

CO6341: Métodos de la Estadística**Repaso: Distribuciones de Probabilidad Discretas y Contínuas Univariadas, Bivariadas y Multivariadas**

- **Distribuciones discretas (univariadas):** Bernoulli, binomial, hipergeométrica, Poisson, binomial negativa y distribución geométrica
- **Distribuciones discretas (multivariadas):** multinomial
- **Distribuciones contínuas (univariadas):** normal, lognormal, gamma, exponencial, y distribución beta
- **Distribuciones contínuas (bivariadas y multivariadas):** normal bivariada y normal multivariada

Modelos de Probabilidad

Concepto de población

Una población es toda la colección de medidas sobre alguna cantidad para la cual queremos tomar alguna conclusión. Por ejemplo: Altura de los varones menores a cinco años.

Concepto de Distribución de Probabilidad

Modelo matemático que se utiliza para modelar la variabilidad inherente en una población. Por ejemplo: No todos los niños varones menores a cinco años miden los mismo.

Vamos a revisar estos modelos conjuntamente con el concepto de *Variable Aleatoria*.

Variable Aleatoria

Es una función de valores reales definida en un espacio muestral S . Las variables aleatorias son las principales herramientas para modelar cantidades desconocidas.

Variables Aleatorias Discretas vs. Variables Aleatorias Contínuas

Ejemplo V.A. discreta: Los resultados de un experimento al lanzar una moneda (toma los valores 0 y 1). Una variable aleatoria discreta está caracterizada por su función de probabilidad (fp), la cuál especifica la probabilidad de que la variable tome valores distintos.

Ejemplo V.A. contínuas: Las mediciones de una cantidad física (están sujetas a la precisión del instrumento de medición). Una variable aleatoria contínua está caracterizada por su función de densidad de probabilidades (fdp). Al integrar la fdp de una variable aleatoria X sobre un conjunto o intervalo se obtiene la probabilidad de que la variable X tome un valor en ese intervalo.

Función de Distribución Acumulada

Si X es una variable aleatoria, la función de distribución F es la función tal que

$$F(x) = Pr(X \leq x)$$

para todos los valores de x . En términos de la función de densidad f la función de distribución acumulada F es:

$$F(x) = Pr(X \leq x) = \int_{-\infty}^x f(t)dt$$

en el caso continuo, y

$$F(x) = Pr(X \leq x) = \sum_{x_i \leq X} f(X_i)$$

en el caso discreto.

Distribuciones Bernoulli y Binomial

Una variable aleatoria X tiene una distribución Bernoulli con parámetro p si la fp de X es $f(x|p) = p^x(1-p)^{1-x}$ para $x = 0, 1$ y $f(x|p) = 0$ si no.

Si X_1, X_2, \dots, X_n son variables aleatorias i.i.d. que tienen distribución Bernoulli (pruebas Bernoulli) con parámetro p , la variable $X = \sum_{i=1}^n X_i$ tiene una distribución Binomial con parámetros n y p . X cuenta el número de éxitos en la n pruebas Bernoulli, donde éxito en la prueba i corresponde a $X_i = 1$ y fracaso corresponde a $X_i = 0$.

Distribución Hipergeométrica

En este caso se tienen pruebas Bernoulli dependientes. Para ellos se muestra sin reemplazo de una población finita.

Supongamos que una población finita consiste de un número conocido de éxitos y fracasos. Si muestreamos un número fijo de unidades de esa población, el número éxitos en esas unidades tiene una distribución de la familia hipergeométrica.

Supongamos que se muestrean n unidades al azar sin reemplazo de una población que tiene T unidades de las cuales A son éxitos y $B = T - A$ son fracasos. Sea X el número de éxitos en la muestra. La distribución de X es hipergeométrica con parámetros A, B y n .

$f(x|A, B, n) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}}$ para $x = 0, 1, \dots, n$ ó $f(x|A, B, n) = 0$ si no.

Distribución Poisson

Se utiliza para modelar datos que se obtienen por conteo. En general sea X una variable aleatoria discreta que toma valores no negativos. Se dice que X tiene una distribución Poisson con media λ ($\lambda > 0$) si la fdp de X es $f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$ para $x = 0, 1, 2, \dots$ ó $f(x|\lambda) = 0$ si no.

Un **Proceso Poisson** con tasa λ es un modelo para las ocurrencias aleatorias con tasa esperada constante λ por unidad de tiempo (o de área).

El número de ocurrencias en intervalos de tiempos disjuntos es independiente y el número de ocurrencias en un intervalo de longitud (o área) t tiene una distribución Poisson con media λt .

Si n es grande y p es pequeño, entonces la distribución binomial con parámetros n y p es aproximadamente la misma que una distribución Poisson con media np .

Distribución Binomial Negativa

Si observamos una secuencia de pruebas Bernoulli con probabilidad de éxito p , el número de fallas hasta el r -ésimo éxito tiene una distribución Binomial Negativa con parámetros r y p . El caso especial $r = 1$ es la **distribución geométrica** con parámetro p .

Estimación de los parámetros de las Distribuciones de Probabilidad

Métodos de Estimación

- Método de Momentos
- Método de Máxima Verosimilitud
- Estimadores insesgados de varianza mínima
- Estimadores L -momentos
- Estimación Bayesiana

Nota: Todos los estimadores utilizan los valores de la muestra de la variable aleatoria $X: x_1, x_2, \dots, x_n$ para estimar los parámetros poblacionales. Entonces si θ es un parámetro poblacional, entonces:

$$\hat{\theta} = h(x_1, x_2, \dots, x_n)$$

Como el estimador es una función de variables aleatorias, este estimador también es una variable aleatoria, por lo que tiene una distribución de probabilidades asociada.

Problemas de Estimación

- **Problema de Inferencia Estadística:** Se analizan datos que provienen de una distribución de probabilidades desconocida y se debe realizar algún tipo de inferencia sobre la distribución.
- **Problemas reales:** Existe un número infinito de distribuciones que podrían haber generado los datos. Analizando los datos se intenta conocer la distribución desconocida para realizar inferencia acerca de la distribución. Se quiere determinar la *verosimilitud relativa* que cada distribución posible tiene de ser la correcta.
- **Parámetros:** Valores que no conocemos y que definen la distribución que generó los datos.

Ejemplo: La distribución exponencial se utiliza a menudo para representar la distribución del tiempo que transcurre antes de la ocurrencia de un suceso, por ejemplo, la duración de un bombillo.

Sea X una variable aleatoria con parámetro β desconocido.

$$X \sim \text{exp}(\beta)$$

La función de densidad de probabilidades (f.d.p.) de X viene dada por:

$$f(X|\beta) = \begin{cases} \beta e^{-\beta x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

donde $\beta > 0$ y $E[X] = 1/\beta$, $V[X] = 1/\beta^2$

Problemas de estimación (Cont.)

- A partir de valores observados de la duración de un conjunto de bombillos, es posible hacer inferencia sobre el parámetro desconocido β por alguna de estas opciones:
 - Especificar el mejor valor para β
 - Especificar un intervalo
 - Especificar una cota superior
- **Otro ejemplo:** Valores de una distribución normal con μ y σ^2 desconocidos. Supongamos que se tienen observaciones de las estaturas de los individuos de una muestra aleatoria de una población que se asume normal (μ, σ^2) donde μ y σ^2 son los parámetros de la distribución. Es posible hacer inferencia sobre los valores de μ y los valores de σ^2 .
- **Espacio paramétrico:** Es el conjunto Ω de todos los valores posibles del parámetro θ o de un vector de parámetros $(\theta_1, \theta_2, \dots, \theta_n)$. En el caso exponencial Ω es el conjunto de todos los reales positivos; en el caso de la distribución normal, $-\infty < \mu < \infty$ y $\sigma^2 > 0$.

Estimadores de Momentos

Momentos de una distribución

El r^{th} -ésimo momento de la distribución de una variable aleatoria X se calcula como

$$\mu'_r = E(X^r)$$

El r^{th} -ésimo momento muestral se calcula como

$$\hat{\mu}'_r = \frac{1}{n} \sum_{i=1}^n x_i^r$$

El r^{th} -ésimo momento central de una variable aleatoria X es

$$\mu_r = E[(X - \mu)^r]$$

El r^{th} -ésimo momento central muestral se calcula como

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

El método de momentos consiste en la simple idea de establecer un sistema de ecuaciones igualando los momentos muestrales con los momentos de la distribución.

Ejemplo para una distribución Gamma

Si $X \sim \text{Gamma}(\alpha, \beta)$, $X > 0$ la media (primer momento) y la varianza (segundo momento central) de la distribución Gamma tienen la forma:

$$E(X) = \frac{\alpha}{\beta}$$

$$V(X) = \frac{\alpha}{\beta^2}$$

Para estimar α y β a partir de una muestra x_1, x_2, \dots, x_n se estiman $E(X)$ y $V(X)$ a partir de los valores muestrales y se conforma un sistema de dos ecuaciones con dos incógnitas de las cuáles se obtienen $\hat{\alpha}$ y $\hat{\beta}$.

Distribución a priori (inicial) y distribución a posteriori (final)

■ Distribución a priori

Supóngase que se tienen o se tendrán a futuro observaciones de $f(x|\theta)$. En muchos casos antes de disponer de observaciones de $f(x|\theta)$ el experimentador o estadístico tiene conocimientos previos acerca de dónde es probable que se encuentre el valor de θ en el espacio paramétrico Ω . Para ello se construye una distribución de probabilidad θ en el conjunto Ω . Esta distribución se denomina distribución inicial o distribución a priori de θ .

1. Hay un grupo de estadísticos (cada vez mayor!) que piensa que siempre se puede elegir una distribución inicial para θ que tiene carácter subjetivo. **Ejemplo:** Suponer una distribución de probabilidades que represente la creencia subjetiva del valor más probable para θ . Este grupo se adhiere a la Estadística Bayesiana.
2. Hay otro grupo de estadísticos que opina que es necesario conocer extensa información previa para proponer una distribución inicial sobre θ .

Ejemplo: Sea $\theta =$ la proporción de artículos defectuosos en un lote de manufacturas. θ es desconocido pero existe información de lotes anteriores que puede ser utilizada para construir una f.d.p para θ . En ambos casos, la estimación usando métodos Bayesianos es aplicable.

Los métodos de estimación por **máxima verosimilitud** no están basados en asignar una distribución inicial para θ .

Distribuciones a priori discretas y continuas

- Si θ sólo puede tomar un número finito de valores la f.d.p. $\xi(\theta)$ será una dist. discreta.
- Si θ puede tomar cualquier valor en la recta real entonces la f.d.p. $\xi(\theta)$ será una dist. continua.

Ejemplos

- 1 Sea $\theta =$ probabilidad de obtener una cara al lanzar una moneda (θ desconocida). Supongamos que se sabe que la moneda es equilibrada ó tiene dos caras. Entonces $\theta = 1/2$ ó $\theta = 1$.
Si la probabilidad inicial de que la moneda sea equilibrada es p , entonces la f.p. inicial de θ es:

$$\xi(1/2) = p$$

$$\xi(1) = 1 - p$$

- 2 Sea $\theta =$ proporción de artículos defectuosos (θ es desconocida). Supongamos que se asigna una distribución inicial uniforme a θ en el intervalo $(0,1)$. La f.d.p. inicial de θ es:

$$\xi(\theta) = \begin{cases} 1 & 0 < \theta < 1 \\ 0 & \text{en otro caso} \end{cases}$$

3 Parámetro de una exponencial: Sa va a observar la duración de cierto tipo de lámparas fluorescentes. Se asume que la distribución de la duración de cualquier lámpara es exponencial con parámetro β donde:

- β es desconocido
- Por experiencia previa se considera que la distribución inicial de β es gamma con media 0.0002 y desviación típica 0.0001.

Supóngase que la distribución inicial gamma tienen parámetros α_0 y β_0 . En general $X \sim \text{gamma}(\alpha, \beta)$ si X tienen una f.d.p. de la forma:

$$f(x|\alpha, \beta) = \begin{cases} \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

donde $\Gamma(\alpha) =$ Función Gamma, la cual se define como:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

Notar que:

- $\Gamma(1) = 1$; si $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$; si n es entero positivo $\Gamma(n) = (n - 1)!$
- Además $\int_{-\infty}^\infty f(x|\alpha, \beta) dx = 1$ por ser una función de densidad.

La esperanza y la varianza de X se calculan como: $E(X) = \frac{\alpha}{\beta}$ y $V(X) = \frac{\alpha}{\beta^2}$.

Si $E(\beta) = \frac{\alpha_0}{\beta_0} = 0,0002$ y $V(\beta)^{1/2} = \frac{\alpha_0^{1/2}}{\beta_0} = 0,0001$

$\frac{\alpha_0}{\beta_0^2} = 0,00000001 \Rightarrow \frac{0,0002}{\beta_0} = 1 \times 10^{-8} \Rightarrow \beta_0 = 20000$ De aquí se deduce

que $\alpha_0 = 4$; por lo tanto la f.d.p. a priori para β es:

$$\xi(\beta) = \begin{cases} \frac{(20000)^4}{3!} \beta^3 e^{-20000\beta} & \beta > 0 \\ 0 & \beta \leq 0 \end{cases}$$

Distribución a posteriori (o distribución final)

Sea x_1, x_2, \dots, x_n una muestra aleatoria de una distribución con f.d.p. $f(x|\theta)$.

- θ es desconocido con f.d.p. a priori $\xi(\theta)$ con $\theta \in \Omega$.
- La f.d.p. conjunta
 $f_n(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta) \dots f(x_n|\theta) = f_n(\mathbf{x}|\theta)$ donde \mathbf{x} es un vector.
- $f_n(\mathbf{x}|\theta)$ es la f.d.p. conjunta condicional para un valor dado de θ .
- Al multiplicar $f_n(\mathbf{x}|\theta) \cdot \xi(\theta)$ se obtiene la f.d.p. conjunta $n + 1$ de \mathbf{x} y θ .
- La distribución marginal de $\mathbf{x} = (x_1, x_2, \dots, x_n)$ se obtiene integrando la conjunta sobre todos los valores de θ :

$$g_n(\mathbf{x}) = \int_{\Omega} \mathbf{f}_n(\mathbf{x}|\theta) \xi(\theta) d\theta.$$
- La distribución de probabilidad condicional de θ dado \mathbf{x} se denota como $\xi(\theta|\mathbf{x}) = \frac{\mathbf{f}_n(\mathbf{x}|\theta) \xi(\theta)}{g_n(\mathbf{x})}$, $\theta \in \Omega$. Notar que $g_n(\mathbf{x})$ no depende de $\theta \rightarrow \xi(\theta|\mathbf{x}) \propto \mathbf{f}_n(\mathbf{x}|\theta) \xi(\theta)$. Además $\int_{\Omega} \xi(\theta|\mathbf{x}) d\theta = \mathbf{1}$, ya que $\xi(\theta|\mathbf{x})$ es una f.d.p. para θ . (**Teorema de Bayes para parámetros y muestras aleatorias**).

Función de verosimilitud

Cuando la f.d.p conjunta $f_n(\mathbf{x}|\theta)$ se considera como función de θ para valores dados de x_1, x_2, \dots, x_n se llama **función de verosimilitud**.

Entonces la distribución a posteriori $\xi(\theta|\mathbf{x})$ es proporcional a la función de verosimilitud y a la f.d.p. a priori $\xi(\theta)$.

Ejemplo 1: Proporción de artículos defectuosos

Sea $\theta =$ proporción de artículos defectuosos en un lote manufacturado (θ es desconocida). Supongamos $\xi(\theta) \sim Uniforme(0, 1)$.

Sean X_1, \dots, X_n n pruebas Bernoulli tales que:

- $X_i = 1$ si el artículo es defectuoso
- $X_i = 0$ si el artículo no es defectuosos

Para cada valor observado x_i

$$f(x|\theta) = \begin{cases} \theta^x(1-\theta)^{1-x} & x = 0, 1 \\ 0 & \text{si no} \end{cases}$$

Sea $y = \sum_{i=1}^n x_i$. La **función de probabilidad conjunta** de $\mathbf{x} = (x_1, x_2, \dots, x_n)$ para $x_i = 0, 1$ ($i = 1, \dots, n$) es:

$$f_n(\mathbf{x}|\theta) = \theta^y(1-\theta)^{n-y}$$

Entonces $f_n(\mathbf{x}|\theta) \xi(\theta) = \theta^y(1-\theta)^{n-y}$. (*)

Ejemplo 1 (Cont.)

Si una variable aleatoria X tienen distribución beta con parámetros α y β ($\alpha > 0, \beta > 0$),

la f.d.p. $f(\mathbf{x}|\alpha, \beta)$ es:

$$f(x|\alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{si no} \end{cases}$$

Se observa que (*) tienen la misma forma si $\alpha = y + 1$ y $\beta = n - y + 1$
 $\Rightarrow \xi(\theta|\mathbf{x}) \propto \text{Beta}(y + 1, n - y + 1)$ con constante de proporcionalidad
 $\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}$

Ejemplo 2: Parámetro de una exponencial

Sea $X =$ duración de lámparas fluorescentes

$$X \sim \text{exp}(\beta)$$

Supongamos que la f.d.p a priori de β es una Gamma(α_0, β_0).

Supongamos que se observa una muestra aleatoria de duraciones x_1, x_2, \dots, x_n de n lámparas.

Para cada X_i

$$f(x|\beta) = \begin{cases} \beta e^{-\beta x} & x > 0 \\ 0 & \text{si no} \end{cases}$$

Ejemplo 2 (Cont.)

Sea $y = \sum_{i=1}^n x_i$. La f.d.p. conjunta de x_1, x_2, \dots, x_n se puede escribir como :

$$f_n(\mathbf{x}|\beta) = \beta^n e^{-\beta y}$$

Como $\xi(\theta) \sim \text{Gamma}(\alpha_0, \beta_0)$ tal que $\alpha_0 = 4$ y $\beta_0 = 20000$, entonces

$$\xi(\beta) = \frac{(20000)^4}{3!} \beta^3 e^{-20000\beta}$$

$$\Rightarrow f_n(\mathbf{x})\xi(\beta) \propto \beta^{n+3} e^{-(y+20000)\beta}.$$

Esta distribución es proporcional a una distribución gamma con parámetros $\alpha = n + 4$ y $\beta = y + 20000$. \Rightarrow para $\beta > 0$

$$\xi(\beta|\mathbf{x}) = \frac{(y + 20000)^{n+4}}{(n + 3)!} \beta^{n+3} e^{-(y+20000)\beta}$$

Observaciones secuenciales y cálculo de la posteriori

Se calcula la f.d.p. a posteriori después de cada observación:

$$\begin{aligned} \xi(\theta|x_1) &\propto f(x_1|\theta)\xi(\theta) \\ \xi(\theta|x_1, x_2) &\propto f(x_2|\theta)\xi(\theta|x_1) \\ &\cdot \\ &\cdot \\ &\cdot \\ \xi(\theta|x_n) &\propto f(x_n|\theta).\xi(\theta|x_1, \dots, x_{n-1}) \end{aligned}$$

Estimadores máximo verosímiles

Sea x_1, x_2, \dots, x_n una muestra aleatoria de una distribución discreta o continua $f(\xi|\theta)$ donde θ es un parámetro real o un vector de parámetros, tal que $\theta \in \Omega$ (espacio de parámetros).

Supongamos que la probabilidad de obtener un vector observado \mathbf{x} es alta si $\theta = \theta_0$.

Para cualquier vector observado \mathbf{x} se escoje θ tal que $f_n(\mathbf{x}|\theta)$ es máximo.

Sea $\xi(\mathbf{x}) \in \Omega$ un valor de $\theta \in \Omega$ tal que $f_n(\mathbf{x})$ es un máximo y sea $\tilde{\theta} = \xi(\mathbf{x})$ el estimador de θ definido de esa manera. $\tilde{\theta}$ es el estimador máximo verosímil de θ .

Notar que:

- Para ciertos valores observados puede no alcanzarse el máximo de $f_n(\mathbf{x}|\theta) \Rightarrow$ el estimador máximo verosímil (EMV) no existe.
- Para otros valores observados el máximo de $f_n(\mathbf{x}|\theta)$ puede alcanzarse en más de un punto \Rightarrow EMV no es único.

Propiedades de los Estimadores Máximo Verosímiles

■ Invarianza

Si $\hat{\theta}$ es un EMV de θ y $g(\theta) = \tau$ es una función biunívoca de θ , tal que $\theta = h(\tau)$ es la función inversa correspondiente, entonces $g(\hat{\theta})$ es un EMV de $g(\theta)$.

■ Consistencia

Se considera una muestra aleatoria de una distribución con parámetro θ . Para todo tamaño muestral n suficientemente grande podemos suponer que existe un EMV de θ . Bajo ciertas condiciones, la sucesión de estimadores máximo verosímiles converge en probabilidad al valor desconocido θ cuando $n \rightarrow \infty$.

Ejemplos de estimación por máxima verosimilitud (DeGroot & Schervish sección 7.5)

- Parámetro de una distribución Bernoulli ($\text{Ber}(\theta)$)
- Media μ de una distribución normal (varianza conocida) ($\text{N}(\mu, \sigma_0^2)$)
- Media μ y varianza σ^2 de una distribución normal ($\text{N}(\mu, \sigma^2)$)
- Parámetro θ de una distribución uniforme $(0, \theta)$

Nota: El EMV es el valor de θ que maximiza la fp o fdp condicional de los datos X dado θ . Por lo tanto el EMV es el valor de θ que tiene el mayor chance de haber producido la muestra observada. Pero éste no es necesariamente el valor del parámetro que es más probable dados los datos. En cambio la distribución de probabilidad a posterior sí serviría para este propósito.

A continuación estudiaremos un método para medir la cantidad de información que una muestra contiene sobre un parámetro desconocido.

Criterio de información de Fisher para una sola variable aleatoria

Sea X una variable aleatoria cuya f.p. ó f.d.p. es $f(x|\theta)$; θ es un parámetro desconocido tal que $\theta \in \Omega \in \mathfrak{R}$.

X toma valores en S (espacio muestral). Además $f(x|\theta) > 0$. Esta condición implica que la distribución uniforme en $(0, \theta)$ no puede ser considerada porque $f(x|\theta) = 0$ si $x > \theta$ (depende de θ).

Sea $\lambda(x|\theta) = \log f(x|\theta)$. Asumimos que $f(x|\theta)$ es una función dos veces diferenciable con respecto a θ .

Sea $\lambda'(x|\theta) = \frac{d}{d\theta} \lambda(x|\theta)$ y $\lambda''(x|\theta) = \frac{d^2}{d\theta^2} \lambda(x|\theta)$.

La *Información de Fisher* $I(\theta)$ en la variable aleatoria X se define como:

$$I(\theta) = E_{\theta}\{[\lambda'(X|\theta)]^2\}$$

$$\Rightarrow I(\theta) = \int_S [\lambda'(x|\theta)]^2 f(x|\theta) dx$$

Por otra parte:

$$\lambda'(x|\theta) = f'(x|\theta)/f(x|\theta)$$

$$\Rightarrow E_{\theta}[\lambda'(x|\theta)] = \int_S \lambda'(x|\theta) f(x|\theta) dx = \int_S f'(x|\theta) dx$$

Esta última expresión coincide con la derivada con respecto a θ de la igualdad:

$$\int_S f(x|\theta) dx = 1$$

lo cual es igual a cero (asumiendo que podemos derivar dentro de la integral!).

De lo anterior concluimos que:

$$E_{\theta}[\lambda'(X|\theta)] = 0$$

Por lo tanto la media de $\lambda'(X|\theta)$ es cero.

Podemos escribir:

$$I(\theta) = \text{Var}_{\theta}[\lambda'(X|\theta)]$$

lo cual es equivalente a:

$$I(\theta) = E_{\theta}[(\lambda'(X|\theta))^2]$$

.

Además, podemos calcular:

$$\lambda''(x|\theta) = \frac{f(x|\theta)f''(x|\theta) - [f'(x|\theta)]^2}{[f(x|\theta)]^2}$$

$$\lambda''(x|\theta) = \frac{f''(x|\theta)}{f(x|\theta)} - [\lambda'(x|\theta)]^2$$

Al tomar valor esperado de la expresión anterior obtenemos:

$$E_{\theta}[\lambda''(X|\theta)] = \int_S f''(x|\theta)dx - I(\theta)$$

La integral es cero porque $\int_S f''(x|\theta)dx = 0$

$$\Rightarrow I(\theta) = -E_{\theta}[\lambda''(x|\theta)]$$

Ejemplos de Cálculo de la Información de Fisher

1 Distribución Bernoulli:

Sea $X \sim Ber(p)$. Se tiene que $f(x|p) = p^x(1-p)^{1-x}$.

Queremos calcular $I(p)$.

$$\lambda(x|p) = \log f(x|p) = x \log(p) + (1-x) \log(1-p)$$

Al derivar con respecto a p obtenemos:

$$\lambda'(x|p) = \frac{x}{p} - \frac{1-x}{1-p}$$

La derivada segunda es:

$$\lambda''(x|p) = -\left[\frac{x}{p^2} + \frac{1-x}{(1-p)^2}\right]$$

Pero $E(X) = p$. Entonces tenemos que al tomar valor esperado a la expresión anterior se obtiene:

$$I(p) = -E[\lambda''(x|p)] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

2 Distribución Normal con media desconocida y varianza conocida:

Sea $X \sim N(\mu, \sigma^2)$ con $-\infty < \mu < \infty$; $-\infty < x < \infty$

Queremos calcular $I(\mu)$. Para ello calculamos $\lambda''(x|\mu)$ y tomamos valor esperado.

$$\lambda(x|\mu) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x - \mu)^2}{2\sigma^2}$$

Al calcular la primera y segunda derivada de $\lambda(x|\mu)$ con respecto a μ obtenemos:

$$\lambda'(x|\mu) = \frac{(x - \mu)}{\sigma^2}; \quad \lambda''(x|\mu) = -\frac{1}{\sigma^2}$$

$$\Rightarrow I(\mu) = -E[\lambda''(X|\mu)] = \frac{1}{\sigma^2}$$

Cálculo de la Información de Fisher en una muestra aleatoria

Sea x_1, x_2, \dots, x_n una muestra aleatoria de una distribución cuya f.d.p. es $f(x|\theta)$. Sea $\lambda_n(\mathbf{x}|\theta) = \mathbf{log} \mathbf{f}_n(\mathbf{x}|\theta)$. Por analogía se define la Información de Fisher para la muestra aleatoria x_1, x_2, \dots, x_n como:

$$I_n(\theta) = E_{\theta}\{[\lambda'_n(\mathbf{X}|\theta)]^2\}$$

Entonces la información de Fisher en toda la muestra viene dada por la integral múltiple:

$$I_n(\theta) = \int_S \int_S \dots \int_S [\lambda'_n(\mathbf{x}|\theta)]^2 \mathbf{f}_n(\mathbf{x}|\theta) \mathbf{d}_{\mathbf{x}_1} \dots \mathbf{d}_{\mathbf{x}_n}$$

Por analogía con el caso de una sola variable aleatoria podemos escribir:

$$I_n(\theta) = Var_{\theta}[\lambda'_n(\mathbf{X}|\theta)]$$

También podemos escribir:

$$I_n(\theta) = -E_{\theta}[\lambda''_n(\mathbf{X}|\theta)]$$

Dado que $f_n(\mathbf{x}|\theta) = \mathbf{f}(\mathbf{x}_1|\theta) \dots \mathbf{f}(\mathbf{x}_n|\theta)$ se puede comprobar que:

$$\lambda_n(\mathbf{x}|\theta) = \sum_{i=1}^n \lambda(\mathbf{x}_i|\theta)$$

$$\lambda''_n(\mathbf{x}|\theta) = \sum_{i=1}^n \lambda''(\mathbf{x}_i|\theta)$$

Tomando esperanzas a ambos lados de la expresión anterior se obtiene:

$$I_n(\theta) = nI(\theta)$$

En otras palabras, *la información de Fisher en una muestra aleatoria es n veces la observación de Fisher en una sola observación.*

Desigualdad de Cramer-Rao

Como una aplicación de los resultados anteriores demostraremos cómo la información de Fisher puede ser utilizada para determinar una cota inferior de la varianza de un estimador arbitrario del parámetro θ en un problema concreto.

Sea X_1, X_2, \dots, X_n una muestra aleatoria de una f.d.p. $f(x|\theta)$.

Sea $T = r(X_1, X_2, \dots, X_n) = r(\mathbf{X})$ un estimador arbitrario de θ con varianza finita.

Vamos a calcular $Cov_\theta[T, \lambda'_n(\mathbf{X}|\theta)] =$ covarianza entre T y la variable aleatoria $\lambda'_n(\mathbf{X}|\theta)$.

Se tiene que:

$$\begin{aligned} \lambda'_n(\mathbf{x}|\theta) &= \mathbf{f}'_n(\mathbf{x}|\theta)/\mathbf{f}_n(\mathbf{x}|\theta) \\ \Rightarrow E_\theta[\lambda'_n(\mathbf{X}|\theta)] &= \int_S \dots \int_S \mathbf{f}'_n(\mathbf{x}|\theta) \mathbf{d}_{\mathbf{x}_1} \dots \mathbf{d}_{\mathbf{x}_n} = \mathbf{0} \end{aligned}$$

porque

$$\int_S \dots \int_S f_n(\mathbf{x}|\theta) \mathbf{d}_{\mathbf{x}_1} \dots \mathbf{d}_{\mathbf{x}_n} = 1$$

$$Cov_\theta[T, \lambda'_n(\mathbf{X}|\theta)] = E_\theta[T \lambda'_n(\mathbf{X}|\theta)] \quad (1)$$

$$= \int_S \dots \int_S r(\mathbf{x}) \lambda'_n(\mathbf{x}|\theta) \mathbf{f}_n(\mathbf{x}|\theta) \mathbf{d}_{\mathbf{x}_1} \dots \mathbf{d}_{\mathbf{x}_n} \quad (2)$$

$$= \int_S \dots \int_S r(\mathbf{x}) \mathbf{f}'_n(\mathbf{x}|\theta) \mathbf{d}_{\mathbf{x}_1} \dots \mathbf{d}_{\mathbf{x}_n} \quad (3)$$

Definamos $E_\theta[T] = m(\theta)$ $\theta \in \Omega$

$$\int_S \dots \int_S r(\mathbf{x}) \mathbf{f}_n(\mathbf{x}|\theta) \mathbf{d}_{\mathbf{x}_1} \dots \mathbf{d}_{\mathbf{x}_n} = \mathbf{m}(\theta)$$

Derivando a ambos lados con respecto a θ se obtiene:

$$\int_S \dots \int_S r(\mathbf{x}) \mathbf{f}'_n(\mathbf{x}|\theta) \mathbf{d}_{\mathbf{x}_1} \dots \mathbf{d}_{\mathbf{x}_n} = \mathbf{m}'(\theta)$$

$$\Rightarrow Cov_\theta[T, \lambda'_n(\mathbf{X}|\theta)] = \mathbf{m}'(\theta)$$

Desigualdad de Schwarz (Sección 4.6)

Para cualquier par de variables aleatorias U y V se cumple:

$$[E(U.V)]^2 \leq E(U^2)E(V^2)$$

Esto es equivalente a decir que:

$$[Cov(U, V)]^2 \leq \sigma_U^2 \sigma_V^2$$

Utilizando este resultado se obtiene:

$$\{Cov_\theta[T, \lambda'_n(\mathbf{X}|\theta)]\}^2 \leq \mathbf{Var}_\theta(\mathbf{T}) \cdot \mathbf{Var}_\theta[\lambda'_n(\mathbf{X}|\theta)]$$

$$\Rightarrow [m'(\theta)]^2 \leq Var_\theta(T) \cdot I_n(\theta)$$

$$\Rightarrow Var_\theta(T) \geq \frac{[m'(\theta)]^2}{nI(\theta)}$$

Esta es la *Desigualdad de Información o Desigualdad de Cramer-Rao*

Si el estimador de θ es insesgado $m(\theta) = \theta$ y $m'(\theta) = 1$

$\Rightarrow Var_\theta(T) \geq 1/[nI(\theta)]$. Esta desigualdad quiere decir que la varianza de cualquier estimador insesgado no puede ser menor que el inverso de la información de Fisher.

Estimadores Inesgados

- Un estimador $\delta(X_1, \dots, X_n)$ es un estimador inesgado de un parámetro θ si $E_\theta[\delta(X_1, \dots, X_n)] = \theta$ para todo valor posible de θ .

Ejemplo: \bar{X}_n es un estimador inesgado de la media de una distribución normal porque $E_\theta(\bar{X}_n) = \theta$, $-\infty < \theta < \infty$; pero el estimador $S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ **no es un estimador inesgado** de σ^2 (**demostrar**).

- Si un estimador δ es inesgado su error cuadrático medio (ECM):

$$E_\theta[(\delta - \theta)^2] = V_\theta(\delta)$$

es decir, **su ECM es igual a su varianza**.

- La búsqueda de un estimador inesgado con varianza pequeña es equivalente a la búsqueda de un estimador inesgado con ECM pequeño.
- Cualquier estimador que no sea inesgado se denomina *sesgado*. La diferencia entre un estimador sesgado y el parámetro de interés θ se llama **sesgo** del estimador.
- En la mayoría de los problemas no existe un estimador inesgado del parámetro o de alguna función del parámetro que se desea estimar.

Estimadores Eficientes

Se dice que T es un estimador eficiente de su esperanza $m(\theta)$, si se da la igualdad en la desigualdad de la información:

$$\text{Var}_\theta(T) \geq \frac{[m'(\theta)]^2}{nI(\theta)}$$

Nota: Puede no existir un estimador de una función particular $m(\theta)$ cuya varianza alcance la cota inferior de la desigualdad de la información.

Se puede demostrar que T será un estimador eficiente si y sólo si existen funciones $u(\theta)$ y $v(\theta)$ que dependen de θ pero que no dependen de X_1, \dots, X_n tal que T puede ser expresado de la siguiente manera:

$$T = u(\theta)\lambda'_n(\mathbf{X}|\theta) + \mathbf{v}(\theta)$$

La expresión anterior implica que T es una función lineal de $\lambda'_n(\mathbf{X}|\theta)$.

Ejemplo: Muestreo de una distribución Poisson

Supongamos que X_1, \dots, X_n es una muestra aleatoria de una distribución Poisson(θ) ($\theta > 0$). Entonces \bar{X}_n es un estimador eficiente de θ (demostrar).

Estimadores insesgados de varianza mínima

Supongamos que en un problema dado, T es un estimador eficiente de su esperanza $m(\theta)$ y sea T_1 cualquier estimador de $m(\theta)$.

Entonces para todo valor de $\theta \in \Omega$, $Var_\theta(T) =$ cota inferior proporcionada por la desigualdad de Cramer-Rao y $Var_\theta(T_1) \geq \frac{m'(\theta)}{nI(\theta)}$.

Por lo tanto $Var_\theta(T) \leq Var_\theta(T_1)$ para $\theta \in \Omega$.

Si T es un estimador eficiente $m(\theta)$, T tendrá la menor varianza para todo valor de θ posible.

Propiedades de la estimadores Máximo Verosímiles para Muestras Grandes

Supongamos que X_1, \dots, X_n constituyen una muestra aleatoria de una distribución cuya f.d.p. es $f(x|\theta)$.

Para cualquier tamaño muestral n , sea $\hat{\theta}_n$ el EMV de θ . *Se demostrará que si n es grande entonces la distribución de $\hat{\theta}_n$ será aproximadamente normal con media θ y varianza $1/nI(\theta)$.*

Distribución asintótica de un estimador eficiente

Consideremos la variable aleatoria $\lambda'_n(\mathbf{X}|\theta)$. Sabemos que:

$$\lambda_n(\mathbf{X}|\theta) = \sum_{i=1}^n \lambda(\mathbf{X}_i|\theta)$$

También se cumple que:

$$\lambda'_n(\mathbf{X}|\theta) = \sum_{i=1}^n \lambda'(\mathbf{X}_i|\theta)$$

Como los X_i, \dots, X_n son independientes e ident. dist. (i.i.d)
 $\Rightarrow \lambda'(X_1|\theta), \dots, \lambda'(X_n|\theta)$ son también i.i.d.

Sabemos que:

$$E_\theta[\lambda'(X_i|\theta)] = 0$$
$$V_\theta[\lambda'(X_i|\theta)] = I(\theta)$$

Por Teorema Central del Límite:

$$\lambda'_n(\mathbf{X}|\theta)/[\mathbf{nI}(\theta)]^{1/2} \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$$

Supongamos que T es un estimador eficiente de θ , entonces $E_\theta = \theta$ y $Var_\theta = 1/nI(\theta)$.

Además existen funciones $u(\theta)$ y $v(\theta)$ tal que:

$$T = u(\theta)\lambda'_n(\mathbf{X}|\theta) + \mathbf{v}(\theta)$$

Como:

$$E_\theta[\lambda'(\mathbf{X}|\theta)] = \mathbf{0}$$
$$V_\theta[\lambda'(\mathbf{X}|\theta)] = \mathbf{I}(\theta)$$

$$\Rightarrow E_\theta(T) = v(\theta) \quad ; \quad V_\theta(T) = [u(\theta)]^2 I(\theta)$$

$$\Rightarrow v(\theta) = \theta \quad ; \quad u(\theta) = 1/[nI(\theta)]$$

Al sustituir los valores de $u(\theta)$ y $v(\theta)$ en la ecuación $T = u(\theta)\lambda'_n(\mathbf{X}|\theta) + \mathbf{v}(\theta)$, se obtiene:

$$[nI(\theta)]^{1/2}(T - \theta) = \frac{\lambda'_n(\mathbf{X}|\theta)}{[nI(\theta)]^{1/2}} \sim N(0, 1)$$

$$\Rightarrow [nI(\theta)]^{1/2}(T - \theta) \sim N(0, 1)$$

Distribución Asintótica de un E.M.V.

Si el E.M.V. $\hat{\theta}_n$ de θ es un estimador eficiente de θ para cada valor de n ; entonces la distribución asintótica de:

$$[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta) \sim N(0, 1)$$

En general supongamos que en un problema arbitrario el E.M.V. $\hat{\theta}_n$ se determina resolviendo la ecuación $\lambda'_n(\mathbf{X}|\theta) = \mathbf{0}$.

Supongamos también que $\lambda''_n(\mathbf{X}|\theta)$ y $\lambda'''_n(\mathbf{X}|\theta)$ existen. **Entonces la distribución asintótica de $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ es aproximadamente $N(0, 1)$.**

Ejercicio 1: Hallar el E.M.V. de la desviación típica σ de una distribución Normal con media cero y calcule su distribución asintótica.

Ejercicio 2: Discutir el ejemplo 8.8.6 sobre la estimación insesgada del parámetro de una distribución exponencial.

Teorema de Bayes (repasso Sección 2.3 DeGroot & Schervish)

Información Preliminar

El teorema de Bayes provee una manera efectiva de racionalizar los procesos que comprenden un diagnóstico médico. También permiten a un prisionero determinar la relevancia de toda nueva información que se obtenga en su contra. En general el teorema permite actualizar probabilidades sobre un evento cuando llega información nueva.

El teorema ha sido aplicado con frecuencia en casos legales y a problemas de diagnóstico médico.

El reverendo Thomas Bayes (1702-1761) está enterrado en el cementerio de Bunhill, Moorgate, London. El teorema que lleva su nombre fué aplicado originalmente a un problema que involucraba bolas de billard. Se dice que ingresó en la universidad de Edinburgo en 1719, y aunque no se graduó, pudo haber sido alumno de James Gregory, jefe de la cátedra de Matemáticas o de Colin MacLaurin, sucesor de Gregory.

Sea \mathcal{E} un experimento estadístico con espacio muestral S . Sea B_1, B_2, \dots, B_r una partición de S . Sea $P(A) : A \subseteq S =$ la distribución de probabilidades definida sobre todos los eventos de S .

Para todo evento A y B en S con $P(A) > 0$,
 $P(B|A) = P(A \cap B)/P(A)$ = Probabilidad condicional de que B ocurra dado que A ocurrió.

Una versión discreta del teorema empleado por Bayes (1763) nos dice:

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{P(A)} \quad i = 1, 2, \dots, r$$

donde por la ley de probabilidades totales:

$$P(A) = \sum_{j=1}^r P(A|B_j)P(B_j)$$

Este teorema provee un argumento efectivo en el diagnóstico médico.

Ejemplo 1: Considere una enfermedad que se piensa ocurre en una proporción $\theta = 0,01$ de la población.

Supongamos que un médico observa que de los pacientes que poseen la enfermedad, 99 % posee el síntoma Z (Ejemplo: Un resultado positivo en una prueba de sangre). El médico podría concluir que un paciente con el síntoma Z posee una alta probabilidad de tener la enfermedad.

Para cualquier persona escogida aleatoriamente dentro de la población:

- Sea $A =$ el evento de que la persona tiene el síntoma Z .
- sea $B =$ el evento de que la persona tiene la enfermedad

Entonces:

$$\begin{aligned} P(B) &= \theta = 0,01 \\ P(A|B) &= 0,99 \end{aligned}$$

Ejemplo 1 (Cont.): Supongamos que la partición del espacio muestral es de tamaño $r = 2$ tal que $B_1 = B$ y $B_2 = B^c$

$$\begin{aligned}
 P(\text{Enfermedad}|\text{sintoma}) &= P(B|A) \\
 &= \frac{P(A|B).P(B)}{P(A|B).P(B) + P(A|B^c).P(B^c)} \\
 &= \frac{0,99 \times 0,01}{0,99 \times 0,01 + 0,99 \times P(A|B^c)} \\
 &= \frac{1}{1 + 100P(A|B^c)}
 \end{aligned}$$

Si $P(A|B^c) = 1/10$ es la probabilidad de que aparezca el síntoma dado que no se tiene la enfermedad, entonces $P(B|A) = 1/11$, lo cuál quiere decir que la probabilidad de diagnóstico es pequeña (resultado contrario a la intuición del médico).

Al calcular:

$$\begin{aligned}
 P(\text{Enfermedad}|no\ hay\ sintoma) &= P(B|A^c) \\
 &= \frac{P(A^c|B).P(B)}{P(A^c|B).P(B) + P(A^c|B^c).P(B^c)} \\
 &= \frac{0,01 \times 0,01}{0,01 \times 0,01 + 0,99 \times (1 - P(A|B^c))} \\
 &= \frac{1}{9901 - 9900P(A|B^c)}
 \end{aligned}$$

El resultado anterior indica que aunque el síntoma no esté presente hay todavía una pequeña probabilidad de que el paciente sufra la enfermedad. **Ejemplo:** Una prueba de SIDA falsamente negativa no es una completa garantía de no padecer la enfermedad.

Teorema de Bayes en casos de crimen

Ejemplo 2: Sea $N =$ número de individuos sospechosos de cometer un crimen. Antes de recibir cualquier evidencia se le asignan iguales probabilidades a los eventos: $\{i\} = \{ \text{el } i\text{-ésimo miembro de la población es culpable} \}$ ($i = 1, 2, \dots, N$). Se tiene que $P(\{i\}) = 1/N$, $i = 1, \dots, N$.

- Sea Ω una pieza de evidencia presentada en el juicio.
- Sea ϕ_i la probabilidad de que la evidencia Ω haya ocurrido dada $\{i\}$.

Se tiene que $\phi_i = P\{\Omega|\{i\}\}$. **Ejemplo:** $\Omega =$ Prueba de ADN resulta positiva.

Queremos calcular la probabilidad $\xi = P(\{i|\Omega) =$ Probabilidad de que el culpable sea i dada la evidencia. Aplicamos teorema de Bayes:

$$\begin{aligned}\xi_i &= P(\{i|\Omega) \\ &= \frac{P(\Omega|\{i\}).P(\{i\})}{P(\Omega)} \\ &= \frac{\phi_i.N^{-1}}{P(\Omega)}\end{aligned}$$

Calculamos:

$$\begin{aligned}P(\Omega) &= \sum_{i=1}^N P(\Omega|\{i\}).P(\{i\}) \\ &= N^{-1} \sum_{i=1}^N \phi_i \\ &= \phi\end{aligned}$$

$\Rightarrow \xi_i = \frac{\phi_i}{N\phi}$, donde ϕ es el promedio de todos los ϕ_i . ϕ puede ser interpretado como la probabilidad de Ω dado que el criminal fué un individuo seleccionado al azar de la población.

Ejemplo 2: Continuación

Supongamos que $\phi_1 = 1$, $\phi_i = \varepsilon$ para $i = 2, 3, \dots, M$ y $\phi_i = 0$ para $i = M + 1, M + 2, \dots, N$.

En este caso $N\phi = \sum_{i=1}^N \phi_i = 1 + (M - 1)\varepsilon$. Entonces $\xi_1 = \frac{1}{1 + (M - 1)\varepsilon}$.

Note que las probabilidades $P(\{i\}) = 1/N$ pueden ser denominadas como las **probabilidades a priori**, mientras que $\xi_i = P(\{i\}|\Omega)$ son denominadas las **probabilidades a posteriori**.

Ejemplo 3: Estimación de un parámetro de una distribución discreta

Una moneda balanceada es lanzada hasta que se registra la r -ésima cara y el número m de lanzamientos es observado y reportado. Supongamos que r no es reportado y queremos obtener inferencias sobre r a partir los datos.

Supongamos que r tiene una distribución binomial negativa. La verosimilitud de r es:

$$l(r/m) \propto \frac{(m-1)!}{(r-1)!(m-r)!} \quad r = 1, \dots, m$$

Supongamos que se tiene una secuencia contable de probabilidades ϕ_1, ϕ_2, \dots que suman 1 llamadas las probabilidades a priori para $r = i, (i = 1, 2, \dots)$ que representan nuestra información a priori sobre r *antes de observar m* .

Dado que uno observa un valor de m , por el Teorema de Bayes se calcula:

$$p(r/m) = \frac{\phi_r \frac{(m-1)!}{(r-1)!(m-r)!}}{\sum_{i=1}^m \phi_i \frac{(m-1)!}{(i-1)!(m-i)!}}$$

Supongamos que:

$$\phi_r \propto r^{-1} \quad (r = 1, \dots, m^*)$$

Entonces:

$$p(r/m) = \frac{\frac{1}{(r)!(m-r)!}}{\sum_{i=1}^m \frac{1}{(i)!(m-i)!}} \propto \frac{1}{r!(m-r)!} \propto \frac{m!}{r!(m-r)!}$$

Esto implica que si $m < m^*$, la distribución a posteriori de r es binomial con probabilidad $1/2$ y tamaño muestral m . La media a posteriori de esta distribución es $E(r|m) = m/2$ y desviación estándar a posteriori $\sqrt{m}/2$

Distribuciones iniciales conjugadas

Muestreo de una distribución Bernoulli

Ciertas distribuciones iniciales son particularmente convenientes para utilizar con otras distribuciones.

Ejemplo:

Supongamos que se selecciona una muestra aleatoria de una distribución Bernoulli con parámetro θ desconocido. Si la distribución inicial de θ es una distribución Beta, la distribución final de θ será de nuevo una distribución Beta.

Teorema:

Supongamos que X_1, X_2, \dots, X_n constituye una muestra aleatoria de una distribución Bernoulli con parámetro θ desconocido. Sea $\theta \sim \text{Beta}(\alpha, \beta)$ ($\alpha > 0, \beta > 0$). Entonces la distribución final de θ es una distribución $\text{Beta}(\alpha + \sum_{i=1}^n X_i, \beta + n - \sum_{i=1}^n X_i)$.

Demostración del Teorema: en clase.

Ejemplo:

Sea $\theta =$ Proporción de artículos defectuosos en un lote de manufactura. Supongamos que se extrae un artículo del lote. ¿Cuáles son las consecuencias de observar un artículo defectuoso y un artículo no defectuoso en la distribución final de θ ?

Supongamos que θ es desconocida con distribución inicial

$$\xi(\theta) \propto \text{Beta}(\alpha, \beta)$$

- Si el primer artículo es defectuoso:

$$\xi(\theta|X = 1) \propto \text{Beta}(\alpha + 1, \beta)$$

- Si el primer artículo no es defectuoso:

$$\xi(\theta|X = 0) \propto \text{Beta}(\alpha, \beta + 1)$$

\Rightarrow

- cada vez que se observa un artículo defectuoso α se incrementa una unidad.
- cada vez que se observa un artículo no defectuoso β se incrementa una unidad.

La familia de distribuciones Beta se denomina familia conjugada de distribuciones iniciales para muestras de una distribución Bernoulli.

Problema:

Determinar la varianza de la distribución final Beta para θ si la distribución inicial es uniforme (0,1). (Resolver en clase)

Muestreo de una distribución Poisson

La familia de distribuciones Gamma es una familia conjugada de distribuciones iniciales para una muestra de una distribución Poisson.

Teorema:

Si X_1, X_2, \dots, X_n es una muestra aleatoria de una distribución de Poisson con media θ desconocida ($\theta > 0$). Supongamos que la distribución inicial de θ es una $\text{Gamma}(\alpha, \beta)$. La distribución final de θ dado $X_i = x_i$ ($i = 1, \dots, n$) es una distribución $\text{Gamma}(\alpha + \sum_{i=1}^n x_i, \beta + n)$.

Demonstración del teorema: En clase:

Teorema sobre la distribución final de la media de una distribución normal con varianza conocida

Supongamos que X_1, X_2, \dots, X_n es una muestra aleatoria de una distribución Normal con media θ y varianza σ^2 conocida. Supongamos que la distribución inicial $\xi(\theta) \sim N(\mu, v^2)$; entonces la distribución final de θ , dado $X_i = x_i$ ($i = 1, \dots, n$) tiene $N(\mu_1, v_1^2)$ tal que:

$$\begin{aligned}\mu_1 &= \frac{\sigma^2 \mu + nv^2 \bar{x}_n}{\sigma^2 + nv^2} \\ v_1^2 &= \frac{\sigma^2 v^2}{\sigma^2 + nv^2}\end{aligned}$$

Demonstración: Se deja como ejercicio

Sugerencia: Utilizar el siguiente lema de completación de cuadrados:

Lema: Para toda constante A,B,a y b:

$$A(\theta - a)^2 + B(\theta - b)^2 = (A + B)(\theta - \theta^*)^2 + (A^{-1} + B^{-1})^{-1}(a - b)^2$$

donde:

$$\theta^* = (A + B)^{-1}(Aa + Bb)$$

Consecuencias del teorema anterior:

- μ_1 es una media ponderada de la media μ de la distribución inicial y la media muestral \bar{x}_n
- Para valores fijos de v^2 y n , a medida que aumenta la varianza de la muestra σ^2 , el peso relativo asignado a \bar{x}_n será menor
- Para valores fijos de v^2 y σ^2 , a medida que aumenta el tamaño muestral n , el peso relativo asignado a \bar{x}_n será mayor
- Para valores fijos de σ^2 y n , a medida que aumenta la varianza v^2 de la distribución inicial, el peso relativo asignado a \bar{x}_n será mayor
- v_1^2 no depende de los valores de la muestra

Ejemplo: Varianza final de la media de una distribución normal

Supongamos que se seleccionan observaciones al azar de una distribución normal con media θ desconocida y varianza 1. y que la distribución inicial de θ es normal con varianza 4. Se desea estimar el número de observaciones a seleccionar tal que la varianza final (a posterior) de θ sea menor o igual que 0,01.

De la ecuación para v_1^2 :

$$v_1^2 = \frac{4}{1 + 4n}$$

$$v_1^2 \leq 0,01 \text{ si y sólo si } n \geq 99,75$$

$$\Rightarrow n \geq 100.$$

Estimadores de Bayes

Naturaleza del Problema de Estimación

Diferencias entre Estimador y Estimación

- Sean X_1, X_2, \dots, X_n variables aleatorias con f.d.p. $f(x|\theta)$. Un **estimador** de θ es $\delta(X_1, X_2, \dots, X_n)$ = función de las variables aleatorias X_1, X_2, \dots, X_n .
- Una **estimación** es un valor específico del estimador $\delta(x_1, x_2, \dots, x_n)$; donde x_1, x_2, \dots, x_n ó $\delta(\mathbf{x})$ es un valor específico de $\delta(\mathbf{X})$.
- Un **buen estimador** es aquel que tiene una alta probabilidad de que $\delta(\mathbf{X}) - \theta \approx 0$.

Función de pérdida y pérdida esperada

Sea $L(\theta, a)$ = Función que mide la pérdida o el costo para el estadístico cuando el verdadero valor del parámetro es θ y su estimación es a . Si la distancia entre a y θ aumenta, $L(\theta, a)$ también aumenta.

Sea $\xi(\theta)$ = distribución inicial de θ . Si elegimos una estimación particular de $\theta = a$, su pérdida esperada es:

$$E(L(\theta, a)) = \int_{\Omega} L(\theta, a)\xi(\theta)d\theta$$

En cualquier problema de estimación, una función L cuya esperanza $E(L(\theta, a))$ va ser minimizada, se denomina función de pérdida.

Estimador de Bayes

Sea \mathbf{x} un valor observado del vector aleatorio \mathbf{X} antes de estimar θ y sea $\xi(\theta|\mathbf{x})$ la f.d.p final de θ sobre el intervalo Ω .

Para cualquier estimación a , la **pérdida esperada** es:

$$E[L(\theta, a)|\mathbf{x}] = \int_{\Omega} L(\theta, a)\xi(\theta|\mathbf{x})d\theta$$

Para cada valor posible del vector aleatorio \mathbf{X} sea $\delta^*(\mathbf{x})$ un valor de a cuya pérdida esperada es mínima. Entonces la función $\delta^*(\mathbf{X})$ evaluada en \mathbf{x} será un estimador de θ y se denomina **estimador de Bayes**.

El estimador de Bayes se elige tal que:

$$E[L(\theta, \delta^*(\mathbf{x})|\mathbf{x})] = \min_{a \in \Omega} E[L(\theta, a)|\mathbf{x}]$$

El estimador de Bayes dependerá de la función de pérdida y de la distribución inicial de θ .

Algunas funciones de pérdida

Función de pérdida más común:

Pérdida del error cuadrático:

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

La estimación de Bayes $\delta^*(\mathbf{x})$ para cualquier valor observado de \mathbf{x} será el valor de a cuya esperanza $E[(\theta - a)^2 | \mathbf{x}]$ es mínima.

Cuando se utiliza la función del error cuadrático medio, el estimador de Bayes es: $\delta^*(\mathbf{X}) = E(\theta | \mathbf{X})$

Ejemplo 1: Caso Bernoulli con distribución inicial $Beta(\alpha, \beta)$, $\alpha, \beta > 0$.

Sea x_1, x_2, \dots, x_n valores observados cualesquiera y sea $y = \sum_{i=1}^n x_i \Rightarrow \delta(\theta | \mathbf{x}) \sim Beta(\alpha + y, \beta + n - y)$.

Esta distribución tiene media $\frac{\alpha + y}{\alpha + \beta + n}$

El valor estimado de Bayes para x_1, x_2, \dots, x_n es $\delta(\mathbf{x}) = \frac{\alpha + y}{\alpha + \beta + n}$.

El estimador de Bayes es:

$$\delta^*(\mathbf{X}) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n}$$

Estimación de la media de una distribución normal

Supongamos que X_1, X_2, \dots, X_n es una muestra de una distribución normal con media θ desconocida y varianza σ^2 conocida.

Supongamos que la distribución inicial de θ $\xi(\theta) \sim N(\mu, v^2)$.

Si usamos la función de pérdida del error cuadrático

$$L(\theta, a) = (\theta - a)^2, \quad -\infty < \theta < \infty, \quad -\infty < a < \infty.$$

Por resultado anterior:

$$\xi(\theta|\mathbf{x}) \sim N(\mu_1, v_1^2)$$

Por lo tanto el estimador de Bayes $\delta^*(\mathbf{X})$ es:

$$\delta^*(\mathbf{X}) = \frac{\sigma^2 \mu + nv^2 \bar{x}_n}{\sigma^2 + nv^2}$$

Función de pérdida del Error Absoluto

La función de pérdida del error absoluto se define como:

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

Para cualquier valor observado de \mathbf{x} , $\delta(\mathbf{x})$ (la estimación de Bayes) será el valor de a tal que $E(|\theta - a|)$ sea mínima.

$E(|\theta - a|)$ es mínima si $a =$ mediana de la distribución de θ . **El estimador de Bayes en este caso, $\delta^*(\mathbf{X})$, es el estimador cuyo valor siempre es igual a la mediana de la distribución final de θ .**

Ejemplos: Función de pérdida valor absoluto

1. Caso parámetro de la Bernoulli:

La estimación de Bayes $\delta^*(\mathbf{x})$ será igual a la mediana de la distribución final $\text{Beta}(\alpha + y, \beta + n_y)$. Este valor se determina por aproximaciones numéricas para cada conjunto de valores observados.

2. Caso de la media de la distribución normal:

Como la media y la mediana de una distribución normal son iguales, $\delta^*(\mathbf{x})$ es igual a la media de la distribución final.

Otras funciones de pérdida

Las funciones de pérdida del error cuadrático y del error absoluto son las más utilizadas. Sin embargo en algunos problemas puede ser más costoso sobreestimar el valor de θ en cierta cantidad, que subestimarlo en la misma cantidad. En este caso una función de pérdida adecuada podría ser:

$$L(\theta, a) = \begin{cases} 3(\theta - a)^2 & \theta \leq a \\ (\theta - a)^2 & \theta > a \end{cases}$$

Existen otras funciones de pérdida que pueden ser relevantes para otros problemas.

Estimadores de Bayes para muestras grandes

Ejemplo: Efectos de las distintas distribuciones iniciales

Sea θ = Proporción de artículos defectuosos (desconocido).

Supongamos que se desea estimar θ y se utiliza la función de pérdida del Error Cuadrático.

Sea n = número total de artículos = 100.

Sea y = número total de artículos defectuosos = 10

Supongamos que $\xi(\theta) \sim Unif(0, 1)$. Como $\xi(\theta) \sim Beta(1, 1)$

$$\Rightarrow \xi(\theta|\mathbf{x}) \sim Beta(1 + y, 1 + n - y)$$

$$\Rightarrow \delta^*(\mathbf{X}) = E(\theta|\mathbf{X}) = \frac{1+y}{1+1+n} = \frac{11}{102} = 0,108$$

Supongamos ahora que $\xi(\theta) = 2(1 - \theta)$ $0 < \theta < 1$. Se puede demostrar que esta es una distribución propia, es decir, su integral es igual a 1.

También podemos decir que $\xi(\theta|\mathbf{x}) \sim Beta(1, 2)$ lo cual implica que $\xi(\theta|\mathbf{x}) \sim Beta(11, 92)$

$$\Rightarrow E(\theta|\mathbf{x}) = \xi^*(\mathbf{x}) = 11/103 = 0,107.$$

Aunque las dos distribuciones iniciales son muy distintas con medias $1/2$ (Caso Uniforme) y $1/3$ (caso Beta) las estimaciones de Bayes son casi iguales debido a que n es grande ($n = 100$).

Además, ambas estimaciones están cerca de la proporción observada. ($\bar{x}_n = 0,10$) de artículos defectuosos.

Consistencia del estimador de Bayes

- Una sucesión de estimadores que converge al valor desconocido del parámetro que se estima cuando $n \rightarrow \infty$, se denomina sucesión consistente de estimadores. Esto en términos prácticos significa que cuando se selecciona un número grande de observaciones, existe una alta probabilidad de que el estimador de Bayes se encuentre muy cerca del valor desconocido del parámetro θ .
- Para muestras aleatorias de las dos familias de distribuciones tratadas anteriormente, si se asigna una distribución inicial conjugada y se utiliza la función de pérdida cuadrática, los estimadores de Bayes son una sucesión consistente de estimadores. Por ejemplo considere el caso de la distribución normal con media θ y el estimador de Bayes dado por la ecuación:

$$\delta^*(\mathbf{X}) = \frac{\sigma^2 \mu + nv^2 \bar{x}_n}{\sigma^2 + nv^2}$$

.

Por ley de grandes números \bar{X}_n convergerá al valor desconocido de la media θ cuando $n \rightarrow \infty$. Al tomar el límite cuando $n \rightarrow \infty$ $\delta^*(\mathbf{X}) \rightarrow \theta$.

Intervalos de confianza

Los intervalos de confianza son una forma alternativa de utilizar un estimador $\hat{\theta}$ cuando deseamos estimar un parámetro desconocido θ . Podemos encontrar un intervalo (A, B) que pensamos tiene una alta probabilidad de contener al parámetro.

Supongamos que X_1, \dots, X_n constituye una muestra aleatoria de una distribución con parámetro θ . Supongamos también que es posible encontrar estadísticos $A(X_1, \dots, X_n)$ y $B(X_1, \dots, X_n)$ tales que:

$$P(A(X_1, \dots, X_n) < \theta < B(X_1, \dots, X_n)) = \gamma$$

donde γ es una probabilidad fija ($0 < \gamma < 1$). Si los valores observados de A y B son a y b, diremos que el intervalo (a, b) es un *intervalo de confianza para θ* con coeficiente de confianza γ .

No es correcto afirmar que θ está en el intervalo (a, b) con probabilidad γ (las variables aleatorias son los extremos del intervalo; una vez observada la muestra, desaparece la aleatoriedad).

Interpretación del Intervalo de Confianza:

Antes de tomar una muestra, hay una probabilidad γ de que el intervalo que se va a construir a partir de la muestra incluya el valor desconocido de θ . Una vez que se observa la muestra, no es posible asignar una probabilidad al suceso $\theta \in (a, b)$ sin considerar a θ como variable aleatoria, lo cual implica que θ tendrá una distribución de probabilidades.

Limitaciones del Intervalo de Confianza

Aunque los valores de la muestra particular que se observan den mayor información al experimentador sobre si el intervalo realmente incluye a θ , no existe un método estándar para ajustar el coeficiente de confianza γ partiendo de esta información.

Intervalo de Confianza para la media de una distribución normal

Sean X_1, \dots, X_n es una muestra aleatoria de una distribución normal con parámetros μ y σ desconocidos.

Es conocido que:

$$\frac{\bar{X}_n - \mu}{\sqrt{s^2/n}} \sim t_{n-1}$$

Sea $g_{n-1}(x)$ la f.d.p de una t_{n-1} .

Sea c una constante tal que :

$$\int_{-c}^c g_{n-1}(x) dx = \gamma$$

$$\Rightarrow \gamma = G_{n-1}(c) - G_{n-1}(-c) = G_{n-1}(c) - (1 - G_{n-1}(c)) = 2G_{n-1}(c) - 1$$

$$\Rightarrow G_{n-1}(c) = \frac{1 + \gamma}{2}$$

Llamando $1 + \gamma = \alpha$; $c = t_{n-1}^{-1}(\alpha/2)$, es decir, c es el cuantil $\alpha/2$ de una distribución t-Student con $n - 1$ grados de libertad.

$$\Rightarrow P\left(-c < \frac{\bar{X}_n - \mu}{\sqrt{s^2/n}} < c\right) = 1 - \alpha$$

$$\Rightarrow P\left(\bar{X}_n - c \frac{s}{n^{1/2}} < \mu < \bar{X}_n + c \frac{s}{n^{1/2}}\right) = \gamma$$

$$\Rightarrow \left(\bar{X}_n - t_{n-1}^{-1}(\alpha/2) \frac{s}{n^{1/2}}, \bar{X}_n + t_{n-1}^{-1}(\alpha/2) \frac{s}{n^{1/2}}\right)$$

es **intervalo de confianza de coeficiente de confianza γ para μ** .

Intervalos de confianza (Continuación)

Nótese que si $P(c_1 < T < c_2) = \gamma$, ($T \sim t_{n-1}$), el intervalo:

$$\left(\bar{X}_n + \frac{c_1 s}{n^{1/2}}, \bar{X}_n + \frac{c_2 s}{n^{1/2}}\right)$$

es también un intervalo de confianza con coeficiente de confianza γ para μ ; sin embargo puede demostrarse que el intervalo simétrico con respecto a \bar{X}_n es el de menor longitud.

Ejemplo: Variables aleatorias uniformes en un intervalo de longitud uno

Sean X_1 y X_2 son variables aleatorias tomadas de una distribución uniforme en el intervalo $[\theta - 1/2, \theta + 1/2]$, donde el valor de θ es desconocido ($-\infty < \theta < \infty$).

Sean $Y_1 = \min(X_1, X_2)$ y $Y_2 = \max(X_1, X_2)$.

$$\begin{aligned} P(Y_1 < \theta < Y_2) &= P(X_1 < \theta < X_2) + P(X_2 < \theta < X_1) \\ &= P(X_1 < \theta)P(X_2 > \theta) + P(X_2 < \theta)P(X_1 > \theta) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \end{aligned}$$

Para valores observados de y_1 y y_2 el intervalo (y_1, y_2) será un intervalo de confianza para θ con coeficiente de confianza $1/2$.

Continuación del ejemplo anterior

Las observaciones X_1 y X_2 deben ser al menos $\theta - (1/2)$ y ambas deben ser máximo $\theta + (1/2)$. Sabemos con certeza que $y_1 \geq \theta - (1/2)$ y $y_2 \leq \theta + (1/2)$. En otras palabras:

$$y_2 - (1/2) \leq \theta \leq y_1 + (1/2) \quad (*)$$

Supongamos que $(y_2 - y_1) > 1/2$; entonces $y_1 < y_2 - (1/2)$ y de (*) se obtiene que $y_1 < \theta$. Además como $y_1 + (1/2) < y_2$, entonces de (*) se obtiene que $\theta < y_2$. Esto nos dice con certeza que $y_1 < \theta < y_2$; es decir, el intervalo (y_1, y_2) incluye al valor desconocido de θ cuando $(y_2 - y_1) > 1/2$, aunque el coeficiente de confianza sea solamente $1/2$.

Notar que el coeficiente de confianza no depende de los valores observados y_1 y y_2 .

Intervalos de Probabilidad

Sean c_1 y c_2 cantidades tales que:

$$P(c_1 < \theta < c_2 | \mathbf{x}) = \gamma$$

En este caso diremos que (c_1, c_2) es un intervalo de probabilidad γ para θ (**Nótese la diferencia de interpretación!**).

Claramente existen muchos valores posibles de c_1 y c_2 , y por lo tanto muchos posibles intervalos de probabilidad. Se suele usar el que tiene mínima longitud, el cuál corresponde al **intervalo de probabilidad posterior máxima (HPD)**.

Idealmente a uno le gustaría reportar una región de valores de θ que sea tan pequeña como sea posible, pero que tenga tanta probabilidad como sea posible.

Análisis Bayesiano de muestras de una distribución normal

Sea X_1, \dots, X_n una muestra aleatoria de una distribución normal con media μ y varianza desconocida.

Sea $\tau = 1/\sigma^2$ la *precisión*. Si X es una v.a. con distribución $N(\mu, \tau)$, entonces

$$f(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^2 \exp\left[-\frac{1}{2}\tau(x - \mu)^2\right] \quad -\infty < x < \infty$$

Si X_1, \dots, X_n es una muestra aleatoria de una distribución normal con media μ y precisión τ , entonces la f.d.p. conjunta $f_n(\mathbf{x}|\mu, \tau)$ con $-\infty < x_i < \infty$ ($i = 1, \dots, n$) es :

$$f_n(\mathbf{x}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{1}{2}\tau \sum_{i=1}^n (x_i - \mu)^2\right]$$

Familia conjugada de distribuciones iniciales

Consideremos la distribución inicial:

$$\xi(\mu, \tau) = \xi_1(\mu|\tau)\xi_2(\tau)$$

- ξ_1 se asume normal con precisión proporcional a τ
- ξ_2 se asume Gamma

Esta familia de distribuciones es una **familia conjugada de distribuciones iniciales conjunta**.

Para cualquier conjunto posible de valores observados, la distribución final conjunta de μ y τ pertenece a la familia.

Este resultado se formaliza en el siguiente teorema:

TEOREMA

Sea X_1, \dots, X_n una muestra aleatoria de una distribución normal con media μ y precisión τ desconocidas ($-\infty < \mu < \infty$) y $\tau > 0$.

Supongamos que $\xi(\mu, \tau) = \xi_1(\mu|\tau)\xi_2(\tau)$ es la distribución inicial conjunta donde:

$$\xi_1(\mu|\tau) \sim N(\mu_0, \lambda_0\tau)$$

y

$$\xi_2(\tau) \sim \text{Gamma}(\alpha_0, \beta_0) \quad (\alpha_0 > 0, \beta_0 > 0)$$

La distribución final conjunta de μ y τ dado $X_i = x_i$ ($i = 1, \dots, n$) es como sigue:

$$\xi(\mu, \tau|\mathbf{x}) = \xi_1(\mu|\tau, \mathbf{x}) \cdot \xi_2(\tau|\mathbf{x})$$

donde:

$$\xi_1(\mu|\tau, \mathbf{x}) \sim N(\mu_1, \lambda_1\tau)$$

con $\mu_1 = \frac{\lambda_0\mu_0 + n\bar{x}_n}{\lambda_0 + n}$ y $\lambda_1 = \lambda_0 + n$.

$$\xi_2(\tau|\mathbf{x}) \sim \text{Gamma}(\alpha_1, \beta_1)$$

con $\alpha_1 = \alpha_0 + n/2$ y $\beta_1 = \beta_0 + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 + \frac{n\lambda_0(\bar{x}_n - \mu_0)^2}{2(\lambda_0 + n)}$

Se dice que μ y τ tienen una distribución conjunta **normal-gamma con parámetros** $\mu_1, \lambda_1, \alpha_1, \beta_1$

Demonstración del Teorema: En clase

Distribución marginal de la media

Para calcular la distribución marginal de la media μ debemos integrar la f.d.p. conjunta con respecto a τ :

$$\xi_3(\mu) = \int_0^{\infty} \xi_1(\mu|\tau) \cdot \xi_2(\tau) d\tau$$

$$\Rightarrow \xi_3 \propto \int_0^{\infty} \tau^{\alpha_0-1/2} e^{-[\beta_0+1/2\lambda_0(\mu-\mu_0)^2]\tau} d\tau$$

Recordar que:

$$\int_0^{\infty} x^{\alpha-1} e^{-\beta x} dx = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

$$\Rightarrow \xi_3(\mu) \propto [\beta_0 + 1/2\lambda_0(\mu - \mu_0)^2]^{-(\alpha_0+1/2)}$$

$$\Rightarrow \xi_3(\mu) \propto \left[1 + \frac{1}{2\alpha_0} \frac{\lambda_0\alpha_0}{\beta_0} (\mu - \mu_0)^2\right]^{-\frac{(2\alpha_0+1)}{2}}$$

Se define la variable aleatoria:

$$Y = \left(\frac{\lambda_0\alpha_0}{\beta_0}\right)^{1/2} (\mu - \mu_0)$$

$$\Rightarrow \mu = \left(\frac{\beta_0}{\lambda_0\alpha_0}\right)^{1/2} \cdot Y + \mu_0$$

Recordar (Pag. 167. Sección 3.8 del DeGroot & Schervish):

Supongamos que X y Y son dos variables aleatorias tales que $Y = r(X)$ y $X = s(Y)$.

$$G(y) = Pr(Y \leq y) = Pr(r(X) \leq y) = Pr(X \leq s(y)) = F(s(y))$$

$$\Rightarrow g(y) = \frac{dG(y)}{dy} = \frac{dF(s(y))}{dy} = f(s(y)) \cdot \frac{ds(y)}{dy}.$$

$$g(y) = \xi_3 \left[\left(\frac{\beta_0}{\lambda_0 \alpha_0} \right)^{1/2} y + \mu_0 \right] \frac{\beta_0^{1/2}}{(\lambda_0 \alpha_0)^{1/2}}$$

(donde aparece μ sustituimos $\frac{\beta_0^{1/2}}{(\lambda_0 \alpha_0)^{1/2}} y + \mu_0$)

$$\Rightarrow g(y) \propto \left(1 + \frac{y^2}{2\alpha_0} \right)^{-\frac{(2\alpha_0+1)}{2}}$$

$g(y)$ es proporcional a una t-student con $2\alpha_0$ grados de libertad.

La distribución marginal de μ es una t-student trasladada μ_0 unidades y con distinto factor de escala

La media y la varianza de la distribución marginal de μ se obtienen así:

Dado que

$$E[Y] = 0 \quad \text{si } \alpha_0 > 1/2$$

y

$$V[Y] = \frac{2\alpha_0}{2\alpha_0 - 2} = \frac{\alpha_0}{\alpha_0 - 1} \quad \text{si } \alpha_0 > 1$$

\Rightarrow

$$\begin{aligned} E[\mu] &= \mu_0 \\ Var[\mu] &= \frac{\beta_0}{\lambda_0 \alpha_0} V[Y] \\ &= \frac{\beta_0}{\lambda_0 (\alpha_0 - 1)} \end{aligned}$$

Notar que:

- La probabilidad de que μ esté en cualquier intervalo específico se puede obtener de una tabla t -student aunque $2\alpha_0$ no sea necesariamente entero.
- La distribución final marginal de μ también es una t -student. Por lo tanto la media y la varianza de esta distribución final marginal se pueden obtener de la distribución t -student correspondiente. En este caso debemos hablar de regiones o intervalos de probabilidad a posterior máxima (HPD)

Ejemplo:

- Encontrar el intervalo de probabilidad inicial para la media μ de una distribución Normal-Gamma tal que:
 $E[\mu] = 10, V[\mu] = 8, E[\tau] = 2, V[\tau] = 2.$
- Si se observa una muestra de tamaño 20 tal que $\bar{x}_n = 7,5$ y $\sum_{i=1}^n (x_i - \bar{x}_n)^2 = 28$, encontrar el intervalo de probabilidad final para μ .

(Ejemplo desarrollado en clase)

Ejemplo: Casas Hogares en New Mexico

- Datos de casas hogares con licencia analizados por Smith, Piland y Fisher (1992).
- Se considera la variable: Total anual de días de hospitalización (X), medida en cientos.
- Antes de observar los datos se modelan los valores de X como una normal para cada casa hogar, con media μ y precisión τ .
- Se usa la siguiente información adicional para estimar μ y τ :
 - Hay en promedio 111 camas con desviación estándar de 43.5 camas. Se asume a priori un 50% de ocupación. Se puede estimar una media (en cientos de pacientes por a no) como $0,5 \times 365 \times 1,11 \approx 200$ y una desviación estándar de $0,5 \times 365 \times 0,435 \approx 6300^{1/2}$.
 - Se atribuye la mitad de la varianza de 6300 a las casas hogares, y la otra mitad es la varianza de μ . Entonces $Var(\mu) = 3150$ y $E(\tau) = 1/3150$. Si se escoge $\alpha_0 = 2$, como $E(\tau) = \alpha_0/\beta_0$ entonces $\beta_0 = 6300$.
 - Usando que $E(\mu) = \mu_0$ y $Var(\mu) = \frac{\beta_0}{\lambda_0(\alpha_0-1)}$, se obtiene $\mu_0 = 200$ y $\lambda_0 = 2$.

- La variable aleatoria Y definida anteriormente tiene una distribución t-student con $2\alpha_0$ grados de libertad. Por lo tanto la variable $0.025(\mu - \mu_0)$ tiene una distribución t-student con cuatro grados de libertad.
- El cuantil de la t-student con dos grados de libertad, correspondiente a una probabilidad de 0.975 es 2.776. Por lo tanto $Pr[-2,776 < 0,025(\mu - 200) < 2,776] = 0,95$, lo cual es equivalente a: $Pr(89 < \mu < 311) = 0,95$. El intervalo (89,311) es el intervalo de probabilidad del 95 % para μ .
- Al observar la muestra:
128, 281, 291, 238, 155, 148, 154, 232, 316, 96, 146, 151, 100, 213, 208, 157, 48, 217,
se obtiene: $\bar{x}_n = 182,17$ y $s_n^2 = 88678,5$.
- Los parámetros de la normal-gamma final son: $\mu_1 = 183,95$,
 $\lambda_1 = 20$, $\alpha_1 = 11$, $\beta_1 = 50925,37$.
- Usando el argumento anterior se llega a un intervalo de probabilidad a posterior del 95 % = (152.38,215.52), el cual es mucho más corto que el intervalo inicial.

Comparación con el Intervalo de Confianza

La variable aleatoria:

$$U = \frac{n^{1/2}(\bar{X}_n - \mu)}{\left(\frac{S_n^2}{n-1}\right)^{1/2}}$$

tiene una distribución t-student con 17 grados de libertad ($n = 18$).

El cuantil 0.975 de la t-student es 2.110. Por lo tanto

$$Pr(-2,110 < U < 2,110) = 0,95$$

En este caso el intervalo con coeficiente de confianza del 95 % para μ es (146.25, 218.09), el cual es cercano al intervalo (152.38, 215.52).

Se obtienen resultados similares aunque la interpretación de los intervalos es distinta!

Contraste de Hipótesis

Sea θ un parámetro desconocido perteneciente al espacio paramétrico Ω . Supongamos que Ω puede descomponerse en los conjuntos Ω_0 y Ω_1 disjuntos.

Sea:

- H_0 : Hipótesis de que $\theta \in \Omega_0$
- H_1 : Hipótesis de que $\theta \in \Omega_1$

Como $\Omega = \Omega_0 \cup \Omega_1$, una de las dos hipótesis debe ser verdadera. Debemos decidir entre si aceptar H_0 ó aceptar H_1 .

Problema de Contraste de Hipótesis

Sólo hay dos decisiones posibles. Para tomar la decisión se lleva a cabo un procedimiento de contraste.

En general H_0 y H_1 se tratan de una forma distinta. A H_0 se le llama la **hipótesis nula** y a H_1 se le llama la **hipótesis alternativa**.

Hipótesis Simples y Compuestas

Supongamos que X_1, \dots, X_n es una muestra aleatoria de una f.d.p. $f(x|\theta)$ tal que $\theta \in \Omega$ y $\Omega = \Omega_0 \cup \Omega_1$. (Ω_0 y Ω_1 son disjuntos)

Consideremos:

- $H_0 : \theta \in \Omega_0$
- $H_1 : \theta \in \Omega_1$
 - Si el conjunto Ω_i sólo contiene un valor de θ se dice que H_i es **simple**.
 - Si Ω_i contiene más de un valor se dice que H_i es **compuesta**.

En el caso $H_0 : \theta = \theta_0$ el tamaño del contraste es $\pi(\theta_0)$.

Nota:

Con una hipótesis simple la distribución de las observaciones queda completamente especificada. Con una hipótesis compuesta se dice que las observaciones pertenecen a una cierta clase de distribuciones.

Región Crítica

Consideremos el problema de contraste:

- $H_0 : \theta \in \Omega_0$

- $H_1 : \theta \in \Omega_1$

Vamos a definir el concepto de Región Crítica.

Región Crítica (Cont.)

Antes de tomar una decisión se observa una muestra X_1, \dots, X_n de una distribución de probabilidades con parámetro desconocido θ .

Sea S el espacio aleatorio n-dimensional de todos los posibles valores de $\mathbf{X} = X_1, \dots, X_n$ (Vector aleatorio n-dimensional).

Dividimos a S en dos subconjuntos S_0 y S_1 :

- S_0 : El que contiene todos los valores de \mathbf{X} para los cuales se acepta H_0 .
- S_1 : El que contienen todos los valores de \mathbf{X} para los cuales se rechaza H_0 (por lo tanto se acepta H_1).

El conjunto para el cuál H_0 es rechazada se denomina **región crítica del contraste**.

En conclusión:

Un procedimiento de contraste se determina especificando la región crítica del contraste (donde H_0 será rechazada). Por lo tanto, en el complemento, H_0 será aceptada.

Nota: Las divisiones del espacio paramétrico Ω_0 y Ω_1 y del espacio muestral S_0 y S_1 están relacionadas entre sí, pero no coinciden. Si la muestra aleatoria X cae en la región crítica S_1 rechazamos la hipótesis nula Ω_0 . Si $X \in S_0$, no rechazamos Ω_0 .

Función de Potencia

Sea δ un procedimiento de contraste. Sea $\pi(\theta|\delta) = \text{Probabilidad de que el procedimiento de contraste } \delta \text{ produzca el rechazo de } H_0$.

Entonces $1 - \pi(\theta|\delta)$ es la probabilidad de que se produzca la aceptación de H_0 .

Sea $C =$ Región crítica del contraste para $\theta \in \Omega$.

$\pi(\theta|\delta)$ especifica para cada valor de θ la probabilidad de que H_0 sea rechazada.

Sería ideal que:

- $\pi(\theta|\delta) = 0 \quad \forall \theta \in \Omega_0$
- $\pi(\theta|\delta) = 1 \quad \forall \theta \in \Omega_1$

En la práctica esta función de potencia es irreal.

Para cualquier valor de $\theta \in \Omega_0$, la decisión de rechazar H_0 es incorrecta; entonces si $\theta \in \Omega_0$, $\pi(\theta|\delta) = \text{Probabilidad de tomar la decisión incorrecta}$.

Errores Tipo I y Tipo II

- Si $\theta \in \Omega_0$ $\pi(\theta|\delta)$ es la probabilidad de cometer un **error tipo I**.
- Si $\theta \in \Omega_1$ $1 - \pi(\theta|\delta)$ es la probabilidad de cometer un **error tipo II**.

Generalmente se especifica una cota superior α_0 para esta probabilidad y se consideran los contrastes tales que $\pi(\theta|\delta) \leq \alpha_0 \quad \forall \theta \in \Omega_0$.

α_0 se denomina **Nivel de Significancia** del contraste.

Tamaño del contraste

Se define como:

$$\alpha(\delta) = \sup_{\theta \in \Omega_0} \pi(\theta|\delta)$$

Por lo tanto α es la máxima probabilidad de tomar una decisión incorrecta entre todos los valores de θ que satisfacen la hipótesis nula.

Ejemplo 1:**Contrate de Hipótesis de una Distribución Uniforme**

Sea X_1, \dots, X_n una muestra aleatoria de una distribución uniforme sobre el intervalo $(0, \theta)$ (θ desconocido).

Se desea contrastar la hipótesis:

$$- H_0 : 3 \leq \theta \leq 4$$

$$- H_1 : \theta < 3 \text{ o } \theta > 4$$

Se sabe que el E.M.V. de θ es $Y_n = \max(X_1, \dots, X_n)$. Aunque Y_n debe ser menor que θ , si el tamaño muestral es alto, existe una probabilidad alta de que Y_n esté cerca de θ .

Supongamos que H_0 es aceptada si el valor observado de $Y_n \in 2,9 \leq Y_n \leq 4$ y H_0 es rechazada si Y_n no pertenece a este intervalo.

Entonces, la Región Crítica del contraste es la región que contiene a todos los valores de X_1, \dots, X_n para los cuales $Y_n < 2,9$ ó $Y_n > 4$.

Función de Potencia

La función de potencia $\pi(\theta|\delta)$ se calcula considerando la probabilidad de que el procedimiento de contraste produzca el rechazo de H_0 . Por lo tanto:

$$\pi(\theta|\delta) = Pr(Y_n < 2,9|\theta) + Pr(Y_n > 4|\theta)$$

Cálculo de la función de potencia para el ejemplo 1

- Si $\theta \leq 2,9$, $Pr(Y_n < 2,9|\theta) = 1$ y $p(Y_n > 4|\theta) = 0 \Rightarrow \pi(\theta|\delta) = 1$
- Si $2,9 < \theta \leq 4$, $Pr(Y_n < 2,9|\theta) = (2,9/\theta)^n$ y
 $Pr(Y_n > 4|\theta) = 0. \Rightarrow \pi(\theta|\delta) = (2,9/\theta)^n$
- Si $\theta > 4$ $Pr(Y_n < 2,9) = (2,9/\theta)^n$ y
 $Pr(Y_n > 4|\theta) = 1 - (4/\theta)^n \Rightarrow \pi(\theta|\delta) = (2,9/\theta)^n + 1 - (4/\theta)^n$

Ejercicio: Graficar la función de potencia.

Tamaño del Contraste

Es la máxima probabilidad de tomar una decisión incorrecta entre todos los valores que satisfacen la hipótesis nula:

$$\Rightarrow \alpha(\delta) = \sup_{3 \leq \theta \leq 4} \pi(\theta|\delta)$$

$$\Rightarrow \alpha(\delta) = \pi(3|\delta) = (2,9/3,0)^n$$

Si por ejemplo $n = 68 \Rightarrow \alpha = 0,0997$. Entonces δ es un contraste a un nivel de significancia $\alpha_0 \geq 0,0997$

Nota: Cuando escojamos un procedimiento de prueba o contraste, debemos examinar la función de potencia. La función de potencia debe ser mayor para $\theta \in \Omega_1$ que para $\theta \in \Omega_0$. Esta debe incrementarse cuando θ se aleja de Ω_0 .

p-Valor

Es el valor más pequeño de α_0 tal que rechazaríamos la hipótesis nula a un nivel de significancia α_0 con los datos observados.

- Un investigador que rechaza la hipótesis nula sí y sólo sí el *p-Valor* si a lo sumo α_0 , está utilizando una prueba con nivel de significancia α_0 .
- Un investigador que requiere una prueba con nivel de significancia α_0 rechazará la hipótesis nula sí y sólo sí el *p-Valor* es a lo sumo α_0

Por esta razón el *p-Valor* es algunas veces llamado el **nivel de significancia observado**.

Si nuestra región crítica es de la forma: $T \geq c$ para alguna prueba estadística T . Supongamos que $T = t$. Entonces el *p-Valor* cuando $T = t$ es observado es el tamaño de la prueba δ_t :

$$\sup_{\theta \in \Omega_0} \pi(\theta | \delta_t) = \sup_{\theta \in \Omega_0} Pr(T \geq t | \theta)$$

Esta probabilidad es el área de la cola de la distribución de T la derecha de t .

Prueba de Hipótesis para el parámetro de una Bernoulli

Sea X_1, \dots, X_n una muestra aleatoria de una distribución Bernoulli con parámetro p .

Supongamos que queremos probar:

$$- H_0 : p \leq p_0 - H_1 : p > p_0$$

Sea $Y = \sum_{i=1}^n X_i$. Y tiene una distribución binomial con parámetros n y p . Si p es grande, Y también lo será.

Definimos el contraste: **Rechazamos H_0 si $Y \geq c$** para alguna constante c . Supongamos que se elige un contraste tan cercano a α_0 como sea posible pero no mayor que α_0 .

Se puede probar que $Pr(Y \geq c|p)$ es una función creciente de p . El tamaño del contraste se calcula como: $Pr(Y \geq c|p_0)$. Entonces c es el número más pequeño tal que $Pr(Y \geq c|p_0) \leq \alpha_0$.

Por ejemplo, si $n = 10$, $p = 0,3$ y $\alpha_0 = 0,1$ se puede calcular:

- $\sum_{y=6}^1 0Pr(Y = y|p = 0,3) = 0,0473$
- $\sum_{y=5}^1 0Pr(Y = y|p = 0,3) = 0,1503$

Para mantener el tamaño de la prueba con un valor máximo de 0,1, tenemos que escoger $c > 5$. Cada valor de c en el intervalo $(5, 6]$ produce el mismo valor de la prueba porque Y sólo puede tomar valores enteros.

Contrastes del cociente de verosimilitudes

Cosideremos

$$- H_0 : \theta \in \Omega_0$$

$$- H_1 : \theta \in \Omega_1$$

El estadístico

$$\Lambda((x)) = \frac{\sup_{\theta \in \Omega_0} f_n((x)|\theta)}{\sup_{\theta \in \Omega} f_n((x)|\theta)}$$

es el **cociente de verosimilitudes**. Se rechaza H_0 si $\Lambda((x)) \leq k$ para alguna constante k .

Nota: El contraste anterior rechaza H_0 si la función de verosimilitud en Ω_0 es suficientemente pequeña cuando se compara con la verosimilitud en todo Ω . Generalmente se escoge k para que la prueba tenga un nivel de significancia α_0 .

TEOREMA: Si Ω es un espacio n -dimensional y se asume que H_0 especifica un subconjunto de l coordenadas de θ fijo. Si se asume que H_0 es cierta y que la función de verosimilitud satisface que el EMV es asintóticamente normal y eficiente, entonces si $n \rightarrow \infty$, $-2\log\Lambda((X))$ converge a una distribución χ^2 con l grados de libertad.

Contraste de Hipótesis Simples

Dos tipos de Errores

Sea X_1, \dots, X_n una muestra aleatoria de una f.d.p. $f(X|\theta)$.

Se supone $\theta = \theta_0$ ó $\theta = \theta_1$. Esto implica que el espacio paramétrico consta sólo de dos puntos: θ_0 y θ_1 .

Se quiere contrastar:

$$- H_0 : \theta = \theta_0$$

$$- H_1 : \theta = \theta_1$$

Ambas hipótesis son **hipótesis simples**. Para $i = 0$ ó $i = 1$ se define:

$$f_i(\mathbf{x}) = f(x_1|\theta_i)f(x_2|\theta_i) \dots f(x_n|\theta_i)$$

Esta es la f.d.p. conjunta si H_i es cierta ($i = 0$ ó $i = 1$).

Podemos describir dos tipos de errores:

Decisión	Estados de la Naturaleza	
	H_0	H_1
Rechazo H_0	Error tipo I	✓
No Rechazo H_0	✓	Error tipo II

Sea δ un procedimiento de contraste.

Definamos:

$$\alpha(\delta) = Pr[\text{Rechazar } H_0 | \theta = \theta_0]$$

$$\beta(\delta) = Pr[\text{NoRechazar } H_0 | \theta = \theta_1]$$

Se quiere construir un procedimiento de contraste tal que $\alpha(\delta)$ y $\beta(\delta)$ sean mínimas.

Se construye entonces un procedimiento para que una combinación lineal $a\alpha(\delta) + \beta(\delta)$ sea mínimo.

Pruebas óptimas

Descripción de un procedimiento para el que el valor de una combinación lineal específica de α y β sea mínimo.

TEOREMA 1:

Sea δ^* un procedimiento de contraste tal que la hipótesis H_0 no se rechaza si $af_0(\mathbf{x}) > bf_1(\mathbf{x})$ y la hipótesis H_0 se rechaza si $af_0(\mathbf{x}) < bf_1(\mathbf{x})$. Cualquiera de las dos hipótesis puede ser rechazada o no si $af_0(\mathbf{x}) = bf_1(\mathbf{x})$. Además para cualquier otro procedimiento δ

$$a\alpha(\delta^*) + b\beta(\delta^*) < a\alpha(\delta) + b\beta(\delta)$$

Demonstración:

Caso en que la muestra aleatoria X_1, \dots, X_n proviene de una distribución discreta.

En este caso $f_i(\mathbf{x})$ es la f.d.p conjunta cuando H_i es cierta ($i = 0, 1$).

Sea R = región crítica para un procedimiento de contraste arbitrario. Entonces R contiene todos los resultados muestrales para los que δ especifica que H_0 debería ser rechazada y R^c contiene los resultados para los que H_0 debería ser aceptada.

$$\begin{aligned}
 a\alpha(\delta) + b\beta(\delta) &= a \sum_{\mathbf{x} \in R} f_0(\mathbf{x}) + b \sum_{\mathbf{x} \in R^c} f_1(\mathbf{x}) \\
 &= a \sum_{\mathbf{x} \in R} f_0(\mathbf{x}) + b \left[1 - \sum_{\mathbf{x} \in R} f_1(\mathbf{x}) \right] \\
 &= b + \sum_{\mathbf{x} \in R} [af_0(\mathbf{x}) - bf_1(\mathbf{x})]
 \end{aligned}$$

Esta última cantidad será mínima si se elige R tal que lo que está entre [] es mínimo. Para ello se elige R tal que $af_0(\mathbf{x}) - bf_1(\mathbf{x}) < 0$.
 $\Rightarrow af_0(\mathbf{x}) < bf_1(\mathbf{x})$ y no se incluyen los puntos \mathbf{x} para los cuales $af_0(\mathbf{x}) - bf_1(\mathbf{x}) > 0$. Si para algún punto \mathbf{x} $af_0(\mathbf{x}) - bf_1(\mathbf{x}) = 0$, entonces es irrelevante si $\mathbf{x} \in R$ puesto que la cantidad anterior es cero.

NOTA: El cociente de verosimilitudes de la muestra es $f_1(\mathbf{x})/f_0(\mathbf{x})$.

Por teorema anterior se afirma que un procedimiento de contraste para el cual $a\alpha(\delta) + b\beta(\delta)$ es mínimo, rechaza H_0 si $f_1(\mathbf{x})/f_0(\mathbf{x}) > a/b$ y no rechaza H_0 si $f_1(\mathbf{x})/f_0(\mathbf{x}) < a/b$

Minimización de la probabilidad de cometer un error tipo II: Lema de Neyman-Pearson

Supongamos que δ^* es un procedimiento de contraste que tiene la siguiente forma para una constante $k > 0$: No se rechaza la hipótesis H_0 si $f_0(\mathbf{x}) > kf_1(\mathbf{x})$ y se rechaza H_0 si $f_0(\mathbf{x}) < kf_1(\mathbf{x})$. No se rechaza H_0 ó H_1 si $f_0(\mathbf{x}) = kf_1(\mathbf{x})$.

Si δ es cualquier otro procedimiento de contraste tal que $\alpha(\delta) < \alpha(\delta^*)$ entonces $\beta(\delta) > \beta(\delta^*)$. Además si $\alpha(\delta) < \alpha(\delta^*)$ entonces $\beta(\delta) > \beta(\delta^*)$.

Ejemplo 1: Muestreo de una distribución Normal (Sección 9.2 DeGroot).

Ejemplo 2: Muestreo de una distribución Bernoulli (Sección 9.2 DeGroot).

Selección del Nivel de Significación

- El lema de Neyman-Pearson describe un procedimiento para el cual dado un nivel de significación α_0 , $\beta(\delta)$ es mínimo para un procedimiento δ con $\alpha(\delta) \leq \alpha_0$.
- Valores tradicionales de α_0 son: 0.10, 0.05, 0.01. Valor más común: $\alpha_0 = 0,05$.
 - Si las consecuencias del error tipo I son poco importantes $\alpha_0 = 0,10$.
 - Si las consecuencias son serias: $\alpha_0 = 0,01$
- Se escoge $\alpha_0 = 0,01$ cuando el experimentador quiere ser conservador, es decir, no rechaza H_0 a menos que los datos muestrales proporcionen fuerte evidencia para ello.
- Sin embargo, si n es grande, seleccionar $\alpha_0 = 0,01$ puede conducir a un procedimiento de contraste que rechazará H_0 para ciertas muestras que proporcionan evidencias de que H_0 es cierta.

Ejemplo

Supongamos que tenemos datos de una distribución normal con media θ desconocida y varianza 1 con:

$$- H_0 : \theta = 0$$

$$- H_1 : \theta = 1$$

Entre todos los procedimientos de contraste para los cuales $\alpha(\delta) \leq 0,01$, el valor $\beta(\delta)$ será mínimo para el procedimiento δ^* que rechaza H_0 cuando $\bar{x}_n > k'$ donde k' se elige de forma que:

$$Pr(\bar{x}_n > k' | \theta = 0) = 0,01$$

. Si $\theta = 0$ $\bar{x}_n \sim N(0, 1/n)$. Por lo tanto de la tabla normal $k' = 2,326n^{-1/2}$.

Este procedimiento es equivalente a rechazar H_0 cuando $f_1(\mathbf{x})/f_0(\mathbf{x}) > k$, donde $k = \exp(2,326n^{-1/2} - 0,5n)$.

En este caso la probabilidad de un **error tipo I** será $\alpha(\delta^*) = 0,01$.

Error tipo II: Probabilidad de no rechazar H_0 cuando H_1 es cierta.

$$\Rightarrow \beta(\delta^*) = Pr(\bar{x}_n < 2,326n^{-1/2} | \theta = 1).$$

Si $\theta = 1$

$$\bar{x}_n \sim N(1, 1/n) \Rightarrow$$

$$z' = n^{1/2}(\bar{x}_n - 1) \sim N(0, 1) \Rightarrow$$

$$\begin{aligned} \beta(\delta^*) &= Pr(n^{1/2}(\bar{x}_n - 1) < (2,326n^{-1/2} - 1)n^{1/2}) \\ &= Pr(z' < 2,326 - n^{1/2}) \\ &= \Phi(2,326 - n^{1/2}) \end{aligned}$$

Dependencia de $\alpha(\delta^*)$ y $\beta(\delta^*)$ del tamaño muestral.

n	$\alpha(\delta^*)$	$\beta(\delta^*)$	k
1	0.01	0.91	6.21
25	0.01	0.0038	0.42
100	0.01	8×10^{-15}	$2,5 \times 10^{-12}$

- Si $n = 1$ H_0 se rechazará si $f_1(\mathbf{x})/f_0(\mathbf{x}) > k = 6,21$. Esto implica que H_0 no se rechazará a menos que los valores observados x_1, \dots, x_n de la muestra sean 6.21 veces más probables con H_1 que con H_0 .
- Si $n = 100$ $\beta(\delta^*)$ es extremadamente pequeño en relación con $\alpha(\delta^*) = \alpha_0$. Por lo tanto δ^* es más conservador con respecto a un error tipo II que con respecto a un error tipo I.
- Un valor de α_0 que es apropiado para un valor pequeño de n , puede ser muy grande para un valor grande de n .

Supongamos que el experimentador considera que el error tipo I es más serio que el error tipo II y desea utilizar un procedimiento de contraste tal que $100\alpha(\delta) + \beta(\delta)$ sea mínimo.

Por teorema 1, se podría rechazar H_0 , si y sólo si, el cociente de verosimilitudes $f_1(\mathbf{x})/f_0(\mathbf{x}) > 100$ independientemente del tamaño muestral. Entonces el procedimiento que minimiza $100\alpha(\delta) + \beta(\delta)$ no rechazará H_0 a menos que los valores observados x_1, \dots, x_n sean 100 veces más probables con H_1 que con H_0 .

Parece razonable minimizar una combinación lineal de la forma $a\alpha(\delta) + b\beta(\delta)$ en lugar de fijar un valor de $\alpha(\delta)$ y minimizar $\beta(\delta)$. Desde un punto de vista Bayesiano es natural utilizar esta alternativa.

Procedimiento de contraste de Bayes

Sea X_1, \dots, X_n una muestra aleatoria de una distribución cuya f.d.p. ó f.p. es $f(\mathbf{x}|\theta)$.

Se desean contrastar las hipótesis simples:

- $H_0 : \theta = \theta_0$

- $H_1 : \theta = \theta_1$

Sea:

- d_0 = Decisión de no rechazar H_0

- d_1 = Decisión de rechazar H_0

Sea w_0 : Pérdida cuando se elije d_1 y H_0 es correcta

Sea w_1 = Pérdida cuando se elije d_0 y H_1 es correcta.

Sea $L(\theta_i, d_j)$ = Función de pérdida cuando θ_i es el verdadero valor y se elige d_j , $j = 0, 1$.

$L(\theta_i, d_j)$	d_0	d_1
θ_0	0	w_0
θ_1	w_1	0

Sea ξ_0 = Probabilidad inicial de que H_0 sea cierta

Sea ξ_1 = Probabilidad inicial de que H_1 sea cierta = $1 - \xi_0$.

Pérdida esperada de cualquier procedimiento de contraste δ

$$r(\delta) = \xi_0 E(P'erdida|\theta = \theta_0) + \xi_1 E(P'erdida|\theta = \theta_1)$$

$$E(P'erdida|\theta = \theta_0) = w_0\alpha(\delta)$$

$$E(P'erdida|\theta = \theta_1) = w_1\beta(\delta)$$

$$\Rightarrow r(\delta) = \xi_0 w_0 \alpha(\delta) + \xi_1 w_1 \beta(\delta)$$

Un procedimiento que minimiza esta pérdida esperada $r(\delta)$ se denomina procedimiento de **contraste de Bayes**.

Por teorema anterior un procedimiento de contraste de Bayes no rechazará H_0 si

$$\xi_0 w_0 f_0(\mathbf{x}) > \xi_1 w_1 f_1(\mathbf{x})$$

y rechazará H_0 si

$$\xi_0 w_0 f_0(\mathbf{x}) < \xi_1 w_1 f_1(\mathbf{x})$$

Cualquiera de las dos hipótesis puede ser rechazada o no si:

$$\xi_0 w_0 f_0(\mathbf{x}) = \xi_1 w_1 f_1(\mathbf{x})$$

Pruebas basadas en la distribución a posterior

En este caso se trata de minimizar la pérdida esperada a posterior.

Volvamos a la situación general donde la hipótesis nula es $H_0 : \theta \in \Omega_0$ y la hipótesis alternativa es $H_1 : \theta \in \Omega_1$.

Sea nuevamente:

- d_0 = Decisión de no rechazar H_0
- d_1 = Decisión de rechazar H_0

Sea w_0 : Pérdida cuando se elije d_1 y H_0 es correcta

Sea w_1 = Pérdida cuando se elije d_0 y H_1 es correcta.

La función de pérdida $L(\theta, d_i)$ puede ser resumida de la siguiente forma:

	d_0	d_1
Si H_0 es cierta	0	w_0
Si H_1 es cierta	w_1	0

Sea $\xi(\theta|\mathbf{x})$ la distribución a posterior de θ .

La pérdida esperada a posterior al escoger la decisión $d_i (i = 0, 1)$ es:

$$r(d_i|(x)) = \int L(\theta, d_i)\xi(\theta|\mathbf{x})d\theta$$

Podemos escribir las siguientes fórmulas para la pérdida esperada a posterior para cada $i = 0, 1$.

$$r(d_0|\mathbf{x}) = \int_{\Omega_1} w_1 \xi(\theta|\mathbf{x}) d\theta = w_1 [1 - Pr(H_0 \text{ verdadera}|\mathbf{x})]$$

$$r(d_1|\mathbf{x}) = \int_{\Omega_0} w_0 \xi(\theta|\mathbf{x}) d\theta = w_0 [Pr(H_0 \text{ verdadera}|\mathbf{x})]$$

El procedimiento de Bayes escoge la decisión que tenga la menor pérdida esperada a posterior, es decir, escoge d_0 si $r(d_0|\mathbf{x}) < r(d_1|\mathbf{x})$; escoge d_1 si $r(d_0|\mathbf{x}) \geq r(d_1|\mathbf{x})$.

La desigualdad $r(d_0|\mathbf{x}) \geq r(d_1|\mathbf{x})$ puede ser reescrita como:

$$Pr(H_0 \text{ verdadera}|\mathbf{x}) \leq \frac{w_1}{w_0 + w_1}$$

El procedimiento de prueba que rechaza H_0 cuando se cumple la desigualdad anterior es la prueba de Bayes en todas las situaciones en las que la función de pérdida está dada por la tabla anterior.

Ejemplo: Ejemplo 9.8.3 DeGroot & Schervish.

Modelos Lineales: El método de mínimos cuadrados

Método para calcular los coeficientes de una función lineal para predecir una variable y en función de otras variables x_1, x_2, \dots, x_n .

TEOREMA:

Sean $(x_1, y_1), \dots, (x_n, y_n)$ un conjunto de n puntos. La línea recta que minimiza la suma de cuadrados de las desviaciones verticales de todos los puntos con respecto a la línea tiene la siguiente pendiente e intercepto:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Demonstración

La suma de cuadrados de las distancias verticales en los n puntos es:

$$Q = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Minimizamos la ecuación anterior con respecto a β_0 y β_1 y resultan las **ecuaciones normales** para β_0 y β_1 . (DeGroot y Schervish, sección 11.1)

Línea de mínimos cuadrados

Es la ecuación definida por:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Ajuste de un polinomio por el método de mínimos cuadrados

En este caso se quiere minimizar:

$$Q = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i + \dots + \beta_k x_i^k)]^2$$

El polinomio de mínimos cuadrados es:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \dots + \hat{\beta}_k x^k$$

Ajuste de una función lineal de varias variables

Cada valor y_i tienen mediciones x_{i1}, \dots, x_{ik} .

En este caso la función lineal tiene la forma:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Se quiere minimizar:

$$Q = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})]^2$$

El función lineal de mínimos cuadrados es:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

Regresión Lineal

En este caso los valores de y son valores observados de una colección de variables aleatorias. En este caso hay un modelo estadístico y el método de mínimos cuadrados produce los estimadores de máxima verosimilitud de los parámetros del modelo.

Respuesta/Predictor/Regresión

Las variables X_1, \dots, X_k son llamadas **predictores**; la variable aleatoria Y es la variable de **respuesta**; la esperanza condicional de Y dados los valores observados x_1, \dots, x_k de X_1, \dots, X_k es la **función de regresión**.

La función de regresión tiene la forma:

$$E(Y|x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Los coeficientes β_0, \dots, β_k son los **coeficientes de regresión**. Estos coeficientes son desconocidos y deben ser estimados.

Los valores $\hat{\beta}_0, \dots, \hat{\beta}_k$ obtenidos por el método de mínimos cuadrados, son los estimadores de mínimos cuadrados.

Regresión Lineal Simple

La variable aleatoria Y puede ser representada de la siguiente forma:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

ε es una variable aleatoria con distribución normal, media 0 y varianza σ^2 .

$E[Y|x]$ es una función lineal de los parámetros β_0 y β_1 . Observamos los pares $(x_1, Y_1), \dots, (x_n, Y_n)$

Suposiciones:

- El predictor es conocido
- Normalidad
- Media Lineal
- Varianza común (homocedasticidad)
- Independencia

Nota: Las suposiciones se pueden generalizar en el caso de tener más de un predictor

Distribución condicional conjunta de Y_1, \dots, Y_n dado el vector $\mathbf{x} = (x_1, \dots, x_n)$:

$$f_n(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right]$$

TEOREMA:

Estimadores de Máxima Verosimilitud de β_0, β_1 y σ^2 . Los EMV de β_0, β_1 son los estimadores de mínimos cuadrados; y el EMV de σ^2 es:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Demonstración: DeGroot y Schervish Sección 11.2

Distribución de los estimadores de mínimos cuadrados

Distribución conjunta de los estimadores $\hat{\beta}_0, \hat{\beta}_1$

Definimos $s_x = (\sum_{i=1}^n (x_i - \bar{x})^2)^{1/2}$

TEOREMA

Bajo las suposiciones anteriores, las distribuciones de $\hat{\beta}_0, \hat{\beta}_1$ son:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}\right)\right) \quad (4)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_x^2}\right) \quad (5)$$

Finalmente la covarianza entre $\hat{\beta}_0$ y $\hat{\beta}_1$ es:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{s_x^2}$$

Esto implica que $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados de β_0, β_1 .

Predicción

Se desea predecir el valor de una observación independiente Y para un valor determinado de x . El predictor natural es:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

TEOREMA

El error cuadrático medio (ECM) de esta predicción es:

$$E[(\hat{Y} - Y)^2] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_x^2} \right]$$

De esta ecuación se observa que el ECM aumenta cuando x se aleja de \bar{x} .

TEOREMA:

La distribución conjunta de $(\hat{\beta}_0, \hat{\beta}_1)$ es una normal bivariada con medias, varianzas y covarianzas definidas anteriormente. También si $n \geq 3$ $\hat{\sigma}^2$ es independiente de $(\hat{\beta}_0, \hat{\beta}_1)$ y $n\hat{\sigma}^2/\sigma^2$ tiene una distribución χ^2 con $n - 2$ grados de libertad.

Inferencia Bayesiana en la Regresión Lineal Simple

Sea $\tau = 1/\sigma^2$.

Necesitamos proponer una distribución a priori para β_0, β_1, τ .

Una opción es suponer una previa impropia: $\xi(\beta_0, \beta_1, \tau) = 1/\tau$

El objetivo es encontrar la distribución posterior de los parámetros β_0, β_1, τ

TEOREMA:

Condicional en τ la distribución a posterior conjunta de β_0, β_1 es una normal bivariada con correlación $-n\bar{x}/(n \sum_{i=1}^n x_i^2)^{1/2}$ con medias y varianzas dadas en la tabla 11.10.

La distribución a posterior de τ es una gamma con parámetros $(n-2)/2$ y $S^2/2$ donde $S^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

La distribución posterior marginal de $c_0\beta_0 + c_1\beta_1$ es una t-student con media $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$, varianza = $\frac{S^2}{(n-2)} / [\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2}]$ y $n-2$ grados de libertad.