APUNTES DE ESTADISTICA

Nancy Lacourly

1996

Con la colaboración de Ernesto San Martín y Felipe Farías

PREFACIO

Este curso de estadística hace parte del plan común de ingeniería [1]. Como para algunas carreras es el único curso que tendrá el alumno de Ingenieria, se ha trata aquí dar una visión de la metodología basica de la Inferencia Estadística y una introducción a los modelos lineales y métodos multidimensionales. Se busca preparar al futuro ingeniero en la aplicación de modelos estadísticos para tratar fenómenos aleatorios en física, mecánica o economía en donde se encuentra errores de medición, errores de muestreo etc., así como grandes volumenes de datos que en la actualidad pueden ser estudiados fácilmente.

Si bien el cálculo de las probabilidades es una teoría matemática abstracta, que deduce consecuencias de un conjunto de axiomas, al contrario la estadistica necesita dar una interpretación concreta a la noción de probabilidad. Varias interpretaciones fueron propuestas por los estadisticos, que se pueden resumir en dos puntos de vista diferentes: la noción frecuentista y la noción intuicionista.

El punto de vista *frecuentista* asocia la noción de probabilidad a la noción empírica de frecuencia, basada en observaciones aleatorias repetidas, mientras que el punto de vista *intuicionista* liga la noción de probabilidad a lo incierto, para definir un grado de creencia.

[1] Este texto fue financiado parcialmente por la Escuela de Ingeniería y Ciencias (Proyecto Docente 139301)

INDICE

1 INTRODUCCION A LA ESTADISTICA

La estadística es una rama del método científico que trata datos empíricos, es decir datos obtenidos contando o midiendo propiedades sobre poblaciones de fenómenos naturales, cuyo resultado es "incierto".

En teoríia de las probabilidades, estudiaron el experimento relativo a tirar un dado y hicieron el supuesto que el dado no está cargado (sucesos elementales equiprobables), lo que permite deducir que la probabilidad de sacar "un número par" es igual a 1/3. A partir de un modelo probabilitico adecuado, se deduce nuevos modelos o propiedades. En Estadíistica tratamos responder a la pregunta "¿el dado no está cargado?", comprobando si el modelo probabilistico de equiprobable subyacente esta en acuerdo con datos experimentales obtenidos tirando el dado un cierto número de veces. Se propone entonces un modelo probabilitico que debe seguir los datos y no lo contrario.

La teoría de las probabilidades permite deducir propiedades a partir de una serie de axiomas, mientras que la Estadítica propone métodos para verificar hipótesis.

Esta introducción se inicia con una breve presentación histórica de la estadística, para seguir con algunos ejemplos de problemas estadísticos. Siguen las etapas del razonamiento que se usa para resolver tales problemas. Terminamos esta introducción con la presentación de la teoría de muestreo, que es la base de la solución de todo problema estadístisco.

1.1 HISTORICO

Antes de la aparición del cálculo de las probabilidades en el siglo 17, la estadística se ha desarrollado poco y se limita a estudio descriptivo, que es la parte de la estadística que no se apoya sobre la noción de probabilidad. En efecto es una actividad bien antigua, aquella de recolectar datos para conocer la situación de los estados: el emperador chino Yao organizó un censo de producciones agricolas en 2238 A.C.; en Egipto ya se hacían catastros y censos en 1700 A.C.; más cerca, los Incas con sus quipus mantenían al día las estadísticas de las cosechas. Durante este período, los censos de poblaciones y recursos naturales son sólo cifras informativas y descriptivas. Es sólo en el siglo 18 que se expande la idea introducida por el ingles John Grant, que las estadísticas demográficas podrían servir de base a *predicciones*. Con Adophe Quetelet se empieza a concebir que la estadística puede ser fundada en el cálculo de las probabilidades. Pero hay que esperar los primeros estadísticos matemáticos ingleses (después de 1900) para ver realmente una metodología estadística como una teoría *inductiva* bien formalizada, que permite inducir a partir de datos observados particulares, conclusiones generales sobre el comportamiento probabilístico de fenómenos observados. Después de la Estadística Matemática, que se desarrolla entre 1900 y 1950, los estadísticos neo-bayesianos proponen hacer inferencia, no sólo a partir de los datos observados, sino tomando también en cuenta el conocimiento *a priori* respecto de los modelos probabilísticos. En la misma época (1950), la aparición de los computadores potentes permite el auge del análisis de grandes

volumenes de datos, con más observaciones y más variables. Un conjunto de técnicas para estudiar datos multidimensionales, que se basan en modelos no probabilísticos, permiten describir, clasificar y simplificar los datos con el objeto de facilitar su interpretación además de sugerir leyes, modelos o explicar fenómenos.

1.2 EJEMPLOS DE PROBLEMAS ESTADISTICOS

- Probar si una moneda está cargada.
- Hacer predicciones demográficas a partir de un censo.
- Controlar de la calidad de un proceso de fabricación.
- Estudiar la confiabilidad de un material.
- Evaluar el efecto de un fertilizante sobre la cosecha del choclo.
- Evaluar la eficacia de una droga para combatir una enfermedad.
- Predecir los resultados de una elección presidencial.
- Evaluar la audiencia de los programas de televisión.
- Evaluar el efecto del consumo de alcohol sobre los reflejos del conductor.
- Evaluar la pobreza en un país.

Todos estos problemas son distintos; algunos se podrán basar en datos censales y otros en datos muestrales. Pero hay una línea general del razonamiento que es la misma para todos.

1.3 EL RAZONAMIENTO ESTADISTICO

Las etapas del razonamiento estadístico son generalmente las siguientes:

- Recolección de los datos.
- Descripción estadística de los datos.
- Análisis de los datos.
- Decisión o predicción.

1.3.1 Recolección de los datos

Se distingue los censos, en que los datos estan recolectados sobre la integralidad de las unidades de la población considerada, de los muestreos, en los cuales se recoge informaciones sobre sólo una parte de la población. La forma de elegir la muestra depende del problema (diseño de muestreo y diseño de experimentos) y puede ser muy compleja, pero generalmente la muestra está obtenida aleatoriamente y llama a usar la teoría de las probabilidades.

1.3.2 Descripción estadística de los datos

La descripción estadística permite resumir, reducir y presentar el contenido de los datos con el objeto de facilitar su interpretación, sin considerar que estos datos provienen de una muestra. Las técnicas dependerán del volumen de las observaciones, de la cantidad de las variables, de la naturaleza de los datos y de los objetivos del problema.

1.3.3 Análisis de los datos

El análisis estadístico es la etapa más importante del razonamiento estadístico, y generalneralmente se basa en un modelo matemático o probabilístico. Tal modelo dependerá de los datos y eventualmente del conocimiento *a priori* que se puede tener sobre el fenómeno estudiado. El modelo no está en general totalmente determinado (es decir, se plantea una familia de modelos de un cierto tipo); por ejemplo, en el caso de modelos probabilísticos podrá ser una distribución normal, una distribución de Poisson o una distribución Beta, o en el caso de modelos matemáticos podrá ser un modelo lineal. Estos modelos tendrán algunos **parametros** indeterminados. Se trata entonces de fijar lo mejor posible tales parámetros desconocidos a partir de datos empíricos obtenidos sobre una muestra: **es un problema de estimación estadística**. Por otro lado, antes o durante el análisis, se tienen generalmente consideraciones teóricas respecto del problema estudiado y se trata entonces de comprobarlas o rechazarlas a partir de los datos empíricos: **es un problema de test estadístico**.

1.3.4 Decisión o predicción

Una vez analizados los datos, se tiene en general que tomar una decisión o proceder a alguna predicción, que dependerá del análisis previo. Por ejemplo, se tiene que decidir, a partir de algunos experimentos, si un tratamiento es eficaz, o bien predecir el IPC del próximo mes.

1.4 TEORIA DE MUESTREO

Una base importante de la estadística está contenida en la teoría de muestreo.

Los datos experimentales son obtenidos sobre conjunto de individuos u objetos, llamado **población**, sobre el cual se quiere conocer algunas características. La población puede ser finita -por ejemplo, en una encuesta de opinión, es la población de un país o una región, los productos fabricados por una maquina- o infinita, cuando la población se define a partir del experimento de tirar un dado, o sacar valores de la distribución de probabilidad de la v.a. $\mathcal{N}(0,1)$ (es el espacio muestral). Como generalmente la población a estudiar es demasiado vasta o incluso infinita, se extrae solamente un subconjunto de la población, llamada **muestra** sobre la cual se observan características llamadas **variables**. ¿Cómo entonces sacar una muestra de una población o de una distribución de probabilidad desconocida para obtener informaciones fidedignas sobre la población de la cual proviene? Es lo que pretende contestar la teoría de muestreo, planteando la pregunta de otra manera: ¿Si la distribución probabilidad de obtener la muestra que se obtuvo? La teoría de muestreo permite de defimir el tamaño de la muestra a tomar pero la forma de seleccionar los elementos de la muestra también. Se tiene varios métodos de muestreo para obtener muestras que, dependiendo del problema, pueden ser muy complejos.

Los valores de las variables obtenidos sobre los elementos de la muestra se llaman **valores muestrales**. Ahora bien, cuando se emiten conclusiones sobre una población a partir sólo de valores muestrales, entonces estos resultados están afectados de **errores** debidos al muestreo. Pero se tiene generalmente errores de medición también que pueden influir sobre la precisión de las conclusiones.

Ahora bien hay que observar que los errores de muestreo decrecen con el tamaño de la muestra, pero los errores de observación crecen con este tamaño. Lo ideal es entonces tener un buen equilibrio entre estos tipos de errores.

Se vió en el curso de probabilidad que el muestreo aleatorio simple (m.a.s.) permite sacar muestras de tamaño dado equiprobables, distinguiendo el m.a.s. con reemplazo del m.a.s. sin reemplazo.

Dado un experimento aleatorio $\mathcal{E}$ y una población (o espacio muestral) Ω de sucesos elementales, el conjunto de n realizaciones del experimento $\mathcal{E}$ es **una muestra de tamaño n**.

- Una muestra aleatoria simple con reemplazo (o con repetición) se obtiene realizando n repeticiones independientes del experimento $\mathcal{E}$, tomando sobre Ω los sucesos elementales equiprobables. Se obtiene entonces una n-tupla de Ω.
- Una muestra aleatoria simple sin reemplazo (o sin repetición) se obtiene de la población Ω realizando el experimento $\mathcal{E}$:
  - sobre Ω. Se obtiene un suceso ω_1 con equiprobabilidad;
  - sobre $\Omega \setminus \{\omega_1\}$. Se obtiene un suceso ω_2 con equiprobabilidad;
  - sobre $\Omega \setminus \{\omega_1, \omega_2\}$. Se obtiene un suceso ω_3 con equiprobabilidad, etc.

  Así se obtienen elementos de Ω, todos distintos.

El muestreo aleatorio simple es un método para obtener muestras de tamaño fijo de tal forma que todas las muestras de mismo tamaño tengan la misma probabilidad de ser seleccionadas. Pero no es la única forma de proceder.

2 DISTRIBUCIONES EN EL MUESTREO

2.1 INTRODUCCION

Los métodos estadísticos permiten confrontar modelos matemáticos o probabilísticos con los datos empíricos obtenidos sobre una muestra:

> *Dadas observaciones obtenidas sobre una muestra de tamaño n, se busca deducir propiedades de la población de la cual provienen.*

Si se tiene una sola variable aleatoria X cuya función de distribución F es desconocida, obteniendo observaciones de esta variable X, buscaremos conocer a la función de distribución F de la población. Los valores $X_1, X_2, ..., X_n$ de una v.a. X obtenidos sobre una muestra de tamaño n son **los valores muestrales**.

Se busca entonces, por ejemplo, **estimar** la media de la distribución F a partir de los valores muestrales. Esto tendrá sentido si la muestra es **representativa** de la población.

2.2 TIPOS DE VARIABLES

La cantidad y la naturaleza de las cacterísticas que se puede medir sobre los elementos de una población Ω son de varios tipos. Supondremos aquí una sola variable que es una función $X: \Omega \longrightarrow Q$. Se distingue la naturaleza de la variable X según el conjunto Q:

- variable cuantitativa (también llamada intervalar) si Q es un intervalo de $\mathbb{R}$ o todo $\mathbb{R}$; es una v.a. real continua.
- variable discreta si Q es un subconjunto de $\mathbb{N}$;
- variable cualitativa (o nominal) si Q es un conjunto finito de atributos (o modalidades) no numéricos;
- variable ordinal si Q es un conjunto de atributos no numéricos que se pueden ordenar.

El tratamiento estadístico depende del tipo de variable considerada.

2.3 FUNCION DE DISTRIBUCION EMPIRICA

2.3.1 Caso de variables numericas (reales o enteras)

Sean $X_1, X_2, ..., X_n$, los valores muestrales obtenidos de un m.a.s..

$F_n(x) = \frac{Card\{X_i / x_i \leq x\}}{n}$ es la proporción de observaciones de la muestra inferiores o iguales a x; $F_n(x)$ tiene las propiedades de una función de distribución: $F - n(x)$ es monotona no decreciente; tiene limites a la derecha y a la izquierda; es continua a la derecha; $F(-\infty) = 0$; $F(+\infty) = 1$. Además sus puntos de discontinuidad son en número finito y son con salto;

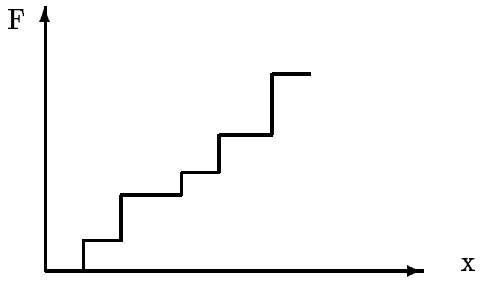


Figura 2.1: Una distribución empírica.

Además para x fijo $F_n(x)$ es una variable aleatoria y $nF_n(x)$ es una v.a. igual a la suma de variables de Bernoulli independientes de mismo parámetro $F(x)$, o sea $nF_n(x) \sim \mathcal{B}(n, F(x))$.

Teorema 2.1 *Para todo x, $F_n(x)$ converge casi-seguramente hacia la distribución teórica F(x) de X.*

Demostración: Como $nF_n(x) \sim \mathcal{B}(n, F(x))$, de la ley de los grandes números se concluye que:

$$P(\lim_n F_n(x) = F(x)) = 1$$

O sea que $F_n(x) \xrightarrow{c.s.} F(x)$

Teorema 2.2 *(Glivenko-Cantelli)*

$$D_n = \sup_x \mid F_n(x) - F(x) \mid \longrightarrow 0$$

Teorema 2.3 *(Kolmogorov)*
La distribución asintótica de D_n es conocida y no depende de X:

$$\lim_{n\to\infty} P(\sqrt{n} D_n < y) = \sum_{-\infty}^{+\infty} (-1)^K exp(-2K^2 y^2)$$

No se demuestran estos dos teoremas.

2.3.2 Caso de variables no son numéricas (nominal u ordinal)

Cuando las variables no son numéricas, Q es un conjunto finito:
$Q = \{q_1, q_2, ..., q_r\}$. La distribución de población está definida por las probabilidades
$I\!P(X = q_k)$ ($\forall$ k = 1,...,r).

Dada una muestra aleatoria simple $X_1, X_2, ..., X_n$ de tamaño n, se define las proporciones en el muestreo $s_j = \frac{Card\{X_i = q_j\}}{n}$, $j = 1, ..., r$.

Consideramos el caso $r = 2$, por ejemplo, una pieza es defectuosa o no es defectuosa; sea p la probabilidad desconocida que una pieza esté defectuosa. Dada una muestra aleatoria simple de tamaño n, si f_n es la proporción de piezas defectuosas encontradas entre las n observadas, nf_n sigue una distribución Binomial(n,p) y además $f_n \longrightarrow \mathcal{N}(p, p(1-p)/n)$.

2.4 DISTRIBUCIONES EN EL MUESTREO Y EN LA POBLACION

Sean $X_1, X_2, ..., X_n$, los valores muestrales.

Definición 2.1 *Las funciones de los valores muestrales son v.a. llamadas* **estadísticos** *y las distribuciones de los estadísticos se llaman* **distribuciones en el muestreo**.

La distribución de la v.a. X, que es generalmente desconocida, se llama **distribución de población**. Se le da en general una expresión **teórica**. Se supone, por ejemplo, que la distribución de población pertenece a una familia de distribuciones, por ejemplo la distribución normal, la distribución beta o la distribución de Poisson. Quedan desconocidas, en este caso, sólo algunas características. Estas características, son los **parámetros** de la distribución de población.

Los estadísticos y sus distribuciones en el muestreo (o sus distribuciones asintóticas cuando n tiende a $+\infty$) permiten **estimar** los parámetros desconocidos de la distribución de población.

2.4.1 Media muestral

Sean $X_1, X_2, ..., X_n$, los valores muestrales independientes e identicamente distribuidos (i.i.d.) de una v.a. X. Se define la media muestral como $\bar{X}_n = \sum X_i/n$. Si la distribución de población tiene como esperanza y varianza μ y σ^2 respectivamente ($E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ para todo i), entonces $E(\bar{X}_n) = \mu$ y $Var(\bar{X}_n) = \sigma^2/n$. Si además la distribución de población es normal entonces la distribución en el muestreo de $\bar{X}_n$ también lo es. Los valores muestrales X_i no provienen necesariamente de una distribución normal pero si son i.i.d., entonces la distribución asintótica de $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ es $\mathcal{N}(0,1)$ (TEOREMA DEL LIMITE CENTRAL).

2.4.2 Varianza muestral

Sea una m.a.s. $\{X_1, X_2, ..., Xn\}$, con $\mathrm{E}(X_i) = \mu$ y $\mathrm{Var}(X_i) = \sigma^2$.

$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X}_n - \mu)^2$

Propiedades:

- $S_n^2 \xrightarrow{c.s.} \sigma^2 \qquad (\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{c.s.} E(X^2)$ y $\bar{X}_n^2 \xrightarrow{c.s.} [E(X)]^2)$.
- $S_n^2 \xrightarrow{m.c.} \sigma^2 \qquad (E((S_n^2 - \sigma^2)^2) \longrightarrow 0)$.
- Cálculo de $E(S_n^2)$
  $E(S_n^2) = E(\frac{1}{n}\sum(X_i^2 - \bar{X}_n)^2) = E(\frac{1}{n}\sum(X_i^2 - \mu)^2 - (\bar{X}_n - \mu)^2)$
  $E(S_n^2) = \frac{1}{n}\sum Var(X_i) - Var(\bar{X}_n) = \frac{1}{n}\sum \sigma^2 - \frac{\sigma^2}{n}$
  $E(S_n^2) = \frac{n-1}{n}\sigma^2 \longrightarrow \sigma^2$.
- Cálculo de $Var(S_n^2)$
  $Var(S_n^2) = \frac{n-1}{n^3}((n-1)\mu_4 - (n-3)\sigma^4)$
  en que $\mu_4 = E((X - \mu)^4)$ es el momento teórico de orden 4 de la v.a. X.
  Se deja este cálculo como ejercicio.
  $Var(S_n^2) \approx \frac{\mu_4 - \sigma^4}{n} \longrightarrow 0$.
- Cálculo de $Cov(\bar{X}_n, S_n^2)$
  $Cov(\bar{X}_n, S_n^2) = E((\bar{X}_n - \mu)(S_n^2 - \frac{n-1}{n}\sigma^2))$
  $Cov(\bar{X}_n, S_n^2) = E((\frac{1}{n}\sum X_i - \mu)(\frac{1}{n}\sum(X_j - \mu)^2 - (\bar{X}_n - \mu)^2 - \frac{n-1}{n}\sigma^2))$
  $Cov(\bar{X}_n, S_n^2) = E((\frac{1}{n}\sum(X_i - \mu))(\frac{1}{n}\sum(X_j - \mu)^2 - (\bar{X}_n - \mu)^2 - \frac{n-1}{n}\sigma^2))$
  $E(X_i - \mu) = 0 \ \forall i$ y $E(X_i - \mu)(X_j - \mu) = 0 \ \forall (i, j)$
  $Cov(\bar{X}_n, S_n^2) = \frac{1}{n^2}E(\sum(X_i - \mu)^3) - E((\bar{X}_n - \mu)^3)$
  $Cov(\bar{X}_n, S_n^2) = \frac{1}{n^2}E(\sum(X_i - \mu)^3) - \frac{1}{n^3}E(\sum X_i^3)$
  $Cov(\bar{X}_n, S_n^2) = \frac{\mu_3}{n} - \frac{\mu_3}{n^2} = \frac{n-1}{n^2}\mu_3$
  si n $\to +\infty$, $Cov(\bar{X}_n, S_n^2) \to 0$ (lo que no significa que hay independencia).
  En particular si la distribución es simétrica ($\mu_3 = 0$), entonces $Cov(\bar{X}_n, S_n^2) = 0$.

2.4.3 Caso de una distribución normal

$X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. $\Longrightarrow \bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$

$S_n^2 = \frac{1}{n}\sum(X_i - \mu)^2 - (\bar{X}_n - \mu)^2$

$\frac{nS_n^2}{\sigma^2} = \sum(\frac{X_i-\mu}{\sigma})^2 - (\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}})^2$

Como las v.a. $(\frac{X_i-\mu}{\sigma})$ son i.i.d. de una $\mathcal{N}(0,1)$, entonces $U = \sum(\frac{X_i-\mu}{\sigma})^2$ es una suma de los cuadrados de n v.a. independientes de $\mathcal{N}(0,1)$ cuya distribución es fácil de calcular y se llama **Ji-cuadrado con n grados de libertad y se denota** χ_n^2. Por otro lado, $(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}})^2$ sigue una distribución χ^2 con 1 grado de libertad.

En efecto recordemos en primer lugar la distribución de $Y = Z^2$, en que $Z \sim \mathcal{N}(0,1)$.
Sea $\Phi(x)$ la función de distribución de $Z \sim \mathcal{N}(0,1)$ y F(y) la de $Y = Z^2$.
$F(y) = P(Y \leq y) = P(Z^2 \leq y) = P(-\sqrt{y} \leq Z \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$.

Se deduce la función de densidad de Y:

$$f(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} exp(-y/2) \quad \forall y > 0$$

Se dice que Y sigue una distribución Ji-cuadrado con 1 grado de libertad, χ_1^2.

Observando que la χ_1^2 tiene una distribución Gamma particular $\Gamma(1/2, 1/2)$, la función generatriz de momentos (f.g.m.) se escribe:

$$\Psi_Y(t) = E(e^{tY}) = (\frac{1}{1-2t})^{1/2} \quad \forall t < \frac{1}{2}$$

Sea $U = \sum_1^n Y_i = \sum_1^n Z_i^2$ en que las Z_i^2 son χ_1^2 independientes, entonces
$\Psi_U(t) = (\frac{1}{1-2t})^{n/2}$, que es la f.g.m. de una distribución $Gamma(\frac{n}{2}, \frac{1}{2})$.

Se deduce así la función de densidad de U la v.a. χ_n^2, una Ji-cuadrado con n g.l.:

$$f(u) = \frac{1}{2^{n/2}} \frac{u^{n/2-1}}{\Gamma(n/2)} exp(-u/2) \quad \forall u > 0$$

Se observa que $E(U) = n$ y $Var(U) = 2n$ y se tiene el siguiente resultado:

Corolario 2.1 *La suma de k v.a. independientes y de distribución* χ^2 *a* r_1, r_2,...,r_k *g.l. respectivamente sigue una distribución* χ^2 *a* $r_1 + r_2 + ... + r_k$ *g.l.*

Aplicamos estos resultados al cálculo de la distribución de S_n^2 cuando $X \sim \mathcal{N}(\mu, \sigma^2)$

Teorema 2.4 *Si* $X_1, X_2, ..., X_n$ *son i.i.d. de la* $\mathcal{N}(\mu, \sigma^2)$, *entonces la v.a.* nS_n^2/σ^2 *sigue una distribución* χ_{n-1}^2

Demostración: Sea $\underline{X}$ el vector de las n v.a. y una transformación ortogonal $\underline{Y} = B\underline{X}$ tal que la primera fila de B es igual a $(1/\sqrt{n}, ..., 1/\sqrt{n})$. Se tiene entonces que:

- $Y_1 = \sqrt{n}\bar{X}_n$
- $\sum Y_i^2 = \sum X_i^2 = \sum (X_i - \bar{X}_n)^2 + n\bar{X}_n^2$
  $Y_2^2 + ... + Y_n^2 = nS_n^2$
- $(Y_1 - \sqrt{n}\mu)^2 + Y_2^2 + ... + Y_n^2 = (X_1 - \mu)^2 + ... + (X_n - \mu)^2$

La densidad conjunta de $Y_1, ..., Y_n$ es entonces proporcional a:

$$exp\{-(y_1 - \mu\sqrt{n})^2 + Y_2^2 + ... + Y_n^2\}/2\sigma^2$$

Luego $Y_1^2, ..., Y_n^2$ son independientes y

$$\sqrt{n}\bar{X}_n = Y_1 \sim \mathcal{N}(\sqrt{n}\mu, \sigma^2)$$

$$nS_n^2/\sigma^2 = Y_2^2 + ... + Y_n^2\}/\sigma^2 \sim \chi_{n-1}^2$$

Además $\bar{X}_n$ y S_n^2 son independientes.

Teorema 2.5 *Sean $X_1, X_2, ..., X_n$ v.a. i.i.d., entonces $\bar{X}_n$ y S_n^2 son independientes si y sólo si las X_i provienen de una distribución normal.*

La demostración se deduce del teorema 2.4 y del corolario 2.1.

Definemos a continuación la distribución t de Student (Student es un seudónimo utilizado por el estadístico inglés W. S. Gosset para publicar), que tiene muchas aplicaciones en inferencia estadística como la distribución χ^2.

Definición 2.2 *Si X e Y son dos v.a. independientes, $X \sim \mathcal{N}(0,1)$ e $Y \sim \chi_n^2$, entonces la v.a. $T = \frac{X}{\sqrt{\frac{Y}{n}}}$ tiene una distribución t de Student a n grados de libertad.*

Buscamos la función de densidad de la v.a. T. Si $f(x,y)$ es la densidad conjunta de (X,Y) y $f_1(x)$ y $f_2(y)$ las densidades marginales de X e Y respectivamente, entonces $f(x,y) = f_1(x)f_2(y)$.

$$f_1(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2}) \quad \forall x \in I\!R$$

$$f_2(y) = \frac{1}{2^{n/2}} \frac{y^{n/2-1}}{\Gamma(n/2)} exp(-y/2) \quad \forall y > 0$$

El jacobiano del cambio de variables $X = T\sqrt{W/n}$ e $Y = W$ es $J = \sqrt{W/n}$. Deducimos la densidad conjunta de (T, W):

$$g(t,w) = \sqrt{\frac{w}{n}} \frac{e^{-\frac{t^2 w}{2n}}}{\sqrt{2\pi}} \frac{w^{\frac{n}{2}-1} e^{-\frac{w}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \quad \forall w > 0, \quad -\infty < t < \infty$$

$$g(t,w) = \frac{w^{\frac{n-1}{2}} e^{-\frac{1}{2}(1+\frac{t^2}{n})w}}{\sqrt{2^{n+1}\pi n}\Gamma(\frac{n}{2})} \quad \forall w > 0, \quad -\infty < t < \infty$$

$$h(t) = \frac{\Gamma(\frac{n+1}{2})(1+\frac{x^2}{n})^{-(\frac{n+1}{2})}}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \quad t \in I\!R$$

Se observa que la función de densidad de T es simétrica y $E(T) = 0$ y $var(T) = \frac{n}{n-1}$ para $n \geq 2$. Ademas para n=1 se tiene la distribución de Cauchy y para n grande se puede aproximar la distribución de T a una $\mathcal{N}(0,1)$.

Aplicando estos resultados, deducimos que la distribución de la v.a.

$$V = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/(n-1)}}$$

es una t de Student con n-1 grados de libertad.

2.4.4 Valores extremos

Es importante estudiar entre que valores podrian estar los valores muestrales.

Si $X_{(1)}, ..., X_{(n)}$ los estadísticos de orden (los valores muestrales ordenados de menor a mayor: $X_{(1)} \leq X_{(2)} ... \leq X_{(n)}$) entonces $X_{(1)} = inf\{X_1, ..., X_n\}$ y $X_{(n)} = sup\{X_1, ..., X_n\}$.

En el curso de Probabilidades se estudio las distribuciones de estos estadísticos de orden en función de la distribución de población F(x) de X. En particular:

- La distribución de $X_{(1)}$ es $1 - (1 - F(x))^n$
- La distribución de $X_{(n)}$ es $(F(x))^n$

El rango $W = X_{(n)} - X_{(1)}$ es otro estadístico interesante a estudiar.

2.4.5 Cuantilas

Definición 2.3 *Dada una función de distribución F(x) de X, se llama cuantila de orden p al valor x_p tal que $F(x_p) = p$.*

Si tomamos p $= 1/2$, entonces $x_{1/2}$ es tal que hay tantos valores por debajo que por arriba de $x_{1/2}$, que se llama **mediana** de la distribución. Se llaman **cuartilas** a $x_{1/4}$ y $x_{3/4}$ y **intervalo intercuartila** a $x_{3/4} - x_{1/4}$.

Se observara que para una distribución discreta o empírica F_n una cuantila para un p dado no es única. Se define entonces como x_p al valor tal que $I\!P(X < x_p) \leq p \leq I\!P(X \leq x_p)$.

3 ESTIMACION PUNTUAL

3.1 INTRODUCCION

En un problema estadístico, si los datos fueron generados a partir de una distribución de probabilidad F(x) desconocida, los métodos de la **Inferencia Estadística** permite decir algo respecto de esta distribución. Cuando se supone que tal distribución no es totalmente desconocida - por ejemplo pertenece a una determinada familia de distribuciones - entonces son desconocidos sólo uno o varios **parámetros** que definen cada distribución de esta familia. En este caso la teoría de estimación tiene por objetivo dar valores a estos parámetros a partir de los valores muestrales.

Por ejemplo, $F(x)$ pertenece a la familia de las distribuciones normales $\mathcal{N}(\mu, 1)$ de varianza igual a 1 y de esperanza μ desconocida. Aquí μ es el único parámetro desconocido de la distribución. Pero si se supone la varianza también desconocida, se tendrán dos parámetros desconocidos, la media μ y la varianza σ^2.

Los parámetros son constantes que toman valores en un espacio llamado **espacio de parámetros** Θ:

$\mathcal{N}(\mu, 1)$	$\Theta = I\!R$
$\mathcal{N}(\mu, \sigma)$	$\Theta = I\!R \times]0, +\infty[$
$Exp(\beta)$	$\Theta =]0, +\infty[$
Binomial(10,p)	$\Theta = [0, 1]$

Sean $X_1, ..., X_n$ los valores muestrales obtenidos sobre una muestra aleatoria simple de una v.a. X de función de densidad $f(x/\theta)$, en que θ es desconocido. Hay varias maneras de decir algo sobre θ. Lo más simple consiste en dar un valor único para θ. Es **la estimación puntual**: se busca elegir un valor para θ a partir de los valores muestrales. Es decir se tiene que definir una función $\delta : I\!R^n \longrightarrow \Theta$, que es un estadístico llamado **estimador** de θ. El valor tomado por esta función sobre una muestra particular de tamaño n es una **estimación**. Otra forma de estimar un parámetro consiste en buscar no un sólo valor para θ, sino un conjunto de valores, un intervalo en general, en el cual se tiene alta probabilidad de encontrar θ. Es la **estimación por intervalo**.

Procediendo así, tratamos de **estimar el valor de los parámetros**, que son considerados como constantes, a partir de estadísticos que son aleatorios. Ahora bien, frecuentemente se sabe algo más sobre los parámetros; este conocimiento obviamente no es preciso, sino no se tendría el problema de estimar estos parámetros; pero se tienen ideas sobre sus posibles valores, que pueden ser traducidas a una **función de distribución a priori** sobre el espacio de parámetro Θ. Los estimadores bayesianos toman en cuenta la distribución a priori y los valores muestrales.

El problema es encontrar métodos que permitan construir estos estimadores.
A continuación daremos los métodos usuales de estimación puntual.

3.2 METODO DE LOS MOMENTOS

Vimos en el capítulo anterior que la media muestral $\bar{X}_n \xrightarrow{c.s.} E(X) = \mu$. Más generalmente si el momento $\mu_r = E(X^r)$ existe, entonces por la ley de los grandes números:

$$m_r = \frac{1}{n}\sum X_i^r \xrightarrow{c.s.} \mu_r \qquad (I\!P(lim_{n\to\infty} m_r = \mu_r) = 1)$$

Luego se puede estimar μ_r como $\hat{\mu}_r = m_r$.

Ejemplo: este método produce como estimador de la media μ, $\hat{\mu} = \bar{X}_n$ y como estimador de la varianza $\sigma^2 = m_2 - \bar{X}_n^2 = S_n^2$

3.3 METODO DE MAXIMA VEROSIMILITUD

Sean $x_1, x_2, ..., x_n$ una muestra aleatoria simple de una v.a. de densidad $f(x/\theta)$ en que $\theta \in \Theta$, el espacio de parámetros.

Definición 3.1 *Se llama* **función de verosimilitud** *a la densidad conjunta del vector de los valores muestrales; para todo vector observado* $\underline{x} = (x_1, x_2, ..., x_n)$ *en la muestra, se denota* $f_n(\underline{x}/\theta)$.

Como los valores son independientes, se tiene:

$$f_n(\underline{x}/\theta) = f_n(x_1, x_2, ..., x_n/\theta) = \prod_{i=1}^{n} f(x_i/\theta)$$

Un estimador del parámetro θ basado en una muestra de tamaño n es una función δ de los valores muestrales $(x_1, x_2, ..., x_n)$ a valores en el espacio de parámetro Θ.

El valor que toma el estimador δ sobre una muestra $(x_1, ..., x_n)$ se llama **estimación** o **valor estimado**.

El estimador de Máxima Verosimilitud es el estimador que hace $f_n(\underline{x}/\theta)$ máxima.

Tal estimador puede entonces no ser único, o bien no existir.

3.4 EJEMPLOS

Ejemplo 1: Una máquina produce diariamente un lote de piezas. Un criterio basado sobre normas de calidad vigente permite clasificar cada pieza fabricada como defectuosa o no defectuosa. El cliente aceptara el lote si la proporción de piezas θ defectuosas contenidas en el lote no sobrepasa el valor θ_o. El fabricante tiene que controlar entonces la proporción θ de piezas

defectuosas contenidas en cada lote que fabrica. Pero si la cantidad de piezas N de cada lote es muy grande, no podrá examinar cada una para determinar el valor de θ. El fabricante efectúa entonces el control de calidad de una muestra aleatoria pequeña con n piezas. Se define la v.a. X que toma el valor 1 si la pieza es defectuosa y 0 en el caso contrario. Sean $x_1, x_2, ..., x_n$ los valores obtenidos sobre la muestra.

$$x_i \sim Bernoulli(\theta) \quad (0 \leq \theta \leq 1)$$

$$f_n(\underline{x}/\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\max_{\theta} f_n(\underline{x}/\theta) \Longleftrightarrow \max_{\theta} Log\, f_n(\underline{x}/\theta)$$

$$Log\, f_n(\underline{x}/\theta) = \sum_{i=1}^{n}[x_i Log\theta + (1-x_i)Log(1-\theta)]$$

$$\frac{dLog\, f_n(\underline{x}/\theta)}{d\theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1-\theta} = 0$$

Luego el estimador de máxima verosimilitud (E.M.V.) $\hat{\theta}$ de θ es la proporción de piezas defectuosas observada $\sum x_i/n$.

Ejemplo 2: El ministerio de la salud quiere conocer la talla promedia μ de las mujeres chilenas adultas. Si $X_1, X_2, ..., X_N$ son las tallas de todas las chilenas adultas, $\mu = \sum X_i/N$. Dado el tamaño grande de esta población, se obtiene la talla de una muestra aleatoria de tamaño pequeño n. Sean $x_1, x_2, ..., x_n$.

Se supone que $x_i \sim \mathcal{N}(\mu, \sigma^2)$ con μ y σ^2 desconocidos.

$$f_n(\underline{x}/\theta) = (1/2\pi\sigma^2)^{n/2} exp\{-\sum(x_i - \mu)^2/2\sigma^2\}$$

$Log\, f_n(\underline{x}/\theta)$ es máximo cuando $\mu = \bar{X}_n$ la media muestral y $\sigma^2 = S_n^2$ la varianza muestral.

Notas:
- Si se supone la varianza poblacional σ^2 conocida, el E.M.V. de μ queda igual a la media muestral $\bar{X}_n$.

- Se puede buscar el estimador de la varianza o bien de su raíz σ. El resultado no cambia.

Ejemplo 3: $x_i \sim Uniforme[0, \theta] \qquad \theta > 0$

$$f_n(\underline{x}/\theta) = 1/\theta^n \quad si \quad 0 \leq x_i \leq \theta \quad \forall i$$

Cuando $\theta \geq x_i$ para todo i, $f_n(\underline{x}/\theta)$ es no nulo y es decreciente en θ; luego $f_n(\underline{x}/\theta)$ es máxima para el valor más pequeño de θ que hace $f_n(\underline{x}/\theta)$ no nulo: el E.M.V. de θ es entonces $\hat{\theta} = max\{x_1, x_2..., x_n\}$

El método de los momentos produce un estimador bien diferente. En efecto, como
E(X) $= \theta/2$, el estimador de los momentos es $\tilde{\theta} = 2\bar{X}_n$.

En este ejemplo, una dificultad se presenta cuando se toma el intervalo $]0,\theta[$ abierto, dado que no se puede tomar como estimador el máximo $\hat{\theta}$; en este caso no existe E.M.V. Puede ocurrir que no es único también: si se define el intervalo $[\theta, \theta+1]$, la función de verosimilitud es:

$$f_n(\underline{x}/\theta) = 1 \quad si \quad \theta \leq x_i \leq \theta + 1 \quad \forall i$$

es decir:

$$f_n(\underline{x}/\theta) = 1 \quad si \quad max\{x_1, ..., x_n\} - 1 \leq \theta \leq min\{x_1, ..., x_n\}$$

Por lo cual todo elemento del intervalo $[max\{x_1, ..., x_n\} - 1, min\{x_1, ..., x_n\}]$ es E.M.V.

Aquí el estimador de los momentos, que es igual a $\bar{X}_n - 1/2$, es bien diferente también.

3.5 PROPIEDADES

¿Cómo elegir un estimador? ¿Cómo decidir si un estimador es aceptable? Para ayudarnos en esta elección se puede estudiar si el estimador cumple ciertas propiedades razonables.

3.5.1 Invarianza

Observamos en las notas del ejemplo 2, que el E.M.V. de σ se puede obtener directamente o como la raiz del E.M.V. de σ^2. Eso se debe de la propiedad de **invarianza** del E.M.V. por transformación funcional:

Proposición 3.1 *Si $\hat{\theta}$ es el E.M.V. del parámetro θ, si $g : \Theta \longrightarrow \Theta$ es biyectiva, entonces $g(\hat{\theta})$ es el E.M.V. de $g(\theta)$*

Demostración: en efecto si $\tau = g(\theta)$, como g es biyectiva, $\theta = g^{-1}(\tau)$; si $f_n(\underline{x}/\theta) = f_n(\underline{x}/g^{-1}(\tau))$ es máxima para $\hat{\tau}$ tal que $g^{-1}(\hat{\tau}) = \hat{\theta}$. $\hat{\tau}$ es necesariamente el E.M.V. y como g es biyectiva, $\hat{\tau} = g(\hat{\theta})$.

3.5.2 Consistencia

Un estimador depende del tamaño de la muestra a traves de los valores muestrales; los estimadores $\hat{\theta}_n$ asociados a muestras de tamaño n ($n \in I\!N$) constituyen sucesiones de v.a.. Un buen estimador deberia converger en algún sentido hacia θ.

Definición 3.2 *Se dice que un estimador $\hat{\theta}_n$ de un parámetro θ es* **consistente** *cuando converge en probabilidad hacia θ:*

$$I\!P(|\hat{\theta}_n - \theta| < \epsilon) \stackrel{n\to\infty}{\longrightarrow} 1$$

Los momentos empíricos de una v.a. real son estimadores consistentes de los momentos teóricos correspondientes. Más aún la convergencia es casi-segura y la distribución asintótica de estos estimadores es normal.

3.5.3 Estimador insesgado

Definición 3.3 *Se dice que un estimador $\hat{\theta}$ de θ es* **insegado** *si $E(\hat{\theta}) = \theta$.*

Vimos que la media muestral $\bar{X}_n$ es un estimador insesgado de la media poblacional si la muestra es aleatoria simple, pero la varianza muestral $S_n^2 = 1/n \sum (x_i - x_n)^2$ no es un estimador insesgado para la varianza poblacional σ^2:

$$E(S_n^2) = \frac{n-1}{n}\sigma^2$$

Pero, la diferencia $|E(S_n^2) - \sigma^2| = \sigma^2/n$, que es el sesgo, tiende a cero.

Definición 3.4 *Se dice que el estimador $\hat{\theta}$ es* **asintoticamente insesgado** *cuando $E(\hat{\theta}) \overset{n\to\infty}{\longrightarrow} \theta$.*

Por otro lado se puede construir un estimador insesgado de σ^2 a partir de S_n^2:
$\tilde{\sigma}^2 = \sum (x_i - \bar{X}_n)^2/(n-1)$. Pero observamos que $\tilde{\sigma}^2 = (\frac{n}{n-1})^2 \sigma^2$, es decir que el estimador insesgado $\tilde{\sigma}^2$ tiene mayor varianza que S_n^2.

Por otro lado observamos que si $\hat{\theta}_n^2$ es un estimador sesgado de θ, se tiene:

$$E(\hat{\theta}_n - \theta)^2 = Var(\hat{\theta}_n) + (sesgo)^2$$

En efecto,

$$E(\hat{\theta}_n - \theta)^2 = E[(\hat{\theta}_n - E(\hat{\theta}_n) + E(\hat{\theta}_n) - \theta)^2]$$
$$E(\hat{\theta}_n - \theta)^2 = E[(\hat{\theta}_n - E(\hat{\theta}_n))^2] + [E(\hat{\theta}_n) - \theta)]^2$$

Si $[E(\hat{\theta}_n) - \theta)]^2 \longrightarrow 0$ entonces $\hat{\theta}_n$ converge en media cuadrática hacia θ. $(\hat{\theta}_n \overset{m.c.}{\longrightarrow} \theta)$.

Proposición 3.2

$$E(\hat{\theta}_n - \theta)^2 \longrightarrow 0 \Longleftrightarrow Var(\hat{\theta}_n) \to 0 \quad y \quad E(\hat{\theta}_n) \to \theta$$

Como la convergencia en media cuadrática implica la convergencia en probabilidad se tiene:

Proposición 3.3 *Si $\hat{\theta}_n$ es un estimador consistente de θ y $E(\hat{\theta}_n)$ es finito entonces $\hat{\theta}_n$ es asintoticamente insesgado.*

Proposición 3.4 *Si $Var(\hat{\theta}_n) \to 0 \quad y \quad E(\hat{\theta}_n) \to \theta$, entonces $\hat{\theta}_n$ es un estimador consistente de θ.*

Nota: Es una condición suficiente pero no necesaria.

3.5.4 Suficiencia

En el ejemplo 1, se busca deducir de las observaciones de una muestra aleatoria de n piezas una información sobre la proporción θ de piezas defectuosas en el lote total. Es más simple considerar el número de piezas defectuosas encontradas en la muestra en vez de la sucesión de resultados $x_1, x_2, ..., x_n$. El conocimiento de los valores individuales no procura ninguna información aditiva para la proporción θ que $\sum_{i=1}^{n} x_i$. Se redujo los n datos a un sólo valor, que es función de estos datos, sin perder información para determinar θ.

En el ejemplo 2, la media muestral $\bar{X}_n$ permite simplificar la información dada por los n valores muestrales. Pero nos preguntamos si se pierde información usando la media muestral para estimar la media μ de la población.

Observamos que si suponemos la varianza conocida, la función de verosimilitud puede escribirse como función unicamente de la media muestral y del tamaño n de la muestra:

$$f_n(\underline{x}/\theta) = (1/\sqrt{2\pi})^n exp\{-n(\bar{X}_n - \theta)^2/2\}$$

Es decir que la única información relevante para estimar θ es dada por la media muestral. En este caso se dice que la media muestral es un estadístico suficiente. Un estadístico suficiente que se toma como estimador del parámetro θ, debería contener toda la información que llevan los valores muestrales sobre θ.

Definición 3.5 *Un estadístico $T(x_1, ..., x_n)$, función de los valores muestrales y con valor en Θ se dice* **suficiente** *para θ si la distribución conjunta de los valores muestrales condicionalmente a $T(x_1, ..., x_n)$ no depende de θ.*

Definición 3.6 *Se dice que un estadístico T es suficiente minimal si no se puede encontrar otro estadístico suficiente que hace una mejor reducción de los datos que T.*

No es siempre fácil detectar si un estadístico es suficiente. Los dos siguientes teoremas permiten enunciar condiciones para que un estadístico sea suficiente.

Teorema 3.1 *Teorema de factorización*
Si $T(\underline{x})$ es suficiente para θ y $g(T(\underline{x})/\theta)$ es la densidad de $T(\underline{x})$, entonces

$$f_n(\underline{x}/\theta) = g(T(\underline{x})/\theta)h(\underline{x}/T(\underline{x}))$$

Teorema 3.2 *Theorema de Darmois-Koopman*
Si X es una variable real cuyo dominio de variación no depende del parámetro θ, una condición necesaria y suficiente para que existe un estadístico suficiente es que la función de densidad de X sea de la forma:

$$f(x, \theta) = b(x)c(\theta)exp\{a(x)q(\theta)\}$$

$T_n(X) = \sum_{i=1}^{n} a(X_i)$ *es un estadístico suficiente minimal.*

Si $X \sim \mathcal{N}(\theta, 1)$ y una muestra aleatoria es $x_1, ..., x_n$ de X,

$$f_n(x_1, ..., x_n/\theta) = \frac{1}{(2\pi)^{n/2}} exp(-\frac{1}{2}\sum x_i^2) exp(-\frac{n\theta^2}{2} + n\theta\bar{X})$$

El término $exp(-\frac{1}{2}\sum x_i^2)$ no depende de θ y el término $exp(-\frac{n\theta^2}{2} + n\theta\bar{x_n})$ depende de θ y $\bar{X}_n$.

$n\bar{X} = \sum x_i$ es un estadístico suficiente; también toda función biyectiva de $\bar{X}_n$ lo es, en particular $\bar{X}_n$.

3.6 ESTIMADORES BAYESIANOS

3.6.1 Distribuciones a priori

En el problema de estimación de un parámetro de una distribución de función de densidad $f(x/\theta)$, es frecuente tener algunas ideas sobre los valores que puede tomar θ; en este caso conviene tomar en cuenta este conocimiento o **creencia** que se puede traducir en una distribución de probabilidad sobre el espacio de parámetros Θ, sea $\pi(\theta)$. Es decir que ahora θ ya no es un parámetro constante, sino una variable aleatoria. Esta distribución no depende de los valores muestrales. Está definida previo al muestreo.

Por ejemplo, en un proceso de fabricación se tiene la proporción θ desconocida de piezas defectuosas. Si no se sabe nada respecto a θ, se puede suponer que todos los valores son equiprobables: $\theta \sim \mathcal{U}(0, 1)$. Pero uno puede sopechar que los valores alrededor de 0.10 son más probables; en este caso se podrá tomar una distribución más concentrada en 0.10.

Definición 3.7 *Se llama* **distribución a priori** *a la distribución atribuida a un parámetro poblacional, antes de tomar alguna muestra.*

3.6.2 Distribuciones a posteriori

Ahora hay que relacionar los valores muestrales con la distribución a priori $\pi(\theta)$.

La función de verosimilitud $f_n(\underline{x}/\theta)$ es ahora una densidad condicional y $h(\underline{x}, \theta) = f_n(\underline{x}/\theta)\pi(\theta)$ es la densidad conjunta de $(\underline{x}, \theta)$. De la cual se puede deducir la distribución condicional de θ dado los valores muestrales $\underline{x}$:

Definición 3.8 *La distribución condicional de θ dada la muestra $(x_1, ..., x_n)$ se llama* **distribución a posteriori** *y su densidad es igual a $\xi(\theta/\underline{x}) = \frac{f_n(\underline{x}/\theta)\pi(\theta)}{g_n(\underline{x})}$, en que $g_n(\underline{x}) = \int_\Theta h(\underline{x}, \theta)d\theta$ es la densidad marginal de $\underline{x}$.*

La distribución a posteriori representa la actualización de la información a priori $\pi(\theta)$ en vista de la información contenida en los valores muestrales, $f_n(\underline{x}/\theta)$. Podemos entonces estudiar esta distribución a posteriori de θ dando la moda, la media, la mediana, la varianza, etc. Un estimador natural en este caso es tomar la moda de $\xi(\theta/\underline{x})$, que aparece como el máximo de la verosimilitud corregida.

Ejemplo 4: Sean $X \sim Bernoulli(p)$ y $p \sim \beta eta(\alpha, \beta)$, con α y β dados.

$$f_n(\underline{x}/p) = p^{n\bar{X}_n}(1-p)^{n-n\bar{X}_n}$$

$$\pi(p) = p^{\alpha-1}(1-p)^{\beta-1}/B(\alpha,\beta) \quad 0 \leq p \leq 1$$

en que $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
La densidad a posteriori de p es entonces:

$$\xi(p/\underline{x}) = p^{\alpha+n\bar{X}_n-1}(1-p)^{\beta+n-n\bar{X}_n-1}/B(\alpha+n\bar{X}_n, \beta+n-n\bar{X}_n)$$

que es la distribución $\beta eta(\alpha+n\bar{X}_n, \beta+n-n\bar{X}_n)$. La moda de esta distribución, cuando está definida, es igual a $(\alpha-1+n\bar{X}_n)/(\alpha+\beta+n)$.

Ejemplo 5: Sean $X \sim \mathcal{N}(\theta, 1)$ y $\theta \sim \mathcal{N}(0, 10)$.
$\xi(\theta/\underline{x}) \propto f_n(\underline{x}/\theta)\pi(\theta)$ ($\propto$ se refiere a la proporcionalidad con respecto a θ).

$\xi(\theta/\underline{x}) \propto exp(-\frac{\sum(x_i-\theta)^2}{2} - \frac{\theta^2}{20})$

$\xi(\theta/\underline{x}) \propto exp(-\frac{-11\theta^2}{20} + n\theta\bar{X}_n)$

$\xi(\theta/\underline{x}) \propto exp(-\frac{-11}{20}(\theta - (10n\bar{X}_n/11))^2)$

La distribución a posteriori de θ es entonces $\mathcal{N}(\frac{10}{11}n\bar{X}_n, \frac{10}{11})$. La moda de la distribución es la media $\frac{10}{11}n\bar{X}_n$.

3.6.3 Funciones de pérdida

Los métodos de estimación propuestos hasta ahora no toman en cuenta un aspecto importante del problema, que son las consecuencias de tales estimaciones.

Dado que los estimadores son la base de una decisión final, es importante poder comparar los procedimientos que conducen a estas decisiones mediente algún criterio de evaluación, que mide las consecuencias de cada estimación en función de los valores del parámetro θ.

Definición 3.9 *Se llama* **función de pérdida** *o* **función de costo** *a la función* L: $\Theta \times \Theta \longrightarrow [0, +\infty[$, *en que* $L(\theta, \delta)$ *es creciente con el error entre el parámetro* θ *y su estimador* δ.

No es siempre fácil definir esta función de pérdida, que es específica de cada problema y puede tener algún aspecto subjectivo (noción de utilidad). Sin embargo, se puede elegir entre diversas funciones de pérdida clásicas, cuando no se puede construir una propia:

- Función de pérdida cuadrática
  Es la función de pérdida más utilizada y más criticada:

$$L(\theta, \delta) = (\theta - \delta)^2$$

  que penaliza demasiado los errores grandes.

- Función de pérdida absoluta
  Una solución alternativa a la función cuadrádica es usar el valor absoluto:

$$L(\theta, \delta) = |\theta - \delta|$$

  o bien una función afín por parte:

$$L(\theta, \delta) = \begin{cases} k_1(\theta - \delta) & \text{si } \theta > \delta \\ k_2(\delta - \theta) & \text{si no} \end{cases}$$

- Función de perdida "0-1"
  Sea $I_\varepsilon(\delta)$ el intervalo de centro δ y largo 2ε.

$$L(\theta, \delta) = \begin{cases} 0 & \text{si } \theta \in I_\varepsilon(\delta) \\ 1 & \text{si no} \end{cases}$$

3.6.4 Estimadores de Bayes

La función de pérdida $L(\theta, \delta)$ es una función de θ considerada como aleatoria con la distribución a posteriori $\xi(\theta/\underline{x})$. Luego es natural de buscar un estimador $\delta(\underline{x})$ de θ tal que la pérdida promedio sea mínima.

Definición 3.10 *El estimador de Bayes es solución de* $\min_\delta E(L(\theta, \delta)/\underline{x})$

- Función de pérdida cuadrática
  Para la función de pérdida cuadrática $L(\theta, \delta) = (\theta - \delta)^2$, el estimador de Bayes es simple de encontrar: $E((\theta - \delta)^2/\underline{x})$ es mńimo para $\delta(\underline{x}) = E(\theta/\underline{x})$.

- Función de pérdida absoluta
  Para la función de pérdida absoluta $L(\theta, \delta) = |\theta - \delta|$, el estimador de Bayes es la mediana de la distribución a posteriori. Mostramos un resultado más general:

  Proposición 3.5 *El estimador de Bayes asociado a la distribución a posteriori ξ y a la función de perdida*

  $$L(\theta, \delta) = \begin{cases} k_1(\theta - \delta) & si\ \theta > \delta \\ k_2(\delta - \theta) & si\ no \end{cases}$$

  es la fractila $\frac{k_1}{k_1+k_2}$ de ξ.

  Demostración: Se tiene

  $$E[L(\theta, \delta)/\underline{x}] = k_2 \int_{-\infty}^{\delta} (\delta - \theta)\xi(\theta/\underline{x})d\theta + k_1 \int_{\delta}^{+\infty} (\theta - \delta)\xi(\theta/\underline{x})d\theta$$

  Derivando con respecto a δ, se obtiene:

  $$k_2 I\!P(\theta < \delta/\underline{x}) - k_1 I\!P(\theta > \delta/\underline{x}) = 0$$

  Es decir:

  $$I\!P(\theta < \delta/\underline{x}) = \frac{k_1}{k_1 + k_2}$$

  En particular si $k_1 = k_2$, se obtiene la mediana de la distribución a posteriori de θ.

- Función de pérdida "0-1"
  $E[L(\theta, \delta)]$ es mínimo cuando $\int_{I_\varepsilon(\delta)} \xi(\theta/\underline{x})d\theta$ es máximo. Si $\varepsilon \longrightarrow 0$, entonces $E[L(\theta, \delta)]$ es mínimo cuando $\xi(\theta/\underline{x})$ es máximo. El estimador de Bayes es la moda de $\xi(\theta/\underline{x})$.

Teorema 3.3 *Theorema de Rao-Blackwell*
Si T(X) es un estadístico suficiente para θ y si b(X) es un estimador insesgado de θ, entonces

$$\delta(T) = E(b(X)/T)$$

es un estimador insesgado de θ basado sobre T mejor que b(X).

Este teorema permite entonces construir estimadores insesgados mejores.

3.6.5 Estimadores de Bayes para muestras grandes

Se muestra aquí, a través de un ejemplo, los efectos de la distribución a priori y de la función de pérdida sobre el estimador de Bayes, para muestras grandes. Sea θ la proporción de defectuosos. Tomamos dos distribuciones a priori y dos funciones de pérdida:

$\pi(\theta) = 1$ para $\theta \in [0,1]$ y $\pi^{'}(\theta) = 2(1-\theta)$ para $\theta \in [0,1]$
$L(\theta,\delta) = (\theta-\delta)^2$ y $L^{'}(\theta,\delta) = |\theta-\delta|$. Las distribuciones a posteriori son respectivamente

$$\xi(\theta/\underline{x}) \propto \theta^{n\bar{X}_n}(1-\theta)^{n-n\bar{X}_n}$$

que es una $\beta eta(1+n\bar{X}_n, n+1-n\bar{X}_n)$ y

$$\xi^{'}(\theta/\underline{x}) \propto \theta^{n\bar{X}_n}(1-\theta)^{n+1-n\bar{X}_n}$$

que es una $\beta eta(1+n\bar{X}_n, n+2-n\bar{X}_n)$.
Los estimadores de Bayes para la pérdida cuadrática son las respectivas esperanzas de la distribución βeta:
$\delta = (1+n\bar{X}_n)/(n+2)$ para ξ y $\delta^{'} = (1+n\bar{X}_n)/(n+3)$ para $\xi^{'}$.
Los estimadores de Bayes para la pérdida absoluta son las respectivas medianas de la distribución βeta, que se obtienen resolviendo la ecuación:

$$K\int_0^{\delta}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = 1/2$$

en que $\alpha = 1+n\bar{X}_n$ y $\beta = n+1-n\bar{X}_n$ para ξ y $\beta = n+2-n\bar{X}_n$ para $\xi^{'}$.
Si n=100 y $n\bar{X}_n = 10$ entonces $\delta = 11/102 = 0.108$ y $\delta^{'} = 11/103 = 0.107$ para la pérdida cuadrática. Se observara cómo la muestra corrige la distribución a priori, con las medias a priori $E(\theta) = 1/2$ con ξ y $E(\theta) = 1/3$ con $\xi^{'}$.
Encontramos ambos estimadores de Bayes a posteriori muy cercanos con n=100 y cercanos de la media muestral $\bar{X}_n = 10/100 = 0.100$.

En este ejemplo observamos que el estimador de Bayes cuadrático es consistente. No se puede siempre asegurar que el estimador de Bayes es consistente, pero bajo condiciones bastante generales es cierto.

3.7 EJERCICIOS

1. Sea X_i, $i = 1,...,n$ una muestra aleatoria simple de una v.a. X de función de distribución Gamma(α,β).
Estime E(X) por Máxima Verosimilitud. Muestre que el estimador resultante es insesgado, convergente en media cuadrática y es consistente.

2. Sea una m.a.s. $x_1,...x_n$ de una v.a. X de función de densidad $f(x/\theta) = \theta x^{\theta-1}\boldsymbol{I}_{[0,1]}$.
Encuentre el estimador de Máxima Verosimilitud $\hat{\theta}$ de θ y pruebe que $\hat{\theta}$ es consistente y asintóticamente insesgado.

3. Sea Y una v.a. de Bernoulli de parámetro θ. Considere una m.a.s. $y_1,...y_n$ y una distribución a priori Beta(a,b) para θ. Obtenga el estimador de Bayes, $\hat{\theta}$ para θ, usando una función de pérdida cuadrática. Muestre que $\hat{\theta}$ es sesgado, asintóticamente insesgado, convergente en media cuadrática y consistente.

4.. Sean dos preguntas complementarias: Q="vota por Pedro" y Q'="no vota por Pedro". Se obtiene una m.a.s. de n personas que contestan a la pregunta Q o Q'; lo unico que se sabe es que cada persona ha contestado a Q con probabilidad θ conocida y Q' con probabilidad $(1-\theta)$. Se definen:
p: la probabilidad que una persona contesta "SI" a la pregunta (Q o Q')
π: la proporción desconocida de votos para Pedro en la población.

a) Dé la proporción π en función de p y θ.
b) Dé el estimador de Máxima Verosimilitud de p y deduzca un estimador $\hat{\pi}$ para π. Calcule la esperanza y la varianza de $\hat{\pi}$.
c) Estudie las propiedades de $\hat{\pi}$; estudie en particular la varianza $\hat{\pi}$ cuando $\theta = 0.5$.

5. Suponga que X tiene una función de densidad $f(x/\theta)$ y que $T(\underline{X})$ es un estimador de Bayes insesgado para θ con la función de pérdida cuadrática y una distribución a priori $\pi(\theta)$.

a) Demuestre que $E(\theta - T(\underline{X}))^2 = 0$
b) Asuma que $f(x/\theta)$ es una $\mathcal{N}(\theta, 1)$. Pruebe que $E(\theta - \bar{X}_n)^2 = \frac{1}{n}$. Concluya si $\bar{X}_n$ puede ser un estimador de Bayes para pérdida cuadrática.

6. Sea $x_1, x_2, ..., x_n$ una m.a.s. de una distribución tal que $I\!P(x_i \in [a,b]) = \theta$.

Se define $y_i = \begin{cases} 1 & \text{si } x_i \in [a,b] \\ 0 & \text{en caso contrario} \end{cases}$

a) Dé la distribución de y_i.
b) Dé el estimador de máxima verosimilitud $\hat{\theta}$ de θ.
c) Dé la esperanza y la varianza de $\hat{\theta}$.
d) Sean las distribuciones a priori de θ:

$\pi_1(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ (Distribución $Beta(\alpha,\beta)$) y $\pi_2(\theta) = 2(1-\theta)$

Dé los estimadores de Bayes y sus varianzas cuando se usa una función de pérdida cuadrática.
e) Aplicación numérica: dé las soluciones a las preguntas anteriores con los valores: n=10, $\alpha = 2$, $\beta = 2$; x_i: 1.2, 3.5, 2.4, 1.5, 6.3, 2.8, 4.2, 4.5, 3.8, 5.1 y [a,b]=[2,4].

7. Sea $\{X_1, X_2, ..., X_n\}$ una m.a.s. de una v.a. X con función de densidad $f(x/\theta)$. Sea $Y = \delta(X_1, ..., X_n)$ un estimador de θ. Se define Y_i el estimador δ calculado sobre la muestra salvo la observación i $(i = 1, 2, ..., n)$, $Y_i^* = nY - (n-1)Y_i$ y $Y^* = (1/n)\sum_i^n Y_i^*$.

a) Calcule la varianza S^{*2} de Y^* cuando $Y = \bar{X}_n$ la media muestral y $E(X) = \theta$.
b) Deducir la distribución de $(Y^* - \theta)/S^*$ cuando $Y = \bar{X}_n$ y $X \sim \mathcal{N}(\theta, \sigma^2)$.

8. Sea X una v.a. real con densidad $f(x/\theta)$, $\theta \in \Theta = \{\theta_1, \theta_2, ..., \theta_N\}$ (finito).
Sean π una distribución de probabilidad a priori sobre Θ y la función de pérdida:

$$L(\theta, \delta) = \begin{cases} 0 & \text{si } \theta = \delta \\ c & \text{si } \theta \neq \delta (c > 0) \end{cases}$$

a) Pruebe que la pérdida esperada se escribe como $E(L(\theta,\delta)) = c(1-\xi(\delta/x))$, en donde ξ es la distribución a posteriori sobre Θ.
b) Deduzca la condición que debe satisfacer δ para ser el estimador de Bayes de θ asociado a π. Pruebe que el estimador no depende de c.
c) Si π es la distribución uniforme sobre Θ, pruebe que el estimador de Bayes de θ y el estimador de máxima verosimilitud coinciden.

9. Se considera la distribución discreta: $I\!P(X=x) = a_x\theta^x/h(\theta)$, con $x = 0, 1, 2, ...$, en donde h es diferenciable y a_x puede ser nulo para algunos x.
Sea $\{x_1, x_2, ..., x_n\}$ una m.a.s. de esta distribución.

a) Dé las expresiones de $h(\theta)$ y $h'(\theta)$.
b) Dé el estimador de máxima verosimilitud de θ en función de h y h'.
c) Muestre que el estimador de máxima verosimilitud es el mismo que el del método de los momentos.
d) Aplique lo anterior para los casos siguientes:
i) $X \sim Binomial(N,p)$ (N conocido)
ii) $X \sim Poisson(\lambda)$.

10. Sean T_i, $i = 1, ..., I$ estimadores del parámetro θ tales que : $E(T_i) = \theta + b_i$, $b_i \in R$
Se define un nuevo estimador T de θ como $T = \sum_{i=1}^{I} \lambda_i T_i$

a) Dé una condición sobre los λ_i para que T sea insesgado.
b) Suponga que $b_i = 0\ \forall i$ (estimadores insesgados). Plantee el problema de encontrar los coeficientes λ_i para que la varianza de T sea mínima.
c) Suponiendo que los T_i son no correlacionados , resuelva el problema planteado antes.
d) Sean X_{ij}, $i = 1 \ldots M$, $j = 1 \ldots n_i$ M m.a.s. independientes entre si, de variables aleatorias X^i con distribuciones normales de varianza común σ^2.
Sea $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$, el estimador insesgado de la varianza calculado en la muestra "i".
Demuestre que $S^2 = \frac{1}{\sum_{i=1}^{M} n_i - M} \sum_{i=1}^{M} (n_i - 1) s_i^2$ es el estimador lineal insesgado de varianza mínima para σ^2.

4 ESTIMACION POR INTERVALO

4.1 INTRODUCCION

Vimos en el capítulo anterior métodos de estimación puntual. Pero no podemos esperar que la estimación que produce coincida exactamente con el verdadero valor del parámetro desconocido θ. Aquí buscamos entonces construir un intervalo $[\theta_1, \theta_2]$ tal que la probabilidad que θ esté en el intervalo sea alta.

Esta probabilidad tiene diferente interpretación según estemos en el caso bayesiano o no. Se tiene entonces dos clases de métodos para construir estos intervalos.

4.2 CASO BAYESIANO

En el bayesiano, el intervalo tiene una interpretación imediata a partir de la distribución a posteriori de θ. Lo único inconviente es la falta de unicidad de tal intervalo. Pero es natural buscar el intervalo de largo mínimo.

Ejemplo: Vimos que si $X \sim Bernoulli(p)$ y $p \sim \beta eta(\alpha, \beta)$, entonces la distribución a posteriori de p es una $\beta eta(\alpha + n\bar{X}_n, \beta + n - n\bar{X}_n)$:

$$\xi(p/\underline{x}) = p^{\alpha+n\bar{X}_n-1}(1-p)^{\beta+n-n\bar{X}_n-1}/B(\alpha+n\bar{X}_n, \beta+n-n\bar{X}_n)$$

Se define entonces un intervalo $[p_1, p_2]$ de probabilidad $1-\alpha$ tal que $I\!P(p_1 \leq p \leq p_2)$, calculada a partir de la distribución ξ.

4.3 INTERVALO DE CONFIANZA DE NEYMANN

En el caso de estimación no bayesiana, el parámetro θ no es una variable aleatoria. En este caso es el intervalo $[\theta_1, \theta_2]$ que es aleatorio, y se habla de la probabilidad de que el parámetro θ cubre el intervalo. Los valores θ_1 y θ_2 son entonces funciones de los valores muestrales.

Sean $X_1, X_2, ..., X_n$ los valores muestrales, se tiene que encontrar dos funciones $\theta_1 = t_1(X_1, X_2, ..., X_n)$ y $\theta_2 = t_2(X_1, X_2, ..., X_n)$ tales que :

$$I\!P(\theta_1 \leq \theta \leq \theta_2) = 1 - \alpha$$

siendo la cantidad $1-\alpha$ fijada a priori y llamada **el nivel de confianza**. Generalmente se determinan las funciones t_1 y t_2 a partir de un estimador de θ.

Ejemplo 1: Intervalo para una media
Sea $X \sim \mathcal{N}(\theta, \sigma^2)$, con la media θ desconocido y la varianza σ^2 conocida y una muestra de tamaño n. Sea $X_1, ...X_n$ los valores muestrales, si $\bar{X}$ es la media muestral, $Z = \frac{\bar{X}-\theta}{\sigma/\sqrt{n}} \sim$

$\mathcal{N}(0,1)$. Si $I\!P(u_1 \leq Z \leq u_2) = 1-\alpha$, $[\bar{X} - u_2\frac{\sigma}{\sqrt{n}}, \bar{X} - u_1\frac{\sigma}{\sqrt{n}}]$ define un intervalo para θ de nivel de confianza $1-\alpha$.

Hay una infinidad de intervalos de mismo nivel de confianza $1-\alpha$. Pero se puede mostrar que el intervalo $[\bar{X}-u, \bar{X}+u]$ simétrico con respecto a $\bar{X}$ tiene el largo mínimo entre los intervalos de mismo nivel de confianza igual a $1-\alpha$. Por ejemplo, para $\alpha = 0.05$, se obtiene el intervalo $[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}]$.

Si no se supone que σ es conocida, se tiene que usar un estadístico cuya distribución muestral no depende de σ. Eso nos lleva a usar el estadístico

$$T = \frac{\bar{X} - \theta}{\sqrt{\sum(X_i - \bar{X})^2/(n-1)}}$$

que sigue una distribución t Student a n-1 g.l..

El estadístico T puede escribirse en función del estimador sesgado $\hat{\sigma}^2$ de σ: $T = \frac{\bar{X}-\theta}{\hat{\sigma}/\sqrt{n}}$.
Si $I\!P(t_1 \leq t \leq t_2) = 1-\alpha$, $[\bar{X} + t_1\hat{\sigma}/\sqrt{n}, \bar{X} + t_2\hat{\sigma}/\sqrt{n}]$ define un intervalo para θ de nivel de confianza $1-\alpha$.

Como en el caso de la distribución normal, el intervalo más corto de nivel de confianza $1-\alpha$ es simétrico con respecto a $\bar{X}$:
$[\bar{X} - t\hat{\sigma}/\sqrt{n}, \bar{X} + t\hat{\sigma}/\sqrt{n}]$ con t tal que $I\!P(-t \leq t_{n-1} \leq t) = 1-\alpha$.

Ejemplo 2: Intervalo para una varianza
Si los valores muestrales $X_1, ..., X_n$ son i.i.d. de la $\mathcal{N}(\theta, \sigma^2)$, $U = \sum(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$.
Un intervalo de nivel de confianza $1-\alpha$ se obtiene a partir de $I\!P(u_1 \leq U \leq u_2) = 1-\alpha$:

$$I\!P\left(\frac{\sum(X_i - \bar{X})^2}{u_2} \leq \sigma^2 \leq \frac{\sum(X_i - \bar{X})^2}{u_1}\right) = 1-\alpha$$

Ejemplo 3: Intervalo para la diferencia de dos medias
Sean dos poblaciones normales $\mathcal{N}(\mu_1, \sigma_1^2)$ y $\mathcal{N}(\mu_2, \sigma_2^2)$. Se consideran una muestra aleatoria de tamaño n_1 de la primera población y una muestra aleatoria de tamaño n_2 de la segunda población, las dos muestras siendo independientes. Si $\bar{X}_1$ y $\bar{X}_2$ son las medias muestrales respectivas, $d = \bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$.

Si las varianzas son conocidas entonces un intervalo para d esta dado por: $[\bar{X}_1 - \bar{X}_2 - u\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + u\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}]$, con u determinado a partir de las tablas de la distribución normal según el nivel de confianza $1-\alpha$.

Si las varianzas no son conocidas, para encontrar un estadístico que nos sirve y cuya distribución no depende de estas varianzas, hay que hacer alguno supuesto suplementario. En efecto si tomamos como estimador de la varianza de la diferencia $\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}$ con $\hat{\sigma}_1^2$ y $\hat{\sigma}_2$ las varianzas muestrales sesgadas, $\frac{n_1\hat{\sigma}_1^2}{\sigma_1^2} + \frac{n_2\hat{\sigma}_2^2}{\sigma_2^2} \sim \chi^2_{n_1+n_2-2}$ y

$$\frac{(\bar{X}_1 - \bar{X}_2 - \mu_1 + \mu_2)/\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}{\sqrt{(\frac{n_1\hat{\sigma}_1^2}{\sigma_1^2} + \frac{n_2\hat{\sigma}_2^2}{\sigma_2^2})/(n_1 + n_2 - 2)}} \sim t_{n_1+n_2-2},$$

que depende de la varianzas desconocidas σ_1^2 y σ_2^2.

Si se supone que estas varianzas son proporcionales: $\sigma_2 = k\sigma_1$, entonces se tiene un estadístico que no depende de σ_1^2 y σ_2^2:

$$\frac{\bar{X}_1 - \bar{X}_2 - \mu_1 - \mu_2}{\sqrt{(\frac{k^2 n_1 \hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{k^2(n_1+n_2-2)})(\frac{k^2 n_1 + n_2}{n_1 n_2})}} \sim t_{n_1+n_2-2}$$

Usualmente si toma $k = 1$.

Ejemplo 4: Intervalo para el cuociente de dos varianzas: la distribución F de Fisher
Sean dos poblaciones normales $\mathcal{N}(\mu_1, \sigma_1^2)$ y $\mathcal{N}(\mu_2, \sigma_2^2)$, nos interesamos al cuociente de las varianzas: $\frac{\sigma_1^2}{\sigma_2^2}$.

El estadístico $n_1\hat{\sigma}_1^2/\sigma_1^2 \sim \chi^2_{n_1-1}$ y el estadístico $n_2\hat{\sigma}_2^2/\sigma_2^2 \sim \chi^2_{n_2-1}$, siendo estos independientes.

Mostramos que si $U \sim \chi_r^2$ y $V \sim \chi_s^2$, y son independientes, entonces $Y = sU/rV$ sigue una distribución de Fisher a r y s grados de libertad con una función de densidad igual a:

$$h(y) = \frac{\Gamma(\frac{r+s}{2})}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \frac{r^{r/2}s^{s/2}y^{(r/2)-1}}{(ry+s)^{(r+s)/2}} \quad \forall y > 0$$

Como U y V son independientes, se puede calcular facilmente la función de densidad conjunta de (U,V):

$$f(u,v) = \frac{u^{(r/2)-1}e^{-u/2}}{2^{r/2}\Gamma(r/2)} \frac{v^{(s/2)-1}e^{-v/2}}{2^{s/2}\Gamma(s/2)}$$

Con el cambio de variables:$(U,V) \longrightarrow (Y,Z)$ con $U = rYZ/s$ y $V = Z$, obtenemos la densidad conjunta de (Y,Z):

$$g(y,z) = \frac{(r/s)z}{2^{(r+s)/2}\Gamma(r/2)\Gamma(s/2)} (r/s)^{(r/2)-1} y^{(r/2)-1} z^{(r+s-1)/2} e^{-1/2(ry/s+1)z}$$

Se deduce la densidad marginal de Y:

$$f(y) = \int_0^\infty g(y,z)dz = \frac{\Gamma(\frac{r+s}{2} r^{r/2}s^{s/2}y^{(r/2)-1}}{\Gamma(r/2)\Gamma(s/2)(ry+s)^{(r+s)/2}}$$

Observamos que si $Y \sim F_{r,s}$ entonces $1/Y \sim F_{s,r}$.

Ejercicio: Muestre que $\frac{rY/s}{1+rW/s} \sim \beta eta((r-2)/2,(s-2)/2)$.

Aquí el estadístico $\frac{n_1\hat{\sigma}_1^2/(n_1-1)\sigma_1^2}{n_2\hat{\sigma}_2^2/(n_2-1)\sigma_2^2} \sim F_{n_1-1,n_2-1}$, lo que permite construir un intervalo de confianza para el cuociente σ_1^2/σ_2^2.

Ejemplo 5: Intervalo para una proporción
Sea la proporción θ de piezas defectuosas en un lote de piezas fabricadas por una industria. El número de piezas defectuosas encontradas en una muestra aleatoria simple de tamaño n sigue una distribución binomial $B(n,\theta)$. Para construir un intervalo de confianza para una proporción es más complicado que para una media o varianza. Cuando n es pequeño hay que recorrer a la distribución binomial (tablas y abacos fueron calculados para determinar valores de θ_1 y θ_2 para los diferentes valores de k y n y del nivel de confianza $1-\alpha$).

Cuando n es grande, se puede usar la aproximación a la distribución normal
$\mathcal{N}(n\theta, n\theta(1-\theta))$, pero la varianza depende también de θ.

Si $\hat{p} = \frac{Y}{n}$, se tiene:

$$I\!P(|\ \frac{\sqrt{n}(\hat{p}-\theta)}{\sqrt{\theta(1-\theta)}}\ | \leq u) = 1-\alpha$$

Lo que equivale a:

$$I\!P(n(\hat{p}-\theta)^2 - u^2\theta(1-\theta) \leq 0) = 1-\alpha$$

Las soluciones de la ecuación:

$$(n+u^2)\theta^2 - (2n\hat{p}+u^2)\theta + n\hat{p}^2 = 0$$

siendo $\frac{2n\hat{p}+u^2 \pm \sqrt{u^4+4n\hat{p}u^2-4nu^2\hat{p}^2}}{2(n+u^2)}$, se obtiene:

$$I\!P(\frac{n}{n+u^2}(\hat{p}+\frac{u^2}{2n}) - u\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{u^2}{4n^2}} \leq \theta \leq \frac{n}{n+u^2}(\hat{p}+\frac{u^2}{2n}) + u\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{u^2}{4n^2}}) = 1-\alpha$$

Para n muy grande, se puede aproximar por:

$$I\!P(\hat{p} - u\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \theta \leq \hat{p} + u\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 1-\alpha$$

4.4 EJERCICIOS

1. Sea una m.a.s. $\{x_1,...x_n\}$ de una distribución normal de media θ desconocida y varianza σ^2 conocida.

a) Dé el número mínimo n del tamaño de la muestra para que un intervalo de confianza I a 95% tenga un largo L a lo más igual a $0.016\ \sigma$.

b) Sea $L = \sigma/5$. Dé el nivel de confianza $1-\alpha$ cuando n=10, 20, 30 y 100.
c) Repetir b) con σ^2 desconocido. Comente.
d) Dé el intervalo de confianza de largo mínimo para θ con un nivel de confianza de 95%, cuando $\sigma^2 = 4$.

2. Una empresa desea estimar el promedio de tiempo que necesita una secretaria para llegar a su trabajo. Se toma una m.a.s. de 36 secretarias y se encuentra que un promedio de 40 minutos. Suponiendo que el tiempo de trayecto proviene de una $\mathcal{N}(\mu, \sigma^2)$, con $\sigma = 12$, dé un intervalo de confianza para la media μ.

3. Se dispone de 10 muestras de sangre tomadas en las mismas condiciones a una misma persona. Se obtiene para cada una la dosis de Colesterol (en gramos) 245, 248, 250, 247, 249, 247, 247, 246, 246, 248. Cada medida puede considerarse como una realización particular de la variable "tasa de Colesterol" $X \sim \mathcal{N}(\mu, \sigma^2)$.

a) Dé un intervalo de confianza para μ al 95% suponiendo $\sigma^2 = 1.5$.
b) Dé un intervalo de confianza para μ al 95% suponiendo σ^2 desconocido.
c) Construya un intervalo de confianza para σ^2 al 95% .

4. En el ejercicio 6 del capitulo 3, muestre que para construir un intervalo de confianza al 95% para θ, en el caso no bayesiano, hay que resolver una inecuación de segundo grado en θ y escriba la inecuación.

5. En el ejercicio 7 del capitulo 3, suponiendo las Y_i^* independientes y n grande, dé un intervalo de confianza para θ a 95%.

6. Se tienen 2 muestras de tamaños n_1 y n_2 de una misma v.a. X medida sobre dos poblaciones distintas. Se asume que para ambas poblaciones X sigue una distribución Normal con medias μ_1, μ_2 y varianzas σ_1^2, σ_2^2, respectivamente.

a) Construya un intervalo de confianza para $\mu_1 - \mu_2$, suponiendo que $\sigma_2^2 = k^2\sigma_1^2$ en que k es una constante conocida.
b) Muestre que los extremos del intervalo anterior convergen en probabilidad si los tamaños de las muestras crecen.
c) Se supone ahora la constante k desconocida. Dé un método para construir un intervalo de confianza para la constante k.
d) ¿ Que inconveniente cree ud. que tiene este método?

7. Se considera una v.a. $X \sim \mathcal{N}(\mu, 1)$ y una m.a.s. de X con una sóla observación x. Dada una constante $a > 0$, se define el intervalo aleatorio: $C_a(x) = [min(0, x-a), max(0, x+a)]$.

a) Muestre que $\mathbb{P}(\mu \in C_a(x)/\mu = 0) = 1 \ \forall x$.
b) Muestre que $C_a(x)$ es un intervalo de confianza para μ de nivel de confianza $1-\alpha = 95\%$, cuando a=1.65.
c) Sea $\pi(\mu) = 1$ $(\forall \mu)$ una distribución a priori para μ. Deducir la distribución a posteriori de μ dado x.
d) Sea Φ la función de distribución de la normal $\mathcal{N}(0,1)$. Muestre que se encuentra una

probabilidad condicional

$$\mathbb{P}(\mu \in C_a(x))/x) = \begin{cases} \Phi(-x) - \Phi(-a) & \text{si } x < -a \\ \Phi(a) - \Phi(-a) & \text{si } -a < x < a \\ \Phi(a) - \Phi(-x) & \text{si } x > a \end{cases}$$

e) Deducir que, para a=1.65, la probabilidad condicional $\mathbb{P}(\mu \in C_a(x)/x) \geq 0.90$ y que $\lim_{a \to \infty} \mathbb{P}(\mu \in C_a(x)/x) = 1$.

5 TESTS DE HIPOTESIS

5.1 GENERALIDADES

En el capítulo 3, se presentaron métodos que permiten encontrar los valores de los parámetros desconocidos de la distribución de población y en el capítulo anterior, la estimación por intervalo permite dar una cierta indicación sobre la **precisión** de la estimación puntual. Tales estimaciones, puntuales y por intervalo, que fueron obtenidas a partir de valores muestrales, permiten formarse una opinión sobre la población y entonces darse una **hipótesis** de trabajo.

Ejemplos:

- Antes de apostar "cara" o "sello" en el lanzamiento de una moneda, se tiene que postular que la moneda está equilibrada. La hipótesis de trabajo es entonces que el parámetro p=probabilidad de sacar "cara" de la Bernoulli es

$$p = 0.5$$

- Un agricultor se compromete a entregar a una fábrica de azúcar remolacha con un cierto porcentaje p_o de glucosa; la hipótesis de trabajo es entonces

$$p = p_o \qquad o \qquad p \geq p_o$$

- Los hombres chilenos pretenden ser más altos que los argentinos en promedio; si μ_1 y μ_2 son las tallas promedias respectivas de los hombres chilenos y argentinos, la hipótesis de trabajo es

$$\mu_1 \geq \mu_2$$

- Cuando se hizo la estimación puntual de la talla promedia μ_1 de los hombres chilenos, se hizo la hipótesis de trabajo que la v.a. X talla de los hombres chilenos sigue una distribución

$$F \sim Normal$$

En los cuatro casos se procedera de la misma manera: se tiene una hipótesis de trabajo y una muestra de observaciones; se trata de decidir si la hipótesis planteada es compatible con lo que se puede aprender del estudio de los valores muestrales. Se tiene que encontrar un procedimiento para decidir si la muestra que se obtuvo esta de acuerdo con la hipótesis de trabajo. Naturalmente no se espera que, para cualquier muestra, el valor empírico obtenido en la muestra coincide con el valor esperado de la hipótesis; el problema es entonces decidir si la desviación encontrada entre el valor esperado y el valor observado en la muestra es demiasiado grande para poner en duda la hipótesis de trabajo. Ahora bien si se pone en duda la hipótesis original, entonces se la rechaza en favor de una **hipótesis alternativa**.

En efecto, en el ejemplo de la moneda, si se encuentra una proporción de 0.45 en 100 lanzamientos, ¿debemos rechazar la hipótesis p=1/2? y si se rechaza, ¿será a favor de la hipótesis $p \leq 1/2$?

Se distingue la hipótesis de trabajo llamandola **hipótesis nula** y una hipótesis nula se confronta a una **hipótesis alternativa**.

> ¿Con qué grado de desacuerdo uno tiene que abandonar la hipótesis nula para la hipótesis alternativa?

Para decidir, se necesita una **regla de decisión**. Cualquier regla de decisión debería tratar de minimizar los errores de decisión. Si δ es la regla de decisión adoptada y $\alpha(\delta)$ la probabilidad de equivocarse cuando la hipótesis nula es cierta y $\beta(\delta)$ la probabilidad de equivocarse cuando la hipótesis alternativa es cierta, uno buscara minimizar ambas probabilidades de error. Pero veremos, a través de un ejemplo, que a tener $\alpha(\delta)$ nula, se hace $\beta(\delta)$ igual a 1 e inversamente.

Dada una hipótesis nula H_o, vimos que $\alpha(\delta)$ es la probabilidad condicional de rechazar la hipótesis H_o con la regla δ cuando H_o es cierta. Ahora bien la regla δ se basa en los valores muestrales; si la muestra es de tamaño n y los valores muestrales en $\mathbb{R}$, una regla de decisión δ consiste en dividir el dominio $\mathbb{R}^n$ del conjunto de todas las muestras de tamaño n en dos partes disjuntas: la parte W en donde se rechaza la hipótesis nula H_o y la parte $\overline{W}$ en donde no se rechaza. La parte W se llama región de rechazo de H_o o **región crítica del test**.

Como la región crítica del test es aquella en donde se rechaza H_o, debería tomar en cuenta la hipótesis alternativa.

Una regla de decisión consiste entonces en determinar la región crítica del test en función de las dos hipótesis.

5.2 HIPOTESIS ESTADISTICAS

Las hipótesis estadísticas son muy precisas: se refieren al comportamiento de variables aleatorias. Pero en los ejemplos expuestos en el parrafo anterior, se observara que las hipótesis no son todas del mismo tipo. En los tres primeros ejemplos, la hipóptesis concierne solamente a los valores de parámetros de una distribución cuya forma no está puesta en duda y es especificada a priori. Tales hipótesis se llaman **hipótesis parámetricas**. En el último ejemplo, es la distribución completa que esta puesta en juicio; se habla de **hipótesis no parámetricas**.

Por ejemplo, sea una v.a. X de distribución $F(x/\theta)$, que depende de un parámetro θ. Si Ω es el espacio del parámetro θ y Ω_o un subconjunto de Ω, entonces

$$H : \theta \in \Omega_o$$

es una hipótesis parámetrica, mientras que

$$H : F \sim Normal$$

es una hipótesis no parámetrica.

Se puede clasificar también las hipótesis parámetricas según su grado de especifidad. Cuando en la hipótesis parámetrica

$$H : \theta \in \Omega_o$$

Ω_o esta reducido a un sólo valor, entonces se habla de **hipótesis simple**, sino se habla de **hipótesis compuesta**.

5.3 TEST DE HIPOTESIS PARAMETRICAS

Trataremos en primer lugar los tests de hipótesis parámetricas para hipótesis simples antes de tratar el caso general apoyandonos en los resultados del caso de las hipótesis simples. Encontrar una regla de decisión es encontrar una región crítica del test. ¿Como hacerlo minimizando los errores de decisión? Para eso usaremos la función de potencia.

5.3.1 Función de potencia

Sea un test de hipótesis sobre el parámetro θ ($\theta \in \Omega$) de la distribución F de una v.a. X.

$$H_o : \theta \in \Omega_o \quad contra \quad H_1 : \theta \in \Omega_1$$

Si una regla de decisión nos condujo a una región crítica W para el test, entonces para cada valor de $\theta \in \Omega$, determinaremos la probabilidad $\pi(\theta)$ que la regla de decisión nos conduce a rechazar H_o cuando el parámetro vale θ.

Definición 5.1 *La función $\pi(\theta) = I\!P(rechazar H_o/\theta)$ se llama FUNCIÓN DE POTENCIA del test.*

¡OJO! aquí θ no es una variables aleatoria.

W es la región crítica del test y $\underline{x}$ el vector de los valores muestrales, entonces

$$\pi(\theta) = I\!P(\underline{x} \in W/\theta) \quad \forall \theta \in \Omega$$

Luego la región crítica ideal es aquella que produce una función de potencia tal que:

$$\pi(\theta) = \begin{cases} 0 & \text{si } \theta \in \Omega_o \\ 1 & \text{si } \theta \in \Omega_1 \end{cases}$$

En efecto, para todo $\theta \in \Omega_o$, la decisión de rechazar H_o es una decisión equivocada, entonces $\pi(\theta)$ es **una probabilidad de error de tipo I** (o riesgo de primer especie). Por otro lado, para todo $\theta \in \Omega_1$, la decisión de rechazar H_o es una decisión correcta, entonces $1 - \pi(\theta)$ es **una probabilidad de error de tipo II** (o riesgo de segundo especie).

Definición 5.2 *Se llama TAMAÑO del test a* $sup\{\pi(\theta)/\theta \in \Omega_o\}$

El problema es que tal región crítica ideal no existe; como lo veremos en el siguiente ejemplo, cuando se disminuye uno de los errores a 0, se aumenta el otro a 1.

Ejemplo: Sea $x_1, x_2, ..., x_n$ una m.a.s. de una v.a. X uniforme en [0,θ] con $\theta > 0$.
Consideramos la hipótesis nula $H_o : 3 \leq \theta \leq 4$ contra la hipótesis alternativa $H_1 : \theta < 3$ o $\theta > 4$. Supongamos que una regla de decisión δ nos llevo a decidir de no rechazar a la hipótesis nula H_o cuando $\max\{x_1, x_2, ..., x_n\}$ de una m.a.s. de la v.a. X esta en el intervalo [2.9,4.1] y a rechazar H_o en el caso contrario. Luego la región crítica del test es un subconjunto $W \subset I\!R^n$ tal que $\max\{x_1, x_2, ..., x_n\} < 2.9$ o $> 4.1\}$. La función de potencia del test es entonces:

$$\pi(\theta) = I\!P(\max\{x_1 + x_2, ..., x_n\} < 2.9/\theta) + I\!P(\max\{x_1, x_2, ..., x_n\} > 4.1/\theta)$$

$$\text{Si } \theta \leq 2.9 \Rightarrow \left\{ \begin{array}{l} I\!P(\max\{x_1, x_2, ..., x_n\} < 2.9/\theta) = 1 \\ I\!P(\max\{x_1, x_2, ..., x_n\} > 4.1/\theta) = 0 \end{array} \right\} \Rightarrow \pi(\theta) = 1$$

$$\text{Si } 2.9 < \theta \leq 4.1 \Rightarrow \left\{ \begin{array}{l} I\!P(\max\{x_1, x_2, ..., x_n\} < 2.9/\theta) = (\frac{2.9}{\theta})^n \\ I\!P(\max\{x_1, x_2, ..., x_n\} > 4.1/\theta) = 0 \end{array} \right\} \Rightarrow \pi(\theta) = (\frac{2.9}{\theta})^n$$

$$\text{Si } \theta > 4.1 \Rightarrow \left\{ \begin{array}{l} I\!P(\max\{x_1, x_2, ..., x_n\} < 2.9/\theta) = (\frac{2.9}{\theta})^n \\ I\!P(\max\{x_1, x_2, ..., x_n\} > 4.1/\theta) = 1 - (\frac{4.1}{\theta})^n \end{array} \right\} \Rightarrow \pi(\theta) = 1 + (\frac{2.9}{\theta})^n - (\frac{4.1}{\theta})^n$$

El tamaño del test es igual a $\alpha = Sup\{\pi(\theta)/3 \leq \theta \leq 4\} = \pi(3) = (\frac{2.9}{3})^n$

En los gráficos 5.1, se muestra la función de potencia para los casos n=10 y 50. Se observa que el tamaño del test $\alpha = 0.10$, es decir que en el intervalo [3, 4] la probabilidad de equivocarse no sobrepasa 10%. Pero el error de tipo II, que es igual a $1 - \pi(\theta)$ cuando $\theta \in \Omega_o$, puede ser muy elevado; entre 3 y 2.9, el error disminuye de 0.10 a 0; pero entre 4 y 4.1 es casi igual a 1.

En este ejemplo si queremos disminuir el tamaño del test α, hay que elegir un intervalo $\overline{W}$ más

grande o una muestra de tamaño mayor. Pero en ambos casos se aumentara el error de tipo II. Para tratar de acercarnos a la situación ideal, se puede, por ejemplo, buscar minimizar una función de los dos errores, o bien fijarse una cota máxima para el error de tipo I y minimizar el error de tipo II.

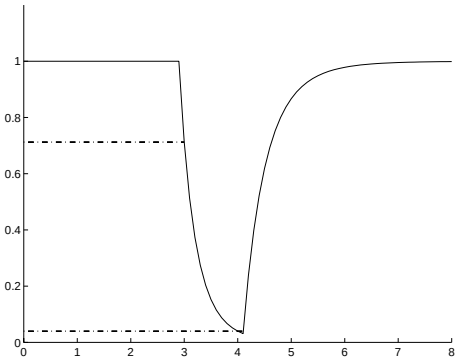


Gráfico 5.1: Función de potencia para la región crítica [2.9,4.1] con n=10

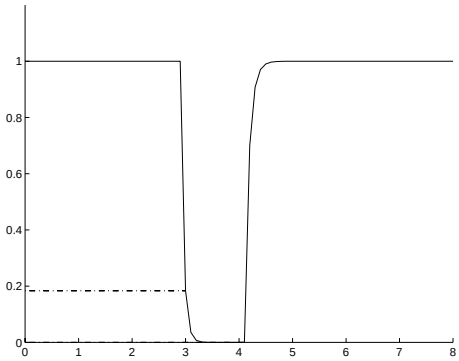


Gráfico 5.2: Función de potencia para la región crítica [2.9,4.1] con n=50

5.3.2 Tests para hipótesis simples

Sean $x_1, x_2, ..., x_n$, los valores muestrales independientes de una v.a. de función de densidad $f(x/\theta)$. Se plantea las hipótesis simples:

$$H_o : \theta = \theta_o \quad contra \quad H_1 : \theta = \theta_1$$

Dada una regla de decisión δ, se tienen los dos errores:

$$\alpha(\delta) = I\!P(rechazar H_o/\theta = \theta_o) \quad (error\ de\ tipo\ I)$$

$$\beta(\delta) = I\!P(no\ \ rechazar H_o/\theta = \theta_1) \quad (error\ de\ tipo\ II)$$

Presentaremos en primer lugar como minimizar una función simple de los dos errores, tomando una función del tipo

$$a\alpha(\delta) + b\beta(\delta)$$

Usaremos la solución anterior para encontrar la forma de construir la región crítica, tal que si uno se fija una cota máxima para el error de tipo I, el error de tipo II sea mínima.

Dados dos escalares a y b, buscamos minimizar la función $a\alpha(\delta) + b\beta(\delta)$. Se denota $f_o(\underline{x})$ y $f_1(\underline{x})$ a las funciones de verosimilitud dado H_o y dado H_1 respectivamente:
$f_o(\underline{x}) = \prod_i^n f(x_i/\theta_o)$ y $f_1(\underline{x}) = \prod_i^n f(x_i/\theta_1)$

Teorema 5.1 *Si δ^* es la regla de decisión tal que:*

$$se\ rechaza\ H_o\ cuando\ af_o(\underline{x}) < bf_1(\underline{x}),$$
$$se\ acepta\ H_o\ cuando\ af_o(\underline{x}) > bf_1(\underline{x}),$$

entonces $a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta) \ \ \forall\delta$

Demostración: Si W es la región crítica asociada a una regla de decisión δ,

$$\alpha(\delta) = \int ... \int_W f_o(\underline{x})dx_1...dx_n$$

$$\beta(\delta) = \int ... \int_{\overline{W}} f_1(\underline{x})dx_1...dx_n$$

$$a\alpha(\delta) + b\beta(\delta) = a\int ... \int_W f_o(\underline{x})dx_1...dx_n + b(1 - \int ... \int_W f_1(\underline{x})dx_1...dx_n)$$

Luego $a\alpha(\delta) + b\beta(\delta)$ es mínimo cuando $\int ... \int_W (af_o(\underline{x}) - bf_1(\underline{x}))dx_1...dx_n$ es mínimo.
Es decir si: $\begin{cases} af_o(\underline{x}) - bf_1(\underline{x}) < 0 \ \ \forall\underline{x} \in W \\ af_o(\underline{x}) - bf_1(\underline{x}) > 0 \ \ \forall\underline{x} \in \overline{W} \end{cases}$

entonces δ^* es óptimo para estos valores a y b dados. Se observará que $f_o(\underline{x}) - bf_1(\underline{x}) = 0$ es irrelevante, dado que no cambia el mínimo.

Definición 5.3 *Se llama RAZÓN DE VEROSIMILITUD de la muestra al cuociente*

$$\frac{f_1(\underline{x})}{f_o(\underline{x})}$$

Sea α_o la cota máxima de error de tipo I que se quiere aceptar.

Definición 5.4 *Se llama NIVEL DE SIGNIFICACIÓN del test a la cota máxima de error de tipo I aceptada.*

Se tiene entonces que buscar una regla de decisión δ que produce un error de tipo I $\alpha(\delta) \leq \alpha_o$ y tal que $\beta(\delta)$ sea mínimo. El siguiente lema, que deriva del teorema anterior, nos da la forma de proceder.

Lema 5.1 *(NEYMAN-PEARSON)*
Si δ^* *es una regla de decisión tal que para algún* $k > 0$ *fijo,*
se rechaza H_o*, si* $\frac{f_1(\underline{x})}{f_0(\underline{x})} > k$
no se rechaza H_o*, si* $\frac{f_1(\underline{x})}{f_0(\underline{x})} < k$*,*
entonces para toda regla δ *tal que* $\alpha(\delta) \leq \alpha(\delta^*)$ *se tiene* $\beta(\delta) \geq \beta(\delta^*)$*.*

Ejemplo: sea $x_1, ..., x_n$ de una muestra aleatoria simple de la v.a. $X \sim \mathcal{N}(\mu, \sigma^2)$, μ desconocido y σ^2 conocido. Se estudia $H_o : \mu = 1$ contre $H_1 : \mu = 2$. La razón de verosimilitud se escribe:

$$\frac{f_1(\underline{x})}{f_o(\underline{x})} = exp\{-\frac{1}{2\sigma^2}[\sum(x_i - 2)^2 - \sum(x_i - 1)^2]\}$$

$$\frac{f_1(\underline{x})}{f_o(\underline{x})} = exp\{-\frac{1}{2\sigma^2}[-2\sum x_i + 3n]\}$$

$$\frac{f_1(\underline{x})}{f_o(\underline{x})} = exp\{\frac{\sum x_i}{\sigma^2} - \frac{3n}{2\sigma^2}\}$$

La regla de decisión que minimiza a $a\alpha(\delta) + b\beta(\delta)$ consiste en rechazar H_o si

$$\frac{f_1(\underline{x})}{f_o(\underline{x})} > \frac{a}{b}$$

es decir: $\bar{X} > \frac{3}{2} + \sigma^2 ln(\frac{a}{b})$

Si $\sigma^2 = 2$ y $n = 20$, la región crítica $\mathcal{R}$, que es de la forma $\{\bar{X} > c\}$ depende de a y b:

si a=b, c=3/2, pero si ($a > b$ y $c > 3/2$) o si ($a < b$ y $c < 3/2$); en particular, si a=2/3 y b=1/3, $\mathcal{R} = \{\bar{X} > 2.88\}$, pero si a=1/3 y b=2/3, $\mathcal{R} = \{\bar{X} > 0.113\}$.

El error de tipo I $\alpha(\delta)$ es $\pi(1) = I\!P(\bar{X} > C/\mu = 1)$. Como $\bar{X} \sim \mathcal{N}(1, \sigma^2/n)$ bajo H_o, $\alpha(\delta) = 1 - \Phi(\frac{c-1}{\sigma/\sqrt{n}})$, en que $\Phi(x)$ es la función de distribución de $\mathcal{N}(0,1)$.

El error de tipo II $\beta(\delta)$ es $1 - \pi(2) = 1 - I\!P(\bar{X} > c/\mu = 1) = I\!P(\bar{X} < c/\mu = 2) = \Phi(\frac{c-2}{\sigma/\sqrt{n}})$

Si a=b, como c=3/2, para n=20, se obtiene $\alpha(\delta) = \beta(\delta) = 1 - \Phi(1.58) = 0.057$, pero con n=100, $\alpha(\delta) = \beta(\delta) = 1 - \Phi(3.53) \simeq 0$.

Si se obtuvo una media muestral $\bar{X} = 1.30$ para una muestra aleatoria de tamaño 20, no se rechaza $H_o : \mu = 1$ con un error de tipo I de 0.057 cuando se toma a=b; si se toma a=0.3 y b=0.7, se rechaza H_o a favor de H_1 con un error de tipo I igual a 0.11.

Si ahora se tiene un nivel de significación fijado a $\alpha_o = 0.05$, entonces se obtiene una región crítica $\mathcal{R} = \{\bar{X} > c\}$ tal que

$$I\!P(\bar{X} > c/\mu = 1) = 0.05$$

Como $\sqrt{n}(\bar{X} - 1) \sim \mathcal{N}(0,1)$

$$I\!P(\bar{X} > c/\mu = 1) = 1 - \Phi(\sqrt{n}(c-1)/\sqrt{2}) = 0.05$$

Como $\Phi(1.65) = 0.95$, se obtiene que $\sqrt{n}(c-1)/\sqrt{2} = 1.65$, es decir que c=1.52 y $\mathcal{R} = \{\bar{X} > 1.52\}$. En este caso no se rechaza H_o.

5.3.3 Tests U.M.P.

Vamos extender ahora los resultados del lema de Neyman-Pearson para hipótesis compuestas.

Sean las hipótesis compuestas $H_o : \theta \in \Omega_o$ contra $H_1 : \theta \in \Omega_1$.

Si nos fijamos un nivel de significación α_o, buscamos una regla de decisión δ tal que la función de potencia cumple:
$\pi(\theta/\delta) \leq \alpha_o \; \forall \theta \in \Omega_o \quad y \quad \pi(\theta/\delta)$ sea máxima $\forall \theta \in \Omega_1$.

Ahora bien no es siempre posible encontrar un test δ que satisfaga esta condición. En efecto si $\Omega = \{\theta_1, \theta_2\}$, un test δ podra tener una potencia máxima para θ_1 pero no necesariamente para θ_2.

Retomando el ejemplo anterior, si tomamos como una hipótesis alternativa con dos valores $H_1 = \{0, 2\}$, entonces para $\theta = 0$ la región crítica más potente sera de la forma $\mathcal{R} = \{\bar{X} < c\}$, que, como lo vimos, no es la región crítica más potente para $\theta = 2$.

Definición 5.5 *Si un test δ maximiza la función de potencia para todo valor θ de la hipótesis alternativa $H_1 : \theta \in \Omega_1$, se dice que el test δ es uniformemente más potente (U.M.P.); es decir que δ^* es un test U.M.P. al nivel de significación α_o si $\alpha(\delta) \leq \alpha_o$ y si para todo otro test δ tal que $\alpha(\delta) \leq \alpha_o$, se tiene $\pi(\theta/\delta) \leq \pi(\theta/\delta^*)\ \forall \theta \in \Omega_1$*

Observamos en el ejemplo que la razón de las verosimilitud dado $\mu = \mu_2$ y $\mu = \mu_1$ se escribe:

$$\frac{f_n(\underline{x}/\mu_2)}{f_n(\underline{x}/\mu_1)} = exp\{\frac{n(\mu_2 - \mu_1)}{\sigma^2}(\bar{X} - \frac{1}{2}(\mu_2 - \mu_1))\}$$

Se observa que $\frac{f_n(\underline{x}/\mu_2)}{f_n(\underline{x}/\mu_1)}$ depende de $\underline{x}$ a traves sólo de la media muestral $\bar{X}$; además crece en función de $\bar{X}$ si $\mu_1 < \mu_2$. Es decir que este cuociente es monótono con respecto a $\bar{X}$.

Definición 5.6 *Se dice que $f_n(\underline{x}/\theta)$ tiene una razón de verosimilitud monótona para un estadístico $g(\underline{x})$ si y sólo si $\forall \theta_1, \theta_2$ tal que $\theta_1 < \theta_2$, el cuociente $\frac{f_n(\underline{x}/\theta_2)}{f_n(\underline{x}/\theta_1)}$ depende del vector $\underline{x}$ a traves de la función $g(\underline{x})$ y el cuociente es una función creciente de $g(\underline{x})\ \forall \underline{x}$.*

En el ejemplo anterior $f_n(\underline{x}/\mu)$ tiene una razón de verosimilitud monótona en $\underline{x}$. Veamos otro ejemplo: una muestra aleatoria de una Bernoulli de parámetro p.
Tomando $y = \sum x_i$, $f_n(\underline{x}/p) = p^y(1-p)^{n-y}$.

Si $0 < p_1 < p_2 < 1$: $\frac{f_n(\underline{x}/p_2)}{f_n(\underline{x}/p_1)} = (\frac{p_2(1-p_1)}{p_1(1-p_2)})^y(\frac{(1-p_2)}{(1-p_1)})^n$
cuociente que depende de $\underline{x}$ a traves de y, y es una función creciente de y; tiene una razón de verosimilitud monótona en $\sum x_i$.

Definición 5.7 *Un test sobre las hipótesis $H_o : \theta \leq \theta_o$ contra $H_1 : \theta > \theta_o$, se dice test unilateral y un test sobre las hipótesis $H_o : \theta = \theta_o$ contra $H_1 : \theta \neq \theta_o$, se dice test bilateral.*

Vamos a mostrar que si $f_n(\underline{x}/\theta)$ tiene una razón de verosimilitud monótona en algún estadístico T, entonces existe un test U.M.P. para las hipótesis $H_o : \theta \leq \theta_o$ contra $H_1 : \theta > \theta_o$

Teorema 5.2 *Si $f_n(\underline{x}/\theta)$ tiene una razón de verosimilitud monótona en el estadístico T y si c es la constante tal que $I\!P(T \geq c/\theta = \theta_o) = \alpha_o$, entonces la regla de decisión que permite rechazar la hipótesis nula si $T \geq c$ es un test U.M.P. para $H_o : \theta \leq \theta_o$ contra $H_1 : \theta > \theta_o$ al nivel de significación α_o.*

Demostración: Sea θ_1 tal que $\theta_1 > \theta_o$
$\alpha(\delta) = I\!P(\text{rechazar } H_o/\theta = \theta_o) = \pi(\theta_o/\delta)$

$\beta(\delta) = I\!P(\text{aceptar } H_o/\theta = \theta_1) = 1 - \pi(\theta_1/\delta)$

Del lema de Neyman-Pearson, se deduce que entre todos los procedimientos tales que el error de tipo I $\alpha(\delta) < \alpha_o$, el valor de $\beta(\delta)$ será mínimo para el procedimiento δ^* que consiste en rechazar H_o cuando $\frac{f_n(\underline{x}/\theta_1)}{f_n(\underline{x}/\theta_o)} \geq k$, k siendo elegido de tal forma que
$I\!P(\text{rechaza } H_o/\theta = \theta_o) \leq \alpha_o$.

Como $\frac{f_n(\underline{x}/\theta_1)}{f_n(\underline{x}/\theta_o)}$ es una función creciente de T, un procedimiento, que rechaza H_o cuando el cuociente es al menos igual a k, es equivalente al procedimiento que rechaza H_o cuando T es al menos igual a una constante c.
La constante c es elegida de tal forma que $I\!P(\text{rechazar } H_o/\theta = \theta_o) \leq \alpha_o$

Ahora bien esto es cierto para todo $\theta_1 > \theta_o$. Luego este procedimiento es U. M. P. para $H_o : \theta = \theta_o$ contra $H_1 : \theta > \theta_o$

Por otro lado, la función de potencia es no decreciente en θ y por lo tanto que si $\pi(\theta_o/\delta) \leq \alpha_o$, entonces $\pi(\theta/\delta) \leq \alpha_o \ \forall \theta \leq \theta_o$.

Cuando $f_n(\underline{x}/\theta)$ no tiene una razón de verosimilitud monótona, el test de razón de verosimilitud permite resolver una gran cuantidad de problemas:

Si $H_o : \theta \in \Theta_o$ contra $H_1 : \theta \in \Theta_1$, se define

$$\lambda(\underline{x}) = \frac{Sup f_n(\underline{x}/\theta \in \Theta_1)}{Sup f_n(\underline{x}/\theta \in \Theta_o)}$$

El test de razón de verosimilitud consiste en rechazar H_o si $\lambda(\underline{x}) > k$ y no rechazar H_o si $\lambda(\underline{x}) < k$.

El problema es encontrar la distribución de $\lambda(\underline{x})$. El siguiente teorema da una solución.

Teorema 5.3 *SI θ es un parámetro de dimensión p y si la hipótesis nula es de la forma $H_o : H\theta = 0$ en que $H \in \mathcal{M}_{nxp}$, entonces $-2ln\lambda(\underline{x})$ tiene una distribución asintotica χ^2_r.*

5.3.4 Tests usuales

Veamos algunos tests usuales que se basan en los resultados anteriores.

Test sobre una media con la varianza conocida
Sea una v.a. $X \sim \mathcal{N}(\mu, \sigma^2)$ en que la varianza σ^2 es conocida y igual a 36^2 y una muestra aleatoria de tamaño n=9.

Sea $H_o : \mu = 180$ contra $H_1 : \mu > 180$ y un nivel de significación de 0.05.

De lo anterior, se deduce que la región crítica más potente es de la forma $\mathcal{R} = \{\bar{X} > c\}$ con c determinado por:

$$I\!P(\bar{X} > c/\mu = 180) = 0.05$$

Como $(\bar{X} \sim \mathcal{N}(\mu, 144)$, $(\bar{X} - 180)/12 \sim \mathcal{N}(0, 1)$ bajo la hipótesis $H_o : \mu = 180$.
$I\!P(\bar{X} - 180)/12 > (c - 180)/12) = 0.05 \Longrightarrow (c - 180)/12 = 1.65 \Longrightarrow c = 200$.

La región crítica $\{\bar{X} > 200\}$ es U. M. P. para todo $\mu > 180$ de la hipótesis alternativa.

El error de tipo II depende de μ. Como lo muestra la tabla 5.1 y el gráfico 5.3, el error de tipo II aumenta cuando el valor de μ es muy cercano al valor 180 de H_o.

μ	180	185	190	200	210	220	230
$\pi(\mu)$	0.05	0.11	0.20	0.50	0.80	0.95	0.994
$1 - \pi(\mu)$	0.95	0.89	0.80	0.50	0.20	0.05	0.006

Tabla 5.1: Potencia y error de tipo II para $H_1 : \mu > 180$

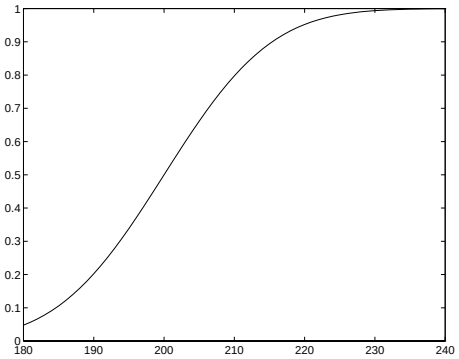


Gráfico 5.3: Función de Potencia para $H_1 : \mu > 180$

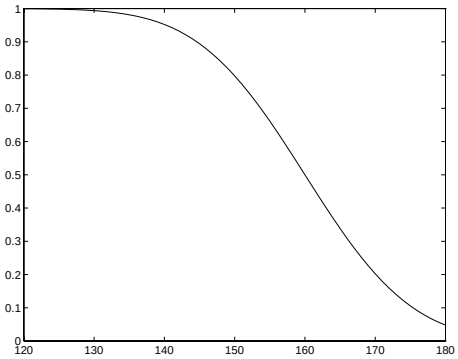


Gráfico 5.4: Función de Potencia para $H_1 : \mu < 180$

Sea ahora $H_o : \mu = 180$ contra $H_1 : \mu < 180$ con un nivel de significación de 0.05.

La región crítica más potente es de la forma $\mathcal{R} = \{\bar{X} < c\}$ con c determinado por:

$$I\!P(\bar{X} < c/\mu = 180) = 0.05$$

La región crítica $\{\bar{X} < 160\}$ es U. M. P. para todo $\mu < 180$ de la hipótesis alternativa. La función de potencia esta dada en la tabla 5.2 y el gráfico 5.4.

μ	180	175	170	160	150	140	130
$\pi(\mu)$	0.05	0.11	0.20	0.50	0.80	0.95	0.99
$1-\pi(\mu)$	0.95	0.89	0.80	0.50	0.20	0.05	0.006

Tabla 5.2: Potencia y error de tipo II para $H_1 : \mu < 180$

Sea finalemente $H_o : \mu = 180$ contra $H_1 : \mu \neq 180$ con un nivel de significación de 0.05.

No existe un test U. M. P. para este test bilateral; se propone como región crítica $\mathcal{R} = \{\bar{X} < a\} \cup \{\bar{X} > b\}$ de tal forma que $I\!P(\bar{X} < a) = 0.025$ y $I\!P(\bar{X} > b) = 0.025$. Obtenemos a=156.5 y b=203.5, que da una función de potencia presentada en la tabla 5.3 y el gráfico 5.5. Se nota que la potencia es siempre inferior o igual a la potencia de la tabla 5.1 o 5.2 para todo μ.

μ	140	150	160	170	175	180	185	190	200	210	220
$\pi(\mu)$	0.91	0.69	0.37	0.12	0.07	0.05	0.07	0.12	0.37	0.69	0.91
$1-\pi(\mu)$	0.09	0.31	0.43	0.88	0.93	0.95	0.93	0.88	0.43	0.31	0.09

Tabla 5.3: Función de Potencia para $H_1 : \mu \neq 180$

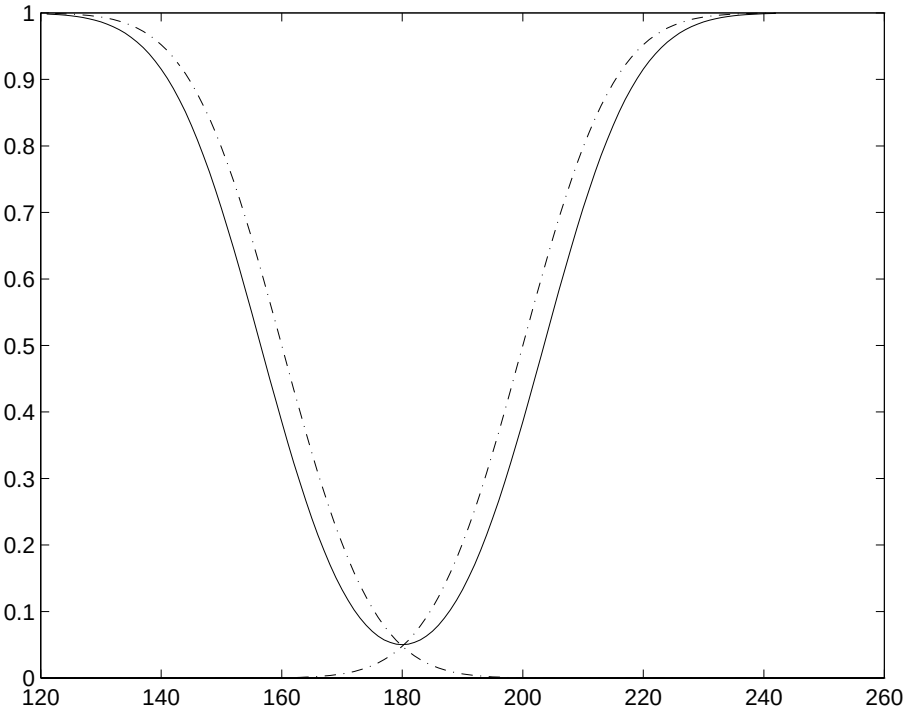


Gráfico 5.5: Función de Potencia para $H_1 : \mu \neq 180$

Se observara que este test se basa en el supuesto de distribución normal de los valores muestrales. Cuando el tamaño de la muestra es grande, este supuesto es aceptable, pero para muestras pequeñas, es importante comprobar si lo es.

Test sobre una media con la varianza desconocida
Si retomamos el problema anterior pero suponemos que la varianza es desconocida. En este caso se procede de manera parecida al caso anterior con la distribución de Student de la variable $\frac{(\bar{X}-\mu)}{S_n/\sqrt{n-1}}$ que es una Student a n-1 g.l.

El problema en este caso es la dificultad que se encuentra para calcular la potencia del test para una hipótesis alternativa.

Test sobre una varianza
Si ahora planteamos las hipótesis:

$$H_o : \sigma^2 \geq \sigma_o^2 \qquad contra \qquad H_1 : \sigma^2 < \sigma_o^2,$$

en donde σ_o^2 es un escalar positivo dado.

A partir del estadístico $\frac{nS_n^2}{\sigma_o^2}$, que sigue una distribución de χ^2 a n-1 grados de libertad bajo H_o, se construye la región critica de nivel de significación α::

$$I\!P(\frac{nS_n^2}{\sigma_o^2} > c) = \alpha$$

Test de comparación de dos medias
Frecuentemente uno esta interesado no en uma sola media, pero en la diferencia entre dos medias. Por ejemplo, la diferencia de sueldos medios μ_1 y μ_2 entre dos poblaciones Ω_1 y Ω_2. Las hipótesis se escriben entonces:

$$H_o : \mu_1 - \mu_2 = d_o$$

$$H_1 : \mu_1 - \mu_2 \neq d_o$$

Es más usual tomar $d_o = 0$ y la hipótesis alternativa H_1 puede ser

$$H_1 : \mu_1 - \mu_2 \neq 0 \qquad o \qquad H_1 : \mu_1 - \mu_2 > 0$$

Sea la v.a. sueldo X $\sim \mathcal{N}(\mu_1, \sigma_1^2)$ en Ω_1 y X $\sim \mathcal{N}(\mu_2, \sigma_2^2)$ en Ω_2. Si se tiene una media muestral $\bar{X}_1$ de X obtenida sobre una muestra de tamaño n_1 en Ω_1 y una media muestral $\bar{X}_2$ de X obtenida sobre una muestra de tamaño n_2 en Ω_2, entonces

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

Si las varianzas σ_1^2 y σ_2^2 son conocidas, entonces se obtiene una región crítica de nivel de significación $\alpha = 0.05$ para $H_o : \mu_1 - \mu_2 = 0$ contra $H_1 : \mu_1 - \mu_2 > 0$:

$$I\!P(\bar{X}_1 - \bar{X}_2 > 1.96\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2})$$

Si las varianzas son desconocidas, pero si se supone que son iguales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), entonces se estima esta varianza y se usa un estadístico que sigue una distribución t de Student. Un estimador insesgado de σ^2 es:

$$S^2 = (n_1 S_1^2 + n_2 S_2^2)/(n_1 + n_2 - 2)$$

en que S_1^2 y S_2^2 son las varianzas empíricas sesgadas de σ_1^2 y σ_2^2. Entonces

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2})(\frac{n_1 + n_2}{n_1 n_2})}}$$

es una Student a $n_1 + n_2 - 2$ grados de libertad.

La región crítica se define entonces como:

$$I\!P(\bar{X}_1 - \bar{X}_2 > t_\alpha \sqrt{(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2})(\frac{n_1 + n_2}{n_1 n_2})}$$

en donde t_α es tal que

$$I\!P(t_{n_1+n_2-2} > t_\alpha) = \alpha$$

Aquí se hizo el supuesto de igualdad de las varianzas y de independencia de las dos muestras.

Test para pares de observaciones

Hay situaciones en donde las muestras no son independientes. Es el caso cuando se toman muestras formadas de pares, es decir cuando cada observación de una muestra es relacionada a una observación de la otra muestra. Por ejemplo, se considera la diferencia de edades de las parejas en un grupo de matrimonios; una muestra esta formada de las esposas y la otra muestra de sus maridos. La dos muestras no son independientes y son del mismo tamaño. Sean (X,Y) las v.a. edades de la mujer y su marido y una muestra de n matrimonios $\{(x_i, y_i), i = 1, 2, ..., n\}$. La diferencia entre las medias empíricas $\bar{X}_n$ y $\bar{y}_n$ es un estimador insesgado de la diferencia poblacional μ_1 y μ_2 en las dos poblaciones apareadas:

$$E(\bar{X}_n - \bar{y}_n) = E(X - Y) = E(X) - E(Y) = \mu_1 - \mu_2$$

Pero debido a la dependencia entre X e Y la varianza de la diferencia X-Y cambia.

$$\sigma_{X-Y}^2 = E(X - Y - \mu_1 - \mu_1)^2) = E(X - \mu_1)^2 + E(Y - \mu_1)^2 - 2E(X - \mu_1)(Y - \mu_2)$$

$$\sigma_{X-Y}^2 = \sigma_1^2 + \sigma_1^2 - 2Cov(X, Y)$$

Como no se conoce en general las varianzas σ_1^2, σ_1^2 y la covarianza Cov(X,Y), lo más simple es estimar la varianza de la diferencia σ_{X-Y}^2 considerando que los valores muestrales son las diferencias $d_i = x_i - y_i$ que provienen de una sola muestra:

$$\hat{\sigma}_{X-Y}^2 = \frac{\sum (d_i - \bar{d}_n)^2}{n}$$

en donde $\bar{d}_n = \frac{\sum d_i}{n} = \frac{\sum x_i - y_i}{n}$

$$\hat{\sigma}^2_{X-Y} = \frac{\sum (x_i - y_i)^2}{n} - \bar{d}_n^2$$

El estimador de la varianza de la media diferencia es entonces $\frac{\hat{\sigma}^2_{X-Y}}{n}$ y $\frac{\bar{X}_n - \bar{y}_n - \mu_1 - \mu_2}{\hat{\sigma}_{X-Y}/(n-1)}$ sigue una t de Student a n-1 g.l.

Test de comparación de dos varianzas: la distribución F
Se quiere comparar las varianzas σ_1^2 y σ_2^2 de dos poblaciones normales a partir de muestras aleatorias independientes de cada población. Si $X_1, ..., X_n$ e $Y_1, ..., Y_m$ son las muestras aleatorias respectivas, $S_1^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$ y $S_2^2 = \frac{1}{n}\sum (Y_i - \bar{Y})^2$ son las varianzas muestrales sesgadas.

$U = nS_1^2/\sigma_1^2 \sim \chi^2_{n-1}$ y $V = mS_2^2/\sigma_2^2 \sim \chi^2_{m-1}$; ademas U y V son independientes.

Vimos en el capítulo anterior que $\frac{U/(n-1)}{V/(m-1)}$ sigue una distribución F de Fisher a n-1 y m-1 grados de libertad.

Consideramos entonces el estadístico

$$\frac{nS_1^2/(n-1)}{mS_2^2/(m-1)}$$

que sigue una distribución $F_{n-1,m-1}$ bajo la hipótesis nula $H_o : \sigma_1 = \sigma_2$.

Se define entonces la región crítica de nivel de significación α para $H_o : \sigma_1^2 = \sigma_2^2$

$$I\!P(\frac{nS_1^2/(n-1)}{mS_2^2/(m-1)} > F_\alpha) = \alpha$$

en donde F_α es calculado a partir de la F de Fisher a $n-1$ y $m-1$ g.l.

5.4 TESTS χ^2

Diversas situaciones pueden describirse a partir de una distribución multinomial. Veremos previamente dos distribuciones de vectores aleatorios, la distribución normal multivariada, y la distribución multinomial con su comportamiento asintótico. Después de presentar un test para un modelo multinomial, veremos aplicaciones para hipótesis no paramétricas.

5.4.1 La distribución normal multivariada

Se tiene dos definiciones equivalentes para la distribución normal multivariada.

Sea $X = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ X_p \end{pmatrix}$ un vector aleatorio

Definición 5.8 *Sea u:* $\mathbb{R}^p \longrightarrow \mathbb{R}$
Se dice que X es un vector normal multivariado de orden p de vector de media μ *y de matriz de varianza-covarianza* Γ $(X \sim \mathcal{N}_p(\mu, \Gamma))$ *si y sólo si* $u(X) \sim \mathcal{N}(u(\mu), \Gamma(u,u))$

Es decir que si X es un vector normal, toda combinación lineal de X es una v.a. normal.

Definición 5.9 *Se dice que* $X \sim \mathcal{N}_p(\mu, \Gamma))$ *si su función caracteristica es*

$$\Psi_X(u) = exp(iu^t\mu - \frac{1}{2}u^t\Gamma u) \qquad \forall u \in \mathbb{R}^p$$

Propiedades:

- Tomando como vector u los vectores canónicos, se obtiene las leyes marginales de X, que son normales; pero la recíproca es falsa: un vector formado de variables normales no es necesariamente un vector normal.
- Sea Y una matriz (pxq).
  $X \sim \mathcal{N}_p(\mu, \Gamma) \Longrightarrow Y = AX \sim \mathcal{N}_q(A\mu, A\Gamma A^t)$
- Las v.a. X_i son independientes $\Longleftrightarrow$ Γ es diagonal
- Γ es semidefinida positiva
  En efecto $\Gamma(u,u) = u^t\Gamma u$ es la varianza de la v.a. $u(X) = u^tX$.
- Si Γ es de rango r, existe Λ una matriz (pxr) tal que $\Gamma = \Lambda\Lambda^t$. Entonces:

  $$X \sim \mathcal{N}_p(\mu, \Gamma) \Longleftrightarrow X = \mu + \Lambda Y \quad Y \sim \mathcal{N}_r(0, I_r)$$

  es decir que las componentes del vector Y son centradas, normalizadas y independientes entre si.
- Si Γ es invertible, Λ es invertible también e $Y = \Lambda^{-1}(X - \mu)$.

Este último resultado permite calcular la densidad del vector X. En efecto se puede calcular la densidad del vector $Y \sim \mathcal{N}_p(0, I_p)$:

$$f(Y) = \prod f(Y_i) = (\frac{1}{2\pi})^{p/2} exp(-\frac{1}{2}\sum Y_i^2)) = (1/2\pi)^{p/2} exp(-\frac{1}{2}YY^t)$$

Como $YY^t = (\Lambda^1(X-\mu)^t\Lambda^1(X-\mu) = (X-\mu)^t\Gamma^{-1}(X-\mu)$, el Jacobiano de la transformación es $|\Gamma|^{-1/2}$, luego la densidad de X es:

$$h(X) = (\frac{|\Gamma|^{-1/2}}{2\pi})^{p/2} exp(-\frac{1}{2}(X-\mu)^t\Gamma^{-1}(X-\mu))$$

Proposición 5.1 *Si $X \sim \mathcal{N}(\mu, \Gamma)$ con Γ de rango r, entonces $\| X-\mu \|^2_{\Gamma^{-1}} \sim \chi^2_r$.*

Demostración: Acordamos que si $Y \sim \mathcal{N}(0, I_r)$, $\| Y \|^2 = \sum Y_i^2 \sim \chi^2_r$. Como $\Gamma = \Lambda\Lambda^t$, existe Y tal que $X = \mu + \Lambda Y$, con $Y \sim \mathcal{N}(0, I_r)$. Pero se puede escribir $Y = (\Lambda^t\Lambda)^{-1}\Lambda^t(X-\mu)$, luego:

$$\| Y \|^2_{I_p} = YY^t = \| X-\mu \|^2_{\Gamma^{-1}} \sim \chi^2_r$$

5.4.2 La distribución multinomial

Es una generalización de la distribución binomial. En vez de tener dos alternativas en cada experimento, se tienen k alternativas (k $\geq$ 2). Por ejemplo, hay seis resultados posibles cuando se tira un dado. Si el "1" tiene probabilidad p_1, el "2" tiene probabilidad p_2,...,el "6" tiene probabilidad p_6, y si hacemos n lanzamientos independientes, los números M_1 de "1", M_2 de "2",..., M_6 de "6" constituyen un vector aleatorio M con una distribución multinonimal de parámetros $n, p_1, p_2, ..., p_6$. Se observa que $\sum M_i = n$.

$$I\!P(M = m) = I\!P(M_1 = m_1, ..., M_6 = m_6) = \frac{n! p_1^{m_1} p_2^{m_2} ... p_6^{m_6}}{m_1! m_2! ... m_6!}$$

Calculamos la esperanza y la varianza de M. Si $p = \begin{pmatrix} p_1 \\ p_2 \\ . \\ p_6 \end{pmatrix}$, entonces $E(M) = np$.

Sea el resultado J_i del lanzamiento i: $J_i = e_h$, el h-esimo vector canonico si el resultado es h. Entonces $M - \sum J_i$, $E(J_i) = p$ y

$$E(J_i J_i^t) = \sum_h e_h e_h^t I\!P(J_i = e_h) = \begin{pmatrix} p_1 & 0 & . & 0 \\ 0 & p_2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & p_6 \end{pmatrix} = Diag(p)$$

$Var(J_i) = E(J_i J_i^t) - E(J_i)[E(J_i)]^t = Diag(p) - pp^t = \Sigma(p)$

Luego $Var(M) = n\Sigma(p)$

Por el Teorema del Límite Central, se tiene:

$$lim_{n \longrightarrow +\infty} I\!P(\frac{M-np}{\sqrt{n}} \leq x) = \Phi(x),$$

en donde Φ es la función de distribución normal multivariada centrada de matriz de covarianza $\Sigma(p)$.

Ejercicio: Muestre que si el vector multinomial es de dimensión k, entonces el rango de la matriz $\Sigma(p)$ es igual a k-1 (Se podra mostrar que el nucleo de $\Sigma(p)$ es de dimensión 1).

Proposición 5.2 *Si M es un vector de distribución multinomial $(n, p_1, ..., p_k)$, entonces $Q = \sum \frac{(M_i - np_i)^2}{np_i}$ tiene una distribución asintótica de χ^2_{k-1}.*

La demostración se basa en el resultado del ejercicio.

5.4.3 Test de ajuste para un modelo multinomial

Sea un dado que se tira n=102 veces. Se obtiene entonces la distribución empírica (tabla 5.4):

M_i	1	2	3	4	5	6	Total
f_i	12	11	22	20	16	21	102

Tabla 5.4

¿Podemos concluir si el dado esta cargado?

Sea la hipótesis nula $H_o : p_i = \frac{1}{6} \quad \forall i$

Entonces calculamos el estadístico Q para construir la región crítica (tabla 5.5).

i	M_i	np_i	$M_i - np_i$	$(M_i - np_i)^2/np_i$
1	12	17	-5	1.471
2	11	17	-6	2.118
3	22	17	5	1.471
4	20	17	3	0.529
5	16	17	-1	0.059
6	21	17	4	0.941
Total	102	102	0	6.589

Tabla 5.5

Se obtiene Q=6.589, y $I\!P(\chi^2_5 > 6.589) > 5\%$, por lo cual no se rechaza H_o. Las diferencias no son suficientemente significativas para concluir que el dado esta cargado.

5.4.4 Test de ajuste para una distribución discreta

Se considera el número de accidentes X observados cada fin de semana en una carretera (tabla 5.6). Se quiere probar la hipótesis que X sigue una distribución de Poisson de parámetro λ a partir de datos obtenidos sobre un año. En un primer tiempo supondremos λ conocido e igual a 1.5. Se tiene entonces $H_o : X \sim \mathcal{P}(1.5)$.

N^o accidentes	0	1	2	3	4	5	6	Total
N^o semanas	17	16	10	5	2	1	1	52

Tabla 5.6

Bajo H_o, los números de semanas M_o con 0 accidente, M_1 con 1 accidente, ..., M_6 con 6 o más accidentes sigue una distribución multinomial de parámetros n=52, y $p_o = I\!P(X = 0)$, $p_1 = I\!P(X = 1)$, ..., $p_6 = I\!P(X \geq 6)$,

Calculamos los $p_i = I\!P(X = i)$, con $X \sim \mathcal{P}(1.5)$.

i	M_i	p_i	$M_i - np_i$	$(M_i - np_i)^2/np_i$
0	17	0.2231	5.3988	2.5124
1	16	0.3347	0.5956	0.0204
2	10	0.2510	-3.0520	0.7137
3	5	0.1255	-1.5260	0.3568
4	2	0.0471	-0.4492	0.0824
5	1	0.0141	0.2668	0.0971
6	1	0.0045	0.7660	3.2735
Total	52	1.0000	0	7.0563

Tabla 5.7

Se obtiene Q=7.0563 (tabla 5.7), y $I\!P(\chi^2_6 > 7.0563) > 5\%$, por lo cual no se rechaza H_o.

Ahora si se supone que no se conoce el parámetro λ, se puede estimar por $\hat{\lambda} = \bar{X}_n = \sum iM_i/52 = 72/52 = 1.385$ y proceder como antes. Pero ahora el estadístico Q pierde un grado de libertad debido a la estimación.

Con el parámetro $\hat{\lambda}$, Q=5.62 y $I\!P(\chi^2_5 > 5.62) > 5\%$.

5.4.5 Test de ajuste para una distribución continua

Si queremos construir un test χ^2 para una hipótesis sobre una distribución continua como $H_o : X \sim \mathcal{N}(1, 0.25)$, hay que transformar la variable en una variable discreta. Se divide el rango de X en k intervalos disjuntos I_1, I_2, ...,I_k y se cuenta los números de observaciones de

la muestra M_i que caen en el intervalo I_i. El vector M de los efectivos de los intervalos sigue una distribución multinomial de parámetros de probabilidad determinados por la hipótesis nula.

Sea por ejemplo, las temperaturas medias X del mes de septiembre en Urbe durante 60 años (tabla 5.8). Se quiere probar la hipótesis nula $H_o : X \sim normal$.

Hay diferentes maneras de definir la partición de intervalos de $\mathbb{R}$. Una vez fijado el número de intervalos, se pueden elegir del mismo largo o de la misma probabilidad. Tomaremos aquí 10 intervalos equiprobables.

Para calcular las probabilidades, hay que estimar previamente los parámetros μ y σ^2 de la normal:

$$\hat{\mu} = \bar{X}_n = 15.76 \quad \hat{\sigma}^2 = S_n^2 = 13.82$$

Luego los intervalos I_j se obtienen de tal forma que (tabla 5.9):

$$\mathbb{P}(X \in I_j) = 0.10 \quad \forall j$$

en donde $X \sim \mathcal{N}(15.76, 13.82)$.

Se obtiene Q=9.35. El estadístico χ^2 tiene aquí 7 g.l. (Se estimaron dos parámetros). Como $\mathbb{P}(\chi_7^2 > 9.35) > 5\%$, no se rechaza la hipótesis de normalidad.

5.2	6.5	7.5	8.2	10.1	10.5	11.6	12.0	12.0	12.8	13.5	13.8
13.9	14.0	14.0	14.2	14.3	14.5	14.7	14.8	15.0	15.0	15.2	15.2
15.3	15.4	15.6	15.8	15.8	15.9	16.0	16.1	16.2	16.4	16.4	16.5
16.5	16.8	16.9	17.0	17.0	17.1	17.1	17.1	17.4	17.6	17.9	18.2
18.5	18.8	18.9	19.4	19.8	20.3	20.9	21.4	21.9	22.5	22.8	23.9

Tabla 5.8: Temperaturas medias

I_i	M_i	np_i	$M_i - np_i$	$(M_i - np_i)^2/np_i$
$]-\infty$,10.96]	6	6	0	0.00
]10.96,12.64]	3	6	-3	1.50
]12.64,13.83]	3	6	-3	1.50
]13.83,14.83]	8	6	2	0.67
]14.83,15.76]	7	6	1	0.17
]15.76,16.69]	10	6	4	2.67
]16.69,17.69]	9	6	3	1.50
]17.69,18.88]	4	6	-2	0.67
]18.88,20.56]	4	6	-2	0.67
]20.56,$+\infty$]	6	6	0	0.00
Total	60	60	0	9.35

Tabla 5.9

5.4.6 Test de independencia en una tabla de contingencia

Cuando dos v.a. discretas con valores en A y B respectivamente son independientes, se tiene:

$$I\!P(X = i \ e \ Y = j) = I\!P(X = i)I\!P(Y = j) \quad \forall (i,j) \in A \times B$$

Si A y B son conjuntos finitos (card(A)=p, card(B)=q), las frecuencias M_{ij} de observaciones obtenidas en una muestra bivariada de tamaño n siguen una distribución multinomial de parámetro n, p en donde p es el conjunto de las probabilidades $p_{ij} = I\!P(X = i \ e \ Y = j)$.

Bajo la hipótesis de independencia de X e Y, se puede estimar estos parámetros p_{ij} a partir de las frecuencias marginales de X e Y:

$$\hat{p}_{ij} = \hat{p}_{i\bullet}\hat{p}_{\bullet j}$$

con $\hat{p}_{i\bullet} = \sum_j M_{ij}/n$ y $\hat{p}_{\bullet j} = \sum_i M_{ij}/n$.

Lo que permite usar el estadístico

$$Q = \sum_{ij} \frac{(M_{ij} - n\hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{n\hat{p}_{i\bullet}\hat{p}_{\bullet j}}$$

que sigue una distribución asintótica χ^2 a (p-1)(q-1) g.l.

Sea un conjunto de consumidores que dan su apreciación sobre una margarina. Se quiere estudiar si existe una relación entre la opinión de los consumidores y su nivel socio-económico (NSE).

Se considera la tabla de contingencia obtenida a partir de une encuesta de estudio de mercado sobre 1600 consumidores (tabla 5.10), que presenta las frecuencias M_{ij} para cada NSE i y apreciación j.

NSE	APRECIACION			TOTAL
	MALA	REGULAR	BUENA	
A	140	100	45	285
B	50	225	350	625
C	15	175	500	690
TOTAL	205	500	895	1600

Tabla 5.10: Tabla de contingencia

Las probabilidades p_{ij} se estiman usando las frecuencias marginales de la tabla; por ejemplo, para el NSE A con la apreciación MALA se obtiene $\hat{p}_{11} = 285 \times 205/1600 = 0.0228$ y $n\hat{p}_{11} = 36.51$.

Se obtiene el valor Q=521.46. Como $I\!P(\chi_4 > 521.46) < 5\%$, se rechaza la hipótesis de independencia entre el NSE y la apreciación.

Nota: Se puede usar el mismo test para probar la independencia de dos variables continuas transformandolas en variables discretas.

5.5 EJERCICOS

1. El cocinero del casino preparó la masa para hacer 500 empanadas. Ese mismo día, en un grupo de 20 alumnos que almorzaron juntos, alguien propuso contar la cantidad de pasas que cada uno encontrase en su empanada, encontrándose la siguiente distribución:

N^o de pasas	N^o de empanadas
0	1
1	3
2	4
3	5
4	4
5	2
8	1

a) Suponiendo que la distribución de la cantidad de pasa X en una empanada sigue una ley de Poisson, estime el parámetro λ de esta ley.
b) Justifique la hipótesis: "H_0: *La distribución de la cantidad de pasas en una empanada sigue una ley de Poisson*" de las dos formas siguientes:
(i) A priori: Buscando la probabilidad de que una empanada tenga exactamente x pasas.
(ii) A posteriori: comparando los resultados esperados bajo la hipótesis con aquellos observados en la muestra.
c) Se decide que las empanadas son *aceptables* si en promedio cada empanada tiene 3.5 pasas; el cocinero afirma que esta se la cantidad de pasas por empanadas. Los alumnos, en cambio, objetan que las empanadas tienen en promedio sólo 2.5 pasas.
¿Qué significa la elección de los test de hipótesis siguientes:

$$H_0\text{: } \lambda = 3.5 \text{ vs. } H_1\text{: } \lambda = 2.5 \qquad H_0'\text{: } \lambda = 2.5 \text{ vs } H_1'\text{: } \lambda = 3.5 \text{ ?}$$

d) Dar la región crítica al test H_0 vs. H_1 al nivel de significación $\alpha = 0.05$. Dar la potencia de este test y concluir si las empanadas son *aceptables*.
e) Misma pregunta tomando H_0' vs H_1'.
f) Comparar las dos decisiones anteriores.

2. Se tienen los pesos de diez parejas antes y después de 6 meses de matrimonio:

	antes	72	69	81	71	88	78	68	76	86	95
Hombres	despues	77	68.5	85	74.5	90.5	76	71	75	87.5	101
	antes	52	56	61	49	57	63	66	59	67	51
Mujeres	después	54	55	58	50	55	61	64	56	70	50

¿Cuál es la influencia del matrimonio sobre el peso de los hombres y de las mujeres?

3. Se quiere probar si hay una diferencia de ingreso entre hombres y mujeres médicos. Se hizo una encuesta a $n = 200$ médicos seleccionados al azar e independientemente. Se obtuvo la siguiente información:

	Ingresos bajos	Ingresos altos	Total
Hombres	20	100	120
Mujeres	70	10	80
Total	90	110	200

a) Sean p_1 y p_2 las proporciones poblacionales de médicos hombres y mujeres; y sean p_1' y p_2' las proporciones poblacionales de médicos con ingresos bajos y altos. Realice los tests

$$H_0 : p_1' = p_2 \text{ vs. } H_1 : p_1' \neq p_2 \qquad H_0' : p_1 = p_2' \text{ vs. } H_1' : p_1 \neq p_2'.$$

b) Estudie la independencia entre sexo e ingreso.

4. Supóngase que $X_1, ..., X_n$ constituyen una m.a.s. de una v.a. X con distribución uniforme sobre $[0, \theta]$ y que se han de contrastar las siguientes hipótesis:

$$H_0 : \theta \geq 2 \text{ vs. } H_1 : \theta < 2.$$

Sea $Y_n = max\{X_1, ..., X_n\}$ y considérese un procedimiento de contraste tal que la región crítica contenga todos los resultados tq. $Y_n \leq 1.5$.

a) Determínese la función de potencia del contraste.
b) Determínese el tamaño del test.

5. Supóngase que se desconoce la proporción p de artículos defectuosos en una población de artículos y se desea probar las hipótesis

$$H_0 : p = 0.2 \text{ vs } H_1 : p \neq 0.2.$$

Supóngase además que se selecciona una m.a.s. de tamaño 20. Sea Y el número de artículos defectuosos en la muestra y considérese un procedimiento para resolver el test tal que la región crítica está dada por $Y \geq 7$ o $Y \leq 1$.

a) Determínese el función de la potencia $\pi(p)$ en los puntos $p = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ y 1.
b) Determine el tamaño del test.

6. Sea $X_1, ..., X_n$ una m.a.s. de una distribución normal de media μ desconocida y varianza 1. Sea μ_0 un real dado. Se tienen las hipótesis

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0.$$

Supongamos que el tamaño de la muestra es 25, y considérese que el procedimiento para no rechazar H_0 está dado por $|\bar{X}_n - \mu_0| < c$. Determínese el valor de c para que el tamaño del test sea 0.05.

7. Sea $X_1, ..., X_n$ una m.a.s. de una distribución de media θ desconocida y varianza 1, y sean las hipótesis

$$H_0 : \theta = 3.5 \text{ vs. } H_1 : \theta = 5.0.$$

a) Entre los procedimientos para resolver el test anterior tal que $\beta(\delta) \leq 0.05$, descríbase un procedimiento para el que $\alpha(\delta)$ sea un mínimo.

b) Para $n = 4$, encuéntrese el valor mínimo descrito en a).

8. Supóngase que se selecciona una observación X de una $U(0, \theta)$, donde θ es desconocido y se plantean las siguientes hipótesis:

$$H_0 : \theta = 1 \text{ vs } H_1 : \theta = 2.$$

a) Demostrar que existe un procedimiento para resolver el test para el cual $\alpha(\delta) = 0$ y $\beta(\delta) < 1$.
b) Entre todas las soluciones del test para las cuales $\alpha(\delta) = 0$, hállese una para el cual $\beta(\delta)$ sea mínimo.

9. Sea $X_1, ..., X_n$ una m.a.s. de una $Poisson(\lambda)$, con λ desconocido. Sean λ_0 y λ_1 dados, con $\lambda_1 > \lambda_0 > 0$. Se tienen las siguientes hipótesis:

$$H_0 : \lambda = \lambda_0 \text{ vs. } H_1 : \lambda = \lambda_1.$$

Demuéstrese que el valor de $\alpha(\delta) + \beta(\delta)$ se minimiza por un procedimiento que rechaza H_0 cuando $\bar{X}_n > c$ y encuéntrese el valor de c.

10. Sea $X_1, ..., X_n$ una m.a.s. de una distribución con parámetro θ cuyo valor es desconocido. Supóngase además que se desea constrastar las siguientes hipótesis:

$$H_0 : \theta \leq \theta_0 \text{ vs } H_1 : \theta > \theta_0.$$

Supóngase además, que el procedimiento que se va a utilizar ignora los valores observados en la muestra y, en vez de ello, depende unicamente de una aleatorización auxiliar en la que se lanza una moneda desequilibrada de forma que se obtendrá cara con probabilidad 0.05 y sello con probabilidad 0.95. Si se obtiene una cara, entonces se rechaza H_0, y si se obtiene sello, no se rechaza H_0. Descríbase la función de potencia de este procedimiento.

11. Sea $X_1, ..., X_n$ una m.a.s. de una distribución con parámetro θ desconocido y una función de densidad conjunta $f_n(x/\theta)$ que tiene cociente de verosimilitud monótona en el estadístico $T = r(X)$. Sea θ_0 un valor específico de θ y supóngase que se quieren constrastar las hipótesis

$$H_0 : \theta \geq \theta_0 \text{ vs } H_1 : \theta < \theta_0.$$

Sea c una constante tal que $P(T \leq c/\theta = \theta_0) = \alpha_0$. Demostrar que el procedimiento que rechaza H_0 si $T \leq c$ es UMP al nivel α_0.

12. Sea $X_1, ..., X_n$ una m.a.s. de una $Poisson(\lambda)$ con λ desconocido. Supóngase que se quiere constrastar las hipótesis

$$H_0 : \lambda \geq 1 \text{ vs } H_1 : \lambda < 1 \ .$$

Supóngase además que el tamaño de la muestra es $n = 20$. ¿Para qué niveles de significación α_0, con $0 < \alpha_0 < 0.03$ existen tests UMP?

13. Consideremos una observación X de una distribución de Cauchy con un parámetro de localización desconocido θ, esto es, una distribución cuya función de densidad está dada por:

$$f(x/\theta) = \frac{1}{\pi[1 + (x - \theta)^2]} \quad (\forall x)$$

Se desean constrastar las hipótesis

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta > 0.$$

Demuestre que no existe un test UMP de estas hipótesis a ningún nivel de significación α_0.

14. Sea $X_1, ..., X_n$ una m.a.s. de una distribución $\mathcal{N}(\mu, 1)$. Supóngase que se desean contrastar las hipótesis

$$H_0 : \mu \leq 0 \text{ vs } H_1 : \mu > 0.$$

Se denota δ^* el test UMP con nivel de significación $\alpha_0 = 0.025$ y $\pi(\mu/\delta^*)$ la función de potencia de δ^*.

15. Sea $X_1, ..., X_n$ una m.a. de una distribución $U(0, \theta)$ con θ desconocido. Supongamos que queremos contrastar las hipótesis

$$H_0 : \theta = 3 \text{ vs } H_1 : \theta \neq 3.$$

Considere que H_0 se rechaza si $c_2 \leq max\{X_1, ..., X_n\} \leq c_1$ y sea $\pi(\theta/\delta)$ la función de potencia de δ. Determine los valores de c_1, c_2 para que $\pi(3/\delta) = 0.05$ y δ sea insesgado.

1 ASOCIACION ENTRE DOS VARIABLES

1.1 Introducción

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 presenta en primer lugar índices descriptivos de asociación y en seguida se hace inferencia sobre estos coeficientes.

1.2 Variables cuantitativas

Si se consideran dos variables X e Y cuantitativas, que toman valores sobre un conjunto I de individuos, una simple representación gráfica en $\mathbb{R}^2$ permitirá detectar la existencia y la forma de una eventual relación entre las dos variables. Una inspección del gráfico es necesaria para asegurarse de interpretar correctamente el coeficiente de correlación lineal, el índice de asociación más usual.
Sea $\{(x_i, y_i)/\text{i=1,...,n}\}$ el conjunto de realizaciones del par (X,Y) de variables sobre una muestra I de tamaño n. 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 de 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 1.1 *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 ariables, se elimina este efecto con el coeficiente de correlación lineal, que toma en cuenta de las varianzas s_x^2 de X y s_y^2 de Y.

Definición 1.2 *Se llama correlación lineal entre X e Y a:* $r_{x,y} = \frac{cov(x,y)}{s_x s_y}$

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

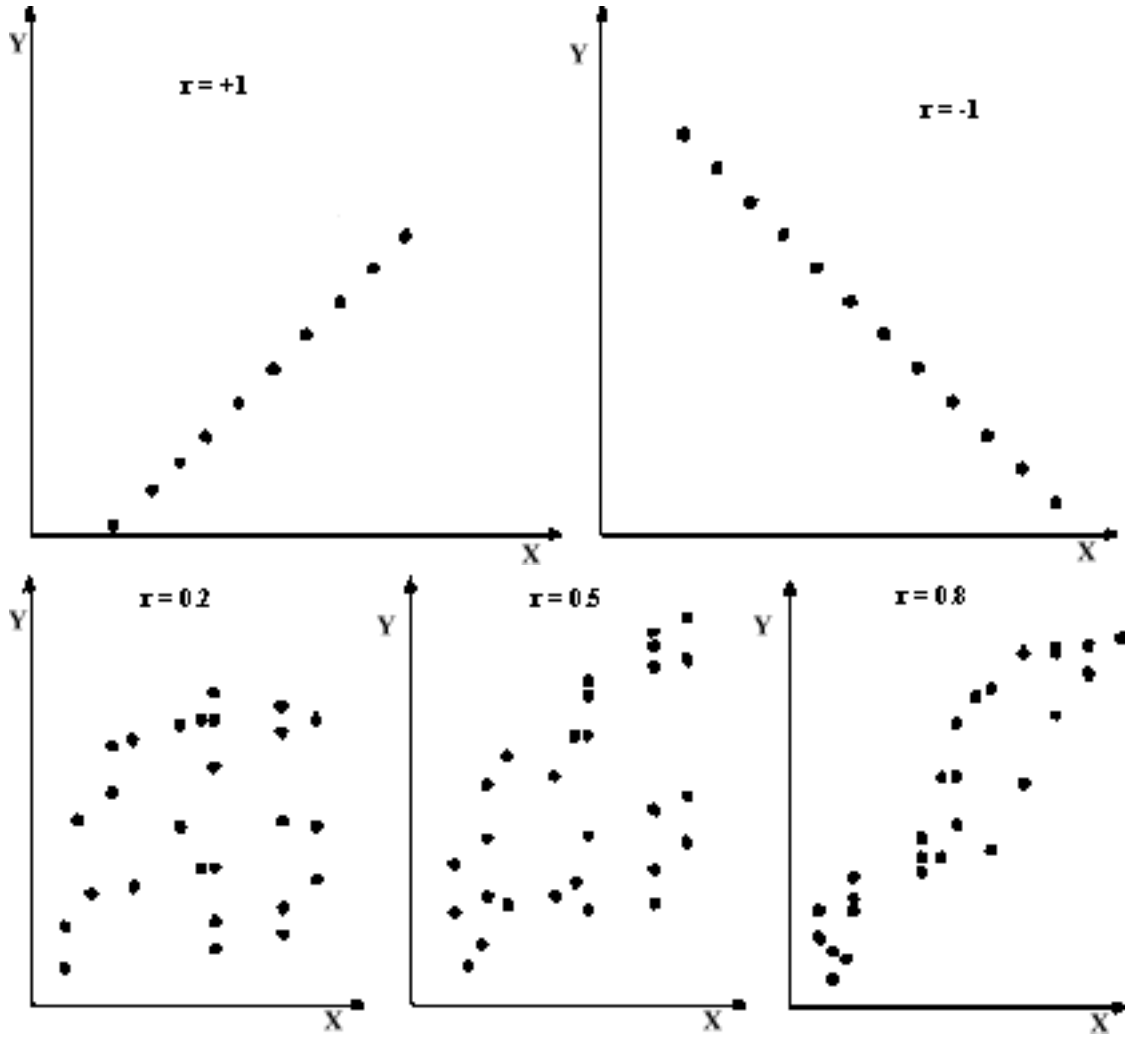
Este coeficiente toma como valores extremos +1 y -1; da el grado de relación de tipo lineal que existe entre X e Y.

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

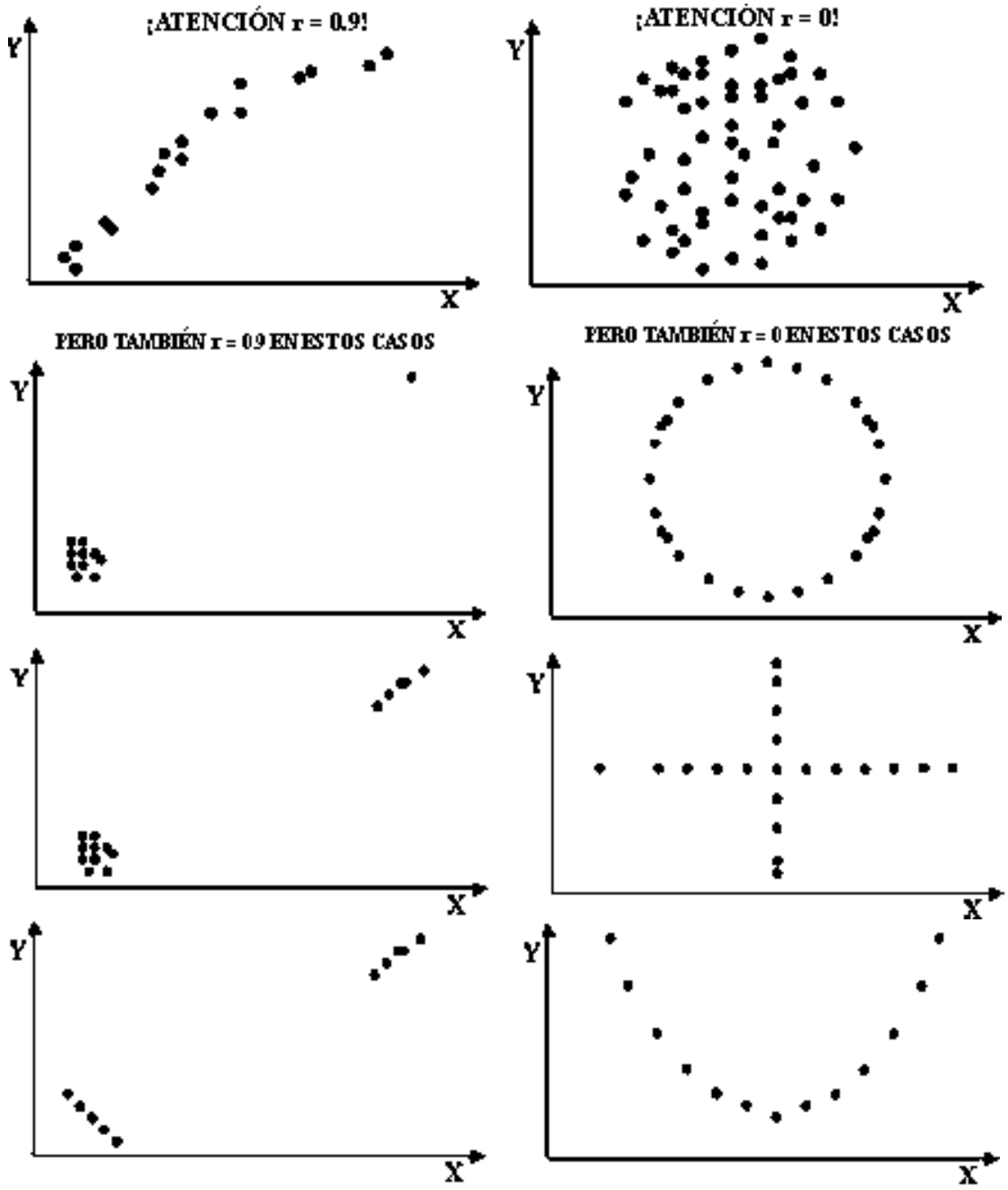
La tendencia lineal aumenta cuando $r_{x,y}$ tiende a ± 1 (ver gráficos 6.1). Pero cuando $r_{x,y} \neq \pm 1$, hay muchos casos muy diferentes que pueden producir el mismo valor del coeficiente $r_{x,y}$. De aquí la importancia de tener mucho cuidado en la interpretación de un coeficiente de correlación. Un dato aberrante, una mezcla de poblaciones, una relación no lineal pueden cambiar totalmente el valor del coeficiente (ver gráficos 6.2).
Cuando se estudia más de dos variables, se puede presentar los coeficientes de correlación en una matriz cuyo termino general r_{ij} es el coeficiente de correlacón relativo a las variables i y j. La matriz de correlaciones asociadas a la tabla 6.1, en la cual se tiene 6 variables observadas sobre 20 paises, esta dada en la tabla 6.2. Por ejemplo, el coeficiente de correlación entre la tasa de natalidad y la fecundidad es igual a 0.972.

PAIS	% POB. URBANA	/ASA NATALIDAD	/ASA MOR /ALIDAD	ESPERAN ZA VIDA	FECUN DIDAD	MORTALIDAD INFANTIL
ARGENTINA	86.2	20.3	8.6	71.0	2.8	28.8
BOLIVIA	51.4	34.4	9.3	54.5	4.6	84.8
BRASIL	76.9	26.1	7.5	65.6	3.2	56.4
COLOMBIA	70.3	25.8	5.9	68.8	2.9	37.0
COSTA RICA	53.6	26.3	3.7	74.9	3.1	13.7
CUBA	74.9	17.4	6.7	75.4	1.9	14.2
CHILE	85.6	22.5	6.4	71.8	2.7	16.9
ECUADOR	56.9	30.9	6.9	66.0	3.9	57.4
EL cALVADOR	44.4	33.5	7.1	64.4	4.0	45.6
GUATEMALA	42.0	38.7	7.6	63.4	5.4	48.5
HAITI	30.3	35.3	11.9	55.7	4.8	86.2
HONDURAS	43.6	37.1	7.2	64.9	4.9	59.7
MÉXICO	72.6	27.9	5.4	69.7	3.2	35.2
NICARAGUA	59.8	40.5	6.9	64.8	5.0	53.1
PANAMÁ	54.8	24.9	5.2	72.4	2.9	20.8
PARAGUAY	47.5	33.0	6.4	67.1	4.3	47.0
PERÚ	70.2	29.0	7.6	63.0	3.6	75.8
R. DOMINICANA	60.4	28.3	6.2	66.7	3.3	56.5
URUGUAY	85.5	17.1	10.3	72.2	2.3	20.0
VENEZUELA	90.5	28.3	5.4	70.0	3.5	33.2

TABLA 6.1: INDICADORES DEMOGRÁFICOS PARA 20 PAISES LATINOAMERICANOS
Fuente: PNUD 1992



GRAFICOS 6.1 PARA INTERPRETAR UN COEFICIENTE DE CORRELACIÓN LINEAL



GRAFICOS 6.2 PARA INTERPRETAR UN COEFICIENTE DE CORRELACIÓN LINEAL

VARIABLES	12	13	14	15	16	17
12 % POB. URBANA	1.0	-.739	-.179	.588	-.735	-.532
13 TASA .NATALIDAD	-.739	1.0	.101	-.723	.972	.682
14 TASA MORTALIDAD	-.179	.101	1.0	-.609	.262	.533
15 ESPERANZA VIDA	.588	-.723	-.609	1.0	-.769	-.951
16 FECUNDIDAD	-.735	.972	.262	-.769	1.0	.709
17 MORTAL. INFANTIL	-.532	.682	.533	-.951	.709	1.0

TABLA 6.2: Matriz de correlaciones asociada a la tabla 6.1

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

- Dada una función f sobre X, calcular el coeficiente de correlación entre f(X) e y. Este método es factible cuando se sospecha de la función f.
- Usar otros índices, como se vera más adelante.

1.3 Una variable cuantitativa y una variable nominal

Cuando una de las dos variables es nominal u ordinal, no se puede calcular el coeficiente de correlación lineal, salvo si se codifica tal variable, atribuyendo un valor numérico a cada una de las modalidades de la variable nominal. El problema esta entonces en la elección de una codificación.

1.3.1 Codificación de la variable nominal

Una forma natural de codificar una variable nominal X para medir su ligazón con una variable cuantitativa Y consiste en buscar la codificación de las modalidades de X que produce la mayor correlación lineal con la variable Y. Si X tiene p modalidades, se le puede asociar p variables indicadores $\{X^1, X^2, ..., X^p\}$ 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), entonces la variable cuantitativa ξ asociada a esta codificación se puede escribir:

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

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

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

Numéricamente el máximo se puede obtener con *la razón de correlación*, comparando varianzas.
Para una variable ordinal, es natural de imponer además que las codificaciones a_j siguen el mismo orden que las modalidades; en este caso hay que tomar en cuenta esta restricción en la maximización.

1.3.2 Razón de correlación

Se puede 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 j de X, se puede escribir y_{1j},...,$y_{n_j j}$ estas n_j observaciones de Y.
Si $\bar{y}$ es la media empírica de la variable Y sobre el total de las n observaciones, la varianza empírica de todas estas observaciones es igual a:

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

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:

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

Podemos comparar la *varianza total* s_y^2 con el promedio de las varianzas de los p grupos.

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

La media ponderada por los efectivos relativos $\frac{n_j}{n}$ de las medias $\bar{y}_j$ es igual a la media total $\bar{y}$ ($\sum \frac{n_j}{n}\bar{y}_j = \bar{y}$). Luego la cantidad $\sum_j \frac{n_j}{n}(\bar{y}_j - \bar{y})^2$ es la varianza ponderada de las medias $\bar{y}_j$.
Sean $b^2 = \sum_j \frac{n_j}{n}(\bar{y}_j - \bar{y})^2$ la varianza entre los grupos, y $w^2 = \sum_j \frac{n_j}{n} w_j^2$ la varianza promedio dentro de los grupos. La varianza total se descompone entonces en:

$$s_y^2 = b^2 + w^2$$

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 podra obtener el valor de un observación sobre la variable Y conociendo su modalidad sobre X. Se detecto en este caso una relación funcional de X hacia Y.
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 podra decir nada sobre el valor de Y conociendo la modalidad de X. No se detecto ninguna relación funcional de X hacia Y.
El índice siguiente permite de medir el grado de asociación de tipo funcional de X hacia Y:

$$\eta_{y/x}^2 = \frac{b^2}{s_y^2}$$

Este coeficiente toma valores entre 0 y 1:

$\eta_{y/x}^2 = 1$	relación funcional estricta
$0 < \eta_{y/x}^2 < 1$	tendancia funcional
$\eta_{y/x}^2 = 0$	ausencia de tendencia funcional

La tendencia funcional aumenta con $\eta_{y/x}^2$.
Se tiene finalmente el resultado propuesto en el párrafo anterior, que permite dar otra interpretación de la razón de correlación.

Proposición 1.1 *El máximo de la* $cor^2(y, \sum_j a_j x^j)$ *es igual a* $\eta_{y/x}^2$.

1.3.3 Relación funcional entre dos variables cuantitativas

Cuando un coeficiente de correlación lineal entre X e Y es bajo, no significa que las variables X e Y no estan ligadas; puede existir otro tipo de relación entre X e Y. Ahora bien, por codificación se puede transformar una variable nominal u ordinal en una variable cuantitativa, inversamente, se puede transformar una variable cuantitativa en una variable ordinal particionando el rango 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 sobre Y. El valor del coeficiente dependerá de la transformación (número de modalidades construidas). Se observa que ahora se tiene un coeficiente que no es símetrico en las variables como lo es el coeficiente de correlación lineal. Por lo cual obtendremos resultados distintos según la variable que transformamos, 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 tranformaciones influyen sobre los coeficientes de asociación.

1.4 Variables nominales

Para estudiar la relación entre dos variables nominales, se puede proceder como en el párrafo anterior: a partir de codificaciones de ambas variables o construyendo un índice.

1.4.1 Codificación de las dos variables nominales

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 empírico de las codificaciones

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

sea máximo.

Esta correlación se usa en análisis factorial de correspondancias y esta relacionada al χ^2 de contingencia.

1.4.2 χ^2 de contingencia

Los datos obtenidos sobre las dos variables nominales pueden resumirse en una tabla de contingencia sin perder información, salvo la identificación de las observaciones.
En la elección de consejales de 1991, se puede 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).

PARTIDO	I	II	III	IV	V	METR.	VI
D.C.	30142	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	13827	1848188
RADICAL	23416	61962	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

TABLA 6.3: Resultados de la elección de consejales de 1991

Sean diversos ejemplos de tablas de contingencia (Tablas 6.4 a 6.7) sobre dos variables X (en fila) e Y (en columna). Se observa en la tabla 6.4, que las columnas 1 y 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 ve 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 6.5. En efecto, si una observación toma la modalidad B_1, tomará la modalidad A_2 de X; dada A_1, entonces se tendrá la modalidad B_3 de Y, pero dada A_2, se tendrá B_1 o B_2. Se tiene entonces una relación funcional de Y hacia X, y existe una relación de X hacia Y, pero no de tipo funcional.
En el caso de la tabla 6.7 existe una relación funcional, pero no hay ninguna en la tabla 6.6.

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

TABLA 6.4

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

TABLA 6.5

	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	

TABLA 6.6

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

TABLA 6.7

Si denotamos n_{ij}, (i=1,..,p, j=1,...,q) los elementos de una tabla de contingencia, se tiene 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:

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

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

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

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

$$\frac{n_{1j}}{n_{1\bullet}} = \frac{n_{2j}}{n_{2\bullet}} = ... = \frac{n_{pj}}{n_{p\bullet}} = \frac{n_{\bullet j}}{n} \quad \text{para todo j=1,...,q}$$

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 son equivalentes a:

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

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 traduce 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 $\times$ Mín{p-1,q-1} cuando hay una relación estricta entre una variable con respecto a la otra.

1.4.3 Relación entre dos variables cuantitativas

Si transformamos las dos variables cuantitativas en variables nominales podremos usar el χ^2 de contingencia que nos permite 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 observará que las transformaciones producen variables menos precisas que las originales, pero con éstas se puede investigar relaciones no lineales.

1.5 Variables ordinales

1.5.1 Codificación de las dos variables ordinales

Como anteriormente vimos se puede construir codificaciones de ambas variables o índices.
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 de Y e Y^i, i=1,...,p, las indicadoras de Y.
Aquí las codificaciones deben respetar el orden definido sobre las modalidades. Entonces se busca las codificaciones que respetan los ordenes y tales que el coeficiente de correlación lineal empírico

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

sea máximo.
Este problema no es fácil de resolver en general.

1.5.2 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.*
Sea $x_1, ..., x_n$ 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} son los rangos asociadas 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 $x_i < x_j$
$S(x_i, x_j) =$ -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 *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 R_S, τ toma valores entre -1 y +1 y vale +1 si los órdenes son idénticos y -1 cuando son totalmente opuestos.

1.5.3 Relación entre dos variables cuantitativas

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

1.6 Inferencia

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

1.6.1 Coeficiente de correlación lineal

¿Cuándo se obtiene un coeficiente de correlación lineal r pequeño podemos admitir que la correlación ρ en la población es nula o si su valor no lo es, podemos concluir a una relación lineal?
Para responder 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. Para eso se requiere la distribución del coeficiente de correlación r.
Cuando $\rho = 0$ y las dos variables X e Y provienen de una distribución normal bivariada, entonces la distribución del coeficiente r de la muestra es fácil de obtener; éste depende del tamaño n de la muestra: existen tablas de la distribución de r en función de n y para n¿100 se puede aproximar 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, entonces vamos a rechazar que ρ=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 a 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 el par de variables no es 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, sólo que no hay ligazon lineal.

1.6.2 Razón de correlación

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 $H_0 : \eta = 0$.
Se supone entonces distribuciones condicionales de Y dada cada modalidad de X normales de misma media y misma varianza. Se considera entonces el estadístico:

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

que sigue una distribución de Fisher a p-1 y n-p grados de libertad bajo la hipótesis de independencia.

1.6.3 χ^2 de contingencia

¿Si dos variables nominales X e Y son independientes, cuales son los valores más probables del χ^2 de contingencia?

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

Si X e Y son independientes, $n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$ para todo par (i,j); en este caso el estadístico χ^2 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.

1.6.4 Coeficiente de correlación de rangos de Spearman

Cuando X e Y resultan de un ordenamiento sin empates, entonces los rangos inducidos por X o Y son valores de 1,..,n y los rangos de uno se obtienen por permutación de los rangos 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 los distintas permutaciones.
- Para n < 100, exite 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 $I\!P(|R_S| > 0.447) = 0.05$, lo que nos lleva a rechazar la independencia entre la Esperanza de Vida y la Tasa de Mortalidad.

1.6.5 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 6.1, 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.

1.7 EJERCICIO

(Se deja propuesto)
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 varian 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 con la partición del intervalo [-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 concluya: $\eta_{Y/X_1} = 0.96$, $\eta_{Y/X_2} = 0.93$, $\eta_{Y/X_3} = 0.86$ 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 $\chi^2_{X_4,Y_4} = 306$. Concluir.

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.

2 MODELO LINEAL

2.1 INTRODUCCIÓN

Estudiamos en el capítulo anterior como detectar una asociación entre dos variables; generalmente los roles de las variables no son simétricos - una variable puede influir sobre la otra y la recíproca no ser cierta - y más de una variable pueden intervenir en esta relación. Aquí nos interesamos no es solamente en evaluar la intensidad de la asociación, pero también, describir esta relación.

Algunas relaciones son fáciles a plantear y verificar - como las relaciones planteadas a partir de leyes físicas o mecánicas - pero cuando la aleatoriedad juega un papel importante, el estudio de las relaciones es más dificil. Se busca aquí descubrir como un conjunto de variables $X^1, X^2, ..., X^p$ - llamadas **variables explicativas o variables independientes o variables exógenas** - influye sobre otra variable Y - llamada **variable a explicar o variable repuesta o variable dependiente o variable endógena**. Cuando las variables son cuantitativas, se busca una función f que permita reconstituir los valores obtenidos sobre una muestra:

$$Y = f(X^1, X^2, ..., X^p)$$

Por rázon histórica, este análisis se llama **regresión**. Preferemos aquí hablar de **modelo**.

Ejemplo 1: La distancia d que una partícula recorre en el tiempo t esta dada por la fórmula:

$$d = \alpha + \beta t$$

en que β es la velocidad promedio y α la posición de la partícula en t=0. Si α y β son desconocidos, observando la distancia d en dos tiempos distintos, la solución del sistema de las 2 ecuaciones lineales permite obtener α y β. Sin embargo es dificil obtener en general la distancia sin error de medición ϵ que es de tipo aleatorio. Por lo cual se observa una variable aleatoria: $Y = d + \epsilon$ en vez de d. En este caso no basta tener dos ecuaciones pero valores de la distancia para varios valores del tiempo y métodos estadísticos basados en la aleatoriedad del error permiten estimar a α, β y d sobre la base de una relación funcional de tipo lineal.

Ejemplo 2: Si consideramos el peso P y la talla T de las mujeres chilenas, esta claro que no existe una relación funcional entre P y T, pero existe una cierta

tendancia. Considerando que P y T son variables aleatorias de distribución conjunta normal bivariada, se plantea el modelo lineal:

$$E(P/T) = \alpha + \beta T$$

en que α y β dependen de los parámetros de la distribución conjunta de P y T. El peso se escribe entonces:

$$P = \alpha + \beta T + \epsilon$$

en que ϵ refleja la variabilidad del peso P entre las chilenas de la misma talla con respecto a la media.
Ejemplo 3: Para decidir la construción de una nueva central eléctrica, ENDESA busca prever el consumo total de electricidad en Chile después del año 2000. Se construye un modelo que liga el consumo de electricidad con variables económicas y demográficas, que se estima en base a datos pasados. Se aplica entonces el modelo para predecir el consumo de electricidad según ciertas evoluciones económicas y demográficas.
Ejempo 4: Para establecer una determinada publicidad a la televisión, se cuantifica el efecto de variables culturales y socio-económicas sobre la audiencia de los diferentes programas.
Ejemplo 5: Ajuste polinomial
El modelo lineal puede ser generalizado tomando funciones de las variables explicativas y/o de la variable a explicar. Es el caso cuando se tiene una variable Y a explicar a partir de una sola variable X en un modelo polinomial:

$$Y = a_o + a_1 X^1 + ... + a_p X^p$$

en donde X^j es la potencia j de X.
Ejemplo 6: Se quiere estimar la constante g de la gravitación; se toma los tiempos de caida t de un objeto desde una distancia d dada del suelo.

$$d = \frac{1}{2} g t^2$$

Dados los errores de mediciones varias observaciones son necesarias y se puede considerar este modelo como lineal tomando como variable t^2.
Nos limitamos aquí a los modelos lineales, es decir que la variable repuesta se escribe como combinación lineal de las variables explicativas.

Observamos en los distintos ejemplos que las variables pueden ser aleatorias o no. En el caso que ninguna de las variables es aleatoria, se tiene un problema de ajuste y se presenta a continuación el método matemático de ajuste de los mínimos cuadrados, que permite estimar los coeficientes del modelo lineal a partir de valores observados.
Para el caso de variables aleatorias, se presenta en el párrafo siguiente, el método de máxima verosimilitud, que basado en un modelo probabilístico normal permite justificar el método de mínimos cuadrados y discutir las propiedades de los estimadores y la precisión del ajuste. Finalmente se usará el modelo para predicciones.
Se enfatizará los aspectos geométricos del problema y como hacer una crítica de los supuestos probabilísticos usuales.

2.2 SOLUCION DE LOS MINIMOS CUADRADOS

Sean $\{(y_i, x_i^1, x_i^2, ..., x_i^p), (i = 1, ..., n)\}$ los valores obtenidos sobre una muestra $p + 1$ dimensional de tamaño n. Se plantea el modelo lineal:

$$y_i = \beta_o + \beta_1 x_i^1 + ... + \beta_p x_i^p$$

Como generalmente no existen coeficientes que cumplen exactamente esta relación para todas las observaciones, se escribe:

$$y_i = \beta_o + \beta_1 x_i^1 + ... + \beta_p x_i^p + \epsilon_i$$

en lo cual ϵ_i es el **error** debido al modelo para la observación i, y se buscará minimizar una función de los errores, como por ejemplos:

- $\sum_i \epsilon_i^2$
- $\sum_i |\epsilon_i|$
- $\sum_i Max\{\epsilon_i\}$

El criterio de los mínimos cuadrados toma como función: $\sum_i \epsilon_i^2$ cuya solución es fácil de obtener y que tiene una interpretación geométrica simple.

Escribiremos matricialmente el modelo aplicado a la muestra de observaciones.

Sea $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & ... & x_1^p \\ 1 & x_2^1 & x_2^2 & ... & x_2^p \\ & ... & ... & ... & \\ 1 & x_n^1 & x_n^2 & ... & x_n^p \end{pmatrix}$, $\underline{\beta} = \begin{pmatrix} \beta_o \\ \beta_1 \\ ... \\ \beta_p \end{pmatrix}$, $\underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{pmatrix}$

Entonces el modelo se escribe:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

El criterio de los mínimos cuadrados consiste entonces en buscar el punto del subespacio vectorial W de $I\!R^n$ generado por las columnas de la matriz X el más cercano al punto $\underline{y}$. La solución es la proyección ortogonal del punto $\underline{y}$ sobre W.

En efecto, $\sum_i \epsilon_i^2$ es igual a $\|\underline{\epsilon}\|^2$, el cuadrado de la norma del vector $\underline{\epsilon}$, es decir el cuadrado de la distancia entre los vectores $\underline{y}$ y $X\underline{\beta}$, siendo $X\underline{\beta}$ un vector del subespacio vectorial W. Si P es el operador lineal de proyección ortogonal sobre el subespacio vectorial W, entonces la solución es $X\underline{\hat{\beta}} = P\underline{y}$.

La expresión matricial de P se puede obtener en función de la matriz X:

Como $\underline{\epsilon} = \underline{y} - P\underline{y}$ es ortogonal a W o sea que $\underline{y} - X\underline{\beta}$ es ortogonal a cada columna de X; si se denotan $X_o, X_1, ..., X_p$ las p+1 columnas de X, se expresa la ortogonalidad en términos de los p+1 productos escalares:

$$< \underline{y} - X\underline{\beta}, X_j > \qquad (j = 0, 1, ...p)$$

Matricialmente se escribe: $X_j^t(\underline{y} - X\underline{\beta}) = 0$ $(\forall j)$, y juntando las p+1 ecuaciones se obtiene las **ECUACIONES NORMALES**:

$$X^t X\beta = X^t y$$

Este sistema de ecuaciones lineales tiene una solución única cuando las columnas de X son linealmente independientes, es decir que forman una base del subespacio vectorial de W, o sea que X es de rango igual a p+1. En este caso la solución de los mínimos cuadrados es igual a:

$$\underline{\hat{\beta}} = (X^t X)^{-1} X^t \underline{y}$$

¡Se deduce que el operador de proyección ortogonal sobre W se escribe matricialmente como:

$$P = X(X^t X)^{-1} X^t$$

Este operador lineal es idempotente ($P^2 = P$) y simétrico ($P^t = P$).
Si el rango de X es inferior a p+1, basta encontrar una base de W entre las columnas de X, y remplazar X por la matriz formada de estas columnas linealmente independientes.

2.3 SOLUCIÓN DE MÁXIMA VEROSIMILITUD

En el párrafo anterior, se uso un criterio matemático para estimar los coeficientes β_j. Aquí usaremos un modelo probabilístico y el método de máxima verosimilitud para estimarlos. El modelo consiste en la esperanza condicional de la variable repuesta Y dadas las variables explicativas $X^1, X^2, ..., X^p$:

$$E(Y/X^1, X^2, ..., X^p) = \beta_o + \beta_1 X^1 + \beta_2 X^2 + ... + \beta_p X^p$$

con

$$Y = E(Y/X^1, X^2, ..., X^p) + \epsilon$$

en donde se supone que el error ϵ sigue una distribución normal de esperanza nula y de varianza σ^2.
Si ahora tenemos una muestra de tamaño n $\{(y_i, x_i^1, x_i^2, ..., x_i^p\}$, (i=1,...,n), en que las observaciones son independientes, entonces la función de verosimilitud condicional se escribe:

$$f_n(y/x^1, ..., x^p, \beta_o, \beta_1, ...\beta_p) = (\frac{1}{2\pi\sigma^2})^{n/2} exp\{-\frac{1}{2\sigma^2}\sum(y_i - \beta_o - \beta_1 x_i^1 - ... - \beta_p x_i^p)^2\}$$

La función es máxima cuando se cumplen las ecuaciones normales:

$$X^t X\underline{\hat{\beta}} = X^t \underline{y}$$

Además el estimador de máxima verosimilitud de σ^2 es:
$\frac{1}{n}\sum e_i^2 = \frac{1}{n}\sum_{i=1}^n (y_i - -\hat{\beta}_0 - \hat{\beta}_1 x_i^1 - ... - \hat{\beta}_p x_i^p)^2$.

2.4 PROPIEDADES DEL ESTIMADOR

Las propiedades del estimador $\underline{\hat{\beta}}$ son ligadas a los supuestos hechos sobre los errores ϵ_i. Supondremos aquí que X es de rango p+1 ($\underline{\hat{\beta}} = (X^t X)^{-1} X^t \underline{y}$).

- El estimador es insesgado: $E(\underline{\epsilon}) = \underline{0} \Longrightarrow E(\underline{\hat{\beta}}) = \underline{\beta}$

- El estimador es consistente.
- El estimador tiene mínima varianza:

  Teorema 2.1 *Teorema de GAUSS MARKOV:*

  Si $E(\underline{\epsilon}) = \underline{0}$ *y* $E(\underline{\epsilon}\underline{\epsilon}^t) = \sigma^2 I_n$, *entonces toda combinación lineal* $a^t\underline{\hat{\beta}}$ *de* $\underline{\hat{\beta}}$ *tiene mínima varianza entre los estimadores insesgados lineales en* $\underline{y}$ *de* $a^t\underline{\beta}$.

  Demostración del teorema de GAUSS MARKOV:

  Hay que comparar las varianzas de $a^t\underline{\hat{\beta}}$ y $a^t\underline{\beta^*}$ en que $\underline{\beta^*}$ es un estimador insesgado de la forma $C\underline{y}$.

  $\underline{\beta^*} = \underline{\hat{\beta}} + D\underline{y}$, en que $D = C - (X^tX)^{-1}X^t$.

  Como los dos estimadores son insesgados, $E(Dy) = 0$ y luego $DX = 0$.

  $Var(\underline{\beta^*}) = Var(\underline{\hat{\beta}}) + Var(D\underline{y}) + 2Cov(\underline{\hat{\beta}}, D\underline{y})$

  $Cov(\underline{\hat{\beta}}, D\underline{y}) = \sigma^2(X^tX)^{-1}X^tD^t = 0$

  $Var(\underline{\beta^*}) = Var(\underline{\hat{\beta}}) + \sigma^2 DD^t$

  Luego, $Var(a^t\underline{\beta^*}) = a^tVar(\underline{\hat{\beta}})a + \sigma^2 a^tDD^ta$

  Como $\sigma^2 a^tDD^ta > 0$, $Var(a^t\underline{\beta^*}) > Var(a^t\underline{\hat{\beta}})$

- La estimación de σ^2 obtenida por máxima verosimilitud es sesgada. En efecto, si $Q = I - P$, entonces $\underline{e} = Q\underline{y} = Q\underline{\epsilon}$. Luego, $\sum e_i^2 = \underline{e}^t\underline{e} = \underline{\epsilon}^tQ^tQ\underline{\epsilon} = \underline{\epsilon}^tQ\underline{\epsilon} = Traza(Q\underline{\epsilon}\underline{\epsilon}^t)$ Luego $E(\underline{e}^t\underline{e}) = Traza(QE(\underline{\epsilon}\underline{\epsilon}^t)) = \sigma^2 Traza(Q)$

  $Traza(Q) = Traza(I - X(X^tX)^1X^t) = n - Traza(I_{p+1}) = n - p - 1$

  Es decir que $E(\underline{e}^t\underline{e}) = (n - p - 1)\sigma^2$

  Se obtiene entonces un estimador insesgado de σ^2 tomando:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}(\underline{y} - X\underline{\hat{\beta}})^t(\underline{y} - X\underline{\hat{\beta}})$$

2.5 CALIDAD DEL MODELO

Los residuos e_i dan la calidad del ajuste para cada observación. Pero es una medida individual que depende de la unidad de medición. Un índice que evita este problema es:

$$\frac{\sum e_i^2}{\sum y_i^2}$$

que representa el cuadrado del coseno del ángulo del vector $\underline{y}$ con el vector $\underline{\hat{y}} = P\underline{y}$ en $\mathbb{R}^n$. Se puede comparar las varianzas:

- varianza residual: $(1/n)\sum e_i^2$
- varianza explicada: $(1/n)\sum(\hat{y}_i - \bar{y})^2$
- varianza total: $(1/n)\sum(y_i - \bar{y})^2$

El índice estadísticamente más interesante es **el coeficiente de correlación múltiple**:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

que compara el varianza explicada con la varianza total. La raíz cuadrada del coeficiente de correlación multiple (R) es el coeficiente de correlación lineal entre $\underline{y}$ e $\underline{\hat{y}}$; El valor de R esta comprendido entre 0 y 1. Cuando R=0, el modelo es E(y)= $\bar{y}$, la media muestral de los valores y_i, y cuando R es igual a 1, el vector $\underline{y}$ pertenece al subespacio vectorial W, es decir que existe un modelo lineal que permite escribir las observaciones y_i como combinación de las variables explicativas. Cuando R es cercano de 1, el modelo es bueno siendo los valores observados y_i vecinos de los valores estimados $\hat{y}_i$.
Si se plantea la hipótesis $H_o : \beta_1 = \beta_2 = ... = \beta_p = 0$, se usa el estadístico

$$F = \frac{\sum_i(\hat{y}_i - \bar{y})^2/(p+1)}{\sum_i e_i^2/(n-p-1)}$$

que bajo la hipótesis nula H_o sigue una distribución de F de Fisher a p+1 y n-p-1 grados de libertad.
Se observará que $F = \frac{R^2/(p+1)}{(1-R^2)/(n-p-1)}$

2.6 PREDICCIÓN

Si se tiene una nueva observación para la cual se conoce los valores de las variables explicativas, sean $x_o^1, x_o^2, ..., x_o^p$, pero se desconoce el valor y_o de la variable repuesta, se puede entonces usar el modelo para inferir un valor para y_o a través de su valor esperado:

$$\mu_o E(y_o) = \underline{x}_o^t \underline{\beta}$$

en que $x_o = (1, x_o^1, ..., x_o^p)$.
Se tiene un estimador insesgado de μ_o aplicando el modelo sobre los valores tomados por las variables explicativas debido a la nueva observación :

$$\hat{y}_o = \widehat{E(y_o)} = \underline{x}_o^t \hat{\underline{\beta}}$$

Se puede calcular un intervalo de confianza para μ_o. La distribución de $\hat{y}_o$ es $\mathcal{N}(\mu_o, \sigma_o^2(\underline{x}_o^t(X^tX)^{-1}x_o)$ luego Como y_o no depende del vector $\underline{y}$ que sirvió a estimar $\underline{\beta}$, y_o no depende de $\hat{y}_o$ tampoco. Además $E(\hat{y}_o) = E(y_o) - \underline{x}_o^t\beta = \mu_o$. Luego $\frac{\hat{y}_o - \mu_o}{\hat{\sigma} sqrt \underline{x}_o^t (X^tX)^{-1} \underline{x}_o} \sim t_{n-p-1}$ se usa este estadístico para construir un intervalo de confianza de nivel $1-\alpha$ para μ_o:

$$I\!P(\hat{y}_o - t_{\alpha/2}\hat{\sigma} sqrt \underline{x}_o^t (X^tX)^{-1} x_o < \mu_o < \hat{y}_o + t_{\alpha/2}\hat{\sigma} sqrt \underline{x}_o^t (X^tX)^{-1} x_o = 1 - \alpha$$

Un problema distinto es el de estimar un intervalo no para μ_o pero si para y_o. Hablaremos de un intervalo para la predicción. En este caso hay que tomar en cuenta la varianza de la v.a. y_o. $y_o = \hat{y}_o + \hat{\epsilon}_o$
La varianza de $\hat{\epsilon}_o$ es igual a:

$$\sigma^2 \underline{x}_o^t (X^tX)^{-1} \underline{x}_o + \sigma^2$$

dado que $\hat{y}_o$ no depende de y_o. Un intervalo de predicción apra y_o se obtiene entonces a partir de

$$\frac{y_o - \hat{y}_o}{\hat{\sigma} sqrt(1 + \underline{x}_o^t (X^tX)^{-1} \underline{x}_o)} \sim t_{n-p-1}$$

. El intervalo es entonces:

$$\hat{y}_o - t_{\alpha/2}\hat{\sigma} sqrt(1 + \underline{x}_o^t (X^tX)^{-1} \underline{x}_o),\ \hat{y}_o + t_{\alpha/2}\hat{\sigma} sqrt(1 + \underline{x}_o^t (X^tX)^{-1} \underline{x}_o)$$

2.7 ANÁLISIS DE LOS RESIDUOS

Dado que las propiedades del estimador dependen de ciertos supuestos, es importante comprobar si estas últimas se cumplen. Las propiedades fundamentales se plantean sobre los errores y la mejor forma de chequear si los errores son aleatorias de medias nulas, independientes y de misma varianza, es estudiando los residuos:

$$\hat{\epsilon}_i = y_i - \sum_j \hat{\beta}_j x_i^j \quad \forall i = 1, ..., n$$

Se puede usar el gráfico $(x_i, \hat{\epsilon}_i)$, que debería mostrar ninguna tendancia de los puntos, o bien test de hipótesis sobre los errores.

2.8 EJERCICIOS

1. Cuatro médicos estudián los factores que hacen esperar a sus pacientes en la consulta. Toman una muestra de 200 pacientes y consideran el tiempo de espera de cada uno el día de la consulta, la suma de los atrasos de los médicos a la consulta este mismo día, el atraso del paciente a la consulta este día (todos estos tiempos en minutos) y el número de médicos que están al mismo tiempo en la consulta este día. Se encuentra un tiempo promedio de espera de 32 minutos con una desviación típica de 15 minutos. Se estudia el tiempo de espera en función de las otras variables mediente un modelo lineal cuyos resultados estan dados a continuación:

VARIABLE	COEFICIENTE	DESV. /IPICA	/ ctudent	$\mathbb{P}(\mid X \mid > T)$
CONSTANTE	22.00	4.42	4.98	0.00
ATRASO MÉDICO	0.09	0.01	9.00	0.00
ATRASO PACIENTE	-0.02	0.05	0.40	0.66
NÚNERO MÉDICOS	-1.61	0.82	1.96	0.05

Coef. oe @orrelacióW lultiple $R^2 = 0.72$; F oe Fisher = 168 ; $\mathbb{P}(X > F) = 0.00$

1. Interprete los resultados del modelo lineal. Comente su validez global y la influencia de cada variable sobre el tiempo de espera. Especifique los grados de libertad de las t de Student y de la F de Fisher.

2. Muestre que se puede calcular la F de Fisher a partir del R^2. Si se introduce una variable explicativa suplementaria en el modelo, ¿ el R^2 sera más elevado?.

3. Dé un intervalo de confianza a 95% para el coeficiente del atraso médico.

4. Predecir el tiempo de espera, con un intervalo de confianza a 95%, para un nuevo paciente que llega a la hora un día que el consultorio funciona con 4 médicos que tienen respectivamente 10, 30, 0 y 60 minutos de atraso.

2. Suponga que tenemos un modelo lineal $Y = X\beta + \epsilon$ con $\epsilon \sim \mathcal{N}_n(0,\ \sigma^2 I)$, $\beta \in I\!R^p$, $X \in \mathcal{M}_{np}(I\!R)$.

1. Escribamos X como: $X = (X_1, X_2)$, con X_1 y X_2 submatrices de X tales que $X_1^t X_2 = 0$ (La matriz nula). El modelo inicial $Y = X\beta + \epsilon$ se escribe $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$ con $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$

   Si $\hat{\beta}_1$ es el estimador de máxima verosimilitud de α_1 en el modelo $Y = X_1\alpha_1 + \epsilon_1$ y $\hat{\beta}_2$ es el estimador de máxima verosimilitud de α_2 en el modelo $Y = X_2\alpha_2 + \epsilon_2$, muestre que el estimador de máxima verosimilitud de $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ en el modelo $Y = X\beta + \epsilon$ es $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$.
   (Ind.: Se usará el siguiente resultado: Si $A \in \mathcal{M}_{nn}(I\!R)$ es una matriz diagonal por bloque, i.e. $A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$, con las submatrices A_1 y A_2 invertibles, entonces A es invertible, y $A^{-1} = \begin{pmatrix} A_1^{-1} & 0 \\ 0 & A_2^{-1} \end{pmatrix}$).

2. Si $X_1^t X_2 \neq 0$ y si se toma $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ como estimador de β, que propiedad pierde bajo el supuesto usual $E(\epsilon) = 0$.

3. Consideramos tres variables Y, X, Z observadas sobre una muestra de tamaño n=40, $\{(y_i, x_i, z_i)\}$. Se busca explicar Y a partir de X y Z.

1. Se presentan los resultados de modelo lineal: $y_i = \alpha + \beta x_i + \epsilon_i$:

VARIABLE	MEDIA	DESV. TIPICA	ESTIMA-CION	DESV. TIPICA	T STUDENT	$I\!P(\mid X \mid > T)$
Y	11.68	3.46				
CONS-TANTE			7.05	1.03	6.84	0.00
X	5.854	2.74	0.79	0.16	4.94	0.00

Coef. de correlación multiple $R^2 = 0.39$; F de Fisher = 24.44 ; $I\!P(X > F) = 0.00$

Interprete estos resultados y efectue el test de hipótesis $H_o : \beta = 0$.

2. Dé una estimación insesgada para σ^2 la varianza de los errores de este modelo.

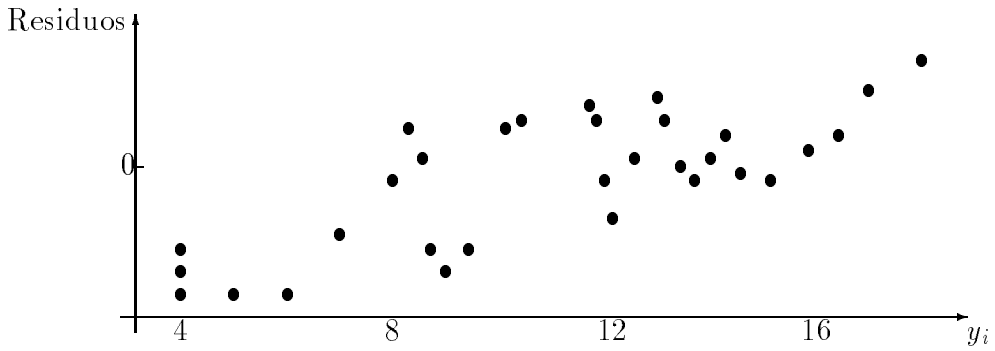
3. Comente el gráfico de los residuos en función de los y_i.

4. Se tiene una nueva observación que toma sobre la variable X el valor $x_o = 6.50$. Dé una estimación $\hat{y}_o$ del valor y_o que toma sobre la variable Y. Dé $Var(\hat{y}_o)$.

5. Se presentan los resultados del modelo lineal: $y_i = \delta + \gamma z_i + \epsilon_i$:

VARIABLE	MEDIA	DESV. TIPICA	ESTIMA-CION	DESV. TIPICA	T STUDENT	$I\!P(\mid X \mid > t)$
Y	11.68	3.46				
CONS-TANTE			11.68	0.36	32.54	0.00
Z	0.00	2.65	1.00	0.14	7.27	0.00

Coef. de correlación múltiple $R^2 = 0.58$; F de Fisher = 52.78 ; $I\!P(X > F) = 0.00$



Se tiene $\sum_i x_i z_i = 0$ y $\sum_i z_i = 0$.

Muestre que si $X_1 = (\boldsymbol{I}\ X)$ es la matriz formada del vector de unos y del vector de los x_i y $X_2 = Z$ el vector formado de los z_i, se tiene $X_1^t X_2 = 0$. Usando los resultados del ejercicio 2 deduzca las estimaciones de los parámetros del modelo $y_i = \beta_0 + \beta_1 x + \beta_2 z + \epsilon$.

4. Se quiere ajustar una función escalón $y = f(t)$ con f constante en los intervalos $I_j = (a_{j-1}, a_j]$ en que j=0, ..., K y $a_j \leq a_{j+1}\ \ \forall$ j. Para ello se observan datos (t_i, Y_i) ,i=1, ..., n. Se asume que los Y_i son independientes y que la distribución de Y_i es $N(f(t_i),\ \sigma^2)$.

1. Formule el problema anterior como un modelo lineal.

2. Obtenga la función ajustada por mínimos cuadrados.

3. Construya un intervalo de confianza para $\int_{a_0}^{a_K} f(t)dt$.

5. Sea Y $\in \mathbb{R}^n$ un vector aleatorio con E(Y)=μ y Var(Y)=σ^2I . Se considera el modelo lineal $Y = X\beta + \epsilon$, en que X=$(1_n, X_1, ..., X_p)$ y X es de rango completo. Llameremos W al subespacio de $\mathbb{R}^n$ generado por las columnas de X e $\hat{Y}$ al estimador de mínimos cuadrados de $\mu = E(Y)$.

1. Sea a $\in$ W y Δ_a la recta generada por a. Se define $H_a = \{z \in W / a^t z = 0\}$ el suplemento ortogonal de δ_a en W. Se tiene entonces la descomposición en suma directa ortogonal de W: $W = H_a \bigoplus \Delta_a$.

Muestre que el estimador de mínimos cuadrados Y^* de μ en H_a se escribe como
$Y^* = \hat{Y} - (\frac{a^t \hat{Y}}{a^t a})a$.

2. Si $b \in I\!R^n$, muestre que $Var(b^t Y^*) = Var(b^t \hat{Y}) - \sigma^2 \frac{(b^t a)^2}{a^t a}$.

3. Suponiendo que los errores son normales, dé la distribución de $\frac{\sum_{i=1}^n \epsilon_i^{*2}}{\sigma^2}$, en que $\epsilon_i^* = Y_i - Y_i^*$.

4. Se considera el caso particular $a = I_n$. Dé la distribución de $\frac{\frac{\sum Y_i^{*2}}{p}}{\frac{\sum \epsilon_i^2}{n-p}}$.

   Muestre que si las variables son centradas, $\hat{Y} = Y^*$.

1 ANALISIS DE DATOS MULTIDIMENSIONALES

1.1 INTRODUCCION

Vimos que es práctico asociar gráficos a la interpretación de los coeficientes de asociación empiricos; permiten visualizar la existencia de ligazon entre dos variables y de posibles tipologias de las observaciones, mientras que los coeficientes permiten medir el grado de relación. Pero la mayoria de los problemas envolucran más de dos variables. En el capitulo anterior, el modelo lineal permitio estudiar la relación de una variable a partir de un conjunto de variables explicativas. Veremos en este capitulo una forma de visualizar observaciones y variables para interpretar la estructura que contienen.

1.2 PLANTEAMIENTO GENERAL

En general un fenomeno se observa en varias dimensiones, lo que hace más complejo el estudio. Se busca entonces sintetizar los multiples aspectos del fenomeno en pocos valores. Es así que el objeto de un índice es reducir una realidad compleja a una sola dimensión, de manera a permitir comparaciones. Esta reducción es imposible sin deformar aspectos del fenomeno.
Sea la tabla de datos (Tabla 6.1) que contiene 6 variables socioeconomicas tomadas sobre 20 países de America Latina. Si queremos comparar los países tomando una o dos variables, se puede ordenar los países o graficarlos. Pero para las 6 variables, es más dificil hacerlo. El análisis en componentes principales permite hacerlo: en este metodo se propone un cambio de base, que permite una mejor descripción de los países y de los coeficientes de correlación entre las variables.
Si tuvieramos dos variables solamente - Esperanza de vida y tasa de mortalidad infantil - con el gráfico 8.1 tendriamos una buena herramienta para interpretar estos datos. Se observa, por ejemplo, que Bolivia tiene una alta mortalidad infantil y una baja esperanza de vida, mientras que en Costa Rica se da lo contrario; además se nota una relación lineal de pendiente negativa entre las dos variables (vimos en la tabla 6.2 del capitulo 6, que el coeficiente de correlación lineal es igual a -0.951).
Con tres o más variables, no se puede hacer tal representación gráfica, que seria en $\mathbb{R}^3$ o mayor dimensión. La idea del metodo es entonces hacer un cambio

de variables y, mediante aproximaciones, llevar a un conjunto de representaciones gráficas. Las nuevas variables -llamadas componentes principales- son índices que permiten interpretar mejor los datos.

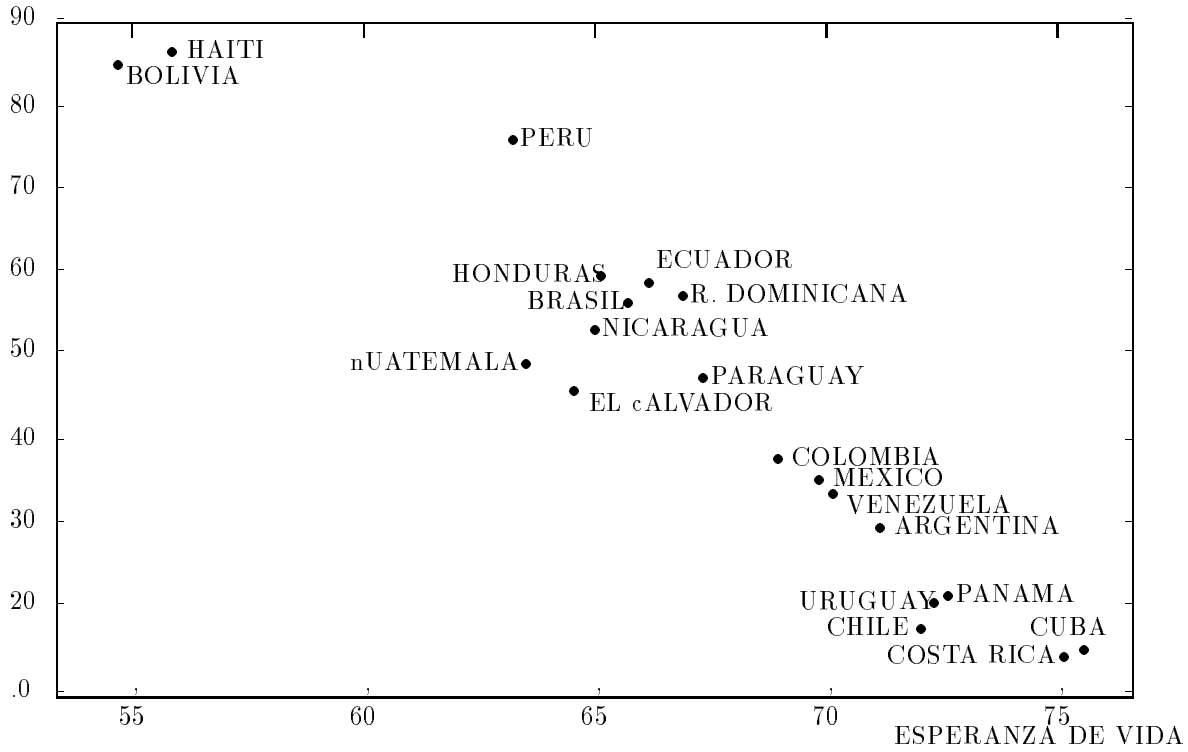


GRAFICO 8.1

1.3 EL ANALISIS EN COMPONENTES PRINCIPALES

Sea X la tabla de datos 6.1. Hay dos maneras de mirarla:

- Mirar las filas, que definen el conjunto de las observaciones

$$\mathcal{M} = \{\underline{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{i6} \end{pmatrix} \in I\!R^6\}$$

- Mirar las columnas, que definen el conjunto de las variables

$$\mathcal{N} = \{\underline{x}^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{20j} \end{pmatrix} \in I\!R^{20}\}$$

Se observara que si X es de rango igual a 1, entonces existe $\underline{c} \in I\!R^{20}$ y $\underline{u} \in I\!R^6$ tal que $X = \underline{c}\underline{u}^t$. En este caso, existe una recta en $I\!R^6$ que pasa por el origen, con $\underline{u}$ de vector director, a la cual pertenecen los puntos de $\mathcal{M}$; si $\underline{u}$ es unitario, las componentes c_i de $\underline{c}$ son las coordenadas de los países sobre esta recta. Sea $\|\underline{c}\|^2 = l$. Simétricamente, existe una recta en $I\!R^{20}$, que pasa por el origen, con $\underline{c}$ de vector director, a la cual pertenecen los puntos de $\mathcal{N}$. Si $\|\underline{c}\|^2 = l$, $u_j/\sqrt{l}$ son las coordenadas de las variables sobre esta recta. Si X es de rango igual a 2, entonces existe $\underline{c}_1$ y $\underline{c}_2 \in I\!R^{20}$ y $\underline{u}_1$ y $\underline{u}_2 \in I\!R^6$ tal que $X = \underline{c}_1\underline{u}_1^t + \underline{c}_2\underline{u}_2^t$. En este caso, los soportes de $\mathcal{M}$ en $I\!R^6$ y de $\mathcal{N}$ en $I\!R^{20}$ son planos. Más generalmente si X es de rango igual a r, entonces existe $\underline{c}_1, ..., \underline{c}_r \in I\!R^{20}$ y $\underline{u}_1, ..., \underline{u}_r \in I\!R^6$ tal que $X = \underline{c}_1\underline{u}_1^t + ... + \underline{c}_r\underline{u}_r^t$. En este caso, los soportes de $\mathcal{M}$ en $I\!R^6$ y de $\mathcal{N}$ en $I\!R^{20}$ son de dimensión r con estos vectores como vectores directores. El problemas es encontrar los vectores de tal descomposición.
Se distinguen las representaciones en $I\!R^6$ y en $I\!R^{20}$.

1.3.1 Representación en R^6

En este espacio los puntos son los 20 países. Para comparar dos países i e i', se considera la distancia entre las filas $\underline{x_i}$ y $\underline{x_{i'}}$ correspondientes:

$$d(i,i') = \sqrt{\sum_{j=1}^{j}(x_{ij} - x_{i'j})^2}$$

El calculo de esta distancia puede tener ciertos inconvenientes: la unidad de medición de las variables tiene un efecto, en el sentido que si multiplico por 10 una variable, por cambio de unidad, la distancia sera multiplicaado por 10 también. Se puede evitar este problema *normalizando* todas las variables, es decir tomandolas de varianza iguales a 1: si σ_j^2 es la varianza de la variable j ($\sigma_j^2 = (1/20)\sum_i(x_{ij} - \bar{x}^j)^2$), se tomara:

$$x_{ij}/\sigma_j$$

Para simplicar la notación, se supone que en la matriz X las variables son normalizadas.
Se busca entonces el vector $\underline{u} \in I\!R^6$ y el vector $\underline{c} \in I\!R^{20}$ tales que

$$X = \underline{c}\underline{u}^t + E$$

de manera que $(1/20)\sum_i^{20} \|\underline{x}_i - c_i\underline{u}\|^2$ sea mínimo con la restricción $\|\underline{u}\| = 1$. Si el rango de X es igual a 1, se obtendran los dos vectores buscados $\underline{c}$ y $\underline{u}$. La restricción $\|\underline{u}\| = 1$ se impone para tener unicidad de la solución y un vector director de la recta unitario. El criterio de optimización usado es un criterio de mínimos cuadrados que consiste a buscar la recta pasando por el origen tal que los puntos del conjunto $\mathcal{M}$ sean en promedio más cercanos a esta recta (Gráfico 8.2).

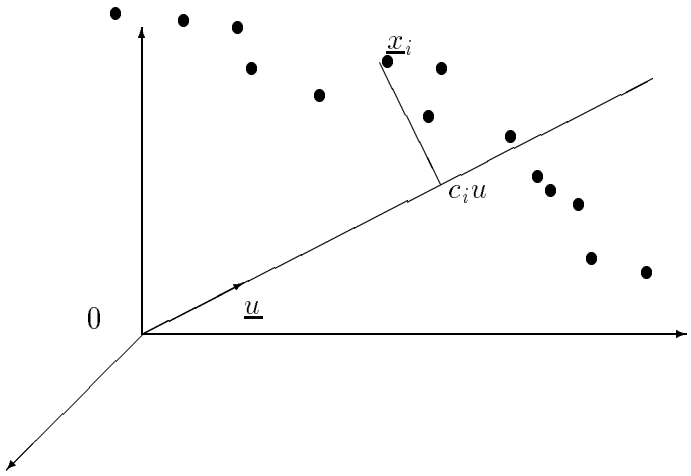


GRAFICO 8.2: Representación de las filas en $I\!R^6$

Ahora bien es un inconveniente de considerar una recta que pasa por el origen. En efecto se observa en el gráfico 8.3 que la recta H' es mejor que la recta H. Si no se impone que la recta pasa por el origen, es facil mostrar que la recta solución pasa por el punto medio $\underline{g} \in I\!R^6\}$ de $calM$: $\underline{g} = \begin{pmatrix} \bar{x}^1 \\ \bar{x}^2 \\ \\ \cdots \\ \bar{x}^6 \end{pmatrix}$, en que $\bar{x}^j$ es

la media de la variable j. En efecto, si δ es una recta pasando por el origen y δ' la recta paralela a δ pasando por $\underline{g}$ y si h_i y h_i' son las proyecciones ortogonales respectivas de $\underline{x}_i$ sobre δ y δ', entonces $\sum \|\underline{x}_i - h_i'\|^2 < \sum \|\underline{x}_i - h_i\|^2$.
De aquí se toma el origen del sistema de referencia en el punto medio, es decir $\underline{g} = \underline{0}$. Se supone entonces que en la matriz X, las columnas suman 0: $\sum_i x_{ij} = 0$ (las medias son todas nulas).

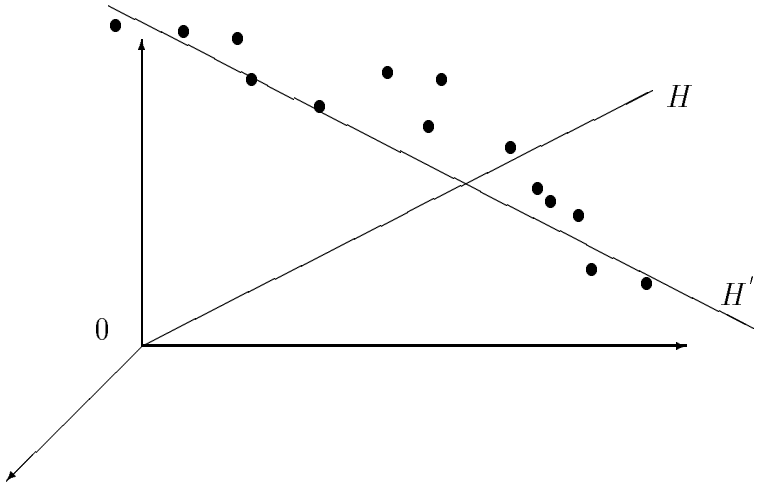


GRAFICO 8.3: Representación de las filas en $\mathbb{R}^6$

En este caso, el criterio de minimos cuadrados es equivalente a maximizar $(1/20)\sum_i \|c_i\underline{u}\|^2$ con $c_i = \underline{x}_i^t\underline{u} = \underline{u}^t\underline{x}_i$. El criterio con la restricción de norma nos da entonces:

$$Q = (1/20)\sum_i \|c_i\underline{u}\|^2 - l(\|\underline{u}\|^2 - 1)$$

$$Q = (1/20)\sum_i c_i^2 - l(\sum_j u_j^2 - 1)$$

$$Q = (1/20)\underline{u}^t(\sum_i \underline{x}_i\underline{x}_i^t)\underline{u} - l(\sum_j u_j^2 - 1)$$

$$Q = (1/20)\underline{u}^t X^t X\underline{u} - l(\sum_j u_j^2 - 1)$$

Sea $V = (1/20)X^tX = (v_{jk})$, $Q = \sum_{jk} u_j u_k v_{jk} - l(\sum u_j^2 - 1)$

$$\frac{\partial Q}{\partial u_j} = 2\sum_k v_{jk}u_k - 2lu_j = 0$$

Se deduce que $V\underline{u} = l\underline{u}$, es decir que el vector $\underline{u}$ es vector propio de la matriz $V = (1/20)X^tX$. Se observara que la matriz V es igual a la matriz de correlaciones asociada a la matriz X (Tabla 8.1) o a la matriz de covarianza cuando las variables no son normalizadas. Esta matriz es simétrica semi-definida positiva: tiene sus valores propios reales no negativos (más aún la suma de los valores propios es igual al número de variables, 6 aquí). Pero no se sabe cual de los vectores propios tomar. Observando que se busca maximizar y que $l = \sum_i c_i^2$, se concluye que hay que tomar un vector propio normalizado asociado al mayor valor propio de V. Llamamos l_1 el mayor valor propio de V, $\underline{u}_1$ el vector propio asociado y $\underline{c}_1 = X\underline{u}_1$. Si X es de rango igual a 1, l_1 es el único valor propio no nulo de V y los puntos $\underline{x}_i$ son alineados en $\mathbb{R}^6$. Si X es de rango mayor que 1, podemos repetir la descomposición a la matriz $Y = X - \underline{c}_1\underline{u}_1^t$. La matriz Y^tY tiene los mismos valores propios no nulos que X^tX salvo l_1. Luego la descomposición solución esta dada por el vector propio normalizado $\underline{u}_2$ asociado a l_2 el segundo mayor valor propio de V, y $\underline{c}_2 = X\underline{u}_2$:

$$X = \underline{c}_1\underline{u}_1 + \underline{c}_2\underline{u}_2^t + E$$

Generalizando, si $l_1 \geq l_2 \geq ... \geq l_r > 0$, $\underline{u}_1, \underline{u}_2, ..., \underline{u}_r$ los vectores propios normalizados asociados y $\underline{c}_k = X\underline{u}_k$ $k = 1, ..., r$, se puede descomponer:

$$X = \underline{c}_1\underline{u}_1^t + \underline{c}_2\underline{u}_2^t + ... + \underline{c}_r\underline{u}_r^t$$

en donde las matrices $\underline{c}_k\underline{u}_k^t$ son de rango 1 y de importancia decreciente en la reconstitución de la matriz X (Tabla 8.2).

La matriz de correlación V siendo simétrica semidefinida positiva, existe una base ortonormal de $\mathbb{R}^6$ formada de vectores propios de V. Luego, $\{\underline{u}_1, \underline{u}_2, ..., \underline{u}_r\}$ es una base ortonormal del espacio que contiene al conjunto $\mathcal{M}$.

Además se observa que los vectores $\underline{c}_k$ son vectores propios de la matriz $(1/20)XX^t$, que tiene los mismos valores propios no nulos que X^tX. En efecto, $\underline{c}_k = X\underline{u}_k$, luego

$$(1/20)X^tX\underline{u}_k = l_k\underline{u}_k$$

$$(1/20)X^t \underline{c}_k = l_k \underline{u}_k$$
$$(1/20)XX^t \underline{c}_k = l_k \underline{c}_k$$

Además $\|\underline{c}_k\|^2 = l_k$ (Se deja mostrarlo como ejercicio).

1.3.2 Representación en $\mathbb{R}^{20}$

En $\mathbb{R}^{20}$, se quiere comparar las columnas de X, que representan las variables, lo que equivale a tomar la matriz X^t en vez de X. El criterio de minimos cuadrados consiste ahora en buscar un vector $\underline{d} \in \mathbb{R}^{20}$ normalizado tal que:

$$(1/20)\sum_j^6 \|\underline{x}^j - v_j \underline{d}\|^2$$

sea mínimo.
Se tiene $v_j = d^t \underline{x}^j$ con $\|d\| = 1$.
Se obtiene que $\underline{d}$ es el vector propio normalizado de XX^t asociado al mayor valor propio l_1. Luego $\underline{d}$ es colineal al vector $\underline{c}_1$ obtenido en el estudio en $\mathbb{R}^6$: $\underline{c}_1 = \sqrt{l_1}\underline{d}$. Los vectores $\underline{u}_1$ y $\underline{v}$ son colineales también: $\underline{v} = \sqrt{l_1}\underline{u}_1$.
Interpretaremos el criterio en el caso de la representación e $\mathbb{R}^{20}$. El criterio de mínimos cuadrados es equivalente a maximizar $\Omega = (1/20)\sum_j \|v_j \underline{d}\|^2 = (1/20)\sum_j v_j^2$. Como $v_j = \underline{d}^t \underline{x}^j$ se obtiene que $\Omega = \sum_j (\underline{d}^t \underline{x}^j)^2$. Como las variables son centradas y normalizadas $\underline{d}^t x^j = Cor(\underline{d}, \underline{x}^j)$, luego el criterio usado aquí consiste en buscar una variable $\underline{d}$ de varianza igual a 1, combinación lineal de las variables $\underline{x}^j$ de tal forma que

$$\sum_j cor^2(\underline{d}, \underline{x}^j)$$

sea máxima. De hecho vimos en el capítulo 6 que el coeficiente de correlación permite comparar dos variables. Muestre como ejercicio que si dos variables son centradas y normalizadas entonces el coeficiente de correlación es igual al coseno del angulo que forman en $\mathbb{R}^{20}$ (Gráfico 8.4).
Además como los vectores $\underline{d}_k$ forman una base ortonormal, se deduce que las nuevas variables, que son las componentes principales no son correlacionadas entre si.

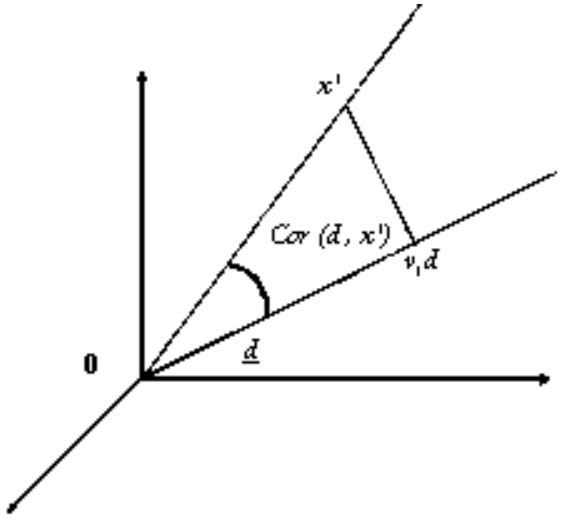


GRAFICO 8.4: Representación de las variables en $\mathbb{R}^{20}$

1.3.3 Interpretación

Veamos como usar estos resultados para interpretar el contenido de la tabla 6.1 (Se centra y normaliza los datos).
$\{\underline{u}_1, \underline{u}_2, ..., \underline{u}_r\}$ forma una base ortonormal de espacio que contiene al conjunto $\mathcal{M}$ de los países. Los ejes definidos por estos vectores se llaman **ejes principales**. Las coordenadas de los países sobre estos ejes son dadas por los vectores $\underline{c}_k$, llamados **componentes principales**, se habla también de **factores**. Si nos limitamos a tomar el primer eje principal definido por $\underline{u}_1$ (tabla 8.2), se obtiene una representación unidimensional de los países; es la mejor representación unidimensional, en el sentido que deforma menos las distancias mutuales entre los países. Aún si es una representación aproximada, tiene la ventaja de permitir una interpretación mucho más simple que la representación original. Las coordenadas de los países sobre este eje constituyen la primera componente principal $\underline{c}_1$ (Tabla 8.3). El valor más elevado lo tiene CUBA y el más bajo HAITI. Observando que

$$\underline{c}_1 = 0.3810\times\underline{x}^1 + -0.4364\times\underline{x}^2 + -0.2361\times\underline{x}^3 + 0.4590\times\underline{x}^4 + -0.4531\times\underline{x}^5 + -0.4389\times\underline{x}^6$$

se ve que la primera componente principal es una combinación lineal de las variables iniciales con algunos coeficientes mayores que otros y algunos positivos y otros negativos. EL PORCENTAJE DE POBLACION URBANA

y LA ESPERANZA DE VIDA tienen un coeficiente positivo, mientras que los otros son negativos. Lo que permite de interpretar la primera componente principal como un índice demografico, que crece con la calidad. Este índice es más manipulable que las seis variables originales. Ahora bien que cantidad de la información contenida en la tabla X perdimos o conservamos en el índice. En la decomposición: $\underline{x}_i = c_i \underline{u}_1 + \underline{e}_i$, $\underline{e}_i$ representa el error de representación de $\underline{x}_i$ sobre el primer eje principal. El valor propio $l_1 = \sum_i c_i^2$ mide la varianza de la componente principal $\underline{c}_1$ y $TrazaV - l_1 = \sum_{k=2}^{r} l_k = 6 - \_1$ mide el error global de la representación sobre el primer eje principal. Como $TrazaV = (1/20) \sum_i \|\underline{x}_i\|^2$ representa la varianza total en $\mathbb{R}^6$, se usa un índice de calidad de la representación de $\underline{c}_1$ con el porcentaje de varianza reproducida por $\underline{c}_1$:

$$100 \frac{l_1}{TrazaV}$$

que aquí vale 69.24%. Se puede considerar 2, 3 o más ejes principales para tener una mejor representación. Por ejemplo, con los dos primeros ejes principales se puede vizualisar los países (Gráfico 8.5) en un sistema cartesiano. En este gráfico cada país i tiene por coordenadas (c_{1i}, c_{2i}) y como los ejes son ortogonales, la varianza reproducida por el plano es igual a

$$100 \frac{l_1 + l_2}{TrazaV}$$

que aquí vale 88.53%.
Se nota en la tabla 8.4 que la representación con 4 ejes principales contiene casi integralmente los países (99.23%). En el gráfico de los dos primeros ejes principales (Gráfico 8.5) se proyectaron además los ejes iniciales, lo que permite explicar las diferencias y semejanzas entre los países. Es así que ARGENTINA y GUATEMALA difieren más por las variables % POBLACION URBANA, TASA NATALIDAD y FECUNDIDAD, que las variables de MORTALIDAD y ESPERANZA DE VIDA. Mientras que PANAMA y HAITI difieren más por la MORTALIDAD.
De la misma manera que se hizo una representación plana aproximada de la representación en $\mathbb{R}^6$, se hace una representación aproximada de las variables en $\mathbb{R}^{20}$, considerando las proyecciones de las variables $\underline{x}^j$ sobre los vectores $\underline{d}_1$ y $\underline{d}_2$ (Gráfico 8.6). Dado que las variables $\underline{x}^j$ y $\underline{d}_1$ y $\underline{d}_2$ son de varianza igual a 1, la proyección de $\underline{x}^j$ sobre $\underline{d}_1$ ($\underline{d}_2$) es igual al coeficiente de correlación entre

$\underline{x}^j$ y $\underline{c}_1$ ($\underline{c}_2$) (Tabla 8.4). Este gráfico permite entonces interpretar las componentes principales. Se observa que la primera componente principal tiene una correlación igual a 0.935 con la ESPERANZA DE VIDA, pero solamente -0.481 con la TASA DE MORTALIDAD, mientras que la segunda componente principal tiene una correlación igual a -0.267 con la ESPERANZA DE VIDA y 0.815 con la TASA DE MORTALIDAD.
Como las variables $\underline{x}^j$ tienen una varianza igual a 1, sus proyecciones en el plano caen al interior de un circulo de centro 0 y de radio 1. Si la proyección de la variable $\underline{x}^j$ es sobre la circunferencia del circulo, significa que x^j pertenece a este plano, es decir que $\underline{x}^j$ puede ser reproducida a partir de $\underline{c}_1$ y $\underline{c}_2$. La distancia de la proyección de una variable al origen mide la calidad de representación de la variable en el plano principal. Más aún es igual al coeficiente de correlación múltiple entre la variable con respecto a $\underline{c}_1, \underline{c}_2$ (Se deja como ejercicio la demostración). Aquí, las seis variables son bastante bien representada en el plano principal.
Como los cosenos de los angulos son iguales al los coeficientes de correlación, se tiene también una vizualisación, aproximada, de la matriz de correlaciones (Tabla 8.1). FECUNDIDAD y TASA DE NATALIDAD hacen un angulo pequeño, son altamente correlacionados (0.972), ESPERANZA DE VIDA y MORTALIDAD INFANTIL, que forman un angulo vecino de π, son altamente correlacionados negativamente (-0.951) y TASA DE MORTALIDAD y TASA DE NATALIDAD, que son casi ortogonal, son muy poco correlacionados (0.101).
Se puede completar el estudio haciendo representaciones planas con otros pares de ejes principales y las componentes principales correspondientes.

VARIABLES	1	2	3	4	5	6
1 % POB. URBANA	1.0	-.739	-.179	.588	-.735	-.532
2 TASA .NATALIDAD	-.739	1.0	.101	-.723	.972	.682
3 TASA MORTALIDAD	-.179	.101	1.0	-.609	.262	.533
4 ESPERANZA VIDA	.588	-.723	-.609	1.0	-.769	-.951
5 FECUNDIDAD	-.735	.972	.262	-.769	1.0	.709
6 MORTAL. INFANTIL	-.532	.682	.533	-.951	.709	1.0

TABLA 8.1: Matriz de correlaciones

	MEDIA	D. TIPICA	$\underline{u}_1$	$\underline{u}_2$	$\underline{u}_3$	$\underline{u}_4$
VALORES PROPIOS			4.15	1.16	0.41	0.24
% POB. URBANA	62.87	17.26	0.3810	0.3203	0.7699	0.3797
TASA NATALIDA	28.86	6.64	-0.4364	-0.3742	0.1920	0.3201
TASA MORTALIDAD	7.11	1.85	-0.2361	0.7567	-0.3904	0.4068
ESPERANZA VIDA	67.11	5.52	0.4590	-0.2479	-0.2093	0.2282
FECUNDIDAD	3.61	0.96	-0.4531	-0.2488	0.0859	0.5245
MORTAL.INFANTIL	44.54	22.28	-0.4389	0.2405	0.3779	0.5102

TABLA 8.2: Tres primeros vectores propios normalizados de la matriz de correlación

	$\underline{u}_1$	$\underline{u}_2$	$\underline{u}_3$	$\underline{u}_4$
VALORES PROPIOS	4.15	1.16	0.41	0.24
ARGENTINA	1.9029	1.3903	-.0092	0.5067
BOLIVIA	-3.1987	1.1181	.4426	-.4150
BRASIL	.2766	.8794	.6930	-.2980
COLOMBIA	1.1429	-.1561	.2443	-.3948
COSTA RICA	1.8937	-1.9713	-.6418	-.3264
CHILE	2.3727	.2178	.2529	.3686
ECUADOR	-.7182	-.1958	.1244	-.2654
EL SALVADOR	-1.1377	-.5743	-.5380	-.1134
GUATEMALA	-2.3926	-.9928	-.3855	.8457
HAITI	-4.0755	1.6465	-1.0408	-.1406
HONDURAS	-2.0627	-.8537	-.1854	.2504
MEXICO	1.0889	-.5733	.4780	-.1113
NICARAGUA	-1.8157	-.9593	.6089	.9066
PANAMA	1.5675	-1.0166	-.7486	-.4131
PARAGUAY	-.8912	-.9583	-.3161	.0197
PERU	-.8602	.8557	.9236	-.6187
REP.DOMINICANA	-.0239	-.1564	.2586	-.7430
URUGUAY	2.3890	2.2318	-.7393	.6895
VENEZUELA	1.3812	-.3757	1.2787	.5232
CUBA	3.1611	.4441	-.7003	-.2704

TABLA 8.3: Tres primeras componentes principales

	MEDIA	D. TIPICA	FACTOR .	FACTOR 2	FACTOR 3	FACTOR 4
VALORES PROPIOS			4.15	..16	'41	0.24
% ACUMULADO DE LA VARIABILIDAD			69.24	88.53	95.22	99.23
% POB. URBANA	62.87	.7.26	0.776	0.345	0.493	0.186
TASA NATALIDA	28.86	6.64	-0.889	-0.403	0.123	0.156
TASA MORTALIDAD	7.11	..85	-0.481	0.815	-0.250	0.199
ESPERANZA VIDA	67.11	5.52	0.935	-0.267	-0.134	0.111
FECUNDIDAD	3.61	0.96	-0.923	-0.268	0.055	0.256
MORTAL.INFANTIL	44.54	22.28	-0.894	0.259	0.242	-0.249

TABLA 8.4: Coordenadas de las variables sobre los 4 primeros factores (r_{jk})

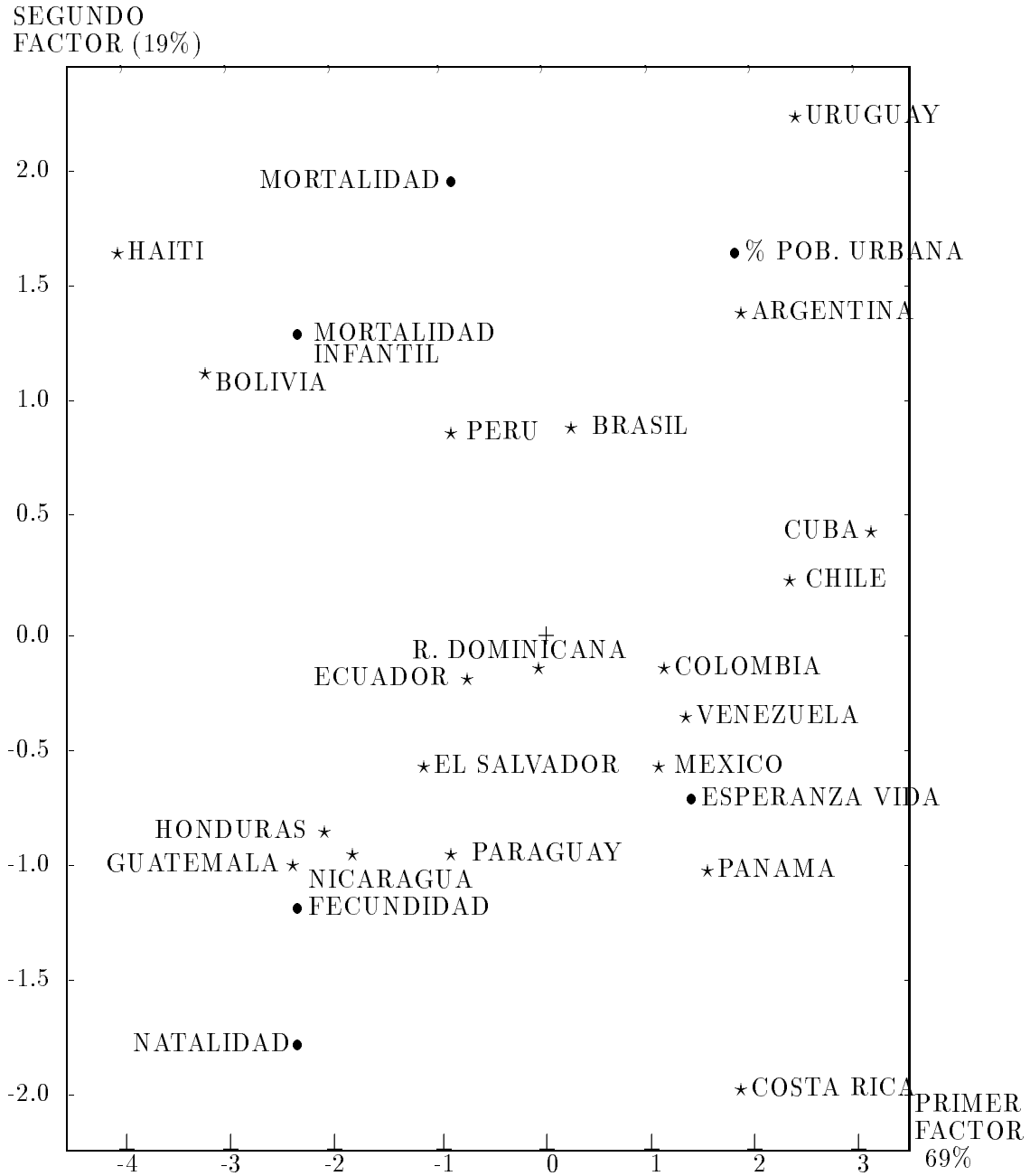


GRAFICO 8.5: Primer plano principal

GRAFICO 8.6: Círculo de correlaciones

1.3.4 Puntos suplementarios

Es interesante de representar a posteriori algunas observaciones o variables que no fueron incluídas en la matriz X originalmente. Sea un país x_o, su proyección sobre el eje principal k es igual a $\underline{x}_o^t \underline{u}_k$. Para una nueva variable $\underline{z}$, su proyeción sobre la componente principal k es igual a $Cor(\underline{z}, \underline{c}_k)$.
Consideramos, por ejemplo, dos países africanos -TUNEZ y EGIPTO- (Tabla 8.5), la coordenada de TUNEZ en el plano son (F_1, F_2) con
$F_1 = 0.3810 \times (-0.514) - 0.4364 \times 0.021 - 0.2361 \times (-0.059) + 0.4590 \times (-0.074) +$
$-0.4531 \times 0.094 + -0.4389 \times 0.155 = -0.335$
$F_2 = 0.3203 \times (-0.514) - 0.3742 \times 0.021 + 0.7567 \times (-0.059) - 0.2479 \times (-0.074) +$
$-0.2488 \times 0.094 + 0.2405 \times 0.155 = -0.18$
Para EGIPTO, se obtiene de la tabla 8.5: $F_1 = -0.18$ y $F_2 = 0.96$. Si se ubican estos dos países en el gráfico 8.5, encontramos TUNEZ cercano de R. DOMINICANA y EGIPTO cercano de BOLIVIA.

Consideramos ahora cuatro nuevas variables cuyos coeficientes de correlación con las dos primeras componentes principales son dados en la tabla 8.6. Las variables GASTO MILITAR y GASTO EN EDUCACION son muy poco correlacionados con estas componentes principales, se podria prever que un modelo lineal de estas variables sobre las seis variables originales no seria bueno. No es el caso de las dos otras variables suplementarias.

	TUNEZ		EGIPTO			
	x	$(x-\bar{x})/\sigma$	x	$(x-\bar{x})/\sigma$	$\bar{x}$	σ
% POB. URBANA	54	-0.514	47	-0.92	62.87	17.26
TASA NATALIDA	29	0.021	33	0.62	28.86	6.64
TASA MORTALIDAD	7	-0.059	10	1.56	7.11	1.85
ESPERANZA VIDA	66.7	-0.074	60.3	-1.23	67.11	5.52
FECUNDIDAD	3.7	0.094	4.3	0.72	3.61	0.96
MORTAL.INFANTIL	48.0	0.155	61.0	0.74	44.54	22.28

TABLA 8.5: Valores de las variables para TUNEZ y EGIPTO

	FACTOR 1	FACTOR 2
PNB	0.814	0.130
GASTO EN EDUCACION	-0.140	0.163
GASTO MILITAR	-0.378	-0.061
ALFABETISMO	0.839	0.021

TABLA 8.6: Coeficientes de correlación

1.4 EJERCICOS

1. Sea X la tabla siguiente: $\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$

Consideremos los seis vectores $\mathcal{M} = \{\underline{x}_1, ..., \underline{x}_6\}$ de $\mathbb{R}^3$ dotado de la métrica euclidiana usual cuyas componentes están dadas por las filas de la matriz X.
a) Muestre que la nube de los 6 puntos en $\mathbb{R}^3$ está centrada en el origen. Calcule $V = (1/6)\sum_i \underline{x}_i \underline{x}_i^t$.

b) Calcule I_0, el momento de inercia de $\mathcal{N}$ con respecto al origen. Compare con $TrazaV$.
c) Determine los diferentes valores propios de V.
d) Dé el vector propio asociado al valor propio nulo de V.
e) Determine dos vectores propios ortonormales de V asociado con los valores propios no nulos de V.
2. Se consideran V_1, V_2, V_3 y V_4, cuatro variables obtenidas sobre 20 observaciones repartidas en 3 clases (A, B y C) (Tabla 8.7).
a) Los resultados del análisis en componentes principales efectuado sobre las variables V_1, V_2 y V_3 con la matriz de correlaciones (tabla 8.9) están dados en el gráfico 1 y la tabla 8.8. Justifique la calidad de la representación en el plano y comente el gráfico 8.7.
b) A partir de la tabla 8.8, dibuje y comente el círculo de correlaciones.
c) En la tabla 8.10, se dan las correlaciones entre las dos componentes principales y la variable V_4. Represente gráficamente V_4 en el círculo de correlaciones.
d) Se quiere efectuar la regresión múltiple de V_4 sobre V_1, V_2 y V_3. ¿Qué problema numérico se va a presentar?
e) Deduzca de la tabla 4 el coeficiente de correlación múltiple de la regresión de V_4 sobre V_1, V_2 y V_3.
f) Deduzca de la tabla 8.10 los coeficientes de la regresión de V_4 sobre las dos componentes principales (la media de V_4 es 242.5 y la desviación tipica es 57.73).

CLASE	V_1	V_2	V_3	V_4	CLASE	V_1	V_2	V_3	V_4
C	45	25	30	160	C	60	27	13	350
C	40	30	30	200	B	38	37	25	240
C	32	32	36	210	B	35	38	27	220
C	35	28	37	250	B	22	38	40	180
C	50	33	17	260	A	18	33	49	190
B	55	45	0	300	B	15	39	46	185
B	58	35	7	320	A	20	40	40	300
C	62	28	10	310	A	25	35	40	220
B	48	32	20	280	A	22	33	45	225
B	52	34	14	300	C	32	26	42	150

TABLA 8.7: Tabla de datos

	MEDIA	DESVIACION TIPICA	FACTOR 1	FACTOR 2
VALORES PROPIOS	1.956	1.044		
% ACUMULADOS DE LOS VALORES PROPIOS	65.20	100.00		
V_1	38.20	14.98	0.997	0.076
V_2	33.40	5.18	-0.189	-0.982
V_3	28.40	14.50	-0.962	0.273

TABLA 8.8: Correlaciones de las variables con las Componentes Principales

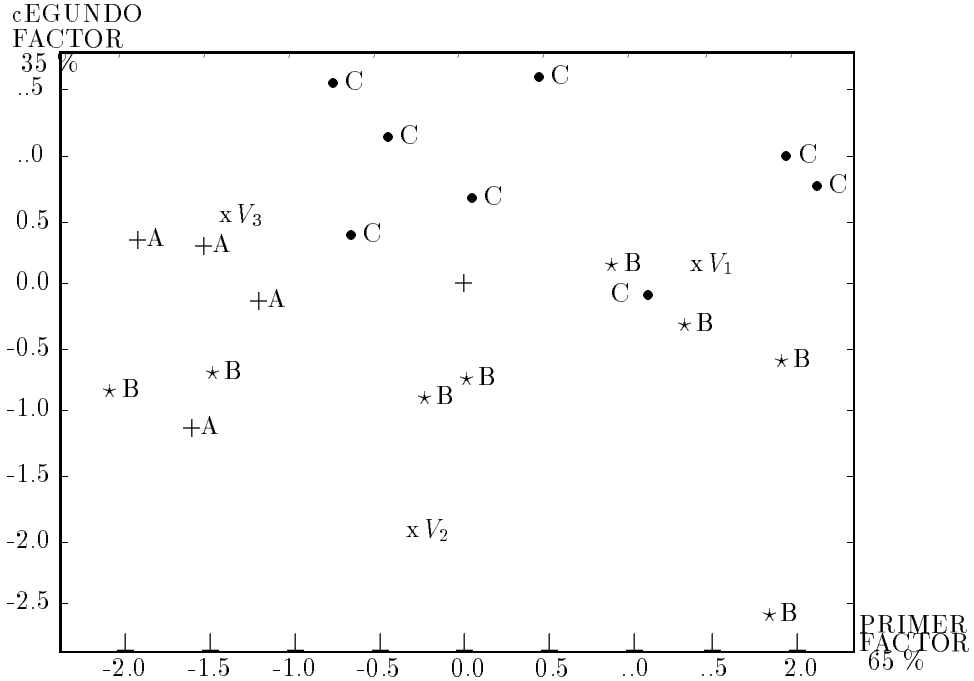


GRAFICO 8.7

	V1	V2	V3
V_1	1.00	-.26	-.94
V_2	-.26	1.00	-.09
V_3	-.94	-.09	1.00

TABLA 8.9: Matriz de correlaciones

	C.P. 1	C.P. 2	V4
C.P. 1	1.00	-.00	.69
C.P. 2	-.00	1.00	-.29
V4	.69	-.29	1.00

TABLA 8.10: Matriz de correlaciones

3. Sea $\mathcal{M} = \{\underline{x}_1, ..., \underline{x}_n\}$ un conjunto de n puntos de $\mathbb{R}^p$. Cada punto $\underline{x}_i$ tiene un peso p_i, con $p_i > 0, \sum p_i = 1$. Se supone que el centro de gravedad de $\mathcal{M}$

es $\underline{g} = 0$ y que la matriz de varianzas-covarianzas asociadas es $V = X^t D_p X$ de rango p con $D_p = diag(p_i)$. Sea $I\!R^p = W_1 \bigoplus W_2$ y sean P_1 y P_2 los proyectores ortogonales sobre W_1 y W_2 respectivamente.
a) Dé las matrices de varianzas-covarianzas V_1 y V_2 de los conjuntos $\mathcal{M}_1 = \{P_1\underline{x}_1, ..., P_1\underline{x}_n\}$ y $\mathcal{M}_2 = \{P_2\underline{x}_1, ..., P_2\underline{x}_n\}$.
b) Muestre que $V = V_1 + V_2 \Longleftrightarrow W_1 \perp_{V-} W_2$
c) Pruebe que: $[W_2 \perp_{V-} W_1] \Longleftrightarrow [V\underline{u} = l\underline{u} \Rightarrow \underline{u} \in W_1 \bigcup W_2]$
4. Sean E y F dos espacios vectoriales de dimensiones respectivas p y n. Se tiene en E y F las métricas euclidianas usuales. Sea S una aplicación lineal de E en F tal que si $\underline{y}_1 = S(\underline{x}_1)$ e $\underline{y}_2 = S(\underline{x}_2)$, entonces $\|\underline{y}_1 - \underline{y}_2\|^2 = \|\underline{x}_1 - \underline{x}_2\|^2$ para todo $\underline{x}_1, \underline{x}_2 \in E$.
a) Dar la relación que cumple S.
b) Sea $E = E_1 \oplus E_2$ con E_2 suplemento ortogonal de E_1. Sea A el proyector ortogonal sobre E_1 y S la aplicación de simetría respecto de E_2: $\underline{y} = S(\underline{x}) = -\underline{x}_1 + \underline{x}_2$.
Dar la expresión de S en función de A y mostrar que S es un isomorfismo.
c) Mostrar que S es simétrica y ortogonal.
d) En el caso de que E_2 tiene dimensión 1, se considera una nube $\mathcal{M}$ de n puntos en E y V la matriz de covarianza asociada. Se supone que hay simetría con respecto a E_2 entre los puntos de $calM$ (si $\underline{x} \in \mathcal{M}$, entonces $S(\underline{x}) \in \mathcal{M}$).
Muestre que E_2 es un eje principal de la nube $\mathcal{M}$ en E.
5. Consideremos el espacio euclidiano $I\!R^p$ dotado de una métrica euclidiana M, y un conjunto de puntos $\mathcal{M} = \{\underline{x}_i : i = 1, 2, ..., n\}$ de $I\!R^p$.
Cada punto $\underline{x}_i$ está dotado de una masa $m_i > 0$, con $\sum m_i = 1$ y suponemos que el centro de gravedad de $\mathcal{M}$ está en el origen ($\underline{g} = \sum_i m_i \underline{x}_i = \underline{0}$) y se define $V = \sum_i m_i \underline{x}_i \underline{x}_i^t$. $I\!R^p$ está descompuesto en una suma directa de dos s.e.v. M-ortogonales: Δ_u, generado por un vector u de $I\!R^p$ pasando por el origen; y el hiperplano $H = \Delta_u^\perp$ M-ortogonal a Δ_u pasando por el origen: $I\!R^p = \Delta_u \bigoplus H$.
a) Exprese el momento de inercia I_0 del conjunto $\mathcal{M}$ con respecto al origen en función de M.
b) Deduzca que $I_0 = tr(VM)$.
c) Muestre que $I_H = \underline{u}^t MVM\underline{u}$ donde I_H es el momento de inercia de $\mathcal{M}$ con respecto a H.

6. EXAMEN DE PRIMAVERA 1994.

PARTE 1

Se considera 6 mediciones hechas sobre 23 peces. Se presenta los resultados de un análisis en componentes principales sobre estos datos.

a) Interprete los porcentajes de los valores propios (Tabla 8.11).

b) Interprete el gráfico 8.8: ?' Que tamaño y forma tienen los peces 1, 5, 8 y 11?

c) Gráfique el circulo de correlación a partir de la tabla 8.11 y comente.

d) Usando la tabla 8.11 dé las expresiones de las primeras componentes principales C_1 y C_2 en función de las 6 mediciones. Interprételas.

e) Usando la matriz de correlaciones (tabla 8.12), ubique las variables suplementarias PESO y RADIOACTIVIDAD en el círculo de de correlaciones.

f) Se quiere hacer el modelo lineal: $PESO = \beta_o + \beta_1 \underline{c}_1 + \beta_2 \underline{c}_2$, en donde $\underline{c}_1$ y $\underline{c}_2$ son las dos primeras componentes principales. Dé el coeficiente de correlación múltiple R^2.

PARTE 2

a) Se quiere hacer el modelo lineal: $RADIOACTIVIDAD\ Y = \beta_o + \beta_1 \underline{c}_1 + \beta_2 \underline{c}_2$. Sea X la matriz (23x3) asociado a este modelo lineal. Calcule la matriz $(X^t X)^{-1}$.

b) Calcule $X^t Y$, en donde Y es el vector a explicar del modelo lineal.

c) Dé los estimadores de mínimos cuadrados de β_o, β_1 y β_2.

d) Dé el coeficiente de correlación múltiple R^2. Deduzca el estimador insesgado de la varianza σ^2 de los errores y la estimación de la varianza de los estimadores.

e) Muestre que los estimadores de β_1 y β_2 son no correlacionados. Haciendo el supuesto de normalidad, encuentre intervalos de confianza de nivel 95% para β_1 y β_2.

f) Efectúe los tests de hipótesis $H_o : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$.

PARTE 3

a) Los 23 peces estan divididos en tres acuarios. Se busca si el acuario tiene un efecto sobre la RADIOACTIVIDAD, usando el modelo: $Y_i = \beta_o + \beta_j + \epsilon_i$ si el pez i esta en el ACUARIO j; el parámetro β_j mide el efecto del ACUARIO j (j=1,..,3) sobre la RADIOACTIVIDAD. Escribe el criterio de los mínimos cuadrados en tres sumas que dependen de los tres acuarios. Usando la tabla 8.13 y tomando la media muestral $\bar{y}$ como estimador de β_o, deduzca el estimador de los mínimos cuadrados de los tres parámetros restantes.

b) Efectúe el test $H_o : \beta_3 = \beta_2 =$. Precise los supuestos que tuvo que hacer.

c) Sea un nuevo pez que toma los valores: LARGO: 180, LARGO SIN CABEZA: 152, ANCHO CABEZA: 40, ANCHO: 38, ANCHO HOCICO: 15, DIAMETRO OJOS: 12. Calcule C_1 y C_2 para este pez. Prediga su RADIOACTIVIDAD y dé un intervalo de confianza.
d) Si se supone que la variable RADIOACTIVIDAD $Y \sim Exp(\mu)$ y una distribución a priori $Exp(\theta)$ para μ ($\pi(\mu) = \theta exp(-\theta\mu)$ para μ positivo), dé la distribución a posteriori de μ dada la muestra de los 23 peces.
e) Tomando la función de perdida cuadratica, dé el estimador de Bayes.

	MEDIA	DESV. TIPICA	FACTOR .	FACTOR 2	FACTOR 3	FACTOR 4
VALORES PROPIOS			4.885	'493	'388	'128
% ACUMULADOS DE LOS VALORES PROPIOS			81.62	89.63	96.10	98.23
LARGO	.90.17	.7.99	0.947	0.226	0.099	0.126
LARGO	.90.17	.7.99	0.947	0.226	0.099	0.126
LARGO cIN CABEZA	.70.70	.5.69	0.939	0.264	0.128	-0.005
ANCHO CABEZA	42.78	4.80	0.959	0.133	0.121	0.045
ANCHO	39.30	4.57	0.922	-0.215	0.144	-0.283
ANCHO HOCICO	.3.57	2.54	0.816	0.071	-0.570	-0.040
DIAMETRO OJOS	9.74	0.96	0.817	-0.550	0.001	0.166

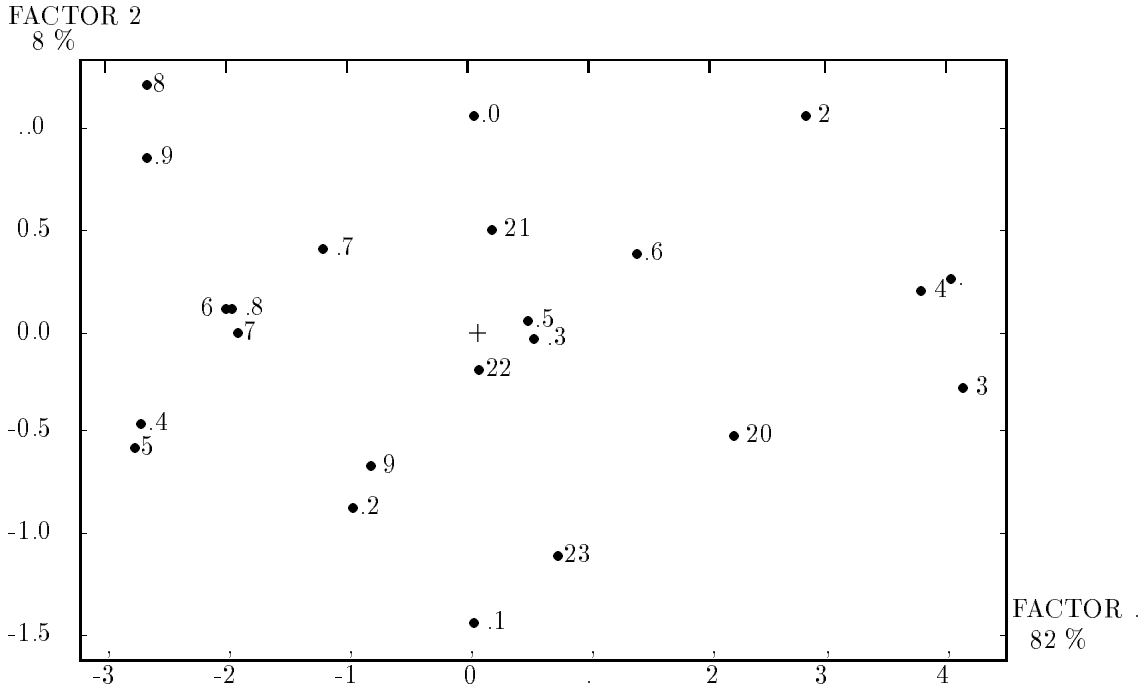
TABLA 8.11: Correlaciones de las variables sobre las 4 primeras componentes principales

VARIABLES	10	2	C	C_2
PESO	1.00	-.44	.98	.00
RADIOACTIVIDAD	-.44	1.00	-.41	.23
C	.98	-.41	1.00	.00
C_2	.00	.23	.00	1.00

TABLA 8.12: Matriz de correlaciones

ACUARIO	RADIOACTIVIDAD 1	2	3	TOTAL	PESO PESO
EFECTIVO	8	8	7	23	23
MEDIA	15.25	33.50	33.71	27.22	82.09
DESVIACION TIPICA	7.13	12.13	21.69	16.47	26.5

TABLA 8.13: Radiactividad


FACTOR 2
8 %
..0
0.5
0.0
-0.5
-1.0
-1.5
-3
-2
-1
0
2
3
4
FACTOR .
82 %
8
.0
2
.9
21
.7
.6
4
6
.8
7
.5
.3
22
3
.4
5
20
9
.2
23
.1
GRAFICO 8.8