

Interpretación del listado informático de un programa de regresión múltiple

En las secciones precedentes de este capítulo se ha insistido en que la regresión múltiple suele calcularse con ayuda de un ordenador, y que prácticamente todos los resultados para el ejemplo que seguimos se obtuvieron de esa forma; por eso se han presentado tan pocos cálculos. Hoy día disponemos de muchos programas capaces de realizar regresiones múltiples, todos ellos ligeramente distintos en lo que se refiere a su listado final. El objetivo de esta sección es mostrar un ejemplo concreto de listado de uno de esos programas: hemos elegido el listado que produjo MINITAB para el ejemplo del tamaño de los yacimientos. Ese listado aparece en la tabla 11.2. Una gran parte está relacionada con cuestiones de inferencia estadística, las cuales no siempre son relevantes para la mayoría de los usos del análisis de regresión en arqueología.

La ecuación de regresión (1) ya la hemos visto y no requiere más comentarios. La siguiente sección (2) repite los valores de los coeficientes de regresión, pero dando más información sobre ellos. En primer lugar, la desviación típica de todos los coeficientes; lo cual nos permite construir los intervalos de confianza para los valores de las estimaciones en la población de la cual se extrajo la muestra usada en este caso en particular, si es que eso nos fuese útil (para el procedimiento, véanse Blalock, 1972, cap. 18; Ryan *et al.*, 1985, pp. 161-162).

Igualmente, la columna siguiente PROPORCIÓN-T = COEF./DESV. TÍP., junto a la cantidad de grados de libertad, dos líneas más abajo, nos proporciona información para descubrir si los diferentes coeficientes son significativamente distintos de cero (véanse de nuevo Blalock, 1972, capítulo 18; Ryan *et al.*, 1985, pp. 161-162).

La línea siguiente (3), la desviación típica de y (o x_0) alrededor de la recta de regresión, se refiere a la variación de los valores de y en torno a la recta (véase el capítulo 9). En este caso, la desviación típica de esta distribución de puntos alrededor de la línea es 8,107. Asumiendo una distribución normal de los residuales en torno a la regresión (véase el capítulo 10) nos dice que aproximadamente un 68 % se encuentra dentro de 8,107 ha a ambos lados del valor predicho.

Cualquier estimación de la cantidad de una población en la regresión no debe basarse en el tamaño de la muestra, sino en el número de grados de libertad asociados con ella (4), si no, la estimación estará deformada. Se pierde un

TABLA 11.2. Traducción del listado de MINITAB para la regresión de los tamaños de los yacimientos sobre la extensión de tierra cultivable y la productividad. Basado en los datos de la tabla 11.1. Las cifras entre paréntesis remiten a las explicaciones en el texto.

LA ECUACIÓN DE REGRESIÓN ES:

Y = −1,87 + 0,416 X1 + 50,3 X2 (1)

	Columna	Coeficiente	Desv. típ. del coef.	Proporción-T coef/Desv. típ.	(2)
		−1,87	11,10	−0,17	
X1	C2	0,41596	0,09286	4,48	
X2	C3	50,33	14,39	3,50	

LA DESVIACIÓN TÍPICA DE Y ALREDEDOR DE LA RECTA DE REGRESIÓN ES:

s = 8,107 (3)

CON (28-3) = 25 GRADOS DE LIBERTAD (4)

R-CUADRADO = 83,0 por ciento (5)

R-CUADRADO = 81,6 por ciento, ajustada para los grados de libertad (6)

ANÁLISIS DE VARIANZA

Debido a:	G.L.	Suma de cuadrados	MS = S.C./G.L.	(7)
REGRESIÓN	2	7.998,80	3.999,40	
RESIDUAL	25	1.643,21	65,73	
TOTAL	27	9.642,01		

DESARROLLO DEL ANÁLISIS DE VARIANZA

Suma de cuadrados explicada por cada variable, cuando han sido entradas en el orden dado

Debido a:	G.L.	Suma de cuadrados	(8)
REGRESIÓN	2	7.998,80	
C2	1	7.194,53	
C3	1	804,27	

FILA	X1 C2	Y C1	Valor Y predicho	Desv. típ. pred. Y	Res.	Res. est.	(9)
21	95,9	76,00	92,87	3,94	−16,87	−2,38 RX	
23	51,0	80,50	81,24	3,90	−0,74	−0,10 X	
25	72,5	88,00	93,21	3,79	−5,21	−0,73 X	
27	89,9	95,30	85,85	4,09	9,45	1,35 X	

R denota una observación con un residual estándar grande
X denota una observación cuyo valor X influye mucho

grado de libertad en la estimación de la media de los valores y, y otro para cada variable independiente usada como predictor del valor y. En nuestro ejemplo del tamaño de yacimientos, había 28 observaciones y se pierden tres grados de libertad: uno por cada variable independiente, y otro para estimar la media.

El valor R^2 (5) ya ha sido tratado ampliamente; la versión ajustada (6) tiene en cuenta el hecho de que ese valor ha sido obtenido de una muestra de un cierto tamaño, y que se han perdido algunos grados de libertad durante el análisis, tal y como se ha dicho en el párrafo anterior.

La sección siguiente (7), titulada análisis de varianza, proporciona la información necesaria para calcular la prueba F de significación de la regresión (Blalock, 1972, capítulo 18). Los grados de libertad ya han sido tratados. El significado de las cantidades que aparecen en la columna SUMA DE CUADRADOS ya han sido adelantadas en el capítulo anterior: la suma *total* de cuadrados es la suma de las desviaciones cuadradas alrededor del valor medio de la variable dependiente: $\sum(y_i - \bar{y})^2$. En este caso adopta un valor de 9.642,01 alrededor de la media del tamaño de los yacimientos. La suma de cuadrados de la *regresión* es la cantidad de ese total explicada por la regresión global, usando ambas variables independientes —en este caso es 7.998,8—. El valor R^2 se obtiene, evidentemente, dividiendo esta cantidad por la suma de cuadrados total. La suma de cuadrados *residual* es la suma de los residuales al cuadrado alrededor de la recta de regresión (o plano, en este caso): $\sum(y_i - \hat{y}_i)^2$. Se trata de la variación no explicada por la regresión.

La columna final MS = S.C./G.L. se explica por sí misma y proporciona las dos cantidades cuya proporción es la prueba F.

La sección siguiente (8), desarrollo del análisis de varianza, proporciona una descomposición de la varianza explicada por la regresión, en términos de la cantidad explicada por cada una de las variables independientes. La línea superior repite la suma de cuadrados de la regresión total; la línea siguiente nos dice que C2 está asociada a una suma de 7.194,53. C2 es la columna en la memoria de trabajo de MINITAB que contiene la variable x_1, en este caso la extensión de la tierra cultivable. Tal como vimos anteriormente, la primera variable de la regresión explica toda la variación que puede, lo que incluye no sólo su propio efecto, sino también cualquier efecto conjunto con el resto de las variables que pueda existir. Se trata de la suma de cuadrados usada para calcular r^2_{01}.

La suma de cuadrados asociada con C3 es la explicada por la variable x_2, la productividad de la tierra, por sí sola, después de haber eliminado todos los efectos conjuntos de x_1 y x_2. En otras palabras, el valor de 804,27 es la suma de cuadrados a partir de la cual se obtiene $r^2_{02.1}$.

La última sección del listado (9) proporciona información acerca de los datos que son, en alguna medida, inusuales, ya sea porque tienen residuales muy grandes, o porque sus valores x provocan una influencia mayor en la pendiente de la recta de regresión. Vimos en la discusión sobre regresión robusta, al final

del capítulo anterior, que resulta posible a un pequeño número de puntos con grandes valores tener un impacto mayor sobre la pendiente de la regresión, de manera que si los eliminamos, la pendiente de la recta cambiará considerablemente (véase fig. 10.19). Es inevitable que los puntos con valores y muy alejados de la media de y produzcan este efecto, pues la regresión se basa en minimizar las diferencias *al cuadrado*. Por eso, al interpretar una regresión es importante apreciar qué confianza puede darse a la validez de unas observaciones desproporcionadamente escasas. Es posible obtener esas informaciones para las observaciones extremas, así como para el resto de la distribución, de ser necesario; sólo se precisa saber la instrucción adecuada en MINITAB.

La información que aquí aparece nos proporciona el valor x_1 de cada observación, su valor actual y, el valor y predicho de acuerdo con la regresión, la diferencia entre el valor observado y el predicho (RESIDUAL) y el valor del residual expresado como cantidad de desviaciones típicas de la distribución de residuales (RES. EST.).

La desviación típica de los valores predichos y (DESV. TÍP. PRED. Y) requiere algo más de comentario, pues aún no la habíamos considerado; está asociada, una vez más, con la inferencia estadística a partir de la regresión. El valor y predicho para una x en particular es el valor medio de las y en ese punto, para una muestra en particular, puesto que los requisitos de la regresión mencionados en el capítulo anterior se han cumplido. Es también la mejor de las estimaciones simples (o estimación puntual) de la media de los valores y para ese valor x en la población. Si, por otro lado, queremos especificar un intervalo dentro del cual se encuentra la media de la población y, con una cierta probabilidad (un intervalo de estimación), necesitaremos saber la desviación típica, o el error típico, de la predicción, de forma que calculemos el intervalo sobre la base del supuesto de que la distribución muestral de valores y predichos sea normal (véase Blalock, 1972, pp. 404-405).

PRESUPUESTOS

En la sección anterior se ha mencionado de paso la cuestión de los presupuestos de la regresión; será conveniente acabar este capítulo haciendo algunos comentarios referidos al caso concreto de la regresión múltiple.

El primer punto en el que hay que insistir es que todos los presupuestos señalados en el capítulo anterior para la regresión simple también han de mantenerse en la regresión múltiple, con la consecuencia adicional de que, en esta última, deben mantenerse en la relación entre la variable dependiente y cada una de las independientes. Además, las variables independientes no deben estar correlacionadas entre sí.

De lo anterior se deduce que, cuando hay muchas variables independientes, resulta bastante complicado saber si los presupuestos del análisis de regresión

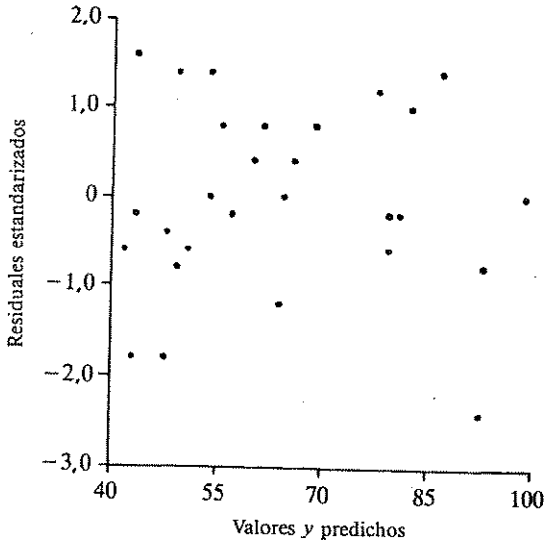


FIGURA 11.7. Diagrama de dispersión de los residuales estandarizados de la regresión múltiple del tamaño del yacimiento sobre la superficie disponible de tierra y la productividad de la misma, representado con relación a los valores del tamaño de los yacimientos predichos por la regresión múltiple, según los datos de la tabla 11.1.

se mantienen. Además, es casi seguro que, en algunas de las relaciones, los presupuestos no se habrán mantenido en su integridad. La pregunta que debemos plantearnos es: ¿hasta qué punto podemos ignorar los presupuestos antes de que el análisis pierda sentido? No hay una respuesta sencilla.

El capítulo anterior sirvió de guía para la detección de problemas y para su solución, y lo que se dijo entonces se aplica también a la regresión múltiple. Es conveniente reiterar, sin embargo, que nunca es satisfactorio limitarse a guardar datos en un ordenador y llevar a cabo un análisis de regresión múltiple, relacionado con cierto problema. El primer paso debe ser estudiar si las distribuciones individuales son aproximadamente normales; a continuación, estudiar los diagramas de dispersión de las diversas relaciones bivariadas para ver si son lineales. Finalmente, los residuales de la regresión múltiple han de ser examinados.

Reiterando de nuevo lo dicho en el capítulo anterior, el que los presupuestos no se hayan mantenido no obliga a prescindir del análisis de regresión, sino que hay que emprender acciones apropiadas. Muy a menudo, esto dará lugar a complejidades que requerirán la ayuda de un estadístico profesional.

En cuanto al ejemplo que hemos seguido en este capítulo, ya se señaló que existía algo de colinealidad, que conduce a la ambigüedad en la afirmación de

los efectos separados de la extensión de la tierra cultivable y la productividad de la misma; pero no llega al grado en que cause los problemas ilustrados por Johnston (1978, pp. 74-77). Las variables del ejemplo estaban construidas de forma que se aproximaban bastante a la distribución normal, mientras que los diagramas de dispersión (figs. 11.4-11.6) indican que las diferentes relaciones bivariadas implicadas son aproximadamente lineales. Finalmente, el diagrama de dispersión de los residuales estandarizados por los valores *y* predichos por la regresión múltiple (fig. 11.7) no señala la existencia de una regularidad que indique que se han violado los presupuestos.

EJERCICIOS

11.1. En un estudio arqueológico de los factores que afectan a la densidad de hallazgos de obsidiana en una serie de grandes yacimientos clásicos en Mesoamérica, se adelanta la hipótesis según la cual la distancia a la fuente y el tamaño del yacimiento (reflejo de su importancia funcional) son las variables más relevantes. Dada la información que aparece en la tabla (datos según Sidrys, 1977), ¿es eso cierto? Usa métodos de regresión y de correlación múltiples y parciales para ayudarte en la conclusión.

Densidad de hallazgos de obsidiana (g/m^3)	Distancia a la fuente (km)	Tamaño del asentamiento (ha)
38	70	32
32	105	16
35	110	24
23	110	14
18	145	33
23	160	30
27	150	29
30	165	40
14	195	65
22	205	44
16	240	37
21	260	48
7	280	59

11.2. En un estudio de los restos de fauna en varias cuevas del pleistoceno se ha decidido estudiar las relaciones entre la cantidad de fragmentos de hue-

sos de lobo y la de huesos de bóvido (datos según Boyle, 1983). ¿Hay alguna relación entre los dos, teniendo en cuenta que el tamaño de cada conjunto es una variable importante?

Conjuntos	Fragmentos de bóvido	Fragmentos de lobo	N.º total de fragmentos en cada conjunto
1	31	1	1.211
2	0	111	618
3	1.622	278	4.260
4	150	63	820
5	13	48	137
6	12	161	2.916
7	0	24	249
8	33	0	128
9	58	0	505
10	107	18	998

11.3. En un estudio de la relación entre la economía de las *villae* romanas y las ciudades adyacentes se ha registrado la proporción de ganado en el registro faunístico de varias *villae*, así como su distancia a la ciudad más cercana y las fechas en las que fueron ocupadas. Las correlaciones son las siguientes:

Distancia de la ciudad-proporción de ganado vacuno, $r = 0{,}72$
Fecha-proporción de ganado vacuno, $r = 0{,}55$
Fecha-distancia a la ciudad, $r = 0{,}60$

Argumenta la relación entre esas variables.

12. CLASIFICACIÓN NUMÉRICA EN ARQUEOLOGÍA

INTRODUCCIÓN HISTÓRICA

Desde los orígenes de la disciplina, las clasificaciones han desempeñado un papel esencial en arqueología. En Europa, por ejemplo, durante el siglo XIX, desde el sistema de las tres edades de Thomsen a Montelius, una gran parte de la investigación se basaba en la agrupación y ordenación del material arqueológico, de forma que esa ordenación tuviera un significado arqueológico. Con el desarrollo del concepto de cultura arqueológica a principios del siglo XX, el estudio de la dimensión espacial se hizo importante: un aspecto clave de la investigación arqueológica fue la definición de unidades temporales y espaciales coherentes, una tarea que sigue siendo importante.

El fundamento para definir esas unidades era una afirmación de las similaridades y diferencias en el material arqueológico. Esta afirmación tiene dos aspectos. Por un lado, la investigación de las agrupaciones y discontinuidades en clases particulares de material, como la cerámica; por el otro, una vez obtenidas las agrupaciones en clases individuales de material, la cuestión de la asociación de materiales entre agrupaciones de material de diferentes clases; por ejemplo: ¿se han encontrado tipos particulares de cerámica junto a tipos particulares de artefactos líticos en el contexto de casas similares o formas de enterramiento similares?

Para decidir cómo agrupar el material arqueológico de cierta manera y no de otra, el arqueólogo se basaba en su experiencia personal. Eran sus conocimientos personales lo que le permitía tratar los materiales arqueológicos a los que tenía acceso, así como aquellos catalogados en grandes *corpora*. Viajar era importante porque permitía una apreciación de primera mano del material en museos extranjeros, y proporcionaba una mayor base de conocimientos para el estudio y la afirmación de semejanzas y diferencias. Entre sus otros muchos méritos, Childe fue uno de los clásicos exponentes de este enfoque entre los prehistoriadores europeos, enfoque que aún tiene gran importancia en la práctica arqueológica cotidiana.

Fue en la década de los cincuenta cuando la situación empezó a cambiar, gracias a las primeras aplicaciones de las técnicas cuantitativas al problema de definir y ordenar las similaridades y diferencias en los materiales arqueológicos. Uno de esos nuevos enfoques fue el de Spaulding (1953), que propuso que los tipos de artefactos se definieran según los esquemas de asociación entre las diferentes variables o atributos que describen los artefactos en cuestión, y que esas asociaciones podían estudiarse por medio de la prueba de χ^2 (véase el capítulo 6).

El otro problema que condujo al uso de los métodos cuantitativos no fue la definición de tipos, sino la ordenación de los artefactos en una secuencia cronológica. De hecho, un problema semejante ya había sido tratado cuantitativamente en un famoso estudio por el egiptólogo sir Flinders Petrie (1901), cuando investigaba los enterramientos del período predinástico en Egipto. Petrie disponía de información acerca del ajuar funerario hallado en una gran cantidad de tumbas, y lo que le interesaba era descubrir el orden cronológico en el que se construyeron esas tumbas, partiendo del supuesto de que uno de los factores más importantes que afectaba la deposición de objetos de ajuar en una tumba en particular era la variación en los tipos de objetos que estaban de moda en una época determinada. Tal y como Kemp (1982) lo describe, la conclusión de Petrie fue que la ordenación de tumbas que más se aproximaría a su secuencia cronológica sería aquella en la que la duración de los tipos individuales fuese lo más breve posible; la idea subyacente es que un tipo de objeto se pone de moda en un momento dado, experimenta por consiguiente un período de popularidad en aumento, estando su uso muy difundido durante un cierto tiempo, y finalmente su popularidad declina, terminando por desaparecer. La solución práctica de Petrie al problema implicaba poner por escrito el contenido de cada tumba en una ficha de cartón, disponiendo todas las fichas en línea, y a continuación las ordenaba para conseguir que todas las ocurrencias de un mismo tipo estuviesen agrupadas. Ciertamente, no es un método muy simple, pues la duración de los distintos tipos se superpone, de manera que la agrupación de las ocurrencias de un mismo tipo puede producir el efecto de dispersar las ocurrencias de otro; sin embargo, Petrie consiguió una ordenación que le satisfizo. Kemp (1982) describe un análisis del material funerario del Egipto predinástico usando técnicas modernas que confirma en gran parte el resultado de Petrie.

A principios de los años cincuenta, Brainerd y Robinson (Brainerd, 1951; Robinson, 1951) decidieron ordenar conjuntos de cerámicas. Partieron del mismo supuesto que Petrie acerca de la forma en que los tipos de objetos están de moda por un tiempo y pierden finalmente popularidad, pero en vez de trabajar con listas en fichas de cartón, ellos basaron sus conjuntos en la comparación de las proporciones relativas de los distintos tipos de cerámica en cada par de unidades; calcularon una media de similaridad entre cada unidad y las demás (véase p. 211). Produjeron una tabla o matriz con esas similaridades y dispusieron el orden de las unidades en la tabla con el propósito de agrupar las

mayores similaridades de todas y producir así una secuencia (véase Doran y Hodson, 1975, pp. 272-274, para una descripción más detallada).

Trataremos más adelante del procedimiento de seriación (pp. 212-215); por el momento señalemos que, desde un punto de vista histórico, hubo un período desde los años cincuenta hasta mediados de los años sesenta en el que los estudios de seriación fueron los más importantes en la investigación cuantitativa en arqueología. Fue entre principios y mediados de los años sesenta cuando empezaron a desarrollarse los enfoques basados en el estudio de la similaridad y las relaciones entre variables. Este enfoque estaba inspirado en la biología, específicamente en la taxonomía biológica.

La clasificación tradicional de plantas y animales en biología no estaba basada en un gran número de las características de los elementos estudiados. Como en arqueología, el taxonomista, gracias a su experiencia personal, seleccionaba un pequeño número de atributos clave que parecían variar significativamente entre los elementos, basando su clasificación en ellos. Algunos advirtieron que obtendrían clasificaciones mucho más satisfactorias, específicamente clasificaciones con mayor significado filogenético, si se usaban muchos de los atributos de las plantas y animales estudiados. Se dieron cuenta, también, que la clasificación sería más satisfactoria si no se asignase *apriorísticamente* ninguna importancia diferenciadora a algunas de las características en particular de los elementos que se pretendía clasificar. El resultado, según los partidarios de este enfoque, serían grupos «naturales».

Dado que es imposible para cualquier taxonomista considerar simultáneamente un gran número de características en un gran número de elementos y sopesarlas todas por igual, se hizo necesario algún procedimiento de automatización, de ahí el auge de la *taxonomía numérica* (Sokal y Sneath, 1963; Sneath y Sokal, 1973).

La taxonomía numérica fue introducida en arqueología por David Clarke (1962; 1970) en el llamado «análisis matricial» de los vasos campaniformes británicos. Clarke pretendía generar una clasificación «natural» de sus vasos campaniformes, usando todos sus atributos y sin dar mayor preeminencia a algunos de ellos; consideraba que esos grupos «naturales» estarían relacionados con las tradiciones sociales del grupo humano (Clarke, 1966), aunque, de hecho, los métodos que usó no eran más que una adaptación de la seriación de Robinson-Brainerd.

A partir de entonces, la dudosa base teórica de los grupos «naturales» ha sido arrojada por la borda y sustituida por un enfoque mucho más sostenible que enfatiza el propósito de una clasificación; así y todo, los métodos de la taxonomía numérica han ido adquiriendo un papel importante en las clasificaciones arqueológicas, a pesar de los innegables problemas que plantean (Doran y Hodson, 1975; Whallon y Brown, 1982).

Los detalles técnicos de los diversos procedimientos se describirán en las secciones siguientes de este capítulo, así como en el próximo, pero no estaría de

más comentar brevemente ciertos aspectos de su uso en arqueología durante las dos últimas décadas, un período en el que se han producido pocos cambios en la forma de utilizar los métodos, si bien se han introducido ciertas mejoras técnicas.

Los métodos de la taxonomía numérica se han usado para agrupar elementos según los valores de las variables o los estados de los atributos que los caracterizan, y no para agrupar variables o atributos a partir de sus esquemas de asociación. Esto se debe a que es el procedimiento anterior el que se considera que corresponde al tradicional enfoque intuitivo para definir tipos de artefactos en arqueología —la agrupación de ciertos elementos físicos, o sus descripciones—. El resultado es que la clasificación numérica no deja de ser arqueología tradicional realizada con ayuda de un ordenador, y que las clasificaciones producidas se evalúan según su aproximación a las clasificaciones tradicionales propuestas por los «expertos». Esto difícilmente parece una justificación para la elaborada metodología que emplea y que es una de las razones por las que los estudiantes (y no sólo ellos) suelen considerar esta área como particularmente aburrida y sin interés.

Spaulding (1977) ha adoptado el punto de vista según el cual este enfoque para la definición de los tipos es erróneo, arguyendo que los tipos se definen en términos de asociaciones de *atributos* estadísticamente significativas; otros autores proponen, sin embargo, la agrupación de *elementos*, y por tanto critican la validez del enfoque de Spaulding (véase Doran y Hodson, 1975). La más reciente expresión de esta polémica puede consultarse en el libro *Essays in Archaeological Typology* (Whallon y Brown, 1982, capítulos 1, 2, 3 y 6), si bien los nuevos desarrollos metodológicos (capítulo 13) dan a entender que la polémica ya no tiene la relevancia que tuvo hace unos años.

Queda por considerar por qué la taxonomía numérica se ha popularizado tanto, incluso entre arqueólogos tradicionalistas que han sido escépticos ante muchos de los avances en el método y la teoría arqueológicos de las dos últimas décadas. Hay varias respuestas posibles a esta cuestión; algunas de ellas ya han sido señaladas, pero creo que lo más importante es que se considera que esos métodos son, en cierto sentido, «objetivos». Es una idea muy arriesgada, y que requiere algunas aclaraciones.

Tal y como veremos en las secciones siguientes de este capítulo, el uso de la taxonomía numérica implica la definición de una medida de similaridad entre los elementos o las variables que queremos agrupar a partir de sus similitudes. Esas medidas y procedimientos tienen distintas propiedades que producen resultados distintos. Son «objetivos» en el sentido en que una vez efectuada la selección pueden ser realizadas automáticamente por un ordenador. Sin embargo, la elección debe basarse en los datos en particular y en el problema investigado: no han de ser arbitrarios, ni considerados objetivos, sino que hay que justificarlos arqueológica y metodológicamente.

Antes que definamos la similaridad, sin embargo, nótese que ya el mero

hecho de poder definir esa medida presupone la existencia de una descripción de los objetos de interés a partir de los cuales puede derivarse la medida. Como ya hemos dicho, la idea según la cual el fundamento de la taxonomía radica en la necesidad de describir cualquier aspecto de los artefactos que pudiera llegar a interesarnos ha sido abandonada. Gardin (1980) ha insistido en la naturaleza problemática de la descripción arqueológica. Describimos siempre con un propósito en mente, implícito o explícito; es mucho mejor que sea explícito, de forma que otorguemos un valor activo a las variables descriptivas elegidas, y la forma en que las construimos en relación con el propósito que tenemos, ya sea la definición de la variabilidad espacial, cronológica o cualquier otro (véanse, por ejemplo, Gardin, 1980; Whallon, 1982; Vierra, 1982). Las decisiones tomadas en este punto determinarán los resultados del análisis.

Esta discusión no significa que en nuestra descripción de los datos y el uso de los métodos analíticos *impongamos* siempre un orden en el mundo. La estructura y el orden en los datos pueden existir o no; su existencia y forma, si es que las hay, han de ser *descubiertas*. Cualquier estructura dependerá de nuestra descripción; nuestros métodos de análisis pueden ocultarla, distorsionarla o revelarla, pero la mera posibilidad de esas alternativas indica la realidad contingente de su existencia.

Nos hemos apartado un tanto del problema de la «objetividad» como justificación del uso de la clasificación numérica, que es el problema más importante en realidad. Si la objetividad es una quimera y las clasificaciones numéricas se justifican por su identidad con las tipologías tradicionales, ¿qué razón podremos aducir para usar tales métodos? En primer lugar, su uso nos ayuda a hacer más explícitas las bases de las decisiones en materia de clasificación. En segundo lugar, dada la frecuente necesidad de buscar un orden en grandes conjuntos de elementos descritos por muchas variables (o bien buscar un orden en grandes conjuntos de variables a partir de su aparición en conjuntos muy numerosos de elementos), el empleo de esas técnicas permite que el procedimiento de agrupación sea consistente (no «objetivo»), y pueda llegar a revelar esquemas presentes en el material que de otro modo no emergerían de la complejidad de los datos originales. Los métodos que se describen a continuación desempeñan un papel indispensable en este último punto.

CLASIFICACIÓN NUMÉRICA: ALGUNAS DEFINICIONES PRELIMINARES

En este capítulo se tratarán los detalles técnicos de algunos aspectos de la clasificación numérica; las cuestiones más específicamente arqueológicas que plantea su uso sólo serán abordadas de paso; puede encontrarse un mejor tratamiento en Doran y Hodson (1975), y en Whallon y Brown (1982), y las referencias allí citadas. Antes de seguir adelante, sin embargo, es necesario aclarar ciertos términos.

La clasificación trata, básicamente, de la identificación de grupos de objetos similares en el conjunto de objetos investigados (los «objetos» pueden ser elementos o variables). Puede considerarse como un procedimiento de simplificación, de forma que puedan plantearse a partir de las similaridades y diferencias dentro de cada grupo y entre los distintos grupos. Esas generalizaciones pueden ser puramente descriptivas, o bien formar la base de hipótesis que habrán de ser contrastadas por otros medios; en este sentido, la clasificación constituye una extensión del análisis de datos exploratorio, presentado en capítulos anteriores.

Es conveniente distinguir entre lo que es *una clasificación propiamente dicha* y ciertos procedimientos similares. En general, una clasificación define unos grupos en un conjunto de datos, basándose en el principio según el cual los miembros de un grupo han de ser más similares entre sí que los no miembros; la similaridad en el interior del grupo es, en cierto sentido, mayor que la existente entre grupos; alternativamente, los grupos (o conglomerados,* como suelen denominarse) deben mostrar cohesión interna y aislamiento externo (Cormack, 1971). El propósito en los estudios de clasificación es, por lo general, descubrir el esquema de las agrupaciones en un conjunto de datos con el menor número posible de requisitos acerca de la naturaleza de las agrupaciones (véase Gordon, 1981, p. 5). El procedimiento suele denominarse *análisis de conglomerados*.**

Puede compararse a un procedimiento de *discriminación* que presuponga la existencia de cierto número de grupos e intente colocar los individuos en los grupos a los que más probablemente pertenezcan. Esos procedimientos se usan, por ejemplo, para situar un nuevo hallazgo en la categoría más apropiada de una clasificación preexistente. Alternativamente sirven para investigar la manera en que la categorización se relaciona con otro conjunto de variables. Por ejemplo, si tenemos un cierto número de vasijas de cerámica procedentes de distintos yacimientos, y esas vasijas se caracterizan por las medidas que describen su forma, entonces ¿acaso difieren las formas en los distintos yacimientos? El problema es el siguiente: dada la división de cerámicas entre yacimientos, ¿se reproduce esa misma división cuando dividimos las vasijas por las variables que definen su forma? (véanse Shennan, 1977; Read, 1982). Una respuesta a esta pregunta implica discriminación y no clasificación.

Otro procedimiento es el de *disección*. En algunos casos podemos saber si nuestros datos no son divisibles en grupos que muestren cohesión interna y aislamiento externo: constituye una nube de puntos continua en la que ninguna división natural es posible. Ahora bien, si para algún propósito en particular hemos de dividirla imperativamente, denominaremos disección a esa división

* El término inglés para conglomerado (*cluster*) es, ciertamente, el que domina, tanto en la bibliografía francesa, como en la alemana, italiana y castellana. El lector ha de tener en cuenta la popularidad del término en inglés cuando consulte otros libros o artículos. (*N. del t.*)

** También está muy difundido en castellano el término original inglés *cluster analysis*, o bien la expresión mixta *análisis de cluster*. (*N. del t.*)

más o menos arbitraria (véase Gordon, 1981, p. 5). La disección no es, realmente, muy importante, si bien algunos de los procedimientos que trataremos al hablar de la ordenación bien pudieran incluirse en este apartado.

Vimos en el capítulo 9 que, si tenemos información sobre una serie de elementos en términos de sus valores en dos variables, podemos representar las relaciones entre los elementos por medio de un diagrama de dispersión, cuyos ejes están definidos por las variables en cuestión. En el capítulo sobre la regresión múltiple vimos que esa representación era problemática con tres variables e imposible con más de tres. Es evidente, por tanto, que no podemos contemplar grupos de elementos similares si están descritos por medio de muchas variables: hay demasiadas dimensiones. El objetivo de los métodos de ordenación es representar y mostrar las relaciones entre los elementos en un espacio de pocas dimensiones —normalmente dos, todo lo más tres—, reteniendo la mayor cantidad posible de la información contenida en las variables descriptivas. Los puntos que representan los objetos estarán más agrupados si su similaridad mutua es alta, y separados si es baja. Un examen visual del diagrama de dispersión indicará si los grupos —definidos como las áreas de elevada densidad de puntos— están presentes. Como etapa posterior en dicha operación, pueden aplicarse los métodos de agrupación (clasificación propiamente dicha) a la distribución de puntos en el diagrama de dispersión.

Ya hemos distinguido la clasificación propiamente dicha de la discriminación, la disección y la ordenación. Dentro de la clasificación podemos hacer aún más distinciones, según las diferentes maneras en que se formen grupos.

Una categoría es la constituida por los métodos de *partición* (Gordon, 1981, pp. 9-10). El uso de los mismos implica tomar una decisión acerca del número de grupos que nos interesa, si bien, a diferencia de la discriminación, no se requiere ninguna especificación del tamaño de los diferentes grupos. Los individuos se agrupan junto a aquellos que, en cierto sentido bien especificado, son semejantes; así se forma la cantidad necesaria de grupos.

La otra categoría principal es la de los métodos *jerárquicos*, divididos a su vez en *aglomerativos* y *divisivos*. Los métodos aglomerativos jerárquicos empiezan con todos los elementos investigados separados, construyendo grupos a partir de ellos y empezando por el más semejante, continuando con los agrupados en niveles de similaridad progresivamente menores y acabando en un único grupo, con un nivel de similaridad interno muy bajo. Los métodos divisivos empiezan con todos los elementos en un único grupo y proceden a dividirlo sucesivamente, de acuerdo con ciertos criterios. En ambos tipos de métodos jerárquicos, las relaciones entre los elementos y los grupos puede representarse bajo la forma de un diagrama en árbol o *dendrograma*.

Todos estos métodos de análisis de conglomerados, pero quizás más los divisivos, imponen en cierta medida su propia estructuración de los datos, como veremos más adelante. Un método divisivo, por ejemplo, impondrá una serie de divisiones en un conjunto de datos, sin tener en cuenta si los grupos resul-

tantes constituyen una distinción auténtica o una disección arbitraria. Por este motivo, la validación de los resultados es muy importante, si bien se trata de algo que ha sido injustificadamente olvidado en arqueología (véase Aldenderfer, 1982). Este tema volverá a considerarse más adelante (pp. 229-234), si bien hay que señalar aquí que los resultados de un único análisis de conglomerados nunca han de aceptarse por sí solos; han de compararse entre sí los resultados de distintos métodos, y emplearse métodos de contrastación alternativos.

Medidas de distancia y similaridad

Ya se ha dicho que antes de usar cualquier método de análisis de conglomerados es necesario tener alguna medida que exprese las relaciones entre los individuos que hemos de analizar. Hemos hablado, en general, de afirmar las similaridades entre los elementos, pero también podemos sustituirlas por distancias, siendo estas últimas el opuesto de las primeras (véase Späth, 1980, pp. 15-16).

Tabla 12.1. Matriz de similaridades entre cuatro conjuntos hipotéticos de cerámica, usando el coeficiente de acuerdo de Robinson.

	1	2	3	4
1	200	14	11	9
2	14	200	147	163
3	11	147	200	157
4	9	163	157	200

Los métodos de clasificación numérica están basados en una matriz $n \times n$ de similaridades o distancias entre n objetos; por ello, el primer paso será calcular esa matriz. Un ejemplo, para las similaridades entre cuatro elementos (conjuntos cerámicos hipotéticos) aparece en la tabla 12.1. En la *diagonal principal* de la matriz aparecen las similaridades de cada elemento consigo mismo, evidentemente el valor máximo, en este caso 200. De hecho, no necesitamos de todas las cifras que aparecen en la matriz, porque sus dos mitades, por encima y por debajo de la diagonal principal, son el reflejo una de la otra. Así, s_{12}, la similaridad entre los elementos 1 y 2, es la misma que s_{21}, en este caso 14. Este tipo de matrices se denominan *simétricas* y sólo se necesita para el análisis una de sus dos mitades. La mayoría de las matrices de distancia o de similaridad son simétricas, aunque más adelante veremos un ejemplo de lo contrario.

Los coeficientes de similaridad o de distancia para construir una de esas matrices son muchos y muy variados (Sneath y Sokal, 1973; Whishart, 1978). Tienen distintas propiedades, siendo algunos de ellos apropiados para datos cuantitativos, y otros para datos cualitativos (presencia/ausencia); la elección

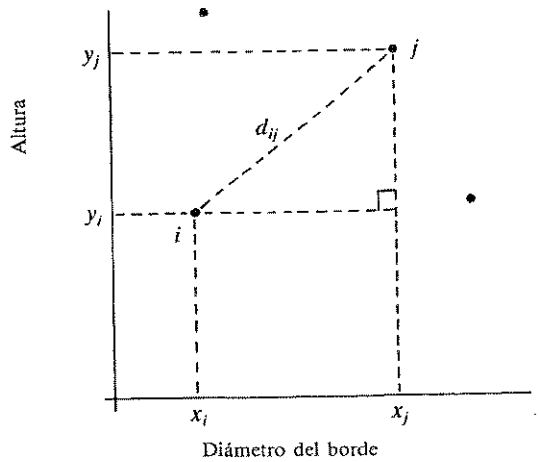


Figura 12.1. Diagrama de dispersión de la altura con relación al diámetro del borde para cuatro vasijas de cerámica, mostrando la definición de la distancia euclídea entre las vasijas i y j.

de los coeficientes, pues, no debe tomarse a la ligera. Aquí sólo nos será posible examinar unos pocos de los más importantes.

La medida más común para los datos en escala interválica o proporcional es la *distancia euclídea*. Dados dos individuos i y j, medidos en términos de una cantidad p de variables, la distancia euclídea (d_{ij}) se define como:

$$d_{ij} = \left(\sum_{k=1}^{p}(x_{ij} - x_{jk})^2\right)^{1/2}$$

Se trata, tan sólo, de la distancia entre dos puntos, tal y como se deduce del teorema de Pitágoras. Veamos un ejemplo sencillo en dos dimensiones: supongamos que hemos de medir la distancia en línea recta entre varias vasijas descritas por su altura y por el diámetro del borde (fig. 12.1). Medimos la distancia entre cada par de objetos i y j, sobre el eje x ($x_i - x_j$) y sobre el eje y (es decir, $y_i - y_j$), elevamos al cuadrado las dos distancias, las sumamos y extraemos la raíz cuadrada. En este caso,

$$d_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2]^{1/2}$$

(Recuérdese que elevar un número a la potencia de 1/2 es otra manera de simbolizar la raíz cuadrada). Efectuando estas operaciones para cada par de puntos obtendremos la matriz de las distancias entre ellos.

Si hay más de dos variables, hemos de añadir términos $(x_{ij} - x_{jk})^2$ extra,

de forma que haya uno para cada variable descriptiva, antes de extraerle la raíz cuadrada.

Cuando los dos puntos están en el mismo sitio —los elementos son iguales entre sí—, d_{ij} es igual a cero; el valor máximo opuesto a ese es infinito —los dos puntos están infinitamente alejados—, con lo que existe entre ambos una disimilaridad total.

El problema se plantea en lo que se refiere a la escala de los ejes. Especialmente cuando las medidas de los elementos investigados están en la misma escala, pero sus límites son muy diferentes. Por ejemplo, al clasificar espadas de bronce según su longitud, anchura y grosor. Evidentemente, la longitud tiene un rango de variación mucho mayor que el grosor, y por tanto afectará más a la clasificación. Si queremos contrarrestar esos efectos es necesario estandarizar las escalas de medida. En ese caso, lo normal es usar las puntuaciones estándar, con lo que todas las variables tendrán igual importancia.

Hay otro problema con la distancia euclídea, ya que presupone que los ejes del espacio definido por las variables son perpendiculares unos a otros, constituyendo un sistema de coordenadas rectangulares. Este hecho se hará mucho más claro en el próximo capítulo, si bien podemos decir ahora que los ejes serán perpendiculares (u *ortogonales*, si usamos el término habitual) cuando las variables sean totalmente independientes unas de otras, lo cual nunca sucede en la práctica. Si las variables están correlacionadas y los ejes no se sitúan en ángulo recto, las d_{ij} serán sobre o subestimadas en un grado que dependerá de la intensidad de la correlación y de si ésta es positiva o negativa.

La solución más habitual a este problema consiste en asegurarse de que los ejes son perpendiculares mediante la definición de la medida de distancia no en términos de las variables originales, sino en los de los ejes definidos por el análisis de componentes principales de las variables, las cuales son perpendiculares por definición, como veremos en el capítulo próximo; también existen métodos alternativos (véanse, por ejemplo, Johnston, 1978, pp. 217-219; Mather, 1976, pp. 313-314; Everitt, 1980, p. 57).

Aunque existen muchas otras medidas de distancia/similaridad, además de la distancia euclídea, y que permiten usar datos en escalas interválicas o proporcionales (véase Sneath y Sokal, 1973), sólo se mencionará aquí una más, basada en la suma de las diferencias absolutas entre los puntos, para cada variable. Así:

$$d_{ij} = \sum_{k=1}^{p} | x_{ij} - x_{jk} |$$

Nos dice que hemos de calcular la diferencia entre los puntos i y j según sus valores en cada variable, para tantas variables como haya en el análisis, sumándolas a continuación sin tener en cuenta si son positivas o negativas (el símbolo « | » afirma que debemos ignorar el signo de las diferencias). Esta distancia recibe el nombre de *métrica city-block*. Podemos ejemplificarla en un caso bi-

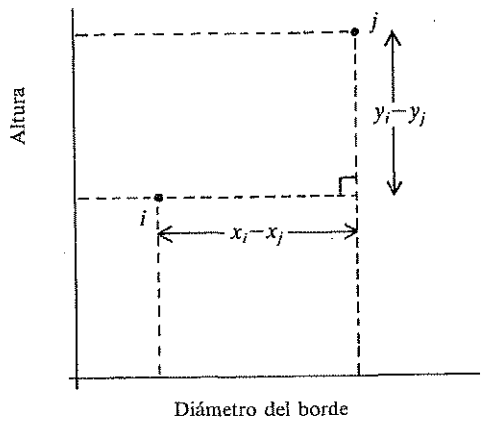


FIGURA 12.2. Diagrama de dispersión de la altura con relación al diámetro del borde para dos vasijas cerámicas (*i* y *j*), mostrando la definición de la distancia de ciudad (*city-block*) entre las dos.

dimensional (fig. 12.2); la fórmula requiere que en este caso calculemos la diferencia entre i y j en la primera variable ($| x_i - x_j |$) y la diferencia entre ellas en la segunda variable ($| y_i - y_j |$) y sumarlas. Esto nos da una medida de distancia compuesta por dos líneas rectas que definen una esquina.

Si nos fijamos ahora en las medidas de similaridad entre elementos utilizables en el caso de datos de presencia/ausencia (o dicotómicos), encontraremos una gran variedad de coeficientes, todos ellos ligeramente distintos. La mayor diferencia entre los mismos estriba en si incluyen o no las *comparaciones negativas*, es decir, la situación en la que ninguno de los individuos en consideración posee el atributo en cuestión. Esta situación puede ilustrarse mediante un ejemplo. Supongamos que tenemos dos tumbas medidas según presenten o no ciertos tipos de objetos en su ajuar (tabla 12.2).

TABLA 12.2. Dos tumbas medidas según la presencia/ausencia de diez tipos de objetos de ajuar distintos.

	Tipos de objetos									
	1	2	3	4	5	6	7	8	9	10
Tumba 1	1	0	1	1	0	0	0	1	1	0
Tumba 2	1	0	0	1	1	0	0	1	0	0

En este ejemplo, hemos asumido que hay 10 objetos diferentes presentes en el conjunto de datos, pero los tipos 2, 6, 7 y 10 no están presentes en ninguna de estas dos tumbas en particular. En un caso como este se podría considerar

que la ausencia de un tipo determinado no tiene la misma importancia que su presencia, y que, en particular, no desearíamos conceder a la ausencia conjunta de un tipo de objeto el mismo valor que la presencia conjunta. Ese sería, quizás, el caso si la ocurrencia de los tipos fuese muy infrecuente. Esas circunstancias, en las que cero y uno tienen una importancia diferente, pueden compararse con aquellas en las que, por ejemplo, codificamos el sexo de los individuos en las tumbas (0 = masculino, 1 = femenino): los dos valores tienen el mismo estatus, pues son etiquetas arbitrarias.

Existen muchos coeficientes apropiados para datos binarios (ausencia/presencia); aquí sólo describiremos dos de los más importantes. Ambos difieren en la forma de tratar las comparaciones negativas. Al principio del análisis debe estudiarse detenidamente cuál de los tratamientos es más conveniente en cada caso.

El coeficiente de comparación simple

Para cada par de elementos, se comparan sus puntuaciones en cada atributo y se registra si son las mismas o no. El número de comparaciones se expresa a continuación como una proporción de la cantidad de atributos. El procedimiento de cálculo puede ilustrarse convenientemente por medio de una tabla 2 × 2 (tabla 12.3). Para cada par de individuos contamos la cantidad de atributos presentes en ambos (a), la cantidad de atributos presentes en j, pero no en i (b), la cantidad de atributos presentes en i, pero no en j (c), y la cantidad de atributos ausentes en ambos (d). Usando los datos de la tabla 12.2, obtenemos la tabla 12.4.

TABLA 12.3. Tabla general para la comparación de dos elementos según la presencia/ausencia de una serie de atributos.

		Individuo i	
		Atributo +	Atributo −
Individuo j	Atributo +	a	b
	Atributo −	c	d

TABLA 12.4. Comparación de tumbas según los datos de la tabla 12.2, según la presencia/ausencia de los diez atributos.

		Tumba 1	
		+	−
Tumba 2	+	3	1
	−	2	4

Por consiguiente, el coeficiente de comparación simple se calcula:

$$S = \frac{a + d}{a + b + c + d}$$

En palabras: las comparaciones positivas más las comparaciones negativas, dividido por la cantidad total de atributos. En el ejemplo de las dos tumbas:

$$S = \frac{3 + 4}{3 + 1 + 2 + 4} = \frac{7}{10} = 0{,}7$$

El coeficiente de Jaccard

Se parte del principio opuesto al anterior, en lo que se refiere a las comparaciones negativas. El que dos elementos sean idénticos al no poseer algunos atributos, no se contabiliza ni como comparación ni en la cantidad total de atributos que constituye el denominador del coeficiente; dado cualquier par de elementos, el divisor es la cantidad de atributos presentes en uno o en otro de los elementos que configuran el par. En los términos de la tabla 2 × 2 (tabla 12.3):

$$S = \frac{a}{a + b + c}$$

Respecto a nuestro ejemplo, tenemos:

$$S = \frac{3}{3 + 1 + 2} = \frac{3}{6} = 0{,}5$$

Como ya hemos indicado, si tuviésemos un conjunto de datos con una gran cantidad de variables que aparecieran con poca frecuencia, de forma que un solo individuo tuviese una pequeña proporción del total, sería preferible, obviamente, el coeficiente de Jaccard. En esa situación, el uso del coeficiente de comparación simple afirmaría un excesivo nivel de similaridad entre todos los casos.

Siempre que sean elegidos apropiadamente, los coeficientes de similaridad proporcionan una definición satisfactoria de la similaridad entre dos elementos; pero supongamos que queremos obtener una medida de la asociación entre *variables* de presencia/ausencia, de la misma forma que el coeficiente de

correlación proporcionaba esa medida en variables continuas. Esto plantea algunos problemas. Las medidas buscadas en tablas de contingencia 2 × 2 suelen ser insatisfactorias porque dependen del valor de la celda *d* en la tabla, la suma de los casos en los que ninguno de los dos atributos en consideración ocurren; en otras palabras, estarán fuertemente afectadas por las comparaciones negativas; y el hecho de que la intensidad de la asociación entre dos atributos esté determinada por el número de casos en los que no aparecen, no parece muy apropiado, especialmente si la cantidad de ocurrencias de algunos de los atributos es baja en relación a la cantidad de casos (véase Speth y Johnson, 1976). Una solución sería volver a usar el coeficiente de Jaccard. Veámoslo por medio de un ejemplo.

Supongamos que estamos estudiando una necrópolis y que no sólo nos interesan las similaridades entre las tumbas, sino también los esquemas de asociación entre distintos tipos de objetos de ajuar. Los detalles de la aparición de dos tipos hipotéticos se muestran en la tabla 12.5. En este ejemplo, el coeficiente de Jaccard adopta el valor $S = 3/6 = 0{,}5$.

Hay dos formas de enfrentarse a esta situación, y de ellas el coeficiente de Jaccard sólo representa una. Nos dice que la mitad de las presencias del tipo 1 están asociadas con el tipo 2. Sin embargo, si lo consideramos desde la perspectiva del menos frecuente de los dos atributos, el tipo 2, pensaremos de distinta manera: podemos decir que muestra una asociación perfecta con el atributo 1, ya que aparece allí donde el tipo 1 también está presente.

TABLA 12.5. Dos tipos de objetos de ajuar medidos según si están o no presentes en una serie de tumbas.

	Tumbas								
	1	2	3	4	5	6	7	8	9
Tipo de objeto 1	1	1	0	1	0	0	1	1	1
Tipo de objeto 2	1	0	0	0	0	0	1	0	1

Doran y Hodson (1975) ven en esta asimetría el motivo para rechazar, más o menos, este tipo de estudios de asociación; no obstante, esto parece una afirmación excesivamente drástica, dada la información útil que puede proporcionarnos la asociación de variables de presencia/ausencia. Su escepticismo concuerda con su preferencia por estudiar la similaridad entre individuos o casos, antes que la asociación entre variables, y con su rechazo hacia esta manera de usar las tablas de contingencia en los estudios de tipología y clasificación. Recomiendan a aquellos que deseen estudiar la asociación entre variables de presencia/ausencia el uso del coeficiente de Jaccard, aceptando sus límites; esto es lo que hizo Hodson en su análisis de la necrópolis de Hallstatt (Hodson, 1977). Una solución alternativa, sin embargo, especialmente si los atributos aparecen con frecuencias muy diferentes entre sí, es usar el siguiente coeficiente:

$$S = \frac{1}{2}\left(\frac{a}{a+c} + \frac{a}{a+b}\right)$$

cuyas letras se refieren a la tabla general 2 × 2 presentada en la tabla 12.3.

Este coeficiente considera las comparaciones positivas como una proporción de las presencias totales del primer atributo (aquí, tipo de objeto de ajuar funerario), y, a continuación, como una proporción del segundo atributo; finalmente se calcula su media aritmética. Es evidente que el atributo menos frecuente recibe un peso mucho mayor con este coeficiente que con el de Jaccard, aunque se pueda considerar que la media es un procedimiento espurio.

La matriz de coeficientes de asociación producida por esta técnica, como todas las matrices de coeficientes que hemos ido viendo, es simétrica: la mitad bajo la diagonal principal es la imagen reflejada de la situada por encima. Otra manera de enfocar el problema planteado por los casos que acabamos de mencionar es produciendo una matriz asimétrica en la que las dos mitades sean distintas; así, una mitad de la matriz estará constituida por los términos de la forma $a/(a + c)$ y la otra por los términos $a/(a + b)$. Se han publicado diversos métodos para analizar estas matrices (Gower, 1977; Constantine y Gower, 1978), si bien parecen tener pocos usos en arqueología.

Hasta ahora hemos distinguido las variables numéricas de un lado y las variables binarias del otro; no obstante, los datos arqueológicos a veces no están caracterizados por ninguna de éstas: son muy frecuentes los atributos *multiestado*. Un ejemplo, procedente del estudio de la cerámica, puede ser el *tipo de borde*. Este es el atributo o variable, y tendrá varios estados; por ejemplo, borde simple, entallado, torneado, con pico, etc., que son mutuamente exclusivos —sólo uno de los estados puede aparecer en una vasija— y exhaustivos —cubren todas las variedades de forma del borde que aparecen en el conjunto de datos estudiado.

Estas variables pueden registrarse como una serie de variables binarias. En el ejempo de los bordes podemos disponer de cuatro variables, una para cada tipo de borde; la que aparece en un caso en particular es codificada como presente, y las otras tres como ausentes. Sin embargo, es importante tener cuidado en la selección del coeficiente de similaridad en esas circunstancias: los que excluyen las comparaciones negativas serán satisfactorios, y los que las incluyan no, porque las variables están interconectadas lógicamente de forma que algunas comparaciones negativas tendrán lugar, simplemente, a causa de su propia definición.

A menudo, un conjunto de datos arqueológicos específico puede describirse en términos tanto de variables cuantitativas como de presencia/ausencia o multiestado. Muchos arqueólogos precisan usarlas todas en un único análisis; y con ese fin se propuso el *coeficiente general de similaridad* de Gower (Gower, 1971). La fórmula es:

$$S = \sum_{k=1}^{p} s_{ijk} \Big/ \sum_{k=1}^{p} w_{ijk}$$

Aquí se comparan dos individuos i y j basándose en una serie de variables; la similaridad s_{ijk} se evalúa para cada variable, y todos los resultados se suman al final. Esta suma se estandariza dividiéndola por la suma de los «pesos» [*weights*] w_{ijk} asociados con cada variable en la comparación con cada par de individuos i y j.

La función de los pesos ya ha sido mencionada anteriormente. Intuitivamente se deducen de la necesidad del uso de alguna medida de la importancia otorgada a ciertas variables en un análisis. En el caso del coeficiente de Gower, los pesos suelen utilizarse de una manera muy simple. El peso se fija en el valor 1 cuando es posible la comparación entre los objetos i y j para la k-ava variable, y en el valor 0 cuando el valor de la variable k es desconocido para los objetos i y j.

Cuando se trata de variables de presencia/ausencia, s_{ijk} está ajustado a 1 en el caso de una comparación positiva, y a 0 en el de una falta de comparación. Dado que en esas circunstancias el peso otorgado a la variable en cuestión será 1, el resultado de esa comparación se tendrá totalmente en cuenta en la evaluación del coeficiente de similaridad final. Si la comparación es negativa, ni i ni j tendrán ese atributo en particular, por lo que tendremos que volver a efectuar la elección de si lo registramos o no. Si la decisión es no, entonces el peso de esa variable en particular será cero, coincidiendo con el coeficiente de Jaccard; de lo contrario, el peso será 1.

Para variables cualitativas multiestado, $s_{ijk} = 1$ si los estados del atributo para las unidades i y j son los mismos, y 0 si difieren; w_{ijk} suele ser 1, a no ser que el atributo no sea aplicable, si bien aquí no hay razón alguna, en principio, para no cambiarlo según las ideas que tenga el analista acerca de la importancia relativa de los distintos estados.

Para variables cuantitativas:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$$

donde R_k es el rango para la variable k. En palabras, tomamos el valor de la variable k, para el objeto j, y lo restamos del valor del objeto i, en la misma variable, ignorando el signo del resultado de dicha resta. Dividimos a continuación ese resultado por el rango de la variable, es decir, la diferencia entre el valor más alto y el más bajo, restando finalmente ese resultado de 1. Obviamente, en el caso específico en el que los objetos i y j sean los que tienen el valor más alto y más bajo, respectivamente, el resultado del término $|x_{ik} - x_{jk}|/R_k$ será 1, el cual, restado de 1, produce una similaridad de 0 para esa variable en particular.

El coeficiente de Gower está descrito en detalle en Doran y Hodson (1975, pp. 142-143), y en los últimos años ha sido aplicado en muchos casos arqueológicos, como hemos señalado más arriba. Esto se debe, a mi juicio, a la tendencia por parte de los arqueólogos de querer volcar toda la información de un

conjunto de elementos en la forma de un análisis de conglomerados o procedimiento de ordenación, y ver qué es lo que resulta de ello. Esta actitud conlleva muchos riesgos en la medida en que sirve de excusa para no estudiar detenidamente las variables del análisis y lo que se supone que representan —las dimensiones subyacentes, como las denomina Whallon (1982)—. En muchos casos puede que no sea ni relevante ni apropiado incluir en un análisis todas las variables registradas en un conjunto de datos en particular. En otros casos puede ser mejor analizar separadamente los conjuntos de las variables, eligiéndolas por razones bien meditadas, y comparando los distintos resultados a continuación.

Para completar esta presentación de las medidas de similaridad, examinaremos un último coeficiente a pesar de sus indudables deficiencias (Doran y Hodson, 1975): el *coeficiente de acuerdo* de Robinson (Robinson, 1951), especialmente diseñado para estudios arqueológicos y, más concretamente, para la medida de la similaridad entre conjuntos de cerámica descritos por los porcentajes de diferentes tipos. Este coeficiente es una variante de la métrica *city-block*. Contabiliza las diferencias porcentuales entre categorías definidas por pares de conjuntos arqueológicos. La diferencia máxima entre dos unidades cualesquiera es el 200 %. Restando cualquier diferencia calculada de 200, se obtiene una medida equivalente de la similaridad o el acuerdo. La fórmula es:

$$S = 200 - \sum_{k=1}^{n} |P_{ik} - P_{jk}|$$

donde P es el porcentaje de representación del atributo o tipo k en los conjuntos i y j.

Puede ser útil demostrar las dos posibilidades extremas mediante ejemplos sencillos:

a)

	Tipo 1	Tipo 2
Conjunto 1	50 %	50 %
Conjunto 2	50 %	50 %

$$\sum |P_{ik} - P_{jk}| = |50 - 50| + |50 - 50| = 0$$
$$S = 200 - 0 = 200$$

b)

	Tipo 1	Tipo 2
Conjunto 1	100 %	0 %
Conjunto 2	0 %	100 %

$$\sum |P_{ik} - P_{jk}| = |100 - 0| + |0 - 100| = 200$$
$$S = 200 - 200 = 0$$

La búsqueda de asociaciones en matrices de similaridad (y distancia)

Una vez obtenida la matriz de coeficientes de similaridad, ¿qué podemos hacer con ella? En general, pretendemos buscar algún tipo de estructura subyacente que sea interpretable arqueológicamente. Cómo se investiga esa estructura depende tanto de lo que se tenga la intención de hacer, como de lo que uno espera encontrar.

Tal y como hemos visto, inicialmente, los arqueólogos que experimentan con métodos cuantitativos están interesados, particularmente, en asociaciones cronológicas, en establecer secuencias de unidades arqueológicas basándose exclusivamente en la comparación entre ellas, dada una situación en la que la datación externa o bien no esté disponible, o bien haya sido dejada de lado. Existe actualmente mucha bibliografía sobre este tema —la seriación— (por ejemplo: Cowgill, 1972; Doran y Hodson, 1975; Marquardt, 1978; Ester, 1981); desde entonces no se han emprendido nuevos enfoques innovadores, dado que se considera que la materia está ya bien tratada.

Presentaremos aquí una técnica peculiar de seriación que proporciona una visión introductoria muy conveniente a la práctica de la investigación de estructuras subyacentes en matrices de coeficientes de similaridad. El método se denomina *análisis de proximidades* (Renfrew y Sterud, 1969); es muy sencillo y no fuerza los datos en un orden lineal.

El procedimiento general es el siguiente: se elige una cualquiera de las unidades que vamos a seriar, se encuentra la unidad más semejante a la que hemos elegido y se coloca a su lado, empezando así una cadena. A continuación se busca la unidad más similar a cualquiera de los extremos de la cadena y se junta a ella, hasta que todas las unidades aparezcan formando una línea continua. El procedimiento en detalle se ofrece más abajo (Renfrew y Sterud, 1969, pp. 266-268), y un ejemplo de aplicación a continuación:

1. Señalar los dos coeficientes de similaridad más altos en cada columna de la matriz completa (a exclusión de los valores a lo largo de la diagonal).
2. Tomar cualquier unidad como punto de partida y registrar sus dos vecinos más próximos, es decir, elegir una columna de la matriz y registrar los dos coeficientes señalados en ella. Unir la unidad inicial a sus vecinos por medio de líneas marcadas con flechas que señalen la dirección de la similaridad; indicar el valor de los coeficientes en las líneas.
3. Tomar una de las unidades con sólo un vecino (en cualquiera de los extremos de la línea), señalar sus dos vecinos más próximos y repetir el procedimiento de la fase 2. Si ya está unida a una de ellas como resultado de ser el vecino más próximo de una de las unidades que ya están en el diagrama, la naturaleza recíproca de la relación se mostrará por medio de flechas en ambas direcciones.
4. Efectuar este procedimiento para cada una de las unidades a su vez, hasta que ya no queden más. Si no se han unido todas las unidades de la matriz, entonces habrá que empezar un nuevo conjunto de uniones con uno de los elementos no vinculados todavía a los demás. Repetir el procedimiento hasta que todas las unidades estén incluidas.
5. Allí donde se detecten bucles, se cortará el vínculo con el menor de los coeficientes, partiendo del requisito de que las cadenas no han de romperse en bloques separados. Usualmente se cortan así los vínculos unidireccionales, siendo preferible cortar los lazos simples antes que los dobles, si es que ambos tienen el mismo valor. Cuando esto se haya hecho así, se habrá obtenido la ordenación final. Hay que señalar que, mientras que el método fuerza la rotura de los bucles, no previene las ramificaciones: las cadenas laterales no se fuerzan en un orden lineal cuando este es inapropiado.
6. Si al final de la fase 4 han aparecido conglomerados totalmente separados del resto, sin uniones entre ellos, se unirán de la forma más apropiada, es decir, encontrando los dos coeficientes de similaridad más altos que unan los miembros de un grupo con los otros. Se tratará de esta forma cada conglomerado por separado. Las uniones entre grupos separados de unidades previamente unidas pueden registrarse mediante líneas punteadas para indicar su debilidad relativa.

Se dijo al principio que este método no fuerza las unidades en una cadena lineal, sino que permite la posibilidad de ramificaciones, si es que la estructura subyacente en los datos no es lineal. Renfrew y Sterud (1969, p. 267) sugieren una forma de afirmar el grado de agrupación en tanto que opuesto a un orden lineal en los datos. Depende de si puede obtenerse una seriación completa, con lo que habría dos coeficientes señalados en cada fila de la matriz, excepto en dos de ellas, que corresponderían a los extremos del orden lineal, los cuales tendrían tres. La fórmula es la siguiente:

$$\text{Coef. de agrupación} = \frac{\sum_{i=1}^{n} N_{i>2} - 2}{2n - 3} \times 100$$

donde $N_{i>2}$ es la cantidad de coeficientes en la fila i más allá de 2. Este coeficiente es cero en el caso de seriación perfecta y 100 en el caso de agrupación máxima.

Todo este procedimiento puede ilustrarse con ayuda de un ejemplo (fig. 12.3), usando el coeficiente de acuerdo de Robinson para representar las similaridades entre cuatro conjuntos cerámicos. En este caso, el coeficiente es $(0 + 1 + + 1 + 0 - 2)/(8 - 3) = 0$. En otras palabras, estamos ante una matriz perfectamente «seriable».

El método del análisis de proximidades es muy bueno cuando existe en los datos una ordenación del tipo de cadena lineal, pero cuando las relaciones en-

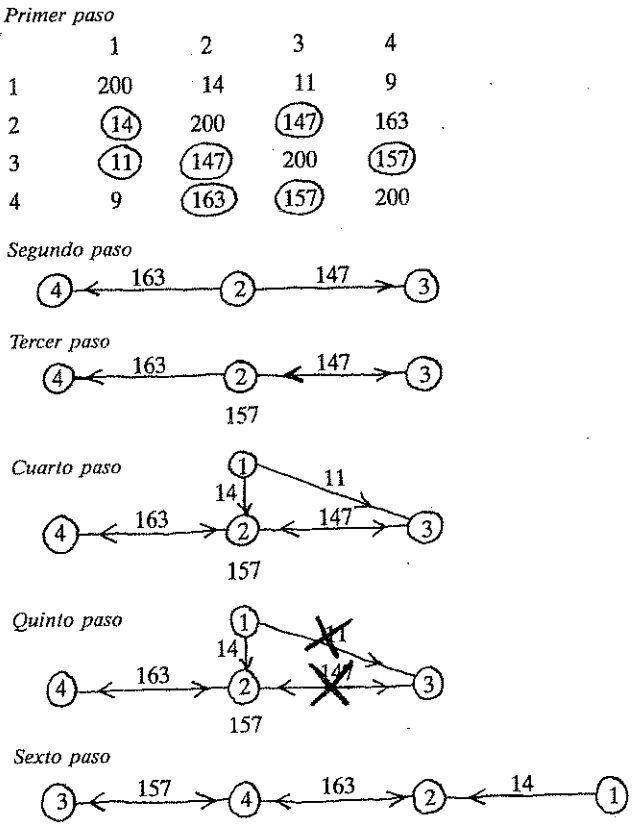


FIGURA 12.3. Análisis de proximidades.

tre las unidades son más complejas y el número de unidades es muy grande, no es muy satisfactorio. Además, su empleo presupone la existencia de la matriz de coeficientes de similaridad. Obtener esa matriz sólo es posible con ayuda de un ordenador; una vez los datos están ya en uno de sus archivos, es mucho más fácil usar una de las técnicas informáticas para buscar estructuras en la matriz que el análisis de proximidades. Esta última técnica, pues, es mucho más útil como forma de ilustrar el análisis de matrices de similaridad que por su uso práctico. Hoy en día, los estudios de seriación suelen realizarse por medio del tipo de métodos de ordenación que veremos en el capítulo siguiente (véase, por ejemplo, Kemp, 1982). En este capítulo, sin embargo, nos interesa examinar las estructuras en la matriz desde el punto de vista de la clasificación propiamente dicha, definida al principio como la definición de agrupaciones en un conjunto de datos, basándose en la idea de que los miembros del grupo

han de ser más similares a otro miembro de ese grupo que a uno que no sea miembro. Necesitaremos, pues, del análisis de conglomerados.

ANÁLISIS DE CONGLOMERADOS («CLUSTER ANALYSIS»)

Se señaló al principio que pueden dividirse en dos categorías, métodos partitivos y métodos jerárquicos, y que dentro de estos últimos podíamos distinguir entre las técnicas divisivas y las técnicas aglomerativas.

Métodos jerárquicos

Ya hemos visto que detrás de este grupo de técnicas está la idea de que los objetos han de ser similares unos con otros a diferentes niveles, de forma que los resultados puedan representarse por medio de un dendrograma: un diagrama en árbol que muestre las relaciones entre individuos y grupos.

Estas técnicas, como en general las de la taxonomía numérica, proceden de los estudios biológicos de clasificación. En ese dominio, la jerarquía de relaciones entre organismos individuales y los grupos que conforman era considerada por sus conexiones filogenéticas, es decir, el árbol de evolución. En los datos arqueológicos, la jerarquización de las interrelaciones entre individuos y grupos que corresponde a la representación jerárquica de la similaridad no es tan obvia, lo que ha llevado en ocasiones a rechazar el uso de las técnicas jerárquicas en arqueología. Clarke (1968) argumentó que, en realidad, era posible definir una jerarquía de entidades arqueológicas en términos de la cual puedan describirse relaciones sustantivas para la historia y la naturaleza de los ejemplos individuales de las entidades implicadas —conjuntos, culturas, grupos culturales, tecnocomplejos—. La postura adoptada en este libro es que la noción de similaridad, en algunos sentidos, pero no en otros, es muy familiar en arqueología, siendo útil y legítimo el conceptualizarla y presentarla de forma jerárquica.

Técnicas aglomerativas

Se empieza con una serie de individuos, constituyendo paulatinamente los grupos a partir de ellos. En otras palabras, lo que se hace en primer lugar es agrupar los elementos más similares entre sí, seguidamente se agregan a esos grupos nuevos elementos, uniendo también los grupos entre sí, a niveles de similaridad progresivamente menores, hasta que finalmente todos están unidos en un único grupo.

La tarea que llevan a cabo los métodos aglomerativos es efectuar la operación de la mejor manera, de acuerdo con cierto criterio predefinido. Existen muchos de esos métodos, porque existe una gran variedad de criterios en cuyos

términos ha de evaluarse la similaridad entre un individuo y un grupo, o entre dos grupos.

Vecino más próximo o enlace simple (nearest neighbour or single link cluster analysis). Es, probablemente, el método más sencillo, y por esa razón es muy útil para ilustrar lo que es en realidad un análisis de conglomerados. El criterio de vinculación, en este caso, es que para unir un individuo en particular a un grupo debe existir un nivel de similaridad específico entre el individuo y cualquiera de los miembros del grupo; para que dos grupos se unan, cualquier miembro de uno de los grupos ha de tener un nivel de similaridad específico con cualquier otro miembro del grupo. En otras palabras, la similaridad o la distancia entre individuos y grupos, o entre grupos y otros grupos, se define como la existente entre sus vecinos más próximos.

TABLA 12.6. Matriz de similaridades entre cinco vasijas cerámicas, basada en los motivos decorativos (Everitt, 1980).

	1	2	3	4	5
1	1,0	0,8	0,4	0,0	0,1
2	0,8	1,0	0,5	0,1	0,2
3	0,4	0,5	1,0	0,6	0,5
4	0,0	0,1	0,6	1,0	0,7
5	0,1	0,2	0,5	0,7	1,0

Ilustraremos este procedimiento realizando el análisis de una pequeña matriz de similaridad entre cinco vasijas cerámicas, basándonos en sus motivos decorativos (véase la tabla 12.6; cf. Everitt, 1980, pp. 9-10). La mayor de todas las similaridades es la existente entre las vasijas 1 y 2, por lo que el primer paso será unirlas. Ya no tendrán identidades separadas en la matriz; serán un grupo y las similaridades de los otros individuos con este grupo se recalcularán de acuerdo con el criterio del vecino más próximo. Lógicamente, con esto se va a obtener una matriz de similaridad revisada.

Por ejemplo, para encontrar la similaridad entre el grupo y la vasija 3, observaremos las similaridades entre 1 y 3 y entre 2 y 3; la mayor de ellas hará las veces de similaridad entre el grupo y el individuo. Aquí, la similaridad entre 1 y 3 es de 0,4, y entre 2 y 3 es de 0,5, por lo que elegiremos esta última. El mismo procedimiento se sigue para unir al grupo las vasijas 4 y 5. La matriz resultante aparece en la tabla 12.7.

TABLA 12.7. Matriz reducida de las similaridades entre cinco vasijas de cerámica, después de una primera agrupación (vasijas 1 y 2) según el criterio del vecino más próximo (según Everitt, 1980).

	(12)	3	4	5
(12)	1,0	0,5	0,1	0,2
3	0,5	1,0	0,6	0,5
4	0,1	0,6	1,0	0,7
5	0,2	0,5	0,7	1,0

En esta matriz, a su vez, buscamos el mayor de los coeficientes, la similaridad entre las vasijas 4 y 5, que es de 0,7, con lo que ambas formarán un nuevo grupo, cuya similaridad con el primer grupo y el resto de los individuos permitirá establecer una tercera matriz. El procedimiento es igual que el anterior. La similaridad entre el primer grupo (vasijas 1 y 2) y la vasija 3 no cambia (0,5). Para encontrar la similaridad entre los dos grupos según el criterio del vecino más próximo, se observa la mayor de las dos similaridades entre el grupo 1 y la vasija 4, y entre el grupo 1 y la vasija 5; vemos que es 0,2 en el último caso. La matriz resultante es la que aparece en la tabla 12.8.

TABLA 12.8. Matriz reducida de las similaridades entre cinco vasijas de cerámica, después de una segunda agrupación según el criterio del vecino más próximo, que ha agrupado las vasijas 4 y 5, además de la 1 y la 2 (según Everitt, 1980).

	(12)	3	(45)
(12)	1,0	0,5	0,2
3	0,5	1,0	0,6
(45)	0,2	0,6	1,0

El siguiente paso consiste en unir el individuo 3 al segundo grupo, con un nivel de 0,6; el último paso será la unión de ambos grupos a un nivel de similaridad de 0,5. La secuencia de las uniones puede representarse como un dendrograma, con la escala de similaridad a un lado (fig. 12.4).

Vecino más alejado o enlace completo (furthest neighbour or complete linkage cluster analysis). El criterio especificado por este método es que, para unir un individuo a un grupo, el individuo ha de tener un grado de similaridad específico con el miembro del grupo más distinto a él; para unir dos grupos, los dos individuos, uno de cada grupo, que sean los más diferentes entre sí, han de tener un grado de similaridad específico. Nuevamente hemos de buscar los valores más altos de similaridad en la sucesión de matrices, pero definiremos los grupos según el criterio del vecino más alejado y no según el criterio del vecino más próximo.

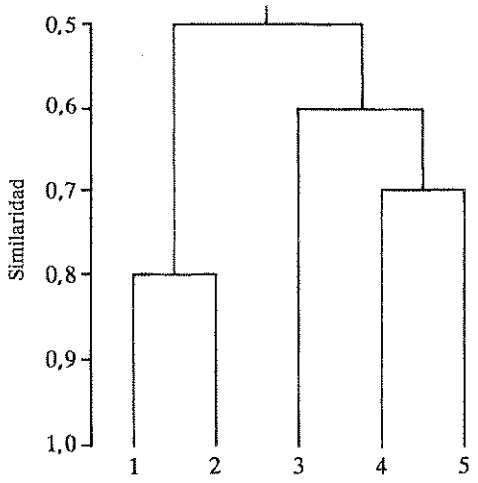


FIGURA 12.4. Dendrograma de los resultados del análisis de conglomerados por enlace simple para la matriz de similaridades entre cinco vasijas cerámicas expuesta en la tabla 12.6 (según Everitt, 1980).

El dendrograma resultante para el mismo ejemplo que en el método anterior aparece en la figura 12.5. En este caso, han cambiado las similaridades relativas, pero la configuración del dendrograma es idéntica a la del enlace simple. Lo más frecuente es que esa configuración sea muy distinta.

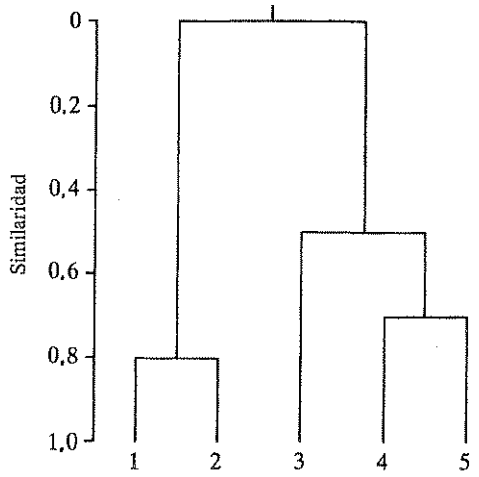


FIGURA 12.5. Dendrograma de los resultados del análisis de conglomerados por enlace del vecino más alejado para la matriz de similaridades entre cinco vasijas cerámicas expuesta en la tabla 12.6 (según Everitt, 1980).

Medias de grupo, o análisis de conglomerados por promedio de las uniones (group average or average-link cluster analysis). A veces es conocido como el método de agrupación por pares «no pesados». Aquí la similaridad o la disimilaridad entre grupos se define como la media aritmética de las similaridades entre pares de miembros, es decir, como:

$$\frac{\sum_{i=1}^{n_i}\sum_{j=1}^{n_j} S_{ij}}{n_i n_j}$$

donde S_{ij} es la similaridad entre un miembro del grupo i y un miembro del grupo j, n_i es la cantidad de individuos en el grupo i, y n_j es la cantidad de individuos en el grupo j. La fórmula nos explica que hemos de tomar el primer individuo del grupo i, que obtengamos la similaridad entre él y todos los miembros del grupo j, que sumemos esas similaridades y que vayamos al segundo miembro del grupo i, repitiendo el procedimiento hasta que se hayan tenido en cuenta todos los miembros del grupo i. El resultado global de la suma de similaridades se divide, a continuación, por el producto de la cantidad de individuos en cada uno de los dos grupos. Gráficamente, este método está ilustrado en la figura 12.6.

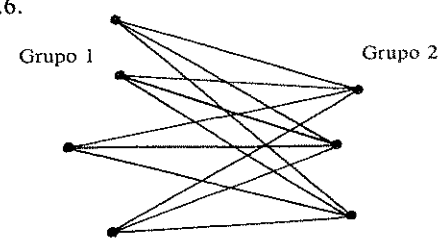


FIGURA 12.6. Diagrama que ilustra el cálculo del promedio de similaridades entre dos grupos.

Aparecen dos grupos, 1 y 2. Las líneas que unen sus respectivos miembros son los S_{ij} (o d_{ij} en este caso), de los cuales hay 12. Se suman y se dividen por $n_1 \times n_2$, es decir, el número de elementos de cada uno de los dos grupos; aquí $4 \times 3 = 12$. En cada fase del análisis de conglomerados por promedios, las similaridades entre los grupos y/o individuos se calculan siguiendo el criterio de promedio del grupo. Se unen entre sí los grupos y/o individuos con la mayor similaridad o la menor distancia en cada uno de los pasos.

Método de Ward. Hay otras muchas técnicas jerárquicas aglomerativas (véanse, por ejemplo, Sneath y Sokal, 1973; Everitt, 1980), si bien muchas de ellas no son sino variaciones de un procedimiento más general (Everitt, 1980, pp. 16-17; Gordon, 1981, pp. 46-49). Aquí sólo describiremos uno más, el mé-

todo de Ward, que ha sido empleado muchas veces en arqueología, sobre todo en el análisis de datos numéricos continuos, tales como los resultados de los elementos traza, y, recientemente, por Whallon (1984) en un interesante análisis microespacial (*intra-site*).

La idea fundamental es que los conglomerados han de ser lo más homogéneos posible. Una manera de definir la homogeneidad es en términos de la distancia de los miembros de un conglomerado a su media. En el método de Ward, la distancia es la *suma de cuadrados del error* (SCE): la suma total de las desviaciones al cuadrado, en donde las desviaciones no son más que las distancias de todos los puntos a las medias de los conglomerados a los que pertenecen. El propósito del método es unir individuos y grupos sucesivamente, de forma que en cada fase del procedimiento de fusión la suma de los errores al cuadrado sea la menor posible; en otras palabras, los conglomerados serán lo más homogéneos posible. Se entenderá mejor este método por medio de un ejemplo (véase Everitt, 1980, pp. 16-17).

TABLA 12.9. Matriz de distancias euclídeas al cuadrado entre cinco puntas de flecha, basadas en las medidas que describen su forma.

	1	2	3	4	5
1	0,0	1,0	36,0	64,0	121,0
2	1,0	0,0	25,0	49,0	100,0
3	36,0	25,0	0,0	4,0	25,0
4	64,0	49,0	4,0	0,0	9,0
5	121,0	100,0	25,0	9,0	0,0

La matriz de las distancias al cuadrado entre cinco puntas de flecha se basa en mediciones cuantitativas de las variables que describen su forma (tabla 12.9). Al principio, cuando todos los individuos están separados unos de otros, la SCE total tiene un valor 0. A continuación se unen aquellos individuos con la menor distancia entre ellos, es decir, aquellos cuya fusión producirá un aumento mínimo en SCE. En este caso, los individuos 1 y 2, separados por una distancia al cuadrado de 1,0. Aquí tratamos sólo con dos individuos; por lo tanto, el incremento en SCE (Gordon, 1981, p. 42) está dado por:

$$I = \frac{1}{2} d_{ij}$$

Aquí:

$$I_{(12)} = \frac{1}{2} 1{,}0 = 0{,}5$$

Al igual que en el ejemplo del análisis de conglomerados por enlace simple hay que calcular una nueva matriz reducida, dadas las distancias entre la media del

grupo y los otros ítems en el análisis. Una fórmula general para obtener las nuevas distancias es la proporcionada por Gordon (1981, p. 42):

$$d_{k(ij)} = \frac{n_k + n_i}{n_k + n_i + n_j} d_{ki} + \frac{n_k + n_j}{n_k + n_i + n_j} d_{kj} - \frac{n_l}{n_k + n_i + n_j} d_{ij}$$

donde $d_{k(ij)}$ es la distancia entre el grupo o elemento k y el nuevo grupo constituido a su vez por los grupos o elementos i y j; n_i es la cantidad de elementos en el grupo i; n_j es la cantidad de elementos en el grupo j; n_k es la cantidad de elementos en el grupo k; d_{ki} es la distancia entre el grupo/elemento k y el grupo/elemento i; d_{kj} es la distancia entre el grupo/elemento k y el grupo/elemento j, y d_{ij} es la distancia entre el grupo/elemento i y el grupo/elemento j. En este ejemplo, el cálculo es el siguiente (en la práctica, por supuesto, se ha realizado por medio de un ordenador):

$$d_{3(12)} = \frac{1+1}{1+1+1} 36 + \frac{1+1}{1+1+1} 25 - \frac{1}{1+1+1} 1 =$$

$$= 24{,}0 + 16{,}666 - 0{,}333 = 40{,}333$$

$$d_{4(12)} = \frac{2}{3} 64 + \frac{2}{3} 49 - \frac{1}{3} 1 =$$

$$= 42{,}666 + 32{,}666 - 0{,}333 = 75{,}0$$

$$d_{5(12)} = \frac{2}{3} 121 + \frac{3}{3} 100 - \frac{1}{3} 1 =$$

$$= 80{,}666 + 66{,}666 - 0{,}333 = 147{,}0$$

El resto de las distancias permanece igual. Los resultados se pueden consultar en la tabla 12.10.

TABLA 12.10. Matriz reducida de las distancias entre cinco puntas de flecha, tras el primer paso de un análisis de conglomerados por el método de Ward, que une los elementos 1 y 2.

	(12)	3	4	5
(12)	0,0	40,333	75,0	147,0
3	40,333	0,0	4,0	25,0
4	75,0	4,0	0,0	9,0
5	147,0	25,0	9,0	0,0

La menor de las distancias es ahora la existente entre los individuos 3 y 4, una distancia de 4,0. Nuevamente hemos de encontrar el incremento de SCE que resulta de la formación del nuevo grupo. Igual que antes calculamos:

$$I_{(34)} = \frac{1}{2}\,4 = 2{,}0$$

y una vez más hemos de generar una nueva matriz (tabla 12.11):

$$d_{(12)(34)} = \frac{2+1}{2+1+1}\,40{,}333 + \frac{2+1}{2+1+1}\,75{,}0 - \frac{2}{2+1+1}\,4 =$$

$$= 30{,}25 + 56{,}25 - 2 = 84{,}5$$

$$d_{5(34)} = 16{,}666 + 6 - 1{,}333 = 21{,}333$$

TABLA 12.11. Matriz reducida de las distancias entre cinco puntas de flecha tras el segundo paso de un análisis de conglomerados por el método de Ward, que une los elementos 3 y 4, además de 1 y 2.

	(12)	(34)	5
(12)	0,0	84,5	147,0
(34)	84,5	0,0	21,333
5	147,0	21,333	0,0

Examinando la nueva matriz, podemos ver que la menor de las distancias es ahora la existente entre el grupo 3 y 4 y el individuo 5 (21,333). Como antes, el incremento en SCE al incluir un nuevo individuo a un grupo es igual a la mitad de la distancia entre ellos. En este caso, 21,333/2 = 10,666.

Queda por evaluar la entrada única en la matriz (tabla 12.12), es decir, la distancia entre el grupo (12) y el grupo (345):

$$d_{(12)(345)} = \frac{2+2}{2+2+1}\,84{,}5 + \frac{2+2}{2+2+1}\,147{,}0 - \frac{2}{2+2+1}\,21{,}333 =$$

$$= 67{,}6 + 88{,}2 - 8{,}532 = 147{,}268$$

TABLA 12.12. Matriz reducida de las distancias entre cinco puntas de flecha en el último paso de un análisis de conglomerados por el método de Ward.

	(12)	(345)
(12)	0,0	147,268
(345)	147,268	0,0

Nuevamente, el incremento del SCE es la mitad de la distancia, proporcionando un valor de 73,65. Los resultados pueden resumirse en forma de tabla (12.13) y las uniones representarse por medio de un dendrograma (fig. 12.7).

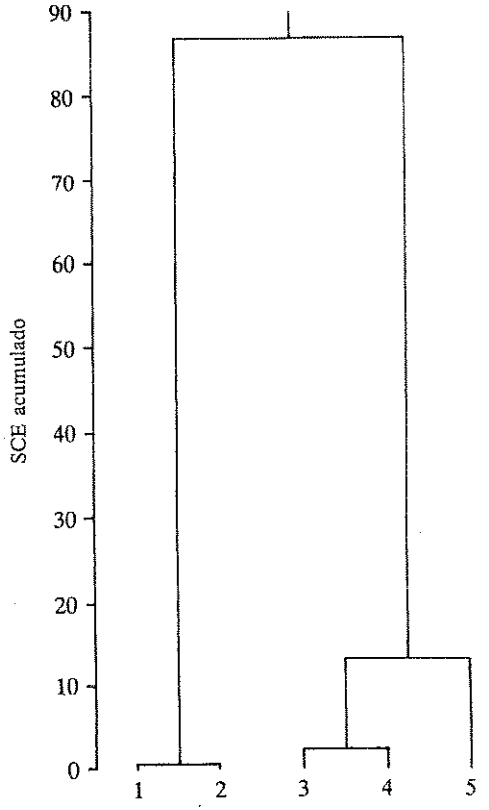


FIGURA 12.7. Dendrograma resultante del análisis según el método de Ward, para la matriz de distancias de la tabla 12.9.

TABLA 12.13. Incremento en la suma de cuadrados del error (SCE) asociado a las uniones sucesivas en el análisis de conglomerados por el método de Ward con los datos de la tabla 12.9.

Fusión	Incremento del SCE	SCE acumulado
1 2	0,5	0,5
3 4	2,0	2,5
(34) 5	10,7	13,2
(12) (345)	73,7	86,9

Técnicas divisivas

Este tipo de métodos suele empezar integrando todos los individuos o unidades en un solo grupo, subdividiéndolo sucesivamente. Hay dos grupos principales de métodos divisivos, politéticos y monotéticos; los primeros consideran los valores de todas las variables en cada fase de decisión, los segundos sólo utilizan los valores de una única variable (véase Everitt, 1980, p. 18). Sólo el enfoque monotético va a ser estudiado aquí, pues es el único usado en arqueología (Tainter, 1975; Peebles, 1972).

Su uso práctico ha estado restringido a los casos en que los datos son del tipo presencia/ausencia, o con valores 1/0; dado un atributo en particular, todos los elementos con valor 1 estarán en un grupo, mientras que los que tienen valor 0 aparecen en el otro. Cuando se necesita una serie de fases divisivas sucesivas, produciendo subdivisiones más y más pequeñas, el resultado, una vez más, será una jerarquía.

Los métodos divisivos han sido usados especialmente por los ecólogos para clasificar áreas según las especies presentes, o las especies según su presencia en áreas determinadas. El método principal es conocido como *análisis asociativo* y corresponde al programa DIVIDE en el paquete de programas informáticos CLUSTAN, específico para el análisis de conglomerados (véase anexo 2).

Habiendo dicho que la serie de divisiones se hace en términos de la presencia o ausencia (valores 1 o 0) de un único atributo en cualquier momento dado, la cuestión que se plantea es la forma de elegir el mejor atributo para efectuar esa división: ¿qué quiere decir aquí «el mejor»? La idea fundamental es que los dos grupos producidos en cualquiera de los pasos de la división han de ser lo más distintos posible el uno del otro, no sólo en términos de su valor en la variable usada para hacer la división, sino globalmente, en términos de todas las variables en el análisis. En otras palabras, la presencia/ausencia del atributo usado para la división ha de estar relacionada con los valores presencia/ausencia de los otros atributos; el atributo cuya presencia/ausencia esté más estrechamente relacionada con los valores de las otras variables en el conjunto de datos que estamos subdividiendo será el que elijamos. Incluso así, este criterio general puede definirse de distintas formas. La que se ha elegido aquí para mostrar el método es de uso muy frecuente, aunque no es precisamente satisfactoria, tal y como veremos. Sin embargo, proporciona un ejemplo sencillo.

Supongamos que hemos codificado diez tumbas según la presencia/ausencia de cuatro tipos de objetos de ajuar funerario (tabla 12.14). Una forma de ver si los valores en una variable están relacionados con los valores en la otra variable es calcular la prueba de χ^2 para la asociación entre las dos variables. La idea subyacente es que los valores χ^2 miden la intensidad de la asociación en un contexto como este porque el tamaño de la muestra es constante en todas las comparaciones.

TABLA 12.14. Matriz de datos para diez tumbas, según la presencia/ausencia de cuatro tipos de objetos de ajuar.

	Tipos de objetos			
Tumba	1	2	3	4
1	0	0	1	1
2	0	0	1	1
3	1	1	1	0
4	1	1	1	0
5	0	1	1	1
6	1	1	0	0
7	1	1	1	1
8	1	0	0	0
9	0	0	0	1
10	0	0	0	1

Si una variable en particular está intensamente relacionada con otras, el valor de cualquier caso en la primera variable será un buen predictor de su valor en las demás. El resultado es que un grupo definido según la presencia o ausencia de una variable será relativamente homogéneo, ya que el estado de esa variable especificará los estados particulares (o atributos) de las otras variables adoptados en los miembros de ese grupo. La variable más estrechamente relacionada con las otras, y por tanto la más apropiada para hacer una división, será aquella con los valores χ^2 más altos en sus relaciones con los demás. Por lo tanto, habremos de calcular el valor χ^2 para todas las asociaciones de atributos, y sumar los resultados para cada variable, con el fin de ver cuál es la que tiene el valor más alto, tal y como lo manifiesta la tabla 12.15.

TABLA 12.15. Tablas de contingencia que muestran las asociaciones entre cada par de tipos de objetos de ajuar para los datos presentados en la tabla 12.14.

(a)	Tipo 1		+	−	Total	
	Tipo 2	+	4	1	5	$\chi^2_{12} = 3{,}6$
		−	1	4	5	
	Total		5	5	10	
(b)	Tipo 1		+	−	Total	
	Tipo 3	+	3	3	6	$\chi^2_{13} = 0$
		−	2	2	4	
	Total		5	5	10	

TABLA 12.15. *Continuación*

(c)	Tipo 1		+	−	Total	
	Tipo 4	+	1	5	6	$\chi^2_{14} = 6{,}666$
		−	4	0	4	
	Total		5	5	10	
(d)	Tipo 2		+	−	Total	
	Tipo 3	+	4	2	6	$\chi^2_{23} = 1{,}666$
		−	1	3	4	
	Total		5	5	10	
(e)	Tipo 2		+	−	Total	
	Tipo 4	+	2	4	6	$\chi^2_{24} = 1{,}666$
		−	3	1	4	
	Total		5	5	10	
(f)	Tipo 3		+	−	Total	
	Tipo 4	+	4	2	6	$\chi^2_{34} = 0{,}278$
		−	2	2	4	
	Total		6	4	10	

Sumamos los valores para cada variable:

Objetos de ajuar tipo 1 = $\chi^2_{12} + \chi^2_{13} + \chi^2_{14} = 3{,}6 + 0{,}0 + 6{,}666 = 10{,}266$

Objetos de ajuar tipo 2 = $\chi^2_{21} + \chi^2_{23} + \chi^2_{24} = 3{,}6 + 1{,}666 + 1{,}666 = 6{,}932$

Objetos de ajuar tipo 3 = $\chi^2_{31} + \chi^2_{32} + \chi^2_{34} = 0{,}0 + 1{,}666 + 0{,}278 = 1{,}944$

Objetos de ajuar tipo 4 = $\chi^2_{41} + \chi^2_{42} + \chi^2_{43} = 6{,}666 + 1{,}666 + 0{,}278 = 8{,}61$

Se deduce que la variable 1 es la más estrechamente relacionada a las demás; es decir, la presencia o ausencia de los otros tipos de objetos de ajuar está más estrechamente relacionada a la presencia o ausencia de objetos del tipo 1. El resultado es que la mejor de todas las divisiones posibles de las tumbas es aquella que se establece entre los grupos en los que el tipo 1 está presente y aquellos en los que está ausente. No se pueden obtener dos grupos más distintos entre sí que los basados en el criterio anterior. En la tabla 12.16 las tumbas están dispuestas según los dos grupos. Se puede ver que los tipos 2 y 4 están distribuidos muy diferentemente en las dos subdivisiones, y que el tipo 3

TABLA 12.16. Tumbas enumeradas en la tabla 12.14, dispuestas de forma que todas aquellas con objetos del tipo 1 aparezcan juntas, diferenciadas de aquellas que no tienen objetos de ese tipo.

		Tipos de objetos de ajuar		
	Tumba	2	3	4
Tipo 1 presente	3	1	1	0
	4	1	1	0
	6	1	0	0
	7	1	1	1
	8	1	0	0
Tipo 1 ausente	1	0	1	1
	2	0	1	1
	5	1	1	1
	9	0	0	1
	10	0	0	1

presenta una distribución idéntica a ambas, pues su presencia/ausencia no está asociada a las del tipo 1, tal y como ponía de manifiesto la prueba de χ^2. Sólo se ha mostrado un paso en esta división; en los análisis de asociación, sin embargo, se suele efectuar una sucesión de divisiones de este tipo.

Hay varios problemas al usar el χ^2 de esta manera (véase Cormack, 1971); uno muy obvio es el que se refiere a la cuestión, ya discutida con anterioridad, de cómo tratar la celda *d* de la tabla de contingencia, esto es, la ausencia conjunta o comparaciones negativas. El método de χ^2 se ha presentado aquí como un ejemplo, y no como una técnica recomendable. Existen otros criterios de división, en particular el conocido como *estadística de información*. Su uso en estudios de clasificación se basa en el concepto de entropía o desorden. Proporciona una medida del desorden en un grupo. Adopta un valor cero cuando todos los elementos del conglomerado son idénticos, y aumenta a medida que el grupo se hace más diverso (véase Sneath y Sokal, 1973, pp. 141-144, 241-244). Esta prueba estadística, disponible en CLUSTAN, ha sido defendida por Peebles (1972) y por Tainter (1975), en tanto que Doran y Hodson (1975, p. 180), también la encuentran satisfactoria.

Métodos partitivos

Todos los métodos de análisis de conglomerados que hemos ido viendo son jerárquicos, por lo que ya va siendo hora de que prestemos atención a los métodos partitivos, a los que se ha hecho referencia al principio. En vez de operar

con múltiples niveles de agrupación a diferentes niveles de similaridad, se toma una decisión previa acerca del número de conglomerados, asignándose los individuos a aquel de los conglomerados que esté más cerca. Este procedimiento de asignación no es nada simple, porque cada vez que se añade un individuo a un grupo la definición del grupo cambia. No se trata de técnicas analíticas que produzcan una única solución correcta, pues el número de variaciones posibles en la asignación de elementos a grupos se hace enormemente grande a medida que aumenta el número de elementos en el análisis; por el contrario, se trata de técnicas que emplean a fondo la velocidad del ordenador para calcular un gran número de operaciones de búsqueda en los datos, asignando individuos a grupos de acuerdo con un conjunto de reglas que se basa en cierto criterio. La asignación producida es lo más próxima posible a la solución deseada, pero ésta no puede garantizarse.

La primera decisión que hay que tomar incumbe al número de conglomerados de partida, si bien en la práctica es posible operar de forma más o menos jerárquica, reduciendo sucesivamente la cantidad de conglomerados. Una vez decidida la cantidad inicial de grupos, es preciso proporcionar un fundamento a los mismos. Los procedimientos sugeridos para definir los puntos de partida incluyen la selección aleatoria de una cantidad específica de casos individuales, que corresponda a la cantidad de grupos restringida, y el uso de los resultados de algún otro método de agrupación, para la cantidad relevante de conglomerados. Cuando el centro de los conglomerados de partida ha sido elegido, los individuos siguientes serán asignados a aquel cuyo centro les sea más próximo.

La idea de situar individuos en los grupos cuyo centro está más próximo es, básicamente, la misma que en el método de Ward; muchas veces se usa la suma de cuadrados del error, basada en la distancia euclídea al cuadrado. En el método de Ward, la mejor fusión de individuos en grupos se consigue mediante un esquema jerárquico de los enlaces, pero, como en todos los demás métodos de análisis de conglomerados, una vez constituido el conglomerado no se puede romper y sus miembros no pueden redistribuirse a otro grupo o grupos. Esto puede conducir a situaciones anómalas en las que la pertenencia de un individuo a un grupo puede ser apropiada cuando este se une al grupo; pero en el momento en que otros individuos son agregados al mismo, la definición del primero cambia hasta el extremo en que bien pudiera llegar a unirse a otro grupo con el que, antes, no tenía muchas afinidades. La idea de recalcular la asignación en cada una de las etapas, y si es preciso cambiar la ubicación de los individuos es, intuitivamente, muy atractiva (Doran y Hodson, 1975, p. 180).

Esto es, precisamente, lo que hacen las técnicas de análisis partitivos, conocidas como *recolocación iterativa* (o bien, *k-medias*). A medida que se añaden individuos al grupo, se recalcula el centro cada vez. La cuestión que se plantea es si todos los elementos están en el grupo más apropiado, considerándose por tanto cada elemento por separado, para ver si debe ser reasignado a otro grupo

o no. Se han propuesto diversos criterios para tomar esas decisiones, si bien la idea fundamental de todos ellos es que ha de reducirse la dispersión de los distintos conglomerados y que las distinciones entre ellos han de maximizarse, nuevamente un concepto muy similar a la idea de minimizar la SCE en el método de Ward.

Evidentemente, una vez que un elemento ha sido trasladado, el centro del conglomerado del que procede y de aquel al que va reasignado se recalculan, por lo que el procedimiento de recolocación es laborioso y se requieren algoritmos informáticos eficaces. El procedimiento continúa hasta que ningún otro movimiento es capaz de mejorar el criterio usado. Una versión muy simplificada aparece en la figura 12.8, en la que sólo se han considerado dos conglomerados.

La solución conseguida al final del procedimiento de recolocación puede representar o no la mejor colocación global posible (u *optimum global*, como en ocasiones se denomina). Una manera de comprobarlo es repitiendo el procedimiento usando unos puntos de partida diferentes, elegidos aleatoriamente o según los resultados de un técnica de agrupación anterior.

Como ya se ha señalado, aunque el procedimiento se efectúe para una cantidad específica de grupos, puede efectuarse también de forma jerárquica. Así, una vez que se encuentra la solución mejor ajustada a un número dado de conglomerados, los dos conglomerados más próximos entre sí pueden unirse, repitiéndose a continuación el procedimiento. Cuando se ha hecho esto, el número de conglomerados puede reducirse otra vez, repitiéndose todo el proceso para tantos conglomerados como se necesite. Es importante señalar que este procedimiento no es jerárquico en el sentido en que lo eran los métodos que hemos visto antes, los cuales producen conglomerados cuyos miembros no son intercambiables, exceptuando aquellos casos en los que se añaden nuevos miembros porque se ha reducido el número de conglomerados. En los métodos de recolocación iterativa descritos en esta sección, los conglomerados pueden cambiar su pertenencia, así como adquirir nuevos miembros o reducir su número.

Al igual que en todos los procedimientos descritos, existe un programa para la recolocación iterativa en el paquete CLUSTAN, especializado en los análisis de conglomerados.

Evaluación del análisis de conglomerados

Ya se apuntó antes que el análisis de conglomerados, en mayor o menor grado, impone su propia estructuración en los datos. Se ha visto claramente, también, a lo largo de la descripción de las distintas técnicas, que hay muchas formas distintas de definir los conglomerados. Inevitablemente, esto significa que muy a menudo se producirán resultados diferentes analizando el mismo conjunto de datos. Dos cuestiones importantes se pueden plantear al respecto: ¿cómo

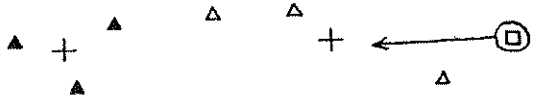
1. Se eligen dos individuos como puntos de partida de los dos grupos; se introduce un tercer individuo, que es situado en el grupo más cercano:

2. Se recalcula la posición del centro del segundo grupo; se introduce un nuevo individuo que también es asignado a uno de los grupos:

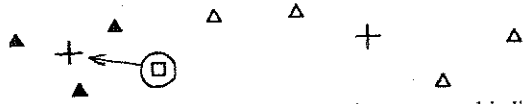
3. Se recalcula la situación del centro del primer grupo; se introduce un nuevo individuo que también es asignado a uno de los grupos:

4. Se recalcula la situación del centro del segundo grupo; se introduce un nuevo individuo que también es asignado a uno de los grupos:

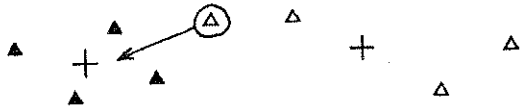
5. Se recalcula la situación del centro del primer grupo; se introduce un nuevo individuo que también es asignado a uno de los grupos:



6. Se recalcula la situación del centro del segundo grupo; se introduce un nuevo individuo que también es asignado a uno de los grupos:



7. Se recalcula la situación del centro del primer grupo; el individuo del segundo grupo situado más a la izquierda está ahora más cerca del centro del primer grupo que del segundo, por lo que pasará a pertenecer al primer grupo:



8. Los centros de ambos grupos son, finalmente, recalculados:

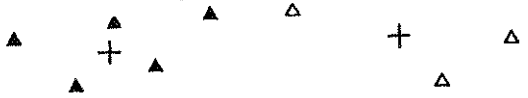


FIGURA 12.8. Fases sucesivas en un procedimiento partitivo de recolocación iterativa para dos conglomerados.

podemos llegar a saber si nuestros grupos representan distinciones auténticas en los datos y no son un mero subproducto del método usado? y ¿qué base tenemos para preferir los resultados de uno de los métodos y no de otro? Una cuestión secundaria relacionada en cierto sentido con la primera es: ¿cómo decidir el número de grupos con el que vamos a trabajar?

Las respuestas a estas cuestiones no son, en modo alguno, simples, pues no hay ninguna forma sencilla de distinguir entre métodos correctos y métodos incorrectos, a no ser que se considere el criterio en cuyos términos una técnica en particular define una buena agrupación y si es apropiada a la estructura de los datos disponibles. Sin embargo, el análisis de conglomerados se usa, precisamente, en aquellas situaciones en las que se sabe muy poco acerca de la estructura de los datos. Igualmente, el fundamento teórico de muchos de los métodos es incierto. Existe mucha bibliografía al respecto, que no podemos detallar aquí; Doran y Hodson (1975), Everitt (1980) y Gordon (1981), entre otros, dan las referencias necesarias.

Cuando estas técnicas empezaron a aplicarse en arqueología, se enfatizó el grado en que los resultados de la clasificación eran comparables con los producidos por los métodos tipológicos tradicionales. Era importante en un período en el que se pretendía justificar la respetabilidad de las nuevas técnicas. Además, tales comparaciones pueden proporcionar información útil, al igual que la comparación de los resultados con información externa, algo que no se hace en los análisis de conglomerados; de otra manera no se pueden considerar válidos los resultados de tales análisis. Por otro lado, si los métodos de clasificación numérica se limitasen a coincidir con las tipologías tradicionales, no habría motivo alguno para emplearlas.

En un artículo reciente, Aldenderfer (1982) ha revisado algunas de las formas más importantes de evaluación de los resultados de un análisis de conglomerados, sugiriendo que no es satisfactorio usarlas por separado. Son las siguientes:

1. El uso de una regla de detención, un medio de comprobar el número de grupos en una secuencia jerárquica (Mojena, 1977); disponible en CLUSTAN.

2. El uso de la prueba lambda de Wilk, una medida basada en la proporción entre la variación dentro del grupo y la variación global en los datos, contrastada por medio de un procedimiento de aleatorización (véase más adelante).

3. El uso de diagramas de dispersión de los datos, los pares de variables a un tiempo, para ver si hay indicaciones de agrupación y la forma que adoptan los conglomerados.

4. Por medio del análisis discriminante, que pretende maximizar la separación existente entre los grupos y proporciona un indicio del grado en que eso es posible (véase el capítulo 13, y Everitt y Dunn, 1983, pp. 106-109).

5. Por medio de la representación gráfica de la distribución de los datos en términos de las puntuaciones en una serie de ejes transformados por técnicas de reducción de datos (análisis de componentes principales) y no en tér-

minos de las variables originales. Se trata de los medios de ordenación a los que nos referíamos al principio del capítulo; se describirán en el capítulo 13.

Otra manera de intentar asegurar la validez de los resultados de un análisis de conglomerados en un conjunto de datos en particular se consigue al analizarlo mediante distintos métodos. Si todas las respuestas fuesen similares en lo que se refiere a la superposición entre los miembros de los conglomerados, la estructura descubierta sería auténtica. Gordon (1981, pp. 132-136) discute algunas formas de sistematizar esta idea. Por otro lado, si los métodos distintos no proporcionan el mismo resultado, no se deduce necesariamente que no haya ninguna estructura relevante, o que la que haya no esté representada adecuadamente. Puede que la estructura de los grupos se pueda identificar convenientemente mediante un método basado en unos supuestos, pero no en otro basado en supuestos distintos.

Pueden usarse diversas variaciones en el tema de la aleatorización cuando emprendemos procedimientos de evaluación. Por ejemplo, un conjunto de datos puede ser dividido aleatoriamente en dos subconjuntos, efectuándose un análisis en cada uno de ellos para ver si son comparables. Un enfoque menos radical consiste en permutar aleatoriamente los valores de las variables entre los distintos casos, destruyendo así cualquier estructura de asociación o similaridad que pueda existir, y comparando los resultados con los de los datos reales, tanto visual como intuitivamente; por ejemplo, en términos de la estructura del dendrograma, o, quizás, usando las medidas de superposición en la pertenencia a los grupos propuestas por Gordon (1981, pp. 132-136). Aldenderfer (1982, p. 66) usa un procedimiento de aleatorización para generar una distribución lambda de Wilk, con la cual se compara la distribución observada, pues la prueba de significación clásica es inapropiada. Asigna aleatoriamente los casos a los grupos y calcula el valor lambda resultante. El procedimiento se repite diez veces, obteniéndose finalmente un valor medio con el que se compara el resultado observado.

Finalmente, otro enfoque de validación considera el grado en que la agrupación de elementos individuales en conglomerados distorsiona el esquema de similaridades o distancias entre los individuos; puede usarse también para comparar la cantidad de distorsión entre distintos métodos de agrupación. El paquete de programas informáticos CLUSTAN dispone de dos de tales medidas: la Δ (delta) de Jardine y Sibson (Jardine y Sibson, 1968) y el llamado coeficiente de correlación cofenético. Ilustraremos la segunda mediante un ejemplo: se comparará la matriz de similaridades usada en el ejemplo anterior del análisis de enlace simple con la agrupación de las similaridades que se deducen de ese análisis.

La matriz original (tabla 12.6) está reproducida a continuación (tabla 12.17); estas similaridades se designan con la notación s_{ij}. El paso siguiente es derivar la estructuración de las similaridades producida por el análisis de conglomerados; estas similaridades se designan con la notación s^*_{ij}. Los valores s^*_{ij} en-

tre cada par de variables pueden ser extraídos directamente del dendrograma (reproducido como fig. 12.9), señalando el valor del coeficiente al que las unidades se vinculan. Así, las unidades 1 y 2 están relacionadas a 0,8; 4 y 5 a 0,7; 3 a 4 y 5 a 0,6; 1 y 2 a 3, 4 y 5 a 0,5. A partir de esas cifras produciremos una nueva matriz s^*_{ij} (tabla 12.18) trazando los elementos correspondientes a esas dos matrices en un diagrama de dispersión (fig. 12.10).

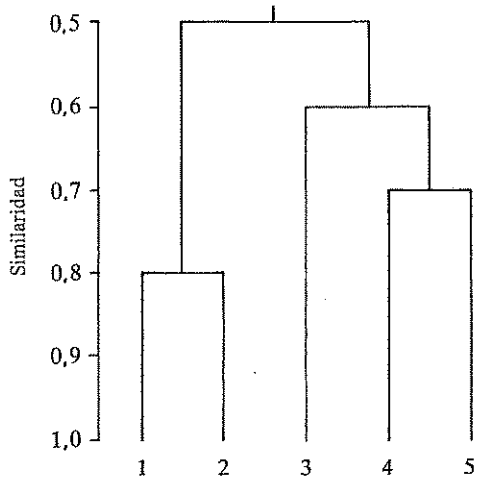


FIGURA 12.9. Dendrograma de los resultados de un análisis de conglomerados por enlace simple de la matriz de similaridades entre cinco vasijas de cerámica, que aparece en la tabla 12.17.

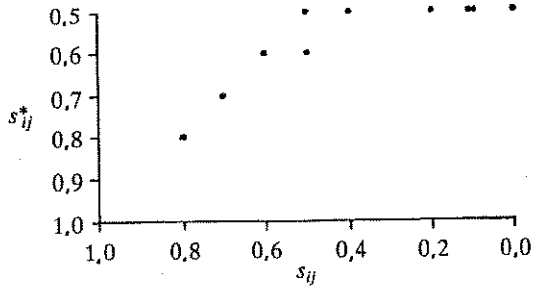


FIGURA 12.10. Diagrama de dispersión de las similaridades s^*_{ij} con relación a las similaridades s_{ij}, basado en las matrices de las tablas 12.18 y 12.17 respectivamente.

TABLA 12.17. Matriz de similaridades (s_{ij}) entre cinco vasijas de cerámica, basada en los motivos decorativos.

	1	2	3	4	5
1	1,0	0,8	0,4	0,0	0,1
2	0,8	1,0	0,5	0,1	0,2
3	0,4	0,5	1,0	0,6	0,5
4	0,0	0,1	0,6	1,0	0,7
5	0,1	0,2	0,5	0,7	1,0

Podemos obtener también el coeficiente de correlación entre las matrices basándonos en sus elementos correspondientes; se calcula de la misma forma que una correlación normal y es llamado, en este contexto, coeficiente de correlación cofenético. En este caso vale 0,44. Como se señaló antes, esta técnica puede usarse para comparar las matrices s_{ij}^* que resultan de los distintos métodos de agrupación.

TABLA 12.18. Matriz de similaridades (s_{ij}^*) entre cinco vasijas de cerámica, derivada de los enlaces en el dendrograma de la figura 12.9.

	1	2	3	4	5
1	1,0	0,8	0,5	0,5	0,5
2	0,8	1,0	0,5	0,5	0,5
3	0,5	0,5	1,0	0,6	0,6
4	0,5	0,5	0,6	1,0	0,7
5	0,5	0,5	0,6	0,7	1,0

Este ejemplo completa el tratamiento de la evaluación de los resultados de un análisis de conglomerados. Es importante tener en cuenta que las decisiones que se han de tomar antes de elegir una técnica de agrupación —basándose en la naturaleza de las variables y en las medidas de similaridad o distancia— pueden afectar al resultado. En la clasificación numérica, quizás más que en otras áreas de aplicación de los métodos cuantitativos en arqueología, es esencial tener bien claros los objetivos, la naturaleza de los datos, las propiedades de la descripción numérica y su análisis, así como los medios apropiados para evaluar los resultados.

EJERCICIOS

12.1. Efectúa un análisis de proximidades para seriar la siguiente matriz de coeficientes de similaridad entre diez conjuntos de artefactos paleolíticos.

	1	2	3	4	5	6	7	8	9	10
1	200	142	124	135	90	78	69	73	70	52
2	142	200	122	131	92	89	79	82	81	63
3	124	122	200	117	95	87	83	85	86	69
4	135	131	117	200	98	110	92	98	75	58
5	90	92	95	98	200	94	95	95	102	77
6	78	89	87	110	94	200	134	132	122	95
7	69	79	83	92	95	134	200	119	129	125
8	73	82	85	98	95	132	119	200	125	103
9	70	81	86	75	102	122	129	125	200	146
10	52	63	69	58	77	95	125	103	146	200

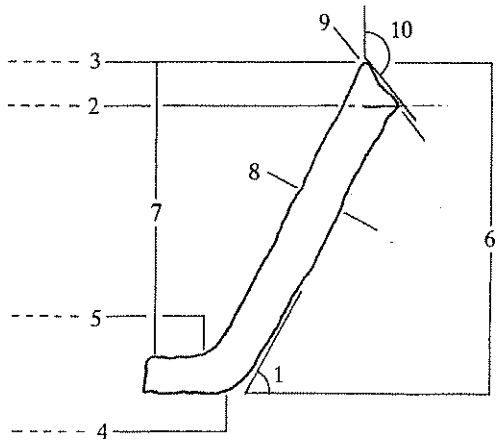
12.2. A causa de la falta de estratigrafía vertical se intenta reconstruir la cronología de un yacimiento seriando un conjunto de fosas por su contenido en cerámica. A continuación se ofrece la matriz de similaridades entre 10 de esas fosas; intenta seriarlas usando el análisis de proximidades. ¿Te parece adecuada la seriación resultante o demasiado compleja como para poder disponerla en una secuencia lineal?

Fosas	1	2	3	4	5	6	7	8	9	10
1	100	36	6	69	48	50	58	83	87	38
2	36	100	74	38	48	93	42	52	62	30
3	6	74	100	38	99	22	28	15	7	75
4	69	38	38	100	36	15	19	90	73	27
5	48	48	99	36	100	57	17	86	57	62
6	50	93	22	15	57	100	93	71	61	68
7	58	42	28	19	17	93	100	32	88	65
8	83	52	15	90	86	71	32	100	92	5
9	87	62	7	73	57	61	88	92	100	95
10	38	30	75	27	62	68	65	5	95	100

12.3. Efectúa un análisis de proximidades de la siguiente matriz de coeficientes de Robinson-Brainerd, representando el grado de similaridad entre los conjuntos cerámicos de 15 yacimientos.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	—	181	179	182	162	164	154	144	148	146	143	147	150	147	132
2	181	—	181	186	156	157	147	141	146	145	149	146	148	146	127
3	179	181	—	184	153	157	155	146	151	147	150	151	153	145	130
4	182	186	184	—	153	157	150	144	149	145	148	147	152	148	130
5	162	156	153	153	—	173	151	145	143	137	133	142	134	160	135
6	164	157	157	157	173	—	159	165	160	155	147	159	152	155	132
7	154	147	155	150	151	159	—	148	149	141	137	155	140	148	144
8	144	141	146	144	145	165	148	—	181	175	167	172	168	137	121
9	148	146	151	149	143	160	149	181	—	178	173	177	177	139	121
10	146	145	147	145	137	155	141	175	178	—	185	175	176	127	123
11	143	149	150	148	133	147	137	167	173	185	—	173	179	122	118
12	147	146	151	147	142	159	155	172	177	175	173	—	177	135	132
13	150	148	153	152	134	152	140	168	177	176	179	177	—	129	120
14	147	146	145	148	160	155	148	137	139	127	122	135	128	—	128
15	132	127	130	130	135	132	144	121	121	123	118	132	120	128	—

12.4. En la tabla de la página siguiente se muestra una serie de medidas que describen la forma de unos cuencos de borde oblicuo del período Uruk en Mesopotamia. El diagrama muestra a qué se refieren las distintas medidas (según Johnson, 1973). Efectúa un análisis de conglomerados con estos datos e intenta establecer grupos. Recuerda que efectuar un solo análisis e interpretar sus resultados es insuficiente. Has de comparar distintos métodos y usar técnicas de evaluación. ¿Puedes detectar algún problema especial en este análisis?



CLAVES: 1. Ángulo de la base; 2. Diámetro del borde (estimado en 0,5 cm); 3. Diámetro interior del borde (en 0,5 cm); 4. Diámetro de la base (en 0,5 cm); 5. Diámetro interior de la base (en 0,5 cm); 6. Altura de la pared (medida en 0,1 cm); 7. Altura interior de la pared (en 0,1 cm); 8. Grosor de la pared; 9. Grosor del borde; 10. Ángulo del borde.

Ejercicio 12.4

	1	2	3	4	5	6	7	8	9	10
1	58	160	150	80	70	73	65	108	145	128
2	57	140	130	70	65	67	62	94	111	137
3	55	175	155	70	70	71	61	107	110	137
4	58	180	170	70	65	84	80	106	121	154
5	62	195	180	80	70	86	72	108	135	150
6	60	165	160	70	65	85	78	111	130	159
7	53	180	170	80	65	85	75	120	123	148
8	68	130	120	60	50	71	65	108	104	150
9	48	150	140	70	60	70	55	133	129	165
10	58	200	190	80	75	96	84	159	141	147
11	47	210	200	85	75	79	74	114	135	163
12	60	160	150	80	70	87	80	110	121	136
13	55	180	170	80	80	88	83	109	118	160
14	65	190	165	80	75	91	79	132	169	150
15	63	190	170	75	70	89	85	137	129	155
16	67	220	210	80	75	118	105	145	138	170
17	44	170	150	80	70	58	44	103	123	154
18	63	185	170	75	80	80	74	117	139	148
19	52	160	150	60	55	75	69	109	126	148
20	62	215	200	90	85	97	81	138	128	133
21	41	175	160	65	60	70	62	110	137	151
22	47	190	170	75	80	69	58	120	129	148
23	50	185	160	70	65	94	80	126	143	152
24	55	195	180	70	65	85	80	130	129	151
25	49	195	180	70	65	77	69	124	102	148
26	58	140	120	65	60	66	54	113	143	130
27	62	170	160	65	60	90	70	94	131	137
28	55	135	120	70	65	73	64	109	102	136
29	53	170	160	70	65	78	64	123	124	135
30	60	175	160	70	60	83	70	112	142	155
31	52	140	120	70	65	73	62	116	126	145
32	59	150	140	75	70	88	76	101	126	135
33	61	140	130	70	60	92	85	116	103	152
34	56	145	130	65	60	72	65	125	134	136
35	60	175	160	75	65	93	78	111	160	130
36	53	165	160	70	60	74	65	111	62	160
37	49	165	150	80	75	75	62	129	147	154
38	60	160	140	70	65	78	66	114	146	143
39	59	170	160	70	60	91	77	138	119	146
40	57	165	160	80	63	77	60	91	124	170
41	55	170	160	80	65	70	66	140	121	149

12.5. A continuación se ofrece información acerca de varios depósitos del período II de la edad del bronce en Dinamarca (según Levy, 1982). Efectúa un análisis de conglomerados de los depósitos para descubrir si se registra o no una clara agrupación de los mismos en tipos específicos. Usa sólo la información que se refiere al contenido de los depósitos y medita cuidadosamente el problema de la codificación de los datos para el análisis de conglomerados. Recuerda de nuevo usar más de una técnica de agrupación, junto con técnicas de validación que incluyan el uso de reglas de detención allí donde fuere necesario. Examina las relaciones entre los resultados de la agrupación y *a*) la región en la que se encontró el depósito (Zelanda, Funen, Jutlandia), *b*) las circustancias del hallazgo y *c*) la categorización que hace Levy en depósitos rituales y no rituales.

Zelanda

1. Dos placas de cinturón, dos torques, cuatro anillos espirales, 33 tutuli (un tipo particular de adorno de bronce), 113 tubos para falda de cordones. Encontrado en un prado anegado, durante la labranza. Suelo de turba. Peso = 1.986 g. Depósito ritual.

2. Dos torques, dos brazaletes, tres brazaletes en espiral, 13 tutuli, un gancho de cinturón. Encontrado en un pantano. Peso estimado = 735 g.

3. Hacha (*palstaves*) de guerra, una punta de espada. Encontrado en un pantano. Peso estimado = 750 g. Depósito ritual.

4. Dos placas de cinturón, dos torques, algunos brazaletes, una hoz. Encontrado en un pantano. Peso estimado = 868 g. Depósito ritual.

5. Dos hachas (*palstaves*) planas, un cuchillo, una placa de cinturón, una placa de bronce (fragmentación anterior a la deposición). Encontrado en un campo. Peso = = 1.211 g. Depósito no ritual.

6. Un torques, un brazalete en espiral, dos brazaletes, aguja de una fíbula. Encontrado en un pantano. Peso = 160 g. Depósito ritual.

7. Tres puntas de lanza, un hacha plana, una hoz. Encontrado en un campo. Peso estimado = 955 g. Depósito no ritual.

8. Tres placas de cinturón, cuatro grandes tutuli, 27 pequeños tutuli, cuatro anillos en espiral, un torques. Han desaparecido actualmente unos fragmentos de hueso y un molino de piedra. Encontrado en un pantano a unos 0,5 metros de profundidad. Peso = 1.530 g. (estimado parcialmente debido a la fragmentación). Depósito ritual.

9. Noventa y cuatro hachas (muchas de ellas sin terminar), un punzón, mango de una espada, lámina de un cuchillo, sesenta puntas de lanza, bloque de metal en bruto (la mayoría de los objetos se han perdido hoy en día). Encontrado en un estrato azulado de arcilla, en una turbera. Peso no estimado. Depósito no ritual.

10. Dos collares, fragmentos de brazalete en espiral. Encontrados juntos en un campo. Peso = 400 g (estimado parcialmente debido a la fragmentación posdeposicional). Depósito ritual.

11. Tres puntas de lanza, fragmentos de una lámina de espada, cincel, útil no identificado con una lámina puntiaguda plana y talón. Encontrado en una zanja de irrigación. Peso = 853 g. Depósito no ritual (?).

12. Dos brazaletes. Hallados juntos en un campo. Peso estimado = 51 g. Depósito ritual.

13. Un punzón, dos hachas planas. Encontrado en un campo. Peso = 900 g. Depósito no ritual.

14. Fragmento de la empuñadura de una espada. Fragmento de un brazalete (posiblemente procedente de la Alemania septentrional), seis hachas planas, fragmentos de hachas, gancho de cinturón, metal en bruto (fragmentación anterior al depósito). Encontrado en un campo. Peso = 4.530 g. Depósito no ritual.

15. Dos puntas de lanza. Encontradas en un pantano. Peso = 227 g. Depósito ritual.

16. Dos espadas con empuñadura de metal. Encontradas juntas en un campo. Peso = 2.446 g. Depósito ritual.

17. Dos torques, una placa de cinturón, un brazalete, fragmentos de brazalete en espiral, un tutulus redondeado, dos hoces, un hacha plana, tres punzones. Encontrado en un hogar con los torques rodeando los objetos de menor tamaño y la placa de cinturón cubriéndolo todo. Peso = 1.150 g (parcialmente estimado debido a la fragmentación posdeposicional). Depósito ritual.

18. Una placa de cinturón, dos brazaletes espirales, un cuchillo. Encontrado en un pantano. Peso = 1.189 g. Depósito ritual.

19. Una espada con empuñadura metálica, una lámina de espada y la extremidad correspondiente del mango. Hallado en un pantano (posibilidad de que no sea un hallazgo cerrado). Peso = 1.169 g. Depósito ritual.

20. Dos hachas planas. Halladas en un pantano de turba a 1,26 metros de profundidad. Peso = 760 g. Depósito ritual.

21. Punta de espada y placa de cinturón. Hallado en un pantano, en el extremo de un delgado estrato de turba. Dudas acerca de su asociación. Peso = 310 g. Depósito ritual.

22. Tres torques, tres placas de cinturón, 21 tutuli, 7-8 anillos en espiral, 3-4 brazaletes en espiral, una hoja de sierra, una hoz. Encontrado en un campo por debajo de un estrato de grava. Peso = 1.760 g (estimado parcialmente debido a la fragmentación posdeposicional). Depósito ritual.

22a. Molde hueco, estatuilla de caballo unida a un disco, cubierto de oro en un lado; ambos fijados en un carrito de seis ruedas. Hallado en un pantano, aparentemente roto después de la deposición. Peso = 4.190 g (según bibliografía; 15 g son de oro). Depósito ritual.

23. Espada con empuñadura de metal, dos láminas de espada, un pomo de espada. Hallado junto a un antiguo curso de agua, en posición horizontal, señalando al suroeste. Peso = 1.457 g. Depósito ritual.

24. Cuatro puntas de lanza y seis hoces (u hojas de sierra). Halladas en un pantano, a 1,25 metros de profundidad, lejos de la tierra seca. Peso = 365 g. Depósito ritual.

25. Dos placas de cinturón, un anzuelo. Hallados juntos en un pantano, a 2 metros de profundidad. Peso = 229 g. Depósito ritual.

26. Una espada con empuñadura de metal, dos hachas planas, una punta de lanza, un brazalete o *anklering*. Hallados juntos en un pantano. Peso = 1.888 g. Depósito ritual.

27. Tres espadas con empuñadura de metal, tres láminas de espada, dos pomos de espada. Hallados juntos, en horizontal, señalando las empuñaduras hacia el este, en el extremo de un estrato de color ocre, justo encima de un estrato de creta; esto señala la presencia antigua de un manantial. Peso = 4.157 g. Depósito ritual.

28. Dos collares. Hallados en un pantano. Peso = 322 g. Depósito ritual.

29. Dos placas de cinturón, un collar, dos tutuli, dos láminas de sierra u hoz, una lámina de cuchillo. Hallado en un suelo calcáreo anegado, a 2 metros de profundidad. Peso = 694 g. Depósito ritual.

30. Un hacha de guerra, un punzón grande. Hallado en un campo inundado durante los trabajos de drenaje, a 1,25 metros de profundidad. Peso = 1.122 g. Depósito ritual.

31. Tres placas de cinturón, 17 tutuli, un punzón, una hoz, al menos cuatro láminas de sierra u hoz. Hallado en el extremo de un estrato de turba, sobre arcilla. Peso = 838 g. Depósito ritual.

32. Cinco hachas planas. Halladas en un jardín, filos hacia el norte. Peso = 2.233 g. Depósito no ritual.

Funen

33. Dieciséis puntas de lanza, 15 hoces, dos hachas planas, dos fragmentos de lámina de espada, dos cuchillos, una varilla de metal. Hallado por una excavadora en el subsuelo arenoso de un campo. Peso no estimado. Depósito no ritual.

34. Cinco puntas de lanza, ocho sierras u hoces, un hacha de guerra, un cuchillo, un cincel. Hallado en un pantano. Peso estimado = 1.600 g. Depósito ritual (?).

35. Dos espadas macizas con empuñadura de metal, una lámina de puñal. Hallado en un pantano. Peso = 2.398 g. Depósito ritual.

36. Tres placas de cinturón, un brazalete, una punta de lanza, dos hachas planas, un cincel, cuatro láminas de sierra u hoz. Hallado en un prado pantanoso. Peso = 1.575 g. Depósito ritual.

37. Tres placas de cinturón, siete tutuli, un torques. Hallado en un pantano. Peso estimado = 759 g. Depósito ritual.

38. Dos torques estriados. Hallados en una zanja. Peso = 77 g. Depósito ritual.

39. Tres espadas macizas con empuñadura de metal. Hallado en la depresión al borde de un curso de agua. Peso = 2.818 g (peso de una de las espadas estimado). Depósito ritual.

40. Tres puntas de lanza, dos sierras u hoces. Hallado en un pantano. Peso estimado = 450 g. Depósito ritual.

41. Siete puntas de lanza. Hallado en el borde de un pantano. Peso = 1.200 g (parcialmente estimado debido a la fragmentación posdeposicional). Depósito ritual.

Jutlandia

42. Un hacha plana, dos hoces. Hallado en un campo. Peso = 513 g (parcialmente estimado, el hacha ha desaparecido). Depósito no ritual.

43. Un torques, cuatro brazaletes distintos. Hallados en un pantano. Peso = 96 g. Depósito ritual.

44. Dos láminas de puñal, una placa de cinturón, un punzón (algo fragmentado, posiblemente en época actual). Hallado en un campo. Peso = 327 g. Depósito ritual (?).

45. Diecisiete tutuli, una aguja, un punzón, cuentas de ámbar. Hallado en un pantano. Peso = 148 g. Depósito ritual.

46. Dos hachas planas, un martillo. Hallado entre piedras. Peso no estimado. Depósito no ritual.

47. *Marstrup*. Dos hachas planas, no usadas. Hallado en un campo. Peso = 425 g. Depósito no ritual.

48. Un torques, una placa de cinturón, seis tutuli, tubos para una falda de cordones. Hallado en un pantano. Peso estimado = 416 g. Depósito ritual.

49. Una placa de cinturón, cuatro tutuli. Hallado cerca de un acantilado. Peso = = 200 g (parcialmente estimado, debido a fragmentación posdeposicional). Depósito ritual.

50. Dos brazaletes en espiral. Hallado en una zanja. Peso = 129 g. Depósito ritual.

51. Una espada maciza con empuñadura de metal, hacha maciza de cubo. Hallado en un pantano, a 1 metro de profundidad. Peso = 2.735 g. Depósito ritual.

52. Un punzón, siete hoces, fragmento de la lámina de una espada (rota en la antigüedad). Hallado junto a una gran piedra. Peso = 804 g. Depósito no ritual.

53. Dos torques fragmentados y aproximadamente dos kilos de ámbar sin trabajar, en muchos trozos. Hallado en un pequeño túmulo natural. Peso del metal = 88 g. Depósito no ritual.

54. Cuatro hachas planas sin usar. Hallado en una colina debajo de una piedra. Peso = 1.403 g. Depósito no ritual (?).

55. Siete hachas planas, un hacha de guerra. Hallado en un pantano. Peso = 2.970 g. Depósito ritual.

56. Cinco tutuli, cinco brazaletes distintos, un torques retorcido, un punzón. Hallado en un campo. Peso estimado = 342 g. Depósito ritual.

57. Dos hachas planas, un punzón, un cincel, el extremo del mango de una espada, Hallado cerca de la superficie de un campo, junto a los restos de un hogar; no hay evidencias claras de asociación entre ambos. Peso = 1.596 g. Depósito no ritual.

58. Punzón macizo, hacha de guerra. Hallado debajo de una gran piedra en un suelo pizarroso. Peso estimado = 1.000 g. Depósito ritual.

59. Cuatro hachas planas, un punzón, un cincel, hoces, lámina de puñal, dos torques, dos placas de cinturón, 18 tutuli, brazaletes en espiral. Tres puntas de espada. Hallado en el borde de un pantano. Peso = 3.331 g. Depósito ritual.

60. Diecinueve hachas planas, dos puntas de lanza, fragmentos de otra punta de lanza. Hallados juntos en la superficie de un campo. Peso = 7.911 g (parcialmente estimado). Depósito no ritual.

61. Una placa de cinturón, dos puntas de lanza. Hallado en un pantano. Peso estimado = 475 g (algunos de los objetos han desaparecido). Depósito ritual.

62. Una espada, un hacha de cubo, seis puntas de lanza, punzón, hacha. Hallado al lado de un túmulo funerario, las puntas de la lanza hincadas verticalmente en el suelo. Peso = 3.280 g. Depósito ritual.

63. Cuatro hachas planas, dos hoces. Hallado en un campo. Peso = 1.677 g. Depósito no ritual.

12.6. Invierte el análisis de los depósitos y en vez de agrupar los depósitos efectúa un análisis de conglomerados de los artefactos, según los depósitos en que aparecen. ¿Se aprecia alguna asociación particularmente marcada?

13. SIMPLIFICACIÓN DE ESPACIOS COMPLEJOS: LA FUNCIÓN DEL ANÁLISIS MULTIVARIANTE

En el capítulo anterior se explicó cómo describir objetos en términos de unas variables, buscando a continuación la estructura adoptada por las similaridades o distancias entre los objetos según su valor respectivo en las variables usadas en su descripción. Se empleaban métodos de análisis de conglomerados para agrupar objetos similares en el mismo conglomerado; vimos distintos métodos que mostraban maneras diferentes de definir un grupo. Al mismo tiempo se señaló que, hasta cierto punto, esos métodos tendían a imponer su propia estructura en los datos, lo cual era un problema que no debía dejarse de lado.

Los métodos que describiremos en este capítulo proceden de un enfoque distinto, ya citado brevemente en el capítulo anterior al mencionar la *ordenación*. Se relaciona con muchos de los conceptos vistos en la presentación del análisis de conglomerados, así como con ideas derivadas del análisis de regresión.

En el caso de la regresión bivariada simple se usaron diagramas de dispersión para ver si existía alguna tendencia en la distribución de las observaciones, si bien podíamos usarla también para ver qué puntos eran similares a otros; al mismo tiempo hubiésemos podido, de haberlo querido, consignar si había o no indicación de que las observaciones estaban dispuestas en grupos. Las variables constituían los ejes del diagrama de dispersión. Nuestro examen de la regresión múltiple mostró, entre otras cosas, que con más de tres variables era imposible representar los datos por medio de un diagrama de dispersión. Si queríamos hacerlo había que dibujarlo de dos en dos variables, o como mucho de tres en tres. Esto puede servir en ocasiones, aunque no proporciona una imagen global: no podemos examinar las tendencias globales para una regresión múltiple, ni las distancias entre los puntos para clasificaciones o agrupaciones.

El propósito de los métodos de ordenación es comprimir la información contenida en un gran número de variables en un número mucho más reducido de nuevas variables, idealmente sólo dos o tres. Así se pueden producir diagramas de dispersión con los datos expresados en esas nuevas variables, que permiten, por tanto, la visualización de grandes cantidades de información. Exa-

minando esos diagramas veremos si hay o no grupos o conglomerados en los datos; los objetos no están forzados en una estructura de agrupación determinada meramente por la técnica de agrupación utilizada. Además, como veremos, el procedimiento de obtener las nuevas variables para crear los diagramas de dispersión produce a su vez una información de gran interés: en ciertos casos, constituirá el principal objetivo.

Un ejemplo arqueológico puede ser útil. Supongamos que estamos analizando las tumbas de un cementerio y que en el curso de nuestro estudio decidimos calcular la matriz de similaridades entre las tumbas por medio de alguno de los métodos descritos en el capítulo anterior, y basándonos en un conjunto de variables descriptivas relevante para la cuestión que queremos investigar. Nos interesa analizar la estructura subyacente a esa matriz. El análisis de conglomerados puede ser útil, pero si el número de tumbas es grande el dendrograma resultante será demasiado confuso; además aún no hemos resuelto el problema de la imposición de una estructura en los datos por parte del método de agrupación elegido. En estas circunstancias, la ordenación de las observaciones se hará del modo siguiente.

Representamos las tumbas como puntos en un espacio, de forma que las similaridades entre las tumbas estén representadas por las distancias entre los puntos. Para representar las relaciones convenientemente, necesitaríamos un espacio de muchas dimensiones. Dentro de este espacio, los puntos no estarán dispersos por igual en todas las direcciones, sino que pueden aparecer concentrados en un espacio muy reducido en una de las direcciones y dispersos en un espacio mucho mayor en otras. Es posible definir la orientación de esas distintas direcciones o ejes, así como la longitud de la distribución de los puntos en ellos. Una vez establecidas la orientación y la longitud del eje más largo, podemos definir el eje que se dirige hacia la más próxima de las mayores partes de la nube de puntos, siempre y cuando se disponga ortogonalmente al primero de los ejes; también podemos calcular la longitud de este segundo eje. Es posible continuar el procedimiento usando todas las dimensiones independientes en el espacio que hemos definido. A menudo estos ejes permiten una interpretación sustantiva, basada en los datos de los cuales derivan. En el caso de las tumbas, por ejemplo, puede que los datos estén ampliamente distribuidos a lo largo de un eje que represente las diferencias cronológicas entre tumbas, de forma que las tumbas más antiguas (según evidencias independientes) son muy distintas de las más tardías (también afirmado a partir de evidencias independientes). Podemos establecer las coordenadas de los puntos (en este caso, tumbas) en relación a esos ejes, usando las nuevas coordenadas para trazar diagramas de dispersión, los cuales serán interpretados según la agrupación de las tumbas y la naturaleza de los ejes, tal y como se acaba de explicar. Antes que fiarnos de la impresión visual en el diagrama de dispersión, puede ser necesario efectuar un análisis de conglomerados de las coordenadas de los objetos en los nuevos ejes, y no de los datos originales.

La aplicación de esta clase de procedimientos al análisis de cementerios (por ejemplo, Shennan, 1983) y muchos otros tipos de datos arqueológicos ha demostrado su utilidad en numerosos casos, pues proporciona una manera de comprender esquemas complejos de variación que, de otra manera, no serían fácilmente asimilables.

Este ejemplo pone de manifiesto los efectos conjuntos de la ordenación. Su relación con las similaridades o distancias y la estructura subyacente a ellas es común al análisis de conglomerados. La idea de buscar las tendencias en la variación es algo que ya hemos visto en la regresión. Sin embargo, hay diferencias importantes entre la ordenación y la regresión, especialmente una: en el análisis de la regresión, el objetivo es modelizar y explicar la variación en la variable dependiente en términos del efecto de una o varias independientes. En el ejemplo anterior de las tumbas, las variaciones originales no reaparecían en el análisis una vez que se calcularon las similaridades. No obstante, tal y como veremos, en los métodos de ordenación que analizan directamente las variables originales no se establece diferencia alguna entre variables dependientes y variables independientes. Simplemente se obtiene una medida de la correlación o covariación entre todas las variables entre sí, analizando la matriz que resulta de ello.

En arqueología se ha optado claramente por los métodos de ordenación antes que por la regresión múltiple, y ello se debe a dos razones. En primer lugar, muchos de los datos arqueológicos suelen ser muy complejos y no resulta obvio en absoluto qué variables han de definirse como dependientes y qué variables han de definirse como independientes. Este es el caso, por ejemplo, cuando tratamos con la descripción cuantitativa de la forma de una vasija o de la punta de un proyectil. Pueden haber correlaciones entre las distintas medidas, cuya comprensión será muy útil para agrupar los objetos en cuestión y vital para explicar por qué varían de la forma que lo hacen. Por otro lado, el tratamiento de una de las variables como dependiente no corresponde a la realidad de la situación; están interrelacionadas unas con otras.

Una segunda razón, quizás menos fiable, de por qué los arqueólogos han preferido los métodos de ordenación es precisamente porque permiten analizar los datos y ver qué estructura emerge sin que sea necesario proponer hipótesis y modelos previos. Tal enfoque corresponde a la filosofía del análisis de datos exploratorio, en el que se ha insistido a lo largo de este libro, pero que, indiscutiblemente, tiene sus riesgos (véase Speth y Johnson, 1976); parece reflejar una tendencia profundamente enraizada entre los arqueólogos que prefieren interpretar estructuras antes que desarrollar y comprobar hipótesis.

ANÁLISIS MULTIVARIANTE

Hasta aquí hemos hablado de la ordenación y la simplificación de espacios complejos en términos muy generales, sin distinguir entre los diferentes méto-

dos. Es necesario que a partir de ahora seamos un poco más específicos y proporcionemos alguna indicación de las distintas técnicas y de la manera que las trataremos en este capítulo. En su mayor parte proceden del área de la estadística conocida como *análisis multivariante*. Difiere de las técnicas usadas en el análisis de conglomerados en que estas últimas son, en muchos sentidos, técnicas heurísticas *ad hoc* sin un fundamento teórico claro, mientras que los análisis multivariantes tienen una base teórica bien fundamentada en las matemáticas y la estadística. Como ya hemos tenido ocasión de señalar (capítulo 11), las matemáticas que fundamentan estos métodos suelen ser muy complejas y, por ese motivo, suelen estar consideradas como algo muy profundo y misterioso. Si bien su conocimiento preciso y el consejo de un experto son esenciales para su aplicación, el propósito de este capítulo es mostrar que, en esencia, son fácilmente comprensibles, hasta el extremo en que debería ser posible a todo el mundo entender y evaluar (con propiedad) los análisis publicados. Es importante una buena comprensión de estas técnicas por lo difundido de sus aplicaciones y debido también a los recientes debates sobre ejemplos particulares de esos análisis.

En las páginas siguientes se presenta una descripción detallada del método del *análisis de componentes principales* y, en menor grado, el método relacionado del *análisis factorial*. Una vez explicados estaremos en disposición de presentar una exposición algo más breve y general del *análisis de coordenadas principales*, las *escalas multidimensionales no métricas* y el *análisis de correspondencias*.

Componentes principales y análisis factorial

El análisis factorial se desarrolló en el dominio de la psicología en los años treinta. Su propósito original fue extraer las medidas fundamentales de la inteligencia a partir de los resultados de las pruebas clásicas.

Se consideraba que ninguna prueba sencilla proporcionaba una medida adecuada de la inteligencia. Los resultados que obtenían en ellas los individuos estaban relacionados con su capacidad mental, pero estaban influidos por las diferencias entre ellos en materia de educación, entorno cultural y las circunstancias de la prueba. Los psicólogos creían que el análisis factorial era capaz de extraer el factor común de la inteligencia a partir de los resultados de los individuos en varias pruebas, aunque ninguna prueba, por sí sola, fuese capaz de medir la inteligencia directamente. Hoy en día, el análisis de componentes principales y el análisis factorial se aplican en una gran variedad de disciplinas distintas, con el propósito de analizar sus datos respectivos.

La idea general es extraer algo que tengan en común diversas variables. Si pudiésemos aislar una dimensión común subyacente en nuestras variables iniciales, seríamos capaces de sugerir su significado en términos de nuestro problema, tal y como hicieron los psicólogos con el factor que se derivaba de las pruebas de inteligencia. Además, si hay un factor común subyacente en la variación de todo un conjunto de variables, podremos olvidarnos de la variación en las variables originales y limitarnos a ver qué es lo que tienen en común. Si examinamos a continuación los resultados de nuestros casos en el reducido número de factores comunes subyacentes en los datos originales, los usaremos como criterio para su ordenación en un espacio de dimensiones reducidas. De hecho, estos métodos no sólo permiten extraer las dimensiones de la variación, sino también conocer su importancia.

Esta presentación se refiere principalmente al análisis factorial. La distinción entre éste y el análisis de componentes principales se expone a continuación. Se dará más relieve al análisis de componentes principales porque es la técnica más simple.

Introducción al análisis de componentes principales

Ya se ha dicho que las matemáticas implicadas son demasiado complejas, por lo que se prescindirá aquí de una presentación matemáticamente rigurosa de esta técnica (véase, por ejemplo, Morrison, 1967). En estas circunstancias, la mejor forma de presentar una exposición intuitiva es mediante el uso de imágenes y de la geometría. La exposición que sigue se basa en la lúcida explicación de Johnston (1978), dirigida a lectores con un nivel similar a los que se dirige ese libro; una buena alternativa es Davis (1973).

En el caso del análisis de componentes principales (y factorial), el punto de partida es la covariación entre las variables. Si un conjunto de variables posee algún factor común subyacente, se deduce que los valores de esas variables estarán correlacionados entre sí. El factor común puede ser visto, entonces, como el promedio del grupo de variables; cuanto más relacionadas estén, más fuerte será el factor común y más significativo por sí mismo, como sustituto de las variables originales.

Para ver cómo funciona el análisis de componentes principales hemos de volver a la manera en que calculábamos la covariación entre dos variables. Vimos en el capítulo 9 que la covariación, en un sentido técnico, estaba definida por $\sum (x - \bar{x})(y - \bar{y})$; si dividimos esa expresión por el tamaño de la muestra, obtendremos la covariación media o covarianza. Si las variables han sido estandarizadas, esto es, transformadas en puntuaciones Z, en las que los valores se expresan en la cantidad de unidades de desviación típica desde la media (véase el capítulo 8), entonces el valor de la covarianza entre dos variables cualesquiera se estandarizará automáticamente y corresponderá al coeficiente de correlación entre las dos variables; se desprende de esta transformación que las varianzas de las variables individuales también están estandarizadas, en un valor de 1,0. Asumiremos, por tanto, que las relaciones entre variables están ex-

presadas como coeficientes de correlación (mejor que covarianzas), aunque tal estandarización no sea necesariamente algo que deseemos adoptar en un análisis real (este punto se argumentará más adelante, p. 262; véanse también Davis, 1973; Everitt y Dunn, 1983, pp. 42 y 47).

Para desarrollar una presentación geométrica del análisis de componentes principales, necesitamos en primer lugar un método geométrico para representar las correlaciones. Si imaginamos que nuestras variables son vectores de idéntica longitud con un origen común, podremos usar la distancia angular entre ellos para representar sus relaciones. Visualmente se entenderá mejor (fig. 13.1). Tenemos cuatro variables, cada una representada como una línea direccionada, y todas ellas con el mismo origen. Según nuestra representación gráfica y las convenciones para su interpretación, x_1 y x_2 están fuertemente correlacionadas, ninguna de las dos se relaciona estrechamente con x_3, si bien x_2 está más próxima a x_3 que a x_1; finalmente, x_4 es más o menos el opuesto de x_1 y de x_2, con muy poca relación con x_3.

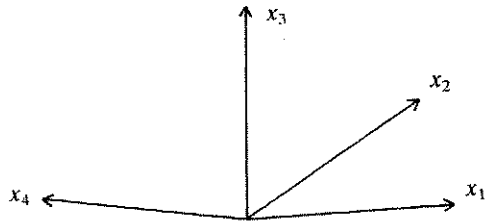


FIGURA 13.1. Representación geométrica de las correlaciones entre cuatro variables.

Esta representación es muy útil porque el tamaño de los ángulos puede relacionarse directamente con los valores de los coeficientes de correlación, ya que corresponde a los cosenos de los ángulos en cuestión. Así, siguiendo la convención gráfica, cuando dos variables estén perfectamente correlacionadas, el ángulo entre ellas será cero: los vectores se superpondrán (fig. 13.2). Obviamente, en este caso, el valor del coeficiente de correlación es 1,0; igualmente, el coseno de un ángulo de cero grados es igual a 1,0.

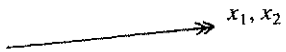


FIGURA 13.2. Representación geométrica de dos variables perfectamente correlacionadas.

Nuevamente, cuando dos variables son diametralmente opuestas, representaremos el ángulo entre ellas como uno de 180° (fig. 13.3). El valor del coeficiente de correlación será $-1{,}0$: el coseno del ángulo de 180° es -1.

Un ángulo de 90° tiene un coseno de 0,0 (fig. 13.4). En estas circunstancias,

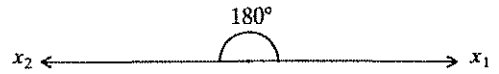


FIGURA 13.3. Representación geométrica de dos variables que muestran una correlación inversa perfecta entre ellas.

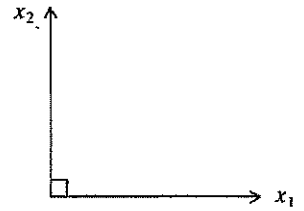


FIGURA 13.4. Representación geométrica de dos variables no correlacionadas.

la correlación entre x_1 y x_2 es también cero, de forma que podremos representar dos variables sin relación alguna entre sí por medio de dos vectores en ángulo recto; se dice entonces que esas variables son *ortogonales*.

Al tratar con dos variables siempre podemos representar sobre el papel la correlación entre ellas por medio de la distancia angular, es decir, en dos dimensiones. Antes hemos visto un diagrama que representaba las correlaciones entre cuatro variables de forma bidimensional; desgraciadamente, eso no siempre es posible. Imaginemos un caso en el que las cuatro variables no estuviesen correlacionadas entre sí; en otras palabras, que cada una tuviese correlación 0 con las demás. Dibujaríamos las dos primeras correctamente (fig. 13.5), pero que intente ahora el lector añadir una tercera. La figura 13.6 es obviamente errónea, aunque x_3 tenga aquí también una correlación 0 (un ángulo de 90°, por consiguiente) con x_2, se sitúa a 180° de x_1, lo cual indicaría una correlación negativa perfecta de $-1{,}0$. La única forma de situar correctamente el tercer vector es saliéndose de la hoja de papel, aunque pueda obtenerse también una representación distorsionada (fig. 13.7): añadir una cuarta variable ortogonal es imposible, pues las relaciones entre todas las variables exigen un espacio tetradimensional.

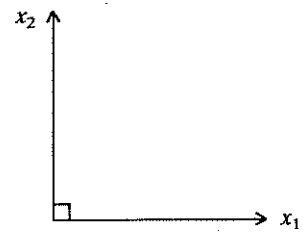


FIGURA 13.5. Dos variables no correlacionadas.

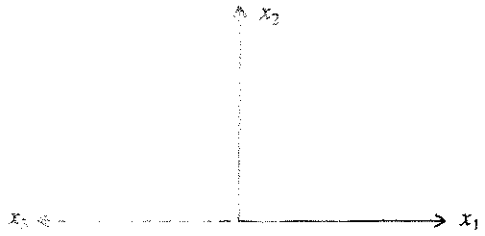


FIGURA 13.6. Intento fallido de representar una tercera variable no correlacionada en una superficie bidimensional.

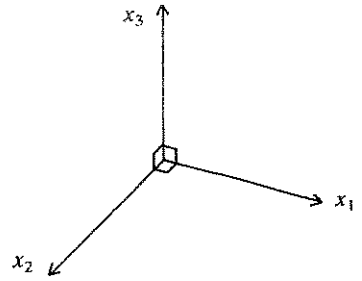


FIGURA 13.7. Representación geométrica de tres variables no correlacionadas.

En general, la cantidad máxima de dimensiones requeridas para representar las correlaciones entre una cantidad específica de variables está dada por la cantidad de variables, aunque puede ser menos. En el caso opuesto, si todas las variables están correlacionadas entre sí, sólo necesitaremos una única dimensión.

En el análisis de componentes principales empezamos con la matriz de coeficientes de correlación (o covarianzas) entre nuestras variables; el objetivo es generar a partir de ellas un nuevo conjunto de variables que no estén correlacionadas entre sí. La manera concreta en que este hecho se relaciona con nuestro propio objetivo, la definición de las dimensiones subyacentes en la variación de nuestros datos, con el fin de poder presentar diagramas de dispersión en dos dimensiones que resuman la información procedente de diez variables, se pondrá de manifiesto cuando examinemos en detalle un ejemplo arqueológico. Sin embargo, podemos adelantar que si es posible representar correctamente las relaciones entre diez variables con ayuda de tan sólo dos dimensiones, estaremos sustituyendo las diez primeras por las dos nuevas variables ortogonales, *las cuales contienen toda la información original.*

La idea no es muy distinta de las clases de resúmenes estadísticos que hemos ido viendo. Podemos tener una gran cantidad de valores de una variable que adopten una distribución normal. Dada su forma, una vez que conozcamos su media y su desviación típica, es mucho lo que explicaremos acerca de

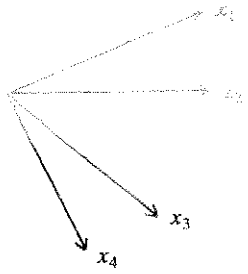


FIGURA 13.8. Representación geométrica de las correlaciones entre cuatro variables.

ésta, basándonos exclusivamente en dos resúmenes estadísticos, sin tener que recurrir a los datos originales.

En este caso, en lugar de obtener un valor promedio, calculamos una variable promedio. Será una variable nueva, artificial, de la misma manera que la media raramente coincide exactamente con cualquiera de las cifras en una distribución. Será también la variable más próxima, globalmente, a todas las variables originales; de nuevo, un concepto muy semejante al de la media de una distribución. Aquí definiremos la proximidad en términos de distancia angular. La variable más próxima globalmente a todas las originales se situará en el punto en que la suma de los ángulos entre ella y las demás sea lo más reducida posible.

Veámoslo por medio de un ejemplo sencillo (fig. 13.8). Tenemos un diagrama que representa las relaciones entre cuatro variables, cuyas correlaciones están representadas por los ángulos entre ellas. Obviamente, la variable promedio que las resuma se situará en algún lugar entre x_2 y x_3. La exactitud de esos promedios en la representación de las variables originales es una cuestión relevante que podemos tratar aquí. Ya lo calcularemos más adelante; por ahora, la cuestión inmediata es cómo descubrir precisamente dónde está ese promedio.

Lo primero que hay que hacer es señalar los valores exactos de todos los ángulos y a continuación los cosenos/correlaciones correspondientes (tablas 13.1-13.2). Habiendo hecho esto, obtenemos la suma total de correlaciones para cada variable (tabla 13.2); recuérdese que la mayor de las sumas de las correlaciones corresponde a la menor de las sumas de los ángulos. Como esperába-

TABLA 13.1. Ángulos entre las variables que aparecen en la figura 13.8.

	x_1	x_2	x_3	x_4
x_1	0	22	61	84
x_2	22	0	40	62
x_3	61	40	0	24
x_4	84	62	24	0

TABLA 13.2. Correlaciones entre las variables que aparecen en la figura 13.8.

	x_1	x_2	x_3	x_4
x_1	1,000	0,927	0,485	0,105
x_2	0,927	1,000	0,766	0,469
x_3	0,485	0,766	1,000	0,914
x_4	0,105	0,469	0,914	1,000
Suma	2;517	3,162	3,165	2,488

mos, la mayor de todas es la existente para x_2 y x_3, que estarán, por tanto, próximas a la media. No obstante, aún no hemos conseguido la situación de esa media. Esto requiere algunos pasos más.

La cantidad total de entradas en la matriz es de 16, la cantidad de variables al cuadrado. Si cada correlación tuviese un valor de 1,0, la suma total de correlaciones en la tabla sería 16,0. En este caso, naturalmente, no lo es: la suma total de correlaciones se calcula efectuando las sumas separadas para cada variable: 2,517 + 3,162 + 3,165 + 2,488 = 11,332.

Volviendo a nuestro ejemplo hipotético por un momento, si la suma total de correlaciones en la tabla fuese de 16,0, el valor máximo posible de cualquier variable única sería de 4,0, la raíz cuadrada de 16,0. Aquí igualmente la suma total de correlaciones es de 11,332; la suma total posible para cualquiera de las variables es $\sqrt{11,332}$, que es 3,366. Esta es la variable con la mayor correlación global con todas las otras variables en términos de distancia angular. En otras palabras, se trata de la variable promedio que estábamos buscando, llamada el primer *componente principal*. Lo que aún no sabemos es dónde se sitúa este componente en relación con las otras variables.

Supongamos que una de las variables originales de este caso en particular coincidió con la variable promedio o primer componente principal, es decir, el ángulo entre ellas era de 0°. También tendría una suma de correlaciones de 3,366 y la proporción correlación/coseno con el componente sería de 1,0; es decir, sería la proporción entre la suma de correlaciones para la variable original y la suma de correlaciones del componente principal; aquí 3,366/3,366 = 1,0, que corresponde a cero grados.

Sucederá lo mismo aunque no coincidan las variables originales con el componente: si dividimos la suma de correlaciones de una variable por la suma de correlaciones del componente, el resultado será la correlación entre los dos, que puede convertirse en un ángulo mediante el coseno. Efectuemos esta operación en nuestro ejemplo (tabla 13.3), y coloquemos el componente de nuestro diagrama original de las relaciones entre las variables (fig. 13.9). Al encontrar este componente hemos obtenido una variable promedio de las otras cuatro, al margen, por el momento, de su relevancia e interés como resumen efectivo de las variables originales.

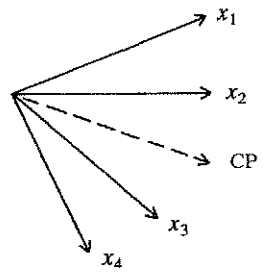


FIGURA 13.9. Representación geométrica de la correlación entre cuatro variables, con el primer componente principal añadido.

TABLA 13.3. Correlaciones y ángulos entre las cuatro variables originales y el primer componente principal. Suma total de correlaciones (ST) = 11,332; $\sqrt{ST}$ = 3,366.

	x_1	x_2	x_3	x_4
Suma	2,517	3,162	3,165	2,488
Suma/$\sqrt{ST}$	0,748	0,939	0,940	0,739
Ángulo	42°	20°	20°	42°

El método funciona de la misma forma en el caso en que haya una correlación negativa importante, como podemos ilustrar brevemente con ayuda del siguiente ejemplo, en el que x_1 y x_2 tienen una elevada correlación mutua, y una fuerte correlación negativa con x_3 (fig. 13.10).

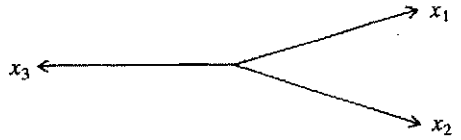


FIGURA 13.10. Representación geométrica de las correlaciones entre tres variables.

Los ángulos aparecen en la tabla 13.4 y las correlaciones/cosenos en la tabla 13.5. En la figura 13.11 se representa la ubicación del componente. En este caso, pues, x_1 y x_2 tienen una fuerte correlación positiva con el componente, y x_3 también una correlación fuerte pero negativa.

TABLA 13.4. Ángulos entre las variables que aparecen en la figura 13.10.

	x_1	x_2	x_3
x_1	0	34	164
x_2	34	0	162
x_3	164	162	0

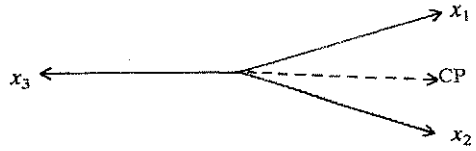


FIGURA 13.11. Representación geométrica de las correlaciones entre tres variables, con el primer componente principal (CP) añadido.

TABLA 13.5. Correlaciones entre las variables que aparecen en la figura 13.10. ST = 0,884; $\sqrt{ST}$ = 0,913.

	x_1	x_2	x_3
x_1	1,000	0,829	−0,961
x_2	0,829	1,000	−0,951
x_3	−0,961	−0,951	1,000
Suma	0,868	0,878	−0,912
Suma/$\sqrt{ST}$	0,950	0,962	−0,999
Ángulo	18°	16°	177°

Estos ejemplos debieran proporcionar al lector una impresión general de lo que es un componente principal, pues muestran una forma de calcularlos. Es preciso añadir que no es así como se obtienen en la práctica mediante los programas informáticos, los cuales llegan a los mismos resultados por medios distintos.

Llegados a este punto, desearía volver a examinar los resultados del primer ejemplo para ver qué más podemos decir, una vez colocado en su sitio el componente principal. Tal y como vimos, el ángulo entre el componente y las variables originales se obtuvo dividiendo la suma de correlaciones de una variable en particular por la raíz cuadrada de la suma total de correlaciones, con lo que se obtenía el valor de la correlación de la variable con el componente, y que correspondía al coseno del ángulo entre ellos.

Son precisamente estas correlaciones con las variables las que definen el componente; se denominan *pesos del componente* y tienen exactamente la misma interpretación que los coeficientes de correlación ordinarios. En particular, podemos usarlos para obtener una respuesta a la cuestión de cuán representativas son de las originales las nuevas variables promedio; elevando al cuadrado la correlación entre las variables y el componente (en otras palabras, los pesos del componente), corresponden precisamente al coeficiente r^2 de valores de determinación que hemos visto al exponer el análisis de regresión. Es decir, al elevar al cuadrado el peso del componente para una variable, podemos obtener el porcentaje de la variación en esa variable que es explicado por el nuevo componente (sin embargo, consúltese la nota de la página siguiente que se refiere al valor propio [o *eigenvalor*]). Las cifras de nuestro ejemplo aparecen en la tabla 13.6. En este caso, se aprecia que el nuevo componente explica el 56 % de la variación en la variable x_1, el 88,2 % de la variación en x_2, y así sucesivamente. Si sumamos todos esos valores, obtendremos la suma total de la variación explicada por el componente. Debido a que la forma en que se calcula esa cantidad en el álgebra matricial suele denominarse *valor propio* o *eigenvalor* (también *latent root*) de la matriz, siendo la matriz en cuestión la matriz original de correlaciones/cosenos que describe las relaciones entre nuestras cuatro variables.[1] La fórmula para el valor propio es:

TABLA 13.6. Correlaciones y correlaciones al cuadrado de las cuatro variables con el primer componente principal; según los datos de la tabla 13.3 y la figura 13.9.

Variable	Peso del componente	Peso del componente al cuadrado (r^2)
x_1	0,748	0,560
x_2	0,939	0,882
x_3	0,940	0,884
x_4	0,739	0,546

$$\lambda_i = \sum_{j=1}^{n} L_{ij}^2$$

donde λ_i es el valor propio para el componente i; L_{ij} es el peso [*loading*] de la variable j en el componente i, y la suma es mayor a todas las variables de la 1 a la n. En este caso tenemos:

$$\lambda_1 = 0{,}56 + 0{,}882 + 0{,}884 + 0{,}546 = 2{,}872$$

Tal como aparece es difícil atribuir mucho significado a la suma de los pesos al cuadrado. Es mucho más útil, para la interpretación de un componente, relacionar ese valor propio con la variación total en las variables. Debido a que tratamos con una matriz de coeficientes de correlación en la que se ha estandarizado la varianza en cada variable, la suma total de la variación en los datos está definida por la cantidad de variables incluidas en el análisis. Para encontrar el porcentaje de la variación en todas las variables conjuntamente y explicadas por el componente, dividimos el valor propio del componente entre la cantidad de variables, multiplicando a continuación por 100:

1. Algunos programas informáticos normalizan los pesos de los componentes de forma que sus valores al cuadrado sumen 1,0 y no el valor propio; en esos casos, los pesos al cuadrado no corresponden a los valores r^2 o los pesos mismos a los coeficientes de correlación. Para producir esa correspondencia, los pesos normalizados han de elevarse al cuadrado y multiplicarse por el valor propio del componente. Esto proporciona el peso al cuadrado corregido que corresponde a un valor r^2; la raíz cuadrada del mismo proporciona, a su vez, un peso que corresponde a un coeficiente de correlación.

$$\text{Porcentaje explicado por} = \frac{\lambda_1}{n} \times 100$$

En este caso:

$$\text{Porcentaje explicado por} = \frac{2{,}872}{4} \times 100 = 71{,}8\ \%$$

Podemos ver que nuestra nueva variable o componente principal explica el 71,8 % de la variación en las cuatro variables originales. La idea es exactamente la misma que la que vimos en el análisis de la regresión. En la regresión múltiple usábamos unas variables independientes para explicar la variación en el conjunto de las variables de partida. En términos de nuestro objetivo —reducir la complejidad en los datos reduciendo la cantidad de variables con las que hemos de tratar— estamos consiguiendo bastante: hemos sustituido cerca del 70 % de la variación en cuatro variables por una sola. En el caso de dos de las variables originales, el componente explica el 88 % de la variación; para las otras dos, x_1 y x_4, no es tan alto, alrededor del 55 %.

La cuestión siguiente es si podemos explicar algo del resto, tanto en las variables individuales como globalmente, obteniendo un segundo componente. Poniéndolo en relación con el tema de la regresión, podemos decir que la variación no explicada por el primer componente es su variación residual —la variación que tiene una correlación cero con él—. La mejor manera de explicar esta variación será en términos de un componente que no esté correlacionado con el primero, es decir, se que se sitúe en ángulo recto u ortogonalmente a él.

Los pesos de las cuatro variables en el segundo componente, junto con su conversión en ángulos, se muestra en la tabla 13.7. Si dibujamos ahora el segundo componente en nuestro diagrama original de las relaciones entre las cuatro variables y entre ellas y el componente, observaremos que el segundo se sitúa en ángulo recto al primero (fig. 13.12).

TABLA 13.7. Correlaciones y ángulos entre las cuatro variables representadas en la figura 13.8 y en las tablas 13.1-13.2, con el segundo componente principal que se deriva de ellas.

Variable	Peso en el segundo componente	Ángulo entre las variables y el segundo componente
x_1	−0,661	131°
x_2	−0,336	110°
x_3	0,335	70°
x_4	0,676	47°

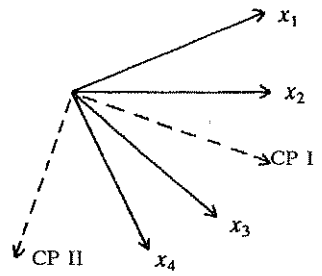


FIGURA 13.12. Representación geométrica de la correlación entre cuatro variables, con los dos primeros componentes principales añadidos.

Podemos seguir adelante, calculando la cantidad de variación en cada una de las variables individuales, y la suma global, explicadas por el segundo componente, simplemente elevando al cuadrado los pesos del componente. Igualmente, estos pesos al cuadrado pueden sumarse para obtener el valor propio del segundo componente. Los resultados de esa operación aparecen en la tabla 13.8, junto con los resultados del componente 1, que se han obtenido previamente.

TABLA 13.8. Correlaciones al cuadrado de las cuatro variables con los componentes principales 1 y 2, basado en la figura 13.12.

Variable	Peso al cuadrado en el componente 2	Peso al cuadrado en el componente 1
x_1	0,437	0,560
x_2	0,113	0,882
x_3	0,112	0,884
x_4	0,457	0,546
Suma	1,118	2,872

Se deduce que el componente 2 explica el 43,7 % de la variación en la variable x_1, el 11,3 % en la x_2, y así sucesivamente. El valor propio global de los componentes es de 1,118; para hallar el porcentaje de la variación de todas las variables juntas explicado por el segundo componente, efectuamos el mismo cálculo que para el primero:

$$\text{Porcentaje explicado por} = \frac{\lambda_2}{n} \times 100$$

Aquí,

$$\text{Porcentaje explicado por} = \frac{1{,}118}{4} \times 100 = 28{,}0\ \%$$

Ahora ya podemos examinar conjuntamente los resultados para ambos componentes. El primero explicaba el 71,8 % de la variación; los dos juntos explican el 99,8 % o el 100 %, dentro de los límites del error de redondeo. Del mismo modo, si examinamos directamente los valores propios, veremos que suman 3,99, mientras que la suma total de la variación en la matriz de correlación, de la cual derivaban los componentes, era 4,0, es decir, la cantidad de variables. Asimismo, si examinamos los resultados de las variables individuales y sumamos los pesos al cuadrado para cada una de ellas, alcanzan, más o menos, y dentro de los límites del error de redondeo, 1,0 o 100 %.

En otras palabras, los dos componentes explican el 100 % de la variación en las cuatro variables originales. Esto nos dice que podemos describir la variación en las cuatro variables, en términos de sólo dos nuevas, sin que perdamos nada de la información original. Nuestros datos, por tanto, pueden simplificarse de inmediato, contribuyendo el procedimiento a detectar y entender las estructuras que pudiese haber entre ellos. Visualmente, en vez de examinar los diagramas de dispersión de las relaciones entre las cuatro variables, nos limitaremos a un diagrama de dispersión en dos dimensiones. De hecho, podemos señalar en este caso que si no hemos sido capaces de explicar toda la variación con ayuda de dos componentes, será imposible representar las relaciones entre las cuatro variables originales correctamente en una hoja de papel. Si hubiésemos podido dibujar un diagrama de las relaciones correctas entre 100 variables en una hoja de papel, sabríamos de antemano que podían reducirse a dos componentes.

Ahora bien, como ya se ha señalado, nuestras nuevas variables pueden proporcionarnos, no sólo una simplificación útil en sí misma, sino que pueden definir las dimensiones subyacentes de variación, con interés sustantivo para nosotros, y que afectan a los valores de las variables que hemos medido. Para ver cómo funciona, habremos de cambiar primero nuestra perspectiva.

Hasta aquí sólo hemos examinado los componentes en relación con una serie de variables abstractas, definidas arbitrariamente para que muestren ciertas relaciones. Cuando tratemos con datos reales, emplearemos casos que adoptan una serie de valores en un conjunto de variables y es precisamente en términos de esos valores y sus relaciones como calcularemos las correlaciones entre las variables. Presumiblemente, por tanto, si sustituimos un conjunto de variables correlacionadas por unas nuevas no correlacionadas, podremos sustituir también los valores de los casos en las variables originales por sus valores en las nuevas, los cuales nos permitirán construir un diagrama de dispersión simplificado; estos nuevos valores se denominan *puntuaciones en los componentes*. Se entenderá mejor la manera en que se obtienen mediante un ejemplo.

Cuando las observaciones en dos variables están correlacionadas, el diagrama de dispersión adopta la apariencia del de la figura 13.13, con el centro de gravedad de la distribución en la intersección de las dos medias. ¿Dónde se sitúan los ejes del espacio que define esa nube de puntos? Son los ejes de la elipse que incluye la nube de puntos; el ángulo entre ellos está definido por la correlación entre las dos variables; el origen es el punto de intersección de las dos medias (fig. 13.14).

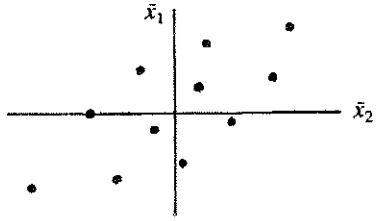


FIGURA 13.13. Diagrama de dispersión de los puntos con valores en ambas variables, x_1 y x_2.

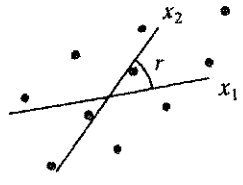


FIGURA 13.14. Ejes de la elipse que define la nube de puntos de la figura 13.13. El ángulo entre los ejes corresponde a la correlación entre las variables.

Al encontrar los componentes principales, definimos diferentes ejes para esa nube de puntos. El primer componente principal corresponde al eje más largo de la elipse, y el segundo componente al eje más corto, situado en ángulo recto al primero. Las longitudes de los nuevos ejes o componentes corresponde a sus valores propios. El resultado se muestra en la figura 13.15.

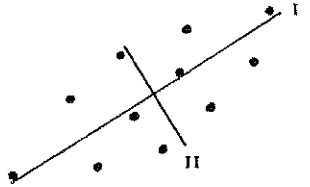


FIGURA 13.15. Los componentes principales de la nube de puntos en la figura 13.13.

Moviéndonos tan sólo de un par de dimensiones a otro, como en este ejemplo, no conseguimos mucho, si bien las posibilidades de reducción del espacio son particularmente atractivas.

Al lector le puede asombrar qué tiene que ver esto con las puntuaciones

en los componentes, pues parece que no hagamos otra cosa que repetir la derivación de los componentes principales desde una perspectiva ligeramente distinta. Las puntuaciones en los componentes aparecen cuando nos centramos en lo que le pasa a un punto en particular cuando los ejes se han transformado: ¿cómo obtenemos, de los valores en las dos variables originales, sus valores en los nuevos componentes? Veámoslo por medio de un diagrama (fig. 13.16).

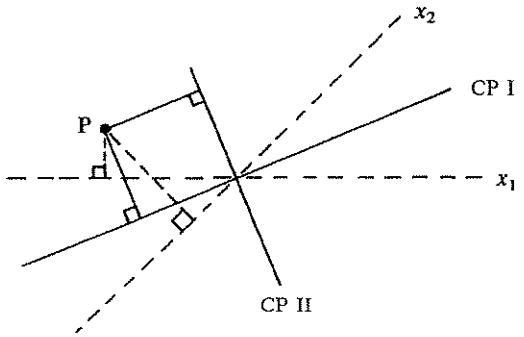


FIGURA 13.16. Las relaciones entre el valor en un punto de las variables originales, x_1 y x_2, y su puntuación en los componentes CP I y CP II.

La posición del punto *P* está inicialmente definida por su valor en las variables x_1 y x_2; las líneas discontinuas proyectadas desde el punto para conectar con los dos ejes en ángulo recto muestran los valores del punto *P* en esas dos variables. Cuando encontramos los componentes principales, no hacemos nada con el punto *P* —permanece donde está—, pero operamos con un nuevo sistema de coordenadas en términos del cual describiremos su posición, la cual está definida por las líneas continuas proyectadas desde el punto *P* a los componentes en ángulo recto. La posición del punto en cada uno de los nuevos ejes recibe el nombre de puntuación en el componente para ese eje en concreto. De la misma forma que trazábamos los diagramas de dispersión de los datos según sus valores en las variables originales, trazaremos los diagramas de dispersión de los puntos según sus puntuaciones en los componentes; ya veremos la utilidad que ello reporta.

Las puntuaciones en los componentes se obtienen a partir de la fórmula:

$$S_{ik} = \sum_{j=1}^{n} x_{ij} L_{jk}$$

donde S_{ik} es la puntuación de la observación *i* en el componente *k*; x_{ij} es el valor estandarizado de la observación *i* en la variable *j*; L_{jk} es el peso de la variable *j* en el componente *k*, y *n* es la cantidad de variables. En palabras, empezamos calculando la puntuación típica de la observación *i* en la variable *j*; los

componentes principales los hemos descrito en términos de una matriz de coeficientes de correlación, por lo que operaremos sobre los coeficientes estandarizados de las variables. A continuación multiplicamos ese valor por el peso de la variable en el componente, es decir, por la correlación de la variable en el componente, que a su vez nos proporciona el ángulo entre la variable y el componente. Obviamente es necesario conocer ese ángulo si queremos pasar de un conjunto de coordenadas a otro. De hecho, hemos de conocer la relación entre cada variable y el componente para poder hacer satisfactoriamente la transformación de las coordenadas. Hemos de registrar cada peso y el valor de los puntos en cada variable, multiplicarlos y a continuación sumar todos los resultados, para todas las variables que haya en el análisis.

Al final del procedimiento de cálculo de las puntuaciones en los componentes —que no ha de hacerse a mano, sino mediante programas informáticos específicos— obtendremos una tabla con las puntuaciones de nuestros casos individuales en cada componente, de la misma forma que al principio teníamos una tabla con sus puntuaciones en las variables que habíamos medido.

Resumen del análisis de componentes principales. En resumen, el ACP (análisis de componentes principales) es extremadamente versátil y hace muchas cosas útiles al mismo tiempo (Doran y Hodson, 1975, p. 196):

1. Proporciona una indicación muy útil de las relaciones entre las variables.
2. Proporciona información acerca de las relaciones entre las unidades.
3. Sugiere si existe alguna tendencia en los datos originales, y qué variables están relacionadas con dicha tendencia.
4. Proporciona una transformación de los datos en la que una gran proporción de la variación entre numerosas variables se comprime en un número menor de variables.
5. La transformación efectuada es tal, que las nuevas variables generadas no están correlacionadas entre sí.

Es esta última propiedad la que hace que el análisis de componentes principales sea apropiado como método para superar el problema de la colinealidad en la regresión múltiple. Es posible usar como datos de entrada en una regresión, no las puntuaciones de los individuos en las variables originales, sino sus puntuaciones en los componentes. El uso de los componentes como variables independientes elimina la posibilidad de desviación o ambigüedad en los coeficientes de regresión. El único problema es que los componentes no siempre son fáciles de interpretar, por lo que puede que no esté claro qué es lo que hace en realidad la regresión.

Finalmente es apropiado decir algo acerca de los presupuestos en el ACP. En sí mismo, es tan sólo un método matemático para extraer los ejes principales de las matrices; puede aplicarse a cualquier matriz simétrica de coeficientes que definan las relaciones entre las variables. Usualmente se aplica a matrices de coeficientes de correlación y, por tanto, es importante que el coeficiente de correlación lineal, *r*, proporcione una imagen satisfactoria de las relaciones en-

tre las variables. Esta cuestión puede investigarse mediante el examen de los diagramas de dispersión bivariados de las relaciones entre pares de variables. Si las distribuciones de variables individuales son normales, el uso del coeficiente de correlación será satisfactorio.

Si la matriz analizada no es una matriz de correlaciones, se tratará generalmente de una matriz de varianzas y covarianzas. El coeficiente de correlación, naturalmente, es la covarianza entre dos variables estandarizadas (puntuaciones *Z*). Por consiguiente, los comentarios relevantes para los coeficientes de correlación también se aplican a las covarianzas.

De hecho, la mayoría de estadísticos profesionales suelen sugerir el uso de la matriz de correlación antes que la covarianza, ya que ésta tiende a destruir la validez de la teoría estadística distribucional; puede hacer que los resultados sean difíciles de interpretar; y permite la posibilidad dudosa de combinar distintos tipos de medidas (Fieller, comunicación personal).

Como se señaló al principio, la decisión de analizar una matriz de covarianza o de correlación no debe tomarse al margen del caso concreto que estudiemos. Además de lo que ya hemos dicho, hay que insistir en que el análisis de dos matrices diferentes puede provocar resultados diferentes —lo que no debiera sorprender a nadie, si pensamos en la estandarización de las variables.

Otro punto que hay que retener es que, como el ACP está diseñado para extraer ejes de las matrices, lo hará al margen de cualquier significado sustantivo que puedan o no tener. En cualquier matriz de correlaciones entre un cierto número no muy grande de variables habrá una gran cantidad de valores. Algunos de ellos serán muy grandes y estadísticamente significativos al azar, incluso aunque los datos originales hayan sido generados aleatoriamente por un ordenador. Basándose en el azar de esos grandes valores, aparentemente significativos, puede parecer que los componentes son interpretables en términos sustantivos para un arqueólogo que esté buscando la estructuración subyacente en sus datos. Este problema ha sido tratado, en un contexto arqueológico, por Vierra y Carlson (1981), que sugieren comprobar la matriz de correlaciones antes de iniciar el ACP, para ver si existe un número significativamente grande de correlaciones significativas. De esta forma pueden evitarse algunos resultados ilusorios.

Un ejemplo arqueológico. Necesitamos de un ejemplo arqueológico para aclarar la exposición metodológica anterior y hacerla comprensible al lector. Analizaremos la variación de la forma de 65 vasijas de cerámica del neolítico final en la Europa central. La forma de las vasijas está descrita por medio de doce medidas, tal y como muestra la figura 13.17.[2] Se han registrado diez de esas medidas a intervalos, desde el borde hasta la base (véase Shennan y Wilcock, 1975); las medidas se obtuvieron desde la línea central de la vasija hasta su superficie exterior. También se registraron otras dos medidas: la altura de la panza del recipiente y la altura del cuello. Para que el tamaño de la vasija,

2. Existen métodos más sofisticados de descripción de la forma de un recipiente; véase, por ejemplo, el uso de los códigos de contornos por Kampffmeyer y Teegen (1986).

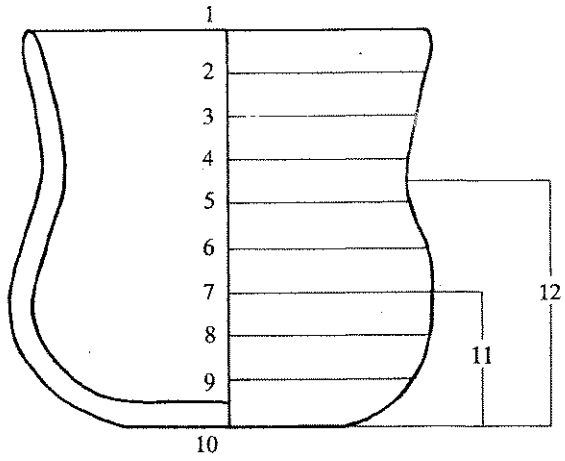


FIGURA 13.17. Las medidas usadas para describir la forma de unas vasijas neolíticas de la Europa central.

medido a partir de la altura global de la misma, no sea un factor mayor de la variación, se estandarizaron todas las medidas, dividiéndolas por la altura global. Se obtuvieron a continuación las correlaciones entre las proporciones resultantes, efectuándose un análisis de componentes principales de la matriz de coeficientes de correlación entre todas las variables. Los valores propios y la varianza explicada aparecen en la tabla 13.9.

TABLA 13.9. Valores propios y varianza explicada por los 12 componentes principales que resultan del análisis de un conjunto de vasijas descritas a partir de su morfometría.

Componente	Valor propio	Porcentaje de varianza	Varianza acumulada
1	7,30	60,87	60,87
2	2,05	17,07	77,93
3	1,41	11,77	89,69
4	0,58	4,85	94,54
5	0,45	3,74	98,27
6	0,08	0,64	98,91
7	0,06	0,48	99,38
8	0,04	0,30	99,68
9	0,02	0,16	99,82
10	0,01	0,09	99,90
11	0,01	0,06	99,96
12	0,00	0,05	100,00

Como puede verse, empezamos con doce variables y hay doce componentes; pero la mayoría de ellos sólo explican proporciones muy pequeñas de la varianza. La cuestión que se plantea es, como en muchos otros casos, cuántos componentes hay que utilizar. No hay ninguna regla fija para ello, si bien la guía que se suele adoptar es considerar los componentes con un valor propio de 1,0 o más. Este valor representa la varianza de una única variable en la matriz de correlación, de forma que si un componente tiene un valor propio menor que este, explicará una menor cantidad de varianza que cualquiera de las variables originales (Johnston, 1978, p. 146). Una alternativa sería trazar los valores propios como eje principal y los componentes como eje horizontal, y observar el punto de declinación de la curva.

Adoptando el primer enfoque al caso que estamos tratando, veremos que sólo hemos de operar con tres componentes, los cuales suman el 90 % de la variación en los datos. Así pues, siguiendo uno de los principales objetivos del ACP, la reducción de los datos y su simplificación, habremos reducido el ejemplo de doce a tres variables, reteniendo la mayor parte de la información original. De las tres, la primera es, naturalmente, la más importante.

Se observa que hay alguna tendencia particular en la inicialmente confusa variación en los datos originales; el paso siguiente será ver qué variables están implicadas en esas tendencias. Para hacerlo necesitamos examinar los pesos de las variables en los componentes, así como sus valores al cuadrado, con el fin de apreciar el porcentaje de la variación en las variables explicado por los componentes. Aparecen en la tabla 13.10.

TABLA 13.10. Pesos y pesos al cuadrado en los tres primeros componentes principales de las 12 variables que definen la forma de una vasija. Proceden del análisis de componentes principales de la matriz de correlaciones entre las 12 variables.

	Componente I		Componente II		Componente III	
Variable	Peso	Pesos al cuadrado	Peso	Pesos al cuadrado	Peso	Pesos al cuadrado
1	0,730	0,532	−0,460	0,211	−0,437	0,191
2	0,811	0,657	−0,395	0,156	−0,393	0,154
3	0,897	0,805	−0,235	0,055	−0,331	0,110
4	0,924	0,854	0,229	0,052	−0,216	0,048
5	0,911	0,829	0,384	0,147	−0,014	0,0002
6	0,919	0,844	0,322	0,104	0,076	0,006
7	0,929	0,864	0,132	0,017	0,223	0,050
8	0,894	0,800	−0,064	0,004	0,384	0,147
9	0,794	0,631	−0,241	0,058	0,521	0,272
10	0,576	0,331	−0,435	0,189	0,215	0,046
11	0,349	0,121	0,727	0,529	−0,454	0,206
12	0,184	0,034	0,726	0,527	0,297	0,085

Si examinamos, en prifner lugar, el componente I, apreciaremos que la gran mayoría de las variables tienen una elevada correlación positiva con él; explica más del 50 % de la varianza de las variables 1 a 9. El componente I define una estructura de variación común a todas ellas. Basándonos en la serie de altas correlaciones, podemos decir que cuando un caso tiene valores elevados en una de ellas, tendrá también valores elevados en las demás, y que cuando adopta valores bajos en una de ellas, adoptará valores bajos en las demás. Este esquema está condensado en una única tendencia definida por el componente I, que explica el 60 % de la varianza en los datos. Precisamente lo que representa esa tendencia se hace más claro al examinar la puntuación de los casos individuales en el componente I, porque es aquí donde la abstracción del análisis puede relacionarse directamente con los datos arqueológicos originales. Por el momento, sin embargo, continuaremos con el examen de los pesos en los componentes.

Un vistazo a los del componente II muestra inmediatamente que, en general, es mucho menos importante que el primero —explica tan sólo una pequeña proporción de la variación de la mayoría de las variables—. La excepción son las variables 11 y 12, por lo que resulta evidente que es con estas dos variables con las que el componente II está asociado, explicando más del 50 % de la variación en cada caso. Se deduce que ambas tienen un esquema de variación común, definido por este componente: cuando la altura de la panza es alta, también lo es la altura del cuello; cuando una es baja, también lo es la otra.

El componente III es, naturalmente, aún menos significativo. En ningún caso explica más del 25 % de la variación en una variable. Examinando los pesos, veremos que las variables 1 y 2 tienen una correlación negativa moderada con el componente, al igual que la variable 11; la variable 9, por su parte, tiene una correlación positiva moderada. Parece que esto sugiera que, en términos de este componente, los valores altos en la variable 9 estuviesen asociados con valores bajos en las variables 1, 2 y 11; pero no está claro lo que esto significa en términos de la forma de la vasija. Para avanzar algo más habremos de fijarnos en la puntuación de los componentes, y no sólo en los pesos.

En la tabla 13.11 se enumeran las puntuaciones de los casos individuales en los tres primeros componentes principales; sólo se han incluido uno de cada tres casos, de un total de 65, pues el objetivo del cuadro es tan sólo ilustrativo. Estas puntuaciones nos proporcionan un medio de obtener una impresión directa del significado arqueológico de los componentes. Señalando los casos con los valores negativos y positivos más altos en los componentes, podremos referirnos a las medidas o a los dibujos de las vasijas en particular y ver qué es lo que tienen en común y qué las diferencia de las demás. El componente representará, por tanto, una tendencia entre esos tipos extremos.

Si examinamos el componente I, vemos que los casos 19, 25, 28 y 58 tienen valores negativos grandes y que 13, 37, 61 y 64 adoptan unos valores positivos bastante elevados. En el componente II, los casos 1, 25, 52 y 64 han adoptado

TABLA 13.11. Puntuaciones de 22 vasijas en los tres primeros componentes principales derivadas de la matriz de correlaciones entre las 12 variables que describen su morfometría.

Caso n.º	Puntuación en el componente I	II	III	Caso n.º	Puntuación en el componente I	II	III
1	2,426	−1,403	−1,550	34	−1,801	3,318	−0,501
4	−2,546	0,586	−0,460	37	4,555	2,965	−1,075
7	−0,362	0,184	0,926	40	−2,593	1,214	−0,974
10	0,871	−0,282	2,436	43	−1,657	0,500	1,281
13	3,296	−0,573	0,676	46	1,505	1,878	2,178
16	−1,517	−0,051	−1,001	49	0,819	0,338	0,468
19	−4,392	−0,061	−0,351	52	−1,759	−2,298	−0,696
22	−2,891	0,496	−0,257	55	−0,544	−0,786	−0,426
25	−3,672	−1,238	−3,329	58	−3,352	0,131	0,445
28	−3,886	−0,627	0,401	61	3,771	0,292	0,557
31	−1,544	4,143	−0,350	64	4,185	−2,577	−2,139

valores negativos importantes, mientras que para 31, 34 y 37 son positivos. Finalmente, en el componente III, los casos 1, 25 y 64 se sitúan en el extremo negativo, y 10, 43 y 46 en el opuesto.

Los valores de los datos originales para todos los casos cuyas puntuaciones en los componentes están dadas —medidas expresadas como porcentajes de la altura de las vasijas— aparecen en la tabla 13.12. Si prestamos atención a los casos con valores negativos altos en el componente I, veremos que suelen adoptar valores bajos en las primeras nueve variables, mientras que los casos opuestos a ellas adoptan valores elevados a esas mismas variables. Recordando que todos esos valores son medidas de anchura en relación a la altura, constataremos que todas aquellas con valores bajos son vasijas estrechas, y aquellas con valores altos, anchas o panzudas. En otras palabras, el componente I representa una tendencia de la anchura a estrecharse en la forma de la vasija, tendencia que resume la mayoría de la covariación entre estas medidas; obviamente, si una de ellas tiende a ser grande o pequeña, las otras medidas tenderán a seguirla, como indicaban los valores de los pesos. Refiriéndonos a estos últimos, la única medida de anchura que no se ajusta con la misma intensidad a ese esquema es la base de la vasija.

Fijémonos ahora en el componente II. Tal y como esperábamos gracias a la información contenida en los pesos, las distinciones que se hacen aquí se basan en la altura de la panza y del cuello (variables 11 y 12). Algunas vasijas presentan una panza y un cuello bastante bajos, si los comparamos con la altura global, mientras que otras son relativamente altas. El componente II representa, pues, una tendencia entre unas y otras. Por definición esta tendencia en la variación es independiente de la definida por el primer componente.

En el caso del componente III el esquema es más complicado y difícil de

TABLA 13.12. Datos originales de las 22 vasijas cuyas puntuaciones en los componentes aparecían en la tabla 13.11.

Caso n.º	Variables 1	2	3	4	5	6	7	8	9	10	11	12
1	60,36	55,86	51,35	48,65	50,45	53,15	54,05	50,45	42,34	27,93	37,84	65,77
4	41,28	37,61	35,78	36,70	39,45	43,12	42,20	38,53	33,03	25,69	38,53	77,06
7	40,96	38,55	37,35	37,35	48,19	53,01	54,22	50,60	43,37	21,69	33,73	68,67
10	34,88	34,88	38,37	40,75	50,00	56,98	59,30	55,81	47,67	33,72	34,88	62,79
13	54,84	50,54	47,31	48,39	53,76	59,14	62,37	58,06	46,24	31,18	34,41	75,27
16	47,62	41,90	39,05	40,00	41,90	44,76	45,71	42,86	36,19	20,00	36,19	70,48
19	38,40	34,40	32,00	32,00	33,60	36,80	39,20	38,40	32,00	16,80	30,40	71,20
22	40,00	36,47	34,12	35,29	36,47	41,18	44,71	42,35	36,47	17,65	36,47	72,94
25	48,24	44,71	38,82	35,29	31,76	38,82	40,00	35,29	25,88	15,29	37,65	51,76
28	37,50	31,94	30,56	31,25	34,72	38,89	42,36	40,28	34,03	24,31	31,25	64,58
31	32,18	27,59	28,74	49,43	52,87	50,57	47,13	41,38	33,33	22,99	55,17	81,61
34	32,86	34,29	37,14	42,86	48,57	51,43	50,00	44,29	34,29	8,57	41,43	84,29
37	50,75	47,76	47,76	64,18	70,15	70,15	64,18	56,72	41,79	20,90	49,25	79,10
40	35,71	34,52	35,71	39,29	44,05	46,43	45,24	39,29	30,95	20,24	41,67	66,67
43	35,29	34,31	33,33	36,27	44,12	49,02	50,98	49,02	41,18	20,59	32,35	70,59
46	37,33	36,00	36,00	45,33	54,67	61,33	62,67	60,00	48,00	21,33	37,33	78,67
49	44,00	42,67	41,33	41,33	50,67	56,00	57,33	54,67	42,67	21,33	36,00	69,33
52	51,39	45,83	38,89	37,50	37,50	40,28	44,44	45,83	37,50	22,22	26,39	59,72
55	46,74	43,48	40,22	41,30	44,57	48,91	52,17	46,74	38,04	22,83	32,61	63,04
58	32,17	32,17	31,30	33,04	39,13	43,48	44,35	42,61	35,65	21,74	34,78	62,61
61	50,53	48,42	48,42	54,74	60,00	62,11	62,11	58,95	48,42	27,37	36,84	73,68
64	66,15	64,62	56,92	52,31	52,31	55,38	55,38	53,85	46,15	33,85	41,54	56,92

apreciar, lo cual no sorprende dados los pesos, generalmente débiles, y la estructura relativamente compleja que estos mismos indican; sólo se señalaba que las variables 1, 2 y 11 mostraban correlaciones negativas moderadas y que la variable 9 adoptaba una correlación positiva moderada con ese componente.

El examen de los valores de esas variables para las vasijas cuya representación se sitúa a cada extremo del componente III muestra que las situadas en el extremo negativo tienen un borde relativamente ancho, lo cual se puede observar en las medidas más próximas al borde, mientras que las vasijas situadas en el extremo positivo son relativamente estrechas de borde; los dos extremos muestran también un marcado contraste en relación con las variables 9 y 11. En el extremo negativo, la anchura cercana a la base (9) y la altura de la panza (11) son bastante similares, pero, en el extremo positivo, las anchuras próximas a la base tienden a ser grandes, y la altura de la panza, pequeña.

Así pues, el examen de la puntuación de los componentes es importante para comprender con precisión qué es lo que nos dice el análisis de componentes principales acerca de nuestros datos. Es a ese nivel donde el análisis abstracto y la evidencia arqueológica se confrontan mutuamente. La puntuación en los componentes es, en esto, mucho más inmediata que los pesos en los mismos componentes. Sin embargo, la identificación de los casos individuales en relación a los componentes es también importante. En este ejemplo hemos definido, simplemente, el significado arqueológico de los componentes en términos de la variación en la forma de la vasija. Puede suceder, no obstante, que esas tendencias en la forma estén relacionadas, por ejemplo, con el cambio a lo largo del tiempo, de forma que las vasijas en uno de los extremos del componente son más antiguas, y las del otro extremo, modernas. Si somos capaces de considerar las vasijas individualmente, sea el caso que sea, tendremos acceso a la información que pueda existir sobre otros aspectos de su contenido arqueológico: si están asociados, por ejemplo, con otros elementos para los que existe una datación independiente. Esta información no se consigue si consideramos tan sólo las correlaciones globales o los pesos.

Ahora bien, hasta aquí hemos examinado las puntuaciones de los casos en cada componente por separado. Obviamente es esencial si lo que pretendemos es definir el significado de los mismos en términos arqueológicos. Sin embargo existe la posibilidad de generar diagramas de dispersión de las puntuaciones en los casos a partir de dos de los componentes. En el ejemplo anterior, un diagrama de dispersión que usara los dos primeros componentes incluiría el 78 % de la variación en los datos, mientras que si usáramos símbolos diferentes para los casos según sus puntuaciones en el componente III, se incluiría un 12 % más, llevando el total a un 90 %. Las posibilidades de definición visual de las estructuras en esas circunstancias, comparada con la que se obtiene de las 12 variables originales, es ciertamente mucho mayor.

¿Qué estructuras estamos buscando? En primer lugar, puede que la evolución cronológica no sea aparente en una sola dimensión, y que se haga más

clara cuando usamos dos. En segundo lugar, como se señaló en el capítulo anterior, esos diagramas de dispersión pueden proporcionar un suplemento o alternativa al análisis de conglomerados. Proporcionan una ordenación de los datos en muy pocas dimensiones que contienen una gran cantidad de información. Podemos decir, por ejemplo, si hay auténticos conglomerados de datos, o si los conglomerados representan la división relativamente arbitraria de un continuo; o bien qué puntos son marginales (*outliers*) y realmente no se relacionan con los demás. De nuevo, y al igual que sucedía con los resultados del análisis de conglomerados de los datos, podemos ver, por ejemplo, si todos los casos de una parte específica del diagrama de dispersión proceden de un yacimiento en particular.

Un diagrama de dispersión de los casos cuyas puntuaciones en los componentes se enumeran aparece en la figura 13.18. Se basa en los dos primeros componentes. Está claro que, si bien hay tendencias muy marcadas en los datos y algunas de las vasijas son distintivamente inusuales, situándose más allá de la distribución principal, hay pocas evidencias de agrupación discreta. Esto no quiere decir que no podamos intentar romper el continuo por alguna razón, pero si lo hacemos hemos de tener en cuenta que estamos rompiendo un continuo y que no reconocemos conglomerados discretos claramente definidos.

Esta exposición de las puntuaciones en los componentes completa la presentación del análisis de componentes principales. Debemos insistir en que mien-

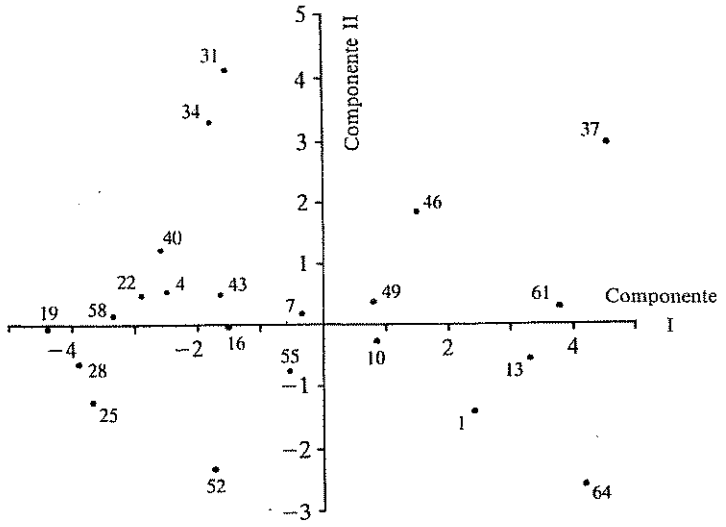


FIGURA 13.18. Diagrama de dispersión de las vasijas cerámicas en los dos primeros componentes principales. Datos a partir de los de la tabla 13.11.

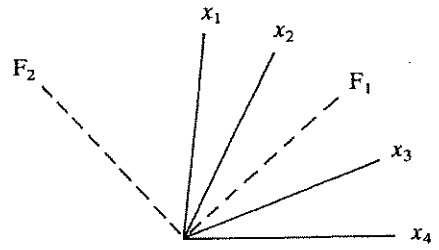


FIGURA 13.19. Representación geométrica de las correlaciones entre cuatro variables, y los dos factores que explican toda la variación entre ellos (según Johnston, 1978).

variables en el espacio simplificado definido por las nuevas dimensiones subyacentes.

Ese es precisamente el paso siguiente que emprende el análisis factorial: los ejes se giran para conseguir esa correspondencia. Las reglas que se han propuesto e implementado para efectuar ese procedimiento son muchas y muy variadas. Algunas de ellas, incluso, prescinden de la exigencia según la cual los ejes han de formar ángulos rectos entre sí; pero no trataremos aquí más que las usuales y que satisfacen esa limitación. Se denominan *rotaciones ortogonales*. Lo que se pretende es girar los ejes hasta una posición que sea lo más próxima posible a un ideal, referido como estructura simple. Este ideal es que cada variable se identifique totalmente con un factor único y no con otro; toda la varianza ha de estar explicada por un único factor, y no dividida entre varios, como sucedía con el análisis de componentes principales. En términos numéricos, el objetivo es que cada variable tenga un peso de 1,0 en un factor y de 0,0 en los demás. Obviamente, eso es imposible en la práctica; afortunadamente existen medios informáticos capaces de aproximarse a ese ideal lo más posible en cada caso en particular. En lo que se refiere al ejemplo de la figura 13.19, podemos hacer la rotación visualmente (fig. 13.20), lo que sería imposible en ejemplos reales.

Como puede verse, una vez efectuada la rotación, no sólo hemos reducido la variación en los datos a una cantidad menor de dimensiones, sino que hemos identificado las nuevas dimensiones, en la medida de lo posible, con conjuntos específicos de variables que pueden ser potencialmente útiles en la comprensión de lo que hay detrás de la variación en los datos.

Ahora bien, la rotación causa sus propios problemas, ya que decidir la cantidad de factores que hay que girar determina el número de dimensiones subyacentes que, supuestamente, están detrás de la variación observada, afectando también a la definición de las dimensiones; si giramos dos factores, se producirán dos grupos de variables; si giramos tres, se obtendrán tres grupos de variables. Además, la introducción de una tercera no añadirá simplemente una dimensión asociada a un grupo de variables al primero de los dos grupos, sino

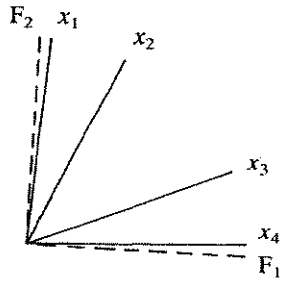


FIGURA 13.20. Los dos factores de la figura 13.19 después de una rotación ortogonal lo más próxima posible a la estructura inicial (según Johnston, 1978).

que cambiará la definición de las dos primeras. Esto es intuitivamente evidente, si se piensa en lo que implica la rotación en una nube de puntos de un conjunto fijo de ejes perpendiculares entre sí; es poco probable que la mejor posición para los tres ejes en términos del criterio especificado antes sea la misma que en el caso de sólo dos ejes. Cuando no sólo el número, sino también la naturaleza, lo que en cierto modo se supone que son «dimensiones subyacentes objetivamente existentes», está afectada, o incluso determinada, por unas selecciones relativamente arbitrarias acerca del número de ejes que hay que rotar, uno siente que se requieren muchas precauciones, a no ser que hayan muy buenas razones para elegir el número de ejes. Si los datos están distribuidos según una distribución normal multivariada, el análisis factorial de máxima verosimilitud (*maximun likelihood*) proporcionará un método estadísticamente riguroso para decidir la cantidad apropiada de factores para la rotación (Everitt y Dunn, 1983, p. 200).

Puntuaciones factoriales. Cuando examinamos las puntuaciones de las observaciones individuales en los nuevos ejes, el análisis factorial presenta nuevos problemas. Vimos que con el análisis de componentes principales podían obtenerse las puntuaciones individuales en los componentes, y que proporcionan información muy útil. Desafortunadamente no pueden obtenerse de la misma forma puntuaciones factoriales precisas para las observaciones individuales, porque el análisis factorial no usa toda la variación en los datos, sino tan sólo la varianza común, «dado que los valores observados en las variables originales combinan elementos comunes y únicos en proporciones desconocidas» (Johnston, 1978, p. 173). A causa de esta discrepancia entre el fundamento del análisis y la naturaleza de las observaciones individuales, las puntuaciones factoriales sólo pueden estimarse por medio de un procedimiento de regresión. Por consiguiente, aunque la mayoría de los programas informáticos de análisis factorial tienen la opción de producir puntuaciones factoriales, y en esencia son lo mismo que las puntuaciones en los componentes, es importante, al interpre-

tras lo usemos con datos apropiados puede proporcionar mucha información arqueológicamente relevante acerca de un conjunto de datos de interés, que no necesariamente ha de ser accesible o aparente a un enfoque intuitivo de los mismos datos, especialmente si son muchos los casos analizados. El ojo y la mente son excelentes para proporcionar una impresión general de un conjunto peculiar de artefactos (si es que estamos estudiando artefactos), en términos de la variación morfológica entre ellos. Para separar la variación global en aspectos distintos, esto es, para analizarla, es mejor usar métodos apropiados y así conseguir una nueva información; después del análisis, el examen visual puede desempeñar todavía un papel importante.

Análisis factorial

Una vez resumida la forma de trabajar con el análisis de componentes principales, hemos de prestar algo de atención al análisis factorial, ya que se ha empleado con más frecuencia en arqueología; la comprensión de un número de artículos no desdeñable depende del conocimiento que tengamos de él. Se trata de un tema que ha provocado una considerable discusión entre los estadísticos profesionales y entre los usuarios de la técnica en diversas disciplinas; sería probablemente cierto decir que las opiniones aún están divididas en lo que se refiere a su utilidad. Esto debiera prevenir al arqueólogo no experimentado. En las páginas que siguen usaremos el enfoque de Johnston (1978); una exposición más técnica es la de Everitt y Dunn (1983, capítulo 1).

La diferencia esencial entre el análisis factorial y el análisis de componentes principales puede parecer insignificante, pero tiene consecuencias bastante importantes. El análisis de componentes principales extrae los componentes de la totalidad de la varianza existente en los datos; el análisis factorial se basa en un principio distinto. Asume que la varianza en una variable puede dividirse en dos segmentos, un segmento que tiene en común con las otras variables y refleja su relación con ellas, y otra parte que le es única y no se relaciona con ninguna otra; estas dos partes se denominan *varianza común* y *varianza única*.

Dado que el análisis factorial pretende definir la estructura subyacente de la variación común a distintas variables, ha de operar tan sólo con la varianza común y dejar la varianza única fuera de la explicación: incluir la totalidad de la varianza sería confuso. La primera cuestión que se plantea es, entonces, cómo estimar la varianza única de las variables para poder eliminarla; la varianza común restante de una variable suele denominarse *comunalidad* en lenguaje técnico.

Comparémoslo con el análisis de componentes principales. Vimos antes que, cuando se extraían los componentes a partir de una matriz de correlación, las entradas individuales sobre la diagonal principal eran siempre 1,0, esto es, la varianza estandarizada de cada variable. En el análisis factorial, el valor a lo

largo de esa diagonal no es 1,0 sino las estimaciones de la comunalidad para las distintas variables. Habitualmente son los valores R^2 múltiples de las variables individuales, variando evidentemente de una a otra variable. Recordará el lector que en el capítulo 11, dedicado a la regresión múltiple, los valores R^2 múltiples daban la proporción de la variación en una variable dependiente explicada por el efecto de todas las variables independientes actuando conjuntamente y por separado. Así pues, para obtener las comunalidades, cada variable es tratada a su vez como dependiente de las demás. La idea es que la cantidad de varianza explicada por las otras variables es una medida de lo que una variable en particular tiene en común con las demás; sea cual sea la fuente del resto de la varianza. La suma de las comunalidades es la cantidad total de variación en el análisis. Si una variable en particular tiene una comunalidad pequeña —en otras palabras, la forma en la que varía tiene poco en común con las otras variables— desempeñará un papel muy pequeño en el análisis.

En términos de la ilustración gráfica de la extracción de los componentes principales, las líneas o vectores que representaban las variables individuales tenían la misma longitud, y cada variable desempeñaba la misma función en la definición de la posición del primer componente; en el análisis factorial, los vectores que corresponden a las variables tendrán longitudes diferentes, las que correspondan al valor de su comunalidad, y la posición de la primera de las variables promedio —el factor en este caso— estará más afectada por aquellas con una comunalidad mayor; una variable con una comunalidad dos veces mayor que otra tendrá dos veces más influencia en la posición del factor.

Como en el caso del análisis de componentes principales, pero usando una matriz diferente, obtenemos los pesos de cada una de las variables en el factor. Esos pesos al cuadrado dan la proporción de la varianza en las variables explicada por el factor. Es aquí donde se aprecia una nueva diferencia entre los dos métodos. En el análisis de componentes principales de una matriz en particular, se obtienen los componentes sucesivamente, de acuerdo a un procedimiento matemático fijo, y se acaba aquí. El análisis factorial, con su énfasis en hallar la estructura de variación subyacente en los grupos de variables, precisa de algunos pasos más.

Lo entenderemos mejor por medio de la figura 13.19, en la que todos los vectores tienen una longitud igual para simplificar la presentación. Aquí disponemos de cuatro variables, junto con dos factores que explican toda la variación dentro de ellas. Al igual que en el análisis de componentes principales, el primero de ellos representa el esquema promedio de los cuatro, basado ahora en sus comunalidades, y el segundo en el promedio de lo que resta, situado por definición en ángulo recto al primero. Pero ¿representan verdaderamente las dimensiones subyacentes a la variación de dos grupos de variables? Tenemos tendencia a creer intuitivamente que si los factores pudieran moverse alrededor, de forma que uno correspondiera a un par de variables y el otro al segundo par, obtendríamos una mejor representación de las relaciones entre las

tarlas, tener en cuenta que se trata tan sólo de estimaciones, mejores en algunos casos que en otros, y no puntuaciones fijas.

Un ejemplo arqueológico. Como ejemplo arqueológico del análisis factorial, consideraremos el descrito por Bettinger (1979). Bettinger pretendía estudiar el sistema de subsistencia y asentamiento en Owens Valley, California. Realizó una prospección de superficie en el área y obtuvo información de 100 yacimientos según la frecuencia de aparición de nueve categorías distintas de rasgos y artefactos. Basándose en esa información, intentó hacer inferencias acerca de la función de los distintos yacimientos.

Bettinger prefirió en este caso un análisis factorial a un análisis de componentes principales, precisamente para saber si en este caso particular tenía algún sentido la distinción entre varianza común y varianza única. Afirmó que en lo que se refería a la distribución regional de rasgos y útiles, la variación idiosincrásica en las variables entre yacimientos procedería de factores tales como la conservación diferencial o el acceso diferencial a las materias primas. La varianza común, por otro lado, los esquemas de variación que las distintas variables tenían en común con las demás de manera sistemática, «procederían verosímilmente de su uso en actividades complementarias sincronizadas por la estructura de la economía indígena» (Bettinger, 1979, p. 457). Dado que era la estructura de la economía indígena lo que le interesaba a Bettinger, sería preferible para la investigación un método analítico que considerase tan sólo la varianza común, esto es, el análisis factorial antes que el de componentes principales.

Su punto de partida fue una matriz de correlación obtenida a partir de las frecuencias de distintos tipos de artefactos y características que le interesaban en los yacimientos individuales; se reproduce en la tabla 13.13. Como señala el mismo Bettinger, la distribución de esas frecuencias es extremadamente asimétrica, algo bastante común con distribuciones de frecuencias de este tipo. En estas circustancias, el coeficiente de correlación producto-momento, *r*, difícilmente proporcionará una buena descripción de la covariación entre las variables. Sin embargo, él lo usa basándose en el hecho de que, en general, tales datos tenderán a producir valores de correlación más bajos de lo que son en realidad; por consiguiente su análisis pecará de conservador. Aunque esto pueda ser cierto hasta cierto punto, es también probable que la descripción de las relaciones halladas por este medio inducirán a error, ya que en muchos de los puntos a lo largo de la recta de regresión las relaciones estarán mal especificadas por la línea —habrá una clara estructuración no aleatoria de los residuales—. Una técnica más apropiada habría sido el análisis de correspondencias (véase p. 281), que no requiere los mismos supuestos, si bien esta técnica era probablemente desconocida por el autor en el momento en que emprendió su análisis.

A pesar de estas dudas acerca de la idoneidad del uso de la correlación, el análisis sigue siendo de gran interés, y especialmente útil para mostrar una interpretación sustantiva de los resultados del análisis factorial.

TABLA 13.13. Matriz de correlación para las variables de los asentamientos (según Bettinger, 1979).

	P	Pm	Cer	Pf	Bi	Mb	Uni	Nu	Rt
Pavimento	—								
Piedra de molino	0,80	—							
Cerámica	0,02	0,39	—						
Punta de flecha	0,51	0,61	0,50	—					
Bifaz	0,40	0,62	0,51	0,83	—				
Modelo basto	0,06	0,21	0,44	0,59	0,58	—			
Unifaz	0,13	0,21	0,17	0,40	0,33	0,60	—		
Núcleo	0,16	0,29	0,16	0,24	0,26	0,36	0,82	—	
Restos de talla	0,08	0,19	0,29	0,35	0,31	0,48	0,84	0,82	—

El examen de la matriz de correlación sugiere que las piedras de molino y la aparición de pavimentos están muy relacionadas entre sí, lo cual también sucede con los bifaces y las puntas de flecha, y con muchos restos de talla y unifaces.

El procedimiento adoptado por Bettinger para analizar esta matriz fue efectuar en primer lugar un análisis de componentes principales para hallar el número de componentes con valores propios mayores que 1,0, obteniendo a continuación y rotando esa cantidad de factores, para lo que usó una matriz que contuviese las estimaciones de la comunalidad a lo largo de la diagonal principal. Tres componentes tenían un valor propio mayor que 1,0, por lo que se obtuvieron tres factores a partir de la matriz basada en las comunalidades, rotándolos a continuación. Los pesos de las variables en los factores rotados aparecen en la tabla 13.14, junto con las comunalidades de las variables. Es posible indicar gracias a estas últimas que la aparición de la cerámica no parece estar relacionada con otras variables, y que lo mismo ocurre, aunque en menor grado, con los modelos bastos.

TABLA 13.14. Rotación varimax de la matriz factorial (según Bettinger, 1979). Los definidores de los factores están indicados entre paréntesis.

Variable	Factor I	Factor II	Factor III	Comunalidad
Pavimento	0,05	0,07	(0,95)	0,91
Piedra de molino	0,11	0,37	(0,81)	0,80
Cerámica	0,10	(0,61)	0,06	0,38
Punta de flecha	0,17	(0,79)	0,43	0,84
Bifaz	0,14	(0,80)	0,37	0,80
Modelo basto	0,40	(0,67)	−0,05	0,61
Unifaz	(0,91)	0,25	0,05	0,89
Núcleo	(0,89)	0,07	0,15	0,82
Restos de talla	(0,88)	0,24	0,02	0,83
% de varianza	59,6	26,5	14,0	

Si examinamos ahora los pesos, veremos que tres de las variables están muy estrechamente asociadas con el factor I: unifaces, núcleos y restos de talla. Bettinger afirma que los unifaces representan la evidencia de trabajo sobre madera y los núcleos de sílex la evidencia de trabajo sobre piedra, mientras que los restos de talla proceden de la manufactura de útiles de piedra y su reparación. Su conclusión, por tanto, es que el factor I «refleja la manufactura de útiles para realizar actividades sobre madera y sobre piedra» (1979, p. 466).

El factor II se identifica con las puntas de flecha, bifaces, modelos bastos y cerámica. Obviamente representa un conjunto de actividades diferentes, y Bettinger concluye que indica «el conjunto básico necesario para iniciar y mantener asentamientos ocupados más de unos pocos días, es decir, asentamientos estables» (1979, p. 466).

El último factor está definido, obviamente, por los pavimentos y las piedras de molino. «Denota viviendas, características de almacenamiento y de preparación de alimentos; estas categorías dejan pocas dudas de que el factor III es un complejo de actividades domésticas que se emplearían a lo largo del año en los yacimientos, y estacionalmente, en otoño e invierno» (1979, p. 466). Es preciso señalar que el análisis factorial de Bettinger formaba parte de un estudio más amplio y que ya había identificado tres distintas categorías funcionales: asentamientos permanentes, asentamientos temporales y asentamientos piñoneros; estos últimos eran pequeños asentamientos ocupados a fines del otoño y en invierno, para facilitar la recolección de los piñones de pino.

Para investigar la relación entre estas categorías, previamente definidas, de yacimientos y los resultados del análisis factorial, Bettinger usó las puntuaciones factoriales de los yacimientos, calculando la media de los mismos para cada categoría de yacimientos, en cada uno de los tres factores (fig. 13.21) y extrayendo algunas conclusiones de los resultados:

> Útiles para hacer manufacturas de envergadura (talla) y sus subproductos; para el mantenimiento del campamento y para procurar recursos; los medios domésticos están muy bien representados en los asentamientos permanentes, y virtualmente ausentes en los asentamientos temporales (1979, p. 467).
>
> Los campamentos piñoneros se caracterizan por los útiles para el mantenimiento del campamento y los que permiten procurarse recursos, así como medios domésticos; lo que coincide con su función como asentamientos de otoño e invierno (1979, p. 467).
>
> Ambos están algo menos representados en los campamentos piñoneros que en los asentamientos; posiblemente a causa del uso estacional de los campamentos piñoneros.
>
> Los útiles para manufacturas de envergadura y sus subproductos son casi insignificantes en los campamentos piñoneros. A diferencia de los asentamientos permanentes, tanto los campamentos piñoneros como los asentamientos temporales eran, ante todo, lugares para procurarse recursos, de la misma manera que las actividades de manufactura ... estaban sistemáticamente ausentes (1979, p. 468).

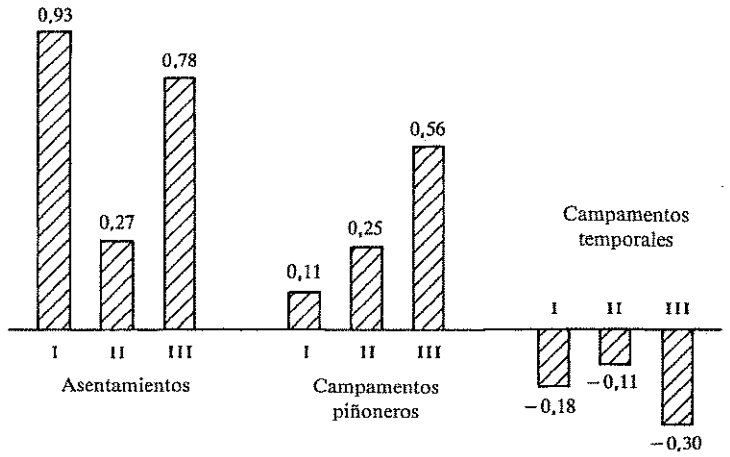


FIGURA 13.21. Puntuación media en los factores para cada categoría de asentamiento: I, factor 1; II, factor 2; III, factor 3 (según Bettinger, 1979).

El análisis de Bettinger proporciona un buen ejemplo de la manera en que el análisis factorial puede usarse en un estudio más amplio para conseguir resultados interpretables y sustantivamente útiles. Se definen conjuntos de variables interrelacionadas y se conectan al problema arqueológico en términos de un modelo claramente formulado. Sin embargo, para mostrar el cuidado con el que han de tratarse los métodos multivariantes para extraer una interpretación sustantiva, se presentan en la tabla 13.15 los pesos de las variables en un conjunto de componentes principales sin rotar y sus valores propios relevantes, derivados por el autor a partir de la matriz de correlaciones calculada por Bettinger.

Ambos conjuntos de resultados son, esencialmente, correctos, pero no son los mismos, dependiendo cada uno del interés de cada investigador. Bettinger estaba interesado en definir grupos de variables relacionadas antes que en resumir las principales dimensiones de variación en los datos; de ahí su selección del método. Otra discusión arqueológica de este tema puede encontrarse en el análisis que hizo Forsberg (1985) sobre los esquemas de subsistencia en asentamientos suecos de cazadores-recolectores.

Estas descripciones y ejemplos relativamente detallados del análisis de componentes principales y del análisis factorial se han expuesto para proporcionar al lector una impresión general del análisis multivariante y del tipo de tareas en las que puede emplearse. Los restantes métodos multivariantes a los que nos referíamos al principio del capítulo se tratarán con mucha más brevedad, por diversas razones. Después de la exposición del análisis de componentes principales y del análisis factorial, se requiere mucho menos espacio para dar una

TABLA 13.15. Resumen de los resultados del análisis de componentes principales: valores propios de los componentes y peso de las variables en los tres primeros componentes principales, derivados de la matriz de correlaciones entre los tipos de artefactos presentados en la tabla 13.13. (NOTA: los pesos no están normalizados, por lo que sus valores al cuadrado suman 1,0. Véase la nota al pie de la página 255 para estudiar la forma de convertirlos de manera que se correspondan a los coeficientes de correlación.)

Valores propios de los componentes	1	2	3	4	5	6	7	8	9
	4,319	1,980	1,228	0,627	0,305	0,233	0,134	0,101	0,074

Variable	Componente I	Componente II	Componente III
Pavimento	−0,230	0,413	0,539
Piedra de molino	−0,323	0,402	0,299
Cerámica	−0,267	0,123	−0,519
Punta de flecha	−0,394	0,262	−0,149
Bifaz	−0,381	0,261	−0,211
Modelo basto	−0,346	−0,122	−0,389
Unifaz	−0,358	−0,409	0,147
Núcleo	−0,322	−0,397	0,314
Restos de talla	−0,345	−0,419	0,112

idea básica de esos otros métodos; la falta de espacio impide presentar ejemplos detallados, los cuales, por otro lado, pueden consultarse en Doran y Hodson (1975) y Bølviken *et al.*, 1982).

Análisis de coordenadas principales

En muchos casos, al tratar con variables cualitativas y multiestado, tendremos una matriz de las similaridades entre una gran cantidad de elementos, del tipo de los presentados en el capítulo anterior. Las similaridades pueden basarse en las variables de una variedad de distintos tipos agrupados, usando el coeficiente general de similaridad de Gower. Un ejemplo como este es el que se usó al principio de este capítulo para explicar lo que eran los procedimientos de ordenación. El procedimiento descrito de forma intuitiva en ese ejemplo y que hacía referencia al estudio de unas tumbas (véase p. 244) es, de hecho, el procedimiento del análisis de coordenadas principales. En otras palabras, para crear un espacio en pocas dimensiones y entender la estructuración presente en los datos, no hemos de definir el espacio en *variables*; podemos hacerlo en términos del espacio definido por las *similaridades entre las unidades*. De la misma forma en que definíamos los componentes principales y sus valores propios asociados, podemos definir los ejes principales y los valores propios en el espacio creado por sus similaridades, es decir, por la matriz de similaridades. El tamaño de los valores propios facilita la información sobre la importancia de cada dimensión, explicando la variación en las distancias entre productos; se puede convertir, de la misma forma, en la cifra del porcentaje de distancia entre puntos en los datos, explicada por una dimensión en particular. De nuevo, son las dos otras primeras dimensiones las que explican la mayor parte. Los ejes principales del espacio nos permiten obtener el equivalente de una puntuación en el componente para cada caso y en cada uno de los ejes ortogonales. Así, pasaríamos de una representación relativa de los casos o puntos en relación con los demás, a su representación en términos de las posiciones determinadas por un conjunto de nuevos ejes. Podemos usar esos ejes para generar los diagramas de dispersión de los puntos, los cuales se examinarán para buscar tendencias o agrupaciones, al igual que hacíamos en los diagramas de dispersión de las puntuaciones en los componentes.

Sin embargo, la consecuencia del hecho de que el análisis de coordendas principales opera sobre las similaridades entre los casos y no sobre las correlaciones entre las variables es que no existen los pesos de las variables en los componentes para usarlos en la interpretación de los resultados. Es sólo el equivalente de las puntuaciones en los componentes lo que produce el análisis de coordenadas principales, y es esto lo que debe interpretarse para alcanzar el significado sustantivo de los nuevos ejes. Ahora bien, ya vimos que esto no suponía problema alguno. Hay que señalar que en algunos casos se sitúan a cada uno de los extremos de uno de los ejes, y hay que volver a examinar los datos originales en los que se basan las coordenadas para ver qué es lo que los diferencia el uno del otro. Por eso, si estudiásemos los resultados de un análisis de coordenadas principales sobre la matriz de similaridades entre unas tumbas, nos referiríamos al listado inicial de los valores de las variables que usamos para caracterizar las tumbas —diferentes tipos de objetos de ajuar, por ejemplo—. Por ese medio, obtendríamos un conocimiento arqueológico sustantivo de los principales factores subyacentes a la variación entre los enterramientos.

Escalas multidimensionales no métricas

Esta técnica trata, esencialmente, el mismo problema que el análisis de coordenadas principales, pero desde un ángulo diferente; para un conjunto de datos en particular, debiera producir unos resultados muy semejantes, en términos de la relación entre puntos en un espacio de pocas dimensiones que el método pretende definir.

La base conceptual de esta técnica es bastante sencilla. El punto de partida es el mismo que en el análisis de coordenadas principales: una medida de similaridad o disimilaridad entre n casos, y una representación de las relaciones entre los casos en un espacio multidimensional, siendo la cantidad de dimensiones menor que el número de casos. A partir de aquí, el método va reduciendo sucesivamente la cantidad de dimensiones en cuyos términos están representa-

dos los puntos, y manteniendo en el mínimo, al mismo tiempo, la distorsión en las relaciones entre los puntos, que empieza a aparecer cuando el número de dimensiones se reduce. El rasgo específico del método es que, a diferencia del análisis de coordenadas principales, no es «métrico»: no opera sobre los valores numéricos de las similaridades/distancias entre los casos, sino sobre su ordenación. Es decir, el método intenta preservar la ordenación de las distancias/similaridades entre los puntos a medida que se reduce el número de dimensiones. Así, en el caso de que la distancia entre el punto x_i y el punto x_j sea la décima menor en la matriz de distancias, ha de seguir siendo la décima menor tras haber reducido la dimensionalidad de la representación multidimensional. Naturalmente no ha de hacer esto sólo para un par de distancias, sino para todas a un tiempo: la ordenación de todas las distancias en el espacio reducido ha de corresponder a la ordenación original entre todas ellas. No es preciso insistir en la dificultad de esta manipulación —de hecho, si los casos son muy numerosos, requiere bastante tiempo de cálculo en el ordenador—, por lo que es casi imposible que la disposición original de los puntos según el orden de sus distancias se mantenga exactamente en un espacio de pocas dimensiones. Un aspecto clave del método de las escalas multidimensionales no métricas es que proporciona una medida del éxito con que se mantiene el orden en la representación dimensionalmente reducida. Esa medida recibe el nombre de «estrés» (¿quizás como reflejo de los círculos psicológicos en los que empezó a usarse el método?), un indicador del grado en que la ordenación de las distancias en una cantidad reducida de dimensiones se diferencia del orden original.

Igual que los valores propios asociados a los ejes en el análisis de componentes o en el de coordenadas principales indican la importancia de aquellos ejes explicando la variación en los datos, de esa misma manera el estrés proporciona una medida del número de dimensiones importante en la representación de los datos en una escala multidimensional no métrica. El estrés se calcula a cada número sucesivamente decreciente de dimensiones, intentando buscar aquella cantidad de dimensiones en particular en la que repentinamente se aprecie un mayor incremento del estrés; esto indicaría un aumento repentino en la cantidad de distorsión en los datos, interpretable como la eliminación de una dimensión importante para la explicación de la variación en los números; algunos programas de ordenador tienen la opción para averiguar la cantidad correcta de dimensiones.

La idea, al igual que en el análisis de componentes y coordenadas principales, es que, si la variación en un conjunto de datos se puede reducir a una pequeña cantidad de tendencias mayores o esquemas (como en el caso de la morfometría de una vasija), el estrés en la cantidad apropiada de dimensiones será inferior que en datos comparables en los que la variación no sea reducible de esa forma. Un valor de estrés bajo corresponderá a aquel caso en el análisis de coordenadas principales en el que el mismo número de dimensiones explique un alto porcentaje de la varianza. Además, el significado arqueológico sustantivo de las dimensiones podrá estudiarse de la misma forma que en el análisis de coordenadas principales, investigando la posición de los diferentes casos en los ejes o dimensiones.

Las escalas multidimensionales no métricas han sido usadas ampliamente en arqueología. El análisis de las necrópolis egipcias predinásticas realizado por Kemp (1982) constituye un ejemplo, mientras que Cherry y otros han usado esta técnica para hacer mapas de los estados egeos en la edad del bronce (por ejemplo, Cherry, 1977; Kendall, 1977). Otros ejemplos aparecen citados por Doran y Hodson (1975), que también exponen esta técnica con cierto detalle. Una presentación introductoria exhaustiva, aún esencial, es la de Kruskal y Wish (1978). La discusión sobre esta técnica se centra en sus ventajas y desventajas relativas a otros medios, como aquellos a los que la hemos comparado. No es posible profundizar más aquí, si bien Doran y Hodson (1975, pp. 214-217) proporcionan una buena exposición de esas cuestiones, al igual que Gordon (1981, pp. 91-101).

Análisis de correspondencias

El uso de esta técnica en arqueología es relativamente reciente, lo cual se refleja en su ausencia en el libro de Doran y Hodson. De hecho, se desarrolló a fines de los años sesenta, especialmente en Francia (no obstante, véase Hill, 1973). En cierto sentido, la relativa lentitud de su adopción por los arqueólogos angloamericanos se debe a la falta de comunicación entre ellos y el mundo arqueológico francés, netamente diferenciado; la velocidad con que esa técnica fue apreciada en la arqueología francesa está indicada por Djindjian (1980).

El método está bien descrito por Bølviken *et al.* (1982),* por lo que no se va a detallar aquí, pues, en cualquier caso, no hay espacio suficiente. El objetivo al incluirlo ha sido tan sólo para llamar la atención sobre una técnica importante y señalar las características especiales que lo hacen particularmente útil.

De las técnicas que hemos examinado hasta aquí, el análisis de correspondencias es la que más se parece al análisis de componentes principales, con el que comparte los mismos principios generales. Sin embargo, ya hemos visto que este último suele tratar una matriz de correlaciones, covarianzas o sumas de cuadrados y productos, a partir de las cuales se derivan las correlaciones;

* Los lectores que no dominen la lengua inglesa o la francesa, pueden consultar con provecho los varios artículos que sobre el análisis de correspondencias —quizás la técnica estadística multidimensional más «popular» en España en estos momentos— han sido incluidos en las *Actas del I Coloquio Español sobre Aplicaciones Informáticas en Arqueología*, compilado por V. M. Fernández Martínez y G. Fernández, publicado en 1990 por la Universidad Complutense de Madrid. (*N. del t.*)

estas medidas sólo son satisfactorias como medidas de asociación si los datos son numéricos y la distribución de las variables no se aparta de la normalidad.

El análisis de correspondencias no tiene esa limitación, puesto que fue diseñado para analizar datos que consistiesen en frecuencias, tales como la frecuencia de aparición de varios tipos de artefactos, analizada por Bettinger en el ejemplo anterior. De hecho, el análisis de coordenadas principales puede usarse para resolver algunos de los problemas que se plantean si las variables no están distribuidas normalmente o no son numéricas. Sin embargo, el análisis de coordenadas principales, como las escalas multidimensionales no métricas, supone el cálculo previo de los coeficientes de similaridad, antes que operar directamente sobre los datos originales, como hace el análisis de componentes principales, y el cálculo de los coeficientes de similaridad implica ya una pérdida de información. Podemos tener un cierto valor para el coeficiente de Jaccard, por ejemplo, que mida la similaridad entre dos casos individuales, pero no sabemos con precisión qué es lo que los dos tienen en común y qué los distingue. El análisis de correspondencias no requiere este paso previo; como el análisis de componentes principales, empieza con el valor de las variables en casos particulares.

Tiene una propiedad adicional, ausente en el análisis de componentes principales (sin embargo, véase la técnica llamada *bitrazado* [*biplot*] en Gabriel [1981]). Como vimos, en el caso del análisis de componentes principales es posible estudiar la distribución de las puntuaciones en el componente en un diagrama de dispersión y ver qué casos son similares entre sí. Igualmente, vimos que las relaciones entre las variables, y entre variables y componentes, pueden representarse también gráficamente. Con el análisis de correspondencias, las relaciones entre casos, las relaciones entre variables y las relaciones entre variables y casos pueden analizarse conjuntamente y representarse en el mismo diagrama de dispersión o serie de diagramas generados al trazar pares de ejes ortogonales. La posibilidad de unir relaciones entre variables particulares directamente con similaridades entre casos particulares es muy significativa desde el punto de vista de la interpretación, y en cierto sentido socava los argumentos que han inundado la literatura arqueológica sobre el tema, especialmente en lo que se refiere a los tipos arqueológicos y si han de tratarse por medio de correlaciones entre variables o similaridades entre casos (véanse los artículos en Whallon y Brown, 1982).

Puede ser útil mostrar los rudimentos del análisis de correspondencias, resumiendo uno de los ejemplos presentados por Bølviken *et al.* (1982). Se disponía de información procedente de la excavación de varios yacimientos mesolíticos acerca de las frecuencias de aparición de 37 tipos de objetos líticos en 14 cabañas. Tras un análisis preliminar se decidió agrupar los 37 tipos en tan sólo 9 categorías basadas en supuestos tales como su función. El objetivo del estudio era ver qué hogares eran similares entre sí, qué categorías funcionales estaban relacionadas en términos de formas de deposición, y la manera en que las similaridades entre los hogares se relacionaban con las categorías funcionales.

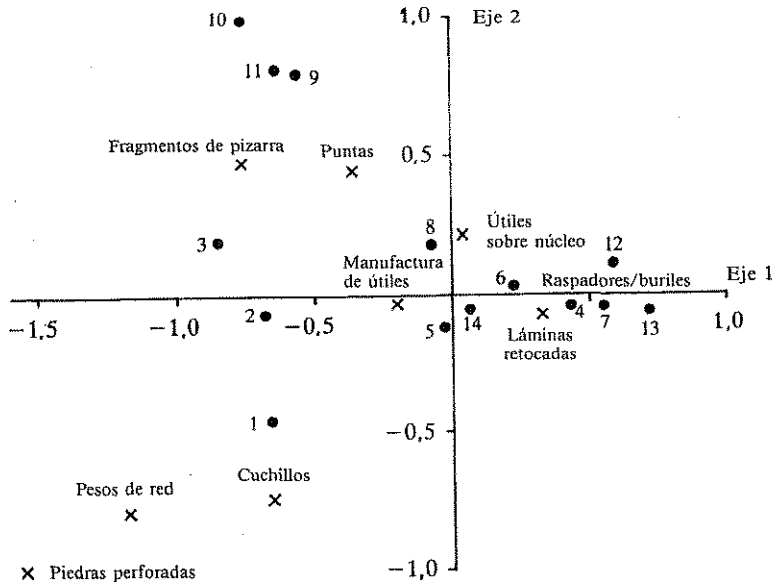


FIGURA 13.22. Análisis de correspondencias de las categorías funcionales y unidades de habitación en Iversfjord (según Bølviken *et al.*, 1982).

Se efectuó un análisis de correspondencias con las 14 cabañas como casos, y las frecuencias en las 9 categorías de útiles como variables.

En la figura 13.22 aparece el diagrama de dispersión basado en los dos primeros ejes, los cuales explican, conjuntamente, el 73 % de la variación. Se aprecia que tanto las variables como los hogares están representados en el diagrama. El primer eje, que explica el 53 % de la variación, opone raspadores, buriles y láminas retocadas, por un lado, y los pesos de red y las piedras perforadas, por otro; los autores lo interpretan como el contraste entre las actividades de mantenimiento y la pesca. El segundo eje, que explica el 20 % de la variación, muestra las puntas de flecha y los fragmentos de pizarra en un extremo, y los cuchillos de pizarra, pesos de red y piedras perforadas en el otro. Según los autores del estudio, representaría el contraste entre caza y pesca.

Interesaba particularmente la cuestión de si los grupos de cabañas excavadas eran o no asentamientos invernales de pesca. Los autores argumentaron que si ese era el caso, todas las cabañas se incluirían en un conglomerado único, basado en los artefactos indicativos de las actividades invernales de pesca. Como señalan, no hay indicios en el diagrama de dispersión de que esto sea así; por el contrario, aparecen tres conglomerados de hogares. Además, como tratamos con los resultados de un análisis de correspondencias, podemos seguir exami-

nando el diagrama para buscar casos asociados con variables. Los cabañas 4, 5, 6, 7, 8, 12, 13 y 14 están en la misma área que los raspadores, buriles y láminas retocadas, con lo que estarán asociadas a actividades de mantenimiento. Las cabañas 9, 10 y 11 aparecen cerca de los fragmentos de pizarra y puntas de flecha, elementos interpretados por los autores de ese estudio como categorías de útiles relacionadas con actividades de caza. Finalmente, las cabañas 1, 2 y 3 aparecían asociadas a actividades de pesca. Los autores concluyen que «este análisis indica una mayor diversidad en la orientación económica (caza pesca y mantenimiento), así como grados de permanencia del asentamiento (actividades diversificadas a largo plazo vs. actividades específicas a corto plazo) distintos de lo que se creía posible para yacimientos prehistóricos costeros» (Bølviken *et al.*, 1982, p. 47).

Es posible discutir esta interpretación de los resultados según el modelo de los autores de la relación entre esquemas de actividad y deposición arqueológica (véase Binford, 1981; 1983), pero la utilidad de la técnica en lo que se refiere al descubrimiento de las relaciones entre casos y variables y entre los dos es innegable.

Análisis discriminante

Antes de abandonar los análisis multivariantes hay una técnica más que debe mencionarse, aunque sólo sea brevemente, si bien es muy distinta a las que hemos visto hasta ahora. Todas ellas investigaban la estructura subyacente a un conjunto de datos, con muy pocos supuestos previos acerca de a qué debe parecerse esa estructura, exceptuando, hasta cierto punto, el análisis factorial. La técnica del *análisis discriminante* supone que podemos dividir nuestras observaciones en grupos, según unos criterios, y a continuación intenta encontrar una forma de distinguir los mismos grupos, basándose en criterios independientes derivados de los mismos datos.

De hecho, el procedimiento ya se ha mencionado en el capítulo 12, donde distinguíamos *clasificación* de *discriminación* y se propuso un ejemplo de lo último. Podemos tener una cantidad de cerámicas sin decorar de un cierto tipo, encontradas en diferentes yacimientos. ¿Difiere su forma, descrita en medidas de proporción, entre los diversos yacimientos? En el análisis discriminante es el usuario el que dice al programa de ordenador qué vasijas proceden de qué yacimientos; el análisis intenta reproducir correctamente, a continuación, la asignación de las cerámicas a los yacimientos, pero basándose tan sólo en las medidas que describen la forma de esas cerámicas. Si tiene éxito en eso, querrá decir que efectivamente las cerámicas son distintas en los diferentes yacimientos; en el caso contrario, tales diferencias no serán significativas. Se trata de un principio, ciertamente, muy útil. Un área de la investigación arqueológica en la que se ha utilizado mucho es en los estudios de caracterización de artefactos, en

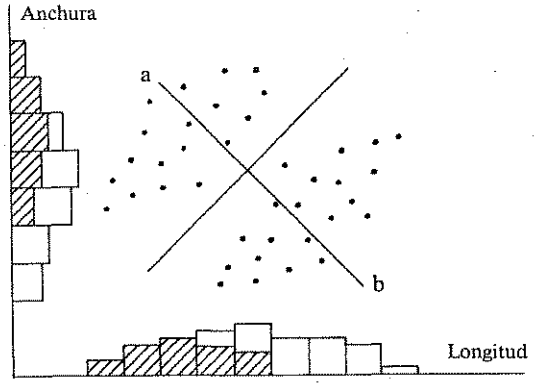


FIGURA 13.23. Discriminación entre dos grupos de restos de talla, a partir de la longitud y anchura de las piezas individuales.

donde las cantidades de elementos-traza en los artefactos líticos o en la cerámica se utilizan para discriminar materiales procedentes de distintas fuentes.

El análisis discriminante se expone en este capítulo y no en el anterior porque es un método de análisis multivariante. Construye un conjunto de variables a partir de las variables originales, al estilo del análisis de componentes principales, pero siguiendo el criterio de que esas variables han de maximizar las diferencias entre los distintos grupos —los distintos yacimientos en el ejemplo anterior.

Su funcionamiento se entenderá mejor usando un diagrama para el caso bivariante (fig. 13.23). Supongamos que tenemos grupos de láminas de sílex de dos fases diferentes en un mismo yacimiento, y que hemos descrito el tamaño y la forma de las láminas en términos de las medidas de longitud y anchura. ¿Es distinta la forma de esas láminas en las dos fases?

Claramente podemos ver que en este caso bivariante en particular lo son (naturalmente en un caso multivariante real no seríamos capaces de apreciarlo). Por otro lado, ninguna de nuestras variables originales, longitud y anchura, distingue los dos grupos. Hay una superposición entre ambas, como revelan los histogramas de los dos grupos, proyectados en cada uno de los ejes. Sin embargo, existe un único eje que distinguiría los dos grupos perfectamente (la línea *a-b* en el diagrama) que está compuesta tanto por la longitud como por la anchura. Se trata de la *función de discriminación* propia de este caso; una línea que la biseccione en ángulo recto entre los dos grupos los divide con éxito.

Los programas informáticos de análisis discriminante efectúan esta operación en el caso multivariante y, de la misma forma que para el análisis de componentes principales, obtenemos el valor propio de cada función y la contribución de cada una de las variables originales a ella.

La exposición anterior no pretende más que dar una idea general de cómo funciona esa técnica, que es demasiado importante como para omitirla. Está descrita ampliamente en muchos libros (por ejemplo, Davis 1971, Norusis 1985) y está disponible en los paquetes de programas estadísticos enumerados en el anexo 2. Como ya se ha indicado, no es una técnica exploratoria al estilo del análisis de conglomerados o el análisis de componentes principales, pero puede ser muy útil al contrastar las hipótesis arqueológicas.

La presentación del análisis discriminante finaliza nuestra introducción al análisis multivariante en arqueología. Resulta obvio que se trata de un tema muy complejo del que sólo hemos arañado su superficie. Doran y Hodson (1975) lo desarrollan mucho más, al igual que Johnston (1978), desde una perspectiva geográfica. El lector no debiera embarcarse directamente en el análisis multivariante de sus propios datos, basándose en lo que ha aprendido en este capítulo, sin la ayuda y el consejo de un experto; no obstante, habrá adquirido alguna idea de por qué se usan esas técnicas, así como una base para entender los ejemplos arqueológicos publicados.

Ejercicios

13.1. Efectúa un análisis de componentes principales con los datos del ejercicio 12.4 (cuencos de Uruk). Comenta todos los aspectos de los resultados que consideres relevantes, prestando especial atención a la interpretación arqueológica de los componentes y a su significabilidad.

13.2. Efectúa un análisis factorial de los mismos datos, distinguiendo entre la varianza común y la única. Rota una cantidad apropiada de factores y vuelve a interpretar los resultados.

13.3. En un estudio de la variación temporal y espacial del arte rupestre prehistórico en Australia, se examinaron 83 yacimientos (Morwood, 1980). En cada yacimiento se registró el color de los distintos símbolos artísticos. Por consiguiente, en cada yacimiento se codificó la cantidad de apariciones de los símbolos de un color en particular, y el porcentaje relativo de la aparición de las distintas categorías de color. Los colores eran:

1. Rojo
2. Púrpura
3. Naranja
4. Amarillo
5. Blanco
6. Negro
7. Marrón
8. Rosa

Se ha realizado un análisis de componentes principales en los dos conjuntos de datos. La figura 13.24 muestra los diagramas de dispersión de las variables en relación a los componentes principales y a los datos relevantes que pertenecen a esos diagramas. Razona los resultados arqueológica y estadísticamente.

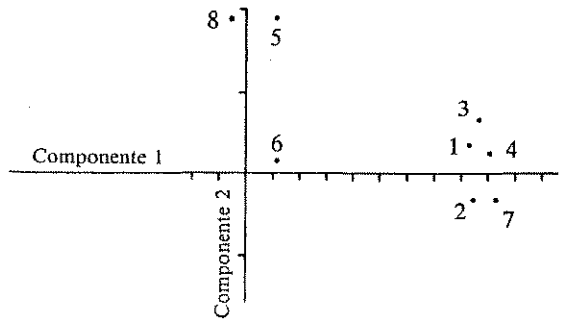
13.4. Las tablas 13.16 y 13.17 muestran el resultado de un análisis de componentes principales de 15 conjuntos de artefactos líticos surafricanos, descritos según el porcentaje de los distintos tipos que intervienen en su composición (Cable, 1984). Razona esos resultados arqueológica y estadísticamente y, con una atención particular, *a*) indica cuáles son a tu juicio los esquemas de variación más importantes, y *b*) describe y explica la relación entre los datos en la matriz de correlaciones y la relación en la matriz factorial.

13.5. Los datos son, otra vez, 15 conjuntos de artefactos líticos procedentes de Suráfrica. La figura 13.25 muestra un gráfico tripolar de las relaciones entre los 15 conjuntos según los datos que implican los tres tipos en la clave explicativa.

La figura 13.26 muestra los resultados de un análisis de conglomerados de esos conjuntos, según la lista completa de tipos de útiles dada antes para el ejercicio 13.4 (se usa el método de Ward). La figura 13.27 muestra los mismos conjuntos que la figura 13.26, basada aún en la lista completa de tipos, trazada para dos componentes principales, tal y como han sido definidos en el ejercicio 13.4.

Nótese que los conjuntos B y D no están presentes en las figuras 13.26 y 13.27; sólo está presente la versión amalgamada del conjunto X. Los conjuntos L y M también han sido eliminados. Razona las relaciones entre los resultados que aparecen en esas figuras; si es necesario, te puedes referir a la respuesta que diste en el ejercicio 13.4. ¿Qué conclusiones arqueológicas son posibles, en particular referentes a los motivos de la variación subyacente?

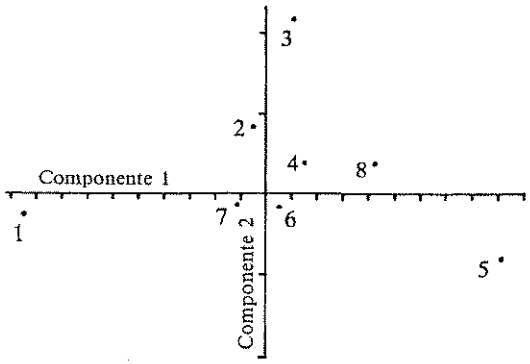
13.6. La excavación de un asentamiento en el norte de Noruega proporcionó gran cantidad de huesos de animal procedentes de catorce estratos (el estrato 1 era el más moderno y el 14 el más antiguo). Las frecuencias de esos huesos aparecen en la tabla 13.18. Para investigar la estructuración aparente en la composición de esos conjuntos a lo largo del tiempo, se efectuó un análisis de correspondencias. El diagrama de dispersión de los estratos y de las especies sobre los dos primeros ejes aparece en la figura 13.28. El primer eje explica el 68,8 % de la variación, el segundo el 29,54 %. Razona los resultados para los estratos y para las especies, así como su relación entre ellos (datos según Mathiesen *et al.*, 1981).



(a)

Componente	Valor propio	Varianza %	Varianza acumulada %
1	4,04	50,6	50,6
2	1,51	18,9	69,4
3	1,02	12,8	82,2

Comunalidades de las variables:
1 0,81 2 0,81 3 0,73 4 0,83 5 0,79 6 0,79 7 0,88 8 0,80



(b)

Componente	Valor propio	Varianza %	Varianza acumulada %
1	1,92	23,9	23,9
2	1,37	17,2	41,1
3	1,08	13,6	54,7

Comunalidades de las variables:
1 0,97 2 0,24 3 0,81 4 0,60 5 0,95 6 0,93 7 0,63 8 0,29

FIGURA 13.24. (a) Análisis de componentes principales de las frecuencias de aparición de los distintos colores. (b) Análisis de componentes principales del porcentaje de las frecuencias de aparición de los distintos colores (según Morwood, 1980).

TABLA 13.16. Estadísticas de los útiles procedentes de 15 yacimientos de Natal (Suráfrica), y sus coeficientes de correlación (según Cable, 1984).

Variable	Nombre	Media %	Desv. típica
P2	Raspador convexo	63,9720	15,0613
P3	Láminas con dorso	6,5337	7,2994
P4	Segmentos	3,1457	4,7784
P5	Puntas	0,6480	0,7627
P6	Piezas con dorso	1,3701	2,0794
P7	Raspadores con muesca	22,6633	18,0055
P8	Punzones	0,0698	0,1853
P9	Perforadores	0,8013	0,9494
P10	Hachas con dorso	0,7960	1,4124

Coeficientes de correlación

	P2	P3	P4	P5	P6	P7	P8	P9
P2	1,00000	−0,21502	−0,26859	−0,03987	−0,24972	−0,66241	0,13254	0,16628
P3	−0,21502	1,00000	0,95950	0,35566	0,81085	−0,57833	−0,20933	0,01052
P4	−0,26859	0,95950	1,00000	0,31807	0,81109	−0,51279	−0,24782	0,00711
P5	−0,03987	0,35566	0,31807	1,00000	0,38298	−0,27684	0,21841	−0,35674
P6	−0,24972	0,81085	0,81109	0,38928	1,00000	−0,46210	−0,24550	0,11475
P7	−0,66241	−0,57833	−0,51279	−0,27684	−0,46210	1,00000	0,01090	−0,19299
P8	0,13254	−0,20933	−0,24782	0,21841	−0,24550	0,01090	1,00000	0,24369
P9	0,16628	0,01052	0,00711	−0,35674	0,11475	−0,19299	0,24369	1,00000
P10	0,80889	−0,11428	−0,27892	0,14748	−0,10471	−0,00265	0,31651	−0,07173

TABLA 13.17. Estadísticas de los útiles procedentes de 15 yacimientos de Natal (Suráfrica): matriz factorial (según Cable, 1984).

Factor	Valor propio	% de varianza	Varianza acumulada
1	3,38154	37,6	37,6
2	1,79047	19,9	57,5
3	1,46956	16,3	73,8
4	1,10192	12,2	86,0
5	0,72836	8,1	94,1
6	0,35722	4,0	98,1
7	0,14690	1,6	99,7
8	0,02400	0,3	100,0
9	−0,00000	−0,0	100,0

Matriz factorial usando el factor principal, sin iteraciones

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
P2	−0,14224	0,85984	−0,20478	−0,44221	−0,21555	0,02352	−0,04362	0,00811
P3	0,95873	−0,02408	0,02118	0,09423	0,08752	−0,20689	0,09981	−0,10247
P4	0,95713	−0,11986	−0,06601	0,08458	−0,03734	−0,20109	0,05742	0,11365
P5	0,45268	0,12302	0,73469	−0,13820	−0,32945	0,30397	0,14199	0,00353
P6	0,89934	−0,06437	0,00905	0,18148	0,11008	0,25582	−0,27644	−0,00641
P7	−0,62763	−0,72569	0,12421	0,23780	−0,05558	0,06618	−0,00204	0,00340
P8	−0,27343	0,47956	0,43771	0,53560	−0,38088	−0,23932	−0,12023	−0,00435
P9	0,00186	0,41236	−0,54669	0,67641	0,01094	0,23200	0,14002	0,00420
P10	−0,20894	0,30053	0,61410	0,19336	0,67076	−0,01033	0,03315	0,02048

	Conjunto	Zona*	*n*	Útiles % R.C.*	% R.M.*	% P.F.*
A	—Umbeli Belli	C.C.	98	69,4	8,2	22,4
B	—Borchers Shelter Layers 1+2	C.C.	123	42,3	8,9	48,8
C	—Borchers Shelter Layer 3	C.C.	173	56,6	11,0	32,4
D	—Borchers Shelter Annexe	C.C.	325	77,2	9,2	13,5
E	—The Falls	C.C.	44	70,5	6,8	22,7
F	—Good Hope Layer 1	T.A.	41	65,9	34,1	0
G	—Good Hope Layer 2	T.A.	497	64,2	31,6	4,2
H	—Bottoms Up Shelter	T.A.	140	67,1	29,3	3,6
I	—Giant's Castle	T.A.	246	67,1	30,9	2,0
J	—Belleview, Spits 1-4	T.A./P.A.	92	56,5	32,6	10,9
K	—Belleview, Spits 5-8	T.A./P.A.	192	87,0	5,2	7,8
L	—Grindstone Shelter	T.A.	49	34,7	65,3	0
M	—Karkloof	C.B.	149	49,0	50,3	0,3
N	—Moshebi's Shelter	P.A.	309	88,7	6,8	4,5
O	—Sehonghong	P.A.	206	80,6	15,0	4,4
X	—Borchers Site Complex (Borchers 1+2 & Annexe)	C.C.	446	67,9	9,2	23,3

* C.C. = Cinturón costero
T.A. = Tierras altas
P.A. = Praderas alpinas
C.B. = Cinturón boscoso

R.C. = Raspadores convexos
R.M. = Raspadores con muesca
P.F. = Puntas de flecha

FIGURA 13.25. Natal y Lesotho oriental: gráfico tripolar de la variabilidad de los conjuntos, y leyenda explicativa (según Cable, 1984).

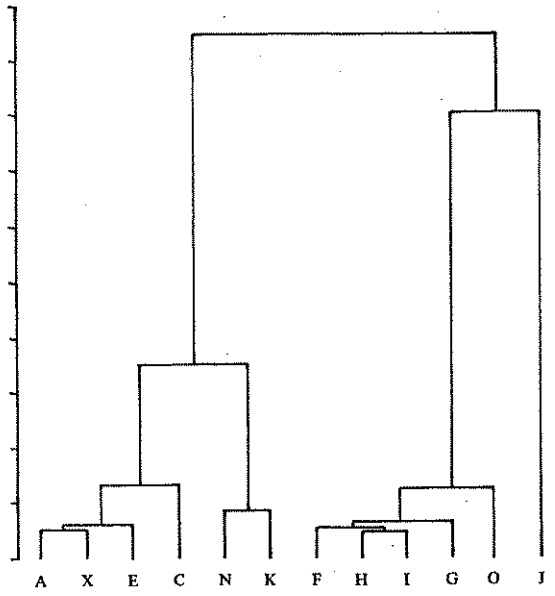


FIGURA 13.26. Análisis de conglomerados de los conjuntos de útiles líticos por medio del método de Ward (según Cable, 1984).

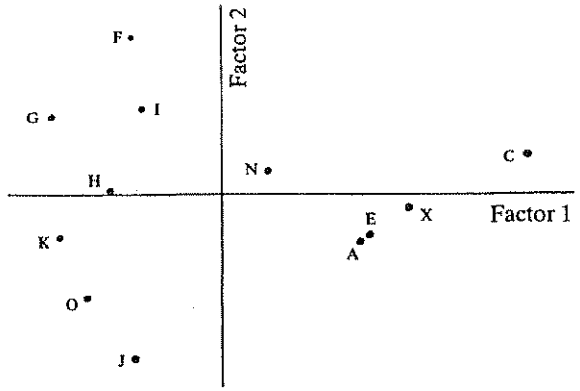
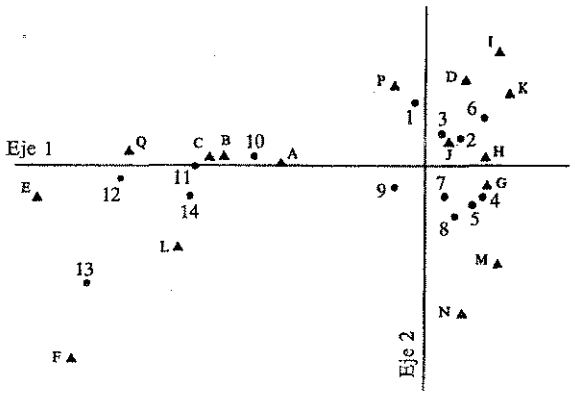


FIGURA 13.27. Los conjuntos de útiles líticos representados con relación a los dos primeros componentes principales (según Cable, 1984).

TABLA 13.18. Material osteológico de los estratos 1-14 del corte 1 de la excavación del túmulo de la isla de Helgøy (véase la leyenda de la figura 13.28 para los detalles) (según Mathiesen *et al.*, 1981).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q	Suma
1	27	42	33	1	3	0	272	5	40	31	3	0	0	0	17	0	474
2	54	122	35	0	4	0	1.080	36	15	73	11	1	0	0	47	0	1.578
3	44	83	54	3	4	0	842	24	71	81	35	12	0	0	32	0	1.284
4	101	151	90	2	6	0	3.247	14	128	81	20	23	3	0	34	1	3.901
5	101	202	58	4	0	4	3.204	95	99	216	24	92	33	0	22	4	4.158
6	43	61	33	6	13	1	1.082	37	170	138	17	1	5	0	4	4	1.615
7	24	40	17	0	23	4	545	23	3	88	3	0	1	3	2	1	777
8	17	24	14	1	30	3	597	22	4	46	4	1	0	0	2	0	765
9	27	42	10	2	14	2	294	8	7	33	2	11	0	0	9	0	461
10	24	53	20	0	6	0	100	6	3	22	0	28	0	0	14	1	277
11	45	78	35	0	30	1	128	4	0	20	0	17	1	0	2	9	379
12	109	367	167	1	142	8	348	1	13	38	1	80	1	0	13	6	1.294
13	15	18	17	0	7	8	25	0	0	4	0	14	0	0	0	1	109
14	42	41	34	0	15	3	93	0	0	0	0	14	0	0	6	0	248
Suma	673	1.324	617	20	297	34	11.857	275	662	871	118	294	44	3	204	27	17.320



Estratos 1-14

A	Bóvidos	*Bos taurus*
B	Ovicápridos	*Ovis aries/Carpa hircus*
C	Porcino	*Sus scrofa dom.*
D	Reno	*Rangifer tarandus*
E	Foca	Fócidos
F	Lagópodo	*Lagopus*
G	Abadejo	*Gadus morrhua*
H	Arenque	*Melanogrammus aeglefinus*
I	Merluza	*Pollachius virens*
J	Gádidos	*Molva molva*
K		*Brosme brosme*
L	Halibut	*Hippoglossus hippoglossus*
M		*Sebastes marinus*
N		*Anarchichas lupus*
P	Pingüino	Álcidos
Q	Gallo	*Gallus gallus f. dom.*

FIGURA 13.28. Análisis de correspondencias de los datos de la tabla 13.18 (según Mathiesen *et al.*, 1981).

14. MUESTREO PROBABILÍSTICO EN ARQUEOLOGÍA

Durante más de dos décadas, el «muestreo» ha configurado una parte importante de los nuevos enfoques en arqueología (Vescelius, 1960). Partidarios y oponentes de los nuevos métodos arqueológicos lo consideran un tema fundamental (Binford, 1964; Hole, 1980). Cuando un método se convierte en parte integrante de la ideología de un enfoque, los resultados suelen ser insatisfactorios, y el muestreo en arqueología no es ninguna excepción (véase Wobst, 1983); ha sido rechazado o aceptado como si de un axioma se tratase y no basándose en argumentos bien razonados en su idoneidad para análisis específicos. ¿Qué es, en realidad, el muestreo y por qué tiene esa importancia?

En un sentido muy general, el muestreo abarca la noción general de usar la información de una parte de algo para hacer inferencias acerca del todo. Ya que los arqueólogos son conscientes de la naturaleza parcial de las evidencias que recuperan —tan sólo recuperan una «muestra»—, la idea de una metodología que pueda ayudarles a resolver los problemas planteados por esa situación tiene una atracción casi mística. Se han desarrollado recientemente algunos modelos matemáticamente muy sofisticados para resolver esta cuestión (por ejemplo, Orton, 1982), si bien no es esa la solución. Desafortunadamente, abundan los errores resultantes de creer que lo es, conduciendo, de un lado, al superoptimismo y, del otro, al rechazo de lo que se consideran exigencias no realistas sobre el uso potencial del muestreo en arqueología.

Es importante aclarar desde el principio que al muestreo le incumben las inferencias acerca de alguna parte específica del registro arqueológico existente, basándose en el estudio de un subconjunto de esa parte del registro arqueológico. *No* le incumben las inferencias que llevan del registro existente (o poblaciones de hallazgos físicos, como Cowgill [1970] las ha denominado) a los resultados materiales de la conducta humana del pasado (poblaciones de consecuencias físicas, según Cowgill), y aún menos le incumben aquellas inferencias que tratan acerca de la naturaleza de la conducta que produjo ese registro.

Naturalmente, en este sentido, los arqueólogos siempre han practicado el muestreo. Han elegido unos yacimientos para excavar en ciertas regiones, y han

hecho trincheras y catas en ellos, sin restringir sus conclusiones a esos sondeos específicos. La transformación que tuvo lugar hace 25 años fue la introducción y la justificación del muestreo *probabilístico*: la selección de una parte del registro arqueológico para investigarlo, de forma que se emplee la teoría de la probabilidad para evaluar las inferencias que partiendo de la parte se refieren al todo del que procede esa parte, y ello basándose en la probabilidad de corrección de las inferencias. Así pues, lo que se pretende es asegurar que la muestra elegida es representativa de la totalidad. La teoría de las probabilidades proporciona las reglas para hacer la selección y plantear las inferencias a partir de esa selección.

Ahora bien, ¿qué inferencias pueden hacerse de ese modo y en cuáles estamos realmente interesados? Esta cuestión debiera llevarnos a plantearnos los objetivos de un proyecto en particular y el diseño de la investigación apropiado para su realización. Desafortunadamente, a esta parte del procedimiento de investigación no se le ha presentado la atención que se merece. Lo más frecuente es que el nivel de la investigación esté determinado por las exigencias financieras de una u otra forma, por lo que ha habido una tendencia a considerar, por ejemplo, que la excavación de urgencia en un yacimiento amenazado, al ser necesariamente parcial, debe seguir un esquema de muestreo estadístico, actitud que conlleva la falta de claridad en los objetivos. El investigador no sabe nada sobre las afirmaciones que pueden hacerse basándose en la muestra; tan sólo que la muestra ha sido estadísticamente seleccionada, lo cual ha de ser, necesariamente, mejor. Este enfoque puede que no sea perjudicial y puede que la distribución de las unidades de excavación o prospección que proceden de él produzcan resultados de interés, si bien no es algo que tenga mucho que ver con la inferencia estadística.

La teoría del muestreo estadístico es relevante cuando el propósito del estudio es usar la muestra elegida para estimar las características de la población de la cual procede. En estas circunstancias, el objetivo es extraer una muestra que sea una «buena representación» de la población y que permita estimar sus características con la mayor precisión posible, dado el coste o esfuerzo invertido en ello (véase Barnett, 1974).

Se considera bien establecido hoy en día que los enfoques de muestreo «prácticos» utilizados en arqueología no son muy satisfactorios, aunque para hacer justicia a los primeros investigadores que los aplicaron hay que señalar que pocas veces lo que se pretendía era una buena representación de la población general. Podemos distinguir dos tipos de muestreo práctico, ambos muy importantes en arqueología:

1. Muestreo por accesibilidad: en este enfoque, el factor clave es la facilidad del acceso a las observaciones; las más fácilmente obtenibles eran las elegidas. Un ejemplo arqueológico podría ser el de las prospecciones de los asentamientos en el Próximo Oriente, que consisten en conducir a lo largo de las carreteras buscando los tells visibles desde ellas. La versión opuesta podría llamarse muestreo por inaccesibilidad, y es la practicada por algunos arqueólogos en el área mediterránea, que buscan colinas prominentes y suben a ellas para ver si hay yacimientos en su cima. No hay nada necesariamente erróneo en este enfoque, siempre y cuando nos proporcione información representativa de la densidad de ocupación o de las proporciones relativas de un tipo de yacimiento en un área.

2. Muestreo guiado por un propósito específico: se trata de investigadores que hacen su propia selección del yacimiento a excavar o del área a prospectar. El objetivo puede ser, o no, obtener una muestra representativa. Si lo es y el investigador dispone de abundantes conocimientos sobre la materia, la muestra podrá ser una representación excelente de la población. El principal problema con este enfoque, si el objetivo fuese lograr la representatividad, no es que las estimaciones que se consigan estén distorsionadas, sino que no se dispone de medios para evaluar la representatividad de la muestra elegida, a no ser por medio de la evaluación del selector y nuestro conocimiento de la situación.

Si usamos un esquema de muestreo probabilístico, consideraremos, al menos de palabra, todas las muestras posibles de un cierto tamaño procedentes de la población, asignaremos una probabilidad a cada muestra de acuerdo con la técnica de muestreo elegida, y extraeremos una muestra en particular de acuerdo con esa misma técnica. El resultado será una estimación cuya fiabilidad y precisión podrán cuantificarse; en otras palabras, una estimación para la que podemos proporcionar un *intervalo de confianza*. Especificando ese intervalo afirmaremos que en un porcentaje en particular de las muestras generadas, la característica de la población (o parámetro) implicada se situará en un intervalo específico, estimado a partir de la muestra. De hecho, también podemos usar nuestro conocimiento teórico, una vez recogidos los datos, para calcular el intervalo de confianza de la estimación que nos interesa, o bien obtener una estimación del tamaño de la muestra necesario para calcular el intervalo de confianza antes de empezar la investigación. Obviamente es mejor seguir este último procedimiento que descubrir demasiado tarde que de los datos recogidos sólo podemos especificar un intervalo sin interés alguno, en el que debe encontrarse la característica de la población. Hay que insistir de nuevo en que todas esas estimaciones están basadas en el registro arqueológico tal y como se ha recuperado y que se refieren al registro que podríamos haber recuperado: hacemos inferencias sobre la «población de hallazgos físicos» a partir de una muestra de ella.

En resumen, debiera de haber quedado claro que el propósito de usar métodos de muestreo probabilístico es obtener estimaciones de unas cantidades cuya fiabilidad y precisión podemos afirmar. En un nivel regional, por ejemplo, puede tener interés estimar la densidad media de yacimientos neolíticos en un área, o bien su cantidad total; al nivel de un yacimiento, podemos necesitar la estimación de la proporción de láminas de sílex retocadas o de fragmentos de borde de un tipo determinado. Si tales medidas son de interés o no para nuestro

estudio dependerá de sus *objetivos*. Por ejemplo, si comparamos yacimientos o áreas en un análisis macroespacial, los ejemplos anteriores serán de gran interés.

En la sección siguiente se describirán los detalles técnicos para obtener esas medidas a partir del muestreo aleatorio; algunos métodos más difíciles serán presentados tan sólo en líneas generales. Finalmente se discutirán algunos de los problemas planteados, argumentándose que hay muchas cuestiones arqueológicas para las que las técnicas clásicas de muestreo probabilístico no son adecuadas, incluso en los casos (la mayoría) en que la evidencia arqueológica sea sólo parcial.

CÁLCULO DE LOS INTERVALOS DE CONFIANZA Y DEL TAMAÑO DE LAS MUESTRAS

Recordemos algunos puntos ya señalados en capítulos anteriores. En primer lugar, la distinción hecha en el capítulo 5 entre las características de una población y las de las muestras: *parámetros* y *estadígrafos*. Para una población dada, el valor de un parámetro cualquiera será fijo, pero desconocido. En el muestreo probabilístico, el objetivo es usar los estadígrafos calculados a partir de la muestra para estimar los parámetros de la población. Dado que no conocemos estos últimos (no necesitaríamos extraer una muestra si los conociésemos), nunca podremos saber con cuánta precisión nuestras estimaciones se aproximan al valor del parámetro. La base para tener una cierta confianza en la estimación ha de basarse en un método de selección de la muestra que tenga fundamentos teóricos seguros con los que justificar las afirmaciones.

En este capítulo examinaremos la estimación de tres características poblacionales: la media de la población (por ejemplo, la densidad media de yacimientos en un área); la población total (por ejemplo, la cantidad total de yacimientos en un área); la proporción de valores en una población que satisfaga cierta condición de interés (por ejemplo, la proporción de yacimientos fortificados en un área). De hecho, esas tres medidas están muy relacionadas entre sí, por lo que, en la exposición preliminar de los conceptos que es necesario entender antes de describir los métodos de estimación, se asumirá que lo que nos interesa verdaderamente es la media de la población.

Si queremos una estimación de la media de la población a partir de una única muestra aleatoria (se definirá más adelante), la media de esa muestra será totalmente satisfactoria ya que no está distorsionada. Desafortunadamente, esto no significa que la media de una única muestra aleatoria siempre corresponda a la media de la población de la que se ha extraído. Por el contrario, si tomamos una serie de muestras aleatorias de una población y obtenemos una media de cada una de esas muestras, cada una de ellas será ligeramente distinta —habrá una distribución de medias—. Es la media de la distribución de las medias muestrales la que corresponderá al parámetro poblacional.

El resultado de todo ello es que si bien la media de una muestra específica es correcta en tanto en cuanto funcione, no tenemos un fundamento para saber si es una buena estimación o no. Si tratamos con la media de una variable continua —por ejemplo, la media de la longitud en milímetros de una muestra de puntas de flecha, calculada con una exactitud de tres decimales—, la probabilidad de que corresponda a la media poblacional es casi infinitamente pequeña.

Si queremos tener un cierto grado de confianza en nuestra estimación de la media poblacional, habremos de tener en cuenta la dispersión de la distribución de medias muestrales a las que nos hemos referido antes. Ahora bien, ¿cómo hemos de hacerlo si en 9 de cada 10 casos lo único que tenemos es una muestra única?

La respuesta es que hemos de empezar con la dispersión de la muestra —su desviación típica—. Esa estimación de la dispersión de la población no nos interesa por sí misma, sino como dispersión de la distribución de las medias muestrales. Sin embargo, cuanto más variable sea la población, más variables serán las medias de una serie de muestras extraídas de ella. La teoría estadística, que no detallaremos aquí, permite pasar de la estimación basada en la muestra de la desviación típica de la población, a una estimación de la distribución teórica de las medias muestrales —denominada *error típico de la media*—. La fórmula es:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

En palabras, podemos obtener una estimación del error típico de la media, dividiendo una estimación de la desviación típica de la población entre la raíz cuadrada de la cantidad de observaciones en la muestra. Intuitivamente significa que la dispersión de una distribución de medias no será tan grande como la dispersión de los valores individuales de la población, y que a medida que la muestra se haga mayor y más representativa de la variabilidad de la población, disminuirá el error típico.

No obstante, hay otro aspecto del tamaño de la muestra que es distinto de la estimación del error típico. Si nuestra muestra es lo suficientemente grande como para incluir toda la población de interés, conoceríamos la media de la población y nuestra estimación no tendría ningún error en absoluto. Por extensión, a medida que nuestra muestra va en aumento, la estimación del error típico puede ajustarse. Esto se hace sumando la fórmula que acabamos de ver a lo que se denomina *factor de corrección de la población finita*. Tenemos ahora:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

donde n es el tamaño de la muestra y N el tamaño de la población.

Habiendo visto cómo calcular una estimación del error típico de la media,

¿qué contribución tiene esto en el cálculo de una estimación de la media de la población, con un intervalo de confianza bien especificado? La respuesta es que depende de ciertas propiedades de la distribución normal que ya hemos visto en un capítulo anterior.

Recapitulando podemos decir que la distribución que nos interesa, si hemos de obtener una estimación, es la distribución teórica de las medias muestrales, cuya desviación típica, es decir, el error típico, ya sabemos calcular. Puede mostrarse nuevamente, por medio de una teoría estadística no detallada aquí, que mientras el tamaño de la muestra sea razonablemente grande —digamos, mayor de 30—, la forma de esa distribución de medias será normal, aunque la forma de la distribución de la *población* sea asimétrica y por tanto no normal. Tal y como vimos antes (capítulo 8), es característico de la distribución normal que haya una proporción constante de las observaciones en una cierta cantidad de desviaciones típicas a ambos lados de la media. Así, si tenemos una distribución normal de medias muestrales podremos decir, por ejemplo, que el 68,2 % de las medias se encontrarán en 1 error típico a ambos lados de la media global; o que el 95 % de las medias se encontrarán a 1,96 desviaciones típicas a ambos lados de la media.

Como esta media global es la media poblacional que nos interesa, podemos decir, por ejemplo, que la media del 95 % de las muestras aleatorias elegidas en esta población se encontrarán a 1,96 errores típicos de la media de la población. Sin embargo, como no conocemos la media de la población, hemos de calcularlo de otra forma: para el 95 % de las muestras extraídas de la población, un intervalo fijado en 1,96 errores típicos a ambos lados de la media muestral incluirá la auténtica media de la población; para el 5 % de las muestras restantes, no. Naturalmente podemos cambiar los porcentajes cambiando el número de errores típicos. Todo ello puede ilustrarse mediante un diagrama (figura 14.1). Este procedimiento nos permite producir un intervalo de confianza: un rango en el que se encontrará nuestro parámetro, con un nivel de probabilidad específico y expresado en términos de la cantidad de errores típicos asociados con el nivel de probabilidad que nos interesa.

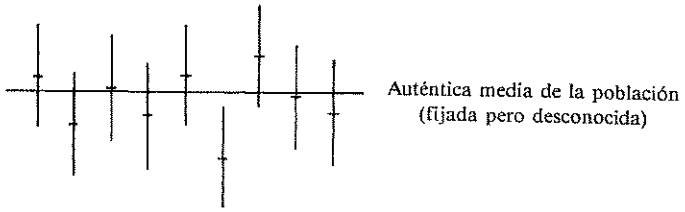


FIGURA 14.1. Relaciones entre la media de las poblaciones y una serie de medias muestrales; cada barra representa una media muestral y el alcance de un número fijo de errores típicos.

Ya es hora de ilustrar esta exposición teórica con un ejemplo sencillo. Supongamos que hemos elegido una muestra aleatoria (ya veremos después cómo hacerlo) de 50 puntas de flecha, de una colección de 2.000, con el propósito de obtener una estimación de la media de la longitud para la colección como un todo, y que queremos que la estimación tenga un 95 % de probabilidades de ser correcta. Supondremos que las medidas han producido una media de 22,6 mm y una desviación típica de 4,2 mm. El primer paso es obtener el error típico de la media. Usando la fórmula:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

tenemos

$$s_{\bar{x}} = \frac{4{,}2}{\sqrt{50}} \sqrt{1 - \frac{50}{2.000}} =$$

$$= (0{,}594)\ (0{,}987) = 0{,}586$$

La media muestral es de 22,6 mm. Para obtener un intervalo que tenga un 95 % de probabilidades de incluir la media de la población y asumir que la distribución teórica de las medias muestrales es normal, sabemos que hemos de definir el intervalo de 1,96 errores típicos a ambos lados de la media muestral. Así, nuestro intervalo estará definido por una media muestral ± 1,96 errores típicos. Aquí:

$$22{,}6 \pm (1{,}96)\ (0{,}586) = 22{,}6 \pm 1{,}15$$

y podremos decir que hay un 95 % de probabilidades de que la longitud media de la colección de puntas de flecha esté en el intervalo 21,45-23,75 mm. Podemos haber tenido mala suerta, claro está, y que la muestra aleatoria elegida sea una de las del 5 % que no incluye la media poblacional correcta. Si nos preocupa esa posibilidad, podemos ampliar la probabilidad de acierto, aunque suponga ampliar el intervalo. Así, si queremos una probabilidad del 99 %, tendremos que ir hasta 2,58 errores típicos a ambos lados de la media muestral. En este caso:

$$22 \pm (2{,}58)\ (0{,}586) = 22{,}6 \pm 1{,}51$$

con lo que habrá una probabilidad del 99 % de que la media de la longitud en la población de puntas de flecha esté entre 21,09 y 24,11 mm.

Para construir, en general, un intervalo de confianza, podemos escribir:

$$\bar{x} \pm Z_\alpha \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

donde Z_α es la puntuación Z, o cantidad de desviaciones típicas asociadas a un nivel de probabilidad en particular.

Hemos de hacer aquí una breve aclaración. Usar la puntuación Z, como hemos hecho en el ejemplo anterior, es perfectamente satisfactorio si el tamaño de la muestra es mayor de 40; si es menor, habremos de tener en cuenta el hecho de que s, la estimación de la desviación típica de la población, está basada en una muestra y que muestras distintas producirán valores de s distintos. Las muestras pequeñas suelen ser más variables, en general, de una a otra, por lo que tendremos en cuenta este hecho usando en las estimaciones la distribución t de Student y no la distribución normal (véase, por ejemplo, Blalock, 1972, pp. 188-193), que varía en la proporción de observaciones en un número dado de desviaciones típicas de la media, de acuerdo con el tamaño de la muestra, convergiendo en la distribución normal a medida que la muestra se hace mayor. Por lo tanto, en lugar de la fórmula anterior, tendremos:

$$\bar{x} \pm t_{\alpha,\ \text{g.l.}} \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

donde $t_{\alpha,\ \text{g.l.}}$ es el valor t para el número específico de grados de libertad ($= n - 1$) asociados a un nivel de probabilidad en particular (véase anexo 1, tabla *C*; nótese que para los límites del intervalo de confianza hay que usar t para la prueba de dos colas, es decir, la fila 2_α en la parte superior de la tabla).

Veamos ahora la principal amenaza a nuestro argumento. En el ejemplo de las puntas de flecha suponemos que ya tenemos una muestra de un cierto tamaño, sobre cuya base fuimos capaces de construir un intervalo de confianza. Eso está muy bien, pero en la situación en la que nos encontramos al principio de una investigación arqueológica para la que disponemos de recursos limitados, una de las cuestiones clave es, precisamente, qué tamaño debe tener la muestra elegida. Si es demasiado pequeña, los intervalos de confianza de nuestras estimaciones de las cantidades que nos interesan serán demasiado amplios para usarlos; si la muestra es mayor de lo que necesitamos, se estarán desperdiciando unos recursos que podrían usarse en otra cosa. ¿Cómo hemos de calcular, pues, el tamaño de la muestra?

Volvamos a la fórmula para construir un intervalo de confianza, asumiendo que tratamos con una distribución normal e ignorando, de momento, la corrección de la población finita. La fórmula para el intervalo alrededor de la media es:

$$\pm Z_\alpha \frac{s}{\sqrt{n}}$$

Designaremos el factor de tolerancia (o $\pm$) como d. Tenemos, pues, que:

$$d = Z_\alpha \frac{s}{\sqrt{n}}$$

que puede transformarse en la siguiente expresión:

$$\sqrt{n} = \frac{Z_\alpha s}{d}$$

o en esta otra:

$$n = \left(\frac{Z_\alpha s}{d}\right)^2$$

Tenemos aquí una fórmula para estimar el tamaño de la muestra en un caso específico, si es que podemos especificar las tres cantidades del lado derecho de la ecuación.

Fijar un valor para Z_α es bastante simple; es cuestión de decidirlo según la probabilidad que queramos que tenga el intervalo que incluirá el parámetro de interés. Especificar la tolerancia, el factor $\pm$, que estamos preparados a aceptar, es también bastante sencillo en principio, aunque no tanto en la práctica. ¿Por qué hemos de decidir un nivel de tolerancia específico? Idealmente ha de originarse en la cuestión específica que investigamos, si bien en la práctica la decisión es bastante arbitraria. Lo fundamental es que si estamos preparados para afirmar una estimación con un factor de error y no insistimos en obtener el valor exacto del parámetro de la población, podemos ahorrar bastante esfuerzo en términos de la cantidad de unidades o elementos que hay que examinar.

La tercera cantidad, s, se ha escrito en minúscula porque, si bien es la desviación típica de la población, S, la que nos interesa, hemos usado su estimación muestral, s, ya que aquélla nos es desconocida. Si quisiéramos calcular el tamaño de la muestra, sin embargo, incluso obtener s sería problemático. En el caso de construir un intervalo de confianza, podemos obtenerlo a partir de la muestra. Pero ¡antes de extraerla no podemos decidir su tamaño! ¿Cómo hacerlo, entonces?

En términos generales hay dos respuestas posibles, ninguna de las cuales es la ideal. Una consiste en efectuar algún estudio piloto sobre la población antes de hacer la investigación definitiva, obteniendo así una estimación preliminar basada en la muestra de la desviación típica poblacional con la que trabajar. Otra consiste en usar los resultados de trabajos anteriores en poblaciones similares que ya han producido estimaciones de su variabilidad. En ambos casos, podemos querer aumentar ligeramente las estimaciones resultantes de la

desviación típica de la población, en el supuesto de que hayan sido subestimadas y que nuestros intervalos de confianza demuestren ser mayores de lo que pretendíamos.

Tras usar la fórmula anterior para obtener una estimación inicial del tamaño de la muestra necesario, hemos de considerar si representa o no una fracción lo suficientemente grande de la población total para la corrección de la población finita que sea necesaria. Para obtener el tamaño necesario teniendo en cuenta la fracción muestral tenemos:

$$n' = \frac{n}{1 + n/N}$$

donde N es el tamaño de la población y n el tamaño de la muestra calculado al principio.

Veamos ahora un ejemplo concreto para entender cómo se calcula el tamaño correcto de una muestra para estimar una media (véase Van der Veen y Fieller, 1982). Volvamos a nuestras 2.000 puntas de flecha hipotéticas y preguntémonos cuántas hay que medir para estimar la media de su longitud, basándonos en el criterio del muestreo aleatorio simple. El conocimiento previo nos dice que tal distribución de longitudes es probable que sea positivamente asimétrica, es decir, con una larga cola positiva, pero que esta asimetría nunca será tan grande como para afectar la normalidad de la distribución de las medias muestrales, que es lo que importa. Aplicando la fórmula

$$n = \left(\frac{Z_\alpha s}{d}\right)^2$$

sólo hemos de sustituir las incógnitas:

Z_α: nos interesa un nivel de probabilidad del 95 %, por lo que su valor será 1,96.

s: asumiremos que hemos elegido un pequeño número de puntas de flecha como muestra piloto, y que nos ha dado el valor 4,00 mm como estimación de la desviación típica de la población. Para asegurarnos, hemos decidido elevar esa estimación en 1,0 mm y operar con una estimación de 5,0 mm.

d: asumiremos que nos interesa estimar la media con una tolerancia de ± 1,0 mm, lo cual es una decisión arbitraria.

$$n = \left(\frac{1{,}96 \times 5{,}0}{1{,}0}\right)^2 = 96$$

Aplicando el factor de corrección de la población finita,

$$n' = \frac{n}{1 + n/N}$$

$$n' = \frac{96}{1 + 96/2.000} = 91{,}6$$

En este caso, como n es una fracción de N, aplicar esta corrección no provoca muchas diferencias.

Para mostrar las diferencias que puede causar una variación en la tolerancia admitida, recalculemos la cifra del tamaño de la muestra; suponiendo que tuviéramos suficiente con una tolerancia de ± 2,0 mm, en lugar de 1,0 mm:

$$n = \left(\frac{1{,}96 \times 5{,}0}{2{,}0}\right)^2 = 24$$

Alternativamente, si lo estableciésemos en ± 0,5 mm:

$$n = \left(\frac{1{,}96 \times 5{,}0}{0{,}5}\right)^2 = 384$$

Obviamente es importante reflexionar antes sobre el nivel de tolerancia, pues afecta enormemente el tamaño requerido de la muestra.

Puede suceder que el tamaño de la muestra se determine simplemente por medio de la cantidad de tiempo y dinero de que dispone el arqueólogo. Basándonos en el conocimiento del coste de recoger la información sobre un aspecto específico, o en experimentos o conocimientos previos acerca del tiempo necesario para recogerla, sabremos de antemano el máximo que se necesita. Si, dada una estimación de la desviación típica de la población, se viera que una muestra de ese tamaño produce tan sólo un intervalo de confianza demasiado amplio para ser de interés, habrá que replantearse todo el proyecto.

Estimación de los totales

Una vez visto el procedimiento para estimar el intervalo de confianza para la media de una población con un nivel de probabilidad dado, y cómo redisponer la información para estimar los tamaños requeridos de la muestra, podemos examinar la cuestión de estimar los totales de la población; lo haremos con más brevedad que antes, pues el procedimiento es semejante al que se acaba de exponer.

Puede desconcertar al lector la idea de estimar un total. Después de todo, ¿no hemos de tener una lista total de la población para poder elegir una muestra? La respuesta es que la lista de elementos a partir de los cuales definimos la muestra no tiene por qué ser la misma lista de elementos que nos interesan. Veamos un ejemplo arqueológico, referido a una prospección regional. El procedimiento habitual sería dividir el área por medio de una cuadrícula y estudiar una muestra de las divisiones de esa cuadrícula: cada una de esas divisiones constituye la lista de elementos a partir de la cual establecemos la muestra; una lista que podemos enumerar por completo. Pero lo que nos interesa no es el número total de cuadrados, sino el número total de yacimientos, quizás de un tipo peculiar o de un período. Para encontrar esta nueva cifra habremos de calcular una estimación muestral de la media de los yacimientos en cada división, y la multiplicaremos por la cantidad total de divisiones. Matemáticamente se expresa del siguiente modo:

$$x_T = Nx$$

donde x_T es el número total de elementos que nos interesan, x la media de elementos de interés en cada unidad muestral, y N el número total de unidades muestrales en la población.

Esto es bastante sencillo, pero, al igual que al estimar la media, no solemos estar interesados en una estimación puntual, sino en un intervalo en el que se encuentre el total que buscamos con un cierto grado de probabilidad. Afortunadamente, esto también es muy fácil. Como podíamos esperar, se trata tan sólo de incluir N en las fórmulas para obtener los intervalos de confianza y los tamaños muestrales que hemos visto en el caso de la media. Así, en el caso del intervalo de confianza, tenemos:

$$\bar{x}_T \pm Z_\alpha N \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

con $t_{\alpha, g.l.}$ sustituida por Z_α en el caso de muestras pequeñas. Si calculamos el tamaño de la muestra, tenemos

$$n = \left(\frac{Z_\alpha s N}{d}\right)^2$$

y, como antes,

$$n' = \frac{n}{1 + n/N}$$

Estimación de la proporción de una población

Lo que ahora vamos a exponer difiere un tanto de lo visto hasta aquí, si bien se trata de una situación que se plantea frecuentemente en arqueología. Podemos necesitar, por ejemplo, estimar la proporción de fragmentos en un conjunto con ciertas características sobre la base de un muestreo aleatorio simple en ese conjunto; quizás la proporción de útiles líticos caracterizados por un cierto tipo de retoque. La diferencia entre este y los ejemplos previos no radica en lo que intentamos hacer, sino en la mera estructura de los valores en la población y en la muestra, que sólo pueden adoptar uno o dos estados: un fragmento en particular, por ejemplo, o tiene esa característica, o no la tiene.

La mejor estimación de la proporción de una población será la correspondiente proporción muestral, si bien de nuevo necesitamos un intervalo de confianza, y para esto hemos de conocer el error típico de la proporción.

La desviación típica de una población de unos y ceros es $[P\ (1 - P)]^{1/2}$, donde P es la proporción de interés. Puede estimarse usando p, la proporción muestral. Para obtener el error típico de la proporción hemos de dividirlo entre $\sqrt{n}$, la raíz cuadrada del tamaño de la muestra, al igual que hicimos con la media. Así, el error típico de la proporción será $[p(1 - p)/n]^{1/2}$.

Obviamente, la forma de esta distribución no estará normalizada. Sin embargo, la distribución teórica de las medias muestrales lo estará, siempre y cuando la muestra sea lo suficientemente grande, digamos, más de 50. En estas circunstancias podemos construir los intervalos de confianza multiplicando el error típico de la proporción por Z_α, el número de desviaciones típicas que corresponden al nivel de probabilidad que nos interesa, ya que la distribución es normal. Finalmente, el factor de corrección de la población finita desempeña el mismo papel que antes. Así, un intervalo de confianza para P, la proporción de la población, estará dado por :

$$p \pm Z_\alpha \sqrt{\left(\frac{p(1 - p)}{n}\right)\left(1 - \frac{n}{N}\right)}$$

donde p es la proporción muestral.

Estos cálculos son muy sencillos cuando tenemos información de una muestra y pretendemos construir un intervalo de confianza para P, basándonos en ella; pero ¿cómo estimar el tamaño de la muestra? Ya vimos antes, al examinar los intervalos de confianza para la media, que si designamos el intervalo a ambos lados de la media que estamos dispuestos a aceptar como d, tenemos (olvidando, de momento, el factor de corrección de la población finita):

$$d = Z_\alpha \frac{s}{\sqrt{n}}$$

Lo podemos transformar en:

$$n = \left(\frac{Z_\alpha s}{d}\right)^2$$

Si lo hacemos ahora en el caso de las proporciones, tendremos

$$d = Z_\alpha \sqrt{\frac{p(1-p)}{n}}$$

y transformándolo:

$$n = \frac{Z_\alpha^2 [p(1-p)]}{d^2}$$

que nos dice que con el objeto de calcular el tamaño de la muestra apropiado para estimar la proporción con un cierto grado de probabilidad necesitamos, en primer lugar, una estimación de esa proporción. Puede parecer surrealista, aunque no es tan malo como se pueda creer.

Para una tolerancia y un nivel de probabilidad dados, el tamaño máximo de la muestra, n, que precisaremos será aquel para el cual $p(1-p)$ alcance su valor máximo. Esto sucederá cuando p = $(1-p) = 1/2$. El producto de los dos, en este caso, será 1/4 y ningún otro par de valores producirá un producto más elevado. En otras palabras, en el caso de proporciones, y a diferencia de medias y totales, siempre podremos encontrar el tamaño máximo de la muestra que necesitemos alcanzar, con una tolerancia y un nivel de probabilidad en particular, tan sólo asumiendo que las proporciones de la población P y $(1-P)$ son 1/2 —no necesitamos estimarlas con ayuda de las estimaciones muestrales.

Para valores actuales de P entre 0,3 y 0,7, este valor máximo no estará muy lejos de la estimación del tamaño de una muestra necesario para un intervalo de confianza dado. Por otro lado, a medida que P se hace mayor o menor que eso, el n necesario disminuirá considerablemente; por lo tanto, asumir $P = 1/2$ implicará mucho trabajo innecesario. En muchos casos, sin embargo, será posible averiguar si la proporción que investigamos es muy escasa o, por el contrario, bastante común, o bien tan sólo razonablemente común.

Al igual que con las otras estimaciones del tamaño de la muestra, podemos corregir para la fracción muestral:

$$n' = \frac{n}{1 + n/N}$$

Resta por considerar un ejemplo de la estimación del tamaño requerido para una muestra. Un investigador que trabajaba en las islas Shetland (Winham, 1978) quería saber cuántos yacimientos conocidos había que visitar para obtener información acerca de sus características locacionales; por ejemplo, la proporción de tipos de suelo en particular. Decidió que necesitaba que las estimaciones de la proporción de yacimientos con unas características en particular tuvieran una tolerancia de ± 7 %, y un 95 % de probabilidades. No tenía información previa y además quería estimar las proporciones para una cierta cantidad de características distintas al mismo tiempo, y había motivo para suponer que éstas tendrían unas frecuencias de aparición distintas entre sí. Adoptó la hipótesis más conservadora, asumiendo $P = 1/2$, lo que produce una muestra del mayor tamaño posible. La población total era de 198 yacimientos. Para encontrar el tamaño necesario de la muestra

$$n = \frac{Z_\alpha^2 [P(1-P)]}{d^2}$$

En este caso

$$n = \frac{Z_\alpha^2 (1/4)}{d^2} = \frac{Z_\alpha^2}{4d^2}$$

sustituyendo por las cifras conocidas:

$$n = \frac{1{,}96^2}{4(0{,}07)^2} = 196$$

que es, aproximadamente, la población total, por lo que habrá que emplear el factor de corrección de la población finita:

$$n' = \frac{n}{1 + n/N} = \frac{196}{1 + 196/198} = \frac{196}{1.989} = 98$$

Así, el tamaño de la muestra necesaria con el grado requerido de precisión y nivel de probabilidad es de 98 yacimientos, seleccionando estos por muestreo aleatorio simple.

SELECCIÓN DE UNA MUESTRA

Ya hemos tratado lo suficiente acerca de la construcción de los intervalos de confianza y la estimación del tamaño necesario de la muestra basándonos

en muestras aleatorias simples; sin embargo, aún no se ha especificado lo que es una muestra aleatoria simple, o bien cómo elegir una.

Para elegir una muestra cualquiera es evidente que necesitamos de un conjunto de *unidades de muestreo*: entidades discretas, definibles, entre las cuales pueden elegirse las muestras. La lista de unidades de muestreo se denomina *marco de muestreo*. Sin un marco de muestreo que contenga la lista de todos los elementos en la población de entre los cuales queremos extraer una muestra, no podemos seguir adelante.

Es importante insistir de nuevo en que las unidades de muestreo que constituyen esa lista no tienen por qué ser los elementos de interés: se limitan a contenerlos. Así, en un contexto arqueológico, el hecho de que antes de excavar un yacimiento no solemos saber qué características tendrá, y que si lo supiéramos no tendríamos que extraer una muestra, es irrelevante en el proceso de extraer la muestra. En todos esos casos podemos extraer muestras de poblaciones conocidas que contengan la que nos interesa. A nivel regional podemos extraer muestras de unidades de superficie, y a nivel de yacimiento, catas o sondeos. Es posible, entonces, argumentar los atributos de esas unidades arbitrarias y estudiar, por ejemplo, la cantidad de yacimientos por km^2, o la cantidad de fragmentos en una cata. O bien podemos estudiar algunos aspectos de los yacimientos o artefactos mismos, en cuyo caso cualquier procedimiento de inferencia estadística ha de tener en cuenta que se opera con grupos de muestras (véase más adelante).

Una *muestra aleatoria simple* es una muestra con la característica de que cualquier individuo, y cualquier combinación de individuos, tienen una oportunidad idéntica de aparecer en la muestra. Ésta ha de obtenerse extrayendo de uno en uno los miembros de la población que se incluirán en la muestra, *sin cambiarlos de sitio*; es decir, una vez que se ha elegido un elemento no tiene una segunda oportunidad de resultar seleccionado.

El mecanismo de selección suele ser una *tabla de números aleatorios* (véase anexo 1, tabla D). Las unidades muestrales en la población se enumeran consecutivamente. Se eligen tantos números aleatorios de la tabla como sea necesario para conseguir una muestra del tamaño decidido previamente, ateniéndose a la limitación de que si se llega a un número que ya ha aparecido anteriormente, la nueva aparición es ignorada y se pasa al siguiente número. ¿Cómo ha de leerse dicha tabla? Supongamos el caso que vimos antes en el que teníamos que elegir 50 puntas de flecha de una población de 2.000; a cada punta se le ha dado un número de 1 a 2.000. Como 2.000 es un número de cuatro cifras, elegiremos números aleatorios de cuatro cifras, que se obtienen leyendo juntos cuatro números consecutivos en la tabla; veamos un pequeño extracto de la misma:

10	09	73	25
37	54	20	48
08	42	26	84
99	01	90	25

No tenemos por qué empezar por la parte superior de la tabla, sino que podemos construir nuestros bloques de cuatro cifras uniendo cualquier serie de cuatro números adyacentes.

Por ejemplo, podemos empezar por la segunda fila, a partir del cuarto número. Obtendremos 4.204, 2.268, 1.902. Como sólo uno de esos números está dentro del rango 1-2.000, sólo se elegirá ese, ignorando los demás. Se sigue leyendo en la tabla, buscando tantos números como sean necesarios, evitando la duplicación.

Si la población sólo contuviese 100 elementos (0-99, o 1-100, con 00 como 100), sólo se necesitarían números de dos cifras.

Si las unidades muestrales son espaciales y están definidas en términos de una cuadrícula, se seleccionarán las unidades individuales por medio de dos números aleatorios que especifiquen las coordenadas de las esquinas de la unidad.

Si el tamaño necesario de la muestra es muy grande, la selección manual de una serie de números es muy laboriosa, por lo que será mejor usar un generador de números aleatorios, disponible en la mayoría de los programas informáticos de estadística. Sin embargo hay que tener cuidado al usarlos, pues muchos de ellos producen siempre el mismo conjunto fijo de números, a no ser que se aleatorice explícitamente el punto de partida.

ALTERNATIVAS AL MUESTREO ALEATORIO SIMPLE

El muestreo aleatorio simple es un procedimiento muy usado y es fácil de emplear, estadísticamente hablando. Sin embargo, muy a menudo habrá circunstancias en las que queramos utilizar algún método más complejo, porque: *a*) son más eficaces que el muestreo aleatorio simple, en el sentido de que se obtiene una estimación más precisa del mismo número de unidades de muestreo; *b*) a menudo son mucho más fáciles de calcular en la práctica que el muestreo aleatorio simple; *c*) nuestros objetivos pueden requerir un método alternativo; *d*) a veces, por razones arqueológicas, no tenemos elección (por ejemplo, hemos de usar muestras agrupadas [véase más adelante]).

Muestreo aleatorio estratificado

En una muestra aleatoria estratificada, la población que se muestrea se divide en unos *estratos* (nada que ver con los arqueológicos) o subdivisiones, extrayéndose una muestra aleatoria independiente de cada uno de ellos. Es el investigador el que establece las subdivisiones, y a menudo las elige porque hay alguna diferencia entre ellas. Por ejemplo, si se realiza una prospección arqueológica en una región, puede que se decida dividir la región en zonas ambienta-

les, basándose en algún criterio, extrayendo muestras independientes en cada zona; o bien, si se excava un yacimiento y de los trabajos preliminares se deduce que está diferenciado funcionalmente, se extraerá una muestra de cada sección funcionalmente distinta. Las subdivisiones serán, sin embargo, más o menos arbitrarias.

Las razones para la estratificación son usualmente las ventajas que tiene este procedimiento sobre el muestreo aleatorio simple. Ello se debe a dos razones. En primer lugar, una estratificación apropiada puede asegurar que se extraigan muestras de todas las partes de nuestra población. El muestreo aleatorio simple no lo garantiza, porque ciertas partes de la lista de las unidades pueden tener más probabilidades que otras de formar parte de la muestra sólo debido al azar. No afecta a la obtención de una estimación para toda la población, pero lo más normal es que nos interesen las comparaciones internas dentro de la muestra, como en los ejemplos anteriores; o bien, por ejemplo, si elegimos una muestra de fragmentos de un yacimiento con varias fases de ocupación para análisis arqueométricos, desearemos, probablemente, asegurar que todos los períodos (o bien sólo algunos de ellos) estén representados. En el muestreo regional, podemos querer asegurarnos de que se examinan todas las partes de la región, sin considerar la variación ambiental o cualquier otra.

La segunda razón para estratificar la población es que si su característica de interés está distribuida con más homogeneidad dentro de los estratos que no entre ellos (es decir, hay variación entre los estratos, pero no dentro de ellos), la precisión de la estimación global obtenida será mayor para un cierto número de unidades muestrales que por medio del muestreo aleatorio simple.

La fórmula para el error típico de la media, basado en una muestra estratificada, es la siguiente (Dixon y Leach, 1978):

$$s_{\bar{x}_{\text{estrat}}} = \sqrt{\frac{\sum_{i=1}^{k} (n_i s_i)^2 (1 - n/N)}{n^2}}$$

donde n_i es el número de unidades en la muestra del estrato i, s_i es la desviación típica en el estrato i, n es el número total de unidades en la muestra, y k es el número total de estratos. Si se extrajesen muestras de distintas fracciones de cada estrato, sin embargo, habría que cambiar la fórmula para tenerlo en cuenta (véase, por ejemplo, Dixon y Leach, 1978, pp. 19-21).

Muestreo sistemático

Con esta técnica (que es un caso especial del muestreo por grupos) el intervalo entre los puntos muestrales está fijado por la relación entre el tamaño de la muestra propuesta y el tamaño de la población. Así, si quisiéramos elegir una muestra de 30 de una población de 300, tendríamos que elegir cada 300/30-avo elemento —uno de cada diez, en otras palabras—. Si es el primero, el undécimo, el vigesimoprimero, etc., los que resultan elegidos, o bien el quinto, el decimoquinto, el vigesimoquinto, etc., o cualquier otro, se determina eligiendo un número aleatorio entre 1 y 10 como punto de partida.

Las razones para elegir una muestra sistemática suelen ser de conveniencia práctica; a menudo es más sencillo de esta manera. El otro motivo, especialmente si nos referimos al muestreo en una prospección arqueológica o en una excavación, es que normalmente necesitaremos que la muestra sirva a distintos propósitos, de los que la estimación de las características de la población sólo será uno. En estas circunstancias, una muestra sistemática constituirá un compromiso entre exigencias conflictivas.

En un sentido estadístico y de estimación estricto, las muestras sistemáticas presentan indudables problemas, en particular para el cálculo de una estimación del error típico del estadígrafo que interesa. Esto es así, a causa de la falta de independencia entre los elementos de la muestra y a causa de la posibilidad de que se produzcan periodicidades en los valores de los elementos de la población que van a integrarse en la muestra, sobre todo en el caso del muestreo de distribuciones espaciales. En la excavación de un asentamiento, por ejemplo, una cuadrícula regular de unidades de muestreo en un intervalo dado puede incluir todas las casas o ninguna de ellas, si es que éstas se distribuyen sistemáticamente. Cualquiera que sea el caso, una estimación del número de casas en el yacimiento basado en el número de casas en la muestra será errónea (véase Winter, 1976).

Diversas reacciones son posibles ante las dificultades presentadas por los muestreos sistemáticos en el cálculo de los errores típicos. Una es prescindir totalmente de ese tipo de muestreo (véase Doran y Hodson, 1975), que es probablemente la mejor solución, si bien podemos excluirla en una situación particular por motivos prácticos y no estadísticos. Otra es emplear una de las fórmulas propuestas por Cochran (1977, pp. 224-227) para obtener el error típico, cada una de las cuales sólo es aplicable en ciertas situaciones. El problema aquí es que su uso implica el conocimiento previo de la estructura de la población de la que queremos extraer una muestra, para poder usar un método apropiado; tal conocimiento puede ser inexistente. En el caso particular en que el orden de los elementos en el marco del muestreo haya sido aleatorizado, o que se asuma que es aleatorio, la muestra sistemática podrá tratarse como un muestreo aleatorio simple (Cochran, 1977, pp. 214-216) y se usará la misma fórmula para el error típico. Sin embargo, esta opción ha de usarse con precaución, pues las periodicidades y las tendencias en la población no siempre son evidentes.

Otro enfoque es el adoptado por Bellhouse en un contexto arqueológico (Bellhouse, 1980; Bellhouse y Finlayson, 1979). Se trata de una extensión de la idea de emplear el conocimiento previo e implica el uso de un programa in-

formático que calcule el resultado de los distintos esquemas de muestreo, incluyendo el muestreo sistemático, en términos del error típico de la estimación para las estimaciones de las medias, totales y proporciones. El problema es que para proporcionar un esquema del muestreo apropiado para un caso en particular, ya sea una región o un yacimiento, el método presupone que se dispone de una información completa para una región similar o yacimiento, con el fin de proporcionar la información sobre la que se fundamentará la realización de los esquemas de muestreo.

La cuestión del muestreo espacial o del área que resulta relevante en este tema (y en el muestreo sistemático en general) se considerará más adelante.

Muestreo por grupos

Lo más sencillo será describir su funcionamiento comparándolo a los demás métodos que hemos visto. Vimos que en un muestreo estratificado la población se subdivide en grupos llamados estratos y que de cada estrato se extrae una muestra. En el muestreo por grupos, la población también se subdivide, si bien algunos de los grupos resultantes se eligen para examinarlos más detenidamente, y se prescinde de otros. Como sería de esperar, los principios en los que se basan las subdivisiones son distintos en cada caso.

Veamos un ejemplo arqueológico de las diferencias entre los dos enfoques, distinto también del muestreo aleatorio simple, y que contribuye a apreciar claramente dichas diferencias. Supongamos que en un yacimiento se ha excavado un gran número de fosas. Nos interesa estudiar la cerámica procedente de cada una de ellas, pero, al faltar los recursos, nos limitaremos a estudiar una muestra.

Si extrajésemos una muestra aleatoria simple de la cerámica, consideraríamos que toda la cerámica de las fosas constituye la población total, y elegiríamos aleatoriamente las muestras sin importarnos de qué fosas procede la cerámica. Es posible que tal procedimiento sea muy difícil de llevar a la práctica, porque el material está almacenado y organizado después de la excavación y porque algunas fosas estarán bien representadas en la muestra, mientras que de otras sólo se habrán elegido unos pocos fragmentos.

Si hubiéramos empleado un muestreo aleatorio estratificado, hubiésemos considerado las fosas individuales como estratos y elegido una muestra aleatoria de cerámica de cada fosa. Aunque este procedimiento sea muy adecuado en cuanto a la estimación de diversas propiedades de la población, será muy complicado de poner en práctica, pues habrá que organizar el procedimiento de elegir una muestra aleatoria de los hallazgos que contenía cada fosa.

El muestreo por grupos consistiría en extraer una muestra aleatoria de las fosas y estudiar la cerámica de las fosas elegidas. Esto tendría la desventaja de que no se examinarían todas las fosas y que las muestras de los grupos siempre darían errores de muestreo mayores que los de un muestreo aleatorio simple

en una muestra del mismo tamaño. Por otro lado, sería el procedimiento más sencillo de realizar, lo que permitiría estudiar más cantidad de cerámica.

En el muestreo por grupos no extraemos las unidades de interés directamente, sino entre grupos de unidades muestrales. El ejemplo que se acaba de exponer es muy característico del tipo de situación que aparece en arqueología; a menudo no hay muchas posibilidades de elección entre las muestras de grupos.

El principal problema con los grupos —por ejemplo, el contenido de las fosas en el caso anterior— es que tienden a ser relativamente homogéneos y que cualquiera de ellos suele representar sólo una fracción de la totalidad de la variación en la población. Esto significa que eligiendo sólo unos cuantos grupos podemos fracasar en definir la variación de esa población, que no estará representada en la muestra; alternativamente, la muestra puede incluir algunos grupos inusuales. En ninguno de esos casos obtendremos una buena estimación del error típico de la característica que nos interesaba. Idealmente, lo que pretendemos de los grupos es que sean heterogéneos, incorporando la mayor cantidad posible del rango de la variación, de forma que omitir alguno de los grupos no provoque muchas diferencias. En el muestreo estratificado, por otro lado, cada estrato está representado por una muestra de unidades; de ahí que nuestros errores de muestreo procedan de la variabilidad *dentro de* los estratos, con lo que los estratos habrán de ser lo más homogéneos posible.

En resumen, los estratos son preferibles en su globalidad, pues su uso puede reducir el error típico de las estimaciones si lo comparamos con el muestreo aleatorio simple. Las muestras de grupos no son posibles muchas veces por razones prácticas; y además suelen proporcionar errores típicos mayores que los obtenidos por medio de un muestreo aleatorio simple. En otras situaciones, sin embargo, los grupos serán más eficaces que los estratos y el muestreo aleatorio simple. Dixon y Leach (1978, p. 22) sugieren que antes de usar la fórmula para el muestreo aleatorio simple y calcular el error típico de las estimaciones en muestras de grupos, hay que aplicar un factor de corrección. Proponen que una forma rápida de hacerlo es empleando un factor de 1,5 como criterio razonable, basado en la experiencia. Así, para un intervalo de confianza, en vez de

$$\bar{x} \pm Z_\alpha \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

tendríamos

$$\bar{x} \pm Z_\alpha \frac{1{,}5s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

y para calcular el tamaño de la muestra:

$$n = \left(\frac{Z_{\alpha}1{,}5s}{d}\right)^2$$

MUESTREO PROBABILÍSTICO DE POBLACIONES CON UNA DIMENSIÓN ESPACIAL

Una vez expuestas brevemente las principales técnicas de muestreo probabilístico y los conceptos sobre los que se basan, es necesario profundizar algo más en su uso en arqueología. Los problemas que se plantean aparecen sobre todo en el contexto de la prospección y excavación de regiones y yacimientos, y pueden dividirse en dos tipos: dificultades técnicas en el muestreo probabilístico y cuestiones más amplias acerca de la relevancia del muestreo probabilístico, según los objetivos del proyecto de prospección y excavación. En esta sección se abordará el primero de esos dos temas.

Como ya hemos señalado antes, el muestreo regional y el muestreo en un yacimiento exigen unidades de muestreo para distribuciones desconocidas. Los problemas proceden de la distribución *espacial* de las observaciones dentro de y entre las unidades de muestreo.

En primer lugar, la distribución de la cantidad de elementos por unidad espacial suele ser muy asimétrica —un modelo típico para esas distribuciones es la distribución de Poisson (véase, por ejemplo, Hodder y Orton, 1976, pp. 33-38). Aunque Ihm (1978, pp. 293-294), en un contexto arqueológico, ha propuesto una fórmula para obtener el intervalo de confianza para la media de una distribución de Poisson, el autor del presente libro nunca la ha visto aplicada en un caso práctico. Los arqueólogos parecen haber usado siempre el procedimiento de estimación descrito en este capítulo y que presupone que la distribución media de las muestras es normal. Vimos antes que, para mantener ese presupuesto, el tamaño de la muestra ha de ser lo suficientemente grande; Thomas (1975) mostró que «suficientemente grande» significa muy grande, aunque la asimetría sea tan grande como es frecuente que lo sea en distribuciones espaciales.

Más difícil es el problema de los elementos que pocas veces se distribuyen aleatoriamente en el espacio, sino que se desvían de ese esquema, usualmente por culpa de su agrupación o agregación. Debido a este problema de la agregación, la mayoría de las investigaciones arqueológicas que usan técnicas de muestreo espacial, ya sea a escala de yacimiento o a escala regional, han sido, esencialmente, investigaciones empíricas *ad hoc* (véase, por ejemplo, Plog, 1976), con todo lo que eso supone en lo que se refiere a límites de su relevancia. No se han fundamentado en la teoría debido a que las consecuencias de la agregación no han sido estudiadas teóricamente.

Algunos de los trabajos más interesantes sobre este y otros problemas asociados con el muestreo espacial son los de Nance (1981; 1983), que obtiene unos resultados perfectamente ajustados a la teoría estadística, pero operando a un nivel de sofisticación estadística mucho mayor que el usado en las primeras aplicaciones y mucho más relevante para los problemas arqueológicos. En particular ha usado la distribución binomial negativa (véase, por ejemplo, Hodder y Orton, 1976, pp. 86-97) para describir distribuciones espaciales, mostrando cómo una agregación en incremento aumenta enormemente el error típico de las estimaciones.

Nance ha examinado también un problema algo distinto, pero igualmente importante, al que se ha dedicado, relativamente, poca atención comparada con la que se ha prestado al tema de la precisión de las estimaciones: la cuestión del descubrimiento de las probabilidades, es decir, la cantidad de unidades de muestreo necesarias en promedio para descubrir la presencia de algún fenómeno entre las unidades de muestreo; o a la inversa, para una muestra de un tamaño dado, la probabilidad de que al menos una unidad de muestreo contenga un ejemplo del fenómeno, para elementos con distintos grados de escasez. También aquí la agregación es importante, pues a medida que aumenta, es decir, a medida que la cantidad de elementos se concentra en un número progresivamente menor de unidades de muestreo, disminuyen las probabilidades de descubrimiento.

Finalmente, Nance ha aplicado también estas ideas a otro problema de considerable importancia en el muestreo regional. La prospección de superficie y las técnicas de muestreo, tal y como las conocemos hoy, se desarrollaron al principio en las regiones áridas del oeste y suroeste de los Estados Unidos, en donde la visibilidad superficial de los hallazgos es generalmente muy buena. Desde entonces también se ha prestado atención a los problemas que plantean esas técnicas allí donde la visibilidad en la superficie es escasa (por ejemplo, Wobst, 1983). En los Estados Unidos y debido a la sustanciosa financiación disponible para la arqueología contractual, se han hecho intentos por resolver las dificultades de la excavación de pequeñas catas de sondeo, por medio de unidades de muestreo mayores. Naturalmente, la cuestión surge respecto a las propiedades estadísticas de tales observaciones y las inferencias que pueden derivarse de ellas (Lovis, 1976; Nance, 1983; Nance y Ball, 1986; Wobst, 1983). El trabajo de Nance sobre las probabilidades de descubrimiento y la estimación de tales situaciones proporciona nuevamente una base para planificar diseños futuros y afirmar las limitaciones de los actuales. McManamon (1981) ha utilizado estas técnicas para estimar la densidad de yacimientos en diferentes estratos ambientales en Cape Cod. En principio, pues, es posible utilizar distribuciones de catas excavadas de manera rigurosa de forma similar a la recolección en superficie. Si los problemas prácticos planteados por la excavación de tales sondeos pueden dejarse de lado o no, es otra cuestión. En particular, las muestras de pequeño tamaño realizadas en la práctica y relativas a la escasez de los elementos que se están buscando constituyen un obstáculo mayor para conseguir estimaciones adecuadas (véase Wobst, 1983).

El enfoque de Bellhouse (Bellhouse, 1980; Bellhouse y Finlayson, 1979) para mejorar los métodos de muestreo espacial ya han sido expuestos durante la presentación del muestreo sistemático, así como los problemas asociados con él. A pesar de ellos, puede ser muy útil en situaciones concretas, y si se usa ampliamente empezarán a surgir algunas generalizaciones.

La insistencia en los aspectos espaciales del muestreo en los párrafos anteriores es un reflejo de la importancia del trabajo de campo en la investigación y de sus problemas inherentes, siempre bajo la perspectiva del muestreo. La manipulación de los datos recogidos no es menos importante, si bien sus características dependen, en última instancia, del trabajo de campo; además, los problemas intrínsecos asociados con el muestreo en conjuntos no son tan grandes, y ello se debe a varias razones. En primer lugar, la dimensión espacial ha sido apartada, excepto como base de la estratificación de la muestra. En segundo lugar, las limitaciones no estadísticas externas, si existen, son menos complejas y menos importantes. Finalmente, suele haber una cantidad considerable de redundancia en las poblaciones, en contraste con la situación a nivel de yacimiento o de área regional, de forma que el muestreo puede marcar una diferencia al ahorrar tiempo y dinero. La redundancia procede de dos fuentes principales: la mayor complejidad inherente y la variabilidad presente a nivel de yacimiento o a nivel regional, y la cantidad absoluta de elementos disponibles para el estudio a nivel del conjunto. Tiende a olvidarse en muchos muestreos arqueológicos que la característica más importante de las muestras, si nos referimos a la estimación, es su tamaño y no la fracción de muestreo; la fracción de una población muy grande necesaria para la extracción de una muestra significativa es a menudo muy pequeña. Es posible que hayamos de examinar 100 artefactos para obtener alguna estimación interesante y que eso pueda hacerse por medio de un pequeño porcentaje de la población de artefactos disponible. En contraste, si algún problema requiere información procedente de 100 yacimientos, podríamos representar una gran proporción de yacimientos en un área. ¡Ni que decir tiene que una investigación sobre 100 artefactos se hace a una escala mucho menor que una investigación referida a 100 yacimientos!

Ahora bien, por otro lado, Nance (1981) ha demostrado que para algunos propósitos, al menos, no es la cantidad total de elementos en la colección lo que importa, sino la cantidad de grupos (unidades de excavación o de prospección de superficie) de los que proceden esos elementos, por lo que estos temas no siempre son sencillos.

Objetivos arqueológicos del muestreo: alternativas a la estimación de los parámetros

Hemos visto que el muestreo probabilístico trata de darnos unas estimaciones cuya fiabilidad y precisión podemos establecer a partir de las caracterís-

ticas de la población. Se señaló también al principio del capítulo que el que tales medidas sean interesantes depende de nuestros propósitos. Lo más frecuente es que nos interese la variabilidad en nuestro yacimiento o área regional, de forma que nunca una medida general de cierta característica será útil por sí misma.

La respuesta a esto, en términos de muestreo, es el procedimiento de estratificación antes descrito. Si nos interesa saber si hay o no diferencias entre ciertas partes de nuestro yacimiento o área, la dividiremos en estratos y extraeremos una muestra de cada uno por separado. Sin embargo, aunque en principio la estratificación sea una buena idea, su empleo puede provocar problemas graves, tanto a nivel de yacimiento como a nivel regional. Por ejemplo, la cuestión de si tenemos suficiente información para generar una estratificación no puede ser marginada, así como tampoco la posibilidad de que estemos perpetuando las deformaciones existentes en la información al idear y aplicar un esquema de estratificación.

En el caso del análisis regional, los problemas se originaron porque pocas veces es viable una prospección regional centrada en un período particular o un tema específico: el trabajo de campo es tan costoso que sólo lo hacemos una única vez, recogiendo información acerca de todos los períodos. No suele haber motivos para creer que los factores que afectan a la densidad de yacimientos o hallazgos será constante para todos los períodos, y de hecho eso es poco probable. Por consiguiente, una estratificación relevante para un período no será relevante para otro, y la aplicación de distintos esquemas de estratificación entrecruzados será imposible en la práctica.

El mismo problema se plantea, naturalmente, en los yacimientos con varias fases de ocupación, que presentan problemas muy difíciles de estudiar si están estratificados. Ahora bien, hay también otro problema a nivel de yacimiento que surge tanto de la naturaleza destructiva de la excavación como del gasto que ésta ocasiona. Al excavar nos interesan, invariablemente, las características de varias poblaciones —cerámica de distintos tipos, huesos animales, semillas, estructuras, útiles líticos, etc.— y es poco probable que esos elementos estén distribuidos espacialmente en el yacimiento de la misma manera: una estratificación relevante para obtener buenas estimaciones de las características de la población para uno de ellos no será necesariamente buena para otros. La integración de propósitos diferentes sigue siendo un grave problema en la excavación.

Incluso cuando pueda llevarse a cabo, un muestreo estratificado no será satisfactorio para obtener ciertos tipos de información espacial. Se ha señalado en muchas ocasiones (por ejemplo, Redman, 1974) que en un caso en particular los estudios tanto en regiones como en yacimientos consisten tanto en un elemento probabilístico como en un elemento no probabilístico, y que para obtener información completa del patrón espacial en un área habrá que cubrir toda el área, ya sea mediante prospecciones o excavaciones, con el fin de con-

seguir un mapa completo. Incluso mediante una excavación en extensión, no podemos dejar de considerar la necesidad de extraer muestras, porque siempre se nos plantea la cuestión de cuán intensiva ha de ser la investigación con el fin de recoger los elementos poco frecuentes en la población, o bien cuán intensivo ha de ser el análisis del suelo para recuperar, por ejemplo, restos vegetales o restos de talla de los útiles líticos (véase la discusión que hemos efectuado antes acerca de los trabajos de Nance).

En su vertiente positiva, sin embargo, una investigación total y continua no siempre es necesaria para hacer inferencias sobre los esquemas espaciales. A pesar de lo que hemos dicho antes sobre el uso limitado de estadísticas descriptivas para explicar la variabilidad interna, existen ciertos estadígrafos globales, calculables mediante los datos de la muestra, que nos explican si la tendencia general del patrón espacial, en particular el grado en que la distribución se divide en grupos, es aleatoria o dispersa. El índice de dispersión de Morisita, por ejemplo (véanse Rogge y Fuller, 1977; Shennan, 1985), a diferencia del análisis del vecino más próximo o las formas más habituales del análisis cuadrático (véase Hodder y Orton, 1976), sólo requiere datos acerca de las frecuencias de yacimientos por unidad de muestra y esas unidades no tienen por qué ser contiguas. Dados los problemas planteados por la conglomerización o la agregación en la estimación de parámetros, estudiar su alcance puede ser un prerrequisito importante para la evaluación de otras estimaciones. Plog (1974) ha tratado otros aspectos útiles del patrón espacial que pueden ser analizados mediante tales índices generales.

Pero incluso cuando queremos información detallada acerca del patrón espacial, no siempre será necesaria una investigación total; además, no siempre sería realizable. En algunos casos de cartografía, el uso de datos muestrales es satisfactorio, si bien no estaremos usando la muestra para estimar parámetros; el criterio para elegir una buena muestra habrá de ser, por consiguiente, distinto. El más importante de esos criterios es la regularidad espacial: ha de haber un elemento sistemático en el esquema muestral, de forma que las observaciones aparezcan registradas en todas las áreas analizadas. El paradigma aquí es la prospección del contorno del yacimiento, en el que se dispone de una cuadrícula regular para las medidas, se toma una medida en cada una y los contornos se interpolan a continuación. En una prospección como esta, no obstante, es posible ver dónde se produce un desfase entre los puntos, con lo que sabremos si son o no representativos y podremos, de ser necesario, considerar nuevas muestras si un punto en particular de la cuadrícula no es representativo, o si hay varianzas locales en el desfase entre puntos no caracterizadas suficientemente por las observaciones en la cuadrícula. Por otro lado, cuando interpolamos los esquemas de variación espacial en las distribuciones arqueológicas nos encontramos a oscuras, como sucede con la estimación de parámetros, y no podemos ver cuándo surgen problemas.

Desde este punto de vista, aunque una distribución regular de datos es esencial, las muestras sistemáticas plantean problemas tan graves como los de la estimación de parámetros: las regularidades de las distribuciones —como, por ejemplo, casas en un asentamiento— son potencialmente catastróficas desde ambos puntos de vista. La respuesta propuesta por los geógrafos que encaran problemas muy parecidos es el uso de un *muestreo estratificado sistemático no alineado* (véase, por ejemplo, Haggett *et al.*, 1977, pp. 272-274). Está ilustrado en la figura 14.2.

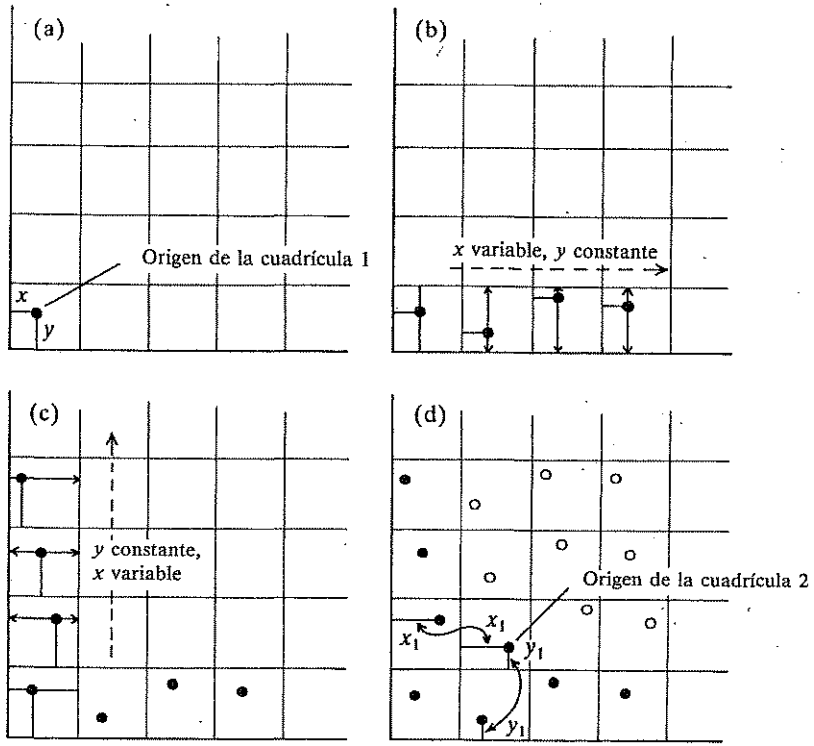


FIGURA 14.2. Fases en la generación de una muestra estratificada sistemática no alineada (según Haggett *et al.*, 1977).

En primer lugar, el área se divide mediante una cuadrícula. Empezamos por el cuadro inferior o el superior izquierdo, usando dos números aleatorios para definir las coordendas de un punto en esa cuadrícula (fig. 14.2a). Todas las cuadrículas a lo largo de la fila inferior o la superior tienen la misma coordenada *x* que la primera de ellas, eligiéndose la coordenada *y* de cada uno mediante números aleatorios (14.2b). Igualmente, todas las cuadrículas en el margen izquierdo del área tienen la misma coordenada *y* que la primera de ellas,

eligiéndose su x respectiva aleatoriamente (14.2c). En todas las demás cuadrículas, el punto de la muestra se obtiene considerando la coordenada x de la fila en la que se encuentra, y la coordenada y de su columna (14.2d).

En efecto, esto es una muestra estratificada con exactitud, con los cuadros haciendo las veces de estratos. El método parece tener la capacidad de evitar el problema de la periodicidad y, al mismo tiempo, proporciona información distribuida regularmente; por otro lado, se ha usado muy poco en la práctica arqueológica para estimar características de poblaciones con intervalos de confianza asociados, el otro propósito para el que se necesitan muestras (con la excepción de Bellhouse, 1980).

Incluso utilizando un método como el anterior, no disponemos aún de un criterio satisfactorio que nos explique la calidad de nuestras interpolaciones. Una forma de resolverlo es por medio del análisis superficial de tendencias para ver si hay alguna tendencia espacial en los valores de las variables que nos interesan. Adopta la forma de una regresión múltiple con dos variables independientes, las dos coordenadas espaciales de los puntos, y una dependiente, el valor de la variable de interés en los puntos (véanse, por ejemplo, Davis, 1973; Orton, 1980). Un método más sofisticado de interpolación espacial se denomina *kriging*: consiste en el uso de los presupuestos de autocorrelación espacial (véase Hodder y Orton, 1976, pp. 174-183, para una exposición arqueológica de esto último) y sirve para estimar el valor de una variable distribuida espacialmente en una localización a partir de sus valores en localizaciones adyacentes; sólo Zubrow y Harbaugh (1978) parecen haber usado esa técnica en una investigación arqueológica. Tal y como muestran, su aplicación no se restringe a la representación espacial; puede usarse de forma predictiva, como en geología, para maximizar la probabilidad de descubrimiento de yacimientos.

Es importante señalar que un ejercicio de prospectiva como ese, que intenta maximizar los resultados, es muy diferente, en sus propósitos, de la obtención de información representativa acerca de, por ejemplo, la densidad de yacimientos en una región. Quizás algunos autores no siempre tengan claro cuál de esos objetivos cumplen o han de cumplir. Si se ha llevado a cabo una prospección del área, puede decirse que maximizar el descubrimiento de yacimientos es mucho más apropiado que estimar la densidad regional de los yacimientos.

Todos los métodos descritos hasta aquí son diferentes enfoques para obtener diferentes tipos de datos espaciales. Como es habitual, serán los objetivos del investigador los que determinarán cuál de ellos es más apropiado. Para tomar esa decisión es importante tener presentes las opciones disponibles y no asumir automáticamente que un esquema simple de muestreo probabilístico será siempre el mejor.

Incluso una conclusión radical de esta discusión del patrón espacial en las distribuciones sería que la estimación de las medias y de los errores típicos es casi totalmente irrelevante: de lo que se trata es, sobre todo, de la forma de la distribución a la que se ha prestado poca atención, aparte de la ya señalada.

Este punto no es distinto de las críticas expresadas por el análisis de datos exploratorio con respecto a las distribuciones estadísticas en general: si una distribución tiene peculiaridades de forma, constituyen lo más importante acerca de ellas, y hemos de conocerlas. Respecto a la agrupación significativa, puede decirse que los promedios son irrelevantes.

Pero podemos ir aún más adelante y afirmar que lo que interesa sobre todo, a diferencia de la estimación, es explicar la variación que observamos. Veamos un ejemplo a escala regional: puede interesarnos menos la estimación de la densidad de yacimientos que la comprensión de los factores subyacentes a la localización de los yacimientos, o, más en general, subyacentes a las variaciones en la densidad de artefactos sobre la superficie del suelo (por ejemplo, Shennan, 1985). En estas circunstancias, antes que adoptar automáticamente una técnica de muestreo probabilístico, hemos de establecer un diseño experimental clásico, eligiendo combinaciones de factores que consideremos relevantes a las variaciones en densidad, y a continuación salir al campo y llevar a cabo prospecciones de superficie en los lugares en los que se dé la combinación de factores elegida. Los datos resultantes pueden analizarse entonces usando técnicas del análisis de varianza (similar a la regresión, pero usando variables independientes a escala nominal; véase, por ejemplo, Blalock, 1972, capítulo 16) y una medida obtenida del porcentaje de variación en la densidad de artefactos explicada por los distintos factores. Naturalmente, habiendo identificado las diversas combinaciones de factores que nos interesan, podríamos, si fuese necesario, extraer probabilísticamente muestras de forma bastante simple.

Aunque no se hayan recogido los datos de manera rigurosa, ni siguiendo un esquema de muestreo probabilístico, podemos usar ese enfoque para analizar la variación en los datos que hemos recogido de manera válida, aunque no seamos capaces de inferir estadísticamente las características de una población mayor. El problema principal, en este caso, se origina probablemente en el hecho de que el diseño resultante pueda estar tan desequilibrado que los efectos de distintas variables se confunden uno con otro, y no pueden distinguirse.

Un argumento similar ha sido adoptado por Wobst (1983), que expresa un escepticismo considerable acerca de los resultados de muchos proyectos basados en el muestreo probabilístico, porque en muchas regiones en las que la visibilidad de los hallazgos es pobre, recoger observaciones suficientes para producir estimaciones con el grado adecuado de precisión es una tarea imposible. La respuesta de Wobst es que hemos de sustituir las estimaciones con un propósito general por la comprobación de hipótesis, pues las observaciones requeridas para la comprobación de una hipótesis pueden especificarse con más precisión y, por consiguiente, tendremos más oportunidades de obtenerlas.

De todo ello se desprende, obviamente, que todo lo que se refiere al muestreo probabilístico es muy complejo y está aún por resolver; tanto si Wobst tiene razón como si no, esas técnicas nunca deben sustituir la reflexión propia del arqueólogo.

EJERCICIOS

14.1. En un estudio de los asentamientos en un área se ha decidido empezar estudiando la localización de los yacimientos conocidos, para ensayar y desarrollar algunos principios predictivos que guíen las prospecciones subsiguientes para buscar nuevos yacimientos. En cada yacimiento se registraron diversas características espaciales y el objetivo es obtener una estimación del porcentaje de yacimientos con características específicas, con una exactitud de ± 5 % y una confianza del 95 %. La cantidad de yacimientos es 291. ¿Cuántos habrá que investigar, asumiendo un muestreo aleatorio simple?

14.2. En un estudio del material procedente de la excavación de un asentamiento neolítico en Hungría, se decidió que era importante investigar la fragmentación de la cerámica, como base para comprender la naturaleza de los distintos depósitos. Para ahorrar tiempo y dinero, se ha decidido elegir una muestra del tipo de pasta de la cerámica en cada depósito, y no pesar todos los fragmentos. La muestra será extraída mediante un muestreo aleatorio simple. Para uno de los tipos de pasta en particular, la selección de una muestra preliminar proporcionó una desviación típica de 25 g. Se ha decidido que una tolerancia de 5 g en el peso estimado es aceptable, con una probabilidad del 95 %. Calcula el tamaño de una muestra para conseguir ese objetivo, que puede ser corregido a continuación para los diversos tamaños de las poblaciones en los distintos depósitos.

14.3. Se ha efectuado una prospección arqueológica en un área de 100 km². La prospección está basada en una muestra aleatoria simple de cuadrículas de una hectárea, que suman 5 km² en total. Las densidades de material en cada cuadrícula han sido registradas; la densidad media de los útiles líticos del neolítico y del bronce antiguo es de 16,95/ha, con una desviación típica de 7,03 y una distribución no muy alejada de la normalidad, aunque asimétrica positiva. Calcula un intervalo de confianza para la cantidad total de artefactos líticos en el total del área prospectada, con una probabilidad del 99 %.

14.4. Se ha excavado un yacimiento de cazadores-recolectores a partir de una cuadrícula con divisiones de 2 × 2 metros. A continuación aparece una lista con la cantidad de piezas líticas halladas en cada una de las 50 cuadrículas excavadas:

2	5	15	17	11	26	25	28	23	22
38	37	35	30	39	48	47	45	48	42
47	45	41	55	50	59	51	59	56	57
53	61	67	64	63	60	79	75	77	72
71	85	82	89	96	93	95	108	103	117

1. Calcula la media y la desviación típica para todo el yacimiento.
2. *a*) Usa números aleatorios para elegir diez muestras aleatorias de 30 cuadrículas; *b*) calcula la media y la desviación típica de cada muestra, junto con la media de las medias; *c*) calcula el error típico de la media para cada muestra, y los intervalos de confianza de la media al 95 % y al 99 %; *d*) ¿cómo se relacionan los intervalos de confianza entre sí y con la media global de la población?
3. Repite 2) para muestras de 10 y de 20 cuadrículas.

NOTA: Ha de usarse la distribución *t* para los intervalos de confianza.

15. CONCLUSIÓN

Nuevos avances

Si el lector ha conseguido llegar hasta aquí, descubrirá que ha alcanzado un nivel de competencia que le capacita para llevar a cabo algunos análisis de datos básicos y para entenderse con estadísticos profesionales; el lector está capacitado también para seguir la mayoría de las argumentaciones estadísticas en la bibliografía y hacerse con la idea general del resto; igualmente, el lector habrá conseguido una base firme para proseguir las lecturas y conocer nuevas técnicas. Ya se han señalado algunos manuales apropiados: Blalock (1972) es muy bueno para lo que se refiere a lo abordado en la primera parte, aunque esté algo anticuado en ciertos aspectos. Hartwig y Dearing (1979) y Tukey (1977) proporcionan un sólido fundamento del análisis de datos exploratorio, así como, en un nivel mucho más avanzado, el libro de Mosteller y Tukey (1977). Johnston (1978) es muy apropiado para ampliar las técnicas más avanzadas.

La principal dificultad, no obstante, es que sin un adecuado conocimiento del cálculo y del álgebra matricial, el lector se verá incapaz de entender muchos de los libros matemáticos sobre estadística y análisis de datos. Davis (1973) proporciona una introducción sencilla al álgebra matricial, mientras que Wilson y Kirby (1980) cubren también el cálculo integral y diferencial. El libro de Everitt y Dunn (1983) es un buen manual de nivel intermedio que requiere algún conocimiento de álgebra matricial, pero no es de gran dificultad.

Si el lector piensa continuar con los análisis cuantitativos de datos, el grado en que lo haga y la dirección en que los emprenda estarán determinados por los problemas específicos con los que se enfrenta: algunos de ellos pueden analizarse por medio de los métodos presentados en este libro, y otros relacionados con ellos, mientras que otros problemas requerirán el desarrollo de nuevas técnicas.*

* En castellano, pueden consultarse textos relativamente sencillos como el de Domènech y Riba (1985), para el análisis de regresión, y el compilado por Sánchez Carrión (1984), para los análisis de conglomerados, análisis de modelos lineales logarítmicos en tablas de contingencia, análisis de componentes principales y análisis de correspondencias. Sobre estos últimos, y en un nivel mucho más avanzado, Cuadras (1981). El libro de Blalock está traducido al castellano; desgraciadamente no existe ningún manual en castellano (ni original ni traducción) sobre el análisis de datos exploratorio. Lo que de él explican los manuales generales es mucho menos de lo que se explica en este mismo libro. En cuanto a las aplicaciones a la arqueología de la península ibérica pueden consultarse en las actas de algunos coloquios recientes (Jornadas de Metodología de la Investigación Prehistórica [Soria, 1981], Arqueología Espacial, Aplicaciones Informáticas de la Arqueología [Madrid, 1990]...), así como en algunos artículos recientes en revistas como *Cuadernos de Prehistoria de la Universidad de Granada, Boletín del Seminario de Arqueología de la Universidad de Valladolid* y *Trabajos de Prehistoria*. Dos libros importantes son los de Lull (1983) y Nocete (1989). (*N. del t.*)

Los trabajos recientes proporcionan ejemplos de estas dos soluciones a los problemas planteados. Los modelos logarítmicos lineales están encontrando cada vez más aplicaciones (por ejemplo, Hietala, 1984; Leese y Needham, 1986), mientras que el análisis de conglomerados ha sido aplicado de forma innovadora en el análisis microespacial (por ejemplo, Whallon, 1984; Kintigh y Ammerman, 1982). Han continuado, asimismo, los trabajos sobre la aplicación de técnicas particulares, como el muestreo probabilístico (por ejemplo, Shott, 1985; Van der Veen, 1985).

A largo plazo, sin embargo, es probable que sean necesarias nuevas técnicas o, incluso, técnicas no clásicas para resolver muchos problemas arqueológicos, y es ahí donde la colaboración entre estadísticos y arqueólogos es especialmente importante. En particular, el uso de técnicas llamadas en inglés de *bootstrap* se irá haciendo, probablemente, más importante; en ellas los análisis matemáticos y estadísticos están dirigidos, en cierta forma, por las complejidades del problema a resolver, consistiendo, por lo general, en la simulación informática de las distribuciones que tengan en cuenta las limitaciones de los datos reales, y que se comparen con los esquemas de relación definidos entre los datos (por ejemplo, Bradley y Small, 1985; Simek, 1984; Berry *et al.*, 1984). Bastante similar en el enfoque son ciertos trabajos de modelización de las relaciones entre la evidencia arqueológica, tal y como ha sido recuperada y tal como fue originalmente depositada (Orton, 1982, en el caso de la cerámica; Fieller y Turner, 1982; Turner y Fieller, 1985, en el caso de los huesos). Estos estudios van mucho más allá del muestreo probabilístico e indican nuevas posibilidades en el desarrollo de modelos rigurosos en los que las matemáticas coinciden con el problema arqueológico.

Muchos de los desarrollos se sitúan fuera del campo del análisis cuantitativo de datos, considerado en su forma clásica. El análisis de conglomerados, por ejemplo, puede ser visto ahora como parte de la disciplina en rápido desarrollo del reconocimiento informático de esquemas y patrones (por ejemplo, Bow, 1984), que en un futuro próximo puede hacer importantes contribuciones al análisis de la forma, ya de objetos o de distribuciones espaciales; algunas de esas técnicas han sido empleadas en arqueología desde hace tiempo para

mejorar el trazado de las lecturas de las prospecciones geofísicas (Scollar, 1969) y que recientemente se han aplicado a las formas (Gero y Mazzullo 1984).

La tendencia general en estos desarrollos es clara: se trata de la creciente integración del conocimiento arqueológico y de la información en los análisis cuantitativos. Los análisis típicos *prêt-à-porter* seguirán siendo apropiados para muchos propósitos, pero no en otros.

Otra línea de investigación en arqueología señala la misma dirección: el desarrollo de sistemas «expertos» y de bases de conocimiento inteligentes para ordenadores (véase, por ejemplo, Huggett 1985). Consiste en definir cuerpos de conocimiento y reglas de inferencia que permitan extraer conclusiones apropiadas al introducir nuevos problemas al sistema; el diagnóstico médico es, posiblemente, su campo de aplicación más conocido hoy en día. Aunque su uso en arqueología plantea indudables problemas, como indica Huggett, la idea subyacente es muy atractiva, sobre todo en el contexto de los desarrollos en el análisis de datos antes señalados.

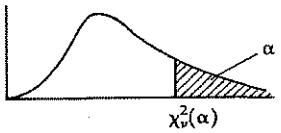
Algunos pueden pensar que desarrollos de esa clase pueden ser interesantes, pero que difícilmente afectarán a algo más que a una pequeñísima área de la profesión arqueológica. Si esto es así o no, depende de dos cosas. La primera es la disponibilidad de los medios, y toda la experiencia de los últimos años sugiere que las grandes capacidades en ordenadores y soportes lógicos seguirán estando disponibles, ¡incluso para los arqueólogos! La segunda es el nivel de educación de los profesionales de la arqueología en el análisis de datos y en informática. Con respecto a esto último, está claro que se está llegando paulatinamente a ese nivel. Si este libro constituye una contribución al primer aspecto, será algo muy positivo, pues ayudará a crear el clima apropiado para la realización del potencial en las grandes cantidades de datos arqueológicos, cuyas posibilidades siguen estando intactas.

ANEXOS

ANEXO 1. TABLAS ESTADÍSTICAS

TABLA A. Puntos porcentuales de la distribución del χ^2.

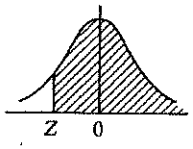
Los valores tabulados son $\chi^2_\nu(\alpha)$, donde
Pr $(\chi^2_\nu > \chi^2_\nu(\alpha)) = \alpha$, para ν grados de libertad.



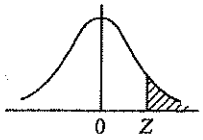
ν	α = 0,995	0,990	0,975	0,950	0,900	0,750	0,500
1	$392704,10^{-10}$	$157088,10^{-9}$	$982069,10^{-9}$	$393214,10^{-8}$	0,0157908	0,1015308	0,454936
2	0,0100251	0,0201007	0,0506356	0,102587	0,210721	0,575364	1,38629
3	0,0717218	0,114832	0,215795	0,351846	0,584374	1,212534	2,36597
4	0,206989	0,297109	0,484419	0,710723	0,063623	1,92256	3,35669
5	0,411742	0,554298	0,831212	1,145476	1,61031	2,67460	4,35146
6	0,675727	0,872090	1,23734	1,63538	2,20413	3,45460	5,34812
7	0,989256	1,239043	1,68987	2,16735	2,83311	4,25485	6,34581
8	1,34441	1,64650	2,17973	2,73264	3,48954	5,07064	7,34412
9	1,73493	2,08790	2,70039	3,32511	4,16816	5,89883	8,34283
10	2,15586	2,55821	3,24697	3,94030	4,86518	6,73720	9,34182
11	2,60322	3,05348	3,81575	4,57481	5,57778	7,58414	10,3410
12	3,07382	3,57057	4,40379	5,22603	6,30380	8,43842	11,3403
13	3,56503	4,10692	5,00875	5,89186	7,04150	9,29907	12,3398
14	4,07467	4,66043	5,62873	6,57063	7,78953	10,1653	13,3393
15	4,60092	5,22935	6,26214	7,26094	8,54676	11,0365	14,3389
16	5,14221	5,81221	6,90766	7,96165	9,31224	11,9122	15,3385
17	5,69722	6,40776	7,56419	8,67176	10,0852	12,7919	16,3382
18	6,26480	7,01491	8,23075	9,39046	10,8649	13,6753	17,3379
19	6,84397	7,63273	8,90652	10,1170	11,6509	14,5620	18,3377
20	7,43384	8,26040	9,59078	10,8508	12,4426	15,4518	19,3374
21	8,03365	8,89720	10,28293	11,5913	13,2396	16,3444	20,3372
22	8,64272	9,54249	10,9823	12,3380	14,0415	17,2396	21,3370
23	9,26043	10,19567	11,6886	13,0905	14,8480	18,1373	22,3369
24	9,88623	10,8564	12,4012	13,8484	15,6587	19,0373	23,3367
25	10,5197	11,5240	13,1197	14,6114	16,4734	19,9393	24,3366
26	11,1602	12,981	13,8439	15,3792	17,2919	20,8434	25,3365
27	11,8076	12,8785	14,5734	16,1514	18,1139	21,7494	26,3363
28	12,4613	13,5647	15,3079	16,9279	18,9392	22,6572	27,3362
29	13,1211	14,2565	16,0471	17,7084	19,7677	23,5666	28,3361
30	13,7867	14,9535	16,7908	18,4927	20,6992	24,4776	29,3360
40	20,7065	22,1643	24,4330	26,5093	20,0505	33,6603	39,3353
50	27,9907	29,7067	32,3574	34,7643	37,6886	42,9421	49,3349
60	35,5345	37,4849	40,4817	43,1880	46,4589	52,2938	59,3347
70	43,2752	45,4417	48,7576	51,7393	55,3289	61,6983	69,3345
80	51,1719	53,5401	57,1532	60,3915	64,2778	71,1445	79,3343
90	59,1963	61,7541	65,6466	69,1260	73,2911	80,6247	89,3342
100	67,3276	70,0649	74,2219	77,9295	82,3581	90,1332	99,3341

ν	α = 0,250	0,100	0,050	0,025	0,010	0,005	0,001
1	1,32330	2,70554	3,84146	5,02389	6,63490	7,87944	10,828
2	2,77529	4,60517	5,99146	7,37776	9,21034	10,5966	13,816
3	4,10834	6,25139	7,81473	9,34840	11,3449	12,8382	16,266
4	5,38527	7,77944	9,48773	11,1433	13,2767	14,8603	18,467
5	6,62568	9,23636	11,0705	12,8325	15,0863	16,7496	20,515
6	7,84080	10,6446	12,5916	14,4494	16,8119	18,5476	22,458
7	9,03715	12,0170	14,0671	16,0128	18,4753	20,2777	24,322
8	10,2189	13,3616	15,5073	17,5345	20,0902	21,9550	26,125
9	11,3888	14,6837	16,9190	19,0228	21,6660	23,5894	27,877
10	12,5489	16,9872	18,3070	20,4832	23,2093	25,1882	28,588
11	13,7007	17,2750	19,6751	21,9200	24,7250	26,7568	31,264
12	14,8454	18,5493	21,0261	23,3367	26,2170	28,2995	32,909
13	15,9839	19,8119	22,3620	24,7356	27,6882	29,8195	34,528
14	17,1169	21,0641	23,6848	26,1189	29,1412	31,3194	36,123
15	18,2451	22,3071	24,9958	27,4884	30,5779	32,8013	37,697
16	19,3689	23,5418	26,2962	28,8454	31,9999	34,2672	39,252
17	20,4887	24,7690	27,5871	30,1910	33,4087	35,7185	40,790
18	21,6049	25,9894	28,8693	31,5264	34,8053	37,1565	42,312
19	22,7178	27,2036	30,1435	32,8523	36,1909	38,5823	43,820
20	23,8277	28,4120	31,4104	34,1696	37,5662	39,9968	45,315
21	24,9348	29,6151	32,6706	35,4789	38,9322	41,4011	46,797
22	26,0393	30,8133	33,9244	36,7807	40,2894	42,7957	48,268
23	27,1413	32,0069	35,1725	38,0756	41,6384	44,1813	49,728
24	28,2412	33,1962	36,4150	39,3641	42,9798	45,5585	51,179
25	29,3389	34,3816	37,6525	40,6465	44,3141	46,9279	52,618
26	30,4346	35,5632	38,8851	41,9232	45,6417	48,2899	54,052
27	31,5284	36,7412	40,1133	43,1945	46,9629	49,6449	55,476
28	32,6205	37,9159	41,3371	44,4608	48,2782	50,9934	56,892
29	33,7109	39,0875	42,5570	45,7223	49,5879	52,3356	58,301
30	34,7997	40,2560	43,7730	46,9792	50,8922	53,6720	59,703
40	45,6160	51,8051	55,7585	59,3417	63,6907	66,7660	73,402
50	56,3336	63,1671	67,5048	71,4202	76,1539	79,4900	86,661
60	66,9815	74,3970	79,0819	83,2977	88,3794	91,9517	99,607
70	77,5767	85,5270	90,5312	95,0232	100,425	104,215	112,317
80	88,1303	96,5782	101,879	106,629	112,329	116,321	124,839
90	98,6499	107,565	113,145	118,136	124,116	128,299	137,208
100	109,141	118,498	124,342	129,561	135,807	140,169	149,449

TABLA B. Áreas de la distribución normal estandarizada.



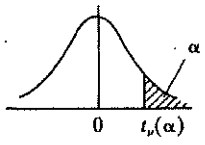
Z	−0,09	−0,08	−0,07	−0,06	−0,05	−0,04	−0,03	−0,02	−0,01	−0,00
−3,9	0,99997	0,99997	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99995	0,99995
−3,8	0,99995	0,99995	0,99995	0,99994	0,99994	0,99994	0,99994	0,99993	0,99993	0,99993
−3,7	0,99992	0,99992	0,99992	0,99992	0,99991	0,99991	0,99990	0,99990	0,99990	0,99989
−3,6	0,99989	0,99988	0,99988	0,99987	0,99987	0,99986	0,99986	0,99985	0,99985	0,99984
−3,5	0,99983	0,99983	0,99982	0,99981	0,99981	0,99980	0,99979	0,99978	0,99978	0,99977
−3,4	0,99976	0,99975	0,99974	0,99973	0,99972	0,99971	0,99970	0,99969	0,99968	0,99966
−3,3	0,99965	0,99964	0,99962	0,99961	0,99960	0,99958	0,99957	0,99955	0,99953	0,99952
−3,2	0,99950	0,99948	0,99946	0,99944	0,99942	0,99940	0,99938	0,99936	0,99934	0,99931
−3,1	0,99929	0,99926	0,99924	0,99921	0,99918	0,99916	0,99913	0,99910	0,99906	0,99903
−3,0	0,99900	0,99896	0,99893	0,99898	0,99886	0,99882	0,99878	0,99874	0,99869	0,99865
−2,9	0,99861	0,99856	0,99851	0,99846	0,99841	0,99836	0,99831	0,99825	0,99819	0,99813
−2,8	0,99807	0,99801	0,99795	0,99788	0,99781	0,99774	0,99767	0,99760	0,99752	0,99744
−2,7	0,99736	0,99728	0,99720	0,99711	0,99702	0,99693	0,99683	0,99674	0,99664	0,99653
−2,6	0,99643	0,99632	0,99621	0,99609	0,99598	0,99585	0,99573	0,99560	0,99547	0,99534
−2,5	0,99520	0,99506	0,99492	0,99477	0,99461	0,99446	0,99430	0,99413	0,99396	0,99379
−2,4	0,99361	0,99343	0,99324	0,99305	0,99286	0,99266	0,99245	0,99224	0,99202	0,99180
−2,3	0,99158	0,99134	0,99111	0,99086	0,99061	0,99036	0,99010	0,98983	0,98956	0,98928
−2,2	0,98899	0,98870	0,98840	0,98809	0,98778	0,98745	0,98713	0,98679	0,98645	0,98610
−2,1	0,98574	0,98537	0,98500	0,98461	0,98422	0,98392	0,98341	0,98300	0,98257	0,98214
−2,0	0,98169	0,98124	0,98077	0,98030	0,97982	0,97932	0,97882	0,97831	0,97778	0,97725
−1,9	0,97670	0,97615	0,97558	0,97500	0,97441	0,97381	0,97320	0,97257	0,97193	0,97128
−1,8	0,97062	0,96995	0,96926	0,96856	0,96846	0,96712	0,96638	0,96562	0,96485	0,96407
−1,7	0,96327	0,96246	0,96164	0,96080	0,95994	0,95907	0,95818	0,95728	0,95637	0,95543
−1,6	0,95449	0,95352	0,95254	0,95154	0,95053	0,94950	0,94845	0,94738	0,94630	0,94520
−1,5	0,94408	0,94295	0,94179	0,94062	0,93943	0,93822	0,93699	0,93574	0,93448	0,93319
−1,4	0,93189	0,93056	0,92922	0,92785	0,92647	0,92507	0,92364	0,92220	0,92073	0,91924
−1,3	0,91774	0,91621	0,91466	0,91308	0,91149	0,90988	0,90824	0,90658	0,90490	0,90320
−1,2	0,90147	0,89973	0,89796	0,89617	0,89435	0,89251	0,89065	0,88877	0,88686	0,88493
−1,1	0,88298	0,88100	0,87900	0,87698	0,87493	0,87286	0,87076	0,86864	0,86650	0,86433
−1,0	0,86214	0,85993	0,85769	0,85543	0,85314	0,85083	0,84850	0,84614	0,84375	0,84134
−0,9	0,83891	0,83646	0,83398	0,83147	0,82894	0,82639	0,82381	0,82121	0,81859	0,81594
−0,8	0,81327	0,81057	0,80785	0,80511	0,80234	0,79955	0,79673	0,79389	0,79103	0,78814
−0,7	0,78524	0,78230	0,77935	0,77637	0,77337	0,77035	0,76731	0,76424	0,76115	0,75804
−0,6	0,75490	0,75175	0,74857	0,74537	0,74215	0,73891	0,73565	0,73237	0,72907	0,72575
−0,5	0,72240	0,71904	0,71566	0,71226	0,70884	0,70540	0,70194	0,69847	0,69497	0,69146
−0,4	0,68739	0,68439	0,68082	0,67724	0,67364	0,67003	0,66640	0,66276	0,65910	0,65542
−0,3	0,65173	0,64803	0,64431	0,64058	0,63683	0,63307	0,62930	0,62552	0,62172	0,61791
−0,2	0,61409	0,61026	0,60642	0,60257	0,59871	0,59483	0,59095	0,58706	0,58317	0,57926
−0,1	0,57535	0,57142	0,56750	0,56356	0,55962	0,55567	0,55172	0,54776	0,54380	0,53983
−0,0	0,53586	0,53188	0,52790	0,52392	0,51994	0,51595	0,51197	0,50798	0,50399	0,50000



Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,49601	0,49202	0,48803	0,48405	0,48006	0,47608	0,47210	0,46812	0,46414
0,1	0,46017	0,45620	0,45224	0,44828	0,44433	0,44038	0,43644	0,43250	0,42858	0,42465
0,2	0,42074	0,41683	0,41294	0,40905	0,40517	0,40129	0,39743	0,39358	0,38974	0,38591
0,3	0,38209	0,37828	0,37448	0,37070	0,36693	0,36317	0,35942	0,35569	0,35197	0,34827
0,4	0,34458	0,34090	0,33724	0,33360	0,32997	0,32636	0,32276	0,31918	0,31561	0,31207
0,5	0,30854	0,30503	0,30153	0,29806	0,29460	0,29116	0,28774	0,28434	0,28096	0,27760
0,6	0,27425	0,27093	0,26763	0,26435	0,26109	0,25785	0,25463	0,25143	0,24825	0,24510
0,7	0,24196	0,23885	0,23576	0,23269	0,22965	0,22663	0,22363	0,22065	0,21770	0,21476
0,8	0,21186	0,20897	0,20611	0,20327	0,20045	0,19766	0,19489	0,19215	0,18943	0,18673
0,9	0,18406	0,18141	0,17879	0,17619	0,17361	0,17106	0,16853	0,16602	0,16354	0,16109
1,0	0,15866	0,15625	0,15386	0,15150	0,14917	0,14686	0,14457	0,14231	0,14007	0,13786
1,1	0,13567	0,13350	0,13136	0,12924	0,12714	0,12507	0,12302	0,12100	0,11900	0,11702
1,2	0,11507	0,11314	0,11123	0,10935	0,10749	0,10565	0,10383	0,10204	0,10027	0,09853
1,3	0,09680	0,09510	0,09342	0,09176	0,09012	0,08851	0,08692	0,08534	0,08379	0,08226
1,4	0,08076	0,07927	0,07780	0,07636	0,07493	0,07353	0,07215	0,07078	0,06944	0,06811
1,5	0,06681	0,06552	0,06426	0,06301	0,06178	0,06057	0,05938	0,05821	0,05705	0,05592
1,6	0,05480	0,05370	0,05262	0,05155	0,05050	0,04947	0,04846	0,04746	0,04648	0,04551
1,7	0,04457	0,04363	0,04272	0,04182	0,04093	0,04006	0,03920	0,03836	0,03754	0,03673
1,8	0,03593	0,03515	0,03438	0,03362	0,03288	0,03216	0,03144	0,03074	0,03005	0,02938
1,9	0,02872	0,02807	0,02743	0,02680	0,02619	0,02559	0,02500	0,02442	0,02385	0,02330
2,0	0,02275	0,02222	0,02169	0,02118	0,02068	0,02018	0,01970	0,01923	0,01876	0,01831
2,1	0,01786	0,01743	0,01700	0,01659	0,01618	0,01578	0,01539	0,01500	0,01463	0,01426
2,2	0,01390	0,01355	0,01321	0,01287	0,01255	0,01222	0,01191	0,01160	0,01130	0,01101
2,3	0,01072	0,01044	0,01017	0,00990	0,00964	0,00939	0,00914	0,00889	0,00866	0,00842
2,4	0,00820	0,00798	0,00776	0,00755	0,00734	0,00714	0,00695	0,00676	0,00657	0,00639
2,5	0,00621	0,00604	0,00587	0,00570	0,00554	0,00539	0,00523	0,00508	0,00494	0,00480
2,6	0,00466	0,00453	0,00440	0,00427	0,00415	0,00402	0,00391	0,00379	0,00368	0,00357
2,7	0,00347	0,00336	0,00326	0,00317	0,00307	0,00298	0,00289	0,00280	0,00272	0,00264
2,8	0,00256	0,00248	0,00240	0,00233	0,00226	0,00219	0,00212	0,00205	0,00199	0,00193
2,9	0,00187	0,00181	0,00175	0,00169	0,00164	0,00159	0,00154	0,00149	0,00144	0,00139
3,0	0,00136	0,00131	0,00126	0,00122	0,00118	0,00114	0,00111	0,00107	0,00104	0,00100
3,1	0,00097	0,00094	0,00090	0,00087	0,00084	0,00082	0,00079	0,00076	0,00074	0,00071
3,2	0,00069	0,00066	0,00064	0,00062	0,00060	0,00058	0,00056	0,00054	0,00052	0,00050
3,3	0,00048	0,00047	0,00045	0,00043	0,00042	0,00040	0,00039	0,00038	0,00036	0,00035
3,4	0,00034	0,00032	0,00031	0,00030	0,00029	0,00028	0,00027	0,00026	0,00025	0,00024
3,5	0,00023	0,00022	0,00022	0,00021	0,00020	0,00019	0,00019	0,00018	0,00017	0,00017
3,6	0,00016	0,00015	0,00015	0,00014	0,00014	0,00013	0,00013	0,00012	0,00012	0,00011
3,7	0,00011	0,00010	0,00010	0,00010	0,00009	0,00009	0,00008	0,00008	0,00008	0,00008
3,8	0,00007	0,00007	0,00007	0,00006	0,00006	0,00006	0,00006	0,00005	0,00005	0,00005
3,9	0,00005	0,00005	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00003	0,00003

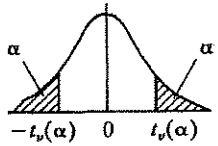
TABLA C. Puntos porcentuales de la distribución t.

Prueba de una cola



Pr $(T_\nu > t_\nu(\alpha)) = \alpha$,
para ν grados de libertad

Prueba de dos colas



Pr $(T_\nu > t_\nu(\alpha)$ o $T_\nu < -t_\nu(\alpha)) = 2\alpha$,
para ν grados de libertad

ν	$\alpha = 0{,}4$ $2\alpha = 0{,}8$	0,25 0,5	0,1 0,2	0,05 0,1	0,025 0,05	0,01 0,02	0,005 0,01	0,0025 0,005	0,001 0,002	0,0005 0,001
1	0,325	1,000	3,078	6,314	12,706	31,821	63,657	127,321	318,309	636,619
2	0,289	0,816	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,599
3	0,277	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,215	12,924
4	0,271	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,265	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,263	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,262	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,261	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,260	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,259	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,259	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,258	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,258	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,257	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,257	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,257	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,257	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,256	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,256	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,256	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,256	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,256	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,256	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,256	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,255	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
60	0,254	0,679	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
120	0,254	0,677	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373
∞	0,253	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

TABLA D. Números aleatorios.

Cada cifra en esta tabla de números aleatorios generados por ordenador es una muestra independiente de una población en la que cada una de las cifras 0 a 9 tiene una probabilidad de ocurrencia de 0,1.

65	23	68	00	77	82	58	14	10	85	11	85	57	11	73	74	45	25	60	46
06	56	76	51	04	73	94	30	16	74	69	59	04	38	83	98	30	20	87	85
55	99	98	60	01	33	06	93	85	13	23	17	25	51	92	04	52	31	38	70
72	82	45	44	09	53	04	83	03	83	98	41	67	41	01	38	66	83	11	99
04	21	28	72	73	25	02	74	35	81	78	49	52	67	61	40	60	50	47	50
87	01	80	59	89	36	41	59	60	27	64	89	47	45	18	21	69	84	76	06
31	62	46	53	84	40	56	31	74	96	52	23	72	95	96	06	58	83	85	22
29	81	57	94	35	91	90	70	94	24	19	35	50	22	23	72	87	34	83	15
39	98	74	22	77	19	12	81	29	42	04	50	62	34	36	81	43	07	97	92
56	14	80	10	76	52	38	54	84	13	99	90	22	55	41	04	72	37	89	33
29	56	62	74	12	67	09	35	89	33	04	28	44	75	01	57	87	45	52	21
93	32	57	38	39	36	87	42	72	55	73	97	98	36	57	41	76	09	11	68
95	69	51	54	43	19	20	49	57	25	90	55	26	20	70	98	43	73	56	45
65	71	32	43	64	67	22	55	65	65	48	86	10	88	20	12	40	18	49	25
20	27	33	43	97	84	20	57	49	91	41	20	17	64	29	60	66	87	55	97
95	29	42	45	61	34	30	13	30	39	21	52	59	28	64	98	08	76	09	27
99	74	06	29	20	55	72	70	11	43	95	82	75	37	90	24	77	43	63	21
87	87	56	91	16	97	51	50	61	36	96	47	76	68	49	11	50	56	51	06
46	24	17	74	97	37	39	03	54	83	34	00	74	61	77	51	43	33	15	67
66	79	81	43	40	92	84	72	88	32	83	24	67	01	41	34	70	19	26	93
36	42	94	58	83	30	92	39	18	40	03	00	12	90	32	27	91	65	48	15
07	66	25	08	99	27	69	48	85	32	16	46	19	31	85	02	86	36	22	96
93	10	05	72	18	26	36	67	68	48	31	69	68	58	93	49	45	86	99	29
49	50	63	99	26	71	47	94	32	71	72	91	34	18	74	06	32	14	40	80
20	75	58	89	39	04	42	73	37	93	11	07	28	77	91	36	60	47	82	62
02	40	62	09	00	71	09	37	80	44	50	37	32	70	20	38	71	86	75	34
59	87	21	38	29	78	72	67	42	83	65	21	54	79	66	42	47	86	31	15
48	08	99	66	43	38	28	13	50	25	47	93	11	15	07	84	28	30	19	07
54	26	86	75	44	15	20	39	20	03	58	54	80	29	62	53	09	67	71	51
35	35	58	45	23	58	63	66	09	62	80	92	14	55	81	41	21	48	87	34
73	84	90	49	01	21	90	29	57	06	68	73	51	10	51	95	61	08	57	99
34	64	78	00	92	59	67	74	58	48	92	09	42	20	40	37	63	80	58	93
68	56	87	47	63	06	24	71	41	98	79	06	07	18	58	29	16	49	67	37
72	47	05	52	88	07	27	55	58	74	82	08	42	28	26	48	25	32	00	31
44	44	96	75	89	57	12	60	42	38	77	36	45	69	21	68	32	70	04	96
28	11	57	47	61	57	89	88	62	18	93	67	57	32	96	72	21	17	13	54
87	22	38	88	91	99	16	08	17	76	27	47	52	14	98	86	35	68	23	85
44	93	14	59	67	40	24	10	11	63	40	47	07	56	14	22	62	74	93	39
81	84	37	25	90	43	58	62	94	58	49	03	84	22	57	22	47	98	86	37
09	75	35	21	04	47	54	08	98	44	08	16	44	86	69	71	20	52	64	94
77	65	05	04	22	18	20	10	81	87	05	69	43	70	96	76	42	05	21	10
19	06	51	61	34	03	61	55	98	58	83	50	01	48	99	85	08	67	15	91
52	91	87	08	19	62	32	28	04	91	42	48	66	24	86	09	87	68	55	51
52	47	25	14	93	91	75	51	49	26	49	41	20	83	30	30	43	22	69	08
52	67	87	40	63	41	91	86	10	47	80	70	56	87	25	86	89	94	21	42
65	25	71	73	78	60	50	62	91	04	95	97	64	16	71	31	32	80	19	61
29	97	56	42	56	90	16	75	74	95	99	26	01	63	25	16	54	18	54	46
15	25	03	68	92	45	53	00	06	29	46	43	46	66	27	12	85	05	22	44
82	08	65	67	64	13	51	14	38	28	24	30	39	62	20	35	23	90	57	36
81	35	03	25	87	24	83	59	04	67	51	52	26	21	69	75	87	28	61	50

TABLA D. Números aleatorios (*continuación*)

67	00	76	07	06	04	17	26	85	10	29	42	93	48	93	46	52	72	77	53
37	41	48	98	99	14	86	78	56	14	20	12	28	86	70	70	66	62	99	86
54	85	60	58	43	58	36	74	44	33	96	38	13	52	98	74	01	27	52	08
82	78	21	26	47	21	31	66	50	67	34	87	78	86	26	32	35	38	94	63
72	32	72	25	83	98	34	31	63	44	31	47	09	57	26	23	89	88	16	10
86	73	37	38	09	68	16	67	81	82	03	42	28	56	09	92	75	20	50	35
54	67	40	72	97	91	06	61	98	95	38	02	94	57	65	32	75	34	64	33
80	86	35	17	08	51	17	12	07	87	75	39	83	43	77	04	66	02	13	46
08	32	44	20	01	13	17	22	42	71	76	76	33	56	94	22	02	67	70	98
96	84	83	43	36	80	18	75	16	54	53	48	71	77	34	88	43	51	41	76
48	67	84	20	48	23	50	47	15	85	24	65	78	93	01	84	02	04	41	31
35	99	47	15	37	62	62	27	35	41	55	57	03	12	74	45	83	25	14	57
13	07	22	58	68	80	91	93	64	68	59	55	19	45	72	83	08	01	28	93
73	15	83	78	75	46	76	36	65	56	34	75	92	58	99	38	51	64	98	42
18	92	29	56	47	99	74	31	42	88	52	71	90	84	23	56	75	22	62	08
50	07	11	21	26	62	94	01	89	32	51	14	17	11	30	31	12	01	18	58
59	50	53	71	99	35	15	56	41	95	71	78	53	15	10	51	86	17	53	81
45	55	85	24	55	08	49	53	00	21	31	67	73	35	42	10	71	12	46	37
90	80	65	04	38	06	30	57	56	62	21	88	30	85	56	89	02	21	43	40
84	51	93	90	28	31	22	31	48	44	45	97	48	85	79	68	78	78	05	18
07	66	01	78	75	25	68	67	31	08	85	38	37	76	01	94	22	20	03	04
19	41	96	21	21	48	53	68	46	91	11	40	98	12	50	26	58	52	74	39
01	38	53	01	20	30	43	53	83	34	87	15	63	52	17	89	43	19	31	11
12	95	21	94	99	72	76	51	69	20	66	93	80	83	88	97	35	52	23	76
25	88	63	69	99	41	89	27	18	92	52	49	56	75	99	20	68	13	04	50
95	89	07	45	38	96	63	61	11	49	98	72	50	67	30	94	93	01	20	20
49	69	36	31	40	43	65	22	63	59	43	94	43	18	76	48	00	90	10	65
47	52	59	03	71	19	04	67	42	38	98	78	36	75	12	62	10	27	23	83
41	89	34	25	98	99	14	49	65	61	20	09	71	32	63	20	88	92	25	40
41	89	18	07	02	57	18	44	53	64	89	51	56	63	63	37	25	64	17	23
46	58	12	07	61	94	29	39	90	76	24	23	64	84	38	61	35	84	78	95
98	42	17	61	53	32	62	34	19	38	05	03	07	09	45	01	61	01	81	34
09	44	61	42	84	40	80	09	25	36	73	61	09	53	51	95	76	09	13	64
41	97	74	05	94	04	57	50	28	49	26	54	91	50	26	20	75	12	91	39
70	42	82	33	21	08	41	30	67	58	46	55	84	19	40	76	47	37	85	59
05	18	96	66	53	07	84	44	17	62	70	43	76	28	64	80	98	32	21	11
69	44	33	07	09	02	87	76	98	50	65	99	36	27	77	23	93	92	15	72
71	95	73	70	09	66	69	55	73	19	20	59	12	95	01	99	75	88	31	13
99	59	52	07	54	56	90	44	75	85	84	35	17	08	97	87	56	04	61	52
97	07	78	13	46	90	10	48	53	29	43	92	58	51	39	39	18	38	47	35
85	04	86	52	92	49	65	46	99	78	99	66	82	34	22	86	79	10	85	86
11	68	36	63	15	84	92	56	31	78	47	49	14	51	34	78	76	47	87	47
12	69	35	64	97	00	63	69	41	06	75	10	94	21	70	74	06	08	90	56
62	72	73	45	26	19	35	75	15	23	75	26	98	66	97	45	31	86	44	80
78	63	02	76	61	95	57	00	30	05	18	52	19	86	40	08	83	32	17	42
65	40	31	04	87	02	46	38	43	16	63	83	76	95	23	06	76	48	54	60
42	68	22	96	29	30	39	32	75	36	64	03	70	64	83	51	61	81	15	96
40	15	54	28	80	30	30	07	53	91	62	62	26	31	75	25	10	23	43	84
51	19	95	91	95	98	92	53	98	08	55	70	68	78	21	13	95	15	87	36
77	55	25	60	17	30	53	23	98	29	52	71	92	10	71	72	52	21	06	21

ANEXO 2. PROGRAMAS INFORMÁTICOS PARA ANÁLISIS ESTADÍSTICOS

Este anexo proporciona alguna información acerca de los programas informáticos más ampliamente difundidos para llevar a cabo análisis estadísticos; más detalles pueden encontrarse en el libro de Richards y Ryan (1985, capítulo 7).* No se ha pretendido dar cuenta de todas las versiones existentes para microordenador, aunque algunos de los programas enumerados tienen una versión en esa plataforma. Igualmente, y a excepción del análisis de conglomerados, no se ha hecho alusión a programas de uso específico: con la excepción de CLUSTAN, todos los programas mencionados son de uso general dentro del análisis estadístico.

En líneas generales, suelen ser sencillos de aprender y de usar, si bien no está de menos recordar que en un primer momento es necesario dedicarles tiempo y esfuerzo, especialmente en el caso de los principiantes. Aunque algunos de ellos cuentan con manuales de introducción, el grueso de la documentación de la mayoría no está tan bien explicada como debería.

La ejecución de análisis particulares está controlada por palabras o mandatos clave, que solicitan los procedimientos adecuados. Esos procedimientos tienen opciones estándar para los análisis habituales, por lo que son muy fáciles de usar; las opciones no estándar son accesibles, no obstante, proporcionando información adicional, lo cual permite un considerable nivel de flexibilidad.

El mayor riesgo que tiene el usario de esos programas es que no dicen si las técnicas se usan de forma correcta o no. El que el programa ejecute una instrucción no significa necesariamente que esa instrucción sea estadísticamente apropiada: la facilidad de empleo de los programas puede llegar a superar los conocimientos estadísticos imprescindibles. Es fundamental tenerlo presente.

* Esa información está dirigida al lector inglés. Con todo, los investigadores que precisen de tales programas en España no tendrán ningún problema: basta con dirigirse al centro de cálculo de su universidad para conseguir el acceso a los programas aquí reseñados, así como cursillos de formación. Aquellos que deseen adquirir y trabajar con las versiones para microordenador tampoco tendrán problema: tanto SPSS, como BMDP y SAS existen en la plataforma PC-compatible e incluso en APPLE Macintosh. A estos programas cabría añadirles otros potentes, flexibles y, sobre todo, algo más fáciles de usar, como SYSTAT o STATGRAPHICS. Puede conseguirse más información en cualquier buen concesionario de microinformática. (*N. del t.*)

MINITAB (Minitab Inc., 215 Pond Laboratory, University Park, PA 16802, EEUU)

El paquete de programas MINITAB proporciona un conjunto de procedimientos muy completo para llevar a cabo analisis estadísticos hasta el nivel del análisis de regresión múltiple. Incluye muchas de las técnicas que no se han expuesto en este libro; por ejemplo, la comparación de medias y el análisis de varianza. Incluye también una serie de procedimientos muy útiles para el tratamiento de matrices, entre ellos la obtención de los vectores y los valores propios.

Quizás sea la facilidad de empleo su característica principal, sobre todo en un nivel introductorio. Es interactivo (los resultados de las instrucciones vuelven a aparecer en la pantalla para que el usuario pueda examinarlos), fácil y flexible, con excelentes capacidades de ayuda en pantalla que pueden solicitarse siempre que el usuario se haya perdido o haya olvidado algo. Además, el *Manual MINITAB* (Ryan *et al.*, 1985) es sencillo e incluye los detalles de los métodos estadísticos y de los procedimientos de MINITAB.

Existe una versión para microordenador.

SPSS-X (SPSS Inc., Suite 3300, 444 North Michigan Avenue, Chicago, Illinois 60611, EEUU)

Es el sucesor de SPSS, un paquete de programas para análisis estadísticos de larga historia y de uso muy difundido, que incorpora tratamiento extensivo de los ficheros, modificación de datos y procedimientos para la preparación e impresión de los informes, además de los análisis estadísticos. Cubre prácticamente todos los métodos descritos en este libro, incluyendo distribuciones de frecuencias y estadísticas descriptivas, clasificaciones cruzadas, regresión y correlación, así como análisis factorial, modelos logarítmicos lineales, y otras muchas técnicas no tratadas aquí.

La documentación es muy completa, con una versión simplificada para principiantes e incorpora la descripción de las técnicas estadísticas y de los procedimientos de SPSS-X.

Existe una versión para microordenador.

SAS (SAS Institute Inc., Box 8000, Cary, North Carolina 27511-8000, EEUU)

El sistema SAS es un sistema completo para análisis de datos. Proporciona procedimientos para guardar la información y acceder a ella, modificación de datos, tratamiento de ficheros y escritura de informes, junto a los análisis estadísticos. Al lado de las estadísticas descriptivas incluye métodos de prueba de significación, regresión, modelos logarítmico-lineales, análisis de conglomerados y análisis multivariante, así como técnicas que no se han expuesto en este libro. Dispone de documentación completa.

En Gran Bretaña no tiene la popularidad de otros paquetes de programas, pero en otros países, especialmente en los Estados Unidos, su uso está muy difundido.

BMDP (BMDP Statistical Software, 1964 Westwood Blvd, Suite 202, Los Ángeles, CA 90025, EEUU)

Otro paquete de programas de uso general, muy parecido a SPSS-X y SAS, aunque

no incluye la escritura de informes no estadísticos que sí aparecen en los otros dos, ni dispone de documentación a nivel de introducción. Tan sólo hay un enorme volumen, demasiado imponente a primera vista. En realidad no es tan malo como parece. Lo que se dice en el manual (Dixon, 1983, p. 15) no está muy lejos de la verdad: los programas en BMDP son fáciles de usar si se ignora lo que no se necesita. Para un principiante, sin embargo, lo que no se ha de saber no es necesariamente evidente.

Son muy numerosas las técnicas incluidas, especialmente las más complejas.

GENSTAT (The Statistical Package Co-ordinator, Numerical Algorithms Group Limited, NAG Central Office, Mayfield House, 256 Banbury Road, Oxford OX27DE, Inglaterra)

GENSTAT es otro paquete de programas de uso general, con la diferencia de que se parece más a un lenguaje de ordenador de alto nivel, como FORTRAN, que a uno de los paquetes anteriores. Esto proporciona mucha flexibilidad, pero complica extraordinariamente su forma de uso. Probablemente sea cierto que es más difícil cometer errores con GENSTAT que con otros paquetes, pues requiere muchos más conocimientos y experiencia. ¡En algunos círculos ha adquirido la reputación del paquete de programas de los expertos! Cubre un amplísimo número de técnicas, entre ellas el análisis de correspondencias.

El manual de GENSTAT no es especialmente fácil de leer; afortunadamente hace algunos años apareció una introducción (Alvey *et al.*, 1982), que permite familiarizarse más con él.

GENSTAT no es muy popular en Estados Unidos, al menos entre los arqueólogos, si hacemos caso a la falta de referencias en la bibliografía.

De la misma estructura y, por tanto, bastante similar, es el paquete de programas de programación lineal GLIM (Baker y Nelder, 1978).

CLUSTAN (Dr. D. Wishart, c/o Dept. of Computational Science, University of St. Andrews, North Haugh, St Andrews KY169SX, Escocia)

Se trata de un conjunto de programas extraordinariamente completo para análisis de conglomerados. Además de los procedimientois descritos en este libro y muchos otros, CLUSTAN dispone de procedimientos para crear matrices de similaridad y distancia usando muchos coeficientes distintos. Actualmente también incluye algunos métodos de validación de grupos, como los descritos en el capítulo 12.

Un punto débil es la falta de procedimientos para tratamiento y modificación de datos, si bien proporciona la posibilidad de unión con BMDP y con SPSS, en particular, que contribuyen a solucionarlo. Un programa asociado llamado CLUSCOM permite la entrada interactiva de las instrucciones de CLUSTAN, pero no siempre es accesible; sin eso, el uso del programa puede llegar a ser laborioso e inducir a error.

Los principiantes tienden a encontrar el manual un tanto difícil de seguir.

Otros programas

IAP: F. R. Hodson y P. Tyers, Dept. of Prehistory, Institute of Archaeology, 31-34 Gordon Sq., Londres WC1H 0PY.

THE BONN SERIATION AND ARCHAEOLOGICAL STATISTICS PACKAGE: The Unkelbach Valley Software Works (dos direcciones):
— 620 Oriole Lane, Mt. Prospect, IL 60056, Estados Unidos.
— In der Au 9, D5480 Remagen 4, Alemania.

MV-ARCH: The Secretary, MV-ARCH, Dept. of Anthropology, University of Sydney, NSW 2006, Australia.

CANOCO: Microcomputer Power, 111 Clover Lane, Ithaca, N.Y. 14850, Estados Unidos.

BIBLIOGRAFÍA

Aldenderfer, M. (1981), «Creating assemblages by computer simulation: the development and uses of ABSIM», en J. A. Sabloff, ed., *Simulations in Archaeology*, University of New Mexico Press, Albuquerque, pp. 67-117.

— (1982), «Methods of cluster validation for archaeology», *World Archaeology*, 14, pp. 61-72.

Alvey, N., N. Galwey y P. Lane (1982), *An introduction to GENSTAT*, Academic Press, Londres.

Baker, R. J., y J. A. Nelder (1978), *The GLIM System, Release 3: Generalised Linear Interactive Modelling*, Royal Statistical Society, Londres.

Barnett, V. (1974), *Elements of Sampling Theory*, English University Press, Londres.

Barth, F. (1966), *Models of Social Organization*, Royal Anthropological Institute Occasional Paper N.° 23, Royal Anthropological Institute, Londres.

Bellhouse, D. (1980), «Sampling studies in archaeology», *Archaeometry*, 22, pp. 123-132.

—, y W. D. Finlayson (1979), «An empirical study of probability sampling designs: preliminary results from the Draper Site», *Canadian J. Archaeology*, 3, pp. 105-123.

Berry, H. J., P. W. Mielke y K. L. Kvamme (1984), «Efficient permutation procedures for analysis of artefact distributions», en H. Hietala, ed., *Intrasite Spatial Analysis in Archaeology*, Cambridge University Press, Cambridge, pp. 54-74.

Bettinger, R. L. (1979), «Multivariate statistical analysis of a regional subsistence-settlement model for Owens Valley», *American Antiquity*, 44, pp. 455-470.

Binford, L. R. (1964), «A consideration of archaeological research dessign», *American Antiquity*, 29, pp. 425-441.

— (1981), *Bones: Ancient Men and Modern Myths*, Academic Press, Nueva York.

— (1983), *In Pursuit of the Past*, Thames and Hudson, Londres (hay trad. cast.: *En busca del pasado*, Crítica, Barcelona, 1988).

Blalock, H. M. (1972), *Social Statistics*, McGraw Hill-Kogakusha, Tokyo (hay trad. cast.: *Estadística Social*, Fondo de Cultura Económica, México).

Bølviken, E., E. Helskog, K. Helskog, I. M. Holm-Olsen, L. Solheim y R. Bertelsen (1982), «Correspondence Analysis: an alternative to principal components», *World Archaeology*, 14, pp. 41-60.

Bow, Sing-Tze (1984), *Pattern Recognition*, Marcel Dekker, Nueva York y Basilea.

Boyle, K. (1983), «The Hunters Nobody Knows», tesis de MSC, inédita, Universidad de Southampton.

Bradley, R., y C. Small (1985), «Looking for circular structures in post-hole distributions: quantitative analysis of two settlements from Bronze Age England», *J. Archaeological Science*, 12, pp. 285-297.

Brainerd, G. W. (1951), «The place of chronological ordering in archaeological analysis», *American Antiquity*, 16, pp. 301-313.
Brumfiel, E. (1976), «Regional growth in the eastern valley of Mexico: a test of the "population pressure" hypothesis», en K. Flannery, ed., *The Early Mesoamerican Village*, Academic Press, Nueva York, pp. 234-249.
Buchvaldek, M., y D. Koutecky (1970), «Vikletice: ein schnurkeramisches Gräberfeld», Universita Karlová, Praga.
Cable, C. (1984), *Economy and Technology in the Late Stone Age of Southern Natal*, Cambridge Monographs in African Archaeology, 9, BAR International Series, 201, British Archaeological Reports, Oxford.
Carothers, J., y A. McDonald (1979), «Size and distribution of the population in Late Bronze Age Mesenia: some statistical approaches», *J. Field Archaeology*, 6, pp. 433-454.
Clark, G. A. (1976), «More on contingency table analysis, decision making criteria and the use of log-linear models», *American Antiquity*, 41, pp. 259-273.
— (1982), «Quantifying archaeological research», en M. Schiffer, ed., *Advances in Archaeological Method and Theory*, vol. 5, Academic Press, Nueva York, pp. 217-273.
Clarke, D. L. (1962), «Matrix Analysis and archaeology with particular reference to British Beaker pottery», *Proc. Prehistoric Society*, 28, pp. 371-382.
— (1966), «A tentative reclassification of British Beaker pottery in the light of recent research», *Palaeohistoria*, 12, pp. 179-198.
— (1968), *Analytical Archaeology*, Methuen, Londres (hay trad. cast.: *Arqueología analítica*, Bellaterra, Barcelona).
— (1970), *Beaker pottery of Great Britain and Ireland*, Cambridge University Press, Cambridge.
Cochran, W. G. (1977), *Sampling Techniques*, John Wiley and Sons, Nueva York, 3.ª ed.
Constantine, A. G., y J. C. Gower (1978), «Graphical representation of assymmetric matrices», *Applied Statistics*, 27, pp. 297-304.
Cormack, R. M. (1971), «A review of classification», *J. Royal Statistical Society A*, 134, pp. 321-367.
Cowgill, G. L. (1970), «Some sampling and reliability problems in archaeology», en J. C. Gardin, ed., *Archéologie et Calculateurs*, CNRS, París, pp. 161-172.
— (1972), «Models, methods and techniques for seriation», en D. L. Clarke, ed., *Models in Archaeology*, Methuen, Londres, pp. 381-424.
— (1977), «The trouble with significance tests and what we can do about it», *American Antiquity*, 42, pp. 350-368.
Cuadras, C. M. (1981), *Métodos estadísticos multivariantes*, Editorial Universitaria, Barcelona.
Chatterjee, S., y B. Price (1977), *Regression Analysis by Example*, John Wiley and Sons, Nueva York.
Cherry, J. F. (1977), «Investigating the political geography of an early state by multidimensional scaling of Linear B tablet data», en J. L. Bintliff, ed., *Mycenaean Geography*, British Association of Mycenaean Studies, Cambridge, pp. 76-83.
Davis, J. C. (1973), *Statistics and Data Analysis in Geology*, John Wiley and Sons, Nueva York.
Dixon, C., y B. Leach (1977), *Sampling Methods for Geographical Research*, Concepts and Techniques in Modern Geography, 17, University of East Anglia, Geo. Abstracts, Norwich.

Dixon, W. J., ed. (1983), *BMDP Statistical Software*, University of California Press, Berkeley.
Djindjian, F. (1980), *Construction des Systèmes d'Aide à la Connaissance en Archéologie Préhistorique. Structuration et Affectation*, tesis doctoral de 3.er ciclo, Université Paris 1, UER d'Art et d'Archéologie.
Doran, J., y F. Hodson (1975), *Mathematics and Computers in Archaeology*, Edinburgh University Press, Edimburgo.
Ester, M. (1981), «A column-wise approach to seriation», *American Antiquity*, 46, pp. 496-512.
Everitt, B. (1980), *Cluster Analysis*, Heinemann Educational Books, Londres, 2.ª ed.
—, y G. Dunn (1983), *Advanced Methods of Data Exploration and Modelling*, Heinemann Educational Books, Londres.
Fieller, N. R. J., y A. Turner (1982), «Number estimation in vertebrate samples», *J. Archaeological Science*, 9, pp. 49-62.
Fienberg, S. E. (1980), *The Analysis of Cross-Classified Categorical Data*, MIT Press, Cambridge, Massachussets.
Forsberg, L. (1985), *An analysis of the Hunter-Gatherer Settlement Systems in the Lule River Valley, 1500-BC/AD*, Dept. Archaelogy, University of Umea, Umea (Suecia).
Gabriel, K. R. (1981), «Biplot display of multivariate matrices for inspection of data and diagnosis», en V. Barnett, ed., *Interpreting Multivariate Data*, John Wiley and Sons, Londres, pp. 147-173.
Gaines, S., ed. (1981), *Databank Applications in Archaeology*, University of Arizona Press, Tucson.
Gamble, C. S. (1982), «Leadership and "surplus" production», en C. Renfrew y S. J. Shennan, eds., *Ranking, Resource and Exchange*, Cambridge University Press, Cambridge, pp. 100-105.
Gardin, J. C. (1980), *Archaeological Constructs*, Cambridge University Press, Cambridge (existe una versión francesa escrita por el propio autor: *Une Archéologie Théorique*, Hachette, París).
GENSTAT (1983), *A General Statistical Program* (release 4.04), Numerical Algorithms Group Ltd., Oxford.
Gero, J., y J. Mazzullo (1984), «Analysis of artefact shape using Fourier Analysis in closed form», *J. Field Archaeology*, pp. 315-322.
Gordon, A. D. (1981), *Classification: methods for the exploratory analysis of multivariate data*, Chapman and Hall, Londres.
Gower, J. C. (1971), «A general coefficient of similarity and some of its properties», *Biometrics*, 27, pp. 857-872.
— (1977), «The analysis of asymmetry and orthogonality», en J. R. Barra, ed., *Recent Developments in Statistics*, North Holland, Amsterdam, pp. 109-123.
Haggett, P., A. D. Cliff y A. Frey (1977), *Locational Analysis in Human Geography*, Edward Arnold, Londres, 2.ª ed.
Hartwig, F., y B. E. Dearing (1979), *Exploratory Data Analysis*, Sage University paper series on Quantitative Applications in the Social Sciences, series n.º 07-016, Sage Publications, Beverly Hills y Londres.
Hawkes, J. (1968), «The proper study of mankind», *Antiquity*, 42, pp. 255-262.
Hietala, H. (1984), «Variations on a categorical data theme: local and global considerations with Near Eastern paleolithic applications», en H. Hietala, ed., *Intrasite Spatial Analysis in Archaeology*, Cambridge University Press, Cambridge, pp. 44-53.

Hill, M. O. (1973), «Reciprocal averaging: an eigenvector method for ordination», *J. Ecology*, 61, pp. 237-249.
Hodder, I. (1978), *Simulation studies in Archaeology*, Cambridge University Press, Cambridge.
— (1982), «Theoretical archaeology: a reactionary view», en I. Hodder, ed., *Structural and Symbolic Archaeology*, Cambridge University Press, Cambridge, pp. 1-16.
—, y C. Orton (1976), *Spatial Analysis in Archaeology*, Cambridge University Press, Cambridge (hay trad. cast.: *Análisis espacial en arqueología*, Crítica, Barcelona, 1990).
Hodson, F. R. (1977), «Quantifying Hallstatt: some initial results», *American Antiquity*, 42, pp. 394-412.
Hole, B. L. (1980), «Sampling in archaeology: a critique», *Annual Review of Anthropology*, 9, pp. 217-234.
Holsler, D., J. Sabloff y D. Runge (1977), «Situation model development: a case study of the Classic Maya colapse», en N. Hammond, ed., *Social Processes in Maya Prehistory*, Academic Press, Nueva York, pp. 553-590.
Huggett, J. (1985), «Expert systems in archaeology», en M. A. Cooper y J. D. Richards, eds., *Current Issues in Archaeological Computing*, BAR International Series, 271, British Archaeological Reports, Oxford, pp. 123-142.
Ihm, P. (1978), *Statistik der Archäeologie*, Archaeo-Physika, 9, Rheinland, Colonia.
Jardine, N., y R. Sibson (1971), *Mathematical Taxonomy*, John Wiley and Sons, Londres.
Johnson, G. A. (1973), *Local Exchange and Early State Development in Southwestern Iran*, University of Michigan Museum of Anthropology, Anthropological Papers n.º 51, University of Michigan, Ann Arbor.
Johnston, R. J. (1978), *Multivariate Statistical Analysis in Geography*, Longman, Londres.
Kampffmeyer, U., y W. R. Teegen (1986), «Untersuchungen zur rechnergestützten Klassifikation von Gefäßformen am Beispiel der eisenzeitlichen Keramik des Gräberfeldes von Veis, Quattro Fontanili», *Die Kunde*, 37, pp. 1-84.
Kemp, B. (1982), «Automatic analysis of Predynastic cemeteries: a new method for an old problem», *J. Egyptian Archaeology*, 68, pp. 5-15.
Kendall, D. (1977), «Computer techniques and the archival map-reconstruction of Mycenaean Messenia», en J. L. Bintliff, ed., *Mycenaean Geography*, British Association of Mycenaean Studies, Cambridge, pp. 83-87.
Kintigh, K. W., y A. J. Ammerman (1982), «Heuristic approaches to spatial analysis in archaeology», *American Antiquity*, 47, pp. 31-63.
Kruskal, J. B., y M. Wish (1978), *Multidimensional Scaling*, Sage University Paper series on Quantitative Applications in the Social Sciences, series n.º 07-011, Sage Publications, Beverly Hills y Londres.
Leese, M. N., y S. P. Needham (1986), «Frequency table analysis: exemples from early Bronze Age axe decoration», *J. Archaeological Science*, 13, pp. 1-12.
Levy, J. (1982), *Social and Religious Organisation in Bronze Age Denmark*, BAR International Series, 124, British Archaeological Reports, Oxford.
Lewis, R. B. (1986), «The analysis of contingency tables in archaeology», en M. Schiffer, ed., *Advances in Archaeological Method and Theory*, vol. 9, Academic Press, Nueva York, pp. 277-310.
Lovis, W. A. (1976), «Quarter sections and forests: an example of probability sampling in the northeastern woodlands», *American Antiquity*, 43, pp. 364-372.
[Lull, V. (1983), *La «cultura» de El Argar*, Akal, Madrid.]
MacIntosh, S. K., y R. J. MacIntosh (1980), *Prehistoric Investigations at Jenne, Mali*, BAR International Series, 89, British Archaeological Reports, Oxford.

Marquardt, W. (1978), «Advances in archaeological seriation», en M. Schiffer, ed., *Advances in Archaeological Method and Theory*, vol. 1, Academic Press, Nueva York, pp. 257-314.
Mather, P. M. (1976), *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley and Sons, Londres.
Mathiesen, P., I. M. Holm-Olsen, T. Sobstad y H. D. Bratein (1981), «The Helgøy Project: an interdisciplinary study of past eco-ethno processes in the Helgøy region, northern Troms, Norway», *Norwegian Archaeological Review*, 14, pp. 77-117.
McDonald, J., y G. D. Snooks (1985), «The determinants of manorial income in Domesday England: evidence from Essex», *J. Economic History*, 45, pp. 541-546.
McManamon, F. P. (1981), «Prehistoric land use on outer Cape Cod», *J. Field Archaeology*, 9, pp. 1-20.
Mellars, P. A., y M. R. Wilkinson (1980), con un apéndice de R. J. Fieller, «Fish otoliths as indicators of seasonality in prehistoric shell middens: the evidence from Oronsay (Inner Hebrides)», *Proc. Prehistoric Society*, 46, pp. 19-44.
Mojena, R. (1977), «Hierarchical grouping methods and stopping rules: an evaluation», *Computer Journal*, 20, pp. 359-363.
Morrison, D. F. (1967), *Multivariate Statistical Methods*, McGraw-Hill, Nueva York.
Morwood, M. J. (1980), «Time, space and prehistoric art: a principal components analysis», *Archaeology and Physical Anthropology in Oceania*, 15, pp. 98-109.
Mosteller, F., y J. W. Tukey (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, Massachussets.
Nance, J. (1981), «Statistical fact and archaeological faith: two models in small-sites sampling», *J. Field Archaeology*, 8, pp. 151-165.
— (1983), «Regional sampling in archaeological survey: the statistical perspective», en M. Schiffer, ed., *Advances in Archaeological Method and Theory*, vol. 6, Academic Press, Nueva York, pp. 289-356.
—, y B. F. Ball (1986), «No surprises? The reliability and validity of test pit sampling», *American Antiquity*, 51, pp. 457-483.
[Nocete, F. (1989), *El espacio de la coerción. La transición al estado en las campiñas del Alto Guadalquivir (España). 3000-1500 a.C.*, BAR International Series, 492, British Archaeological Reports, Oxford.]
Norusis, M. J. (1983), *SPSS-X: Introductory Statistics Guide*, McGraw-Hill, Nueva York.
— (1985), *SPSS-X: Advanced Statistics Guide*, McGraw-Hill, Nueva York.
Orton, C. (1980), *Mathematics in Archaeology*, Collins, Londres (hay trad. cast.: *Matemáticas para arqueólogos*, Alianza, Madrid, 1987).
— (1982), «Computer simulation experiments to assess the performance of measures of quantity of pottery», *World Archaeology*, 14, pp. 1-20.
Ottaway, B. (1973), «Dispersion diagrams: a new approach to the display of carbon-14 dates», *Archaeometry*, 15, pp. 5-12.
Peacock, D. P. S. (1971), «Petrography of certain coarse pottery», en B. W. Cunliffe, ed., *Excavations at Fishbourne*, 2, Report of the Research Committee of the Society of Antiquaries 27, Society of Antiquaries, Londres, pp. 255-259.
Peebles, C. S. (1972), «Monothetic-divisive analysis of the Moundville burials —an initial report», *Newsletter of Computer Archaeology*, 8, pp. 1-13.
Petrie, W. M. F. (1901), *Diospolis Parva*, Londres.
Plog, F. T. (1974), «Settlement patterns and social history», en M. Leaf, ed., *Frontiers of Anthropology*, Van Nostrand, Nueva York, pp. 68-92.

Plog, S. (1976), «Relative efficiencies of sampling techniques for archaeological surveys», en K. V. Flannery, ed., *The Early Mesoamerican Village*, Academic Press, Nueva York, pp. 136-158.
— (1980), *Stylistic Variation in Prehistoric Ceramics*, Cambridge University Press, Cambridge.
Read, D. W. (1982), «Towards a theory of archaeological classification», en R. Whallon y J. A. Brown, eds., *Essays in Archaeological Typology*, Center for American Archaeology Press, Evanston, Illinois, pp. 56-92.
Redman, C. (1974), *Archaeological sampling Strategies*, Addison-Wesley Modular Publications in Anthropology, n.º 55, Addison-Wesley, Reading, Massachusetts.
Renfrew, C. (1977), Alternative models for exchange and spatial distribution», en T. Earle y J. E. Ericson, eds., *Exchange Systems in Prehistory*, Academic Press, Nueva York, pp. 71-90.
—, y K. L. Cooke, eds. (1979), *Transformations: Mathematical Approaches to Culture Change*, Academic Press, Nueva York.
—, y G. Sterud (1969), «Close-proximity analysis: a rapid method for the ordering of archaeological materials», *American Antiquity*, 34, pp. 265-277.
Richards, J. D., y N. S. Ryan (1985), *Data Processing in Archaeology*, Cambridge University Press, Cambridge.
Robinson, W. S. (1951), «A method for chronologically ordering archaeological deposits», *American Antiquity*, 16, pp. 293-301.
Rogge, A. E., y S. L. Fuller (1977), «Probability survey sampling: making parameter estimates», en M. B. Schiffer y G. J. Gumerman, eds., *Conservation Archaeology*, Academic Press, Nueva York, pp. 227-238.
Ryan, B. F., B. L. Joiner y T. A. Ryan (1985), *The MINITAB Student Handbook*, Duxbury Press, Boston, 2.ª ed.
Sabloff, J. A., ed. (1981), *Simulations in Archaeology*, University of New Mexico Press, Albuquerque.
[Sánchez Carrión, J. J., ed. (1984), *Introducción a las técnicas de análisis multivariante aplicadas a las ciencias sociales*, Centro de Investigaciones Sociológicas, Madrid.]
Scollar, I. W. (1969), «Some techniques for the evaluation of archaeological magnetometer surveys», *World Archaeology*, 1, pp. 77-89.
Schoknecht, U., ed. (1980), *Typentafeln zur Ur- und Frügeschichte der DDR*, Kulturbund der DDR, Weimar.
Service, E. R. (1962), *Primitive Social Organisation*, Random House, Nueva York.
Shanks, M., y C. Tilley (1982), «Ideology, symbolic power and ritual communication: a reinterpretation of neolithic mortuary practices», en I. Hodder, ed., *Structural and Symbolic Archaeology*, Cambridge University Press, Cambridge, pp. 129-154.
Shennan, S. J. (1977), «Bell Beakers and their context in Central Europe: a New Approach», tesis doctoral inédita, Universidad de Cambridge.
— (1983), «Disentangling data», en G. Howson y R. McLone, eds., *Maths at Work*, Heinemann Educational Books, Londres, pp. 109-126.
— (1985), *Experiments in the Collection and Analysis of Archaeological Survey Data*, Department of Prehistory and Archaeology, University of Sheffield, Sheffield.
—, y J. Wilcock (1975), «Shape and style variation in Central German Bell Beakers: a computer-assisted study», *Science and Archaeology*, 15, pp. 17-31.
Shott, M. (1985), «Shovel-test sampling as a site discovery technique: a case study from Michigan», *J. Field Archaeology*, 12, pp. 457-468.

Sidrys, R. (1977), «Mass-distance measures for the Maya obsidian trade», en T. Earle y J. E. Ericson, eds., *Exchange Systems in Prehistory*, Academic Press, Nueva York, pp. 91-107.
Simek, J. F. (1984), «Integrating pattern and context in spatial archaeology», *J. Archaeological Science*, 11, pp. 405-420.
Sneath, P., y R. Sokal (1973), *Numerical Taxonomy*, Freeman, San Francisco.
Sokal, R., y P. Sneath (1963), *Principles of Numerical Taxonomy*, Freeman, San Francisco.
Späth, H. (1980), *Cluster Analysis Algorithms*, Ellis Horwood, Chichester.
Spaulding, A. C. (1953), «Statistical techniques for the discovery of artefact types», *American Antiquity*, 18, pp. 305-313.
— (1977), «On growth and form in Archaeology: multivariate analysis», *J. Anthropological Research*, 33, pp. 1-15.
Speth, J., y G. Johnson (1976), «Problems in the use of correlation for the investigation of tool kits and activity areas», en C. E. Cleland, ed., *Cultural Change and Continuity: Essays in Honour of James Bennett Griffin*, pp. 33-57.
Tainter, J. A. (1975), «Social inferences and mortuary practices: an experiment in numerical classification», *World Archaeology*, 7, pp. 1-15.
Thomas, D. H. (1975), «Non-site sampling in archaeology: up the creek without a site?», en J. W. Mueller, ed., *Sampling in Archaeology*, University of Arizona Press, Tucson, pp. 61-81.
— (1976), *Figuring Anthropology*, Holt, Rinehart and Winston, Nueva York.
— (1978), «The awful truth about statistics in archaeology», *American Antiquity*, 43, pp. 231-244.
Tilley, C. (1984), «Ideology and the legitimation of power in the Middle Neolithic of southern Sweden», en D. Miller y C. Tilley, eds., *Ideology, Power and Prehistory*, Cambridge University Press, Cambridge, pp. 111-146.
Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut.
Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachussets.
— (1980), «We need both exploratory and confirmatory», *American Statistician*, 34, pp. 23-25.
Turner, A., y N. R. J. Fieller (1985), «Consideration of minimum numbers: a response to Horton», *J. Archaeological Science*, 12, pp. 477-483.
Van der Veen, M. (1985), «Carbonised seeds, sample size and on-site sampling», en N. R. J. Fieller, D. D. Gilbertson y N. G. A. Ralph, eds., *Paleoenvironmental Investigations: Research dessign methods and Data Analysis*, Symposium 5 (ii) of the Association for Environmental Archaeology, BAR International Series, 258, British Archaeological Reports, Oxford, pp. 165-174.
—, y N. Fieller (1982), «Sampling seeds», *J. Archaeological Science*, 9, pp. 287-298.
Vescelius, G. S. (1960), «Archaeological sampling: a problem of statistical inference», en G. E. Dole y R. L. Carneiro, eds., *Essays in the Science of Culture, in Honour of Leslie A. White*, Thomas Y. Crowell, Nueva York, pp. 457-470.
Vierra, R. (1982), «Typology, classification and theory building», en R. Whallon y J. Brown, eds., *Essays in Archaeological Typology*, Centre for American Archaeology Press, Evanston, Illinois, pp. 162-175.
—, y D. L. Carson (1981), «Factor Analysis, random data and patterned results», *American Antiquity*, 46, pp. 272-283.
Wainwright, G. J. (1979), *Mount Pleasant, Dorset: Excavations 1970-71*, Society of Antiquaries Research Report 37, Thames and Hudson, Londres.

Whallon, R. (1982), «Variables and dimensions: the critical step in quantitative typology», en R. Whallon y J. Brown, eds., *Essays in Archaeological Typology*, Centre for American Archaeology Press, Evanston, Illinois, pp. 127-161.
— (1984), «Unconstrained clustering for the analysis of spatial distributions in archaeology», en H. Hietala, ed., *Intra-site Spatial Analysis in Archaeology*, Cambridge University Press, Cambridge, pp. 242-247.
—, y J. Brown, eds. (1982), *Essays in Archaeological Typology*, Centre for American Archaeology Press, Evanston, Illinois.
White, J., A. Yeats y G. Skipworth (1979), *Tables for Statisticians*, Stanley Thornes (Publishers) Ltd.
Wilson, A. G., y M. J. Kirkby (1980), *Mathematics for Geographers and Planners*, Clarendon Press, Oxford, 2.ª ed.
Winham, P. (1978), «Sampling populations of sites: a case-study from Shetland», en J. F. Cherry, C. S. Gamble y S. J. Shennan, eds., *Sampling in Contemporary British Archaeology*, BAR British Series, 50, British Archaeological Reports, Oxford, pp. 105-120.
Winter, M. (1976), «Excavating a shallow community by random sampling quadrats», en K. V. Flannery, ed., *The Early Mesoamerican Village*, Academic Press, Nueva York, pp. 62-67.
Wishart, D. (1978), *CLUSTAN User Manual* (3.ª ed. con un suplemento: 1982), Inter-University/Research Council Series Report n.º 47, Program Library Unit, Edimburgo.
Wobst, M. (1983), «We can't see the forest for the trees: sampling and the shapes of archaeological distributions», en J. A. Moore y A. Keene, eds., *Archaeological Hammers and Theories*, Academic Press, Nueva York, pp. 37-85.
Zubrow, E., y J. Harbaugh (1978), «Archaeological prospecting: kriging and simulation», en I. Hooder, ed., *Simulation Studies in Archaeology*, Cambridge University Press, Cambridge, pp. 109-122.

Adiciones

Adams, W. Y. (1988), «Archaeological classification: theory versus practice», *Antiquity*, 62, pp. 40-56.
Allen, K. M. S., S. W. Green y E. B. W. Zubrow (1990), *Interpreting Space: GIS and Archaeology*, Taylor and Francis, Londres.
Bertelsen, R. (1988), «Correspondence analysis as explorative tool», en C. L. N. Ruggles y S. P. Q. Rahtz, eds., *Computer and Quantitative Methods in Archaeology 1987*, BAR International Series, 393, British Archaeological Reports, Oxford, pp. 25-28.
Buck, C. E., y C. D. Litton (1991), «A Bayes approach to some archaeological problems», en K. Lockyear y S. P. Q. Rahtz, eds., *Computer and Quantitative Methods in Archaeology 1990*, Tempus Reparatum, BAR International Series, 565, British Archaeological Reports, Oxford, pp. 93-100.
Cowgill, G. L. (1990), «Artifact classification and archaeological purposes», en A. Voorrips, ed., *Mathematics and Information Sciences in Archaeology: A Flexible Framework*, Holos Verlag, Studies in Modern Archaeology, 3, Bonn, pp. 61-78.
Chippindale, C., ed. (1986), *Form and Design in Archaeology: A Grammatical Approach*, Papers for Design Grammar Symposium, Annual Meeting of The Society for American Antiquity, mayo de 1986, Nueva Orleans.

Djindjian, F. (1989), «Fifteen years of contribution of the French school of data analysis», en S. P. Q. Rahtz y J. Richards, eds., *Computer and Quantitative Methods in Archaeology 1989*, BAR International Series, 548, British Archaeological Reports, Oxford, pp. 193-204.
Gob, A. (1988), «Multivariate analysis of lithic industries», en C. L. N. Ruggles y S. P. Q. Rahtz, eds., *Computer and Quantitative Methods in Archaeology 1987*, BAR International Series, 393, British Archaeological Reports, Oxford, pp. 15-24.
Litton, C. D., y M. N. Leese (1991), «Some statistical problems arising in radiocarbon calibration», en K. Lockyear y S. P. Q. Rahtz, eds., *Computer Applications and Quantitative Methods in Archaeology 1990*, Tempus Reparatum, BAR International Series, 565, British Archaeological Reports, Oxford, pp. 101-110.
Madsen, T., ed. (1988), *Multivariate Archaeological, Numerical Approaches to Scandinavian Archaeology*, Aarhus University Press, Jutland Archaeological Society Publications, 21, Aarhus.
— (1989), «Seriation and Multivariate Statistics», en S. P. Q. Rahtz y J. Richards, eds., *Computer and Quantitative Methods in Archaeology 1989*, BAR International Series, 548, British Archaeological Reports, Oxford, pp. 205-214.
Orton, C., y P. Tyers (1989), «Error structures of ceramic assemblages», en S. P. Q. Rahtz y J. Richards, eds., *Computer and Quantitative Methods in Archaeology 1989*, BAR International Series, 548, British Archaeological Reports, Oxford, pp. 275-286.
— (1991), «A technique for reducing the size of sparse contingency tables», en K. Lockyear y S. P. Q. Rahtz, eds., *Computer and Quantitative Methods in Archaeology 1990*, Tempus Reparatum, BAR International Series, 565, British Archaeological Reports, Oxford, pp. 121-126.
Ringrose, T. (1988), «Correspondence analysis for stratigraphic abundance data», en C. L. N. Ruggles y S. P. Q. Rahtz, eds., *Computer and Quantitative Methods in Archaeology 1987*, BAR International Series, 393, British Archaeological Reports, Oxford, pp. 3-14.
Rosch, E., y B. Lloyd, eds. (1978), *Cognition and Categorization*, Erlbaum, Hilldale, N.J.
Sharp, N. D. (1988), «Style and substance: a reconsideration of the Lapita decorative system», en P. V. Kirch y J. L. Hunt, eds., *Archaeology of the Lapita Cultural Complex: A Critical Review*, Washington State Museum, Seattle, pp. 61-81.
Tyers, P., y C. Orton (1991), «Statistical analysis of ceramic assemblages», en K. Lockyear y S. P. Q. Rahtz, eds., *Computer and Quantitative Methods in Archaeology 1990*, Tempus Reparatum, BAR International Series, 565, British Archaeological Reports, Oxford, pp. 117-120.
Voorrips, A., ed. (1990), *Mathematics and Information Sciences in Archaeology: A Flexible Framework*, Holos Verlag, Studies in Modern Archaeology, 3, Bonn.
Washburn, D. K., y D. Crowe (1988), *Symmetries of Culture: Theory and Practice of Plane Pattern Analysis*, University of Washington Press, Seattle.

ÍNDICE ALFABÉTICO

ÍNDICE