Capítulo 6

ASOCIACIÓN ENTRE DOS VARIABLES

6.1. INTRODUCCIÓN

Generalmente un problema estadístico envolucra más de una variables. En una encuesta de opinión para un estudio de mercado o política se hacen varias preguntas cuyas respuestas son interesantes de relacionar. Por ejemplo, se interroga a los votantes no solamente sobre su candidato preferido, sino que también su edad, género, profesión, etc... El análisis de tal encuesta permitirá eventualmente determinar el perfil del electorado de un candidato, lo que orientaá su campaña electoral. El psicólogo querrá comparar las aptitudes mentales (CI) y el rendimiento escalar de un grupo de estudiantes. Estos problemas llaman a medir y describir relaciones entre variables.

Una asociación entre variables expresa el grado de influencia que puede tener una variable sobre otra. Los índices que se pueden definir dependen del tipo de relación que se estudia y de la naturaleza de las variables consideradas. Se presentan en primer lugar índices descriptivos de asociación y en seguida se hacen inferencia sobre estos coeficientes.

6.2. EL COEFICIENTE DE CORRELACIÓN

Si se consideran dos variables X e Y cuantitativas y mediciones sobre un conjunto de individuos $\mathcal{P}$, con valores en $\mathbb{R}$ ó un intervalo de $\mathbb{R}$, una simple representación gráfica en $\mathbb{R}^2$ con un gráfico de dispersión permitirá detectar la existencia y la forma de una eventual relación entre las dos variables. El coeficiente de correlación lineal es el índice de asociación más usual entre dos variables numéricas. Una inspección del gráfico de dispersión es necesaria para asegurarse de que la interpretación es correcta.

Sea $\{(x_i, y_i)|i = 1, ..., n\}$ una muestra aleatoria bivariada del par (X, Y) de variables. Se denotan $\bar{x} = \frac{1}{n}\sum x_i$ y $\bar{y} = \frac{1}{n}\sum y_i$ a las medias empíricas respectivas de x=$(x_1, x_2, ...x_n)$ e y=$(y_1, y_2, ...y_n)$, y $s_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ y $s_y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$ a las varianzas empíricas respectivas de x e y.

Definición 6.2.1 *Se llama covarianza empírica entre X e Y a:*

$$cov(x,y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$$

Como la covarianza es sensible a los cambios de escala de las dos variables, se elimina este efecto con el coeficiente de correlación lineal, que toma en cuenta de las varianzas s_x^2 de los valores x y s_y^2 de y.

Definición 6.2.2 *Se llama correlación lineal entre x e y a la cantidad:*

$$r_{x,y} = \frac{cov(x,y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Este coeficiente, que toma como valores extremos $+1$ y -1, mide el grado de relación de tipo lineal que existe entre x e y.

$r_{x,y} = -1$	relación estrictamente lineal de pendiente negativa
$-1 < r_{x,y} < 0$	tendencia lineal negativa
$r_{x,y} = 0$	ausencia de tendencia lineal
$0 < r_{x,y} < +1$	tendencia lineal positiva
$r_{x,y} = +1$	relación estrictamente lineal de pendiente positiva

La tendencia lineal aumenta cuando $r_{x,y}$ tiende a ± 1 (ver gráficos 6.1). Pero cuando $r_{x,y} \neq \pm 1$, hay muchos casos diferentes que pueden producir el mismo valor del coeficiente $r_{x,y}$. De aquí la importancia de tener cuidado en la interpretación de un coeficiente de correlación por que un dato atípico ó aberrante, una mezcla de poblaciones, una relación no lineal pueden cambiar totalmente el valor del coeficiente (ver gráficos 6.2).

Cuando se estudia en conjunto más de dos variables, se presentan los coeficientes de correlación relativos en una matriz cuyo termino general r_{ij} es el coeficiente de correlación lineal de las variables i y j. La matriz de correlación asociada a los datos de 6 variables recolectados sobre 20 países de América Latina (tabla 6.1), se presentan en la tabla 6.2. Acá, el coeficiente de correlación entre la tasa de natalidad y la fecundidad es igual a 0,972.

Si se quiere estudiar otro tipo de relación, se tiene dos alternativas:

- Dada una función f de X, calcular el coeficiente de correlación entre $f(X)$ e Y. Este método es factible cuando se conoce la función f.
- Usar otros índices, como veremos más adelante.

6.3. LA RAZÓN DE CORRELACIÓN

Cuando una de las dos variables es nominal u ordinal, no se puede calcular el coeficiente de correlación lineal. Si Y es la variable cuantitativa, por ejemplo el PNB de todos países y X

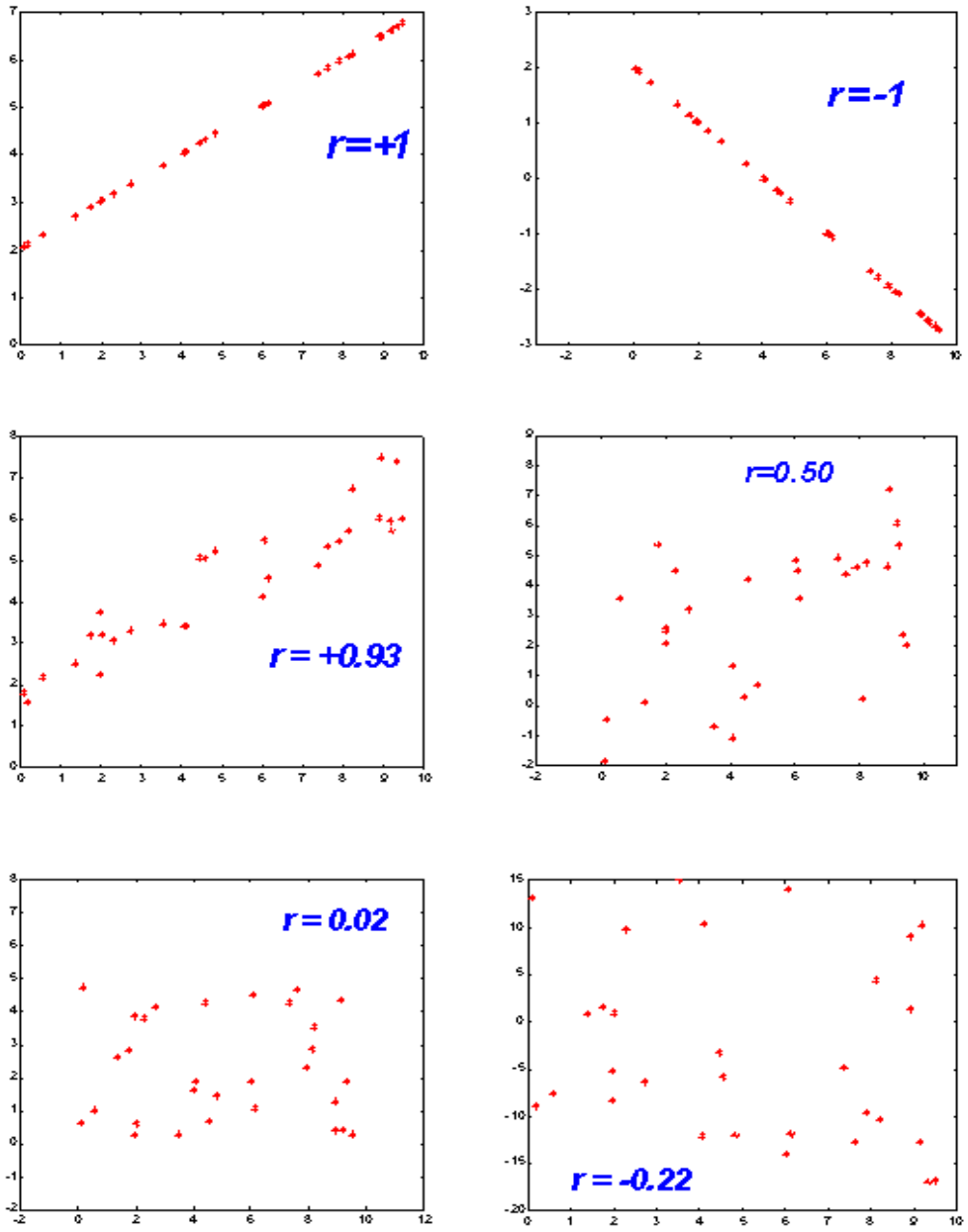

Figura 6.1: Gráfico y coeficiente de correlación lineal

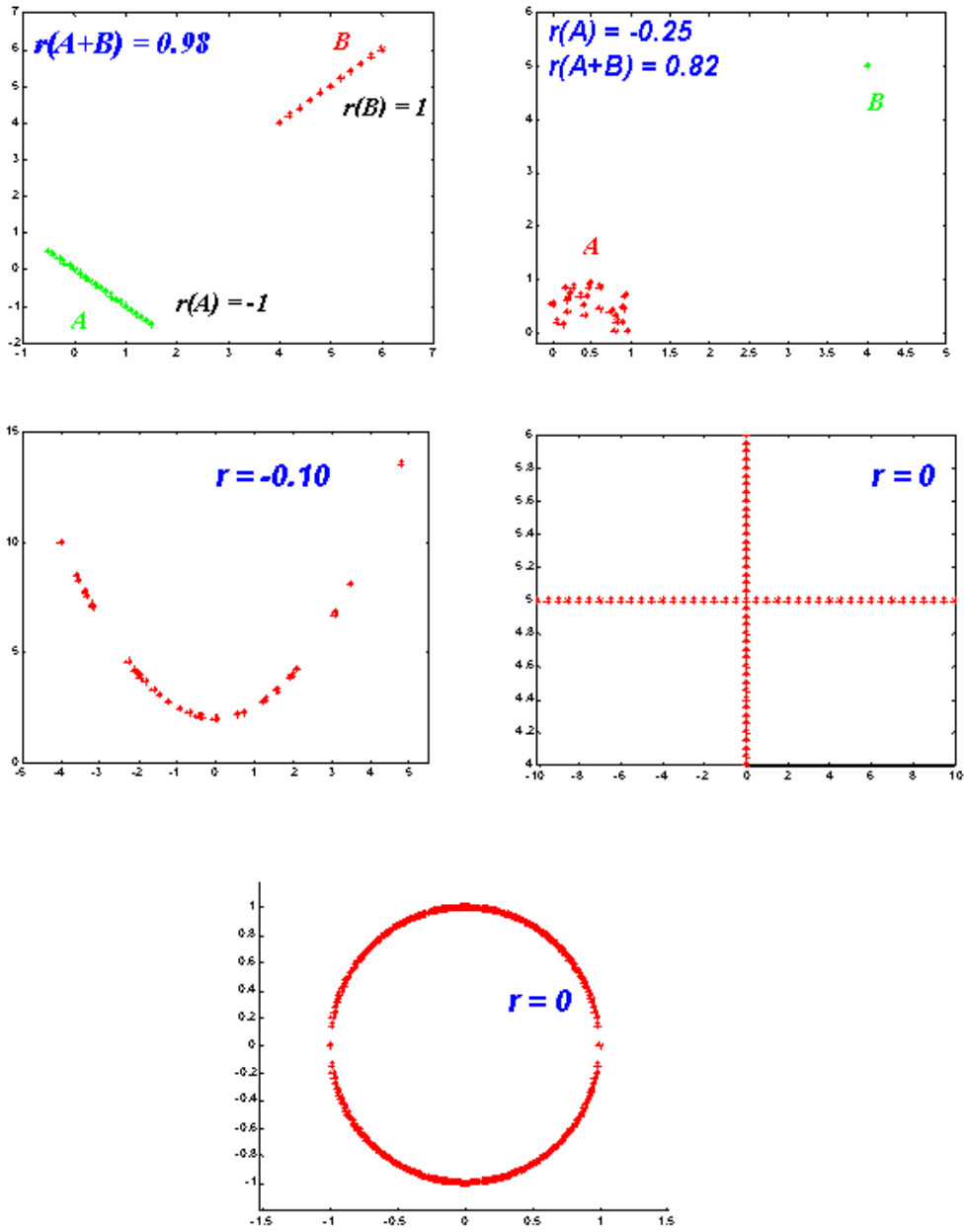

Figura 6.2: ¡OJO! al interpretar el coeficiente de correlación lineal

País	% Pob. Urbana	Tasa natalidad	Tasa mortalidad	Esperanza vida	Fecundidad	Mortalidad infantil
ARGENTINA	86.2	20.3	8.6	71.0	2.8	95.3
BOLIVIA	51.4	34.4	9.3	54.5	4.6	77.5
BRASIL	76.9	26.1	7.5	65.6	3.2	81.1
COLOMBIA	70.3	25.8	5.9	68.8	2.9	86.7
COSTA RICA	53.6	26.3	3.7	74.9	3.1	92.8
CHILE	85.6	22.5	6.4	71.8	2.7	93.4
ECUADOR	56.9	30.9	6.9	66.0	3.9	85.8
EL SALVADOR	44.4	33.5	7.1	64.4	4.0	73.0
GUATEMALA	42.0	38.7	7.6	63.4	5.4	55.1
HAITI	30.3	35.3	11.9	55.7	4.8	53.0
HONDURAS	43.6	37.1	7.2	64.9	4.9	73.1
MEXICO	72.6	27.9	5.4	69.7	3.2	87.3
NICARAGUA	59.8	40.5	6.9	64.8	5.0	81.0
PANAMA	54.8	24.9	5.2	72.4	2.9	88.1
PARAGUAY	47.5	33.0	6.4	67.1	4.3	90.1
PERU	70.2	29.0	7.6	63.0	3.6	85.1
R. DOMINICANA	60.4	28.3	6.2	66.7	3.3	83.3
URUGUAY	85.5	17.1	10.3	72.2	2.3	96.2
VENEZUELA	90.5	28.3	5.4	70.0	3.5	88.1
CUBA	74.9	17.4	6.7	75.4	1.9	94.0

Cuadro 6.1: Indicadores demográficos de 20 países de A.L.(PNUD 1992)

Variables	% Pob. % Urbana	Tasa natalidad	Tasa mortalidad	Esperanza vida	Fecundidad	Mortalidad infantil
% Pob. urbana	1.0	-0.739	-0.179	0.588	-0.735	-0.532
T. natalidad	-0.739	1.0	0.101	-0.723	0.972	0.682
T. mortalidad	-0.179	0.101	1.0	-0.609	0.262	0.533
Esperanza V.	0.588	-0.723	-0.609	1.0	0.769	0.951
Fecundidad	-0.735	0.972	0.262	0.769	1.0	0.709
M. infantil	0.532	0.682	0.533	-0.951	0.709	1.0

Cuadro 6.2: Matriz de correlación

la variables nominal, el clima con p categorías o modalidades, cada modalidad de X, es decir cada tipo de clima define un grupo o subpoblación de países y los grupos son disjuntos entre si. Conviene aquí usar notaciones que permitan distinguir los valores de la variable Y según la modalidad que toman las observaciones sobre la variable nominal X. Si n_j observaciones toman la modalidad o categoría j de X, se puede escribir y_{1j},...,$y_{n_j j}$ estas n_j observaciones de Y.

Decir que el PNB está relacionado al tipo de clima significa que conociendo el tipo de clima de un país se podrá inferir su PNB. Esto podrá darse si los valores del PNB difieren muchos de un grupo a otro.

Construiremos un índice basado en esta variabilidad del PNB.

Si $\bar{y}$ es la media empírica de la variable Y sobre el total de las n observaciones, la varianza

de todas estas observaciones es igual a:

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{n_j}(y_{kj} - \bar{y})^2$$

Como se puede distinguir las observaciones según la modalidad que toman sobre la variable X, se puede calcular medias y varianzas en los p grupos inducidos por las modalidades de X.

Si $\bar{y}_j$ es la media de la variable Y sobre las observaciones que toman la misma modalidad j, la varianza de las observaciones de este grupo es igual a:

$$w_j^2 = \frac{1}{n_j}\sum_{k=1}^{nj}(y_j - \bar{y}_j)^2$$

La variabilidad total de los valores proviene de dos fuentes: la variabilidad al interior de los grupos y la variabilidad entre los grupos.

Consideramos ω^2 la media ponderada de las varianzas de los p grupos: $\omega^2 = \sum_j \frac{n_j}{n} w_j^2$ y b^2 la varianza entre las medias $\bar{y}_j$ de los grupos: $b^2 = \sum_j \frac{n_j}{n}(\bar{y}_j - \bar{y})^2$. Considerando que la media ponderada por los efectivos relativos $\frac{n_j}{n}$ de las medias $\bar{y}_j$ es igual a la media total $\bar{y}$ ($\sum \frac{n_j}{n}\bar{y}_j = \bar{y}$) se puede mostrar que:

$$s^2 = \beta^2 + \omega^2 = \sum_j \frac{n_j}{n}(\bar{y}_j - \bar{y})^2 + \sum_j \frac{n_j}{n} w_j^2$$

Si w^2 es nula, todas las varianzas w_j^2 son nulas y todas las observaciones en un mismo grupo j toman el mismo valor sobre la variable Y, que es igual a la media $\bar{y}_j$ del grupo, y en consecuencias se podrá obtener el valor de un observación sobre la variable Y conociendo su modalidad sobre X. Se observa en este caso una relación funcional de X hacia Y. Esta relación permite estimar el valor de Y para una nueva observación conociendo su valor sobre X.

Al contrario si la varianza entre los grupos b^2 es nula, entonces todas las medias $\bar{y}_j$ son iguales a $\bar{y}$: no se podrá decir nada sobre el valor de Y conociendo la modalidad de X. No se detecto ninguna relación funcional de X hacia y. Se deduce el siguiente índice que permite medir el grado de asociación de tipo funcional de X hacia Y:

$$\eta^2_{Y|X} = \frac{b^2}{s_y^2}$$

Este coeficiente toma valores entre 0 y 1:

$\eta^2_{y\|x} = 1$	relación funcional estricta
$0 < \eta^2_{y\|x} < 1$	tendencia funcional
$\eta^2_{y\|x} = 0$	ausencia de tendencia funcional

La tendencia funcional aumenta con $\eta^2_{y|x}$.

6.3.1. Codificación óptima de una variable nominal

Si se codifica la variable X, atribuyendo un valor numérico a cada una de sus modalidades, podremos usar el coeficiente de correlación lineal como índice de asociación. Pero no se puede codificar de cualquier manera. Una forma natural de hacerlo consiste en buscar la codificación de las modalidades de X que produzca la mayor correlación lineal con la variable Y.

Si X tiene p modalidades, se le pueden asociar p variables indicativas $\{X^1, X^2, ..., X^p\}$ tales que

$$X^j(k) = \begin{cases} 1 & \text{si el individuo k toma la modalidad j de X} \\ 0 & \text{sino} \end{cases}$$

Se observa que $\sum_{j=1}^{p} X^j(k) = 1 \quad (\forall k)$.

Entonces si a_j es la codificación de la modalidad j $(j = 1, ..., p)$, la variable cuantitativa ξ asociada a esta codificación puede escribirse:

$$\xi(k) = \sum_j a_j X^j(k)$$

Dada $\{(x_i, y_i)|i = 1, ..., n\}$ una muestra de (X, Y), se define la codificación $\{a_j|j = 1, ...p\}$ que máximiza

$$cor(y, \sum_j a_j x^j)$$

Numéricamente el máximo de $cor^2(y, \sum_j a_j x^j)$ es igual a $\eta^2_{y|x}$.

6.3.2. Relación funcional entre dos variables cuantitativas

Cuando un coeficiente de correlación lineal entre X e Y es bajo, significa que las variables X e Y no están ligadas linealmente pero, puede existir otro tipo de relación entre ellas. Ahora bien, vimos que por codificación se puede transformar una variable nominal en una variable cuantitativa, inversamente, se puede transformar una variable cuantitativa en una variable ordinal, por lo tanto nominal particionando el recorrido de los valores de la variable en p intervalos.

Si se transforma X en variable nominal, se puede calcular la razón de correlación $\eta^2_{y|x}$, que permitirá detectar la existencia de una relación funcional de X hacia Y. El valor del coeficiente dependrá de la transformación (número de modalidades construidas). Se observa que ahora se tiene un coeficiente que no es simétrico en las variables como en el caso del coeficiente de correlación lineal. Por lo cual obtendremos resultados distintos según la variable que transformemos, salvo si existe una relación biyectiva entre las dos variables. Además, la

razón de correlación es más general que el coeficiente de correlación lineal, y se tiene que $cor^2(x,y) \leq \eta^2_{y|x}$.

Se ilustra en el ejercicio al final del capítulo como estas trasformaciones influyen sobre los coeficientes de asociación.

6.4. VARIABLES NOMINALES

6.4.1. Tabla de contingencia

Los datos obtenidos sobre las dos variables nominales pueden resumirse en una tabla de contingencia. Una tabla de contingencia contiene las frecuencias absolutas conjuntas de las dos variables, es decir las frecuencias obtenidas al cruzar las modalidades de una variable con las modalidades de la otra. En la elección de concejales de 1991, se pueden asociar a cada votante la lista votada y la región. Se puede resumir los resultados en una tabla de frecuencias (Tabla 6.3), que es la única información que se conoce realmente en este caso (por el anonimato de la elección). Esta es una tabla de contingencia. Al pasar de la información individual de los votantes relativa a las dos variables a la tabla de contingencia no se pierde información, salvo la identificación de cada individuo.

Se puede buscar en la tabla si hay mayor concentración de votantes en una región para un partido dado. Vamos a construir un índice que permite medir la existencia de una relación entre las dos variables.

6.4.2. Ji-cuadrado de contingencia

Veamos como leer una tabla de contingencia con ejemplos sencillos (Tablas 6.4 y 6.5 con la variable X en fila y la variable Y en columna. Se observa en la tabla 6.4(a), que las columnas B_1 y B_2 son proporcionales, lo que significa que reparten sus totales en las mismas proporciones entre las modalidades A_1 y A_2. Las modalidades B_1 y B_2 tienen los mismos perfiles. Al observar esta tabla no se ven muchas relaciones entre las dos variables (conociendo una modalidad de una variable, no se puede decir nada sobre la otra variable). No es el caso de la tabla 6.4(b). En efecto, si una observación toma la modalidad B_1, tomará la modalidad A_2 de X; dada A_1, entonces se tendra la modalidad B_3 de Y, pero dada A_2, se tendra B_1 ó B_2. Se tiene entonces una relación funcional de Y hacia X y existe una relación de X hacia Y, pero no es de tipo funcional.

En el caso de la tabla 6.5(d) existe una relación funcional, pero en la tabla 6.5(c) no hay ninguna.

Si denotamos n_{ij}, $(i = 1,..,p, j = 1,...,q)$ los elementos de una tabla de contingencia, se tienen los márgenes-filas: $n_{i\bullet} = \sum_j n_{ij}$, $i = 1,...,p$, y los márgenes-columnas $n_{\bullet j} = \sum_i n_{ij}$, $j = 1,...q$. Se define los perfiles condicionales como:

- Los perfiles condicionales-filas: $\frac{n_{ij}}{n_{i\bullet}}$

PARTIDO	I	II	III	IV	V	METR.	VI
D.C.	30412	63020	16793	58345	226333	759639	90521
RADICAL	19268	19265	9282	14336	39941	59767	21249
A.H. VERDE	2186	0	0	0	1680	43284	784
SOCIALDEMO	596	0	225	562	2817	6351	0
INDEP	346	73	55	86	1608	16493	2383
PPD	5165	11800	7390	21429	56405	295474	30714
SOCIALISTA	3405	15341	18339	28041	33282	177570	35779
INDEP	385	0	0	0	0	0	122
COMUNISTA	36648	13951	11588	21614	51135	171715	19312
LIBERAL	0	248	378	0	512	0	328
R.N.	12236	12424	16795	54648	96224	311801	54439
NACIONAL	0	0	0	0	422	2325	0
INDEP	3971	11669	4202	9385	40126	88614	7877
U.D.I.	8631	17464	8474	14495	71573	314984	33869
INDEP	587	980	47	0	6905	32008	6340
U.C.C.	6460	15428	5623	10671	73163	181913	26395
INDEP	105	5582	6007	1337	12263	37898	12797
IND IQUIQUE	24757	0	0	0	0	0	0
TOTAL	153888	187245	105198	234979	714389	24999836	342909

PARTIDO	VII	VIII	IX	X	XI	XII	TOTAL
D.C.	114070	223287	118841	121815	11555	13287	1848188
RADICAL	23416	61692	14420	26815	1602	2209	313562
A.H. VERDE	0	2931	1069	585	0	0	52519
SOCIALDEMO	7076	1110	6761	1291	0	0	26789
INDEP	1211	2631	1942	3572	45	27	30472
PPD	27759	64167	25498	29682	1250	8739	585472
SOCIALISTA	38338	94626	18987	51485	3715	20786	539694
INDEP	0	0	0	0	0	0	507
COMUNISTA	18379	50121	9824	13202	2342	2546	421377
LIBERAL	0	0	13842	0	241	0	15549
R.N.	60524	87849	56951	77702	8760	5807	856160
NACIONAL	0	1467	0	0	0	0	4214
INDEP	13644	29665	45384	23587	277	723	279124
U.D.I.	45905	75230	18194	32183	2065	8273	651340
INDEP	4794	16420	2358	3385	1598	731	76153
U.C.C.	47112	72049	21566	50650	1478	4237	516745
INDEP	13977	26376	9356	9066	119	1443	136326
IND IQUIQUE	0	0	0	0	0	0	24757
TOTAL	416205	809891	364993	445020	35047	69348	6378948

Cuadro 6.3: Resultados de la elección de consejales de 1991

(a)

	B_1	B_2	B_3	
A_1	50	100	10	160
A_2	100	200	50	350
	150	200	60	

(b)

	B_1	B_2	B_3	
A_1	0	0	50	50
A_2	10	12	0	22
	10	12	50	

Cuadro 6.4: Ejemplos de tablas de contingencias

(c)

	B_1	B_2	B_3	
A_1	20	10	7	37
A_2	40	20	14	74
A_3	80	40	28	148
	140	70	49	

(d)

	B_1	B_2	B_3	
A_1	0	20	0	20
A_2	30	0	0	30
A_3	0	0	0	25
	30	20	25	

Cuadro 6.5: Más ejemplos de tablas de contingencia

- Los perfiles condicionales-columnas: $\frac{n_{ij}}{n_{\bullet j}}$

La variable Y no influye sobre la variable X si y solo si los perfiles condicionales-columnas son todos iguales:

$$\frac{n_{i1}}{n_{\bullet 1}} = \frac{n_{i2}}{n_{\bullet 2}} = ... = \frac{n_{iq}}{n_{\bullet q}} = \frac{n_{i\bullet}}{n} \quad (i = 1, ..., p)$$

De la misma manera la variable X no influye sobre la variable Y si y solo si los perfiles condicionales-filas son todos iguales:

$$\frac{n_{1j}}{n_{1\bullet}} = \frac{n_{2j}}{n_{2\bullet}} = ... = \frac{n_{pj}}{n_{p\bullet}} = \frac{n_{\bullet j}}{n} \quad (j = 1, ..., q)$$

Luego las dos variables X e Y serán independientes si y solo si se cumplen a la vez las dos condiciones anteriores. Se puede demostrar que equivalen a:

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n} \quad \forall (i, j)$$

Considerando las diferencias $n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n}$, se puede evaluar cuán lejos está la relación entre X e Y de la independencia. Se puede construir un índice que traduzca estas diferencias, tomando en cuenta la importancia de cada una, ponderando por la magnitud de n_{ij} o $\frac{n_{i\bullet} \times n_{\bullet j}}{n}$. Es el índice χ^2 de contingencia:

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n})^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{n}}$$

Este índice es nulo cuando X e Y son independientes y crece al alejarse de la independencia hasta un valor máximo igual a $n \times Min\{p-1, q-1\}$ cuando hay una relación estricta entre los dos variables.

6.4.3. Codificación de las dos variables nominales

Ahora tendremos que codificar ambas variables. Sean a_i, $i = 1, ..., p$ las codificaciones de las modalidades de X y X^i, $i = 1, ..., p$ las variables indicadoras de X; sean b_i, $i = 1, ..., q$ las codificaciones de las modalidades Y e Y^i, $i = 1, ..., q$ las indicadoras de Y.

Se busca codificaciones respectivas de X e Y tales que el coeficiente de correlación lineal de las codificaciones

$$cor(\sum_i a_i x^i, \sum_j b_j y^j)$$

sea máximo.

Esta correlación se usa en una técnica llamada análisis factorial de correspondencias y esta relacionada al ji-cuadrado de contingencia.

6.4.4. Relación entre dos variables cuantitativas

Si transformamos las dos variables cuantitativas en variables nominales podremos usar el χ^2 de contingencia que nos permita detectar una relación de cualquier tipo, no solamente lineal o funcional.

Para hacer las transformaciones se requiere un gran número de observaciones para tener una cantidad suficiente de elementos en cada celda de la tabla de contingencia.

Se observara que las transformaciones producen variables menos precisas que las originales, pero con estas se puede investigar otras relaciones que las lineales.

6.5. VARIABLES ORDINALES

6.5.1. Coeficientes de correlación de rangos

A partir de una variable ordinal, se pueden ordenar las observaciones de manera creciente y deducir una nueva variable que es *el rango*, que indica la posición de cada observación según el orden.

Sean $x_1, ..., x_n$ las realizaciones de la variable ordinal X y $R_{x_1}, ..., R_{x_n}$ los rangos asociados:

$$R_{x_i} < R_{x_j} \Longleftrightarrow x_i < x_j$$

Si R_{x_i} y R_{y_i}, $i = 1, 2, ..., n$, son los rangos asociados a X e Y respectivamente, se define entonces **el coeficiente de rangos de SPEARMAN** R_S de x e y como el coeficiente de correlación lineal empírico entre R_x y R_y.

Si $D_i = R_{x_i} - R_{y_i}$, se obtiene una expresión más práctica:

$$R_S = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}$$

Se observa entonces que si los rangos inducidos por X e Y son idénticos $R_S = 1$; si son totalmente opuestos R_S = -1.

Si en vez de definir los rangos, se define dos nuevas variables sobre los pares de observaciones:

$S(x_i, x_j) = $ -1	si $x_i \geq x_j$
$S(x_i, x_j) = 1$	si $x_i < x_j$
$S(x_i, x_j) = $ -1	si $x_i \geq x_j$
$S(y_i, y_j) = 1$	si $y_i < y_j$
$S(y_i, y_j) = $ -1	si $y_i \geq y_j$

Se define entonces **el coeficiente de correlación de rangos de KENDALL**:

$$\tau = \frac{\sum_{i,j} S(x_i, x_j) S(y_i, y_j)}{n(n-1)}$$

El numerador es igual al número de pares de observaciones con el mismo orden menos el número de pares de observaciones con orden contrario. El numerador es igual al número total de pares. Como el coeficiente R_S de Spearman, τ toma valores entre -1 y $+1$ y vale $+1$ si los ordenes son idénticos y -1 cuando son totalmente opuestos.

6.5.2. Relación entre dos variables cuantitativas

A partir de una variable cuantitativa se pueden ordenar las observaciones, y por lo tanto, construir los rangos. Puede ser útil especialmente cuando los valores de las variables no son muy precisos o bien, si se busca la existencia de una relación monótona no lineal entre X e Y. Se pueden aplicar entonces los coeficientes de correlación de rangos anteriores.

6.6. INFERENCIA

Suponiendo que un coeficiente de asociación fue correctamente calculado, es decir que fue calculado sobre una muestra aleatoria simple de una sola población, queremos saber a partir de qué valor se puede decidir la existencia o ausencia de una relación. Para esto se procede mediante un test de hipótesis sobre el valor del coeficiente v de asociación desconocido de la población: $H_0 : v = v_0$, o bien se puede calcular un intervalo de confianza para v. Para eso se requiere conocer la distribución del coeficiente de asociación v en la muestra.

6.6.1. Coeficiente de correlación lineal

¿Cuando se obtiene un coeficiente de correlación lineal r pequeño podemos admitir que la correlación ρ en la población es nula ó si r es grande, podemos concluir que existe una relación lineal?

Para responder a la pregunta se procede mediante un test de hipótesis sobre el valor del coeficiente de correlación ρ desconocido de la población: $H_0 : \rho = \rho_0$, o bien se puede calcular un intervalo de confianza para ρ. El problema es que la distribución del coeficiente de correlación r no es siempre fácil de establecer.

Cuando $\rho = 0$ y las dos variables X e Y provienen de una distribución normal bivariada, la distribución del coeficiente r de la muestra es fácil de obtener ye depende del tamaño n de la muestra: existen tablas de la distribución de r en función de n y para $n > 100$ y se puede aproximara a la normal $\mathcal{N}(0, \frac{1}{\sqrt{n-1}})$.

Por ejemplo, si un coeficiente de correlación lineal r es igual a 0,38 sobre una muestra de $n = 52$ observaciones, vamos a rechazar que $\rho = 0$ al nivel de significación de 5 % o incluso 1 %, dado que $I\!P(|r| > 0{,}27) = 0{,}05$ y $I\!P(|r| > ,35) = 0{,}01$, pero si $r = 0{,}32$ con el mismo tamaño $n = 52$, entonces se rechaza al nivel de 5 % pero no al nivel de 1 %.

Cuando ρ no es nulo, la distribución exacta de r es mucho más complicada de determinar, sin embargo se puede usar una aproximación a partir de $n = 25$: si $Z = 1/2 \ln(\frac{1+r}{1-r})$, la distribución de Z se aproxima a una normal $\mathcal{N}(1/2 \ln(\frac{1+\rho}{1-\rho}), \frac{1}{\sqrt{n-3}})$.

Finalmente, si las dos variables no siguen una distribución normal, se puede usar los resultados anteriores cuando n es mayor que 30, pero si ρ es nulo, no se puede decir que hay independencia, pero sólo que no hay ligazón lineal.

6.6.2. Razón de correlación: ANOVA a un factor

Para estudiar la significatividad de una razón de correlación empírica obtenida sobre n observaciones entre la variable cuantitativa Y con la variable nominal X a p modalidades, se plantea la hipótesis nula $H_0 : \gamma^2 = 0$ donde γ^2 es la razón de correlación en la población. El problema es que no conocemos la distribución de la razón de correlación observado $\eta^2_{Y|X}$.

Si $Y \sim \mathcal{N}(\mu_j, \sigma_j^2)$ cuando X toma la modalidad j, entonces la hipótesis nula puede escribirse $H_o : \mu_1 = \mu_2 = ... = \mu_p$.

Sabemos que si w_j^2 es la varianza observada del grupo j con un efectivo n_j, $\frac{n_j s_j^2}{\sigma_j^2} \sim \chi^2_{n_j-1}$. Suponiendo que el muestreo es aleatorio simple, los grupos son independientes entre si y suponiendo que todas las varianzas σ_j^2 son iguales a σ^2, $\frac{nw^2}{\sigma^2} \sim \chi^2_{n-p}$ donde w^2 es el promedio de las varianzas dentro los grupos.

Por otra parte si b^2 es la varianza observada entre los p grupos, $\frac{nb^2}{\sigma^2} \sim \chi^2_{p-1}$. Además b^2 y w^2 son independientes, dado que cada media $\bar{y_j}$ es independiente de la varianza w_j^2 del grupo y que los grupos son independientes entre si.

Se considera entonces el estadístico que es el cociente de los dos χ^2 divididos respectivamente por sus grados de libertad. Como se cancelan n y σ^2, se obtiene:

$$\frac{b^2/(p-1)}{w^2/(n-p)}$$

que sigue una distribución F de Fisher a $p-1$ y $n-p$ grados de libertad bajo la hipótesis H_o de ausencia de relación de X hacia Y.

Se observará que

$$\frac{b^2/(p-1)}{w^2/(n-p)} = \frac{\eta^2/(p-1)}{1-\eta^2/(n-p)}$$

Consideremos el ejemplo debido a Ronald Fisher (párrafo **??**) sobre 3 especies de flores de la familia de los "iris": setosa, versicolor y virgínica (variable X con 3 modalidades). Se busca verificar si las 3 especies se distinguen por algunas mediciones. Se usan dos variables: Y_1 el largo del pétalo e Y_2 el ancho del sépalo (tabla 6.6).

Especie	efectivo	Y_1		Y_2	
		Media	Varianza	Media	Varianza
Setosa	50	34.280	3.39	14.620	4.30
Versicolor	50	27.700	3.39	42.600	4.30
Virginica	50	29.740	3.39	55.520	4.30

Cuadro 6.6: medias y varianzas

Entre X e Y_1 la razón de correlación es igual a $\eta^2_{Y_1|X} = 0{,}40$ y entre X e Y_2 la razón de correlación es igual a $\eta^2_{Y_2|X} = 0{,}94$. Claramente el largo del pétalo es diferentes de una especie a otra pero no se puede afirmar nada para el ancho del sépalo. Los resultados del ANOVA para ambos casos se encuentran en las tablas 6.7 y 6.8. La varianza b^2 proviene de la especie y la varianza w^2 se interpreta como un error cuando se supone que las medias de las 3 especias son iguales. Si bien en ambos casos se rechaza la hipótesis nula (p-valor nulo), el valor del F es mucho más pequeño en el caso del ancho del sépalo, lo que indica una relación menos clara.

6.6.3. Ji-cuadrado de contingencia

¿Si dos variables nominales X e Y son independientes, cuales son los valores más probables del χ^2 de contingencia? Como vimos en el párrafo **??**,

$$Q = \sum_{ij} \frac{(n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n})^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{n}}$$

Si X e Y son independientes, $n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$ para todo par (i, j); en esta caso el estadístico Q sigue una distribución aproximada de χ^2 a $(p-1)(q-1)$ grados de libertad, si p y q son los números de modalidades de X e Y respectivamente.

6.6.4. Coeficiente de correlación de rangos de Spearman

Cuando X e Y resultan de un ordenamiento sin empates, entonces los rangos inducidos por X ó Y son valores de $\{1, .., n\}$ y los rangos de uno se obtienen por permutación de los rangos

Fuente	n*varianza	g.l.	n*varianza/g.l.	F	p-valor
Especie (b^2)	1134.493	2	567.247	49.160	0.000
Error (w^2)	1696.2	147	11.54		
Total (s^2)	2830.693	149			

Cuadro 6.7: Ancho del sépalo

Fuente	n*varianza	g.l.	n*varianza/g.l.	F	p-valor
Especie (b^2)	43710.28	2	21855.14	1180.16	0.000
Error (w^2)	2722.26	147	18.518776		
Total (s^2)	46432.54	149			

Cuadro 6.8: Largo del pétalo

del otro.

Si X e Y son independientes, cualquiera sean las leyes de X e Y, las dos permutaciones inducidas son independientes. En este caso, si el ordenamiento de X esta fijado, las $n!$ permutaciones de este ordenamiento son equiprobables. Se tiene tres maneras de obtener la distribución de R_S bajo la hipótesis de independencia:

- Si n es muy pequeño, se puede obtener empíricamente la distribución de R_S, calculando los $n!$ valores asociados a las distintas permutaciones.
- Para $n < 100$, existen tablas de la distribución en función de n.
- Para n grande se puede usar la aproximación a la normal $\mathcal{N}(0, \frac{1}{\sqrt{n-1}})$.

El coeficiente de Spearman entre la Esperanza de Vida y la Tasa de Mortalidad de la tabla 1 vale 0,48.

En las tablas de la distribución del coeficiente de Spearman encontramos que $P(|R_S| > 0{,}447) = 0.05$, lo que nos lleva a rechazar la independencia entre la Esperanza de Vida y la Tasa de Mortalidad.

6.6.5. Coeficiente de correlación de rangos de Kendall

Como en el caso del coeficiente de correlación de Spearman, se puede construir empíricamente la distribución del τ de Kendall cuando n es muy pequeño. Pero a partir de $n > 8$, se puede aproximar a una distribución normal $\mathcal{N}(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}})$

Para las variables Esperanza de Vida y Tasa de Mortalidad de la Tabla 1, obtenemos $\tau = 0{,}326$.

$$P(|\tau| < 1{,}96\sqrt{\frac{90}{180 \times 19}}) = P(|\tau| < 0{,}317) = 0{,}05$$

Nuevamente encontramos significativa la relación entre las dos variables.

6.7. EJERCICIO

Sea un conjunto I de $n = 300$ individuos, y cuatro variables cuantitativas X, Y, Z y T observadas sobre los 300 individuos. X varia entre -100 y 100, Y varia entre 0 y 10000, Z y T varían entre -1100 y 1100.

1. Los coeficientes de correlación lineal calculados sobre los 300 individuos son: $R_{X,Y} = -0{,}057$, $R_{Z,T} = 0{,}991$. Interprete estos coeficientes.

2. Se transforma la variable X en una variable nominal particionando [-100,100] en q intervalos iguales; se llama X_1, X_2, X_3 y X_4 a las variables nominales obtenidas para q=10, 8, 6 y 4 respectivamente. Interprete las razones de correlación obtenidas y concluir: $\eta_{Y/X_1} = 0{,}96$, $\eta_{Y/X_2} = 0{,}93$, $\eta_{Y/X_3} = 0{,}86$ y $\eta_{Y/X_4} = 0{,}74$,

3. Se transforma la variable Y en una variable nominal con la partición del intervalo [0,10000] en q intervalos iguales; se llama Y_1, Y_2, Y_3 y Y_4 a las variables nominales obtenidas para q=10, 8, 6 y 4 respectivamente. Interprete las razones de correlación obtenidas y concluya: $\eta_{X/Y_1} = 0{,}038$, $\eta_{X/Y_2} = 0{,}027$, $\eta_{X/Y_3} = 0{,}024$ y $\eta_{X/Y_4} = 0{,}015$,

4. Se calcula los χ^2 de contingencia entre las variables nominales asociadas a X e Y: $\chi^2_{X_1,Y_1} = 853$, $\chi^2_{X_2,Y_2} = 679$, $\chi^2_{X_3,Y_3} = 450$ y $\chi^2_{X_4,Y_4} = 306$. Concluya.

5. Interprete el coeficiente de correlación parcial de Z y T dado X $R_{Z,T|X} = 0{,}027$. Compare con $R_{Z,T}$ y interprete.