Estadística

José Manuel Corcuera.

Índice

1 Estimación puntual **3**

1.1 Modelos estadísticos 3

1.2 Estadísticos y estimadores. 5

1.3 Sesgo y error cuadrático. 6

1.4 Métodos de estimación 8

1.4.1 Método de máxima verosimilitud 8

1.4.2 Método de los momentos 10

1.5 Modelos estadísticos regulares. Cota de Cramer-Rao. 15

1.6 Modelos exponenciales 20

1.7 Propiedades asintóticas de los métodos de estimación. 25

1.8 Muestras de una población normal 29

2 Intervalos de confianza **35**

2.1 Construcción de intervalos a partir de la verosimilitud 35

2.2 Construcción de intervalos a partir de funciones pivotantes. . . . 37

2.3 Problemas de dos muestras. 39

2.3.1 Muestras independientes 39

2.3.2 Muestras relacionadas 41

2.4 Algunos métodos para obtener pivotes 41

2.4.1 Un método bastante general 41

2.4.2 Familias de posición y escala 42

2.4.3 Métodos aproximados 43

2.4.4 Un método especial 45

2.4.5 Regiones de confianza, intervalos simultáneos 46

3 Test de hipótesis **49**

3.1 Test de hipótesis simples 50

3.2 Hipótesis compuestas 55

3.2.1 Construcción de test a partir de intervalos de confianza . 58

3.2.2 Test de la razón de verosimilitudes. 60

4 Test Ji-cuadrado **69**

4.1 El modelo multinomial 69

4.2 Test de ajuste en el modelo multinomial 70

4.2.1 Ajuste a una mutinomial concreta 70
4.2.2 Ajuste a una familia de mutinomiales 73
4.2.3 Test de independencia de dos multinomiales 75
4.3 Test Ji-cuadrado de ajuste, independencia y homogeneidad. . . . 79
4.4 Test de ajuste y homogeneidad no paramétricos. 82
4.4.1 La distribución empírica 82
4.4.2 Test de Kolmogorov-Smirnov 83

5 El modelo lineal 87
5.1 Estimación de los parámetros 88
5.1.1 Propiedades del estimador mínimo-cuadrático 89
5.1.2 Estimación mínimo cuadrática de σ^2. 90
5.2 El modelo lineal normal 92
5.2.1 Estimación máximo verosímil 93
5.2.2 Predicción 93
5.2.3 Contraste de hipótesis 94

Capítulo 1

Estimación puntual

1.1 Modelos estadísticos

En Estadística nos encontraremos con observaciones o datos de la forma: $x = (x_1, x_2, ..., x_n)$ que supondremos corresponden a un cierto objeto aleatorio $X = (X_1, X_2, ..., X_n)$. El propósito será conocer algún valor relacionado con X. Supondremos que X toma valores en un espacio medible $(\mathcal{X},\mathcal{F})$ y que de la ley de X sólo se sabe que es una posible dentro de un conjunto de leyes: $\{P_\theta, \theta \in \Theta\}$. Todo esto modelará un situación de incertidumbre.

Ejemplo 1.1.1 *Suponemos que queremos saber cuál es la probabilidad de que salga cara al lanzar una moneda. Podemos suponer que el fenómeno se puede describir con variables $X_i, i = 1, .., n$ Bernoullis independientes de parámetro p. Queremos saber p.*

Definición 1.1.1 *Un modelo estadístico es una terna $(\mathcal{X}, \mathcal{F}, \{P_\theta, \theta \in \Theta\})$.*

Observación 1.1.1 *$\mathcal{X}$ se denomina* espacio muestral. *n será el* tamaño muestral. *$x = (x_1, x_n, ..., x_n)$ se dirá que es una muestra concreta de tamaño n. $X = (X_1, X_2, ..., X_n)$ se dirá que es muestra genérica y a veces hablaremos de* modelo inducido por X. *Normalmente $\mathcal{X} = \mathbb{R}^n$. Si las X_i son independientes e idénticamente distribuidas (iid) diremos que la muestra es una muestra aleatoria simple. Θ se denomina* espacio de parámetros.

Observación 1.1.2 *Notemos que*

$$(\Omega, \mathcal{A}) \xrightarrow{X} (\mathcal{X}, \mathcal{F}).$$

La ley de X normalmente se denota P_X para distinguirla de la ley P en $(\Omega, \mathcal{A})$ pero salvo que dé lugar a confusión omitiremos este subíndice, en cambio como la ley de X depende de θ lo resaltaremos escribiendo P_θ.

Ejemplo 1.1.2 *(Modelo Bernoulli) Modelo inducido por n Bernoullis independientes de parámetro p:*

$$\begin{aligned}\mathcal{X} &= \{0,1\}^n, \\ \mathcal{F} &= \mathcal{P}(\{0,1\}^n) \\ \theta &= p \\ P_p(x_1, x_n, ..., x_n) &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ \Theta &= (0,1)\end{aligned}$$

Definición 1.1.2 *Diremos que un modelo estadístico es paramétrico si $\Theta \subseteq \mathbb{R}^d$ para algún d.*

Observación 1.1.3 *El modelo Bernoulli es paramétrico con $d = 1$.*

Ejemplo 1.1.3 *Modelo Poisson: modelo inducido por n Poissons independientes de parámetro λ.*

$$\begin{aligned}\mathcal{X} &= (\mathbb{Z}^+)^n, \\ \mathcal{F} &= \mathcal{P}((\mathbb{Z}^+)^n) \\ \theta &= \lambda \\ P_\lambda(x_1, x_n, ..., x_n) &= e^{-\lambda n}\frac{\lambda^{\sum_{i=1}^n x_i}}{\Pi_{i=1}^n x_i!} \\ \Theta &= (0,+\infty)\end{aligned}$$

Ejemplo 1.1.4 *Modelo normal con media y varianza desconocidas: modelo inducido por n normales independientes de parámetros (μ, σ^2).*

$$\begin{aligned}\mathcal{X} &= \mathbb{R}^n, \\ \mathcal{F} &= \mathcal{B}(\mathbb{R}^n) \\ \theta &= (\mu,\sigma) \\ f_{(\mu,\sigma)}(x_1, x_n, ..., x_n) &= \frac{1}{\left(\sqrt{2\pi}\right)^n \sigma^n}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\} \\ \Theta &= (-\infty,+\infty)\times(0,+\infty)\end{aligned}$$

Observación 1.1.4 *Los modelos anteriores son obviamente paramétricos. Nos centraremos en ellos. Notemos también que al ser las observaciones iid $(\mathcal{X}, \mathcal{F}, \{P_\theta, \theta \in \Theta\}) = (\mathcal{X}_1^n, \mathcal{F}_1^n, \{P_{1\theta}^n, \theta \in \Theta\})$*

Ejemplo 1.1.5 *(Modelo no paramétrico) Las X_i son observaciones iid con densidad. En tal caso $\Theta = \{f : f \geq 0, \int_{\mathbb{R}} f(x)dx = 1\}$.*

1.2 Estadísticos y estimadores.

Definición 1.2.1 *Un estadístico T es una aplicación de $(\mathcal{X},\mathcal{F})$ en $\left(\mathbb{R}^k,\mathcal{B}(\mathbb{R}^k)\right)$*

$$T:\mathcal{X}\rightarrow\mathbb{R}^k,$$

k es la dimensión *del estadístico.*

Ejemplo 1.2.1 *Media muestral:*

$$T(x_1,x_2,...,x_n)=\frac{\sum_{i=1}^n x_i}{n}:=\bar{x}$$

Ejemplo 1.2.2 *Varianza muestral:*

$$T(x_1,x_2,...,x_n)=\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n}:=s^2$$

Ejemplo 1.2.3 *Varianza muestral corregida:*

$$T(x_1,x_2,...,x_n)=\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n-1}:=\tilde{s}^2$$

Ejemplo 1.2.4 *Si ordenamos la muestra $(x_1,x_2,...,x_n)$ de menor a mayor podemos escribir la muestra ordenada $(x_{(1)},x_{(2)},...,x_{(n)})$, tenemos así el estadístico de orden i-ésimo*

$$T(x_1,x_2,...,x_n)=x_{(i)}$$

Ejemplo 1.2.5 *Rango*

$$T(x_1,x_2,...,x_n)=x_{(n)}-x_{(1)}$$

Observación 1.2.1 *En todos estos ejemplos el estadístico es de dimensión 1, que será lo habitual.*

Observación 1.2.2 *Fijada la ley de probabilidad de X, T es un variable o vector aleatorio que por construcción tiene la misma ley que $T(X)$:*

$$\begin{aligned}P_T(B) &= P_X(T^{-1}(B))=P(X^{-1}(T^{-1}(B))) \\ &= P((T(X))^{-1}(B))=P_{T(X)}(B).\end{aligned}$$

Por ejemplo si las X_i son iid Bernoullis de parámetro p, y T es la suma de las observaciones $T(X)=\sum_{i=1}^n X_i$, de manera que T tendrá una ley Binomial de parámetros n y p.

El propósito de la estimación puntual es dar un valor razonable del *verdadero* valor de θ o de una función del parámetro $g(\theta)$, dada la muestra x.

Definición 1.2.2 *Un estimador de $g(\theta)$ es un estadístico que se usa para estimar (dar valores razonables) $g(\theta)$.*

Ejemplo 1.2.6 *En modelo Bernoulli de parámetro p, podemos utilizar la media muestral para estimar el parámetro p,*

$$T(x) = \frac{\sum_{i=1}^{n} x_i}{n} := \hat{p}(x) \approx p$$

sabemos que la aproximación es buena si n es grande.

1.3 Sesgo y error cuadrático.

Definición 1.3.1 *Diremos que un estimador T (integrable) de $g(\theta) \in \mathbb{R}$, es un estimador* insesgado *si*

$$E_\theta(T) = g(\theta), \forall \theta \in \Theta$$

Definición 1.3.2 *El* sesgo *de un estimador de $g(\theta) \in \mathbb{R}$ se define como la funcion de θ*

$$Sesgo_\theta(T) = E_\theta(T) - g(\theta), \theta \in \Theta$$

Ejemplo 1.3.1 *Consideremos el modelo Bernoulli, $\hat{p}$ es un estimador insesgado de p:*

$$\begin{aligned} E_p(\hat{p}) &= E_p(\hat{p}(X)) = E_p\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E_p(X_i) = \frac{np}{n} = p \end{aligned}$$

Definición 1.3.3 *Denominaremos* error cuadrático medio *del estimador T (de cuadrado integrable) de $g(\theta) \in \mathbb{R}$, a la función de θ.*

$$ECM_\theta(T) = E_\theta(T - g(\theta))^2, \theta \in \Theta$$

Ejemplo 1.3.2

$$\begin{aligned} ECM_p(\hat{p}) &= E_p(\hat{p} - p)^2 = Var_p(\hat{p}) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var_p(X_i) = \frac{np(1-p)}{n^2} \\ &= \frac{p(1-p)}{n} \end{aligned}$$

Proposición 1.3.1 *Sea T un estimador (con momento de segundo orden)*

$$ECM_\theta(T) = Var_\theta(T) + Sesgo_\theta^2(T), \theta \in \Theta$$

Definición 1.3.4 *Diremos que un estimador T (de cuadrado integrable), de $g(\theta) \in \mathbb{R}$, es uniformemente de mínima varianza (UMV) si minimiza el error cuadrático medio en la clase de estimadores insesgados.*

Proposición 1.3.2 *El estimador UMV es único casi seguramente para todo $\theta \in \Theta$. Esto es, si T_1, T_2 son estimadores UMV de $g(\theta) \in \mathbb{R}$ entonces*

$$P_\theta\{T_1 \neq T_2\} = 0, \forall \theta \in \Theta.$$

Demostración. Bastará ver que $E_\theta((T_1 - T_2)^2) = 0, \forall \theta \in \Theta$. En efecto, por Chebichef

$$P_\theta\{|T_1 - T_2| > \varepsilon\} \leq \frac{E_\theta((T_1 - T_2)^2)}{\varepsilon^2} = 0$$

y

$$P_\theta\left\{|T_1 - T_2| \neq 0\right\} = P_\theta(\cup_{n=1}^{\infty}\{|T_1 - T_2| > \frac{1}{n}\} \leq \sum_{n=1}^{\infty} P_\theta(\{|T_1 - T_2| > \frac{1}{n}\} = 0.$$

Veamos entonces que $E_\theta((T_1 - T_2)^2) = 0$. Primero notemos que $\frac{T_1+T_2}{2}$ será un estimador insesgado de $g(\theta)$:

$$E_\theta\left(\frac{T_1 + T_2}{2}\right) = \frac{E_\theta(T_1) + E_\theta(T_2)}{2} = \frac{g(\theta) + g(\theta)}{2}.$$

Por tanto

$$\begin{aligned}
Var_\theta(T_1) &\leq Var_\theta(\frac{T_1 + T_2}{2}) \\
&= \frac{1}{4}(Var_\theta(T_1) + Var_\theta(T_2) + 2Cov(T_1, T_2)) \\
&= \frac{1}{4}(2Var_\theta(T_1) + 2Cov_\theta(T_1, T_2)).
\end{aligned}$$

Por la desigualdad de Cauchy-Schwarz

$$|Cov_\theta(T_1, T_2)| \leq \sqrt{Var_\theta(T_1)}\sqrt{Var_\theta(T_2)} = Var_\theta(T_1),$$

por tanto

$$Var_\theta(T_1) \leq Var_\theta(\frac{T_1 + T_2}{2}) \leq \frac{1}{4}(2Var_\theta(T_1) + 2Var_\theta(T_1)) = Var_\theta(T_1).$$

De manera que $Var_\theta(\frac{T_1+T_2}{2}) = Var_\theta(T_1)$ y por tanto $Cov_\theta(T_1, T_2) = Var_\theta(T_1)$. Finalmente

$$\begin{aligned}
E_\theta((T_1 - T_2)^2) &= Var_\theta(T_1 - T_2) \\
&= Var_\theta(T_1) + Var_\theta(T_2) - 2Cov_\theta(T_1, T_2) \\
&= 2Var_\theta(T_1) - 2Var_\theta(T_1) = 0
\end{aligned}$$

■

Desigualdad de Cauchy-Schwarz

Sean X, Y dos variables aleatorias, no constantes, de cuadrado integrable, entonces $\forall\lambda \in \mathbb{R}$

$$0 \leq \text{var}(\lambda Y + X) = \lambda^2\text{var}(Y) + 2\lambda\text{cov}(X, Y) + \text{var}(X)$$

implica que el discriminante de la ecuación de segundo grado en λ

$$\lambda^2\text{var}(Y) + 2\lambda\text{cov}(X, Y) + \text{var}(X) = 0$$

es menor o igual que cero, esto es

$$4\text{cov}(X, Y)^2 - 4\text{var}(Y)\text{var}(X) \leq 0,$$

con lo que

$$|\text{cov}(X, Y)| \leq \sqrt{\text{var}(X)}\sqrt{\text{var}(Y).}$$

1.4 Métodos de estimación

Supongamos un modelo estadístico paramétrico $(\mathcal{X}, \mathcal{F}, \{P_\theta, \theta \in \Theta\})$. El propósito será encontrar métodos para estimar $g(\theta)$.

1.4.1 Método de máxima verosimilitud

Supongamos que el modelo tiene densidad y sea $f_\theta(x)$ la densidad correspondiente.

Definición 1.4.1 *Dada una muestra $x \in \mathcal{X}$, se denomina función de verosimilitud a la aplicación*

$$\begin{array}{rcl} L(x;) : \Theta & \rightarrow & \mathbb{R} \\ \theta & \mapsto & L(x; \theta) = f_\theta(x) \end{array}$$

Ejemplo 1.4.1 *Supongamos que tenemos un modelo de n observaciones iid Exponenciales de parametro $\lambda > 0$.Dada $x = (x_1, x_2, ...x_n) \in \mathbb{R}^n_+$ la densidad será*

$$f_\lambda(x) = \Pi_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i},$$

de manera que

$$L(x; \lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}, x \in \mathbb{R}^n_+$$

Ejemplo 1.4.2 *Supongamos que tenemos un modelo de n observaciones iid Uniformes en $(0, \theta)$. Dada $x = (x_1, x_2, ...x_n) \in \mathbb{R}^n_+$ la densidad será*

$$f_\theta(x) = \Pi_{i=1}^n \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\mathbf{1}_{[0,\theta]}(x_{(n)})$$

de manera que

$$\begin{aligned} L(x;\theta) &= \frac{1}{\theta^n}\mathbf{1}_{[0,\theta]}(x_{(n)}) \\ &= \frac{1}{\theta^n}\mathbf{1}_{[x_{(n)},+\infty)}(\theta) \end{aligned}$$

Definición 1.4.2 *Se denomina estimador máximo verosímil (EMV) de θ a $\hat{\theta}$: $\mathcal{X} \to \times \subseteq \mathbb{R}^d$ tal que*

$$L(x;\hat{\theta}(x)) = \sup_{\theta\in\Theta} L(x;\theta), \forall x \in \mathcal{X}\backslash\mathcal{N}$$

donde $P_\theta(\mathcal{N}) = 0, \forall\theta \in \Theta$.

Ejemplo 1.4.3 *Supongamos que tenemos un modelo de n observaciones iid Exponenciales de parametro $\lambda > 0$. Para hallar $\hat{\lambda}$ hay que hallar el máximo de la función de verosimilitud*

$$L(x;\lambda) = \lambda^n e^{-\lambda\sum_{i=1}^n x_i}$$

o lo que es lo mismo el máximo de la log*-verosimilitud*

$$l(x;\lambda) := \log L(x;\lambda) = n\log\lambda - \lambda\sum_{i=1}^n x_i.$$

Como

$$\frac{\partial l(x;\lambda)}{\partial\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0$$

el extremo es

$$\hat{\lambda}(x) = \frac{n}{\sum_{i=1}^n x_i}.$$

Es un máximo local:

$$\frac{\partial^2 l(x;\lambda)}{\partial\lambda^2} = -\frac{n}{\lambda^2} < 0.$$

Como la función es derivable y el máximo local único tenemos un máximo global.

Ejemplo 1.4.4 *Supongamos que tenemos un modelo de n observaciones iid Uniformes en $(0,\theta)$. La verosimilitud es*

$$L(x;\theta) = \frac{1}{\theta^n}\mathbf{1}_{[x_{(n)},+\infty)}(\theta).$$

Es inmediato que

$$\hat{\theta}(x) = x_{(n)}.$$

Proposición 1.4.1 *Sea*

$$\begin{aligned} g:\Theta &\to \Lambda \\ \theta &\to \lambda = g(\theta) \end{aligned}$$

una biyección (medible). Entonces $\hat{\lambda} = g(\hat{\theta})$.

Demostración. Tenemos que

$$L(x;\theta) = L(x; g^{-1}(\lambda)) := \tilde{L}(x;\lambda).$$

Por definición

$$\tilde{L}(x;\hat{\lambda}) = \sup_{\lambda\in\Lambda} \tilde{L}(x;\lambda).$$

Ahora bien

$$\begin{aligned} \tilde{L}(x; g(\hat{\theta})) &= L(x; g^{-1}(g(\hat{\theta}))) = L(x;\hat{\theta}) \\ &= \sup_{\theta\in\Theta} L(x;\theta) = \sup_{\lambda\in\Lambda} L(x; g^{-1}(\lambda)) \\ &= \sup_{\lambda\in\Lambda} \tilde{L}(x;\lambda). \end{aligned}$$

■

Ejemplo 1.4.5 *Supongamos que tenemos un modelo de n observaciones iid Exponenciales de parámetro $\lambda > 0$. Sea $\mu = \frac{1}{\lambda}$, entonces*

$$\hat{\mu} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}.$$

1.4.2 Método de los momentos

Definición 1.4.3 *Supongamos que nuestro modelo es el inducido por n variables iid: $(X_1, X_2, ..., X_n)$. Sean $\mu_i(\theta) = E_\theta(X_1^i)$, esto es, los momentos de orden i-ésimo de las variables. Supongamos que quiero estimar $g(\theta) \in \mathbb{R}$ y que puedo escribir*

$$g(\theta) = h(\mu_1(\theta), \mu_2(\theta), ..., \mu_k(\theta)),$$

entonces el estimador por el método de los momentos de $g(\theta)$, que escribiremos T, viene dado por

$$T(x) = h\left(\frac{1}{n}\sum_{i=1}^n x_i, \frac{1}{n}\sum_{i=1}^n x_i^2, ..., \frac{1}{n}\sum_{i=1}^n x_i^k\right).$$

Ejemplo 1.4.6 *Sea un modelo de n observaciones iid Uniformes en $(0,\theta)$.*

$$\mu_1(\theta) = E_\theta(X_1) = \frac{\theta}{2}.$$

Por tanto

$$\theta = 2\mu_1(\theta),$$

y el estimador de θ por el método de los momentos será

$$T(x) = 2\frac{1}{n}\sum_{i=1}^n x_i = 2\bar{x}.$$

Recordemos que el EMV era $\hat{\theta}(x) = x_{(n)}$. *En primer lugar*

$$f_{X_{(n)}}(x) = n(\frac{x}{\theta})^{n-1}\frac{1}{\theta}\mathbf{1}_{[0,\theta](x)},$$

entonces

$$E_\theta(\hat{\theta}) = E_\theta(X_{(n)}) = \int_0^\theta xn(\frac{x}{\theta})^{n-1}\frac{1}{\theta}dx = \frac{n}{n+1}\theta,$$

de manera que el estimador tiene sesgo, pero podemos corregirlo si consideramos el estimador

$$\theta^* = \frac{n+1}{n}\hat{\theta}.$$

La varianza de θ^* *será*

$$\begin{aligned} var(\theta^*) &= E(\theta^{*2}) - \theta^2 \\ &= \left(\frac{n+1}{n}\right)^2 E(\hat{\theta}^2) - \theta^2 \\ &= \left(\frac{n+1}{n}\right)^2 \frac{n}{n+2}\theta^2 - \theta^2 \\ &= \frac{\theta^2}{n(n+2)}, \end{aligned}$$

que es más pequeña, para todo valor de θ, *que la varianza de* T,

$$var(T) = \frac{4\,Var(X_1)}{n} = \frac{\theta^2}{3n}$$

Observación 1.4.1 *Es inmediato que el estimador por el método de los momentos de* $\mu_k(\theta)$ *es* $\frac{1}{n}\sum_{i=1}^n x_i^k$ *su llamado* estimador empírico. *En general (en el contexto iid) el estimador empírico de* $E_\theta(f(X_1))$ *se define como* $\frac{1}{n}\sum_{i=1}^n f(x_i)$.

Ejemplo 1.4.7 *Consideremos un modelo de observaciones iid* $N(\mu, \sigma^2)$ *con* (μ, σ) *desconocidos.*

$$L(x;\mu,\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\},$$

de manera que la log-*verosimilitud viene dada por*

$$l(x;\mu,\sigma) = \log L(x;\mu,\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 + Cte.$$

Vamos a buscar el máximo de manera iterativa. Fijado σ, *buscamos el máximo en* μ:

$$\partial_\mu l(x;\mu,\sigma) = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu) = 0$$

implica que

$$\hat{\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Es claro que es un máximo local

$$\partial_{\mu}^{2} l(x;\mu,\sigma) = -\frac{n}{\sigma^2} < 0$$

y por su unicidad y la derivabilidad de $l(x,\mu,\sigma)$ *resulta un máximo absoluto, de manera que para todo* σ

$$l(x;\mu,\sigma) \leq l(x;\bar{x},\sigma).$$

Buscamos ahora el máximo de

$$l(x;\bar{x},\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + Cte.$$

$$\partial_{\sigma} l(x;\mu,\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0,$$

de manera que

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Además

$$\begin{aligned}\partial_{\sigma}^{2} l(x;\mu,\sigma)|_{\sigma=\hat{\sigma}} &= \frac{n}{\hat{\sigma}^2} - \frac{3}{\hat{\sigma}^4}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= -\frac{2n}{\hat{\sigma}^2} < 0\end{aligned}$$

con lo que se trata de un máximo local y absoluto. Así tenemos

$$l(x;\mu,\sigma) \leq l(x;\bar{x},\sigma) \leq l(x;\bar{x},\hat{\sigma})$$

con lo que $T(x) = (\bar{x},\hat{\sigma})$ *es el estimador máximo verosímil de* (μ,σ). *Notemos que* $\hat{\sigma}^2$, *estimador máximo verosímil de* σ^2, *tiene sesgo. Primero notemos que :*

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2 \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x}) \\ &\quad +n(\bar{x}-\mu)^2 \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2,\end{aligned}$$

ya que $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. *De manera que*

$$\begin{aligned}E_{(\mu,\sigma)}(\hat{\sigma}^2) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu)^2 \\ &\quad -E(\bar{X} - \mu)^2 \\ &= \frac{n\sigma^2}{n} - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.\end{aligned}$$

Podemos corregir el sesgo si tomamos como estimador de σ^2.

$$\tilde{S}^2(x) = \frac{n-1}{n}\hat{\sigma}^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Notemos también que este resultado es general, para cualquier modelo (de observaciones iid) con varianza desconocida, el estimador $\hat{\sigma}^2$ *(varianza muestral), tiene sesgo y* $\tilde{S}^2$ *(varianza muestral corregida).*

Método de mínimos cuadrados

Definición 1.4.4 *Supongamos que las observaciones (genéricas) en nuestro modelo son de la forma*

$$X_i = g_i(\theta) + \varepsilon_i$$

donde las ε_i son variables aleatorias. Entonces el estimador por mínimos cuadrados de θ se define como $\tilde{\theta}(x)$ tal que

$$\sum (x_i - g_i(\tilde{\theta}))^2 = \min_{\theta \in \Theta} \sum (x_i - g_i(\theta))^2.$$

Ejemplo 1.4.8 *Supongamos que las observaciones son normales de media $g_i(\alpha, \beta) = \alpha + \beta r_i$ (r_i números conocidos y no iguales) y varianza σ^2. Podemos escribir*

$$X_i = \alpha + \beta r_i + \varepsilon_i, \tilde{\beta}$$

con ε_i normales de media cero y varianza σ^2. Entonces

$$\sum_{i=1}^{n} (x_i - (\tilde{\alpha} + \tilde{\beta} r_i))^2 = \min_{\alpha, \beta} \sum_{i=1}^{n} (x_i - (\alpha + \beta r_i))^2 = 0.$$

$$\begin{aligned}
\partial_\alpha \sum_{i=1}^{n} (x_i - (\alpha + \beta r_i))^2 &= -2 \sum_{i=1}^{n} (x_i - (\alpha + \beta r_i)) \\
\partial_\alpha^2 \sum_{i=1}^{n} (x_i - (\alpha + \beta r_i))^2 &= 2n > 0, \bar{r}
\end{aligned}$$

Por tanto el mínimo, fijado β, se alcanza para

$$\tilde{\alpha} = \bar{x} - \beta \bar{r}.$$

Ahora podemos tratar de minimizar

$$\sum_{i=1}^{n} (x_i - (\tilde{\alpha} + \beta r_i))^2 = \sum_{i=1}^{n} (x_i - \bar{x} - \beta(r_i - \bar{r}))^2.$$

$$\partial_\beta \sum_{i=1}^{n} (x_i - (\tilde{\alpha} + \beta r_i))^2 = -2 \sum_{i=1}^{n} (x_i - \bar{x} - \beta(r_i - \bar{r}))(r_i - \bar{r}) = 0,$$

de manera que

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(r_i - \bar{r})}{\sum_{i=1}^{n} ((r_i - \bar{r})^2},$$

ya que

$$\partial_\beta^2 \sum_{i=1}^{n} (x_i - (\tilde{\alpha} + \beta r_i))^2 = 2 \sum_{i=1}^{n} ((r_i - \bar{r})^2 > 0.$$

1.5 Modelos estadísticos regulares. Cota de Cramer-Rao.

Consideraremos que Θ es un intervalo abierto de $\mathbb{R}$.

Definición 1.5.1 *Diremos que un modelo estadístico es regular si se verifican las tres condiciones siguientes:*

i) $L(x;\theta) > 0, \forall x \in \mathcal{X} \backslash \mathcal{N}$ donde $P_\theta(\mathcal{N}) = 0, \forall \theta \in \Theta$.

ii) $\partial_\theta \int_{\mathcal{X}} L(x;\theta)dx = \int_{\mathcal{X}} \partial_\theta L(x;\theta)dx$ (en el caso absolutamente contínuo) $\partial_\theta \sum_{\mathcal{X}} L(x;\theta)dx = \sum_{\mathcal{X}} \partial_\theta L(x;\theta)dx$ (en el caso discreto), $\forall \theta \in \Theta$.

iii) (Información de Fisher) $0 < I_n(\theta) := E((\partial_\theta \log L)^2) < \infty$, $\forall \theta \in \Theta$ (n indica el tamaño muestral).

Proposición 1.5.1 *La condición ii) de regularidad es equivalente a*

$$E_\theta(\partial_\theta \log L) = 0, \forall \theta \in \Theta$$

Demostración. Puesto que $\int_{\mathcal{X}} L(x;\theta)dx = 1$, $\partial_\theta \int_{\mathcal{X}} L(x;\theta)dx = 0$ y la condición ii) equivale a

$$\begin{aligned} 0 &= \int_{\mathcal{X}} \partial_\theta L(x;\theta)dx = \int_{\mathcal{X}} \frac{\partial_\theta L(x;\theta)}{L(x;\theta)} L(x;\theta)dx \\ &= \int_{\mathcal{X}} \partial_\theta \log L(x;\theta) L(x;\theta)dx = E_\theta(\partial_\theta \log L). \end{aligned}$$

■

Corolario 1.5.1

$$I_n(\theta) = var_\theta(\partial_\theta \log L), \forall \theta \in \Theta.$$

Observación 1.5.1 *A la función del parámetro y la muestra $\partial_\theta \log L(x;\theta)$ se le suele denominar función "score". Fijado θ, $\partial_\theta \log L$ es una variable aleatoria de esperanza cero y varianza la información de Fisher.*

Supongamos la condición de regularidad adicional

iv) $\partial_\theta^2 \int_{\mathcal{X}} L(x;\theta)dx = \int_{\mathcal{X}} \partial_\theta^2 L(x;\theta)dx$ (en caso absolutamente continuo) $\partial_\theta^2 \sum_{\mathcal{X}} L(x;\theta)dx = \sum_{\mathcal{X}} \partial_\theta^2 L(x;\theta)dx$ (en el caso discreto), $\forall \theta \in \Theta$.

Proposición 1.5.2 *Si el modelo cumple i) ii) y iii), iv) es equivalente a*

$$I_n(\theta) = E_\theta(-\partial_\theta^2 \log L)$$

Demostración. La condición iv) es equivalente a

$$0 = \int_{\mathcal{X}} \partial_\theta^2 L(x;\theta)dx$$

y como

$$\partial_\theta^2 L(x;\theta) = \partial_\theta^2 \log L(x;\theta) L(x;\theta) + \left(\partial_\theta \log L(x;\theta)\right)^2 L(x;\theta)$$

el resultado se sigue integrando ambos miembros y de que

$$I_n(\theta) = \int_{\mathcal{X}} \left(\partial_\theta \log L(x;\theta)\right)^2 L(x;\theta)dx < \infty$$

. ■

Proposición 1.5.3 *Si el modelo corresponde a n observaciones iid y el modelo correspondiente a una observación es regular, el modelo correspondiente a n observaciones es regular y*

$$I_n(\theta) = nI_1(\theta)$$

donde $I_1(\theta)$ es la información del modelo con una observación.

Demostración. $L_n(x_1, ..., x_n;\theta) = \Pi_{i=1}^n L_1(x_i;\theta)$. Con lo que la condición i) de regularidad es evidente. La condición ii) se deduce de que

$$\log L_n(x_1, ..., x_n;\theta) = \sum_{i=1}^n \log L_1(x_i;\theta),$$

con lo que

$$\partial_\theta \log L_n(x_1, ..., x_n;\theta) = \sum_{i=1}^n \partial_\theta \log L_1(x_i;\theta),$$

$$E_\theta(\partial_\theta \log L_n(X)) = \sum_{i=1}^n E_\theta(\partial_\theta \log L_1(X_i)) = 0.$$

Finalmente

$$\begin{aligned} I_n(\theta) &= \mathrm{var}_\theta(\partial_\theta \log L_n) = \mathrm{var}_\theta(\partial_\theta \log L_n(X_1, ..., X_n;\theta)) \\ &= \mathrm{var}_\theta(\sum_{i=1}^n \partial_\theta \log L_1(X_i;\theta)) = \sum_{i=1}^n \mathrm{var}_\theta(\partial_\theta \log L_1(X_i;\theta)) \\ &= nI_1(\theta). \end{aligned}$$

■

Ejemplo 1.5.1 *Sea el modelo de n X_i iid con ley $Exp(\lambda), \lambda > 0$.*

$$L_1(x;\lambda) = \lambda \exp\{-\lambda x\}$$

$$\log L_1(x;\lambda) = \log \lambda - \lambda x,$$

$$\partial_\lambda \log L_1(x;\lambda) = \frac{1}{\lambda} - x.$$

$$E_\lambda(\partial_\lambda \log L_1) = \frac{1}{\lambda} - E_\lambda(X).$$

Pero

$$\begin{aligned} E_\lambda(X) &= \int_0^\infty x\lambda \exp\{-\lambda x\}dx = [-x\exp\{-\lambda x\}]_0^\infty + \int_0^\infty \exp\{-\lambda x\}dx \\ &= \int_0^\infty \exp\{-\lambda x\}dx = [-\frac{1}{\lambda}\exp\{-\lambda x\}]_0^\infty = \frac{1}{\lambda} \end{aligned}$$

con lo que

$$E_\lambda(\partial_\lambda \log L_1) = 0.$$

Además

$$\begin{aligned} I_1(\lambda) &= var_\lambda(\partial_\lambda \log L_1) = var_\lambda(X) \\ &= E_\lambda(X^2) - E_\lambda(X)^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \end{aligned}$$

Con lo que el modelo es regular y

$$I_n(\lambda) = \frac{n}{\lambda^2}.$$

Además

$$\partial_\lambda^2 \log L_1(x;\lambda) = -\frac{1}{\lambda^2}.$$

y el modelo también cumple iv).

Definición 1.5.2 *Diremos que T es un estadístico regular si*

$$\partial_\theta \int_{\mathcal{X}} T(x)L(x;\theta)dx = \int_{\mathcal{X}} T(x)\partial_\theta L(x;\theta)dx, \forall\theta \in \Theta.$$

Teorema 1.5.1 *(Cota de Crámer-Rao). Consideremos un modelo regular y sea T un estimador insesgado de $g(\theta)$ y regular, esto es entonces,*

$$var_\theta(T) \geq \frac{g'(\theta)^2}{I_n(\theta)}, \forall\theta \in \Theta.$$

Demostración.

$$\partial_\theta \int_{\mathcal{X}} T(x)L(x;\theta)dx = \partial_\theta E_\theta(T) = g'(\theta)$$

$$\begin{aligned}\int_{\mathcal{X}} T(x)\partial_\theta L(x;\theta)dx &= \int_{\mathcal{X}} T(x)\frac{\partial_\theta L(x;\theta)}{L(x;\theta)}L(x;\theta)dx \\ &= \int_{\mathcal{X}} T(x)\partial_\theta \log L(x;\theta)L(x;\theta)dx \\ &= cov_\theta(T, \partial_\theta \log L).\end{aligned}$$

Ya que $E_\theta(\partial_\theta \log L) = 0$. *Por tanto*

$$\partial_\theta \int_{\mathcal{X}} T(x)L(x;\theta)dx = \int_{\mathcal{X}} T(x)\partial_\theta L(x;\theta)dx$$

es equivalente a escribir

$$g'(\theta) = cov_\theta(T, \partial_\theta \log L).$$

Si ahora aplicamos la desigualdad de Cauchy-Schwarz

$$g'(\theta)^2 \leq var_\theta(T) var_\theta(\partial_\theta \log L)$$

■

Definición 1.5.3 *En el contexto anterior si* T *es un estimador insesgado de* $g(\theta)$ *tal que*

$$var_\theta(T) = \frac{g'(\theta)^2}{I_n(\theta)},$$

se dice que es un estimador eficiente de $g(\theta)$.

Proposición 1.5.4 *Si* T *es un estimador eficiente entonces es UMV (en la clase de estimadores regulares).*

Observación 1.5.2 *La condición de eficiencia es equivalente a la condición de igualdad en la desigualdad de Cauchy-Schwarz y por tanto a la "co-linealidad" entre* $T - E_\theta(T)$ *y* $\partial_\theta \log L$, P_θ*- casi seguramente. Esto es, tendremos eficiencia si y sólo si,* $\forall\theta \in \Theta$, *existe* $\lambda(\theta)$ *tal que*

$$T - E_\theta(T) = \lambda(\theta)\partial_\theta \log L, \forall x \in \mathcal{X}\backslash\mathcal{N}_\theta, P_\theta(\mathcal{N}_\theta) = 0.$$

Notemos que

$$(T - E_\theta(T))\partial_\theta \log L = \lambda(\theta)(\partial_\theta \log L)^2,$$

y por tanto

$$\begin{aligned} g'(\theta) &= E_\theta((T - E_\theta(T))\partial_\theta \log L) = \\ &= \lambda(\theta)E_\theta((\partial_\theta \log L)^2) = \lambda(\theta)I_n(\theta).\end{aligned}$$

En definitiva

$$\lambda(\theta) = \frac{g'(\theta)}{I_n(\theta)}.$$

Ejemplo 1.5.2 *Sean n X_i iid Bernoulli(p). Entonces*

$$L(x_1, x_2, ..., x_n; p) = p^{\sum x_i}(1-p)^{n-\sum x_i}, p \in (0,1).$$

El modelo es regular y

$$\begin{aligned} \partial_p \log L &= \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} \\ &= \frac{\sum x_i}{p(1-p)} - \frac{n}{(1-p)}, \end{aligned}$$

con lo que

$$\frac{\sum x_i}{n} - p = \frac{p(1-p)}{n} \partial_p \log L.$$

Resulta así que

$$T(x) = \frac{\sum x_i}{n}$$

es un estimador eficiente de p y

$$\lambda(p) = \frac{p(1-p)}{n}$$

con lo que

$$I_n(p) = \frac{n}{p(1-p)}.$$

Ejemplo 1.5.3 *Sean n X_i iid Exp(λ), ya vimos que el modelo era regular,*

$$L(x_1, x_2, ..., x_n; \lambda) = \lambda^n \exp\{-\lambda \sum x_i\}$$

de manera que

$$\partial_\lambda \log L = \frac{n}{\lambda} - \sum x_i,$$

y

$$\frac{\sum x_i}{n} - \frac{1}{\lambda} = -\frac{1}{n} \partial_\lambda \log L.$$

Por tanto $\frac{\sum x_i}{n}$ es un estimador eficiente de $\frac{1}{\lambda}$ y

$$-\frac{1}{n} = \frac{g'(\lambda)}{I_n(\lambda)} = \frac{-\frac{1}{\lambda^2}}{I_n(\lambda)}$$

de manera que

$$I_n(\lambda) = \frac{n}{\lambda^2}.$$

1.6 Modelos exponenciales

Supongamos que para todo $x \in \mathcal{X}$ se cumple la condición de eficiencia

$$T(x) - g(\theta) = \lambda(\theta)\partial_\theta \log L(x, \theta), \forall\theta \in \Theta$$

integrando la ecuación anterior obtenemos que

$$L(x, \theta) = \exp\{a(\theta)T(x) + b(\theta) + s(x)\},$$

con

$$a(\theta) = \int \frac{1}{\lambda(\theta)} d\theta, b(\theta) = -\int \frac{g(\theta)}{\lambda(\theta)} d\theta$$

y $s(x)$ una constante de integración que puede depender de x.

Definición 1.6.1 *Un modelo con verosimilitud de la forma*

$$L(x, \theta) = \exp\{a(\theta)T(x) + b(\theta) + s(x)\},$$

se dice que es un modelo exponencial. $T(x)$ se denomina estadístico privilegiado.

Observación 1.6.1 *Es inmediato ver que si el modelo correspondiente a una observación es exponencial el correspondiente a n observaciones iid también lo es. En efecto*

$$\begin{aligned} L_n(x_1, x_2, ...x_n; \theta) &= \Pi_{i=1}^n L_1(x_i; \theta) \\ &= \Pi_{i=1}^n \exp\{a_1(\theta)T_1(x_i) + b_1(\theta) + s_1(x_i)\} \\ &= \exp\{a_1(\theta)\sum_{i=1}^n T_1(x_i) + nb_1(\theta) + \sum_{i=1}^n s_1(x_i)\} \\ &= \exp\{a(\theta)T(x) + b(\theta) + s(x)\} \end{aligned}$$

con

$$a(\theta) = a_1(\theta), T(x) = \sum_{i=1}^n T_1(x_i), b(\theta) = nb_1(\theta), s(x) = \sum_{i=1}^n s_1(x_i).$$

Ejemplo 1.6.1 *Supongamos un modelo de n, observaciones iid Bernoulli(p). Entonces*

$$\begin{aligned} L(x_1, x_2, ...x_n; p) &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ &= \exp\{\log\frac{p}{1-p}\sum_{i=1}^n x_i + n\log(1-p)\}, \end{aligned}$$

con lo que podemos escribir

$$a(p) = \log\frac{p}{1-p}, T(x) = \sum_{i=1}^n x_i, b(p) = n\log(1-p), s(x) = 0,$$

Ejemplo 1.6.2 *Supongamos un modelo de n, observaciones iid Exp(λ). Entonces*

$$\begin{aligned} L(x_1, x_2, ...x_n; \lambda) &= \lambda^n \exp\{-\lambda \sum_{i=1}^n x_i\} \\ &= \exp\{-\lambda \sum_{i=1}^n x_i + n \log \lambda\} \end{aligned}$$

con lo que podemos tomar

$$a(\lambda) = -\lambda, T(x) = \sum_{i=1}^n x_i, b(\lambda) = n \log \lambda, s(x) = 0,$$

Ejemplo 1.6.3 *Supongamos un modelo de n, observaciones iid $N(\mu, \sigma^2), \sigma$ conocida. Entonces*

$$\begin{aligned} L(x_1, x_2, ...x_n; \mu) &= \frac{1}{\left(\sqrt{2\pi}\right)^n \sigma^n} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\} \\ &= \exp\{\frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \\ &\quad -n \log \sigma - \frac{n}{2} \log(2\pi)\} \end{aligned}$$

con lo que podemos tomar

$$\begin{aligned} a(\mu) &= \frac{\mu}{\sigma^2}, T(x) = \sum_{i=1}^n x_i, b(\mu) = -\frac{n\mu^2}{2\sigma^2}, \\ s(x) &= -\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 - n \log \sigma - \frac{n}{2} \log(2\pi). \end{aligned}$$

Si el modelo es exponencial obviamente, se cumple la condición i) de regularidad. Para que se cumplan las condiciones ii) y iii) vamos a exigir que $a(\theta)$ sea derivable con $a'(\theta) \neq 0$. Necesitamos un resultado que nos permita intercambiar límites e integrales (o sumas):

Teorema 1.6.1 *(convergencia dominada) Sea $\{f_n(x)\}_{n \geq 1}$ un sucesión de funciones en $\mathbb{R}^d$ tal que $| f_n(x)| \leq g(x)$ $\forall n \geq 1$, donde $g(x)$ es integrable, supongamos que $f(x) = \lim_{n \to \infty} f_n(x)$. Entonces*

$$\lim_{n \to \infty} \int_{\mathbb{R}^d} f_n(x) dx = \int_{\mathbb{R}^d} f(x) dx.$$

y análogamente para sumas.

Observación 1.6.2 *El resultado sigue siendo cierto si cambiamos dx por otra medida en $\mathbb{R}^d$.*

Teorema 1.6.2 *Un modelo exponencial con $a(\theta)$ derivable y $a'(\theta) \neq 0$ es regular.*

Demostración. Veamos que podemos intercambiar la derivada respecto a θ y la integral.

$$\begin{aligned} & \partial_\theta \int_{\mathcal{X}} \exp\{a(\theta)T(x) + b(\theta)\} \exp\{s(x)\}dx \\ = \; & \exp\{b(\theta)\}a'(\theta)\partial_{a(\theta)} \int_{\mathcal{X}} \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx \\ & +b'(\theta). \end{aligned}$$

Entonces si demostramos que

$$\begin{aligned} & \partial_{a(\theta)} \int_{\mathcal{X}} \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx \\ = \; & \int_{\mathcal{X}} \partial_{a(\theta)} \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx \\ = \; & \int_{\mathcal{X}} T(x) \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx, \end{aligned} \tag{1.1}$$

con

$$\int_{\mathcal{X}} |T(x)| \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx < \infty. \tag{1.2}$$

ya estará. Nótese que $b'(\theta)$ estará también bien definido ya que como $\partial_\theta \int_{\mathcal{X}} L(x;\theta)dx = 0$, resultará que

$$b'(\theta) = -\exp\{b(\theta)\}a'(\theta)\partial_{a(\theta)} \int_{\mathcal{X}} \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx$$

y último que el término de la derecha esta bien definido ya estará.

Para ver que (1.1) y (1.2) son ciertos notemos en primer lugar que podemos tomar, sin pérdida de generalidad, $s(x) \equiv 0$. Sea $\tilde{\theta}$ de un entorno de θ, por la desigualdad triangular

$$\begin{aligned} & \int_{\mathcal{X}} |T(x)| \exp\{a(\theta)T(x)\}dx \\ \leq \; & \int_{\mathcal{X}} \left| \frac{\exp\{a(\tilde{\theta})T(x)\} - \exp\{a(\theta)T(x)\}}{a(\tilde{\theta}) - a(\theta)} - T(x)\exp\{a(\theta)T(x)\} \right| dx \\ & + \int_{\mathcal{X}} \left| \frac{\exp\{a(\tilde{\theta})T(x)\} - \exp\{a(\theta)T(x)\}}{a(\tilde{\theta}) - a(\theta)} \right| dx. \end{aligned}$$

Veamos que

$$\left| \frac{\exp\{a(\tilde{\theta})T(x)\} - \exp\{a(\theta)T(x)\}}{a(\tilde{\theta}) - a(\theta)} - T(x)\exp\{a(\theta)T(x)\} \right|$$

está acotado para todo $\tilde{\theta}$ en un entorno de θ, suficientemente pequeño, por una función integrable:

$$\begin{aligned}
&\left|\frac{\exp\{a(\tilde{\theta})T(x)\}-\exp\{a(\theta)T(x)\}}{a(\tilde{\theta})-a(\theta)}-T(x)\exp\{a(\theta)T(x)\}\right| \\
=\;&\exp\{a(\theta)T(x)\}\left|\frac{\exp\{(a(\tilde{\theta})-a(\theta))T(x)\}-1}{a(\tilde{\theta})-a(\theta)}-T(x)\right| \\
\leq\;&\exp\{a(\theta)T(x)\}\sum_{k=2}^{\infty}\frac{|a(\tilde{\theta})-a(\theta)|^{k-1}|T(x)|^k}{k!}.
\end{aligned}$$

De manera que

$$\begin{aligned}
&\left|\frac{\exp\{a(\tilde{\theta})T(x)\}-\exp\{a(\theta)T(x)\}}{a(\tilde{\theta})-a(\theta)}-T(x)\exp\{a(\theta)T(x)\}\right| \\
\leq\;&\frac{1}{h}\exp\{a(\theta)T(x)\}\sum_{k=2}^{\infty}\frac{h^k|T(x)|^k}{k!} \\
\leq\;&\frac{1}{h}(\exp\{(a(\theta)+h)T(x)\}+\exp\{(a(\theta)-h)T(x)\}),
\end{aligned}$$

siempre que

$$|a(\tilde{\theta})-a(\theta)|\leq h.$$

Si tomamos h suficientemente pequeño, como $a'(\theta)\neq 0$, existirán $\theta_1,\theta_2\in\Theta$ tales que

$$\begin{aligned}
a(\theta)+h &= a(\theta_1) \\
a(\theta)-h &= a(\theta_2)
\end{aligned}$$

y como

$$\exp\{-b(\theta_i)\}=\int_{\mathcal{X}}\exp\{a(\theta_i)T(x)\}dx, i=1,2,$$

resultará que $\frac{1}{h}(\exp\{(a(\theta)+h)T(x)\}+\exp\{(a(\theta)-h)T(x)\})$ es integrable y esta cota no depende de $\tilde{\theta}$ en el entorno de θ. Tenemos entonces que se cumple (1.2) y podemos aplicar ahora convergencia dominada en

$$\begin{aligned}
&\left|\lim_{\tilde{\theta}\to\theta}\left(\int_{\mathcal{X}}\frac{\exp\{a(\tilde{\theta})T(x)\}-\exp\{a(\theta)T(x)\}}{a(\tilde{\theta})-a(\theta)}dx-\int_{\mathcal{X}}T(x)\exp\{a(\theta)T(x)\}dx\right)\right| \\
=\;&\lim_{\tilde{\theta}\to\theta}\left|\int_{\mathcal{X}}\frac{\exp\{a(\tilde{\theta})T(x)\}-\exp\{a(\theta)T(x)\}}{a(\tilde{\theta})-a(\theta)}dx-\int_{\mathcal{X}}T(x)\exp\{a(\theta)T(x)\}dx\right| \\
\leq\;&\lim_{\tilde{\theta}\to\theta}\int_{\mathcal{X}}\left|\frac{\exp\{a(\tilde{\theta})T(x)\}-\exp\{a(\theta)T(x)\}}{a(\tilde{\theta})-a(\theta)}-T(x)\exp\{a(\theta)T(x)\}\right|dx \\
=\;&\int_{\mathcal{X}}\lim_{\tilde{\theta}\to\theta}\left|\frac{\exp\{a(\tilde{\theta})T(x)\}-\exp\{a(\theta)T(x)\}}{a(\tilde{\theta})-a(\theta)}-T(x)\exp\{a(\theta)T(x)\}\right|dx=0.
\end{aligned}$$

De manera que

$$\partial_{a(\theta)} \int_{\mathcal{X}} \exp\{a(\theta)T(x)\}dx = \int_{\mathcal{X}} \partial_{a(\theta)} \exp\{a(\theta)T(x)\}dx.$$

Si volviéramos a derivar otra vez respecto a $a(\theta)$ la expresión

$$\int_{\mathcal{X}} T(x) \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx$$

deduciríamos que

$$\begin{aligned} & \partial_{a(\theta)} \int_{\mathcal{X}} T(x) \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx \\ = & \int_{\mathcal{X}} T(x)\partial_{a(\theta)} \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx \\ = & \int_{\mathcal{X}} T^2(x) \exp\{a(\theta)T(x)\} \exp\{s(x)\}dx < \infty \end{aligned} \tag{1.3}$$

De manera que

$$\begin{aligned} I_n(\theta) &= \text{var}_\theta(\partial_\theta \log L) = \text{var}_\theta(a'(\theta)T + b'(\theta)) \\ &= a'(\theta)^2 \text{var}_\theta(T(x)) < \infty. \end{aligned}$$

Como $a'(\theta) \neq 0$ y $T(x)$ no es constante también se cumple que $I_n(\theta) > 0$. Con lo que el modelo cumple las condiciones i), ii) y iii) de regularidad. ■

Observación 1.6.3 *Si $a(\theta)$ es dos veces derivable entonces se puede ver, siguiendo el mismo procedimiento que en la demostración anterior, que se cumple la condición iv) de regularidad de manera que*

$$\begin{aligned} I_n(\theta) &= E_\theta(-\partial_\theta^2 \log L) \\ &= E_\theta\left(-a''(\theta)T - b''(\theta)\right) \\ &= -a''(\theta)E_\theta(T) - b''(\theta). \end{aligned}$$

Como

$$0 = E_\theta(\partial_\theta \log L) = a'(\theta)E_\theta(T) + b'(\theta),$$

resultará finalmente que

$$I_n(\theta) = a''(\theta)\frac{b'(\theta)}{a'(\theta)} - b''(\theta).$$

Observación 1.6.4 *Como*

$$\partial_\theta \log L = a'(\theta)T(x) + b'(\theta),$$

tendremos que

$$T(x) - \frac{-b'(\theta)}{a'(\theta)} = \frac{1}{a'(\theta)}\partial_\theta \log L,$$

con lo que T será un estimador eficiente de $\frac{-b'(\theta)}{a'(\theta)}$.

1.7 Propiedades asintóticas de los métodos de estimación.

Por simplicidad vamos a considerar modelos paramétricos correspondientes a observaciones iid y con Θ un intervalo abierto de $\mathbb{R}$. Si tenemos muestras de tamaño n tendremos una espacio muestral $\mathcal{X}$ y dado un método de estimación tendremos el correspondiente estimador que denotaremos T_n:

$$\begin{aligned} T_n : \mathcal{X} &\to \mathbb{R} \\ (x_1, x_2, ..., x_n) &\to T_n(x_1, x_2, ..., x_n). \end{aligned}$$

El propósito será estudiar el comportamiento de T_n cuando $n \to \infty$. A tal efecto estudiaremos el comportamiento de la sucesión de variables

$$\{T_n(X_1, X_2, ..., X_n)\}_{n\geq 1}.$$

Definición 1.7.1 *Diremos que T_n es un estimador (asintóticamente)* fuertemente *consistente de $g(\theta)$ si*

$$T_n(X_1, X_2, ..., X_n) \underset{n\to\infty}{\overset{c.s.}{\to}} g(\theta), \forall\theta \in \Theta$$

si la convergencia es en probabilidad se dice que es débilmente *consistente.*

Definición 1.7.2 *Diremos que T_n es un estimador asintóticamente insesgado de $g(\theta)$ si*

$$E(T_n(X_1, X_2, ..., X_n)) \underset{n\to\infty}{\to} g(\theta), \forall\theta \in \Theta$$

Definición 1.7.3 *Diremos que T_n es un estimador asintóticamente normal de $g(\theta)$ y varianza asintótica $\frac{\sigma^2}{n}$ si*

$$\sqrt{n}(T_n(X_1, X_2, ..., X_n) - g(\theta)) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} N(0, \sigma^2), \forall\theta \in \Theta$$

a veces se denota $T_n \sim AN(g(\theta), \frac{\sigma^2}{n})$

Observación 1.7.1 *Notemos que si T_n es asintóticamente normal de $g(\theta)$ esto significa*

$$Ley(T_n) \approx N(g(\theta), \frac{\sigma^2}{n})$$

y la aproximación es tanto más cierta cuanto más grande es n.

Definición 1.7.4 *Si $T_n \sim AN(g(\theta), \frac{\sigma^2}{n})$ con*

$$\sigma^2 = \frac{g'(\theta)^2}{I_1(\theta)}$$

se dice que es un estimador asintóticamente eficiente de $g(\theta)$.

Teorema 1.7.1 *Supongamos que el modelo cumple las condiciones de regularidad i) ii) iii) y iv). Y que además v):* $\partial_\theta^2 \log L_1(x;\theta)$ *es una función continua en* θ *y vi)* $|\ \partial_\theta^2 \log L_1(x;\tilde{\theta})| < h_\theta(x)$ *para todo* $\tilde{\theta}$ *de un entorno de* θ*, con* $\int_{\mathcal{X}} h_\theta(x) L_1(x;\theta)dx < \infty$*. Entonces si* $\hat{\theta}_n(x)$ *es una solución fuertemente consistente de las ecuaciones de verosimilitud:*

$$\partial_\theta \log L_n(x;\theta)_{|\theta=\hat{\theta}_n(x)} = 0, \partial_\theta^2 \log L_n(x;\theta)_{|\theta=\hat{\theta}_n(x)} < 0$$

resultará que

$$\sqrt{n}(\hat{\theta}_n(X_1, X_2, ..., X_n) - \theta) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} N(0, \frac{1}{I_1(\theta)}).\forall\theta \in \Theta$$

Demostración.

$$\partial_\theta \log L_n(x;\theta) = \partial_\theta^2 \log L_n(x;\theta)_{|\theta=\theta_n^*}(\theta - \hat{\theta}_n),$$

donde θ_n^* es un punto intermedio entre θ y $\hat{\theta}_n$. De esta manera

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta) &= \sqrt{n}\frac{\partial_\theta \log L_n(x;\theta)}{-\partial_\theta^2 \log L_n(x;\theta)_{|\theta=\theta_n^*}} \\ &= \frac{\frac{1}{\sqrt{n}}\partial_\theta \log L_n(x;\theta)}{-\frac{1}{n}\partial_\theta^2 \log L_n(x;\theta)_{|\theta=\theta_n^*}}. \end{aligned} \tag{1.4}$$

Ahora bien

$$\frac{1}{\sqrt{n}}\partial_\theta \log L_n(X_1, X_2, ..., X_n;\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \partial_\theta \log L_1(X_i;\theta).$$

Para todo $\theta \in \Theta$ las $\partial_\theta \log L_1(X_i;\theta)$ son variables aleatorias iid de media cero y varianza $E_\theta((\partial_\theta \log L_1(X_i;\theta))^2) = I_1(\theta)$. Entonces por el teorema central del límite

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \partial_\theta \log L_1(X_i;\theta) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathrm{N}(0, I_1(\theta)).$$

Por otro lado,

$$-\frac{1}{n}\partial_\theta^2 \log L_n(X_1, X_2, ..., X_n;\theta) = \frac{1}{n}\sum_{i=1}^n -\partial_\theta^2 \log L_1(X_i;\theta),$$

y las variables $-\partial_\theta^2 \log L_1(X_i;\theta)$ son iid de esperanza $I_1(\theta)$, de manera que por la ley fuerte de los grandes números

$$-\frac{1}{n}\partial_\theta^2 \log L_n(X_1, X_2, ..., X_n;\theta) \underset{n\to\infty}{\overset{c.s.}{\to}} I_1(\theta).$$

Entonces si demostramos que

$$\frac{1}{n}\left|\partial_\theta^2 \log L_n(X_1, X_2, ..., X_n;\theta) - \partial_\theta^2 \log L_n(X_1, X_2, ..., X_n;\theta_n^*)\right| \underset{n\to\infty}{\overset{P}{\to}} 0 \tag{1.5}$$

tendremos que

$$-\frac{1}{n}\partial_\theta^2 \log L_n(X_1, X_2, ..., X_n; \theta_n^*) \underset{n\to\infty}{\xrightarrow{P}} I_1(\theta)$$

y aplicando el teorema de Slutsky a (1.4) tendremos

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\frac{1}{\sqrt{n}}\partial_\theta \log L_n(x;\theta)}{-\frac{1}{n}\partial_\theta^2 \log L_n(x;\theta)_{|\theta=\theta_n^*}} \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \frac{\mathrm{N}(0, I_1(\theta))}{I_1(\theta)} = \mathrm{N}(0, \frac{1}{I_1(\theta)}). \quad (1.6)$$

Para probar (1.5), notemos que

$$\begin{aligned}
& \frac{1}{n}\left|\partial_\theta^2 \log L_n(X_1, X_2, ..., X_n; \theta) - \partial_\theta^2 \log L_n(X_1, X_2, ..., X_n; \theta_n^*)\right| \\
\leq \; & \frac{1}{n}\sum_{i=1}^n \left|\partial_\theta^2 \log L_1(X_i;\theta) - \partial_\theta^2 \log L_1(X_i;\theta_n^*)\right| \\
\leq \; & \frac{1}{n}\sum_{i=1}^n \sup_{|\theta-\theta_1|<\delta}\left|\partial_\theta^2 \log L_1(X_i;\theta) - \partial_\theta^2 \log L_1(X_i;\theta_1)\right|, \text{(con probabilidad uno)}
\end{aligned}$$

ya que como $\hat{\theta}_n$ es fuertemente consistente y θ_n^* está entre θ y $\hat{\theta}_n$, para todo $\delta > 0$ si n es suficientemente grande

$$|\theta - \theta_n^*| < \delta$$

con probabilidad uno. Tendremos entonces que

$$\begin{aligned}
& \limsup \frac{1}{n}\left|\partial_\theta^2 \log L_n(X_1, X_2, ..., X_n; \theta) - \partial_\theta^2 \log L_n(X_1, X_2, ..., X_n; \theta_n^*)\right| \\
\leq \; & \lim \frac{1}{n}\sum_{i=1}^n \sup_{|\theta-\theta_1|<\delta}\left|\partial_\theta^2 \log L_1(X_i;\theta) - \partial_\theta^2 \log L_1(X_i;\theta_1)\right| \\
= \; & E_\theta\left(\sup_{|\theta-\theta_1|<\delta}\left|\partial_\theta^2 \log L_1(X_i;\theta) - \partial_\theta^2 \log L_1(X_i;\theta_1)\right|\right),
\end{aligned}$$

donde hemos podemos aplicar la ley fuerte de los grandes números ya que

$$\begin{aligned}
& E_\theta\left(\sup_{|\theta-\theta_1|<\delta}\left|\partial_\theta^2 \log L_1(X_i;\theta) - \partial_\theta^2 \log L_1(X_i;\theta_1)\right|\right) \\
\leq \; & 2E_\theta(h_\theta(X_i)).
\end{aligned}$$

El resultado se sigue haciendo que $\delta \to 0$, aplicando convergencia dominada y la continuidad de las derivadas segundas de la verosimilitud. ■

Proposición 1.7.1 *Supongamos que nuestro modelo paramétrico, correspondiente a observaciones iid, y tal que Θ es un intervalo abierto de $\mathbb{R}$, es localmente* identificable, *esto es la aplicación*

$$\theta \mapsto P_\theta$$

es inyectiva en el entorno de cualquier punto y que

$$\int_{\mathcal{X}} |\log \frac{L_1(x;\tilde{\theta})}{L_1(x;\theta)}| L_1(x;\theta)dx < \infty$$

para todo $\tilde{\theta}$ de un entorno de θ, $\forall \theta \in \Theta$. Entonces las ecuaciones de verosimilitud tienen a partir de un cierto valor de n una solución fuertemente consistente.

Demostración. Consideremos la función

$$\lambda_n(X_1, X_2, ..., X_n; \tilde{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{L_1(X_i;\tilde{\theta})}{L_1(X_i;\theta)}.$$

Por la ley fuerte de los grandes números

$$\lambda_n(X_1, X_2, ..., X_n; \tilde{\theta}) \underset{n\to\infty}{\overset{c.s.}{\to}} E_\theta(\log \frac{L_1(X_1;\tilde{\theta})}{L_1(X_1;\theta)})$$

Sea $k \in \mathbb{N}$ suficientemente grande para que $L_1(X_1;\tilde{\theta})$ sea una variable diferente de $L_1(X_1;\theta)$ si $\tilde{\theta} \in [\theta + \frac{1}{k}, \theta - \frac{1}{k}]$. Esto queda garantizado por la hipótesis de identificabilidad local. Entonces por la desigualdad de Jensen como log es una función estrictamente cóncava y $\frac{L_1(X_1;\tilde{\theta})}{L_1(X_1;\theta)} \neq 1$, resultará que

$$\begin{aligned} E_\theta(\log \frac{L_1(X_1;\tilde{\theta})}{L_1(X_1;\theta)}) &< \log E_\theta(\frac{L_1(X_1;\tilde{\theta})}{L_1(X_1;\theta)}) \\ &= \log \int_{\mathcal{X}} \frac{L_1(x;\tilde{\theta})}{L_1(x;\theta)} L_1(x;\theta)dx \\ &= \log \int_{\mathcal{X}} L_1(x;\tilde{\theta})dx \\ &= \log 1 = 0 \end{aligned}$$

Fijado k, si n es suficiente grande tendremos que

$$\begin{aligned} \lambda_n(X_1, X_2, ..., X_n; \theta + \frac{1}{k}) &< 0 \\ \lambda_n(X_1, X_2, ..., X_n; \theta - \frac{1}{k}) &< 0 \end{aligned}$$

con probabilidad uno salvo en un conjunto N_k tal que $P_\theta(N_k) = 0$. Por definición $\lambda_n(X_1, X_2, ..., X_n; \theta) = 0$. Por tanto en el intervalo $[\theta+\frac{1}{k}, \theta-\frac{1}{k}]$ habrá un máximo local, podemos tomar este valor como $\hat{\theta}_n$, es obvio que entonces $| \hat{\theta}_n - \theta| < \frac{1}{k}$. El resultado se sigue tendiendo k a infinito. ■

1.8 Muestras de una población normal

En esta sección vamos a estudiar las distribuciones de determinados estadísticos cuando los datos que observamos siguen una distribución normal. Estos resultados nos serán de gran ayuda en los cálculos que hagamos en el siguiente capítulo.

Sea A una matriz cuadrada no singular $n \times n$ y $b \in \mathbb{R}^n$. Sea X un vector aleatorio n-dimensional con densidad. Consideremos el vector $Y = AX + b$, sabemos que entonces Y tiene densidad

$$f_Y(y) = f_X(g^{-1}(y))|J_{g^{-1}}(y)|$$

donde $g(x) = Ax + b$, esto es

$$f_Y(y) = f_X(A^{-1}(y - b))|\det A^{-1}|$$

ya que

$$J_{g^{-1}}(y) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} = \det A^{-1}.$$

Si ahora $X_1, X_2, ..., X_n$ son iid $N(0, 1)$

$$\begin{aligned} f_X(x) &= \frac{1}{\left(\sqrt{2\pi}\right)^n} \exp\{-\frac{1}{2}\sum_{i=1}^n x_i^2\} \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^n} \exp\{-\frac{1}{2}x'x\}, \end{aligned}$$

donde $x = (x_1, x_2, ..., x_n)'$ y la prima indica *traspuesta*, entonces

$$f_Y(y) = \frac{1}{\left(\sqrt{2\pi}\right)^n |\det A|} \exp\{-\frac{1}{2}(y - b)'(A^{-1})'A^{-1}(y - b)\}$$

y si escribimos $\Sigma = AA'$, resultará que

$$f_Y(y) = \frac{1}{\left(\sqrt{2\pi}\right)^n |\det \Sigma|^{1/2}} \exp\{-\frac{1}{2}(y - b)'\Sigma^{-1}(y - b)\}.$$

Se dice entonces que Y tiene una distribución normal n-dimensional de media b y matriz de varianzas-covarianzas Σ, se escribe $Y \sim N_n(b, \Sigma)$. Notemos que en efecto

$$\begin{aligned} E(Y) &= AE(X) + b = b \\ Cov(Y) &= E((Y - b)(Y - b)') \\ &= AE(XX')A' \\ &= AA' = \Sigma \end{aligned}$$

y que Σ es definida positiva:

$$y'\Sigma y = y'AA'y = ||A'y||^2 > 0 \text{ para todo } y \neq 0,$$

ya que A es no singular.

Proposición 1.8.1 *Si $Y \sim N_n(b, \Sigma)$, B es una matriz $n \times n$ no singular y $c \in \mathbb{R}^n$ entonces $Z = BY + c \sim N_n(Bb + c, B\Sigma B')$.*

Demostración. Podemos suponer que $Y = AX + b$ con A no singular, $\Sigma = AA'$ y X un vector de $N(0,1)$ independientes, entonces

$$Z = BAX + Bb + c$$

con lo que la nueva media será $Bb+c$ y la nueva matriz de varianzas-covarianzas $BAA'B' = B\Sigma B'$. ■

Corolario 1.8.1 *Si X es un vector n-dimensional de $N(0,1)$ independientes y T es una matriz ortogonal $n \times n$, $Y := TX$ es también un vector n-dimensional de $N(0,1)$ independientes.*

Demostración. $X \sim N_n(0, I_n)$ y por la proposición anterior $Y := TX \sim N_n(0, TT')$ y $TT' = I_n$ ya que T es ortogonal. ■

Proposición 1.8.2 *Si $Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix} \begin{matrix} \}m \\ \}n-m \end{matrix}$ tiene una ley normal n-dimensional y $Cov(Y_{(1)}, Y_{(2)}) = 0$, entonces $Y_{(1)}$ e $Y_{(2)}$ son independientes y con ley normal.*

Demostración. Es inmediata a partir de la factorización de la densidad conjunta. ■

Proposición 1.8.3 *Cualquier* marginal *de una normal multidimensional es normal.*

Demostración. Sea $Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix} \begin{matrix} \}m \\ \}n-m \end{matrix}$ $, Y \sim N_n(b, \Sigma)$, con $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{matrix} \}m \\ \}n-m \end{matrix}$. Hagamos el cambio

$$Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix} \to Z = \begin{pmatrix} Z_{(1)} \\ Z_{(2)} \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ -\Sigma_{12}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix} \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix}.$$

Entonces $Cov(Z_{(1)}, Z_{(2)}) = 0$, de hecho

$$Cov(Z) = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix},$$

y podemos aplicar la proposición anterior. ■

Proposición 1.8.4 *Si L es una matriz $m \times n$ con $m < n$ y m filas linealmente independientes e $Y \sim N_n(b, \Sigma)$ entonces $LY \sim N_m(Lb, L\Sigma L')$.*

Demostración. Basta completar L con $n - m$ filas linealmente independientes y aplicar la proposición anterior. ■

Corolario 1.8.2 *Si $X = (X_1, ..., X_n)'$ es tal que las X_i son iid con ley $N(\mu, \sigma^2)$ entonces $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.*

Demostración. $X \sim N_n(\mu\mathbf{1}, \boldsymbol{\sigma}^2 I_n)$, con $\mathbf{1} = (1, \overset{n)}{\ldots}, 1)'$. Por tanto si tomamos $L = (1/n, ..., 1/n)$ tendremos, por la proposición anterior, que

$$\bar{X} = LX \sim N(\mu L\mathbf{1}, \boldsymbol{\sigma}^2 LL'),$$

y

$$L\mathbf{1} = (1/n, ..., 1/n) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 1$$

$$LL' = (1/n, ..., 1/n) \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix} = \frac{1}{n}$$

■

Definición 1.8.1 *Sean $Z_1, ..., Z_n$ iid $N(0,1)$, entonces $Z_1^2 + .. + Z_n^2$ sigue una ley que se conoce como Ji-cuadrado con n-grados de libertad, se escribe $\mathcal{X}_n^2$. Esto es*

$$\mathcal{X}_n^2 \overset{Ley}{=} Z_1^2 + .. + Z_n^2, \quad Z_i \text{ iid } N(0,1).$$

Observación 1.8.1 *Si $Z_1 \sim N(0,1)$*

$$\begin{aligned} f_{Z_1^2}(u) &= f_{Z_1}(\sqrt{u})\frac{1}{2\sqrt{u}}\mathbf{1}_{(0,\infty)}(u) + f_{Z_1}(-\sqrt{u})\frac{1}{2\sqrt{u}}\mathbf{1}_{(0,\infty)}(u) \\ &= \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{u}{2}}}{\sqrt{u}}\mathbf{1}_{(0,\infty)}(u) = \frac{1}{\sqrt{2\pi}}u^{1/2-1}e^{-\frac{u}{2}}\mathbf{1}_{(0,\infty)}(u), \end{aligned}$$

esto es

$$Z_1^2 \sim Gamma(1/2, 1/2).$$

Es fácil ver que si $W_1 \sim Gamma(\alpha_1, \beta)$ y $W_2 \sim Gamma(\alpha_2, \beta)$ y son independientes entonces $W_1 + W_2 \sim Gamma(\alpha_1 + \alpha_2, \beta)$, de esta manera la ley $\mathcal{X}_n^2$ es una ley $Gamma(n/2, 1/2)$.

Teorema 1.8.1 *(Fisher) Si $X_1, ..., X_n$ son iid con ley $N(\mu, \sigma^2)$ entonces $U := \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \mathcal{X}_{n-1}^2$ independiente de $\bar{X}$.*

Demostración. Tomemos $Y_i = \frac{X_i - \mu}{\sigma}$, entonces $U = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ con Y_i iid $N(0,1)$. Tenemos que,

$$U = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = Y'Y - \left(\sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}}\right)^2,$$

donde $Y = (Y_1, ..., Y_n)'$. Sea T una matriz ortogonal $n \times n$ con su primera fila $(\frac{1}{\sqrt{n}}, \overset{n)}{\ldots}, \frac{1}{\sqrt{n}})$ esto es

$$T = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ . & \cdots & . \\ . & \cdots & . \end{pmatrix},$$

Sea $Z = TY$, entonces $Z_1 = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}}$ y $Z'Z = Y'T'TY = Y'Y$, así

$$\begin{aligned} U &= Y'Y - \left(\sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}}\right)^2 \\ &= Z'Z - Z_1^2 = Z_2^2 + .. + Z_n^2, \end{aligned}$$

y sabemos por el corolario (1.8.1) que Z es un vector de $N(0,1)$ independientes, de manera que U es independiente de $Z_1 = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ y $U \sim \mathcal{X}_{n-1}^2$. ■

Definición 1.8.2 *Si $X \sim N(0,1)$ e $Y \sim \mathcal{X}_n^2$ y X e Y son independientes entonces la ley de*

$$\frac{X}{\sqrt{\frac{Y}{n}}}$$

se conoce como ley t de Student con n grados de libertad, la escribiremos t_n . Esto es

$$\frac{X}{\sqrt{\frac{Y}{n}}} \stackrel{Ley}{=} t_n$$

Observación 1.8.2 *Se puede ver que*

$$f_{t_n}(x) = \frac{1}{\sqrt{\pi}} \frac{1}{n} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} (1 + \frac{x^2}{2})^{-\frac{n+1}{2}}.$$

Notemos también que por la ley de los grandes números y el teorema de Slutsky

$$\frac{X}{\sqrt{\frac{Y}{n}}} \underset{n\to\infty}{\stackrel{Ley}{\to}} N(0,1),$$

la t_n es una distribución simétrica como la $N(0,1)$ pero con las colas más pesadas.

Corolario 1.8.3 *(del teorema de Fisher) Si $X_1, ..., X_n$ son iid con ley $N(\mu, \sigma^2)$ entonces*

$$\sqrt{n}\frac{\bar{X} - \mu}{\tilde{S}} \sim t_{n-1}$$

donde

$$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Demostración. $\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim N(0,1)$ y $\sum_{i=1}^{n}\frac{1}{\sigma^2}(X_i - \bar{X})^2 \sim \mathcal{X}_{n-1}^2$ y son independientes, entonces por la definición de t de Student

$$\frac{\sqrt{n}\frac{\bar{X}-\mu}{\sigma}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\frac{1}{\sigma^2}(X_i - \bar{X})^2}} \sim t_{n-1}$$

pero

$$\frac{\sqrt{n}\frac{\bar{X}-\mu}{\sigma}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\frac{1}{\sigma^2}(X_i - \bar{X})^2}} = \sqrt{n}\frac{\bar{X} - \mu}{\tilde{S}}$$

■

Capítulo 2

Intervalos de confianza

Dado un modelo paramétrico $(\mathcal{X}, \mathcal{F}, \{P_\theta, \theta \in \Theta\})$, con $\Theta \subseteq \mathbb{R}^d$ la idea es dar una región de $\mathbb{R}^d$ en la que se pueda encontrar el verdadero valor de θ. Vamos a concentrarnos en el caso $d = 1$ con lo que nuestras regiones serán "intervalos".

2.1 Construcción de intervalos a partir de la verosimilitud

Sea $x = (x_1, x_2, ..., x_n)$ una muestra de tamaño n y $L(x; \theta)$ la verosimilitud. En estimación puntual utilizamos $\hat{\theta}(x)$ como estimación de θ, sin embargo sabemos que $\hat{\theta}(x)$ no es exactamente θ sino que oscila, al cambiar la muestra, alrededor de θ. Entonces una manera alternativa o complementaria de localizar θ sería dar un conjunto de puntos en lugar de un punto. Podríamos así decir que θ se encuentra en $[\tilde{\theta}_1(x), \tilde{\theta}_2(x)]$, con cierta seguridad o verosimilitud. Notemos que $\tilde{\theta}_1$ y $\tilde{\theta}_2$ estiman los extremos del intervalo y por tanto este va cambiando de muestra a muestra, podemos escribir

$$I(x) = [\tilde{\theta}_1(x), \tilde{\theta}_2(x)].$$

Es evidente que nos gustaría tener un intervalo de longitud $\tilde{\theta}_2(x) - \tilde{\theta}_1(x)$ pequeña (más preciso) frente a uno de longitud grande (impreciso) pero si lo tomamos más pequeño la seguridad de que el intervalo contenga a θ también disminuirá, por tanto debemos buscar un compromiso entre ambas cosas.

Para controlar la seguridad o verosimilitud del "intervalo" podemos tomar el conjunto de valores de θ tales que

$$\frac{L(x; \hat{\theta})}{L(x; \theta)} \leq K$$

para un valor de K no muy grande. Si embargo queda por saber como fijar K.

Una manera podría ser tomar K tal que

$$P_\theta^X\{x, \frac{L(x;\hat{\theta})}{L(x;\theta)} \leq K\} \geq \gamma$$

con un γ próximo a 1.

Definición 2.1.1 *Diremos que los intervalos $I(x)$ para $\theta \in \mathbb{R}$ tienen (coeficiente) de confianza γ si*

$$P_\theta^X(\theta \in I) \geq \gamma$$

esto es $P_\theta^X\{x, \theta \in I(x)\} \geq \gamma$.

Observación 2.1.1 *Habitualmente se suele tomar 0.95 ó 0.99 (si la muestra es muy grande).*

Observación 2.1.2 *Notemos que $P_\theta^X(\theta \in I) = P_\theta(\theta \in I(X))$*

Observación 2.1.3 *Análogamente diríamos que los intervalos $I(x)$ para $g(\theta) \in \mathbb{R}$ tienen (coeficiente) de confianza γ si*

$$P_\theta^X(g(\theta) \in I) \geq \gamma$$

y θ podría ser multidimensional.

Ejemplo 2.1.1 *Sean n observaciones iid con distribución $N(\mu, \sigma^2)$, σ conocida.*

$$\begin{aligned} L(x;\mu) &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\} \\ &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2\} \exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\} \\ &= h(x)\exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\} \end{aligned}$$

entonces, puesto que $\hat{\mu}(x) = \bar{x}$

$$\frac{L(x;\hat{\mu})}{L(x;\mu)} \leq K \Longleftrightarrow \exp\{\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\} \leq K$$

o equivalentemente

$$|\bar{x}-\mu| \leq C_K \frac{\sigma}{\sqrt{n}}$$

donde C_K es una constante que depende del valor de "verosimilitud" elegido, K, pero no de σ ó n. Finalmente obtenemos

$$\bar{x} - C_K\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + C_K\frac{\sigma}{\sqrt{n}}.$$

Sin embargo, no hay una manera clara de como fijar K *o* C_K*. Podemos mirar cómo se comporta el intervalo* $[\bar{x} - C_K \frac{\sigma}{\sqrt{n}}, \bar{x} + C_K \frac{\sigma}{\sqrt{n}}]$ *cuando cambiamos la muestra, en particular*

$$P_\mu(\mu \in [\bar{X} - C_K \frac{\sigma}{\sqrt{n}}, \bar{X} + C_K \frac{\sigma}{\sqrt{n}}])$$

Ahora bien

$$\begin{aligned} P_\mu\{\bar{X} - C_K \frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{X} + C_K \frac{\sigma}{\sqrt{n}}]\} \\ &= P_\mu \left\{ \left| \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right| \leq C_K \right\}. \end{aligned}$$

Sabemos que $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ *de manera que*

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

y la probabilidad anterior No depende de μ *y así podemos encontrar* C_K *independiente de* μ *tal que*

$$P_\mu \left\{ \left| \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right| \leq C_K \right\} \geq \gamma$$

para un valor de γ *prefijado. Por ejemplo, si tomamos* $\gamma = 0.95$*, como (muy aproximadamente)*

$$P\{|N(0,1)| \leq 1,96\} = 0.95$$

bastará tomar $C_K = 1,96$*. Tendremos así dada una muestra* x *el intervalo*

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}},$$

de coeficiente de confianza 0.95.

2.2 Construcción de intervalos a partir de funciones pivotantes.

Como hemos visto en el ejemplo anterior obtener el intervalo para μ de coeficiente γ ha sido posible por la circunstancia de que

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

es decir la distribución de $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ no depende del valor de μ. Esto nos lleva a la siguiente definición general.

Definición 2.2.1 *Sea* $(\mathcal{X}, \mathcal{F}, \{P_\theta^X, \theta \in \Theta\})$, *con* $\Theta \subseteq \mathbb{R}^d$ *un modelo paramétrico, una función pivotante para* $g(\theta) \in \mathbb{R}$ *es una aplicación medible*

$$\begin{array}{rcl} \pi : \mathcal{X} \times g(\Theta) & \to & \mathbb{R} \\ (x, g(\theta)) & \mapsto & \pi(x, g(\theta)) \end{array}$$

y tal que la ley de $\pi(\cdot, g(\theta))$ *no depende de* θ.

Ejemplo 2.2.1 *Sea un modelo de* n *observaciones iid con distribución* $N(\mu, \sigma^2)$ *entonces*

$$\pi(x, \mu) = \sqrt{n}\frac{\bar{x} - \mu}{\tilde{s}}$$

donde

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

es una función pivotante para μ. *En efecto sabemos que (por el teorema de Fisher) que*

$$\pi(\cdot, \mu) \sim t_{n-1} \text{ de Student.}$$

Asimismo

$$\pi(x, \sigma) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$$

es una función pivotante ya que (también por el teorema de Fisher)

$$\pi(\cdot, \sigma) \sim \mathcal{X}_{n-1}^2$$

Proposición 2.2.1 *Sea* $\pi(x, g(\theta))$ *una función pivotante para* $g(\theta)$ *y* $B \in \mathcal{B}(\mathbb{R})$ *tal que*

$$P_\theta^X\{\pi(\cdot, g(\theta)) \in B\} \geq \gamma$$

entonces

$$I(x) = \{g(\theta), \pi(x, g(\theta)) \in B\}$$

es una región de coeficiente de confianza γ.

Demostración. Tenemos que ver que

$$P_\theta^X\{x, g(\theta) \in I(x)\} \geq \gamma,$$

pero decir que $g(\theta) \in I(x)$ es lo mismo que decir que $\pi(x, g(\theta)) \in B$, entonces

$$\begin{array}{rcl} P_\theta^X\{x, g(\theta) & \in & I(x)\} = P_\theta^X\{x, \pi(x, g(\theta)) \in B\} \\ & = & P_\theta^X\{\pi(\cdot, g(\theta)) \in B\} \geq \gamma. \end{array}$$

■

Observación 2.2.1 *Normalmente* B *será un intervalo* $[a, b]$ *de manera que* $\pi(x, g(\theta)) \in B$ *equivaldrá a* $a \leq \pi(x, g(\theta)) \leq b$, *entonces si* $\pi(x, \cdot)$ *es una función estrictamente monótona creciente tendremos que esto equivale a decir que* $\pi^{-1}(x, \cdot)(a) \leq g(\theta) \leq \pi^{-1}(x, \cdot)(b)$.

Ejemplo 2.2.2 *(Intervalo para μ, σ desconocida) Sea un modelo de n observaciones iid con distribución $N(\mu, \sigma^2)$ entonces $\pi(x, \mu) = \sqrt{n}\frac{\bar{x}-\mu}{\tilde{s}}$ es un pivote para μ con distribución t_{n-1} de Student entonces si tomamos el intervalo $[-t_{n-1;\alpha/2}; t_{n-1;\alpha/2}]$ donde $t_{n-1;\alpha/2}$ es un valor tal que*

$$P(t_{n-1} > t_{n-1;\alpha/2}) = \alpha/2$$

con $\gamma = 1 - \alpha$, y donde t_{n-1} denota una variable con ley t_{n-1} de Student, resultará que

$$P_{(\mu,\sigma)}(-t_{n-1;\alpha/2} \leq \sqrt{n}\frac{\mu - \bar{X}}{\tilde{s}} \leq t_{n-1;\alpha/2}) = \gamma$$

y de aqui,

$$-t_{n-1;\alpha/2} \leq \sqrt{n}\frac{\mu - \bar{x}}{\tilde{s}} \leq t_{n-1;\alpha/2}$$

proporcionará un intervalo de confianza γ. Si aislamos μ tendremos que lo anterior equivale a

$$\bar{x} - t_{n-1;\alpha/2}\frac{\tilde{s}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1;\alpha/2}\frac{\tilde{s}}{\sqrt{n}}$$

Ejemplo 2.2.3 *(Intervalo para σ) Sea de nuevo un modelo de n observaciones normales independientes $N(\mu, \sigma^2)$ entonces $\pi(x, \sigma) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2}$ es una función pivotante para σ con distribución $\mathcal{X}^2_{n-1}$ entonces si tomamos los puntos $\mathcal{X}^2_{n-1;1-\alpha/2}$ y $\mathcal{X}^2_{n-1;\alpha/2}$ (con la misma notación de antes) tendremos que*

$$\mathcal{X}^2_{n-1;1-\alpha/2} \leq \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \leq \mathcal{X}^2_{n-1;\alpha/2}$$

proporciona un intervalo de confianza γ. En efecto, de aquí aislando σ^2 obtenemos

$$\frac{n\hat{\sigma}^2}{\mathcal{X}^2_{n-1;\alpha/2}} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{\mathcal{X}^2_{n-1;1-\alpha/2}}$$

con $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$.

2.3 Problemas de dos muestras.

2.3.1 Muestras independientes

Supongamos un modelo correspondiente a observaciones $X_1, X_2, ..., X_{n_1}, Y_1, Y_2, ..., Y_{n_2}$ donde las X son iid con distribución $N(\mu_1, \sigma_1^2)$ y las Y son iid con ley $N(\mu_2, \sigma_2^2)$. Asumimos también que ("las dos muestras") las X y las Y son independientes.

Intervalo de confianza para la diferencia de medias.

Supongamos que $\sigma_1 = \sigma_2$. Entonces, como (por le teorema de Fisher y la independencia de X e Ys)

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$$

y

$$\frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2}{\sigma^2} + \frac{\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}{\sigma^2} \sim \mathcal{X}^2_{n_1+n_2-2}$$

y además son independientes, resultará que

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1+n_2-2}(\frac{1}{n_1} + \frac{1}{n_2})(\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2)}} \sim t_{n_1+n_2-2}.$$

De aquí tendremos que

$$\pi(x, y, \mu_1 - \mu_2) = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1+n_2-2}(\frac{1}{n_1} + \frac{1}{n_2})(\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2)}}$$

es una función pivotante para $\mu_1 - \mu_2$. Y obtendremos un intervalo de confianza γ para $\mu_1 - \mu_2$ dado por

$$\begin{aligned} & \bar{x} - \bar{y} - t_{n_1+n_2-2;\alpha/2}\sqrt{\frac{n_1+n_2}{n_1+n_2-2}(\frac{\hat{\sigma}_1^2}{n_2} + \frac{\hat{\sigma}_2^2}{n_1})} \leq \mu_1 - \mu_2 \\ \leq \quad & \bar{x} - \bar{y} + t_{n_1+n_2-2;\alpha/2}\sqrt{\frac{n_1+n_2}{n_1+n_2-2}(\frac{\hat{\sigma}_1^2}{n_2} + \frac{\hat{\sigma}_2^2}{n_1})}. \end{aligned}$$

Si $\sigma_1 \neq \sigma_2$ lo anterior no proporciona un pivote para $\mu_1 - \mu_2$ ya que no hay manera de "eliminar" σ_1^2 y σ_2^2. Hay diversas soluciones aproximadas, la más sencilla se basa en el hecho de que

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

de manera que aproximadamente

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \sim N(0, 1)$$

y tenemos el intervalo de confianza aproximadamente γ

$$\bar{x} - \bar{y} - k_{\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + k_{\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

donde $P(N(0, 1) > k_{\alpha/2}) = \alpha/2$ y, como siempre, $\gamma = 1 - \alpha$.

Intervalo para el cociente de varianzas

Como
$$\frac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2}{\sigma_1^2}=\mathcal{X}^2_{n_1-1}$$
y
$$\frac{\sum_{i=1}^{n_1}(Y_i-\bar{Y})^2}{\sigma_2^2}=\mathcal{X}^2_{n_2-1}$$
son independientes, si definimos la distribución F_{n_1,n_2} de Fisher (con n_1 y n_2 grados de libertad) como la que se obtiene al hacer el cociente
$$\frac{\mathcal{X}^2_{n_1}/n_1}{\mathcal{X}^2_{n_2}/n_2}$$
donde numerador y denominador son independientes, tendremos que
$$\frac{\tilde{s}_1^2/\sigma_1^2}{\tilde{s}_2^2/\sigma_2^2}\sim F_{n_1-1,n_2-1}$$
y esto nos dará el intervalo de confianza γ
$$F_{n_1-1,n_2-1;1-\alpha/2}\frac{\tilde{s}_2^2}{\tilde{s}_1^2}\leq\frac{\sigma_2^2}{\sigma_1^2}\leq F_{n_1-1,n_2-1;\alpha/2}\frac{\tilde{s}_2^2}{\tilde{s}_1^2}.$$

2.3.2 Muestras relacionadas

Supongamos un modelo correspondiente a observaciones $X_1,X_2,...,X_n,Y_1,Y_2,...,Y_n$ donde las X e Y son dependientes, pero X_i-Y_i son iid con distribución $N(\mu_1-\mu_2,\sigma^2)$ (siendo μ_1 la media común de las X y μ_2 la media común de las Y) entonces utilizando el teorema de Fisher, un intervalo de confianza de coeficiente γ para $\mu_1-\mu_2$ vendrá dado por
$$\bar{x}-\bar{y}-t_{n-1;\alpha/2}\frac{\tilde{s}}{\sqrt{n}}\leq\mu_1-\mu_2\leq\bar{x}-\bar{y}+t_{n-1;\alpha/2}\frac{\tilde{s}}{\sqrt{n}}$$
con $\tilde{s}^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-y_i-(\bar{x}-\bar{y}))^2$.

2.4 Algunos métodos para obtener pivotes

2.4.1 Un método bastante general

Supongamos un modelo asociado a observaciones iid $X_1,X_2,...,X_n$ con función de distribución F_θ continua, $\theta\in\mathbb{R}$. Sabemos que entonces $F_\theta(X_i)\sim U(0,1)$ y consecuentemente $-\log\,F_\theta(X_i)$ $\sim$Exp(1) =Gamma$(1,1)$. Finalmente por la independencia,
$$\sum_{i=1}^n-\log F_\theta(X_i)\sim\text{Gamma}(n,1)=\text{Gamma}(\frac{2n}{2},1)$$

y como 2×Gamma($\frac{2n}{2}$, 1) =Gamma($\frac{2n}{2}$, $\frac{1}{2}$) = $\mathcal{X}^2_{2n}$, resultará que

$$-2\sum_{i=1}^{n}\log F_\theta(X_i) \sim \mathcal{X}^2_{2n}$$

y por tanto

$$\pi(x,\theta) = -2\sum_{i=1}^{n}\log F_\theta(x_i)$$

será una función pivotante para θ con una distribución que está tabulada.

2.4.2 Familias de posición y escala

Supongamos un modelo asociado a observaciones iid $X_1, X_2, ..., X_n$ con densidad de la forma

$$f(x;\mu) = f_0(x-\mu)$$

se dice que en tal caso es una *familia de posición* generada por f_0. En tal caso $Y_i = X_i - \mu$ tiene densidad $f_0(y)$ y por tanto

$$\pi(x;\mu) = g(x_1-\mu, x_2-\mu, ..., x_n-\mu)$$

es una función pivotante para μ cualquiera que sea la función g.

Si las X_i tienen densidad

$$f(x;\sigma) = \frac{1}{\sigma}f_0(\frac{x}{\sigma})$$

se dice que es una *familia de escala* generada por f_0. En tal caso $Y_i = X_i/\sigma$ tiene densidad $f_0(y)$ y por tanto

$$\pi(x;\mu) = g(\frac{x_1}{\sigma}, \frac{x_2}{\sigma}..., \frac{x_n}{\sigma})$$

es una función pivotante para σ cualquiera que sea la función g.

Por último si las X_i tienen densidad

$$f(x;\sigma) = \frac{1}{\sigma}f_0(\frac{x-\mu}{\sigma})$$

se dice que es una *familia de posición y escala* generada por f_0. En tal caso $Y_i = (X_i-\mu)/\sigma$ tiene densidad $f_0(y)$ y por tanto

$$\pi(x;\mu,\sigma) = g(\frac{x_1-\mu}{\sigma}, \frac{x_2-\mu}{\sigma}..., \frac{x_n-\mu}{\sigma})$$

es una función pivotante para (μ,σ) cualquiera que sea la función g.

Ejemplo 2.4.1 *Algunas familias de posición y escala*

Distribución	*Densidad*
Normal: $N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\}$
Laplace: $La(\mu,\sigma)$	$\frac{1}{2\sigma}\exp\{-\vert\frac{x-\mu}{\sigma}\vert\}$
Cauchy: $Cau(\mu,\sigma)$	$\frac{1}{\pi\sigma}\frac{1}{1+(\frac{x-\mu}{\sigma})^2}$
Exponencial: $Exp(\mu,\sigma)$	$\frac{1}{\sigma}\exp\{-\frac{x-\mu}{\sigma}\}\mathbf{1}_{(0,\infty)}(\frac{x-\mu}{\sigma})$
Uniforme: $U(\mu,\mu+\sigma)$	$\frac{1}{\sigma}\mathbf{1}_{(0,1)}(\frac{x-\mu}{\sigma})$

2.4.3 Métodos aproximados

Utilizando la desigualdad de Chebichef.

Supongamos que queremos obtener un intervalo de confianza para $g(\theta) = E_\theta(T)$, T un estadístico de cuadrado integrable. Por al desigualdad de Chebichef

$$P_\theta^X(|T - g(\theta)| > \varepsilon) \leq \frac{Var_\theta(T)}{\varepsilon^2},$$

por tanto si tomamos $1-\gamma = \frac{Var_\theta(T)}{\varepsilon^2}$ resultará que

$$P_\theta^X(|T - g(\theta)| \leq \varepsilon) \geq \gamma$$

con

$$\varepsilon = \sqrt{\frac{Var_\theta(T)}{1-\gamma}}$$

de manera que

$$|T(x) - g(\theta)| \leq \sqrt{\frac{Var_\theta(T)}{1-\gamma}}$$

proporciona un intervalo de confianza γ, despejando $g(\theta)$ obtenemos

$$T(x) - \sqrt{\frac{Var_\theta(T)}{1-\gamma}} \leq g(\theta) \leq T(x) + \sqrt{\frac{Var_\theta(T)}{1-\gamma}}.$$

El problema es que no conocemos $Var_\theta(T)$ ya que depende de θ podemos entonces estimarla por $Var_{\hat{\theta}}(T)$ donde $\hat{\theta}$ es un estimador de θ. El intervalo obtenido será ahora solo de confianza aproximadamente γ. En algunos casos esto se puede evitar.

Ejemplo 2.4.2 *Supongamos que nuestras observaciones son Bernoullis*(p) *y queremos dar un intervalo de confianza para* p. *Podemos tomar* $T(x) = \bar{x}$. *Sabemos que entonces*

$$\begin{aligned} Var_p(T) &= Var_p(T(X)) = \frac{1}{n^2}Var_p(\sum_{i=1}^n X_i) \\ &= \frac{nVar_p(X_1)}{n^2} = \frac{p(1-p)}{n} \end{aligned}$$

de manera que el valor máximo que puede tomar $Var_p(T)$ *es* $\frac{1}{4n}$. *Así*

$$\bar{x} - \frac{1}{2\sqrt{n(1-\gamma)}} \leq p \leq \bar{x} + \frac{1}{2\sqrt{n(1-\gamma)}}$$

tiene confianza γ, *aunque es de longitud demasiado grande. Para* $\gamma = 0.95$ *el radio es aprox* $2.24/\sqrt{n}$

Método basado en el comportamiento asintótico del EMV

Sabemos que en condiciones de regularidad y en el esquema iid, aproximadamente

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\frac{1}{I_1(\hat{\theta})}}} \sim N(0,1)$$

con lo que

$$\hat{\theta} - \frac{k_{\alpha/2}}{\sqrt{nI_1(\hat{\theta})}} \leq \theta \leq \hat{\theta} + \frac{k_{\alpha/2}}{\sqrt{nI_1(\hat{\theta})}}$$

será un intervalo de confianza aproximadamente γ, donde $P_\theta\{N(0,1) > k_{\alpha/2}\} = \alpha/2$, y, como siempre, $\gamma = 1 - \alpha$.

Ejemplo 2.4.3 *Supongamos, como antes, que nuestras observaciones son Bernoullis*(p) *y queremos dar un intervalo de confianza para* p. *Sabemos que el EMV es* $\hat{p}(x) = \bar{x}$, *también sabemos que* $I_1(p) = \frac{1}{p(1-p).}$. *De manera que*

$$\bar{x} - k_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \leq p \leq \bar{x} + k_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}$$

sera un intervalo para de confianza aproximada γ. *En lugar de estimar* $p(1-p)$ *podemos acotarlo por* $\frac{1}{4}$ *y tendríamos*

$$\bar{x} - k_{\alpha/2}\frac{1}{2\sqrt{n}} \leq p \leq \bar{x} + k_{\alpha/2}\frac{1}{2\sqrt{n}}.$$

Si $\gamma = 0.95$, $k_{0.025} = 1.96$ *de manera que aproximadamente*

$$\bar{x} - \frac{1}{\sqrt{n}} \leq p \leq \bar{x} + \frac{1}{\sqrt{n}}$$

nos da un intervalo de confianza 0.95 *(aprox.). Su radio es la mitad del obtenido por Chebichef!. Notemos que si queremos un error inferior al 3% necesitaremos que*

$$\frac{1}{\sqrt{n}} \leq 0.03$$

esto es

$$n \geq \left(\frac{1}{0.03}\right)^2 = 1111,11.$$

En la práctica, por ejemplo, para saber la proporción teórica de individuos con una determinada característica de una población muy grande, se suele tomar un tamaño muestral de 1200, 1500.

2.4.4 Un método especial

El método se basa en el siguiente resultado:

Proposición 2.4.1 *Supongamos un modelo dependiente de un parámetro* $\theta \in \mathbb{R}$, *un estadístico* $T(x)$ *y funciones* $a_1(\theta)$ *y* $a_2(\theta)$, monótonas estrictamente crecientes*, tales que, fijado* $\gamma \in (0,1)$

$$P_\theta^X\{x, a_1(\theta) \leq T(x) \leq a_2(\theta)\} \geq \gamma, \forall \theta \in \Theta$$

entonces

$$a_2^{-1}(T(x)) \leq \theta \leq a_1^{-1}(T(x))$$

nos proporciona un intervalo de confianza γ *para* θ.

Demostración. *Si* x *verifica*

$$a_2^{-1}(T(x)) \leq \theta \leq a_1^{-1}(T(x))$$

resultará que

$$\begin{array}{rcl} T(x) & \leq & a_2(\theta) \\ a_1(\theta) & \leq & T(x) \end{array}$$

o lo que es lo mismo $a_1(\theta) \leq T(x) \leq a_2(\theta)$, *con lo que*

$$P_\theta^X\{x, a_2^{-1}(T(x)) \leq \theta \leq a_1^{-1}(T(x))\} = P_\theta^X\{x, a_1(\theta) \leq T(x) \leq a_2(\theta)\} \geq \gamma$$

■

Ejemplo 2.4.4 *Supongamos el modelo del ejemplo anterior y* $T(x) = \bar{x}$, *tenemos que buscar* $a_1(p)$ *y* $a_2(p)$ *tales que*

$$P_p^X\{x, a_1(p) \leq \bar{x} \leq a_2(p)\} \geq \gamma$$

o equivalentemente

$$P_p\{a_1(p) \leq \bar{X} \leq a_2(p)\} \geq \gamma$$

donde $n\bar{X}$ $\sim$*Binomial*(n,p). *Para determinar* $a_1(p)$ *buscamos, con ayuda de la distribución Binomial*(n,p), *el valor más grande que verifique*

$$P_p\{n\bar{X} < na_1(p)\} \leq \alpha/2$$

se puede ver que $a_1(p)$ *es estrictamente creciente con* p, *y análogamente para* $a_2(p)$, *buscamos el valor más pequeño tal que*

$$P_p\{n\bar{X} > na_2(p)\} \leq \alpha/2.$$

Existen gráficos para $a_1(p)$ *y* $a_2(p)$ *(haces de Clopper-Pearson) para diferentes valores de* γ *y* n *que nos permiten obtener los intervalos. Si miramos el ejemplo anterior también teníamos que aproximadamente*

$$P_p^X\{x, \frac{n(\bar{x}-p)^2}{p(1-p)} \leq k_{\alpha/2}^2\} = \gamma,$$

Ahora bien

$$\begin{aligned}\frac{n(\bar{x}-p)^2}{p(1-p)} &\leq k^2_{\alpha/2} \Longleftrightarrow n(\bar{x}^2 - 2p\bar{x} + p^2) \leq k^2_{\alpha/2}(p-p^2) \\ &\Longleftrightarrow (1 + \frac{k^2_{\alpha/2}}{n})p^2 - (2\bar{x} + \frac{k^2_{\alpha/2}}{n})p + \bar{x}^2 \leq 0\end{aligned}$$

Las soluciones de la ecuación que se obtiene igualando a cero son

$$\begin{aligned}d(\bar{x}) &= (n\bar{x} + k^2_{\alpha/2}/2 - k_{\alpha/2}\sqrt{n\bar{x}(1-\bar{x}) + k^2_{\alpha/2}/4})/(n + k^2_{\alpha/2}) \\ u(\bar{x}) &= (n\bar{x} + k^2_{\alpha/2}/2 + k_{\alpha/2}\sqrt{n\bar{x}(1-\bar{x}) + k^2_{\alpha/2}/4})/(n + k^2_{\alpha/2})\end{aligned}$$

de manera que un intervalo aproximado de confianza γ *viene dado por*

$$d(\bar{x}) \leq p \leq u(\bar{x}).$$

Si np *y* $n(1-p)$ *valen al menos 5 estos intervalos son satisfactorios en la práctica.*

2.4.5 Regiones de confianza, intervalos simultáneos

Región para la media y la desviación típica de observaciones normales

Supongamos que el modelo corresponde a n observaciones iid $N(\mu, \sigma^2)$, μ y σ desconocidos, queremos dar una región de confianza γ para (μ, σ), es decir un subconjunto $B(x)$ (dependerá de la muestra x) $\mathbb{R} \times \mathbb{R}_+$ de manera que

$$P^X_{(\mu,\sigma)}\{x, (\mu, \sigma) \in B(x)\} = \gamma.$$

Por el teorema de Fisher

$$\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \sim N(0,1), \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \mathcal{X}^2_{n-1}$$

y son independientes. Entonces, utilizando la notación habitual, tendremos que

$$\begin{aligned}P_{(\mu,\sigma)}\{-k_{1-\frac{\sqrt{\gamma}}{2}} &\leq \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \leq k_{1-\frac{\sqrt{\gamma}}{2}}\} = \sqrt{\gamma}, \\ P_{(\mu,\sigma)}\{\mathcal{X}^2_{n-1;\frac{\sqrt{\gamma}}{2}} &\leq \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \leq \mathcal{X}^2_{n-1;1-\frac{\sqrt{\gamma}}{2}}\} = \sqrt{\gamma}\end{aligned}$$

y por la independencia

$$P_{(\mu,\sigma)}\left\{\begin{aligned}&-k_{1-\frac{\sqrt{\gamma}}{2}} \leq \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \leq k_{1-\frac{\sqrt{\gamma}}{2}}\ , \\ &\mathcal{X}^2_{n-1;\frac{\sqrt{\gamma}}{2}} \leq \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \leq \mathcal{X}^2_{n-1;1-\frac{\sqrt{\gamma}}{2}}\end{aligned}\right\} = \gamma$$

de manera que

$$
\begin{aligned}
-k_{1-\frac{\sqrt{\gamma}}{2}} &\leq \sqrt{n}\frac{(\bar{x}-\mu)}{\sigma} \leq k_{1-\frac{\sqrt{\gamma}}{2}}, \\
\mathcal{X}^2_{n-1;\frac{\sqrt{\gamma}}{2}} &\leq \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{\sigma^2} \leq \mathcal{X}^2_{n-1;1-\frac{\sqrt{\gamma}}{2}}
\end{aligned}
$$

nos da una región de confianza γ. Más explícitamente,

$$
\begin{aligned}
\bar{x} - k_{1-\frac{\sqrt{\gamma}}{2}}\frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{x} + k_{1-\frac{\sqrt{\gamma}}{2}}\frac{\sigma}{\sqrt{n}}, \\
\frac{\sqrt{n}\hat{\sigma}}{\sqrt{\mathcal{X}^2_{n-1;1-\frac{\sqrt{\gamma}}{2}}}} &\leq \sigma \leq \frac{\sqrt{n}\hat{\sigma}}{\sqrt{\mathcal{X}^2_{n-1;\frac{\sqrt{\gamma}}{2}}}}
\end{aligned}
$$

de manera que obtenemos una región trapezoidal.

El principio de unión-intersección

El llamado "principio de unión-intersección" consiste en un método general, basado en una desigualdad elemental, para obtener intervalos simultáneos, es decir regiones rectangulares, para diferentes parámetros. Sea una muestra de observaciones $X_1, X_2, ..., X_n$ cuya ley depende de dos parámetros θ_1 y θ_2 y supongamos que sabemos encontrar intervalos de confianza γ_1 y γ_2 para θ_1 y θ_2 respectivamente, y sean estos $[S_1(X), S_2(X)]$ y $[T_1(X), T_2(X)]$, esto es

$$
P_{\theta_1,\theta_2}(S_1(X) \leq \theta_1 \leq S_2(X)) \geq \gamma_1, \quad P_{\theta_1,\theta_2}(T_1(X) \leq \theta_2 \leq T_2(X)) \geq \gamma_2.
$$

Pero cuánto vale

$$
P_{\theta_1,\theta_2}(S_1(X) \leq \theta_1 \leq S_2(X), T_1(X) \leq \theta_2 \leq T_2(X)).
$$

Dados dos eventos A_1y A_2

$$
\begin{aligned}
P(A_1 \cap A_2) &= 1 - P(A_1^c \cup A_2^c) \geq 1 - P(A_1^c) - P(A_2^c) \\
&\geq P(A_1) + P(A_2) - 1.
\end{aligned}
$$

Por tanto

$$
P_{\theta_1,\theta_2}(S_1(X) \leq \theta_1 \leq S_2(X), T_1(X) \leq \theta_2 \leq T_2(X)) \geq \gamma_1 + \gamma_2 - 1.
$$

Entonces si tomamos $\gamma_1 = \gamma_2 = \frac{1+\gamma}{2}$ tendremos intervalos simultáneos de coeficiente γ. Si escribimos $\alpha_i = 1 - \gamma_i$, $i = 1, 2$ y $\alpha = 1 - \gamma$ la condición anterior se puede escribir $\alpha_1 = \alpha_2 = \alpha/2$. La generalización es evidente para el caso de d parámetros: $\alpha_1 = \alpha_2 = ... = \alpha_d = \alpha/d$.

Capítulo 3

Test de hipótesis

Hasta ahora hemos utilizado la muestra x para sugerir un valor o un conjunto de valores posibles de θ, el parámetro de nuestro modelo $(\mathcal{X}, \mathcal{F}, \{P_\theta, \theta \in \Theta\})$, pero ahora lo que queremos es validar o rechazar un valor o un conjunto de valores de θ. Por ejemplo, sean Θ_0 y Θ_1 una partición de Θ. Queremos saber si $\theta \in \Theta_0$ y nos ayudaremos de la muestra para decidir si es razonable suponer que $\theta \in \Theta_0$ ó no. Lo formularemos en términos de hipótesis, la hipótesis que llamaremos *nula,* que denotaremos H_0, y que será que $\theta \in \Theta_0$ y la hipótesis que llamaremos *alternativa*, que denotaremos H_1 y que será que $\theta \in \Theta_1$. Escribiremos que nuestro problema es contrastar

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1. \end{aligned}$$

Un test (o contraste) de hipótesis será una aplicación

$$\begin{aligned} \delta : \mathcal{X} &\rightarrow \{H_0, H_1\} \\ x &\mapsto \delta(x). \end{aligned}$$

Todo test de hipótesis establece entonces una partición de $\mathcal{X}$ en dos conjuntos: $A_0 = \delta^{-1}(H_0)$ y $A_1 = \delta^{-1}(H_1)$. A_0 es el conjunto de muestras que nos llevarían a aceptar H_0 según δ y A_1 el conjunto de muestras que nos llevarían a rechazar H_0 (y aceptar H_1 entonces). A A_0 se le llama *región de aceptación* y a A_1 *región crítica.* Como vemos todo test δ queda definido por su región de aceptación (o su región crítica).

Buscaremos test o regiones críticas óptimos en el sentido que nos equivoquemos lo menos posible al apostar por una u otra hipótesis. Pero ¿qué significa equivocarse?, pues rechazar H_0 cuando es cierta (lo llamaremos *error de primera especie* o tipo I) o aceptar H_0 cuando es falsa (lo llamaremos *error de segunda especie* o tipo II). En síntesis, dada una muestra x, puede ocurrir lo siguiente

	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$x \in A_0$	acertamos	error tipo II
$x \in A_1$	error tipo I	acertamos.

Es evidente que, salvo casos triviales no podremos eliminar los dos errores a la vez y si tratamos de disminuir la frecuencia de uno la del otro en general aumentará. Como veremos vamos a dar más importancia a los errores del tipo I de manera que sólo rechacemos la hipótesis nula cuando haya mucha evidencia en este sentido.

3.1 Test de hipótesis simples

Consideremos la situación sencilla donde $\Theta_0 = \{\theta_0\}$ y $\Theta_1 = \{\theta_1\}$. Cuando, como aquí, sólo hay un posible valor en las hipótesis se dice que las hipótesis son *simples* si no se dice que son *compuestas*. En el caso de hipótesis simples podemos calcular las probabilidades de error de tipos I y II:

$$\begin{array}{rcl} \text{Prob de rechazar } H_0 \text{ cuando es cierta} & = & P_{\theta_0}(A_1) \\ \text{Prob de aceptar } H_0 \text{ cuando es falsa=} & & P_{\theta_1}(A_0). \end{array}$$

Adoptaremos el llamado enfoque de Neyman que consiste en tomar A_1de manera que $P_{\theta_0}(A_1) \leq \alpha$ (normalmente $\alpha = 0.05$ ó 0.01 (si la muestra es grande)) y buscar entre todas estas regiones cíticas A_1 las que minimizan $P_{\theta_1}(A_0)$. Como

$$P_{\theta_1}(A_0) = 1 - P_{\theta_1}(A_1),$$

esto equivale a maximizar $P_{\theta_1}(A_1) := \beta$.

Definición 3.1.1 *Dado un test con región crítica A_1 diremos que es de* tamaño *α si $P_{\theta_0}(A_1) = \alpha$.*

Definición 3.1.2 *Dado un test con región crítica A_1 diremos que es de* nivel (de significación) *α si $P_{\theta_0}(A_1) \leq \alpha$.*

Definición 3.1.3 *Dado un test con región crítica A_1 diremos que es de* potencia *β si $P_{\theta_1}(A_1) = \beta$.*

El propósito es buscar entre los test de nivel α el de mayor potencia. Sea $L(x;\theta)$ la verosimilitud de nuestro modelo, en el caso que consideramos θ sólo puede tomar dos valores θ_0 y θ_1. Escribamos

$$\begin{array}{rcl} L_0(x) & = & L(x;\theta_0) \\ L_1(x) & = & L(x;\theta_1), \end{array}$$

Una región crítica razonable sería

$$A_1 = \{x, L_1(x) > KL_0(x)\},$$

con $K > 0$ y suficientemente grande, es decir nos quedamos con la hipótesis de que $\theta = \theta_1$ si θ_1 es mucho más verosímil que θ_0. En realidad esto nos va a dar test óptimos. Tenemos el siguiente teorema conocido más bien como *Lema de Neyman-Pearson*

Teorema 3.1.1 *(Lema de Neyman-Pearson) Sea un test con región crítica A_1 tal que*

$$\begin{aligned} \{x, L_1(x) &> KL_0(x)\} \subseteq A_1 \\ \{x, L_1(x) &< KL_0(x)\} \subseteq A_0, \end{aligned} \tag{3.1}$$

donde $K \geq 0$, supongamos que $P_{\theta_0}(A_1) = \alpha$, entonces es de máxima potencia entre los test de nivel de significación α.

Demostración. *Vamos a suponer que nuestro modelo es de observaciones con densidad. Sea $\tilde{A}_1$ otra región crítica, tenemos que*

$$\int_{\mathcal{X}} (\mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x))(L_1(x) - KL_0(x))dx \geq 0$$

ya que el integrando es siempre no negativo:

$$\begin{aligned} \text{Si } x &\in A_1, \\ L_1(x) - KL_0(x) &\geq 0 \text{ y } \mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x) = 1 - \mathbf{1}_{\tilde{A}_1}(x) \geq 0 \end{aligned}$$

$$\begin{aligned} \text{Si } x &\in A_0, \\ L_1(x) - KL_0(x) &\leq 0 \text{ y } \mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x) = 0 - \mathbf{1}_{\tilde{A}_1}(x) \leq 0. \end{aligned}$$

Por tanto

$$\int_{\mathcal{X}} (\mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x))L_1(x)dx \geq K \int_{\mathcal{X}} (\mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x))L_0(x)dx,$$

equivalentemente

$$P_{\theta_1}(A_1) - P_{\theta_1}(\tilde{A}_1) \geq K(P_{\theta_0}(A_1) - P_{\theta_0}(\tilde{A}_1)) \geq 0$$

ya que $K \geq 0$ y $P_{\theta_0}(A_1) - P_{\theta_0}(\tilde{A}_1) = \alpha - P_{\theta_0}(\tilde{A}_1) \geq 0$ ya que $\tilde{A}_1$ es de nivel α.
■

Observación 3.1.1 *Los test anteriores, que verifican (3.1), los llamaremos* test de Neyman.

También tenemos cierta unicidad de los test de máxima potencia.

Teorema 3.1.2 *Todos los test de máxima potencia son* test de Neyman.

Demostración. Sea un test de máxima potencia a nivel α con región crítica $\tilde{A}_1$. Sea A_1 la región crítica de un test de Neyman. Entonces sabemos, por la demostración anterior, que

$$\int_{\mathcal{X}} (\mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x))(L_1(x) - KL_0(x))dx \geq 0$$

pero al ser $\tilde{A}_1$ de máxima potencia, la desigualdad no puede ser estricta ya que llegaríamos a que

$$P_{\theta_1}(A_1) - P_{\theta_1}(\tilde{A}_1) > 0.$$

Por tanto se debe verificar que

$$\int_{\mathcal{X}} (\mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x))(L_1(x) - KL_0(x))dx = 0$$

pero el integrando es no negativo (lo hemos visto en la demostración anterior) y por tanto, salvo conjuntos de medida de Lebesgue cero (longitud cero)

$$(\mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x))(L_1(x) - KL_0(x)) = 0,$$

de manera que si $L_1(x) - KL_0(x) > 0, \mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x) = 0$ y como entonces $x \in A_1$ x también está en $\tilde{A}_1$. Así que

$$\tilde{A}_1 \supseteq \{x, L_1(x) - KL_0(x) > 0\}.$$

Análogamente si $L_1(x) - KL_0(x) < 0$, $\mathbf{1}_{A_1}(x) - \mathbf{1}_{\tilde{A}_1}(x) = 0$ y como entonces $x \notin A_1$ (es decir $x \in A_0$) x tampoco está en $\tilde{A}_1$ (es decir $x \in \tilde{A}_0$), así

$$\tilde{A}_0 \supseteq \{x, L_1(x) - KL_0(x) < 0\}$$

■

Ejemplo 3.1.1 *Supongamos un modelo de observaciones iid con distribución $N(\mu, \sigma^2)$ donde σ es conocida y μ puede valer μ_0 ó μ_1 ($\mu_1 > \mu_0$). Queremos test óptimos para contrastar*

$$\begin{aligned} H_0 &: \mu = \mu_0 \textit{ versus} \\ H_1 &: \mu = \mu_1. \end{aligned}$$

Tenemos que

$$\begin{aligned} L_1(x) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_1)^2\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2\} \exp\{-\frac{n}{2\sigma^2}(\bar{x} - \mu_1)^2\} \\ L_0(x) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} . \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2\} \exp\{-\frac{n}{2\sigma^2}(\bar{x} - \mu_0)^2\} \end{aligned}$$

Entonces, como

$$\begin{aligned} L_1(x) > KL_0(x) &\iff \\ \frac{\exp\{-\frac{n}{2\sigma^2}(\bar{x} - \mu_1)^2\}}{\exp\{-\frac{n}{2\sigma^2}(\bar{x} - \mu_0)^2\}} > K &\iff (\bar{x} - \mu_0)^2 - (\bar{x} - \mu_1)^2 > Cte \\ (\mu_1 - \mu_0)\bar{x} > Cte &\iff \bar{x} > C \end{aligned}$$

Una región crítica óptima vendrá dada por

$$A_1 = \{x, \bar{x} > C\}.$$

Ahora la constante C se puede determinar de manera que

$$P_{\mu_0}\{\bar{X} > C\} = \alpha.$$

Como, bajo μ_0,

$$\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$$

$$\begin{aligned} P_{\mu_0}\{\bar{X} &> C\} = P_{\mu_0}\{\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > \frac{\sqrt{n}(C - \mu_0)}{\sigma}\} \\ &= P_{\mu_0}\{N(0,1) > \frac{\sqrt{n}(C - \mu_0)}{\sigma}\} = \alpha \end{aligned}$$

Y utilizando la notación habitual

$$\frac{\sqrt{n}(C - \mu_0)}{\sigma} = k_\alpha,$$

con lo que finalmente

$$C = \mu_0 + k_\alpha \frac{\sigma}{\sqrt{n}},$$

y

$$A_1 = \{x, \bar{x} > \mu_0 + k_\alpha \frac{\sigma}{\sqrt{n}}\}$$

Podemos ahora calcular la potencia del test

$$\begin{aligned} \beta &= P_{\mu_1}(A_1) = P_{\mu_1}\{\bar{X} > \mu_0 + k_\alpha \frac{\sigma}{\sqrt{n}}\} \\ &= P_{\mu_1}\{\frac{\sqrt{n}(\bar{X} - \mu_1)}{\sigma} > k_\alpha + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}\} \\ &= P_{\mu_1}\{N(0,1) > k_\alpha + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}\} \\ &= 1 - \Phi(k_\alpha + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}) \end{aligned}$$

donde Φ es la función de distribución de una normal estándar. Notemos que

$$\lim_{n \to \infty} \beta = 1,$$

cuando esto ocurre se dice que el test en cuestión es consistente.

Observación 3.1.2 *Notar que si hacemos el contraste anterior pero con $\mu_1 < \mu_0$ entonces la región crítica óptima de nivel α es*

$$A_1 = \{x, \bar{x} < \mu_0 - k_\alpha \frac{\sigma}{\sqrt{n}}\}$$

Observación 3.1.3 *Una vez construido un test para ver si aceptamos o rechazamos H_0 se toma una muestra x y se mira si está en la región de aceptación o en la región crítica correspondientes a un nivel de significación prefijado α. Otra manera de indicar que tan* significativa *es la muestra observada para rechazar la hipótesis nula es dar su p-valor, consiste en el nivel de significación del test a partir del cual la muestra observada estaría en la región crítica. Si el p-valor es más pequeño que* 0.05 *nuestra muestra estaría en la región crítica de nivel* 0.05. *Esto es, cuánto más pequeño es su p-valor menos se valida la hipótesis nula.*

Ejemplo 3.1.2 *Sea un modelo de observaciones iid con densidad $U(0,\theta)$ donde θ puede ser θ_0 ó θ_1. $\theta_1 > \theta_0$. Queremos construir test óptimos para contrastar*

$$\begin{aligned} H_0 &: \theta = \theta_0 \textit{ versus} \\ H_1 &: \theta = \theta_1. \end{aligned}$$

$$\begin{aligned} L_1(x) &= \frac{1}{\theta_1^n}\mathbf{1}_{[0,\theta_1]}(x_{(n)}) \\ L_0(x) &= \frac{1}{\theta_0^n}\mathbf{1}_{[0,\theta_0]}(x_{(n)}) \end{aligned}$$

de manera que los test con región crítica

$$A_1 \supseteq \{x, \frac{1}{\theta_1^n}\mathbf{1}_{[0,\theta_1]}(x_{(n)}) > \frac{K}{\theta_0^n}\mathbf{1}_{[0,\theta_0]}(x_{(n)})\} \tag{3.2}$$

$$A_0 \supseteq \{x, \frac{1}{\theta_1^n}\mathbf{1}_{[0,\theta_1]}(x_{(n)}) < \frac{K}{\theta_0^n}\mathbf{1}_{[0,\theta_0]}(x_{(n)})\}, \tag{3.3}$$

son óptimos. En particular si tomamos

$$A_1 = \{x, x_{(n)} > C\}$$

con $C \leq \theta_0$, será un test óptimo. En efecto, basta tomar $K = \frac{\theta_0^n}{\theta_1^n}$, entonces si si $x_{(n)} \leq C$ resultará que

$$\mathbf{1}_{[0,\theta_1]}(x_{(n)}) = \mathbf{1}_{[0,\theta_0]}\left(x_{(n)}\right),$$

con lo que

$$A_0 \subseteq \{x, \frac{1}{\theta_1^n}\mathbf{1}_{[0,\theta_1]}(x_{(n)}) = \frac{K}{\theta_0^n}\mathbf{1}_{[0,\theta_0]}(x_{(n)})\},$$

lo que implica que

$$A_1 \supseteq \{x, \frac{1}{\theta_1^n}\mathbf{1}_{[0,\theta_1]}(x_{(n)}) > \frac{K}{\theta_0^n}\mathbf{1}_{[0,\theta_0]}(x_{(n)})\}.$$

Podemos bucar ahora C para un nivel de significación α. Se debera cumplir que

$$P_{\theta_0}(X_{(n)} > C) = \alpha,$$

ahora bien

$$\begin{aligned}P_{\theta_0}\{X_{(n)} > C\} &= 1 - P_{\theta_0}\{X_{(n)} \leq C\} \\ &= 1 - \Pi_{i=1}^n P_{\theta_0}\{X_i \leq C\} \\ &= 1 - \left(\frac{C}{\theta_0}\right)^n = \alpha\end{aligned}$$

con lo que

$$C = \theta_0 \sqrt[n]{1-\alpha}.$$

Podemos ahora calcular la potencia del test

$$\begin{aligned}\beta &= P_{\theta_1}(X_{(n)} > C) = 1 - P_{\theta_1}\{X_{(n)} \leq C\} \\ &= 1 - \Pi_{i=1}^n P_{\theta_1}\{X_i \leq C\} = 1 - \left(\frac{C}{\theta_1}\right)^n \\ &= 1 - (1-\alpha)\left(\frac{\theta_0}{\theta_1}\right)^n\end{aligned}$$

Notemos que

$$\lim_{n\to\infty} \beta = 1,$$

con lo que el test es consistente.

3.2 Hipótesis compuestas

En el caso de hipótesis compuestas tenemos probabilidades de error de primera especie, en plural:

$$P_\theta(A_1), \theta \in \Theta_0$$

y lo mismo para el error de segunda especie

$$P_\theta(A_1), \theta \in \Theta_1.$$

Definición 3.2.1 *Llamaremos* tamaño *del test (con región crítica A_1) al valor*

$$\sup_{\theta\in\Theta_0} P_\theta(A_1)$$

Definición 3.2.2 *Diremos que un test es de* nivel (de significación) α *si*

$$\sup_{\theta\in\Theta_0} P_\theta(A_1) \leq \alpha$$

Definición 3.2.3 *Llamaremos* función de potencia *del test a*

$$\beta(\theta) = P_\theta(A_1), \theta \in \Theta_1$$

Siguiendo el enfoque de Neyman, buscaremos entre los test de nivel α prefijado los que son *uniformemente más potentes* (UMP), es decir que para todo valor de $\theta \in \Theta_1$ su potencia es superior.

Desafortunadamente no tenemos un resultado general que nos permita construir test UMP, pero veremos que en muchos casos los test de Neyman nos permiten construir test UMP.

Proposición 3.2.1 *Fijemos un nivel $\alpha \in [0,1]$, y sea A_1 un test de Neyman de nivel α para contrastar*

$$\begin{aligned} H_0 &: \theta = \theta_0 \textit{ versus} \\ H_1 &: \theta = \theta_1. \end{aligned}$$

y que A_1 no cambia al variar $\theta_1 \in \Theta_1$. *Supongamos también que* $\sup_{\theta \in \Theta_0} P_\theta(A_1) = P_{\theta_0}(A_1)$ *(donde $\theta_0 \in \Theta_0$). Entonces el test con región crítica A_1 es UMP al nivel α para contrastar*

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \textit{ versus} \\ H_1 &: \theta \in \Theta_1. \end{aligned}$$

Demostración. Sea A_1 una región crítica en el contraste

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \text{ versus} \\ H_1 &: \theta \in \Theta_1. \end{aligned}$$

y que tiene las propiedades anteriores.

Supongamos que NO es UMP a nivel α, esto implicará que existe algún valor de la alternativa, pongamos θ_1, tal que su potencia es superada por otra región crítica (otro test), pongamos $\tilde{A}_1$, de nivel α. Si restringimos ambos test a constrastar

$$\begin{aligned} H_0 &: \theta = \theta_0 \text{ versus} \\ H_1 &: \theta = \theta_1. \end{aligned}$$

resultará que son de nivel α y que $\beta(\tilde{A}_1) > \beta(A_1)$ pero A_1 es un test de Neyman por tanto llegamos a una contradicción. ■

Ejemplo 3.2.1 *Consideremos un modelo de observaciones iid con distribución $N(\mu, \sigma^2)$ donde σ es conocida. Queremos contrastar*

$$\begin{aligned} H_0 &: \mu \le \mu_0 \\ H_1 &: \mu > \mu_0. \end{aligned}$$

Sabemos por el ejemplo (3.1.1) que un test de Neyman de nivel α para contrastar

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu = \mu_1. \end{aligned}$$

con $\mu_1 > \mu_0$ *viene dado por la región crítica*

$$A_1 = \{x, \bar{x} > \mu_0 + k_\alpha \frac{\sigma}{\sqrt{n}}\}$$

y esta región crítica No depende de μ_1*. Por otro lado si tomamos* $\mu \leq \mu_0$ *tenemos*

$$\begin{aligned} P_\mu(A_1) &= P_\mu\{\bar{X} > \mu_0 + k_\alpha \frac{\sigma}{\sqrt{n}}\} \\ &= P_\mu\{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > \frac{\sqrt{n}}{\sigma}(\mu_0 - \mu) + k_\alpha\} \\ &= P_\mu\{N(0,1) > \frac{\sqrt{n}}{\sigma}(\mu_0 - \mu) + k_\alpha\} \\ &\leq P_\mu\{N(0,1) > k_\alpha\} \text{ (ya que } \mu_0 - \mu > 0\text{)} \\ &= \alpha. \end{aligned}$$

Con lo que $\sup_{\mu \leq \mu_0} P_\mu(A_1) = P_{\mu_0}(A_1)$*. Por tanto* A_1 *es una región crítica (de un test) UMP.*

Observación 3.2.1 *Notemos que No podemos utilizar el procedimiento anterior para construir un test UMP para el contraste*

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0, \end{aligned}$$

con observaciones iid $N(\mu, \sigma^2)$ *donde* σ *es conocida. De hecho podemos deducir que no existe test UMP, ya que el test UMP, a nivel* α*, para contrastar*

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0, \end{aligned}$$

tiene región crítica

$$A_1 = \{x, \bar{x} > \mu_0 + k_\alpha \frac{\sigma}{\sqrt{n}}\}$$

y el test UMP para contrastar

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu < \mu_0, \end{aligned}$$

tiene región crítica

$$\tilde{A}_1 = \{x, \bar{x} < \mu_0 - k_\alpha \frac{\sigma}{\sqrt{n}}\}.$$

No coinciden, de manera que si quisiéramos máxima potencia "por la derecha" no la tendríamos por la izquierda (de la unicidad del test de máxima potencia en simple contra simple) y viceversa.

A priori no sabemos cómo construir una región crítica en le caso de hipótesis compuestas. En algunos casos la metodología de intervalos de confianza nos permite construir test de hipótesis.

3.2.1 Construcción de test a partir de intervalos de confianza

Supongamos que tenemos que contrastar

$$\begin{aligned} H_0 &: g(\theta) = \lambda_0 \text{ versus} \\ H_1 &: g(\theta) \neq \lambda_0. \end{aligned}$$

Notemos que en este tipo de contrastes tanto la hipótesis nula como la alternativa pueden ser compuestas. Sean $I(x)$, $x \in \mathcal{X}$ intervalos de confianza γ para $g(\theta)$. Definamos la región crítica

$$A_1 = \{x, \lambda_0 \notin I(x)\}$$

entonces se trata de una región crítica de nivel $\alpha = 1 - \gamma$. En efecto, por construcción

$$P_\theta\{x, g(\theta) \in I(x)\} \geq \gamma, \forall \theta \in \Theta$$

Esto es

$$P_\theta\{x, g(\theta) \notin I(x)\} \leq 1 - \gamma = \alpha, \forall \theta \in \Theta$$

en particular

$$P_\theta\{x, g(\theta) \notin I(x)\} \leq \alpha, \forall \theta \text{ tal que } g(\theta) = \lambda_0$$

Por tanto

$$P_\theta(A_1) = P_\theta\{x, \lambda_0 \notin I(x)\} \leq \alpha, \forall \theta \text{ tal que } g(\theta) = \lambda_0$$

y al región crítica A_1 es de nivel $\alpha = 1 - \gamma$. Notemos que en principio no podemos concluir nada acerca de la potencia del test correspondiente, necesitaríamos saber

$$P_\theta\{x, \lambda_0 \notin I(x)\} \text{ para } g(\theta) \neq \lambda_0.$$

Ejemplo 3.2.2 *Supongamos un modelo n de observaciones iid $N(\mu, \sigma^2)$, σ conocida. Queremos contrastar*

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0, \end{aligned}$$

En este caso la hipótesis nula es simple. Ya vimos que los intervalos de confianza γ para μ, venían dados por

$$\bar{x} - k_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + k_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

De manera que una crítica de nivel α vendrá dada por

$$A_1 = \{x, \mu_0 < \bar{x} - k_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ ó } \mu_0 > \bar{x} + k_{\alpha/2} \frac{\sigma}{\sqrt{n}}\}.$$

En este caso podemos calcular fácilmente la función de potencia $\beta(\mu)$.

$$\begin{aligned}
P_\mu(A_1) &= P_\mu(\bar{X} - k_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu_0) + P_\mu(\bar{X} + k_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu_0) \\
&= P_\mu\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} > \frac{\sqrt{n}}{\sigma}(\mu_0-\mu) + k_{\alpha/2}\right) \\
&\quad + P_\mu\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} < \frac{\sqrt{n}}{\sigma}(\mu_0-\mu) - k_{\alpha/2}\right) \\
&= P_\mu\left(N(0,1) > \frac{\sqrt{n}}{\sigma}(\mu_0-\mu) + k_{\alpha/2}\right) \\
&\quad + P_\mu\left(N(0,1) < \frac{\sqrt{n}}{\sigma}(\mu_0-\mu) - k_{\alpha/2}\right) \\
&= 1 - \Phi(\frac{\sqrt{n}}{\sigma}(\mu_0-\mu) + k_{\alpha/2}) + \Phi(\frac{\sqrt{n}}{\sigma}(\mu_0-\mu) - k_{\alpha/2})
\end{aligned}$$

Ejemplo 3.2.3 *Supongamos un modelo n de observaciones iid $N(\mu,\sigma^2)$, σ desconocida. Queremos contrastar*

$$\begin{aligned}
H_0 &: \mu = \mu_0 \\
H_1 &: \mu \neq \mu_0,
\end{aligned}$$

En este caso la hipótesis nula es compuesta. Ya vimos que los intervalos de confianza γ para μ, venían dados por

$$\bar{x} - t_{n-1;\alpha/2}\frac{\tilde{s}(x)}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1;\alpha/2}\frac{\tilde{s}(x)}{\sqrt{n}}.$$

De manera que una crítica de nivel α vendrá dada por

$$A_1 = \{x, \mu_0 < \bar{x} - t_{n-1;\alpha/2}\frac{\tilde{s}(x)}{\sqrt{n}} \text{ ó } \mu_0 > \bar{x} + t_{n-1;\alpha/2}\frac{\tilde{s}(x)}{\sqrt{n}}\}.$$

Si intentamos calcular la función de potencia $\beta(\mu)$ sólo obtenemos in valor aproximado:

$$\begin{aligned}
P_{\mu,\sigma}(A_1) &= P_{\mu,\sigma}(\bar{X} - t_{n-1;\alpha/2}\frac{\tilde{s}}{\sqrt{n}} > \mu_0) + P_{\mu,\sigma}(\bar{X} + t_{n-1;\alpha/2}\frac{\tilde{s}}{\sqrt{n}} < \mu_0) \\
&= P_{\mu,\sigma}\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\tilde{s}} > \frac{\sqrt{n}}{\tilde{s}}(\mu_0-\mu) + t_{n-1;\alpha/2}\right) \\
&\quad + P_{\mu,\sigma}\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\tilde{s}} < \frac{\sqrt{n}}{\tilde{s}}(\mu_0-\mu) - t_{n-1;\alpha/2}\right) \\
&= P_{\mu,\sigma}\left(t_{n-1} > \frac{\sqrt{n}}{\tilde{s}}(\mu_0-\mu) + t_{n-1;\alpha/2}\right) \\
&\quad + P_{\mu,\sigma}\left(t_{n-1} < \frac{\sqrt{n}}{\tilde{s}}(\mu_0-\mu) - t_{n-1;\alpha/2}\right) \\
&\approx 1 - F_{t_{n-1}}(\frac{\sqrt{n}}{\tilde{s}}(\mu_0-\mu) + t_{n-1;\alpha/2}) + F_{t_{n-1}}(\frac{\sqrt{n}}{\tilde{s}}(\mu_0-\mu) - t_{n-1;\alpha/2}),
\end{aligned}$$

donde $F_{t_{n-1}}$ es la función de distribución de una t_{n-1} de Student.

Observación 3.2.2 *Podríamos utilizar los intervalos que vimos en los problemas de dos muestras para obtener los test de hipótesis correspondientes.*

Si las hipótesis no tienen la forma anterior no podemos utilizar los intervalos de confianza para obtener test. Sin embargo vamos a ver que existe un procedimiento general para construir test cuando las hipótesis son compuestas.

3.2.2 Test de la razón de verosimilitudes.

Supongamos el problema de contrastar

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \text{ versus} \\ H_1 &: \theta \in \Theta_1. \end{aligned}$$

el *test de la razón de verosimilitudes* (abreviaremos TRV) es el que tiene por región crítica

$$A_1 = \{x, \lambda(x) \leq K\}$$

donde

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(x;\theta)}{\sup_{\theta \in \Theta} L(x;\theta)},$$

y $0 < K < 1$. Notemos que $\lambda(x) \leq 1$ para todo x y que si $\hat{\theta}$ es el estimador máximo verosímil también podemos escribir

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(x;\theta)}{L(x;\hat{\theta}(x))},$$

incluso si $\tilde{\theta}$ es el estimador máximo verosímil *con la restricción* de que $\theta \in \Theta_0$ podemos escribir más simplemente

$$\lambda(x) = \frac{L(x;\tilde{\theta}(x))}{L(x;\hat{\theta}(x))},$$

Proposición 3.2.2 *Si las hipótesis son simples el test de la razón de verosimilitudes es un test de Neyman.*

Demostración.

$$\lambda(x) = \frac{L(x;\theta_0))}{L(x;\theta_0) \vee L(x;\theta_1)} = \frac{L_0(x)}{L_0(x) \vee L_1(x)},$$

de manera que

$$\begin{aligned} \lambda(x) &\leq K \Longleftrightarrow \frac{L_0(x)}{L_0(x) \vee L_1(x)} \leq K \\ L_0(x) &\leq K(L_0(x) \vee L_1(x)) \Longleftrightarrow L_0(x) \leq KL_1(x)), \end{aligned}$$

ya que $K < 1$. En efecto si x es tal que $L_0(x) \leq L_1(x)$ la equivalencia es inmediata y si x es tal que $L_0(x) > L_1(x)$ no se puede cumplir ni la condición de la izquierda ni la de la derecha. Por tanto

$$A_1 = \{x, L_1(x)) \geq \frac{1}{K} L_0(x)\}$$

con lo que se trata de un test de Neyman. ■

Ejemplo 3.2.4 *Consideremos un modelo de observaciones iid con distribución* $N(\mu, \sigma^2)$ *donde* σ *es conocida. Queremos contrastar*

$$\begin{aligned} H_0 &: \mu \leq \mu_0 \\ H_1 &: \mu > \mu_0. \end{aligned}$$

como en el ejemplo (3.2.1). Vamos a construir el test de la razón de verosimilitudes.

$$\begin{aligned} L(x;\mu) &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\} \\ &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2\} \exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\} \end{aligned}$$

de manera que el denominador del TRV

$$L(x;\hat{\mu}) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2\}$$

y el numerador

$$\sup_{\mu \leq \mu_0} L(x;\mu) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2\} \sup_{\mu \leq \mu_0} \exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\}.$$

de manera que

$$\lambda(x) = \sup_{\mu \leq \mu_0} \exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\}.$$

Ahora bien

$$\sup_{\mu \leq \mu_0} \exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\} = \exp\{-\frac{n}{2\sigma^2}\inf_{\mu \leq \mu_0}(\bar{x}-\mu)^2\}$$

y

$$\inf_{\mu \leq \mu_0}(\bar{x}-\mu)^2 = \begin{cases} (\bar{x}-\mu_0)^2 & si\ \bar{x} \geq \mu_0 \\ 0 & si\ \bar{x} < \mu_0 \end{cases}$$

de manera que

$$\lambda(x) = \begin{cases} exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu_0))^2\} & si\ \bar{x} \geq \mu_0 \\ 1 & si\ \bar{x} < \mu_0 \end{cases}.$$

Así la región crítica vendrá dada por

$$\begin{aligned} A_1 &= \{x, \lambda(x) \leq K\} \\ &= \{x, \bar{x} \geq \mu_0 \ y \ exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2\} \leq K\} \\ &= \{x, \bar{x} \geq \mu_0 \ y \ \bar{x} \geq C\} \end{aligned}$$

donde C es una constante mayor que μ_0 ya que $K < 1$. de manera que finalmente

$$A_1 = \{x, \ \bar{x} \geq C\}.$$

Obtenemos las mismas regiones que en el ejemplo (3.2.1) donde ya vimos que eran óptimas.

Ejemplo 3.2.5 *Supongamos un modelo n de observaciones iid $N(\mu, \sigma^2)$, μ, σ desconocidas. Queremos contrastar*

$$\begin{aligned} H_0 &: \sigma = \sigma_0 \\ H_1 &: \sigma \neq \sigma_0, \end{aligned}$$

Vamos a construir el TRV.

$$\begin{aligned} L(x;\mu,\sigma) &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\} \\ &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2\} \exp\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\} \end{aligned}$$

El numerador será

$$\begin{aligned} \sup_\mu L(x;\mu,\sigma_0) &= \frac{1}{\left(\sqrt{2\pi}\sigma_0\right)^n} \exp\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i-\bar{x})^2\} \\ &= \frac{1}{\left(\sqrt{2\pi}\sigma_0\right)^n} \exp\{-\frac{n\hat{\sigma}^2}{2\sigma_0^2}\} \end{aligned}$$

y el denominador

$$\begin{aligned} L(x;\hat{\mu},\hat{\sigma}) &= \frac{1}{\left(\sqrt{2\pi}\hat{\sigma}\right)^n} \exp\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (x_i-\bar{x})^2\} \\ &= \frac{1}{\left(\sqrt{2\pi}\hat{\sigma}\right)^n} \exp\{-\frac{n}{2}\} \end{aligned}$$

de manera que

$$\lambda(x) = \left(\frac{\hat{\sigma}}{\sigma_0}\right)^n \exp\{-\frac{n\hat{\sigma}^2}{2\sigma_0^2} + \frac{n}{2}\}$$

y $\lambda(x) > K$ (región de aceptación) equivale a que

$$\frac{\hat{\sigma}^2}{\sigma_0^2} \exp\{-\frac{\hat{\sigma}^2}{\sigma_0^2}\} \geq C$$

y resolviendo la inecuación tendremos

$$c_1 < \frac{\hat{\sigma}^2}{\sigma_0^2} < c_2$$

donde c_1 *y* c_2 *son las soluciones de la ecuación*

$$u \exp\{-u\} = C.$$

Notemos que obtendríamos lo mismo a partir de intervalos de confianza. Sin embargo aquí dado α, c_1 *y* c_2 *quedan fijados, en cambio los intervalos de coeficiente* $\gamma = 1 - \alpha$ *se pueden escoger de muchas maneras, de hecho ya vimos que el procedimiento mas simple era dejar* $\alpha/2$ *en cada "cola" y en la práctica esto es lo que se utiliza.*

Comportamiento asintótico del test de la razón de verosimilitudes

Consideremos los siguiente ejemplos

Ejemplo 3.2.6 *Sea un modelo de* n *observaciones iid* $N(\mu, \sigma^2)$, σ *conocida y queremos constrastar*

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0, \end{aligned}$$

$$\begin{aligned} \lambda(x) &= \frac{L(x;\mu_0)}{L(x;\hat{\mu})} \\ &= \frac{\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\{-\frac{1}{2\sigma^2}\sum(x_i - \bar{x})^2\} \exp\{-\frac{n}{2\sigma^2}(\bar{x} - \mu_0)^2\}}{\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\{-\frac{1}{2\sigma^2}\sum(x_i - \bar{x})^2\}} \\ &= \exp\{-\frac{n}{2\sigma^2}(\bar{x} - \mu_0)^2\} \end{aligned}$$

de manera que

$$W(x) := -2\log\lambda(x) = \left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right)^2$$

y consecuentemente bajo H_0

$$W(X) \sim (N(0,1))^2 = \mathcal{X}_1^2.$$

Ejemplo 3.2.7 *Sea un modelo asociado a variables iid*

$$X_1, X_2, ..., X_{n_1} \sim N(\mu_1, \sigma_1^2),$$

$$X_{n_1+1}, X_{n_1+2}, ..., X_{n_1+n_2} \sim N(\mu_2, \sigma_2^2), ...,$$

$$X_{n_1+n_2+...+n_{m-1}+1}, X_{n_1+n_2+...+n_{m-1}+2}, ..., X_{n_1+n_2+...+n_{m-1}+n_m} \sim N(\mu_m, \sigma_m^2)$$

es decir m muestras independientes de variables normales independientes con medias diferentes y varianzas conocidas. Supongamos el contraste

$$\begin{aligned} H_0 &: (\mu_1, \mu_2, ..., \mu_m) = (\mu_{01,}\mu_{02}, ..., \mu_{0m}) \\ H_1 &: (\mu_1, \mu_2, ..., \mu_m) \neq (\mu_{01,}\mu_{02}, ..., \mu_{0m}). \end{aligned}$$

Es fácil ver que

$$-2 \log \lambda(x) = \sum_{i=1}^{m} \left(\frac{\sqrt{n_i}(\bar{x}_i - \mu_{0i})}{\sigma_i} \right)^2 \sim \mathcal{X}_m^2(\textit{bajo } H_0).$$

Si ahora hacemos el contraste

$$\begin{aligned} H_0 &: (\mu_1, \mu_2) = (\mu_{01,}\mu_{02}) \\ H_1 &: (\mu_1, \mu_2) \neq (\mu_{01,}\mu_{02}). \end{aligned}$$

obtenemos

$$-2 \log \lambda(x) = \sum_{i=1}^{2} \left(\frac{\sqrt{n_i}(\bar{x}_i - \mu_{0i})}{\sigma_i} \right)^2 \sim \mathcal{X}_2^2(\textit{bajo } H_0)$$

notemos que los grados de libertad de la Ji cuadrado coinciden con la dimension del espacio total de parámetros con la dimensión del espacio de parámetros bajo la hipótesis nula:

$$\begin{aligned} \dim \Theta &= m \\ \dim \Theta_0 &= m - 2. \end{aligned}$$

Lo mismo ocurre en el caso anterior. De manera que la regla es:

$$-2 \log \lambda(x) \sim \mathcal{X}^2_{\dim \Theta - \dim \Theta_0}.$$

El resultado general que vamos a ver a continuación es que si el tamaño muestral es grande esto se cumple aproximadamente.

Proposición 3.2.3 *(Wilks) Supongamos el problema de contraste*

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \textit{ versus} \\ H_1 &: \theta \in \Theta_1. \end{aligned}$$

y que el modelo es regular (en el sentido del comportamiento del EMV). Sea $\dim \Theta = d$ *y* $\dim \Theta_0 = l$*. Vamos a suponer asimismo que cualquier punto* $\theta \in \Theta$ *se puede coordenar de la forma*

$$\theta = (\theta_1, \theta_2, ..., \theta_l, \theta_{l+1}, ..., \theta_d)$$

de manera que

$$\Theta_0 = \{\theta \in \Theta, \theta_{l+1} = k_{l+1}, .., \theta_d = k_d\}$$

para ciertas constantes k_i. Sea n el tamaño de la muestra iid x y $\lambda_n(x)$ la razón de verosimilitudes correspondiente entonces, bajo H_0,

$$-2\log\lambda_n(X) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{X}^2_{\dim\Theta-\dim\Theta_0}$$

Demostración. *Sea $\hat{\theta}$ el EMV de θ,*

$$\partial_{\theta_i}\log L(X;\theta) = -\sum_{j=1}^{d}\partial_{\theta_i\theta_j}\log L(X;\theta^*(X))(\hat{\theta}_j(X)-\theta_j), \quad i=1,...,d$$

donde $\theta^*(X) = \alpha\hat{\theta}(X)+(1-\alpha)\theta$, $0\le\alpha\le 1$, esto es

$$\begin{aligned}\sqrt{n}(\hat{\theta}_j(X)-\theta_j) &= \sum_{k=1}^{d}\left(-\frac{1}{n}\partial_{\theta\theta}\log L(X;\theta^*(X))\right)^{-1}_{jk}\frac{1}{\sqrt{n}}\partial_{\theta_k}\log L(X;\theta),\\ j &= 1,...,d\end{aligned}$$

que el modelo sea regular va a significar que $\hat{\theta}$ es un estimador consistente de θ y que

$$-\frac{1}{n}\partial_{\theta\theta}\log L(X;\theta^*(X)) \underset{n\to\infty}{\overset{P}{\to}} I(\theta)$$

donde

$$I(\theta) = E_\theta(-\partial^2_{\theta\theta}\log L_1)$$

matriz $d\times d$ que se conoce como matriz de información de Fisher. Como consecuencia y ya que, también debido a la regularidad,

$$\frac{1}{\sqrt{n}}\partial_\theta\log L(X;\theta) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} N_d(0,I(\theta))$$

tendremos que

$$\sqrt{n}(\hat{\theta}(X)-\theta) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} N_d(0,I^{-1}(\theta))$$

Escribamos

$$\xi = (\theta_1,\theta_2,...,\theta_l), \eta = (\theta_{l+1},...,\theta_d), \eta_0 = (k_{l+1},..,k_d)$$

de manerna que $\Theta_0 = \{(\xi,\eta), \eta=\eta_0\}$, y

$$I(\theta) = \begin{pmatrix} I_{\xi\xi} & I_{\xi\eta} \\ I_{\eta\xi} & I_{\eta\eta}\end{pmatrix},$$

vamos a suponer también, sin pérdida de generalidad, que la parametrización es ortogonal *de manera que*

$$I_{\xi\eta} = 0, I_{\eta\xi} = 0$$

y por tanto

$$I^{-1}(\theta) = \begin{pmatrix} I^{-1}_{\xi\xi} & 0 \\ 0 & I^{-1}_{\eta\eta}\end{pmatrix}.$$

Entonces, sea $\hat{\xi}_{\eta_0}(x) = (\tilde{\theta}_1(x), \tilde{\theta}_2(x), ..., \tilde{\theta}_l(x))$ *el EMV bajo* H_0*, tendremos*

$$\begin{aligned} \partial_{\theta_i} \log L(X; \theta_1, ..., \theta_l, \eta_0) &= -\sum_{j=1}^{l} \partial^2_{\theta_i \theta_j} \log L(X; \tilde{\theta}^*(X), \eta_0)(\tilde{\theta}_j(X) - \theta_j), \\ i &= 1, ..., l \end{aligned}$$

donde $\tilde{\theta}^*_j(X) = \alpha\theta_j + (1-\alpha)\tilde{\theta}_j(X), j = 1, ..., l,\ 0 \leq \alpha \leq 1,$ *esto es*

$$\begin{aligned} & \sqrt{n}(\tilde{\theta}_j(X) - \theta_j) \\ &= \sum_{k=1}^{l} \left(-\frac{1}{n} \partial^2_{\theta\theta} \log L(X; \tilde{\theta}^*(X), \eta_0) \right)^{-1}_{jk} \frac{1}{\sqrt{n}} \partial_{\theta_k} \log L(X; \theta_1, ..., \theta_l, \eta_0), \\ j &= 1, ..., l, \end{aligned}$$

por la regularidad del modelo, bajo H_0

$$\left(-\frac{1}{n} \partial^2_{\theta\theta} \log L(X; \tilde{\theta}^*(X), \eta_0) \right) \underset{n\to\infty}{\overset{P}{\to}} I_{\xi\xi}(\xi, \eta_0)$$

y

$$\sqrt{n}(\hat{\xi}_{\eta_0}(X) - \xi) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} N_l(0, I^{-1}_{\xi\xi}).$$

Ahora bien

$$\begin{aligned} \log L(X; \xi, \eta_0) &= \log L(X; \hat{\xi}(X), \hat{\eta}(X)) \\ &\quad + \frac{1}{2} \sum_{i,j=1}^{d} \partial^2_{\theta_i \theta_j} \log L(X; \theta^{**}(X))(\hat{\theta}_j(X) - \theta_j)(\hat{\theta}_i(X) - \theta_i) \\ \log L(X; \xi, \eta_0) &= \log L(X; \hat{\xi}_{\eta_0}(X), \eta_0) \\ &\quad + \frac{1}{2} \sum_{i,j=1}^{l} \partial^2_{\theta_i \theta_j} \log L(X; \tilde{\theta}^{**}(X))(\tilde{\theta}_j(X) - \theta_j)(\tilde{\theta}_i(X) - \theta_i) \end{aligned}$$

where θ^{**} *and* $\tilde{\theta}^{**}$ *son puntos intermedios, de aquí obtenemos restando las dos expresiones, que bajo* H_0

$$\begin{aligned} -2\log\lambda_n(X) &= -2(\log L(X; \hat{\xi}_{\eta_0}(X), \eta_0) - \log L(X; \hat{\xi}(X), \hat{\eta}(X))) \\ &= \sum_{i,j=1}^{d} -\partial^2_{\theta_i \theta_j} \log L(X; \theta^{**}(X))(\hat{\theta}_j(X) - \theta_j)(\hat{\theta}_i(X) - \theta_i) \\ &\quad - \sum_{i,j=1}^{l} -\partial^2_{\theta_i \theta_j} \log L(X; \tilde{\theta}^{**}(X))(\tilde{\theta}_j(X) - \theta_j)(\tilde{\theta}_i(X) - \theta_i) \\ &\quad \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \sum_{i,j=1}^{d} I_{ij} Z_i Z_j - \sum_{i,j=1}^{l} (I_{\xi\xi})_{ij} Z_i Z_j \\ &= \sum_{i,j=l+1}^{d} (I_{\eta\eta})_{ij} Z_i Z_j \end{aligned}$$

y donde $(Z_i)_{l+1\leq i,j\leq d} \sim N_d(0, I_{\eta\eta}^{-1}(\theta))$*, por tanto*

$$\sum_{i,j=l+1}^{d} (I_{\eta\eta})_{ij}\, Z_i Z_j \sim \mathcal{X}^2_{d-l}$$

■

Observación 3.2.3 *Notemos que del resultado anterior para calcular la región crítica del TRV lo haremos en función del estadístico de Wilks* $W(x) = -2\log\lambda(x)$*, así*

$$\begin{aligned} A_1 &= \{x, \lambda(x) \leq K\} \\ &= \{x, -2\log\lambda(x) \geq C\} \\ &= \{x, W(x) \geq C\}, \end{aligned}$$

entonces hallaremos C *con la condición* $\sup_{\theta\in\Theta_0} P_\theta(A_1) = \alpha$*. Pero*

$$\sup_{\theta\in\Theta_0} P_\theta(A_1) \approx P\{\mathcal{X}^2_{\dim\Theta-\dim\Theta_0} \geq C\} = \alpha,$$

de manera que se toma como región crítica

$$A_1 = \{x, W(x) \geq \mathcal{X}^2_{\dim\Theta-\dim\Theta_0;\alpha}\}$$

que tendrá aproximadamente nivel α*.*

Capítulo 4

Test Ji-cuadrado

4.1 El modelo multinomial

Sea X una variable aleatoria que puede tomar m valores distintos, $z_1, z_2, ..., z_m$ con probabilidades respectivas $p_1, p_2, ..., p_m$, tales que $p_i > 0$ y por supuesto $\sum_{i=1}^{m} p_i = 1$, diremos que X es una variable multinomial. Sea ahora el modelo correspondiente a n observaciones iid con la ley X, diremos que nuestro modelo es un *modelo multinomial*. La verosimilitud, dada una muestra x vendrá dada por

$$\begin{aligned}
& L(x; p_1, p_2, ..., p_m) \\
= \; & P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \\
= \; & \Pi_{i=1}^{n} P(X_i = x_i) = \Pi_{i=1}^{n} \Pi_{j=1}^{m} p_j^{\mathbf{1}_{\{z_j\}}(x_i)} = \Pi_{j=1}^{m} p_j^{N_j(x)}.
\end{aligned}$$

Notemos que $N_j(x) = \sum_{i=1}^{n} \mathbf{1}_{\{z_j\}}(x_i)$ cuenta el número de veces que ha ocurrido el resultado z_j en los n experimentos. Es fácil ver que

$$P(N_1 = n_1, ..., N_m = n_m) = \frac{n!}{n_1! n_2! ... n_m!} p_1^{n_1} p_2^{n_2} ... p_m^{n_m},$$

esta distribución se conoce como *distribución multinomial.* Observemos asimismo que el espacio de parámetros del modelo multinomial

$$\Theta = \{(p_1, p_2, ..., p_m), p_i > 0, \sum_{i=1}^{n} p_i = 1\}$$

de manera que su dimensión es $m - 1$.

4.2 Test de ajuste en el modelo multinomial

4.2.1 Ajuste a una mutinomial concreta

En el modelo multinomial queremos hacer el contraste

$$\begin{aligned} H_0 &: (p_1, p_2, ..., p_m) = (p_{01,}p_{02}, ..., p_{0m}) \\ H_1 &: (p_1, p_2, ..., p_m) \neq (p_{01,}p_{02}, ..., p_{0m}). \end{aligned}$$

Vamos a construir el TRV,

$$\lambda(x) = \frac{L(x; p_0)}{\sup L(x, p)} = \frac{L(x; p_0)}{L(x, \hat{p})}$$

donde utilizamos notación vectorial: $p = (p_{1,}p_2, ..., p_m)$. Tenemos que

$$L(x; p) = \Pi_{i=1}^m p_i^{N_i(x)} = \Pi_{i=1}^m p_i^{n_i}$$

de manera que, tomando como variables libres, $p_{1,}p_2, ..., p_{m-1}$ y como $p_m = 1 - p_{1,} - p_2 ... - p_{m-1}$, las condiciones de extremo son

$$\partial_{p_i} \log L(x; p) = \frac{n_i}{p_i} - \frac{n_m}{p_m} = 0, i = 1, ..., m$$

esto da

$$\hat{p}_i = \frac{n_i}{n}, i = 1, ..., m.$$

Como, por la ley fuerte de los grandes números,

$$\frac{n_i}{n} \stackrel{c.s.}{\to} p_i > 0$$

podemos suponer que todos los n_i son positivos, entonces $L(x; p) = \Pi_{i=1}^m p_i^{n_i}$ será cero si algún p_i es cero con lo que el máximo estará en el interior de Θ y corresponderá a la única solución que hemos encontrado. Tenemos entonces que

$$\lambda(x) = \frac{\Pi_{i=1}^m p_{0i}^{n_i}}{\Pi_{i=1}^m \hat{p}_i^{n_i}} = \Pi_{i=1}^m \left(\frac{p_{0i}}{\hat{p}_i}\right)^{n_i} = \Pi_{i=1}^m \left(\frac{p_{0i}}{\hat{p}_i}\right)^{n\hat{p}_i}$$

y, bajo H_0

$$W(x) = -2 \log \lambda(x) = -2n \sum_{i=1}^m \hat{p}_i \log \frac{p_{0i}}{\hat{p}_i} \sim \mathcal{X}_{m-1}^2 (\text{aprox.})$$

ya que $\dim \Theta = m - 1$ y $\dim \Theta_0 = 0$. La región crítica de nivel α será

$$A_1 = \{x, W(x) \geq \mathcal{X}_{m-1;\alpha}^2\}$$

Sin embargo en la práctica se utiliza otro estadístico para construir la región critica. Tenemos que

$$\begin{aligned} W(x) &= -2n\sum_{i=1}^{m}\hat{p}_i \log\frac{p_{0i}}{\hat{p}_i} \\ &= 2n\sum_{i=1}^{m}p_{0i}\frac{\hat{p}_i}{p_{0i}}\log\frac{\hat{p}_i}{p_{0i}}, \end{aligned}$$

sabemos que $\frac{\hat{p}_i}{p_{0i}} \underset{n\to\infty}{\overset{c.s}{\to}} 1$ y si desarrollamos la función $u\log u$ en torno del valor $u=1$, tenemos

$$u\log u = u - 1 + \frac{1}{2}(u-1)^2 - \frac{1}{6}\frac{1}{u^{*2}}(u-1)^3$$

con u^* un punto entre 1 y u. Tenemos así que

$$\begin{aligned} &= 2n\sum_{i=1}^{m}p_{0i}\left(\frac{\hat{p}_i}{p_{0i}}-1\right) + n\sum_{i=1}^{m}p_{0i}\left(\frac{\hat{p}_i}{p_{0i}}-1\right)^2 \\ &\quad -\frac{n}{3}\sum_{i=1}^{m}p_{0i}\frac{1}{u_i^{*2}}\left(\frac{\hat{p}_i}{p_{0i}}-1\right)^3, \end{aligned}$$

donde los u_i^* son puntos entre 1 y $\frac{\hat{p}_i}{p_{0i}}$. El primer sumando se anula

$$\sum_{i=1}^{m}p_{0i}\left(\frac{\hat{p}_i}{p_{0i}}-1\right) = \sum_{i=1}^{m}(\hat{p}_i - p_{0i}) = \sum_{i=1}^{m}\hat{p}_i - \sum_{i=1}^{m}p_{0i} = 1-1=0,$$

el tercero se va a cero cuando n tiende a infinito:

$$n\sum_{i=1}^{m}p_{0i}\frac{1}{u_i^{*2}}\left(\frac{\hat{p}_i}{p_{0i}}-1\right)^3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{m}\frac{(1-p_{0i})^{3/2}}{u_i^{*2}p_{0i}^{1/2}}\left(\frac{\sqrt{n}(\hat{p}_i-p_{0i})}{\sqrt{p_{0i}(1-p_{0i})}}\right)^3 \underset{n\to\infty}{\overset{P}{\to}} 0$$

ya que

$$u_i^* \overset{c.s.}{\to} 1 \text{ y } \frac{\sqrt{n}(\hat{p}_i-p_{0i})}{\sqrt{p_{0i}(1-p_{0i})}} \overset{\mathcal{L}}{\to} N(0,1).$$

En definitiva $W(x)$ y

$$D_n(x) := n\sum_{i=1}^{m}p_{0i}\left(\frac{\hat{p}_i}{p_{0i}}-1\right)^2$$

tienen el mismo comportamiento asintótico, se acercan a una $\mathcal{X}^2_{m-1}$. Este último estadístico se llama estadístico de Pearson, que se puede escribir

$$\begin{aligned} D_n(x) &= n\sum_{i=1}^{m} p_{0i}\left(\frac{\hat{p}_i}{p_{0i}}-1\right)^2 \\ &= n\sum_{i=1}^{m}\frac{(\hat{p}_i-p_{0i})^2}{p_{0i}} \\ &= \sum_{i=1}^{m}\frac{(n\hat{p}_i-np_{0i})^2}{np_{0i}} \\ &= \sum_{i=1}^{m}\frac{(n_i-np_{0i})^2}{np_{0i}} \\ &= \sum_{i=1}^{m}\frac{(O_i-E_i)^2}{E_i}, \end{aligned}$$

donde O_i es la frecuencia observada y E_i la esperada bajo H_0. Entonces la región crítica de nivel α (aproximado) será

$$A_1 = \{x, D_n(x) \geq \mathcal{X}_{m-1;\alpha}\}.$$

La aproximación funciona bien si $np_{0i} \geq 5$ para todo i.

Ejemplo 4.2.1 ***Ejemplo 4.2.2*** *Se lanza un dado 2000 veces con los siguientes resultados:*

1	*2*	*3*	*4*	*5*	*6*
388	*322*	*314*	*316*	*344*	*316*

Se puede pensar que el dado está equilibrado? Es decir queremos hacer el contraste:

$$\begin{aligned} H_0 &: p = (\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}) \\ H_1 &: p \neq (\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}) \end{aligned}$$

Calculamos

$$D_n(x) = n\sum_{i=1}^{6}\frac{(\hat{p}_i-\frac{1}{6})^2}{\frac{1}{6}} \tag{4.1}$$

donde

$$\begin{aligned} \hat{p}_1 &= \frac{388}{2000} = 0.194,\ \hat{p}_2 = \frac{322}{2000} = 0.161,\ \hat{p}_3 = \frac{314}{2000} = 0.157, \\ \hat{p}_4 &= \frac{316}{2000} = 0.158,\ \hat{p}_5 = \frac{344}{2000} = 0.172,\ \hat{p}_6 = \frac{316}{2000} = 0.158. \end{aligned}$$

Esto da un valor

$$D_n(x) = 2000 \times \sum_{i=1}^{6}(\hat{p}_i - \frac{1}{6})^2 \times 6 = 12.6. \tag{4.2}$$

La aproximación asintótica funciona bien ya que $np_{0i} = \frac{2000}{6} \gg 5$. *Si miramos las tablas de una* χ^2_5 *obtenemos que* $\chi^2_{5;0.05} = 11.07$. *Por tanto* $D_n(x) > \chi^2_{5;0.05}$ *y rechazamos* H_0.*Los datos habían sido simulados con una distribución teórica* $p_{01} = 0.2$, $p_{02} = p_{03} = p_{04} = p_{05} = p_{06} = 0.16$. *Notemos que* $1/6 = 0.167$.

4.2.2 Ajuste a una familia de mutinomiales

En el modelo multinomial anterior podemos contrastar si nuestra multinomial pertenece a una familia de mutinomiales. Es decir podemos contrastar las hipótesis

$$\begin{aligned} H_0 &: (p_1, p_2, ..., p_m) = (p_1(\delta), p_2(\delta), ..., p_m(\delta)) \\ H_1 &: (p_1, p_2, ..., p_m) \neq (p_1(\delta), p_2(\delta), ..., p_m(\delta)). \end{aligned}$$

donde $\delta \in \Delta \subseteq \mathbb{R}^l, l < m-1$ donde $p(\Delta)$ es una subvariedad de dimensión l (de manera que podemos coordenar todos los valores de p con los valores de δ y $m-1-l$ coordenadas adicionales), sabemos que en tal caso y bajo H_0 el estadístico de Wilks $W(x) = -2\log\lambda(x)$ tiene una distribución aproximada (si n es grande) $\mathcal{X}^2_{m-l-1}$. Tenemos que

$$\lambda(x) = \frac{\sup_{\delta\in\Delta} \Pi_{i=1}^m p_i(\delta)^{n_i}}{\sup_{p\in\Theta} \Pi_{i=1}^m p_i^{n_i}} = \frac{\Pi_{i=1}^m p_i(\hat{\delta})^{n_i}}{\Pi_{i=1}^m \hat{p}_i^{n_i}} = \Pi_{i=1}^m \left(\frac{p_i(\hat{\delta})}{\hat{p}_i}\right)^{n\hat{p}_i}$$

con $\hat{p}_i = \frac{n_i}{n}$. Haciendo desarrollos de Taylor análogos a los anteriores, se puede ver que

$$W(x) = -2\log\lambda(x) = n\sum_{i=1}^m \frac{\left(\hat{p}_i - np_i(\hat{\delta})\right)^2}{p_i(\hat{\delta})} + o_P(1)$$

donde $o_P(1)$ indica términos que convergen en probabilidad a cero cuando n va a infinito. En la practica se utiliza entonces el estadístico de Pearson

$$D_n(x) = n\sum_{i=1}^m \frac{\left(\hat{p}_i - np_i(\hat{\delta})\right)^2}{p_i(\hat{\delta})}$$

para construir la región crítica del test, y como $D_n(x) \sim \mathcal{X}^2_{m-l-1}$ (aprox.) si n es grande tendremos que una región crítica de nivel aproximadamente α vendrá dada por

$$A_1 = \{x, D_n(x) \geq \mathcal{X}^2_{m-l-1;\alpha}\}$$

y la aproximación funciona bien si $np_i(\approx n_i) \geq 5$.

Ejemplo 4.2.3 *Queremos contrastar:*

$$\begin{aligned} H_0 &: p(\delta) = \left(\frac{2+\delta}{4}, \frac{1-\delta}{4}, \frac{1-\delta}{4}, \frac{\delta}{4}\right), \delta \in (0,1) \\ H_1 &: p \neq p(\delta) \end{aligned}$$

en la muestra se obtuvieron los siguientes resultados:

n_1	n_2	n_3	n_4
1997	*906*	*904*	*32*

Necesitamos calcular $\hat{\delta}$. *La verosimilitud bajo* H_0 *viene dada por*

$$L(x;\delta) = \left(\frac{2+\delta}{4}\right)^{n_1} \left(\frac{1-\delta}{4}\right)^{n_2} \left(\frac{1-\delta}{4}\right)^{n_3} \left(\frac{\delta}{4}\right)^{n_4} \tag{4.3}$$

de manera que

$$\partial_\delta \log L(x;\delta) = \frac{n_1}{2+\delta} - \frac{n_2+n_3}{1-\delta} + \frac{n_4}{\delta} = 0 \tag{4.4}$$

implica

$$(n_1+n_2+n_3+n_4)\delta^2 - (n_1 - 2n_2 - 2n_3 - n_4)\delta - 2n_4 = 0. \tag{4.5}$$

Esto es:

$$3839\delta^2 + 1665\delta - 64 = 0. \tag{4.6}$$

Las soluciones son: $\delta = 0.0355281$ *y* $\delta = -0.469235$. *Por tanto* $\hat{\delta} = 0.0355281$ *y*

$$\begin{aligned} np_1(\hat{\delta}) &= 3839 \times \frac{2+0.0355281}{4} = 1953.77 \\ np_2(\hat{\delta}) &= np_3(\hat{\delta}) = 3839 \times \frac{1-0.0355281}{4} = 925.652 \\ np_4(\hat{\delta}) &= 3839 \times \frac{0.0355281}{4} = 34.0981 \end{aligned}$$

Entonces:

$$\begin{aligned} & D_n(x) \\ = \; & \sum_{i=1}^{4} \frac{(n_i - np_i(\hat{\delta}))^2}{np_i(\hat{\delta})} \\ = \; & \frac{(1997-1953.77)^2}{1953.77} + \frac{(906-925.652)^2}{925.652} + \frac{(904-925.652)^2}{925.652} + \frac{(32-34.0981)^2}{34.0981} \\ = \; & 2.00931, \end{aligned}$$

y si miramos las tablas de una χ^2_2 *obtenemos* $\chi^2_{2;0.05} = 5.9914$ *por tanto estamos en la región de aceptación.*

4.2.3 Test de independencia de dos multinomiales

Sean X e Y dos variables con r y s valores distintos respectivamente. Supongamos n observaciones independientes del par (X,Y) queremos saber si X e Y son independientes y esto lo traduciremos en un problema de contraste de hipótesis. En primer lugar notemos que podemos indicar que el resultado de la observación de (X,Y) ha sido (x_i, y_j) de la siguiente forma

$X\diagdown Y$	y_1	$\ldots$	y_j	$\ldots$	y_s
x_1	0		0		0
$\vdots$					
x_i	0		1		0
$\vdots$					
x_r	0		0		0

de manera que podemos pensar que observamos una muestra de tamaño n de una Bernoulli $r \times s$-dimensional, con probabilidades

$$p_{ij} = P(X = x_i, Y = y_j), \quad i = 1,..,r; j = 1,...,s.$$

El resultado de las n-observaciones independientes se puede escribir en la forma de lo que se llama una tabla de *contingencia*

$X\diagdown Y$	y_1	$\ldots$	y_j	$\ldots$	y_s	
x_1	n_{11}		n_{1j}		n_{1s}	$n_{1\cdot}$
$\vdots$						
x_i	n_{i1}		n_{ij}		n_{is}	$n_{i\cdot}$
$\vdots$						
x_r	n_{r1}		n_{rj}		n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$		$n_{\cdot j}$		$n_{\cdot s}$	n

Escribamos

$$p_{i\cdot} = P(X = x_i), p_{\cdot j} = P(Y = y_j), \quad i = 1,..,r; j = 1,...,s.$$

Entonces X e Y serán independientes si y sólo si

$$p_{ij} = p_{i\cdot} \times p_{\cdot j}, \quad i = 1,..,r; j = 1,...,s.$$

de manera que contrastar esta condición de independencia es equivalente a considerar en el modelo multinomial $r \times s$-dimensional asociado, el contraste

$$\begin{aligned} H_0 &: (p_{ij})_{\substack{1\le i\le r\\ 1\le j\le s}} = (p_{i\cdot} \times p_{\cdot j})_{\substack{1\le i\le r\\ 1\le j\le s}} \\ H_1 &: (p_{ij})_{\substack{1\le i\le r\\ 1\le j\le s}} \neq (p_{i\cdot} \times p_{\cdot j})_{\substack{1\le i\le r\\ 1\le j\le s}} \quad . \end{aligned}$$

Notemos que se trata de un *test de ajuste* a una *familia* de multinomiales. Utilizaremos entonces el estadístico de Pearson

$$D_n(x,y) = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n\widehat{p_{i\cdot}}\widehat{p_{\cdot j}})^2}{n\widehat{p_{i\cdot}}\widehat{p_{\cdot j}}}$$

donde $\widehat{p_{i\cdot}}$ y $\widehat{p_{\cdot j}}$ son los EMV de $p_{i\cdot}$ y $p_{\cdot j}$ respectivamente. Para calcularlos hay que escribir la verosimilitud bajo H_0 que vendrá dada por

$$\begin{aligned} L(x,y;(p_{ij})) &= \Pi_{i=1}^{r}\Pi_{j=1}^{s} p_{ij}^{n_{ij}} = \Pi_{i=1}^{r}\Pi_{j=1}^{s}\left(p_{i\cdot}p_{\cdot j}\right)^{n_{ij}} \\ &= \Pi_{i=1}^{r} p_{i\cdot}^{n_{i\cdot}}\,\Pi_{j=1}^{s} p_{\cdot j}^{n_{\cdot j}} \end{aligned}$$

donde $n_{i\cdot} = \sum_j n_{ij}$ y $n_{\cdot j} = \sum_i n_{ij}$. Entonces es inmediato que

$$\widehat{p_{i\cdot}} = \frac{n_{i\cdot}}{n}, \widehat{p_{\cdot j}} = \frac{n_{\cdot j}}{n}.$$

De manera que finalmente

$$D_n(x,y) = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}.$$

Este estadístico, cuando n es grande (la regla es que $n_{ij} \geq 5$) se comportará como una Ji-cuadrado con grados de libertad $\dim\Theta - \dim\Theta_0$,pero

$$\dim\Theta = rs - 1$$

y

$$\dim\Theta_0 = r - 1 + s - 1,$$

de manera que

$$\begin{aligned} \dim\Theta - \dim\Theta_0 &= rs - 1 - r + 1 - s + 1 \\ &= rs - r - s + 1 \\ &= (r-1)(s-1) \end{aligned}$$

y la region crítica de nivel α (aprox.) vendrá dada por

$$A_1 = \{(x,y), D_n(x,y) \geq \mathcal{X}^2_{(r-1)(s-1);\alpha}\}$$

Ejemplo 4.2.4 *La siguiente tabla nos da el número de mujeres, de un grupo de 7477, de edades comprendidas entre 30 y 40 años con un grado de visión entre 1 y 4 en cada uno de los ojos. Queremos contrastar, a nivel* $\alpha = 0.05$ *si hay independencia entre la visión de ambos ojos.*

Ojo izquierdo→ *Ojo derecho↓*	*1*	*2*	*3*	*4*	*Totales*
1	$n_{11} = 1520$	$n_{12} = 266$	$n_{13} = 124$	$n_{14} = 66$	$n_{1\cdot} = 1976$
2	$n_{21} = 234$	$n_{22} = 1512$	$n_{23} = 432$	$n_{24} = 78$	$n_{2\cdot} = 2256$
3	$n_{31} = 117$	$n_{32} = 362$	$n_{33} = 1772$	$n_{34} = 205$	$n_{3\cdot} = 2456$
4	$n_{41} = 36$	$n_{42} = 82$	$n_{43} = 179$	$n_{44} = 492$	$n_{4\cdot} = 789$
Totales	$n_{\cdot 1} = 1907$	$n_{\cdot 2} = 2222$	$n_{\cdot 3} = 2507$	$n_{\cdot 4} = 841$	$n = 7477$

Queremos hacer el contraste de hipótesis:

$$
\begin{aligned}
H_0 &: \; p_{ij} = p_{i\cdot} \times p_{\cdot j} \text{ (independencia)} && (4.7)\\
H_1 &: \; p_{ij} \neq p_{i\cdot} \times p_{\cdot j} && (4.8)
\end{aligned}
$$

Por tanto

$$
\begin{aligned}
& D_n(x,y) \\
= \; & \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{(n_{ij} - n\widehat{p_{i\cdot}}\widehat{p_{\cdot j}})^2}{n\widehat{p_{i\cdot}}\widehat{p_{\cdot j}}} = \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n} \\
= \; & \frac{(1520 - \frac{1976\times1907}{7477})^2}{\frac{1976\times1907}{7477}} + \frac{(266 - \frac{1976\times2222}{7477})^2}{\frac{1976\times2222}{7477}} + \frac{(124 - \frac{1976\times2507}{7477})^2}{\frac{1976\times2507}{7477}} + \frac{(66 - \frac{1976\times841}{7477})^2}{\frac{1976\times841}{7477}} \\
& + \frac{(234 - \frac{2256\times1907}{7477})^2}{\frac{2256\times1907}{7477}} + \frac{(1512 - \frac{2256\times2222}{7477})^2}{\frac{2256\times2222}{7477}} + \frac{(432 - \frac{2256\times2507}{7477})^2}{\frac{2256\times2507}{7477}} + \frac{(78 - \frac{2256\times841}{7477})^2}{\frac{2256\times841}{7477}} \\
& + \frac{(117 - \frac{2456\times1907}{7477})^2}{\frac{2456\times1907}{7477}} + \frac{(362 - \frac{2456\times2222}{7477})^2}{\frac{2456\times2222}{7477}} + \frac{(1772 - \frac{2456\times2507}{7477})^2}{\frac{2456\times2507}{7477}} + \frac{(205 - \frac{2456\times841}{7477})^2}{\frac{2456\times841}{7477}} \\
& + \frac{(36 - \frac{789\times1907}{7477})^2}{\frac{789\times1907}{7477}} + \frac{(82 - \frac{789\times2222}{7477})^2}{\frac{789\times2222}{7477}} + \frac{(179 - \frac{789\times2507}{7477})^2}{\frac{789\times2507}{7477}} + \frac{(492 - \frac{789\times841}{7477})^2}{\frac{789\times841}{7477}} \\
= \; & 8096.88
\end{aligned}
$$

Sabemos que D_n tiene aproximadamente una distribución χ^2_9 y si vamos a las tablas obtenemos que $P\{\chi^2_9 > 16.92\} = 0.05$ y el valor de D_n es muy superior a 16.92, *de hecho sólo con calcular el primer sumando hubiéramos obtenido* 2048.32 *y esto ya nos hubiera conducido a rechazar la hipótesis nula, esto es la hipótesis de independencia.*

Antes de continuar adelante notemos que la razón de verosimilitudes en el caso del contraste anterior de independencia viene dado por

$$
\begin{aligned}
\lambda(x,y) &= \frac{\Pi_{i=1}^r \widehat{p_{i\cdot}}^{n_{i\cdot}} \Pi_{j=1}^s \widehat{p_{\cdot j}}^{n_{\cdot j}}}{\Pi_{i=1}^r \Pi_{j=1}^s \hat{p}_{ij}^{n_{ij}}} = \frac{\Pi_{i=1}^r \left(\frac{n_{i\cdot}}{n}\right)^{n_{i\cdot}} \Pi_{j=1}^s \left(\frac{n_{\cdot j}}{n}\right)^{n_{\cdot j}}}{\Pi_{i=1}^r \Pi_{j=1}^s \left(\frac{n_{ij}}{n}\right)^{n_{ij}}} \\
&= n^{-n} \frac{\Pi_{i=1}^r n_{i\cdot}{}^{n_{i\cdot}} \Pi_{j=1}^s n_{\cdot j}{}^{n_{\cdot j}}}{\Pi_{i=1}^r \Pi_{j=1}^s n_{ij}{}^{n_{ij}}}
\end{aligned}
$$

Test de homogeneidad de varias multinomiales.

Supongamos r variables multinomiales independientes, $Z_1, Z_2, ..., Z_r$ que pueden tomar cada una s valores $z_1, z_2, ..., z_s$ con probabilidades (desconocidas obviamente)

$$
P(Z_i = z_j) = p_{ij}, \quad p_{ij} > 0, \sum_j p_{ij} = 1,
$$

queremos saber si las variables Z_i tienen la misma distribución, esto es si

$$
p_{ij} = p_j, \forall i, j
$$

de manera que la probabilidad de observar el valor z_j es la misma para todas las variables Z_i esto es, $P(Z_i = z_j)$ sólo depende del índice j. Supongamos que tenemos una muestra con n_i observaciones de la variable Z_i. Escribiremos $n_{i\cdot}$ en lugar de n_i; nuestro modelo estadístico será el de r mutinomiales s-dimensionales independientes con parámetros p_{ij}, $j = 1, ..., s$, y tamaños muestrales $n_{i\cdot}$, respectivamente. Se trata de un problema de r muestras. La verosimilitud del modelo será

$$L(x; (p_{ij})) = \Pi_{i=1}^{r} \Pi_{j=1}^{s} p_{ij}^{n_{ij}}$$

donde $\sum_j p_{ij} = 1$ y $\sum_j n_{ij} = n_{i\cdot}$. El resultado de las observaciones se puede escribir también en forma de tabla de contingencia

$Variable \diagdown resultado$	z_1	$\ldots$	z_j	$\ldots$	z_s	
Z_1	n_{11}		n_{1j}		n_{1s}	$n_{1\cdot}$
$\vdots$						
Z_i	n_{i1}		n_{ij}		n_{is}	$n_{i\cdot}$
$\vdots$						
Z_r	n_{r1}		n_{rj}		n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$		$n_{\cdot j}$		$n_{\cdot s}$	n

En este modelo queremos hacer el contraste

$$\begin{aligned} H_0 &: \quad (p_{ij})_{1 \leq j \leq n} = (p_j)_{1 \leq j \leq n}, i = 1, ..., r \\ H_1 &: \quad \text{para algún } i \ , \ (p_{ij})_{1 \leq j \leq n} \neq (p_j)_{1 \leq j \leq n} . \end{aligned}$$

La verosimilitud bajo la hipótesis nula es

$$\begin{aligned} L(x; (p_j)_{1 \leq j \leq n}) &= \Pi_{i=1}^{r} \Pi_{j=1}^{s} p_j^{n_{ij}} \\ &= \Pi_{j=1}^{s} p_j^{\sum_i n_{ij}} = \Pi_{j=1}^{s} p_j^{n_{\cdot j}} . \end{aligned}$$

Podemos construir el TRV y obtenemos

$$\begin{aligned} \lambda(x) &= \frac{\Pi_{j=1}^{s} \widehat{p_j}^{\,n_{\cdot j}}}{\Pi_{i=1}^{r} \Pi_{j=1}^{s} \widehat{p_{ij}}^{\,n_{ij}}} \\ &= \frac{\Pi_{j=1}^{s} \left(\frac{n_{\cdot j}}{n}\right)^{n_{\cdot j}}}{\Pi_{i=1}^{r} \Pi_{j=1}^{s} \left(\frac{n_{ij}}{n_{i\cdot}}\right)^{n_{ij}}} \\ &= n^{-n} \frac{\Pi_{i=1}^{r} n_{i\cdot}^{n_{i\cdot}} \Pi_{j=1}^{s} n_{\cdot j}^{\,n_{\cdot j}}}{\Pi_{i=1}^{r} \Pi_{j=1}^{s} n_{ij}^{n_{ij}}} \end{aligned}$$

la misma expresión que en el caso del contraste de independencia!. Si aplicamos el teorema de Wilks tendremos que $-2 \log \lambda(X)$, si los tamaños muestrales son grandes, tendrá un distribución ji-cuadrado con grados de libertad $\dim \Theta - \dim \Theta_0$,pero

$$\dim \Theta = r(s - 1)$$

y

$$\dim \Theta_0 = s - 1,$$

de manera que

$$\begin{aligned} \dim \Theta - \dim \Theta_0 &= r(s-1) - (s-1) \\ &= (r-1)(s-1), \end{aligned}$$

igual que en el caso del contraste de independencia!. Al igual que en el caso de independencia se utiliza el estadístico de Pearson en lugar de $-2\log\lambda(x)$, por comparación con el caso de independencia, tendremos que

$$D_n(x) = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}.$$

seguirá aproximadamente una ji-cuadrado con $(r-1)(s-1)$ grados de libertad.

Ejemplo 4.2.5 *Disponemos de dos muestras de estudiantes, la primera de escuelas privadas, la segunda de escuelas públicas. Los resultados son relativos a un determinado test de conocimientos.*

Puntuación del test→ Tipo de escuela↓	*0-275*	*276-350*	*351-425*	*426-450*	*Totales*
Privada	*6*	*14*	*17*	*9*	*46*
Pública	*30*	*32*	*17*	*3*	*82*
Totales	*36*	*46*	*34*	*12*	*128*

Como hemos visto, contrastar la hipótesis de homogeneidad corresponde a contrastar la independencia entre las variables "Puntuación del test" y "Tipo de escuela". Por tanto hay que calcular

$$\begin{aligned} & D_n(x) \\ = \; & \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n} \\ = \; & \frac{(6-\frac{46\times 36}{128})^2}{\frac{46\times 36}{128}} + \frac{(14-\frac{46\times 46}{128})^2}{\frac{46\times 46}{128}} + \frac{(17-\frac{46\times 34}{128})^2}{\frac{46\times 34}{128}} + \frac{(9-\frac{46\times 12}{128})^2}{\frac{46\times 12}{128}} \\ & + \frac{(30-\frac{82\times 36}{128})^2}{\frac{82\times 36}{128}} + \frac{(32-\frac{82\times 46}{128})^2}{\frac{82\times 46}{128}} + \frac{(17-\frac{82\times 34}{128})^2}{\frac{82\times 34}{128}} + \frac{(3-\frac{82\times 12}{128})^2}{\frac{82\times 12}{128}} \\ = \; & 17.2858 \end{aligned}$$

Entonces como $P\{\chi^2_3 > 7.815\} = 0.05$ *resulta que estamos en la región crítica y rechazamos la hipótesis de homogeneidad.*

4.3 Test Ji-cuadrado de ajuste, independencia y homogeneidad.

Si partimos de la observación de variables iid cualesquiera siempre podemos substituirlas por variables multinomiales de manera aproximada agrupando los datos cuando sea necesario. Una vez aproximadas las variables por multinomiales podemos hacer los test anteriores y tendremos los llamados test Ji-cuadrado. Vamos a ver un ejemplo de test Ji-cuadrado de ajuste a una familia.

Ejemplo 4.3.1 *Disponemos de los tiempos de vida de 50 componentes electrónicos del mismo tipo, queremos ver si se ajusta a una distribución exponencial.*

262.8	1.0	36.4	4.0	59.4	35.3	70.5	22.6	3.7	5.8
32.1	0.5	17.4	77.6	46.7	182.4	76.7	3.5	13.4	29.7
6.1	15.1	110.5	45.9	31.7	22.4	27.8	10.0	33.0	26.7
8.0	6.8	63.0	70.9	30.0	12.2	29.6	3.3	32.2	12.3
128.2	24.6	7.0	39.8	71.1	19.4	5.4	4.4	54.4	24.8

Hacemos una partición del intervalo $[0, \infty)$ mirando que el número de observaciones en cada intervalo sea superior o igual a 5. Una posible es

$$I_1 = [0, 20), I_2 = [20, 40), I_3 = [40, 70), I_4 = [70, \infty) \tag{4.9}$$

entonces el número de observaciones en cada intervalo es, respectivamente,

$$n_1 = 20, n_2 = 16, n_3 = 5, n_4 = 9. \tag{4.10}$$

Bajo la hipótesis nula de que los datos siguen una exponencial de parámetro μ de manera que la probabilidad de observar resultados en un intervalo (a, b) viene por

$$p((a, b)) = \int_a^b \mu e^{-\mu x} dx = e^{-a\mu} - e^{-b\mu} \tag{4.11}$$

Así las probabilidades de observar resultados en los intervalos I_1, I_2, I_3,y I_4, bajo H_0, serán, respectivamente,

$$p_1(\mu) = 1 - e^{-20\mu}, p_2(\mu) = e^{-20\mu} - e^{-40\mu}, p_3(\mu) = e^{-40\mu} - e^{-70\mu}, p_4(\mu) = e^{-70\mu}. \tag{4.12}$$

Para estimar μ utilizaremos los datos sin agrupar, esto es calcularemos la estimación máximo-verosímil de μ suponiendo que los 50 datos son exponenciales en lugar de que los cuatro datos, $n_i, i = 1, 2, 3, 4$, son multinomiales con parámetros $p_i(\mu), i = 1, 2, 3, 4$. Esto no cambia el comportamiento asintótico del estadístico de Pearson:

$$D_n(x) = \sum_{i=1}^{4} \frac{(n_i - np_i(\hat{\mu}))^2}{np_i(\hat{\mu})}, \tag{4.13}$$

esto es $D_n \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \chi^2_2$, donde $n = \sum_{i=1}^4 n_i = 50$ es el tamaño muestral. La estimación máximo-verosímil de μ es $\hat{\mu} = \frac{n}{\sum_{i=1}^{50} x_i} = 0.0243$. Esto hace que

$$np_1(\hat{\mu}) = 19.25, p_2(\hat{\mu}) = 11.8, p_3(\hat{\mu}) = 6.35, p_4(\hat{\mu}) = 12.6. \tag{4.14}$$

y

$$
\begin{aligned}
& D_n(x) \\
= & \sum_{i=1}^{4} \frac{(n_i - np_i(\hat{\mu}))^2}{np_i(\hat{\mu})} \\
= & \frac{(20-19.25)^2}{19.25} + \frac{(16-11.8)^2}{11.8} + \frac{(5-6.35)^2}{6.35} + \frac{(9-12.6)^2}{12.6} \\
= & 2.80162.
\end{aligned}
$$

Cómo $P\{\chi_2^2 > 5.99\} = 0.05$, *resulta que* $\chi^2_{2;0.05} = 5.99$ *con lo que* $D_n(x) < \chi^2_{2;0.05}$ *y estamos en al región de aceptación. Así es razonable pensar que los datos siguen una distribución exponencial.*

4.4 Test de ajuste y homogeneidad no paramétricos.

4.4.1 La distribución empírica

Los test de ajuste y homogeneidad anteriores se basan en discretizar la variable original, entonces uno se pregunta si no podríamos crear test que no requieran este paso. Una función que caracteriza la ley de las observaciones es la función de distribución y tenemos el equivalente "empírico".

Definición 4.4.1 *Sean n observaciones iid denominaremos* función de distribución empírica *a la función de distribución, que depende de la muestra $x = (x_1, x_2, ..., x_n)$, dada por*

$$\hat{F}_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, y]}(x_i), y \in \mathbb{R}$$

Es una función escalonada continua por la derecha, que pega un salto $\frac{1}{n}$ en $x_{(1)}$, otro del mismo tamaño en $x_{(2)}$ y así sucesivamente hasta llegar al valor 1.

Si las n observaciones iid tienen función de distribución F, tenemos las siguientes propiedades de $\hat{F}_n(x, y)$.

Proposición 4.4.1

$$\hat{F}_n(X, y) \underset{n\to\infty}{\overset{c.s.}{\to}} F(y), \forall y \in \mathbb{R}$$

Demostración. Por la ley fuerte de los grandes números

$$\hat{F}_n(X, y) \underset{n\to\infty}{\overset{c.s.}{\to}} E_F(\mathbf{1}_{(-\infty, y]}(X_1))) = P_F(X_1 \leq y) = F(y).$$

■

Observación 4.4.1 *Notemos que también tenemos que*

$$\hat{F}_n(X, y-) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, y)}(X_i) \underset{n\to\infty}{\overset{c.s.}{\to}} E_F(\mathbf{1}_{(-\infty, y)}(X_1))) = P_F(X_1 < y) = F(y-)$$

Incluso la convergencia es uniforme. Definamos el estadístico

$$\Delta_n(x) = \sup_{y\in\mathbb{R}} \left| \hat{F}_n(x, y) - F(y) \right|,$$

tenemos el siguiente resultado,

Teorema 4.4.1 *(Glivenko-Cantelli)*

$$\Delta_n(X) \underset{n\to\infty}{\overset{c.s.}{\to}} 0$$

Demostración. Utilizamos el convenio habitual $F(\pm\infty) := \lim_{y\to\pm\infty} F(y)$. Escribamos

$$F^{-1}(z) := \inf\{x, F(x) \geq z\}$$

Sea $m \geq 2$ entero y consideremos los puntos

$$y_k^{(m)} = F^{-1}(\frac{k}{m}), k = 0, ..., m.$$

Notemos que $y_0^{(m)} = -\infty$ y que $\hat{F}_n(X, y) = F(y) = 1$ (c.s.) si $y \geq y_m^{(m)}$. Sea entonces $y \in [-y_k^{(m)}, y_{k+1}^{(m)})$, $k = 1, .., m-1$. Tenemos

$$\begin{aligned}
\hat{F}_n(x,y) - F(y) &\leq \hat{F}_n(x, y_{k+1}^{(m)}-) - F(y_k^{(m)}) \\
&\leq \hat{F}_n(x, y_{k+1}^{(m)}-) - F(y_{k+1}^{(m)}-) + F(y_{k+1}^{(m)}-) - F(y_k^{(m)}) \\
&\leq \hat{F}_n(x, y_{k+1}^{(m)}-) - F(y_{k+1}^{(m)}-) + \frac{k+1}{m} - \frac{k}{m} \\
&\leq \hat{F}_n(x, y_{k+1}^{(m)}-) - F(y_{k+1}^{(m)}-) + \frac{1}{m}.
\end{aligned}$$

y

$$\begin{aligned}
\hat{F}_n(x,y) - F(y) &\geq \hat{F}_n(x, y_k^{(m)}) - F(y_{k+1}^{(m)}-) \\
&\geq \hat{F}_n(x, y_k^{(m)}) - F(y_k^{(m)}) + F(y_k^{(m)}) - F(y_{k+1}^{(m)}-) \\
&\geq \hat{F}_n(x, y_k^{(m)}) - F(y_k^{(m)}) + \frac{k}{m} - \frac{k+1}{m} \\
&\geq \hat{F}_n(x, y_k^{(m)}) - F(y_k^{(m)}) - \frac{1}{m}
\end{aligned}$$

En definitiva

$$\begin{aligned}
\left|\hat{F}_n(x,y) - F(y)\right| &\leq \max_{0\leq k\leq m-1} \max\left\{\left|\hat{F}_n(x, y_k^{(m)}) - F(y_k^{(m)})\right|, \left|\hat{F}_n(x, y_{k+1}^{(m)}-) - F(y_{k+1}^{(m)}-)\right|\right\} \\
&\quad + \frac{1}{m}.
\end{aligned}$$

De manera que, con probabilidad uno,

$$\limsup \sup_y \left|\hat{F}_n(X, y) - F(y)\right| \leq \frac{1}{m},$$

el resultado se sigue tendiendo m a infinito. ■

4.4.2 Test de Kolmogorov-Smirnov

A la vista de estas propiedades parece razonable comparar la función de distribución empírica con la que suponemos es la teórica para ver si realmente lo es. Por ejemplo, supongamos que en un modelo de observaciones iid con función de distribución F queremos contrastar

$$\begin{aligned} H_0 &: \quad F = F_0 \\ H_1 &: \quad F \neq F_0 \end{aligned}$$

donde F_0 representa una función de distribución concreta. Entonces una región crítica razonable consistiría en

$$A_1 = \{x, \Delta_n(x) \geq K\}$$

donde

$$\Delta_n(x) = \sup_y \left| \hat{F}_n(x, y) - F_0(y) \right|$$

es decir, si la función de distribución empírica difiere mucho de la teórica rechazamos la hipótesis nula. Pero para poder aplicar este test necesitamos conocer la distribución de Δ_n al menos bajo F_0 y así poder calcular K para que

$$P_{F_0}(\Delta_n \geq K) = \alpha$$

y así poder tener un tets de nivel α. Afortunadamente tenemos la proposición:

Proposición 4.4.2 *Si tenemos una muestra iid de tamaño n de una variable con una distribución continua F_0, Δ_n tiene una distribución que no depende de F_0.*

Demostración. Como $F_0(y)$ y $\hat{F}_n(x, y)$ son ambas crecientes

$$\begin{aligned} \sup_y \left| \hat{F}_n(X, y) - F_0(y) \right| &= \sup_i \left(\left| \hat{F}_n(X, X_{(i)}) - F_0(X_{(i)}) \right|, \left| \hat{F}_n(X, X_{(i)}-) - F_0(X_{(i)}) \right| \right) \\ &= \sup_i \left(\left| \frac{i}{n} - F_0(X_{(i)}) \right|, \left| \frac{i-1}{n} - F_0(X_{(i)}) \right| \right) \\ &= \sup_i \left(\left| \frac{i}{n} - F_0(X)_{(i)} \right|, \left| \frac{i-1}{n} - F_0(X)_{(i)} \right| \right), \end{aligned}$$

pero $F_0(X) \sim U(0, 1)$ de manera que

$$\sup_y \left| \hat{F}_n(y) - F_0(y) \right| \sim \sup_i \left(\left| \frac{i}{n} - U_{(i)} \right|, \left| \frac{i-1}{n} - U_{(i)} \right| \right)$$

donde U_i son observaciones iid de una variable U con ley uniforme en $(0, 1)$. ■

Esta distribución está tabulada y permite construir lo que se llama el test de ajuste de Kolmogorov-Smirnov.

Ejemplo 4.4.1 *Queremos contrastar si el conjunto de datos siguiente se ajusta a una distribución $N(2, 1)$,*

$$0.3, 0.7, 0.9, 1.2, 1.3, 1.4, 1.5, 1.6, 1.9, 2.0, 2.1, 2.2, 2.3, 2.5, 2.6,$$
$$2.7, 3.0, 3.8, 3.9, 4.0.$$

En este caso $n = 20$. *Escribamos*

$$\delta(x_{(i)}) = \max\left(\left|\hat{F}_n(x, x_{(i)}) - F_0(x_{(i)})\right|, \left|\hat{F}_n(x, x_{(i)}-) - F_0(x_{(i)})\right|\right)$$

y Φ *para indicar la función de distribución de una* $N(0,1)$. *Tendremos*

$x-2$	-1.7	-1.3	-1.1	-0.8	-0.7	-0.6	-0.5	-0.4	-0.1	0
$10\times\hat{F}_n(x_{(i)})$	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
$10\times\Phi(x_{(i)})$	0.4	1	1.3	2.1	2.4	2.7	3.1	3.4	4.6	5
$100\times\delta(x_{(i)})$	4	5	3	6	4	3	4	6	6	5

$x-2$	0.1	0.2	0.3	0.5	0.6	0.7	1	1.8	1.9	2
$10\times\hat{F}_n(x_{(i)})$	5.5	5.6	6.5	7	7.5	8	8.5	9	9.5	10
$10\times\Phi(x_{(i)})$	5.4	5.8	6.2	6.9	7.3	7.6	8.4	9.6	9.7	9.8
$100\times\delta(x_{(i)})$	4	3	3	4	3	4	4	11	7	3

Por tanto $\Delta_{20}(x) = 0.11$, *si vamos a las tablas obtenemos que* $\Delta_{20;0.05} = 0.291$ *por tanto estamos en la región de aceptación.*

También se puede contrastar la homogeneidad si tenemos dos muestras independientes $X_1, X_2, ..., X_{n_1}$ y $Y_1, Y_2, ..., Y_{n_2}$ con distribuciones continuas F_1, F_2, esto es

$$\begin{aligned} H_0 &: \quad F_1 = F_2 \\ H_1 &: \quad F_1 \neq F_2 \end{aligned}$$

entonces la región crítica que propone el test de Kolmogorov-Smirnov es

$$A_1 = \{(x,y), \Delta_{n_1,n_2}(x,y) \geq K\}$$

donde

$$\Delta_{n_1,n_2}(x,y) = \sup_z \left|\hat{F}^{(1)}_{n_1}(x,z) - \hat{F}^{(2)}_{n_2}(y,z)\right|$$

Ejemplo 4.4.2 *Los datos siguientes corresponden a las longitudes en cm de plantes obtenidas en dos cultivos de manera independiente, fertilizadas con dos tipos de adobe. Se quiere contrastar la hipótesis de que la distribución de las longitudes es las misma en ambos casos.*

Ejemplo 4.4.3

Adobe A	7	8	10	12	10	13	9	10
Adobe B	11	10	12	13	10	15	17	16

Tenemos entonces

y	7	8	9	10	11	12	13	14	15	16	17
$\hat{F}^{(1)}_{n_1}(y)$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{6}{8}$	$\frac{6}{8}$	$\frac{7}{8}$	1	1	1	1	1
$\hat{F}^{(2)}_{n_2}(y)$	0	0	0	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{4}{8}$	$\frac{5}{8}$	$\frac{5}{8}$	$\frac{6}{8}$	$\frac{7}{8}$	1
$\left\|\hat{F}^{(1)}_{n_1}(y) - \hat{F}^{(2)}_{n_2}(y)\right\|$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{4}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0

por tanto $\Delta_{8,8}(x,y) = \frac{4}{8}$. *En las tablas se obtiene que* $\Delta_{8,8;0.05} = \frac{6}{8}$, *con lo que estamos en la región de aceptación y podemos decir que ambos fertlizantes producen el mismo efecto.*

Capítulo 5

El modelo lineal

Supongamos un modelo estadístico que corresponde a n observaciones $Y_1, Y_2, ..., Y_n$ independientes y tales que

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{ip-1}$$

y $Var(Y_i) = \sigma^2, i = 1, ..., n (n > p)$. Donde los β_j son parámetros desconocidos (coeficientes de regresión, parámetros,...) y los x_{ij} (regresores, covariables, factores,...) son valores conocidos. Esto es

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{ip-1} + \varepsilon_i, \\ E(\varepsilon_i) &= 0, Var(\varepsilon_i) = \sigma^2, i = 1, ..., n \end{aligned}$$

Diremos que tal modelo es un *modelo lineal.* Con notación vectorial

$$Y = X\beta + \varepsilon$$

donde

$$\begin{aligned} Y &= \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{np-1} & \cdots & x_{np-1} \end{pmatrix}, \\ \beta &= \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \end{aligned}$$

Si la *variable respuesta*, Y, y las covariables x, son cuantitativas, diremos que se trata de un *modelo de regresión,* si las covariables son *cualitativas* (0 ó 1) el modelo se suele llamar de *análisis de la varianza,* si en las covariables hay variables cuantitativas y cualitativas se suele llamar modelo de *análisis de la covarianza.*

Estos modelos fueron introducidos por Legendre en 1805 (obtuvo lo que se llama las ecuaciones normales) estudiados por Gauss (1823) y Fisher (1922) que introdujo los modelos de análisis de la varianza.

5.1 Estimación de los parámetros

Notemos que no hay verosimilitud ya que sólo está definida la estructura de la media de las Y pero no su ley (modelo semiparamétrico). Sin embargo podemos buscar β de manera que las observaciones y_i difieran de $\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{ip-1}$ lo mínimo posible. Más concretamente

$$\sum (y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{ip-1}))^2 = \min$$

También podemos escribir, la condición anterior como

$$||y - X\beta||^2 = \min,$$

donde $|| \cdot ||$ representa la norma euclídea en $\mathbb{R}^n$.Notemos que

$$X\beta = \beta_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} + ... + \beta_{p-1} \begin{pmatrix} x_{1p-1} \\ \vdots \\ x_{np-1} \end{pmatrix}$$

De manera que $\{X\beta, \beta \in \mathbb{R}^n\}$ es el subespacio vectorial de $\mathbb{R}^n$ generado por las columnas de $X := \mathcal{R}(X)$. Entonces $\hat{\beta}$ es una solución a nuestro problema si $X\hat{\beta}$ es la proyección ortogonal de y en $\mathcal{R}(X)$, en efecto

$$y - X\hat{\beta} \perp \mathcal{R}(X)$$

implica que

$$\begin{aligned} ||y - X\beta||^2 &= ||y - X\hat{\beta} + X\hat{\beta} - X\beta||^2 \\ &= ||y - X\hat{\beta}||^2 + ||X\hat{\beta} - X\beta||^2 + 2\langle y - X\hat{\beta}, X\hat{\beta} - X\beta|\rangle \\ &= ||y - X\hat{\beta}||^2 + ||X\hat{\beta} - X\beta||^2 \geq ||y - X\hat{\beta}||^2. \end{aligned}$$

Si X es de rango pleno, es decir sus p columnas son linealmente independientes, $\hat{\beta}$ se puede determinar de manera única:

$$y - X\hat{\beta} \perp \mathcal{R}(X)$$

equivale a

$$X'(y - X\hat{\beta}) = 0$$

donde la prima indica traspuesta. De manera que

$$X'X\hat{\beta} = X'y$$

y como $X'X$ es no singular al ser X de rango pleno, resulta

$$\hat{\beta} = (X'X)^{-1} X'y.$$

Esta estimación se denomina *mínimo-cuadrática.*

5.1.1 Propiedades del estimador mínimo-cuadrático

$$\hat{\beta} = (X'X)^{-1} X'Y$$

es lineal en Y :

$$\hat{\beta} = BY, \text{ con } B = (X'X)^{-1} X',$$

es insesgado:

$$\begin{aligned} E_\beta(\hat{\beta}) &= E_\beta((X'X)^{-1} X'Y) \\ &= (X'X)^{-1} X'E_\beta(Y) \\ &= (X'X)^{-1} X'X\beta = \beta. \end{aligned}$$

Su matriz de varianzas-covarianzas viene dada por

$$\begin{aligned} Var_\beta(\hat{\beta}) &= Var_\beta(BY) = BVar_\beta(Y)B' \\ &= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1}. \end{aligned}$$

Teorema 5.1.1 *(Gauss-Markov) Sea $c \in \mathbb{R}^p$, consideremos la función lineal de $\beta, c'\beta$, donde la prima indica traspuesta. Entonces $c'\hat{\beta}$ es el mejor estimador lineal insesgado (BLUE: Best linear unbiased estimator).*

Demostración. *Sea $\lambda'Y$ un estimador lineal insesgado de $c'\beta$, se deberá cumplir que*

$$E_\beta(\lambda'Y) = c'\beta, \ \forall\beta,$$

de manera que

$$\lambda' E_\beta(Y) = \lambda'X\beta = c'\beta, \forall\beta$$

de manera que

$$\lambda'X = c'.$$

Comparemos su varianza con la de $c'\hat{\beta}$,

$$\begin{aligned} Var_\beta(\lambda'Y) &= \lambda' Var_\beta(Y)\lambda \\ &= \sigma^2\lambda'\lambda, \end{aligned}$$

por otra parte

$$\begin{aligned} Var_\beta(c'\hat{\beta}) &= c'Var_\beta(\hat{\beta})c \\ &= \sigma^2 c' (X'X)^{-1} c. \end{aligned}$$

Por tanto

$$\begin{aligned} Var_\beta(\lambda'Y) - Var_\beta(c'\hat{\beta}) &= \sigma^2(\lambda'\lambda - c' (X'X)^{-1} c) \\ &= \sigma^2(\lambda'\lambda - \lambda'X (X'X)^{-1} c) \\ &= \sigma^2 ||\lambda - X (X'X)^{-1} c||^2 \geq 0, \end{aligned}$$

en efecto

$$\begin{aligned}
||\lambda - X\left(X'X\right)^{-1}c||^2 &= (\lambda - X\left(X'X\right)^{-1}c)'(\lambda - X\left(X'X\right)^{-1}c) \\
&= \lambda'\lambda - \lambda' X\left(X'X\right)^{-1}c - \left(X\left(X'X\right)^{-1}c\right)'\lambda \\
&\quad + \left(X\left(X'X\right)^{-1}c\right)' X\left(X'X\right)^{-1}c \\
&= \lambda'\lambda - 2\lambda' X\left(X'X\right)^{-1}c + c'\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}c \\
&= \lambda'\lambda - 2\lambda' X\left(X'X\right)^{-1}c + c'\left(X'X\right)^{-1}c \\
&= \lambda'\lambda - 2\lambda' X\left(X'X\right)^{-1}c + \lambda' X\left(X'X\right)^{-1}c \\
&= \lambda'\lambda - \lambda' X\left(X'X\right)^{-1}c.
\end{aligned}$$

■

Observación 5.1.1 *Notemos que habrá igualdad si y sólo si* $\lambda - X\left(X'X\right)^{-1}c = 0$, *de manera que* $\lambda' Y = c'\left(X'X\right)^{-1}X'Y = c'\hat{\beta}$.

5.1.2 Estimación mínimo cuadrática de σ^2.

Vamos a suponer que σ es también un parámetro a estimar. Notemos primero ciertas propiedades. En primer lugar $\hat{e} := Y - X\hat{\beta}$, es una estimación de los errores, y como $X\hat{\beta}$ es la proyección de Y sobre $\mathcal{R}(X)$, tenemos en particular que

$$Y - X\hat{\beta} \perp \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

con lo que

$$\sum \hat{e}_i = \sum \left(Y_i - \left(X\hat{\beta}\right)_i\right) = 0.$$

Asimismo

$$X\hat{\beta} = X\left(X'X\right)^{-1}X'Y = PY$$

donde

$$P = X\left(X'X\right)^{-1}X'$$

Proposición 5.1.1 $P = X\left(X'X\right)^{-1}X'$ *es la matriz proyección ortogonal sobre* $\mathcal{R}(X)$. *Es por tanto una matriz, simétrica, idempotente y existirá una matriz ortogonal* Q, $(n \times n)$ *tal que*

$$P = Q\begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}Q'$$

donde I_p *es la matriz identidad de dimensión* p.

Demostración. Sea $u \in \mathbb{R}^n$, entonces

$$Pu = X\left(X'X\right)^{-1}X'u = Xh \in \mathcal{R}(X)$$

con $h = \left(X'X\right)^{-1}X'u \in \mathbb{R}^p$, además

$$\begin{aligned} X'(u - Pu) &= X'(I_n - X\left(X'X\right)^{-1}X')u \\ &= \left(X' - X'X\left(X'X\right)^{-1}X'\right)u \\ &= \left(X' - X'\right)u = 0. \end{aligned}$$

Por tanto es la matriz de proyección ortogonal sobre $\mathcal{R}(X)$.

Sea Q una matriz ortogonal cuyas p primeras columnas son vectores ortonormales que constituyen una base de $\mathcal{R}(X)$. Obviamente tendremos

$$Q'PQ = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

■

Observación 5.1.2 *Es obvio que $I_n - P$ será la matriz proyección ortogonal sobre $\mathcal{R}(X)^{\perp}$ (el complementario ortogonal) y existirá una matriz ortogonal R tal que*

$$I_n - P = R\begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix}R'$$

Lema 5.1.1 *Sea A una matriz $n \times n$ determinista y Z un vector aleatorio n-dimensional con $E(Z) = \mu$ y matriz de varianzas-covarianzas $Var(Z) = \Sigma$. Entonces*

$$E(Z'AZ) = traza(A\Sigma) + \mu' A\mu$$

Demostración.

$$\begin{aligned} E(Z'AZ) &= E((Z-\mu)'A(Z-\mu)) + \mu'A\mu \\ &= E(traza((Z-\mu)'A(Z-\mu)) + \mu'A\mu, \end{aligned}$$

ahora bien

$$\begin{aligned} traza((Z-\mu)'A(Z-\mu) &= \sum_{i,j}(Z-\mu)_{1i}A_{ij}(Z-\mu)_{j1} \\ &= \sum_{i,j}A_{ij}(Z-\mu)_{j1}(Z-\mu)_{1i} \\ &= \sum_{i,l}\delta_{il}\sum_{j}A_{lj}(Z-\mu)_{j1}(Z-\mu)_{1i} \\ &= traza(A(Z-\mu)(Z-\mu)'), \end{aligned}$$

de manera que

$$\begin{aligned} E(Z'AZ) &= E(traza(A\,(Z-\mu)\,(Z-\mu)') + \mu' A\mu \\ &= traza\left(AE((Z-\mu)\,(Z-\mu)')\right) + \mu' A\mu \\ &= traza\,(A\Sigma) + \mu' A\mu. \end{aligned}$$

■

Proposición 5.1.2 *(Gauss)*

$$\tilde{\sigma}^2 := \frac{||Y - X\hat{\beta}||^2}{n-p}$$

es un estimador insesgado de σ^2.

Demostración.

$$\begin{aligned} E\left(||Y - X\hat{\beta}||^2\right) &= E(||(I_n - P)Y||^2) \\ &= E(Y'(I_n-P)'(I_n-P)Y) \\ &= E(Y'(I_n-P)Y) \\ &= traza(\sigma^2(I_n-P)I_n) + \beta' X'(I_n-P)X\beta \\ &= \sigma^2(n-p) \end{aligned}$$

ya que al ser $X\beta \in \mathcal{R}(X)$, $(I-P)X\beta = 0$. ■

5.2 El modelo lineal normal

Hasta ahora las hipótesis en el modelo lineal sobre ε eran $E(\varepsilon) = 0$ y matriz de varianzas-covarianzas $V(\varepsilon) = \sigma^2 I_n$. Si añadimos que ε tiene distribución Gaussiana tendremos el modelo lineal normal. Notemos que como $Y = X\beta + \varepsilon$, resultará que $Y \sim N_n(X\beta, \sigma^2 I_n)$. Tenemos una generalización del teorema de Fisher.

i) $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$

ii) $(\hat{\beta}-\beta)'\frac{X'X}{\sigma^2}(\hat{\beta}-\beta) \sim \mathcal{X}_p^2$

iv) $\hat{\beta}$ y $\tilde{\sigma}^2$ son independientes

v) $\frac{(n-p)\tilde{\sigma}^2}{\sigma^2} \sim \mathcal{X}_{n-p}^2$

Demostración. i) es trivial. ii):

$$\begin{aligned} (\hat{\beta}-\beta)'\frac{X'X}{\sigma^2}(\hat{\beta}-\beta) &= (\hat{\beta}-\beta)'V(\hat{\beta})^{-1}(\hat{\beta}-\beta) \\ &= (\hat{\beta}-\beta)'V(\hat{\beta})^{-1/2\prime}V(\hat{\beta})^{-1/2}(\hat{\beta}-\beta) \\ &= Z'Z \end{aligned}$$

con

$$Z = V(\hat{\beta})^{-1/2}(\hat{\beta} - \beta) \sim N_p(0, I_p).$$

iii): Si demostramos que $\hat{\beta}$ y $\hat{\varepsilon} = Y - X\hat{\beta}$ son independientes ya está entonces como la conjunta de $\hat{\beta}$ y cualquier subconjunto de las $\hat{\varepsilon}_i$ es normal, bastará ver que la covarianza es cero.

$$\begin{aligned} Cov(\hat{\beta}, Y - X\hat{\beta}) &= Cov((X'X)^{-1} X'Y, (I-P)Y) \\ &= (X'X)^{-1} X'Cov(Y,Y)(I-P)' \\ &= \sigma^2 (X'X)^{-1} X'(I-P)' = 0. \end{aligned}$$

iv)

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{||Y - X\hat{\beta}||^2}{n-p} = \frac{Y'(I-P)Y}{n-p} \\ &= \frac{(Y - X\beta)'(I-P)(Y-X\beta)}{n-p}, \end{aligned}$$

ahora bien

$$I - P = R \begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix} R'$$

con R ortogonal y

$$U := \frac{1}{\sigma}(Y - X\beta) \sim N_n(0, I_n)$$

por tanto

$$\frac{(n-p)\tilde{\sigma}^2}{\sigma^2} = \sum_{i=1}^{n-p} U_i^2 \sim \mathcal{X}^2_{n-p}$$

■

5.2.1 Estimación máximo verosímil

Es inmediato comprobar que el EMV de β y el estimador mínimo-cuadrático coinciden. Y que el EMV de σ^2 es

$$\hat{\sigma}^2 = \frac{1}{n}||Y - X\hat{\beta}||^2$$

y que por tanto tiene sesgo.

5.2.2 Predicción

Sabemos que $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$ y $\frac{(n-p)\tilde{\sigma}^2}{\sigma^2} \sim \mathcal{X}^2_{n-p}$ independiente de $\hat{\beta}$. De esta manera ai queremos dar un intervalo para la *media* de las observaciones cuando los regresores son $(1, x_1, x_2, .., x_{p-1}) := x'$ bastará tener en cuenta que

$$x'\hat{\beta} \sim N(x'\beta, \sigma^2 x'(X'X)^{-1}x),$$

y así

$$\frac{\frac{x'\hat{\beta}-x'\beta}{\sqrt{\sigma^2 x'(X'X)^{-1}x}}}{\sqrt{\frac{(n-p)\tilde{\sigma}^2}{(n-p)\sigma^2}}} \sim t_{n-p}$$

esto es

$$\frac{x'\hat{\beta}-x'\beta}{\sqrt{\tilde{\sigma}^2 x'(X'X)^{-1}x}} \sim t_{n-p}.$$

Esto nos dará intervalos de confianza para $x'\beta$:

$$x'\hat{\beta} \pm t_{n-p;\alpha/2}\sqrt{\tilde{\sigma}^2 x'(X'X)^{-1}x}.$$

Si quisieramos dar un intervalo para *una observación futura* $Y_{n+1} = x'\beta + \varepsilon_{n+1}$, entonces

$$Y_{n+1} - x'\hat{\beta} = x'\beta - x'\hat{\beta} + \varepsilon_{n+1}$$

de manera que

$$Y_{n+1} - x'\hat{\beta} \sim N(0, \sigma^2(x'(X'X)^{-1}x + 1))$$

tendremos así que un intervalo de *predicción para* Y_{n+1} vendrá dado por

$$x'\hat{\beta} \pm t_{n-p;\alpha/2}\sqrt{\tilde{\sigma}^2(1 + x'(X'X)^{-1}x)}.$$

5.2.3 Contraste de hipótesis

Sea H una matriz $q \times p$ de rango q y $c \in \mathbb{R}^q$. Consideremos el problema de contrastar

$$\begin{aligned} H_0 &: \quad H\beta = c \\ H_1 &: \quad H\beta \neq c. \end{aligned}$$

Vamos a calcular el TRV.

$$L(y; \beta, \sigma) = \frac{1}{\left(\sqrt{2\pi\sigma}\right)^n} \exp\{-\frac{1}{2\sigma^2}\sum(y_i - (X\beta)_i)^2\}.$$

entonces

$$\begin{aligned} \lambda(x) &= \frac{\sup_{\beta,\sigma;H\beta=c} L(y;\beta,\sigma)}{L(y;\hat{\beta},\hat{\sigma})} \\ &= \frac{L(y;\hat{\beta}_H,\hat{\sigma}_H)}{L(y;\hat{\beta},\hat{\sigma})} \end{aligned}$$

donde $\hat{\beta}_H$ es el EMV con la restricción $H\beta = c$.Es inmediato que

$$\hat{\sigma}_H^2 = \frac{1}{n}||Y - X\hat{\beta}_H||^2,$$

y que

$$\lambda(x) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_H^2}\right)^{\frac{n}{2}}.$$

Se acostumbra a escribir

$$RSS = ||Y - X\hat{\beta}||^2 \text{ y } RSS_H = ||Y - X\hat{\beta}_H||^2 \text{ (RSS: residual sum of squares)},$$

de manera que

$$\begin{aligned}
\lambda(x) &= \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_H^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{RSS}{RSS_H}\right)^{\frac{n}{2}} = \left(\frac{1}{1 + \frac{RSS_H - RSS}{RSS}}\right)^{\frac{n}{2}} \\
&= \left(\frac{1}{1 + \frac{q}{n-p}\frac{(RSS_H - RSS)/q}{RSS/(n-p)}}\right)^{\frac{n}{2}}
\end{aligned}$$

Entonces

$$\lambda(x) \leq K \Leftrightarrow \frac{(RSS_H - RSS)/q}{RSS/(n-p)} \geq C$$

De manera que la región crítica del TRV sería

$$A_1 = \{x, \frac{(RSS_H - RSS)/q}{RSS/(n-p)} \geq C\},$$

resulta que C lo podemos calcular fácilmente si fijamos el nivel de significación α por le siguiente resultado.

Proposición 5.2.1 *Bajo* H_0

$$\frac{(RSS_H - RSS)/q}{RSS/(n-p)} \sim F_{q,n-p},$$

Lema 5.2.1 *Si* $U \sim \mathcal{X}_r^2$ *y* $V \sim \mathcal{X}_l^2$ *con* $r > l$ *y* V *y* $U - V$ *son independientes, entonces* $U - V \sim \mathcal{X}_{r-l}^2$.

Demostración. Sea $\varphi_U(\lambda)$ la función generatriz de momentos de U. Entonces $\varphi_U(\lambda) = h(\lambda)^r$ para una cierta función h. Entonces $\varphi_V(\lambda) = h(\lambda)^l$.Por otro lado, al ser V y $U - V$ independientes $\varphi_U(\lambda) = \varphi_{U-V}(\lambda)\varphi_V(\lambda)$ con lo que $\varphi_{U-V}(\lambda) = h(\lambda)^{r-l}$. ■

Demostración. (De la proposición). Sabemos que $RSS/\sigma^2 \sim \mathcal{X}_{n-p}^2$ (esto se cumple bajo la hipótesis general en particular bajo H_0). Veamos que bajo $H_0, RSS_H/\sigma^2 \sim \mathcal{X}_{n-(p-q)}^2$. Basta considerar el caso $c = 0$, si $c \neq 0$ podemos tomar β_0 tal que $H\beta_0 = c$ y tomar

$$\begin{aligned}
\tilde{Y} &: = Y - X\beta_0 = X(\beta - \beta_0) + \varepsilon \\
&= X\tilde{\beta} + \varepsilon
\end{aligned}$$

y ahora $H_0 : H\tilde{\beta} = 0$, además $||\tilde{Y} - X\widehat{\tilde{\beta}}_H||^2 = ||Y - X\beta_0 - X(\widehat{\beta}_H - \beta_0)||^2 = ||Y - X\widehat{\beta}_H||^2 = RSS_H$. Notemos ahora que $\mathcal{H} := \{X\beta, H\beta = 0\}$ es un subespacio de dimensión $p - q$ de $\mathcal{R}(X)$, por tanto el estimador mínimo-cuadrático)(=EMV) será un $\widehat{\beta}_H$ tal que $X\widehat{\beta}_H = P_{\mathcal{H}}Y$ y tendremos

$$\begin{aligned}
&= ||(I_n - P_{\mathcal{H}})Y||^2 \\
&= Y'(I_n - P_{\mathcal{H}})Y \\
&= (Y - X\beta_H)'(I_n - P_{\mathcal{H}})(Y - X\beta_H) \\
&= (Y - X\beta_H)'L\begin{pmatrix} I_{n-(p-q)} & 0 \\ 0 & 0 \end{pmatrix}L'(Y - X\beta_H),
\end{aligned}$$

con L ortogonal, y como bajo H_0, $\frac{Y - X\beta_H}{\sigma} \sim N(0, I_n)$, resultará que

$$\frac{||Y - X\widehat{\beta}_H||^2}{\sigma^2} \sim \mathcal{X}^2_{n-(p-q)}.$$

Por último,

$$\begin{aligned}
RSS_H &= ||Y - X\widehat{\beta}_H||^2 = ||Y - X\widehat{\beta}||^2 + ||X\widehat{\beta} - X\widehat{\beta}_H||^2 \\
&= RSS + ||(I_n - P_{\mathcal{H}})X\widehat{\beta}||^2
\end{aligned}$$

de manera que

$$RSS_H - RSS = ||(I_n - P_{\mathcal{H}})X\widehat{\beta}||^2$$

independiente de RSS (por el teorema de Fisher) de manera que por el lema anterior $RSS_H - RSS \sim \mathcal{X}^2_q$ y como es independiente de RSS,

$$\frac{(RSS_H - RSS)/q}{RSS/(n-p)} \sim F_{q,n-p}$$

■

Observación 5.2.1 *Particularmente interesante es el llamado test de la regresión*

$$\begin{aligned}
H_0 &: \beta_1 = \beta_2 = ... = \beta_{p-1} = 0 \\
H_1 &: \textit{no todos los } \beta_i, i \geq 1 \textit{ son cero.}
\end{aligned}$$

Es un caso particular de los anteriores con $c = 0$ y $H = (0, I_{p-1})$ (toda la primera columna ceros). De esta manera $q = p - 1$. Bajo H_0 el modelo es

$$Y = \beta_0 + \varepsilon$$

y $\hat{\beta}_{0_H} = \bar{Y}$, de manera que $RSS_H = ||Y - \bar{Y}||^2$ y $RSS_H = RSS + ||X\widehat{\beta} - \bar{Y}||^2$. Una medida del ajuste de la regresión es

$$R^2 = \frac{RSS_H - RSS}{RSS_H}.$$