



Instituto Nacional de Estadística

OPOSICIONES AL CUERPO SUPERIOR DE
ESTADÍSTICOS DEL ESTADO

BOE NÚM. 270, DE 12 DE OCTUBRE DE 2020, PÁG. 87165

Inferencia y modelización estadísticas

Grupo de Materias Comunes

INE
10 de agosto de 2021

Índice general

1	Propiedades de una muestra aleatoria	1
1.1	Conceptos básicos de muestras aleatorias	1
1.1.1	Introducción	1
1.1.2	Distribución de la población. Muestra aleatoria simple. Inferencia estadística.	3
1.1.3	Inferencia estadística paramétrica y no paramétrica	6
1.1.4	Estadístico muestral	10
1.1.5	Distribución empírica de la muestra	11
1.1.6	Momentos muestrales	14
1.1.7	Cuantiles muestrales	18
1.2	Sumas de variables aleatorias de una muestra aleatoria	20
1.3	Muestreo de una distribución normal: propiedades de la media y varian-za muestrales y las distribuciones t de Student y F de Snedecor.	20
1.3.1	Introducción	20
1.3.2	Distribuciones asociadas a la normal	21
1.3.3	Distribuciones en el muestreo para una población normal	25
1.3.4	Distribuciones en el muestreo para dos poblaciones normales	27
1.4	Estadísticos de orden	30
2	Principios de reducción de datos	1
2.1	Introducción	1
2.2	Estadísticos suficientes	2
2.3	Estadísticos auxiliares y completos	8
2.3.1	Estadísticos auxiliares	8
2.3.2	Estadísticos completos	9
2.3.3	Familias paramétricas de tipo exponencial	11
2.4	El principio de verosimilitud	13
2.5	El principio de equivarianza	14
3	Estimación puntual I	1
3.1	Introducción	1
3.2	Estimadores puntuales	2
3.3	Método de los momentos para la obtención de estimadores	3
3.3.1	Método de máxima verosimilitud	5
3.4	Estimadores de Bayes	12
3.4.1	Introducción a la inferencia bayesiana	12
3.4.2	Estimación puntual bayesiana	20
4	Estimación puntual II	1
4.1	Introducción	1
4.2	Estimadores consistentes e insesgados	1
4.3	Error cuadrático medio como medida de la bondad de un estimador	4

4.4	Estimador centrado de uniformemente mínima varianza	8
4.5	Cota de Frechet-Cramer-Rao para la varianza de un estimador	11
4.6	Perdida final esperada como medida de la bondad de un estimador bayesiano	12
5	Tests de hipótesis	1
5.1	Introducción	1
5.2	Conceptos básicos	2
5.2.1	Hipótesis nula y alternativa, simple o compuesta	2
5.2.2	Definición de contraste de hipótesis. Región crítica y región de aceptación.	2
5.2.3	Errores de tipo I y II	3
5.2.4	Nivel de significación	3
5.2.5	Nivel crítico o p -valor	5
5.3	Contrastes de hipótesis paramétricos	8
5.3.1	Contrastes unilaterales y bilaterales	8
5.3.2	Función de potencia	8
5.3.3	Contraste de hipótesis simple frente a simple	9
5.4	Teoría de contrastes paramétricos óptimos	12
5.4.1	Contrastes unilaterales	12
5.4.2	Contrastes bilaterales	14
5.5	Contraste de razón de verosimilitudes	14
5.6	Contrastes para poblaciones normales	26
5.6.1	Distribución teórica $N(\mu, \sigma)$ y muestra de tamaño n	26
5.6.2	Dos poblaciones normales independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ y muestras de tamaños n y m respectivamente	27
5.6.3	Una muestra pareada de tamaño n	28

Índice general

6	Estimación por intervalos I	1
6.1	Intervalos de confianza. Introducción	1
6.2	Método de la función pivote o cantidad pivotal	2
6.2.1	Intervalo de confianza para la media de una distribución normal con varianza conocida	2
6.2.2	Intervalo de confianza para la media de una normal con varianza desco- nocida	4
6.2.3	Intervalo de confianza para la varianza de una normal conocida la media	6
6.2.4	Intervalo de confianza para la varianza de una normal. Media desconocida	7
6.2.5	Intervalo de confianza para la diferencia de medias en poblaciones normales. Varianzas conocidas	8
6.2.6	Intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas desconocidas pero iguales	9
6.2.7	Intervalo de confianza para el cociente de varianzas en poblaciones normales	10
6.3	Método general o de Neyman	11
6.4	Intervalos para parámetros de distribuciones discretas	15
6.4.1	Intervalo de confianza aproximado para p en una población $B(1, p)$ si n es grande ($n \geq 30$).	15
6.4.2	Intervalo de confianza aproximado para λ en el modelo de Poisson . . .	16
6.4.3	Intervalo de confianza para la diferencia de proporciones	17
6.5	Inversión del estadístico de un test	18
6.6	Inferencia bayesiana. Intervalos bayesianos	20
6.6.1	Familias conjugadas de distribuciones	20
6.7	Intervalos de confianza bayesianos	23
7	Estimación por intervalos II	1
7.1	Introducción	1
7.2	Métodos para evaluar estimadores de intervalos: tamaño y probabilidad de cobertura	1
7.3	Intervalo de confianza de longitud más corta	4
7.4	Intervalos de confianza insesgados y equivariantes	9
7.5	Optimalidad bayesiana. Optimalidad de función de pérdida	13

Tema 1

Propiedades de una muestra aleatoria. Conceptos básicos de muestras aleatorias. Sumas de variables aleatorias de una muestra aleatoria. Muestreo de una distribución normal: propiedades de la media y varianzas muestrales y las distribuciones t de Student y F de Snedecor. Estadísticos de orden.

Los contenidos de este tema corresponden a un curso básico de Estadística Matemática, y pueden ser consultados y completados por ejemplo con los libros de la bibliografía. El libro de Rohatgi y Ehsanes es el más completo y también con una mayor formalización matemática.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

1.1 Conceptos básicos de muestras aleatorias

1.1.1 Introducción

Si se desea obtener información sobre una población que no puede ser estudiada exhaustivamente, resulta útil obtener una muestra, una parte de la población, que sea representativa de alguna manera del colectivo completo. Se considera alguna característica de interés que se desea estudiar, y se obtienen datos para cada uno de los elementos de la muestra. Con estos datos observados, mediante un tratamiento adecuado, se puede obtener información sobre la característica de interés.

Consideremos por ejemplo la población de los árboles frutales en una plantación. Se desea hacer una valoración sobre cuál será la cosecha recogida, antes de recogerla en su totalidad y pesarla. Para ello, se pueden cosechar en primer lugar algunos de los árboles y pesar la fruta, y se pueden hacer cálculos aproximados para predecir la cosecha total basados en los datos obtenidos con los árboles de esta muestra (por ejemplo mediante proporcionalidad del número de árboles en la muestra y en la población).

Resulta conveniente que el método de selección de la muestra sea aleatorio (probabilístico), con una distribución de probabilidad preestablecida, para que no intervengan en ello elementos subjetivos. Esto permite extrapolar las propiedades de la muestra a la población completa, y permite también medir el grado de error que podemos estar cometiendo en la extrapolación. Una vez obtenida la muestra aleatoria se pueden obtener, por ejemplo, valores aproximados, estimaciones, para algunas características de la población, como la cosecha total del ejemplo, la media o la varianzas, y se puede

medir el grado de error que podemos estar cometiendo en las estimaciones. En adelante, cuando se haga referencia al muestreo o a la muestra, se sobreentiende que son aleatorios aunque no se mencione.

En una población finita puede considerarse muestreo (aleatorio) con reemplazamiento, en el que las observaciones muestrales son independientes, o muestreo sin reemplazamiento, donde no se cumple esta independencia, pero que evita repeticiones de las observaciones, lo que lleva a una ganancia en precisión en las estimaciones (menor error de muestreo). Además, puede ser útil considerar otros diseños muestrales, como el muestreo estratificado o el muestreo por conglomerados, para conseguir una mayor ganancia en precisión y otros objetivos. Un campo destacado de aplicación de los procedimientos de estimación que se desarrollan para estas técnicas de muestreo es el de los estudios demográficos y sociales sobre poblaciones humanas.

En otras situaciones se consideran muestras asociadas a repeticiones de un experimento aleatorio en los que no hay una población física de la que se seleccionan individuos, sino un conjunto de posibles observaciones, al que también se denomina población. Por ejemplo, el conjunto de posibles mediciones obtenidas con un aparato de medida sobre un objeto, con errores de medida aleatorios, o el conjunto de posibles tiempos entre llegadas a un detector de dos partículas sucesivas en la desintegración radiactiva de un determinado material.

En estos casos, las mediciones son independientes e idénticamente distribuidas (antes de realizar el experimento) si el mecanismo generador que las produce permanece en idénticas condiciones de incertidumbre en las sucesivas observaciones. Así ocurre en los ejemplos anteriores si el aparato de medida y el detector se mantienen en condiciones fijas (ajustes del aparato, temperatura, etc), supuesto que el objeto y el material no se modifican (despreciando que haya alguna partícula menos en el material). En muchos casos la distribución con la que se generan los datos no es conocida, y se desea estudiar esta distribución y sus características (esperanza y varianza del tiempo entre llegadas de las partículas por ejemplo).

Los datos ofrecen información sobre el mecanismo que los ha generado. La Inferencia Estadística, o Estadística Matemática, se ocupa del estudio de la distribución de probabilidad a partir de la cual se han producido los datos observados, del mecanismo aleatorio generador de esos datos, y de sus características. Dedicaremos los próximos temas a su estudio básico.

En este tema presentaremos algunas propiedades básicas de las muestras aleatorias y de funciones de ellas. La “distribución empírica de la muestra” y sus propiedades permiten una primera aproximación estructurada al estudio de la distribución que genera los datos (distribución teórica) y sus características numéricas, como momentos y cuantiles: es posible estudiar estas características mediante las correspondientes características de la distribución empírica. Esto proporciona unas posibles herramientas para ese propósito de realizar inferencias, pero falta concretar el problema y desarrollarlo, lo que se hará en los próximos temas.

También presentaremos en este tema algunas distribuciones de probabilidad asociadas

al muestreo en poblaciones normales: la distribución χ^2 , la t de Student y la F de Snedecor. Se dice que una muestra corresponde a una población normal si la distribución (antes del muestreo) de los datos generados por las repeticiones del experimento aleatorio es normal $N(\mu, \sigma)$. Si el valor de uno o los dos parámetros μ y σ es desconocido, se puede estudiar su valor mediante las técnicas de inferencia que desarrollaremos en los próximos temas, y las distribuciones asociadas a la normal antes mencionadas serán útiles para ello.

Con estos apuntes se pretende facilitar la asimilación de las nociones básicas sobre inferencia estadística. Se han extraído elementos de todos los libros de la bibliografía. Los comentarios son abundantes, en la línea de (VelGar12), con el propósito de ofrecer al lector un tránsito de entrada suave a esta disciplina. No se presentan la mayoría de las demostraciones, que pueden ser consultadas en la bibliografía, aunque sí algunas de resultados destacados, y se dan indicaciones para otras que se obtienen de un modo sencillo.

Las herramientas del Cálculo de Probabilidades, o Teoría de la Probabilidad, son fundamentales para realizar este estudio, que está basado en la suposición de que estamos observando realizaciones de variables aleatorias. Para una adecuada asimilación de los conceptos que se desarrollarán se requieren conocimientos de probabilidad al nivel de un curso básico sobre variable aleatoria unidimensional y multidimensional, y sobre convergencia de sucesiones de variables aleatorias.

1.1.2 Distribución de la población. Muestra aleatoria simple. Inferencia estadística.

La *Inferencia Estadística*, o *Estadística Matemática*, desarrolla métodos para obtener información acerca de la ley de probabilidad de un fenómeno aleatorio mediante la observación del mismo.

Consideremos una variable aleatoria X asociada a un experimento aleatorio, que corresponde a una cierta medición asociada al resultado del experimento aleatorio. Supondremos principalmente que X es unidimensional, aunque también se considerará algún caso bidimensional.

Antes de realizar el experimento, X es una variable aleatoria con su estructura de probabilidad, dada por ejemplo mediante su función de distribución F , o también por su función de masa o densidad f , en los casos discreto o continuo. Después de realizar el experimento se obtiene un valor observado (realización, dato) x , y decimos que X ha tomado el valor x , denotado por $X = x$.

Es habitual en la práctica que F sea desconocida, al menos parcialmente, y se desea obtener información sobre ella, sobre el mecanismo aleatorio que genera los datos. Desarrollaremos métodos de inferencia estadística para obtener información sobre F a partir de una muestra.

La *distribución teórica* o *distribución de la población* es la distribución real de la variable aleatoria de interés asociada al fenómeno, fija pero desconocida. La denotaremos por F . Pretendemos obtener información sobre F .

El símbolo F hará referencia, según el contexto, a la distribución de probabilidad teórica o a su función de distribución, cuando se opere con ella. Abreviamos variable aleatoria como v.a.

Ejemplo 1. Supongamos que lanzamos una moneda mediante un dispositivo mecánico, de modo que el resultado es independiente de la posición inicial de la moneda antes del lanzamiento. Sea X el resultado del lanzamiento, que sabemos que puede tomar los valores cara o cruz. Representamos cara mediante $x = 1$ y cruz mediante $x = 0$ para que el resultado sea numérico; no es necesario, pero simplifica algunas expresiones, la de $f(x)$ por ejemplo, mas abajo, porque permite definir X como “número de caras” en el lanzamiento de la moneda.

En este caso, la distribución de X es parcialmente conocida, es la distribución de Bernoulli, *Bernoulli* (p) o $B(1, p)$ (binomial), para algún valor del parámetro p . Su función de distribución es

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - p & \text{si } 0 \leq x < 1 \\ 1 & \text{si } 1 \leq x \end{cases},$$

y su función de masa es $f(x) = p^x (1 - p)^{1-x}$ para $x = 0, 1$, siendo $f(x) = P\{X = x\}$. En este caso lo único desconocido es el valor del parámetro p de la $B(1, p)$, que será un valor en el intervalo $(0, 1)$, $0 < p < 1$. Entonces, la información sobre F se obtiene obteniendo información sobre p .

Cuando se habla genéricamente de la distribución teórica es habitual utilizar el símbolo F , tanto para la función de distribución como para la correspondiente distribución de probabilidad, pero cuando se utiliza en expresiones, la distribución de probabilidad se denota por P .

Denominamos *soporte* de P , o de F , al mínimo conjunto cerrado que tiene probabilidad 1: si C , cerrado, es el soporte de P , entonces $P(C) = P\{X \in C\} = 1$ y $P(C') < 1$ para todo cerrado $C' \subset C$ con $C' \neq C$.

Para las distribuciones discretas, el soporte es el conjunto de valores que tienen probabilidad positiva, por ejemplo, el conjunto $\{0, 1\}$ para la distribución de Bernoulli y $\{0, 1, 2, \dots\}$ para la distribución de Poisson.

Para las distribuciones continuas (con función de densidad, absolutamente continuas respecto de la medida de Lebesgue), el soporte es un intervalo cerrado acotado o no, caso habitual, o una unión finita o numerable de ellos. Para la distribución beta el soporte es el intervalo $[0, 1]$, para la distribución gamma la semirrecta $[0, \infty)$, y para la normal \mathbb{R} . Puesto que la frontera de estos cerrados se limita a conjuntos finitos de puntos, que tienen probabilidad 0, podemos ampliar el uso del término soporte también a los abiertos correspondientes: $(0, 1)$ para la beta y $(0, \infty)$ para la gamma por ejemplo.

Para obtener información sobre la distribución teórica F se llevan a cabo repeticiones independientes del experimento aleatorio, obteniéndose unos valores (realizaciones del

experimento)

$$x_1, \dots, x_n.$$

Con la Estadística Descriptiva se manejan estos datos por si mismos, sin introducir la suposición de una distribución F subyacente a los datos.

En la Estadística Matemática (inferencia) se supone que las observaciones x_1, \dots, x_n son realizaciones de una variable aleatoria con distribución F desconocida. Además F es fija, característica del fenómeno aleatorio es cuestión (salvo en un enfoque bayesiano, que introduciremos en el tema 3).

De este modo, se considera que los datos están relacionados con una estructura de probabilidad, y por ello se pueden utilizar las herramientas de cálculo de probabilidades para analizarlos. En primer lugar, introducimos las variables aleatorias X_1, \dots, X_n de las que suponemos que x_1, \dots, x_n son una realización. Abreviaremos con “vaid” la expresión “variables aleatorias independientes e idénticamente distribuidas”.

Definición 2. Una *muestra aleatoria simple de tamaño n* de una variable aleatoria X con distribución F , es una secuencia de n vaid

$$X_1, \dots, X_n$$

con distribución común F . Lo abreviamos como *m.a.s.* También a una realización x_1, \dots, x_n de X_1, \dots, X_n se la denomina m.a.s. de X , o de la distribución F , cuando la identificación de una u otra, variable aleatoria (X_1, \dots, X_n) o realización muestral (x_1, \dots, x_n) , sea clara por el contexto.

Expresaremos la muestra indiferentemente con paréntesis, cuando queramos indidir en el carácter unitario de la v.a. n -dimensional (X_1, \dots, X_n) , o sin él. Cometiendo abuso de notación, denotamos por F también a la función de distribución conjunta $F(x_1, \dots, x_n)$, y por $f(x_1, \dots, x_n)$ a la función de masa o densidad teórica conjunta (en los casos discreto o continuo respectivamente, lo que en adelante se sobreentenderá cuando se hable de f). Los argumentos x o (x_1, \dots, x_n) que utilicemos especificarán a qué función nos referimos en cada caso.

Se denomina *muestreo* al procedimiento y al proceso de obtención de la muestra x_1, \dots, x_n .

El procedimiento de muestreo de la definición, con vaid, es el que seguiremos principalmente, y cuando se haga referencia a una muestra se entenderá que es una m.a.s. Para poblaciones finitas, las observaciones también son vaid cuando se obtienen mediante un muestreo aleatorio con reemplazamiento.

Denotamos por Ω al *soporte de (X_1, \dots, X_n)* , que incluye al conjunto de posibles realizaciones muestrales, y es el producto cartesiano del soporte de X por si mismo n veces, con $\Omega \subset \mathbb{R}^n$. Para una distribución teórica beta, $\Omega = [0, 1]^n$, y para una población normal, $\Omega = \mathbb{R}^n$. Los elementos de Ω son de la forma (x_1, \dots, x_n) .

Para todo $(x_1, \dots, x_n) \in \Omega$ se tiene que

$$F(x_1, \dots, x_n) = F(x_1) \cdots F(x_n) \quad \text{y}$$

$$f(x_1, \dots, x_n) = f(x_1) \cdots f(x_n) .$$

La igualdad para la función de distribución se obtiene de un modo inmediato a partir de la independencia de X_1, \dots, X_n . También la igualdad para la función de masa en el caso discreto. Para la función de densidad en el caso continuo, en ambos miembros de la igualdad para la función de distribución considérese la derivada de orden n respecto de x_1, \dots, x_n .

(Téngase en cuenta que en el caso continuo f solo está definida casi seguro, de modo que puede ser modificada en un conjunto con probabilidad 0 sin cambio en la distribución correspondiente, y entonces la igualdad para f es solo casi segura. Pero esta es una cuestión que no afecta a los desarrollos que se realizarán y que no se volverá a mencionar).

1.1.3 Inferencia estadística paramétrica y no paramétrica

En algunos casos la distribución teórica F es totalmente desconocida, y P podría ser cualquier distribución de probabilidad sobre (\mathbb{R}, \mathbb{B}) , si X es unidimensional, siendo \mathbb{B} la σ -álgebra de Borel sobre \mathbb{R} . (Considerar como sucesos solamente a los elementos de una σ -álgebra proporciona la consistencia necesaria para poder definir una distribución de probabilidad. La σ -álgebra \mathbb{B} consta de una amplia clase de sucesos que incluye los intervalos y la unión finita o numerable de ellos).

En otras ocasiones hay alguna información sobre F que restringe la clase de las posibles distribuciones teóricas. Se podría conocer que F es discreta, o continua, o simétrica. Para los casos sin información o con esta información limitada, la Inferencia Estadística No Paramétrica se ocupa del estudio de F y sus características. Este nombre en estos términos se utiliza para contraponerlo al caso paramétrico, en el que tenemos información sobre F de tal modo que sabemos que está restringida a alguna familia paramétrica, como comentamos a continuación.

En ocasiones es conocido que F pertenece a alguna familia paramétrica de distribuciones, como la familia de las distribuciones de Bernoulli $B(1, \theta)$ para $0 < \theta < 1$, y entonces el único elemento de incertidumbre sobre F es el valor del parámetro θ , que especifica F con unicidad. Del estudio de F en este caso se ocupa la Inferencia Estadística Paramétrica.

Supongamos que sabemos que F es una distribución con forma funcional fija y conocida, de modo que conocemos la expresión concreta de F (o la de f) salvo el valor que toma un parámetro θ , que se sabe que debe ser algún valor $\theta \in \Theta$ en un subconjunto $\Theta \subset \mathbb{R}^k$, denominado *espacio paramétrico*. Se dice en este caso que la distribución teórica F es k -paramétrica, con $k \geq 1$ parámetros, o simplemente que F es paramétrica.

La distribución específica F correspondiente a un valor $\theta \in \Theta$ del parámetro se suele denotar como F_θ . Y f como f_θ , y la distribución de probabilidad teórica P como P_θ , y la

esperanza y varianza E y V para variables aleatorias funciones de la muestra como E_θ y V_θ . Con esta notación se incide en el hecho de que los cálculos hechos con esos elementos tienen resultados que son función de θ , que dependen de cuál sea el verdadero valor de θ .

Entonces, en este caso paramétrico sabemos que la distribución teórica F pertenece a una determinada familia paramétrica:

$$F \in \{F_\theta : \theta \in \Theta\}, \text{ con } \Theta \subset \mathbb{R}^k.$$

Por ejemplo, si la distribución teórica es de Bernoulli, $B(1, \theta)$, el espacio paramétrico es el intervalo

$$\Theta = (0, 1) \subset \mathbb{R},$$

con dimensión $k = 1$, y

$$F \in \{B(1, \theta) : 0 < \theta < 1\}.$$

Si la distribución teórica es normal $N(\mu, \sigma)$ con ambos parámetros desconocidos, el espacio paramétrico es el producto cartesiano

$$\Theta = \mathbb{R} \times (0, \infty) \subset \mathbb{R}^2,$$

puesto que $-\infty < \mu < \infty$ y $0 < \sigma < \infty$, con dimensión $k = 2$, y

$$F \in \{N(\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}.$$

El estudio de la distribución teórica se reduce al estudio del verdadero valor de θ .

Algunos casos en los que la distribución teórica es paramétrica son los siguientes.

Cuando se clasifican los resultados de un experimento aleatorio dos clases, denominadas éxito y fracaso por ejemplo, la distribución teórica es de Bernoulli.

Cuando se observa el número de éxitos en pruebas de Bernoulli independientes, la distribución teórica es binomial.

La distribución del número de llegadas en un tiempo dado en un proceso de Poisson, como por ejemplo en las llegadas a un detector en una desintegración radiactiva, tiene distribución de Poisson. La distribución del tiempo entre llegadas es exponencial, que es la única distribución continua que "no tiene memoria", esto es, que cumple $P\{X > x + t | X > x\} = P\{X > t\}$ para $x, t > 0$, siendo X el tiempo entre llegadas.

La distribución normal es frecuente por ejemplo asociada a todo tipo de mediciones (como longitud, resistencia eléctrica, concentración química, biométricas). Y también funciones de ella, como la distribución lognormal, que es la distribución de e^Y para Y normal.

A continuación recordamos la expresión de f_θ y su soporte para algunas distribuciones paramétricas conocidas, el espacio paramétrico, los momentos $\mu = E_\theta[X]$ y $\sigma^2 =$

$V_\theta[X] = E_\theta[(X - \mu)^2]$, y presentamos la expresión de $f_\theta(x_1, \dots, x_n)$ (con el abuso de notación habitual para $f = f_\theta$) dejando su obtención como ejercicio.

Para el caso discreto, f_θ es la función de masa de probabilidad (también llamada función de probabilidad o función de densidad de probabilidad en algunos textos), y para el caso continuo f_θ es la función de densidad de probabilidad, tanto para la distribución teórica como para la de la muestra. Se presenta el valor de f_θ en su soporte, para x y para (x_1, \dots, x_n) , y se sobreentiende que f_θ es nula fuera de su soporte aunque no se mencione.

- Para la distribución teórica de Bernoulli $B(1, \theta)$ se tiene que $f_\theta(x) = \theta^x (1 - \theta)^{1-x}$ para $x = 0, 1$, con $\Theta = (0, 1)$. Se tiene que $\mu = \theta$ y $\sigma^2 = \theta(1 - \theta)$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = \theta^{x_1 + \dots + x_n} (1 - \theta)^{n - (x_1 + \dots + x_n)} = \theta^t (1 - \theta)^{n-t}$$

si $x_1, \dots, x_n = 0, 1$, siendo $t = \sum_{i=1}^n x_i$.

- Para una distribución teórica de Poisson $\mathcal{P}(\theta)$ se tiene que $f_\theta(x) = e^{-\theta} \frac{\theta^x}{x!}$ si $x = 0, 1, 2, \dots$, con $\Theta = (0, \infty)$. Se tiene que $\mu = \theta$ y $\sigma^2 = \theta$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = e^{-n\theta} \frac{\theta^t}{x_1! \dots x_n!}$$

si $x_1, \dots, x_n = 0, 1, 2, \dots$, siendo $t = \sum_{i=1}^n x_i$.

- Para una distribución teórica geométrica $Geom(\theta)$ se tiene que $f_\theta(x) = \theta(1 - \theta)^{x-1}$ si $x = 1, 2, \dots$, con $\Theta = (0, 1)$. Se tiene que $\mu = \frac{1}{\theta}$ y $\sigma^2 = \frac{1-\theta}{\theta^2}$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = \theta^n (1 - \theta)^{t-n}$$

si $x_1, \dots, x_n = 1, 2, \dots$, siendo $t = \sum_{i=1}^n x_i$. (Corresponde al número de repeticiones hasta el primer éxito en pruebas de Bernoulli, aunque en ocasiones se considera el número de fracasos antes del primer éxito, una unidad menor, con una expresión de $f_\theta(x)$ algo diferente).

- Para una distribución teórica normal $N(\mu, \sigma)$ se tiene que $f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$ para $-\infty < x < \infty$, con $\theta = (\mu, \sigma)$ y $\Theta = \mathbb{R} \times (0, \infty)$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = (2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}$$

si $-\infty < x_1, \dots, x_n < \infty$.

- Para una distribución teórica exponencial $Exp(\theta)$ se tiene que $f_\theta(x) = \theta e^{-\theta x}$ para $x > 0$, con $\Theta = (0, \infty)$. Se tiene que $\mu = 1/\theta$ y $\sigma^2 = 1/\theta^2$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = \theta^n e^{-\theta t}, \quad \text{con } t = \sum_{i=1}^n x_i$$

si $x_1, \dots, x_n > 0$. (A veces se reparametriza expresándose como $f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}$, con $\mu = \theta$ en este caso).

- Para una distribución teórica gamma $\gamma(p, a)$ se tiene que $f_\theta(x) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax}$ para $x > 0$, con $\theta = (p, a)$ y $\Theta = (0, \infty)^2$. Se tiene que $\mu = p/a$ y $\sigma^2 = p/a^2$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = \frac{a^{np}}{\Gamma(p)^n} (\prod_{i=1}^n x_i)^{p-1} e^{-\theta t}, \quad \text{con } t = \sum_{i=1}^n x_i$$

si $x_1, \dots, x_n > 0$.

- Para una distribución teórica beta $Be(\alpha, \beta)$ se tiene que $f_\theta(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$ para $0 < x < 1$, con $\theta = (\alpha, \beta)$ y $\Theta = (0, \infty)^2$. Se tiene que $\mu = \frac{\alpha}{\alpha+\beta}$ y $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = B(\alpha, \beta)^{-n} (\prod_{i=1}^n x_i)^{\alpha-1} (\prod_{i=1}^n (1-x_i))^{\beta-1}$$

si $0 < x_1, \dots, x_n < 1$.

- Para una distribución teórica uniforme $U(0, \theta)$ se tiene que $f_\theta(x) = \frac{1}{\theta} I_{(0, \theta)}(x)$ si $x > 0$, con soporte $(0, \theta)$ que depende de θ , siendo $I_A(x) = 1$ si $x \in A$ y $I_A(x) = 0$ si $x \notin A$. Con $\Theta = (0, \infty)$. Se tiene que $\mu = \theta/2$ y $\sigma^2 = \theta^2/12$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = \theta^{-n} I_{(0, \theta)}(x_{(n)}) \quad \text{para } x_1, \dots, x_n > 0,$$

siendo $x_{(n)}$ el máximo $x_{(n)} = \max\{x_1, \dots, x_n\}$

- Para una distribución teórica uniforme $U(\theta_1, \theta_2)$ se tiene que

$$\begin{aligned} f_\theta(x) &= \frac{1}{\theta_2 - \theta_1} I_{(\theta_1, \theta_2)}(x) = \frac{1}{\theta_2 - \theta_1} I_{(\theta_1, \infty)}(x) I_{(-\infty, \theta_2)}(x) \\ &= \frac{1}{\theta_2 - \theta_1} I\{x > \theta_1\} I\{x < \theta_2\}, \quad \text{para } -\infty < x < \infty, \end{aligned}$$

denotando $I_A(x)$ por $I\{x \in A\}$. El soporte es el intervalo (θ_1, θ_2) , y el espacio paramétrico

$$\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : \theta_1 < \theta_2\}.$$

Se tiene que $\mu = \frac{\theta_1 + \theta_2}{2}$ y $\sigma^2 = \frac{(\theta_2 - \theta_1)^2}{12}$. Se obtiene:

$$f_\theta(x_1, \dots, x_n) = (\theta_2 - \theta_1)^{-n} I\{x_{(1)} > \theta_1\} I\{x_{(n)} < \theta_2\} \quad \text{para } -\infty < x_1, \dots, x_n < \infty,$$

siendo $x_{(1)}$ el mínimo $x_{(1)} = \min\{x_1, \dots, x_n\}$

(Aquí I es la función indicatriz, con $I_A(y) = 1$ si $y \in A$ y $I_A(y) = 0$ si $y \notin A$).

El estudio del verdadero valor de θ , fijo pero desconocido, puede afrontarse de distintas maneras utilizando la información proporcionada por la muestra:

- Estimación puntual: se obtiene un valor numérico único cercano al verdadero valor; el grado de error cometido puede ser medido.

- Contrastes o tests de hipótesis: se pretende corroborar o rechazar cierta afirmación sobre la distribución de probabilidad del fenómeno estudiado, tales como comprobar si un dado está equilibrado o si cierto fármaco aumenta la supervivencia ante una enfermedad. La afirmación (hipótesis) se expresa en términos de un subconjunto Θ_0 de Θ , y se pretende estudiar si el verdadero valor de θ pertenece a Θ_0 o no.
- Estimación por intervalos de confianza: se obtiene un intervalo numérico en el que pueda razonablemente afirmarse que está el verdadero valor del parámetro.

1.1.4 Estadístico muestral

La información proporcionada por la muestra puede ser resumida mediante funciones de la muestra. En el próximo tema estudiaremos cuándo se pierde o no información con ello (estadísticos suficientes).

Definición 3. Se denomina *estadístico* a cualquier función medible de la muestra

$$T : \Omega \rightarrow \mathbb{R}^k .$$

La medibilidad impone que la imagen inversa de los conjuntos de Borel en \mathbb{R}^k sean elementos de la σ -álgebra de sucesos considerada en Ω , para que esté definida la probabilidad sobre \mathbb{R}^k a partir de la probabilidad sobre Ω . Si T es continua o continua a trozos, caso habitual, es medible, y podemos soslayar entonces esta cuestión de consistencia para el formalismo matemático.

Ejemplos de estadísticos son la media $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, o el mínimo, o el máximo de la muestra ($k = 1$), o el vector formado por el mínimo y el máximo ($k = 2$). La misma muestra (x_1, \dots, x_n) es un estadístico, con $k = n$ y siendo T la identidad. La dimensión k del espacio imagen se denomina *dimensión del estadístico*.

El estadístico $T(x_1, \dots, x_n)$ toma distintos valores según (x_1, \dots, x_n) varía. Puesto que suponemos que las observaciones son realizaciones de X_1, \dots, X_n , variables aleatorias, podemos considerar T como una transformación (cambio de variable aleatoria) de la muestra.

Se denomina *distribución (en el muestreo) de un estadístico T* a la distribución de la variable aleatoria.

$$T(X_1, \dots, X_n) .$$

En estadística descriptiva solo se utiliza el valor numérico $T(x_1, \dots, x_n)$. En estadística matemática estamos suponiendo que (x_1, \dots, x_n) es una realización de (X_1, \dots, X_n) , lo que permite considerar el estadístico $T = T(X_1, \dots, X_n)$ como v.a., y entonces podemos utilizar las herramientas del cálculo de probabilidades para estudiar su distribución. Esto permite el diseño de métodos de inferencia con los que se logra extraer la información que ofrecen los datos x_1, \dots, x_n sobre el mecanismo que los genera.

1.1.5 Distribución empírica de la muestra

La *distribución empírica de la muestra* (x_1, \dots, x_n) , denotada por P_n^* , es la distribución de probabilidad discreta con soporte $\{x_1, \dots, x_n\}$ que asigna probabilidad $1/n$ a cada valor x_i . Si hay datos repetidos se acumula la probabilidad.

Es una distribución artificial, construida por conveniencia, y no se pretende realizar ningún experimento aleatorio para obtener una realización de P_n^* dada una muestra (x_1, \dots, x_n) (salvo métodos de remuestreo, que no se estudian aquí). La distribución empírica resume la información muestral (ignorando el orden en que fueron obtenidos x_1, \dots, x_n), y tiene utilidad por sus propiedades probabilísticas, que se obtendrán a partir de la incorporación a la formulación de la distribución en el muestreo.

Veremos que con n suficientemente grande es muy probable que P_n^* sea muy cercana a la distribución teórica P . De este modo, P se puede aproximar mediante P_n^* para hacer cálculos de probabilidad y también de características numéricas de P . En particular, veremos que los momentos y los cuantiles de P se pueden aproximar (estimar, con la terminología del tema 3) mediante los correspondientes momentos y cuantiles de P_n^* . Así, para obtener información sobre la esperanza μ de la distribución teórica, se puede utilizar la esperanza de la distribución empírica que es (compruébese) \bar{x} , aunque no siempre es el mejor estadístico para ello, como veremos en los próximos temas.

A la correspondiente función de distribución se la denomina *función de distribución empírica*, denotada por F_n^* , de modo que $F_n^*(x)$ es la proporción de los x_i que son menores o iguales que x , que se puede expresar como

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^n I\{x_i \leq x\} \quad \text{para } -\infty < x < \infty,$$

con $I\{x_i \leq x\} = 1$ si $x \geq x_i$ e $I\{x_i \leq x\} = 0$ si $x < x_i$.

Ejemplo 4. Se obtuvo la muestra $(x_1, x_2, x_3, x_4) = (3.9, 6.7, 3.9, 2.1)$. Se presentan a continuación la función de masa de la distribución empírica P_4^* (en su soporte), y la función de distribución empírica:

$$P_4^*\{2.1\} = 1/4, P_4^*\{3.9\} = 1/4 + 1/4 = 2/4 \text{ y } P_4^*\{6.7\} = 1/4$$

$$F_4^*(x) = \begin{cases} 0 & \text{si } x < 2.1 \\ 1/4 & \text{si } 2.1 \leq x < 3.9 \\ 3/4 & \text{si } 3.9 \leq x < 6.7 \\ 1 & \text{si } x \geq 6.7 \end{cases}$$

Para todo x , $F_n^*(x)$ es un estadístico, puesto que es función de la muestra. Vamos a obtener su distribución en el muestreo. Utilizamos la misma notación $F_n^*(x)$ para la función de distribución empírica como realización y como v.a., y ahora consideramos, para cada x , la v.a.

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\}.$$

Sea $Z_i = I\{X_i \leq x\}$, de modo que $F_n^*(x) = \bar{Z}$, con $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$.

Es inmediato comprobar que Z_1, \dots, Z_n son v.a. i.i.d. $B(1, p)$, con $p = F(x)$. Entonces, $\sum_{i=1}^n Z_i = nF_n^*(x)$ tiene distribución binomial, $nF_n^*(x) \sim B(n, F(x))$, y entonces la distribución de $F_n^*(x)$ (su función de masa) viene dada por:

$$\begin{aligned} P\{F_n^*(x) = k/n\} &= P\{nF_n^*(x) = k\} = P\{B(n, F(x)) = k\} \\ &= \binom{n}{k} F(x)^k [1 - F(x)]^{n-k} \quad \text{para } k = 0, \dots, n. \end{aligned}$$

Con el símbolo " \sim " denotamos la igualdad en distribución, entre dos variables aleatorias, o igualdad de la distribución de la v.a. que hay a la izquierda con la distribución que se nombra a la derecha, como en $nF_n^*(x) \sim B(n, F(x))$.

(Cometeremos abuso de notación en expresiones de este tipo, permitiendo escribir $P\{B(n, F(x)) = k\}$, de modo que en esa expresión, $B(n, F(x))$ se refiere a una v.a., $nF_n^*(x)$, con esa distribución, y no a la distribución).

A partir de la esperanza y varianza de la $B(n, F(x))$ y simplificando, se obtiene para cada x

$$E[F_n^*(x)] = F(x) \quad \text{y} \quad V[F_n^*(x)] = F(x)[1 - F(x)]/n.$$

Entonces, $F_n^*(x)$ se distribuye alrededor de $F(x)$, con una dispersión que tiende a cero cuando n tiende a infinito (pequeña para n grande). Entonces $F_n^*(x)$ proporciona un valor aproximado para $F(x)$.

Ahora consideramos, para cada x , a $F_n^*(x)$ como v.a. función de n y estudiamos la convergencia cuando n tiende a infinito. También enunciamos un resultado que involucra a $F_n^*(x)$ para todos los valores reales de x simultáneamente.

Puesto que $F_n^*(x) = \bar{Z}$ y $E[Z_i] = F(x)$, por la ley fuerte de los grandes números se tiene que

$$F_n^*(x) \xrightarrow{\text{c.s.}} F(x). \quad (1.1)$$

Recuérdese que la convergencia casi segura implica la convergencia en probabilidad, y entonces $F_n^*(x) \xrightarrow{p} F(x)$. Además, puesto que $F(x)$ es constante, la convergencia en probabilidad es a una distribución degenerada en $F(x)$, y ésto es equivalente a la convergencia en distribución (en ley): se tiene que $F_n^*(x) \xrightarrow{d} F(x)$. Por tanto, la distribución de $F_n^*(x)$ se va acercando a una distribución degenerada en $F(x)$ según n crece, aunque esto se podría haber obtenido también a partir de la desigualdad de Chebichev y los dos primeros momentos de $F_n^*(x)$, que hemos obtenido antes.

El siguiente resultado establece que esta convergencia se cumple además uniformemente, para todo x simultáneamente.

Teorema 5. (de Glivenko-Cantelli) Si X_1, X_2, \dots son v.a. i.i.d. con función de distribución F se tiene que

$$\sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)| \rightarrow 0 \quad \text{c.s.} \quad (1.2)$$

Por el resultado (1.1), para cada x se tiene que $F_n^*(x)$ tiende a concentrar la probabilidad cada vez mas cerca de $F(x)$ según crece n . Por el teorema 5, esto ocurre para todo x simultaneamente, según se explica a continuación.

Consideremos la v.a. $\Delta_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)|$. (La distribución del estadístico Δ_n , que no se presenta aquí, fue estudiada por A.N. Kolmogorov, y se utiliza en problemas de inferencia no paramétrica). Se tiene que

$$P \{ \Delta_n \leq \varepsilon \} = P \{ |F_n^*(x) - F(x)| \leq \varepsilon \text{ para todo } x \} .$$

Puesto que la convergencia casi segura implica la convergencia en probabilidad, entonces se tiene por (1.2) que $\lim_{n \rightarrow \infty} P \{ \Delta_n \leq \varepsilon \} = 1$ para todo $\varepsilon > 0$. Entonces, con n suficientemente grande se tiene que

$$F_n^*(x) - \varepsilon < F(x) < F_n^*(x) + \varepsilon , \text{ para todo } x \text{ simultaneamente,}$$

con una probabilidad (en el muestreo) tan grande como se desee, lo que ofrece una acotación para $F(x)$ para todo x con una amplitud de 2ε , con probabilidad grande.

Una m.a.s. es una secuencia de vaaid, y por tanto este teorema establece que la muestra no solo aporta información sobre la distribución teórica F desconocida, sino que esta información tiende a ser completa cuando cuando el tamaño de muestra crece.

Podemos utilizar el teorema central del límite (TCL) para obtener para cada x una aproximación de cómo de cercana a $F(x)$ es la v.a. $F_n^*(x)$ para n dado no demasiado pequeño. Puesto que $E[Z_i] = F(x)$, $V[Z_i] = F(x)[1 - F(x)]$ y $F_n^*(x) = \bar{Z}$, por el TCL se tiene que $F_n^*(x)$ es asintóticamente normal $N \left(F(x), \sqrt{F(x)[1 - F(x)]/n} \right)$, denotado por

$$F_n^*(x) \sim AN \left(F(x), \sqrt{F(x)[1 - F(x)]/n} \right) ,$$

que quiere decir que la variable aleatoria

$$\frac{F_n^*(x) - F(x)}{\sqrt{F(x)[1 - F(x)]/n}}$$

converge en ley a una (v.a. con) distribución $N(0, 1)$.

(O que $(F_n^*(x) - F(x)) \sqrt{n}$ converge en ley a una $N \left(0, \sqrt{F(x)[1 - F(x)]} \right)$).

Puesto que, a través de la distribución empírica, la muestra tiende a dar información completa sobre la distribución teórica, las características de la distribución empírica deberían tender a dar información completa sobre las correspondientes características de la distribución teórica. Así ocurre cuando estas características son los momentos y los cuantiles, según estudiamos en los dos próximos apartados. En todos los casos, cuando n tiende a infinito, los momentos y cuantiles muestrales convergen (con su distribución en el muestreo, como variables aleatorias) casi seguro a (una v.a. degenerada en) los correspondientes momentos y los cuantiles poblacionales, y la distribución asintotica es normal.

1.1.6 Momentos muestrales

Definición. Media y varianza muestral

El momento muestral de orden k respecto del origen es

$$a_k = \frac{1}{n} \sum_{i=1}^n x_i^k ,$$

El momento muestral de orden k respecto de la media (o central) es

$$b_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k , \text{ con } \bar{x} = a_1 .$$

Estos estadísticos, los momentos muestrales, son justamente los correspondientes momentos de la distribución empírica. Lo comprobamos para a_1 y b_2 , y el caso general se comprueba del mismo modo.

Sea ξ una v.a. cuya distribución sea la distribución empírica P_n^* de una muestra dada (x_1, \dots, x_n) . Entonces, por ejemplo, el momento de la v.a. ξ de orden 1 respecto del origen es

$$E[\xi] = \sum_{i=1}^n \frac{1}{n} x_i = \bar{x} ,$$

que coincide con la cantidad que hemos definido como momento muestral de orden 1 respecto del origen a_1 . El momento de ξ de orden 2 central es la varianza

$$V[\xi] = E[(\xi - E[\xi])^2] = \sum_{i=1}^n \frac{1}{n} (x_i - \bar{x})^2 ,$$

que coincide con la cantidad que hemos definido como momento muestral de orden 2 central b_2 .

Entonces, “momento muestral” es equivalente a “momento de la distribución empírica”.

Se usan principalmente con k igual a 1 y 2. Denotamos a_1 como \bar{x} y b_2 como s^2 :

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

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

Las dos siguientes expresiones para s^2 son útiles en algunas ocasiones. Desarrollando el cuadrado $(x_i - \bar{x})^2 = ((x_i - a) + (a - \bar{x}))^2$ y simplificando se obtiene, para toda constante a , $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2 - (\bar{x} - a)^2$. Esta expresión se suele utilizar con $a = 0$ y $a = \mu$:

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

Esta expresión también se puede obtener utilizando las propiedades de la varianza, teniendo en cuenta que $s^2 = V[\xi]$ es la varianza de la distribución empírica:

$$\begin{aligned} s^2 &= V[\xi] = V[\xi - a] = E[(\xi - a)^2] - E[\xi - a]^2 \\ &= E[(\xi - a)^2] - (E[\xi] - a)^2 = \sum_{i=1}^n \frac{1}{n} (x_i - a)^2 - (\bar{x} - a)^2 . \end{aligned}$$

Al estadístico \bar{x} se le denomina *media muestral* y al estadístico s^2 *varianza muestral*.

A partir de una realización muestral se pueden calcular los momentos muestrales, que son valores numéricos. Pero antes del muestreo, estos estadísticos son variables aleatorias, con su distribución en el muestreo, que estudiamos a continuación.

Distribución y momentos de los momentos muestrales

Utilizamos la misma notación para considerar los momentos muestrales como v.a. que como realizaciones (con la excepción de \bar{x} , que la denotaremos por \bar{X} como v.a.).

Sean entonces

$$a_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad \text{y} \quad b_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k ,$$

y en particular

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{y} \quad s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 .$$

La distribución de estos estadísticos no admite una expresión general, y debe ser estudiada en cada caso particular, según la forma funcional de la distribución teórica. Aún así, su obtención puede no ser sencilla. Mas adelante en este tema nos ocuparemos del caso de poblaciones normales, para el que el manejo de la distribución exacta de los principales momentos sí tiene una solución satisfactoria, e involucra las distribuciones χ^2 , t de Student y F de Snedecor.

Se dice que una *familia paramétrica* de distribuciones, con parámetro θ , es *reproductiva* respecto de θ , si la distribución de la suma de dos variables independientes X e Y de esa familia pertenece también a la familia, siendo el valor del parámetro para $X + Y$ la suma de los valores para X y para Y .

Para comprobar la reproductividad, se puede determinar la función característica de $X + Y$ mediante la fórmula $\varphi_{X+Y}(t) = \varphi_X(t) \varphi_Y(t)$, y comprobar si corresponde también a la familia, y a qué valor del parámetro. Por ejemplo, si la función característica de la distribución teórica es de la forma $\varphi(t) = h(t)^\theta$, y X e Y , independientes, tienen distribución en esa familia con parámetro θ_1 y θ_2 respectivamente, se obtiene $\varphi_{X+Y}(t) = h(t)^{\theta_1 + \theta_2}$, que corresponde a esa misma familia con parámetro $\theta_1 + \theta_2$, y entonces la familia paramétrica considerada es reproductiva.

En caso de que la distribución teórica sea paramétrica y reproductiva, la distribución de $n\bar{X} = \sum_{i=1}^n X_i$ (lo expresamos así por brevedad) se obtiene de un modo inmediato.

Por ejemplo, si $X \sim \mathcal{P}(\theta)$ (Poisson), reproductiva respecto de θ , entonces $n\bar{X} \sim \mathcal{P}(n\theta)$, y a partir de aquí se obtiene la distribución de \bar{X} . (Itérese la propiedad de reproductividad, definida para dos sumandos, para obtener el resultado en este caso con n sumandos de $n\bar{X}$).

Si $X \sim B(k, p)$ (binomial), reproductiva respecto de $\theta = k$ para p fijo, se tiene que $n\bar{X} \sim B(nk, p)$.

Si $X \sim N(\mu, \sigma)$ (normal), reproductiva respecto de $g(\theta) = (\mu, \sigma^2)$, se tiene que $n\bar{X} \sim N(n\mu, \sigma\sqrt{n})$.

Si $X \sim \gamma(p, a)$ (gamma), reproductiva respecto de $\theta = p$ para a fijo, se tiene que $n\bar{X} \sim \gamma(np, a)$.

Y lo mismo se aplica a otros momentos a_k si la distribución de X^k , o una función sencilla de ella, es reproductiva. Por ejemplo, si $X \sim N(0, \sigma)$, se obtiene (lo comprobaremos mas adelante) que $\frac{X^2}{\sigma^2} \sim \gamma(1/2, 1/2)$, y por la reproductividad de la gamma respecto de $\theta = p$ se obtiene $\frac{na_2}{\sigma^2} \sim \gamma(n/2, 1/2)$.

Los momentos muestrales, como variables aleatorias, tienen sus propios momentos. Por ejemplo, podemos estar interesados en determinar la esperanza y la varianza (momentos de orden 1 y 2) de la media muestral (momento muestral de orden 1). A diferencia de lo que ocurre con la distribución de los momentos muestrales, los momentos de los momentos muestrales sí admiten una expresión general. Ésto nos permite obtener su comportamiento asintótico, que resulta útil cuando la distribución exacta es poco manejable.

Denotamos los momentos poblacionales (de X , de F) por

$$\alpha_k = E[X^k] \quad \text{y} \quad \beta_k = E[(X - \mu)^k] .$$

en particular, $\mu = \alpha_1 = E[X]$ y $\sigma^2 = \beta_2 = E[(X - \mu)^2] = V[X]$ son la esperanza y la varianza poblacionales (esto es, de la distribución teórica). A esta esperanza μ también se la denomina media poblacional, pero el término "poblacional" es frecuentemente omitido, y debe ser identificado por el contexto si la media a la que se hace referencia es media muestral o poblacional, esperanza.

Es sencillo comprobar que $E[a_k] = \alpha_k$ y $V[a_k] = (\alpha_{2k} - \alpha_k^2)/n$. Con $k = 1$ se obtiene

$$E[\bar{X}] = \mu , \tag{1.3}$$

$$V[\bar{X}] = \frac{\sigma^2}{n} . \tag{1.4}$$

Para la varianza muestral s^2 se verifica:

$$E[s^2] = \frac{n-1}{n}\sigma^2 \quad \text{y} \tag{1.5}$$

$$V[s^2] = \frac{(n-1)^2}{n^3}\beta_4 - \frac{(n-1)(n-3)}{n^3}\beta_2^2 .$$

La varianza muestral s^2 sirve para estudiar la varianza poblacional σ^2 . Sin embargo, la esperanza de $E[s^2]$ no es exactamente σ^2 (aunque sí asintóticamente, véase (1.5)). Resulta conveniente disponer de un estadístico cuya esperanza coincida con σ^2 , y por ello se define la *cuasivarianza muestral* (también llamada *varianza muestral corregida*, o *varianza muestral*):

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

que tiene la siguiente propiedad.

$$E[S^2] = \sigma^2.$$

Comprobamos que $E[s^2] = \frac{n-1}{n} \sigma^2$, y de aquí que $E[S^2] = \sigma^2$. Se tiene que:

$$\begin{aligned} E[s^2] &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X} - \mu)^2 \right] \\ &= \frac{1}{n} \sum_{i=1}^n E[(X_i - \mu)^2] - E[(\bar{X} - \mu)^2] \\ &= \frac{1}{n} \sum_{i=1}^n V[X_i] - V[\bar{X}] = \frac{1}{n} n \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n-1}{n} \sigma^2. \end{aligned}$$

De aquí se obtiene

$$E[S^2] = E \left[\frac{n}{n-1} s^2 \right] = \frac{n}{n-1} E[s^2] = \sigma^2.$$

Comportamiento asintótico de los momentos muestrales

Aquí seguimos considerando los momentos muestrales como variables aleatorias, y estudiamos la convergencia cuando n tiende a infinito. Indicamos de un modo explícito la dependencia de n con la notación $a_k(n) = a_k$ y $b_k(n) = b_k$. Los siguientes resultados solo son válidos si todos los momentos poblacionales α_k y β_k involucrados son finitos, que es el caso habitual.

Puesto que las X_i^k son vauid podemos aplicar las leyes de los grandes números y el teorema central del límite a $a_k(n)$.

Tomando $Z_i = X_i^k$, se tiene que Z_1, \dots, Z_n son vauid con $E[Z_i] = \alpha_k$ y $V[Z_i] = \alpha_{2k} - \alpha_k^2$. Se tiene que $a_k(n) = \bar{Z}$, y entonces se verifica que

$$a_k(n) \xrightarrow{c.s.} \alpha_k \quad \text{y} \quad a_k(n) \sim AN \left(\alpha_k, \sqrt{\frac{\alpha_{2k} - \alpha_k^2}{n}} \right). \quad (1.6)$$

Se verifica también que $b_k(n) \xrightarrow{c.s.} \beta_k$ y que $b_k(n)$ es asintóticamente normal, aunque la demostración requiere mas elaboración que la de (1.6), que es aplicación directa de los resultados de convergencia mencionados.

En particular

$$\bar{X} \sim AN\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{y} \quad s^2 \sim AN\left(\sigma^2, \sqrt{\frac{\beta_4 - \beta_2^2}{n}}\right). \quad (1.7)$$

En el caso de poblaciones normales, no solamente \bar{X} tiene distribución asintótica $N(\mu, \sigma/\sqrt{n})$, sino que la distribución es exactamente normal, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$, puesto que es combinación lineal de v.a. normales independientes, con distribución conjunta normal por tanto, y entonces \bar{X} es normal, con los mismos momentos que los obtenidos en (1.3) y (1.4).

Ejercicio 6. Consideremos una m.a.s. de tamaño n de una distribución $N(\mu, 1.5)$.

- Calcula la probabilidad de que la media muestral difiera de μ (en valor absoluto) en menos de 0.1, con $n = 100$.
- ¿Cuanto debe valer n para que la probabilidad de que la media muestral difiera de μ en menos de 0.1 valga 0.9.

Solución 7. En este caso $\bar{X} \sim N(\mu, 1.5/\sqrt{n})$ y entonces $\frac{\bar{X}-\mu}{1.5/\sqrt{n}} \sim N(0, 1)$.

a) Se tiene que

$$\begin{aligned} P\{|\bar{X} - \mu| < 0.1\} &= P\left\{\left|\frac{\bar{X}-\mu}{1.5/\sqrt{n}}\right| < \frac{0.1}{1.5/\sqrt{n}}\right\} \\ &= P\{|N(0, 1)| < 0.67\} = 1 - 2 \cdot 0.251 = 0.497 \end{aligned}$$

(con el abuso de notación habitual en $P\{|N(0, 1)| < 0.67\}$).

b) Se tiene que

$$\begin{aligned} P\{|\bar{X} - \mu| < 0.1\} &= P\left\{\left|\frac{\bar{X}-\mu}{1.5}\sqrt{n}\right| < \frac{0.1}{1.5}\sqrt{n}\right\} \\ &= 1 - 2 \cdot P\{N(0, 1) > \frac{1}{15}\sqrt{n}\} = 0.9, \end{aligned}$$

y entonces $P\{N(0, 1) > \frac{1}{15}\sqrt{n}\} = 0.05$. Puesto que $P\{N(0, 1) > 1.64\} = 0.0505$, debe ser (aproximadamente, utilizando tablas y redondeando, y mejor aproximación con software) $\frac{1}{15}\sqrt{n} = 1.64$, y se obtiene $n = 605$.

También se obtienen resultados similares para los cuantiles muestrales.

1.1.7 Cuantiles muestrales

Los *cuantiles muestrales* son los cuantiles de la distribución empírica. Convergen casi seguro a los correspondientes cuantiles poblacionales y son asintóticamente normales.

Recordamos la definición de cuantil y enunciamos los resultados de convergencia, que sirven para realizar aproximaciones cuando n es suficientemente grande.

Consideremos una v.a. Y con función de distribución G . Para $0 < p < 1$, un *cuantil de orden p* de Y (o de G) es un valor y_p tal que $G(y_p^-) \leq p \leq G(y_p)$ (con $G(y_p^-) =$

$\lim_{y \uparrow y_p} G(y)$). Si existe un único y tal que $G(y) = p$ entonces $y_p = y = G^{-1}(p)$, y el cuantil y_p es único (es el único valor que deja a su izquierda una probabilidad p).

En el caso discreto, para todo p , o no existe y con $G(y) = p$ o si existe no es único. Si g es la función de masa se obtiene $G(y_p^-) = G(y_p) - g(y_p)$, y entonces un cuantil y_p cumple

$$G(y_p) - g(y_p) \leq p \leq G(y_p) .$$

De aquí se obtiene que si y es un punto de salto de G (si y está en el soporte de G), con salto $g(y) > 0$, para todo p con $G(y) - g(y) < p \leq G(y)$ se tiene que $y_p = y$ es el cuantil de orden p , y es único para cada uno de estos valores de p . Para el resto de valores y (no en el soporte), si $G(y) = p$ para $a < y < b$, entonces cualquiera de estos valores y es un cuantil de orden p . Se suele tomar $y_p = (a + b) / 2$. Las propiedades de convergencia que se estudiarán no dependen de esta elección.

A continuación enunciamos los resultados de convergencia.

Sean X_1, X_2, \dots v.a.i.i.d con distribución teórica F . Suponemos que F tiene un único cuantil de orden p , que denotamos por x_p . Sea $c_p(n)$ el cuantil de orden p (considerado como v.a.) para la distribución empírica de la m.a.s. X_1, \dots, X_n . Entonces se tiene que

$$c_p(n) \xrightarrow{c.s.} x_p .$$

Si, además, la distribución teórica F es continua con función de densidad f , se verifica que

$$c_p(n) \sim N \left(x_p, \frac{\sqrt{p(1-p)}}{\sqrt{n}f(x_p)} \right) \quad (1.8)$$

Ejercicio 8. Consideremos una m.a.s. de tamaño n de una distribución $N(\mu, 1.5)$.

- Calcula la probabilidad de que la mediana muestral difiera de μ (en valor absoluto) en menos de 0.1, con $n = 100$.
- ¿Cuanto debe valer n para que la probabilidad de que la mediana muestral difiera de μ en menos de 0.01 valga 0.9.

Solución 9. La mediana es el cuantil de orden $p = 1/2$. La distribución normal es simétrica, y por tanto $x_{1/2} = \mu$. Vamos a utilizar la aproximación dada por (1.8) ($n = 100$ es suficientemente grande). En este caso $f(x_{1/2}) = f(\mu) = 1 / (1.5\sqrt{2\pi})$. Sea $M = c_{1/2}(n)$. Por (1.8) se tiene que $M \sim N \left(\mu, \frac{1.5\sqrt{2\pi}}{2\sqrt{n}} \right)$ ($\frac{\sqrt{2\pi}}{2} = 1.253$).

a) Se tiene que

$$\begin{aligned} P \{|M - \mu| < 0.1\} &= P \left\{ \left| \frac{M - \mu}{\frac{1.5\sqrt{2\pi}}{2\sqrt{n}}} \right| < \frac{0.1}{\frac{1.5\sqrt{2\pi}}{2\sqrt{n}}} \right\} \\ &\simeq P \{|N(0, 1)| < 0.53\} = 1 - 2 \cdot 0.2981 = 0.404 \end{aligned}$$

b) Se tiene que

$$\begin{aligned} P\{|M - \mu| < 0.01\} &\simeq P\left\{|N(0, 1)| < \frac{1}{75\sqrt{2\pi}}\sqrt{n}\right\} \\ &= 1 - 2 \cdot P\left\{N(0, 1) > \frac{1}{75\sqrt{2\pi}}\sqrt{n}\right\} = 0.9, \end{aligned}$$

y entonces $P\left\{N(0, 1) > \frac{1}{75\sqrt{2\pi}}\sqrt{n}\right\} = 0.05$. Puesto que $P\{N(0, 1) > 1.64\} = 0.0505$, debe ser $\frac{1}{75\sqrt{2\pi}}\sqrt{n} = 1.64$, y se obtiene $n = 95058$.

1.2 Sumas de variables aleatorias de una muestra aleatoria

Véase el apartado 1.1.6.

1.3 Muestreo de una distribución normal: propiedades de la media y varianzas muestrales y las distribuciones t de Student y F de Snedecor.

1.3.1 Introducción

En muchas situaciones la distribución teórica es normal $N(\mu, \sigma)$. Los estadísticos básicos para el estudio de los parámetros de esta distribución, que se realizará en los próximos temas, son:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ y } s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \text{ o } S^2 = \frac{n}{n-1} s^2.$$

Como se comentó, la obtención de la distribución de los momentos muestrales, en particular la de \bar{X} y s^2 , no es un problema sencillo de resolver, y no hay una solución satisfactoria para muchas familias paramétricas. Sin embargo, en poblaciones normales la distribución de \bar{X} y s^2 sí admite una expresión exacta y manejable, e igualmente para otros estadísticos función de ellos que serán herramientas útiles, como el estadístico de Student.

Además, aunque la distribución teórica no sea normal, para muestras grandes (n grande), la distribución de \bar{X} es aproximadamente la misma que en el caso normal, lo que resulta de utilidad para poder aplicar los métodos de inferencia estadística: \bar{X} tiene distribución asintótica $N(\mu, \sigma/\sqrt{n})$ para cualquier distribución teórica con esperanza μ y varianzas σ^2 finitas, por el TCL.

Comenzamos este apartado introduciendo las distribuciones χ^2 , t de Student y F de Snedecor, asociadas a variaciones normales. La distribución de los principales estadísticos de interés en el muestreo de poblaciones normales se expresa en términos de estas distribuciones. Estudiaremos la distribución de estos estadísticos, como por ejemplo la distribución conjunta de (\bar{X}, s^2) . También consideraremos los casos en los que hay, o dos muestras normales, o muestras de una población normal bidimensional.

1.3.2 Distribuciones asociadas a la normal

Distribución χ^2

En primer lugar, comprobamos que la distribución del cuadrado de una v.a. $N(0, 1)$ es $\gamma(1/2, 1/2)$.

Sea $X \sim N(0, 1)$ e $Y = X^2$. La función $h(x) = x^2$ no es inyectiva para x en el soporte de X , y no se puede aplicar el teorema de cambio de variable aleatoria para obtener la distribución de Y . Hay que operar de otro modo, por ejemplo de la siguiente manera. Sean F_X, f_X, F_Y, f_Y las funciones de distribución y de densidad de las correspondientes variables. Para $y \geq 0$ se tiene que:

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\{X^2 \leq y\} = P\{|X| \leq \sqrt{y}\} \\ &= P\{-\sqrt{y} \leq X \leq \sqrt{y}\} = 2 \cdot P\{0 \leq X \leq \sqrt{y}\} \\ &= 2(F_X(\sqrt{y}) - F(0)), \end{aligned}$$

teniendo en cuenta que la distribución $N(0, 1)$ es simétrica. De aquí se obtiene:

$$\begin{aligned} f_Y(y) &= F'_Y(y) = 2 \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) = y^{-1/2} \frac{1}{\sqrt{2\pi}} e^{-\sqrt{y}^2/2} \\ &= \frac{2^{-1/2}}{\sqrt{\pi}} y^{\frac{1}{2}-1} e^{-y/2} \quad \text{para } y > 0, \end{aligned}$$

que es la función de densidad de una distribución gamma $\gamma(1/2, 1/2)$, puesto que $\Gamma(1/2) = \sqrt{\pi}$. Entonces, $Y = X^2 \sim \gamma(1/2, 1/2)$.

La distribución γ es reproductiva respecto del primer parámetro, y por tanto la distribución de la suma de los cuadrados de n variables $N(0, 1)$ independientes es $\gamma(n/2, 1/2)$: si X_1, \dots, X_n son v.a.i.d $N(0, 1)$, se tiene que

$$\sum_{i=1}^n X_i^2 \sim \gamma(n/2, 1/2).$$

Esta distribución se define a partir de variables $N(0, 1)$, pero también está relacionada con variables $N(\mu, \sigma)$ en general, puesto que las podemos transformar en $N(0, 1)$ mediante la tipificación: si X_1, \dots, X_n son v.a.i.d $N(\mu, \sigma)$ se tiene que

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \gamma(n/2, 1/2).$$

También está relacionada con la distribución de s^2 (y de S^2) para muestras de una $N(\mu, \sigma)$ (a partir del teorema de Fisher, más adelante). La varianza muestral s^2 sirve para estudiar la varianza poblacional σ^2 , y por ello esta distribución $\gamma(n/2, 1/2)$ es útil en este estudio. Por su amplio uso, se le ha puesto nombre propio.

Se dice que una v.a. X tiene *distribución χ^2 con n grados de libertad*, denotada por χ_n^2 , si tiene distribución $\gamma(n/2, 1/2)$. La función de densidad es

$$f(x) = \frac{2^{-n/2}}{\Gamma(n/2)} x^{n/2-1} e^{-x/2} \quad \text{para } x > 0.$$

La distribución χ_2^2 es la $\gamma(1, 1/2)$, que es la $Exp(1/2)$.

La esperanza y la varianzas son

$$E[X] = n, \quad V[X] = 2n.$$

La distribución χ_n^2 es reproductiva: si $X \sim \chi_n^2$ e $Y \sim \chi_m^2$ y son independientes, entonces $X + Y \sim \chi_{n+m}^2$. Esta propiedad se obtiene a partir de la reproductividad de la gamma.

No se dispone de una expresión sencilla de la función de distribución, y por ello se hace necesario el uso de procedimientos informáticos o tablas para efectuar cálculos con esta distribución. La tabla usada habitualmente comprende valores de n desde 1 hasta 30.

Para valores de n grandes, podemos obtener la distribución aproximada (asintótica) utilizando el TCL, aplicándolo a Z_1, \dots, Z_n , con $Z_i = X_i^2$. Se obtiene que $\chi_n^2 \sim AN(n, \sqrt{2n})$, con el abuso de notación de escribir χ_n^2 en lugar de $\sum_{i=1}^n X_i^2$, siendo X_i variid $N(0, 1)$, que tiene distribución χ_n^2 .

Sin embargo, la siguiente aproximación, que se obtiene aplicando el TCL y otros desarrollos, tiene una convergencia mas rápida: se tiene que.

$$\sqrt{2\chi_n^2} - \sqrt{2n-1} \sim AN(0, 1). \tag{1.9}$$

Entiéndase aquí que χ_n^2 se refiere a una v.a. con esa distribución. Se considera que la aproximación es buena para $n > 30$.

Se denota por $\chi_{n,\alpha}^2$ al valor numérico tal que la χ_n^2 deja a su derecha una probabilidad α , esto es, $P\{\chi_n^2 > \chi_{n,\alpha}^2\} = \alpha$. Por ejemplo, buscando en las tablas se obtiene $\chi_{5,0.7}^2 = 3$, $\chi_{9,0.1}^2 = 14.68$ y $\chi_{25,0.01}^2 = 44.31$.

Ejemplo 10. Obtenemos el valor $\chi_{30,0.05}^2$ buscando en la tabla y mediante la aproximación (1.9). De la tabla se obtiene $\chi_{30,0.05}^2 = 43.77$. Sea $c = \chi_{30,0.05}^2$, que sabemos que vale 43.77, pero cuyo valor vamos a aproximar ahora mediante (1.9) como ejercicio de aplicación del resultado. Se tiene que

$$\begin{aligned} 0.05 &= P\{\chi_{30}^2 > c\} = P\left\{\sqrt{2\chi_{30}^2} - \sqrt{2 \cdot 30 - 1} > \sqrt{2c} - \sqrt{2 \cdot 30 - 1}\right\} \\ &\simeq P\left\{N(0, 1) > \sqrt{2c} - \sqrt{59}\right\}. \end{aligned}$$

Entonces $\sqrt{2c} - \sqrt{59} = z_{0.05} = 1.64$, y de aquí se obtiene $c = \frac{1}{2}(z_{0.05} + \sqrt{59})^2 = 43.44$, que es un valor cercano al valor correcto (con dos decimales de aproximación) 43.77.

Distribución t de Student

Consideremos $Z \sim N(0, 1)$ e $Y \sim \chi_n^2$ independientes. A la distribución de

$$T = \frac{Z}{\sqrt{Y/n}}$$

se la denomina *distribución t de Student con n grados de libertad*, y se la denota por t_n . De este modo, $T \sim t_n$. La función de densidad es

$$f_n(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi} \cdot \Gamma(n/2)} (1 + t^2/n)^{-(n+1)/2} \quad \text{para } -\infty < x < \infty.$$

Se tiene que $t_n \sim AN(0, 1)$. Se puede comprobar a partir de la obtención del siguiente límite: $\lim_{n \rightarrow \infty} (1 + t^2/n)^{-(n+1)/2} = e^{-t^2/2}$. Para $n > 30$ se considera que la aproximación de la t_n por la $N(0, 1)$ es buena.

La esperanza y la varianzas son

$$E[T] = 0 \text{ si } n > 1 \text{ y } V[T] = \frac{n}{n-2} \text{ si } n > 2.$$

Si X, X_1, \dots, X_n son vaaid $N(0, \sigma)$ se tiene que

$$\frac{X}{\sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2}} \sim t_n,$$

para cualquier valor de σ .

Esta distribución es útil para el estudio de μ en poblaciones normales para las que σ es también desconocida.

No hay una expresión sencilla para la función de distribución, y también se hace necesario el uso de procedimientos informáticos o tablas.

Denotaremos por $t_{n,\alpha}$ al valor tal que la t_n deja a su derecha una probabilidad α , esto es, $P\{t_n > t_{n,\alpha}\} = \alpha$. Por ejemplo, buscando en las tablas se obtiene $t_{10,0.1} = 1.372$, $t_{16,0.05} = 1.746$ y $t_{20,0.01} = 2.528$.

La fila de la tabla con $n = \infty$ corresponde a la $N(0, 1)$, según la normalidad asintótica de t_n mencionada antes. Por ejemplo, buscando en ambas tablas se obtiene $t_{\infty,0.025} = z_{0.025} = 1.96$.

Distribución F de Snedecor

Consideremos $X \sim \chi_n^2$ e $Y \sim \chi_m^2$ independientes. A la distribución de

$$T = \frac{X/n}{Y/m}$$

se la denomina *distribución F de Snedecor con n, m grados de libertad*, y se la denota por $F_{n,m}$. La función de densidad es

$$g(t) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{n}{m}\right) \left(\frac{n}{m}t\right)^{\frac{n}{2}-1} \left(1 + \frac{n}{m}t\right)^{-\frac{n+m}{2}} \quad \text{para } t > 0.$$

La esperanza y la varianzas son

$$E[T] = \frac{m}{m-2} \quad \text{si } m > 2 \quad \text{y} \quad V[T] = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)} \quad \text{si } n > 4.$$

Si $X_1, \dots, X_n, Y_1, \dots, Y_m$ son vaaid $N(0, \sigma)$ se tiene que

$$\frac{\frac{1}{n} \sum_{i=1}^n X_i^2}{\frac{1}{m} \sum_{i=1}^m Y_i^2} \sim F_{n,m},$$

para cualquier valor de σ .

Si $T \sim F_{n,m}$ se tiene que $(1/T) \sim F_{m,n}$. Expresado con los abusos de notación que estamos cometiendo, se tiene que $\frac{1}{F_{n,m}} \sim F_{m,n}$, que se demuestra teniendo en cuenta que $\left(\frac{\chi_n^2/n}{\chi_m^2/m}\right)^{-1} = \frac{\chi_m^2/m}{\chi_n^2/n}$, siendo χ_n^2 y χ_m^2 independientes.

También se hace necesario el uso de procedimientos informáticos o tablas. Denotaremos por $F_{n,m,\alpha}$ al valor tal que la $F_{n,m}$ deja a su derecha una probabilidad α , esto es, $P\{F_{n,m} > F_{n,m,\alpha}\} = \alpha$.

Por ejemplo, buscando en las tablas se obtiene $F_{4,7,0.005} = 10.050$, $F_{24,14,0.025} = 2.7888$ y $F_{12,28,0.1} = 1.7895$.

En las tablas solamente se incluyen valores que dejan a su derecha una probabilidad pequeña (de 0.005 a 0.1). También se necesitan en las aplicaciones estadísticas valores que dejan a su derecha una probabilidad grande. Obtenemos una relación que permite expresar los unos en función de los otros.

Consideremos α con $0 < \alpha < 1$. Por ejemplo, sea α pequeño (≤ 0.1), de modo que $1 - \alpha$ es grande (≥ 0.9). Se tiene que

$$\begin{aligned} \alpha &= P\{F_{n,m} > F_{n,m,\alpha}\} = P\left\{\frac{1}{F_{n,m}} < \frac{1}{F_{n,m,\alpha}}\right\} \\ &= P\left\{F_{m,n} < \frac{1}{F_{n,m,\alpha}}\right\} = 1 - P\left\{F_{m,n} > \frac{1}{F_{n,m,\alpha}}\right\}. \end{aligned}$$

Entonces $P\left\{F_{m,n} > \frac{1}{F_{n,m,\alpha}}\right\} = 1 - \alpha$, y por tanto debe ser

$$F_{m,n,1-\alpha} = \frac{1}{F_{n,m,\alpha}}.$$

Por ejemplo, $F_{2,5,0.9} = \frac{1}{F_{5,2,0.1}} = \frac{1}{9.2926} = 0.107$.

1.3.3 Distribuciones en el muestreo para una población normal

En este apartado se estudia la distribución de ciertos estadísticos para muestras normales, que serán útiles para los procedimientos de inferencia que se desarrollarán en los próximos temas. Se demostrará el teorema de Fisher y se presentará el estadístico de Student. Ampliamos el significado del término “estadístico”, función de la muestra, permitiendo ahora que también intervengan en él parámetros desconocidos, como en $\frac{\bar{X}-\mu}{\sigma} \sqrt{n}$, que es un estadístico en este sentido ampliado.

Consideramos que X_1, \dots, X_n es una m.a.s. de una distribución teórica normal $N(\mu, \sigma)$.

Distribución de la media y la varianzas muestral

El siguiente teorema permite obtener la distribución conjunta de (\bar{X}, s^2) . El resultado será de gran utilidad para el estudio de los parámetros en poblaciones normales.

Teorema 11. (de Fisher) Sea X_1, \dots, X_n una m.a.s. de una distribución teórica $N(\mu, \sigma)$. Se tiene que \bar{X} y S^2 son independientes, y que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Recuérdese que $(n-1)S^2 = ns^2$. Presentamos una demostración basada en el siguiente resultado sobre transformaciones lineales de variables aleatorias normales con vectores de coeficientes ortonormales. Los vectores considerados son vectores fila, y el superíndice t se utiliza para denotar la transposición.

Lema 12. Sean X_1, \dots, X_n v.a.i.d con distribución $N(0, \sigma)$ y $X = (X_1, \dots, X_n)$. Sean c_1, \dots, c_p vectores fila de \mathbb{R}^n ortonormales, de modo que $c_i c_j^t = 0$ y $c_i c_i^t = 1$ para todo i y $j \neq i$; con $p < n$. Para $i = 1, \dots, p$ consideremos la v.a. unidimensional $Y_i = c_i X^t$. Se tiene que la v.a.

$$Z = \sum_{i=1}^n X_i^2 - \sum_{i=1}^p Y_i^2$$

es independiente de Y_1, \dots, Y_p y que Z/σ^2 tiene distribución χ_{n-p}^2 .

Demostración. Podemos añadir $n-p$ vectores más c_{p+1}, \dots, c_n a los vectores c_1, \dots, c_p de modo que todos juntos formen una base ortonormal de \mathbb{R}^n . Estos vectores existen y no son únicos. Su elección es arbitraria y no hay que especificarla, y solo tienen utilidad como elemento auxiliar en la demostración; lo que importa aquí es, que existen, y su distribución, que como veremos no depende de la elección de c_{p+1}, \dots, c_n . Sea C la matriz ortogonal $n \times n$

$$C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix},$$

con $CC^t = I$ y $C^tC = I$, siendo I la matriz identidad. Sea $Y^t = CX^t$, de modo que sus p primeras componentes son los vectores Y_1, \dots, Y_p del enunciado. Puesto que X tiene distribución conjunta normal, e Y se obtiene como transformación lineal no singular de X , la distribución de Y también es normal n -dimensional. Calculamos el vector de esperanzas y la matriz de covarianzas de Y :

$$\begin{aligned} E[Y] &= E[XC^t] = E[X]C^t = \mathbf{0}C^t = \mathbf{0} \text{ y} \\ V[Y] &= E[Y^tY] = E[CX^tXC^t] = CE[X^tX]C^t \\ &= CV[X]C^t = C(\sigma^2I)C^t = \sigma^2CC^t = \sigma^2I, \end{aligned}$$

siendo $\mathbf{0}$ el n -vector fila de ceros, puesto que $E[X] = \mathbf{0}$ y la matriz de covarianzas de X es $V[X] = \sigma^2I$. Entonces, Y_1, \dots, Y_n son vaíid con distribución $N(0, \sigma)$ (igual que X_1, \dots, X_n).

Puesto que $Y = XC^t$ se tiene que $X = YC$, y de aquí:

$$Z = XX^t - \sum_{i=1}^p Y_i^2 = YCC^tY^t - \sum_{i=1}^p Y_i^2 = YY^t - \sum_{i=1}^p Y_i^2 = \sum_{i=p+1}^n Y_i^2.$$

Entonces, Z es independiente de Y_1, \dots, Y_p , y Z/σ^2 tiene distribución χ_{n-p}^2 , puesto que se puede expresar como suma de $n - p$ cuadrados de vaíid $N(0, 1)$. ■

Demostración. (teorema de Fisher) Desarrollando el cuadrado

$$(n - 1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n ((X_i - \mu) - (\bar{X} - \mu))^2,$$

y simplificando se obtiene

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

Aplicamos el lema anterior a $X_1 - \mu, \dots, X_n - \mu$, que son vaíid $N(0, \sigma)$, tomando $p = 1$ y $c_1 = \frac{1}{\sqrt{n}}(1, \dots, 1)$. Obsérvese que $c_1c_1^t = 1$. Se obtiene $Y_1 = \sqrt{n}(\bar{X} - \mu)$ y

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

Entonces, $(n - 1)S^2$ es independiente de $\sqrt{n}(\bar{X} - \mu)$, y por tanto S^2 es independiente de \bar{X} . Además, $\frac{(n-1)S^2}{\sigma^2}$ tiene distribución χ_{n-1}^2 . ■

Distribución del estadístico de Student

Si σ es desconocida, la distribución de \bar{X} , $N(\mu, \sigma/\sqrt{n})$, tiene varianza desconocida, σ^2/n .

Para estudiar μ en este caso en que σ es desconocida, será útil considerar el estadístico de Student, dado por

$$T = \frac{\bar{X} - \mu}{s} \sqrt{n - 1} = \frac{\bar{X} - \mu}{S} \sqrt{n}. \quad (1.10)$$

Teorema 13. (de Student) Se verifica que

$$T \sim t_{n-1} .$$

Demostración. Puesto que $Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$, $Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$, y son independientes (por el teorema de Fisher), se tiene que $\frac{Z}{\sqrt{Y/(n-1)}} \sim t_{n-1}$. Es inmediato comprobar que $\frac{Z}{\sqrt{Y/(n-1)}} = T$, lo que concluye la demostración. ■

Las propiedades de este estadístico que lo harán útil son que en él interviene $\bar{X} - \mu$, no interviene σ , y su distribución, t_{n-1} , no depende del parámetro desconocido (μ, σ) .

1.3.4 Distribuciones en el muestreo para dos poblaciones normales

Introducción

Consideramos ahora otras situaciones en las que se amplía el contexto considerado hasta ahora de una muestra de v.a. unidimensionales. Por una parte se consideran dos muestras de variables normales unidimensionales (problema de tipo 1, para una referencia rápida), y por otro una muestra de variables normales bidimensionales (problema de tipo 2).

- **Problema 1** Dos poblaciones normales independientes:

$$\begin{aligned} X_1, \dots, X_n & \text{ m.a.s. de una } N(\mu_1, \sigma_1) \text{ y} \\ Y_1, \dots, Y_m & \text{ m.a.s. de una } N(\mu_2, \sigma_2) , \end{aligned}$$

siendo las X_i independientes de las Y_j .

Por ejemplo, utilizando los fertilizantes A ó B la producción de cierto cereal es $N(\mu_A, \sigma_A)$ ó $N(\mu_B, \sigma_B)$ respectivamente. Un estudio de $\mu_A - \mu_B$ basado en $n + m$ cultivos en un vivero, n con A y m con B, nos permite comparar el rendimiento de ambos fertilizantes.

- **Problema 2** Muestras pareadas:

$$(X_1, Y_1), \dots, (X_n, Y_n) \text{ m.a.s. de una } N(\mu, \Sigma) ,$$

donde $\mu = (\mu_1, \mu_2)$ es el vector esperanzas y $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ es la matriz de covarianzas, con $\sigma_{12} = \rho\sigma_1\sigma_2$.

Los datos son bidimensionales. Por ejemplo, el nivel de colesterol de n pacientes antes y después de un tratamiento médico. Si $\mu_1 - \mu_2 > 0$ el tratamiento ha reducido el nivel.

Denotamos por \bar{X} , s_1^2 y S_1^2 a la media, varianza y cuasivarianza de las X_i , y por \bar{Y} , s_2^2 y S_2^2 a las de las Y_j .

Distribución del estadístico diferencia de medias

En este apartado obtenemos estadísticos que serán útiles para el estudio de $\mu_1 - \mu_2$. Puesto que el estudio de la media poblacional en una población normal está basada en la media muestral, el estudio de $\mu_1 - \mu_2$ estará basado en $\bar{X} - \bar{Y}$.

- Problema 1:

Se tiene que

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim N(0, 1) . \quad (1.11)$$

Ejercicio: Comprueba que $\bar{X} - \bar{Y}$ tipificada es justamente el estadístico en (1.11), con lo que queda comprobado (1.11).

En caso de ser σ_1, σ_2 desconocidas será útil considerar estadísticos en los que no inter vengan estos valores desconocidos.

Si σ_1, σ_2 son desconocidas, pero podemos admitir que $\sigma_1 = \sigma_2$, entonces será útil considerar el siguiente estadístico:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2} , \text{ con } S_p = \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}} . \quad (1.12)$$

Comprobamos que, efectivamente, la distribución de este estadístico es t_{n+m-2} . Supongamos que $\sigma_1 = \sigma_2 = \sigma$. Por el teorema de Fisher, se tiene que $\frac{(n-1)S_1^2}{\sigma^2} \sim \chi_{n-1}^2$ y $\frac{(m-1)S_2^2}{\sigma^2} \sim \chi_{m-1}^2$. Puesto que estas v.a. son independientes y que la distribución χ^2 es reproductiva, se tiene que la suma, $Y = \frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2}$, tiene distribución χ_{n+m-2}^2 . Sea $Z \sim N(0, 1)$ el estadístico en (1.11). Teniendo en cuenta que una muestra es independiente de la otra y el teorema de Fisher, se tiene que \bar{X} e \bar{Y} son independientes de S_1^2 y S_2^2 , y por tanto Z e Y son independientes. Entonces, $T = \frac{Z}{\sqrt{Y/(n+m-2)}} \sim t_{n+m-2}$. Simplificando, se obtiene que T es justamente el estadístico en (1.12).

Cuando σ_1 y σ_2 son desconocidas y no podemos suponer que sean iguales, sería útil disponer de estadísticos que involucren $\bar{X} - \bar{Y} - (\mu_1 - \mu_2)$ pero no σ_1 ni σ_2 , y cuya distribución no dependa de los parámetros desconocidos. Este problema ha sido objeto de estudio, pero sin soluciones completas para m.a.s. La distribución de los estadísticos disponibles depende "poco" de los parámetros para n grande, y se maneja de un modo aproximado.

Uno de estos estadísticos es (1.13). Por las propiedades de convergencia de S_1^2 y S_2^2 cuando n y m tienden a infinito, los valores que tomen en sus realizaciones deberían ser cercanos a σ_1^2 y σ_2^2 , para n y m "grandes" y entonces el estadístico (1.13) podría servir como aproximación del estadístico en (1.11), con distribución conocida $N(0, 1)$. Por tanto,

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n + S_2^2/m}} \quad (1.13)$$

tiene distribución aproximadamente $N(0, 1)$. Se suele considerar que la aproximación es buena con $n, m \geq 15$.

Para n ó m pequeños una aproximación es la siguiente: el estadístico dado en (1.13) tiene distribución aproximadamente t_k , siendo k el entero más próximo a

$$\frac{(S_1^2/n + S_2^2/m)^2}{\frac{1}{n+1} (S_1^2/n)^2 + \frac{1}{m+1} (S_2^2/m)^2} - 2.$$

- Problema 2: Buscamos estadísticos en los que intervenga $\bar{X} - \bar{Y} - (\mu_1 - \mu_2)$.

Consideramos

$$Z_i = X_i - Y_i \text{ para } i = 1, \dots, n.$$

Se tiene que Z_1, \dots, Z_n son vauid $N(\mu_Z, \sigma_Z)$ (puesto que Z_i es combinación lineal de normales), con

$$\begin{aligned} \mu_Z &= E[Z_i] = \mu_1 - \mu_2, \text{ y} \\ \sigma_Z^2 &= V[Z_i] = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}. \end{aligned}$$

Se obtiene $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i = \bar{X} - \bar{Y}$ y

$$\bar{Z} - \mu_Z = \bar{X} - \bar{Y} - (\mu_1 - \mu_2),$$

y entonces el estudio de $\mu_1 - \mu_2$ se puede realizar mediante el estudio de μ_Z a partir de las variables unidimensionales $Z_i = X_i - Y_i$.

Distribución del estadístico cociente de varianzas

En este apartado obtenemos un estadístico que será útil para el estudio de σ_2^2/σ_1^2 cuando μ_1 y μ_2 son desconocidos. Puesto que el estudio de la varianza poblacional en una población normal está basada en la varianza (o cuasivarianza) muestral, el estudio de σ_2^2/σ_1^2 estará basado en S_2^2/S_1^2 .

- Problema 1.

Consideramos un estadístico basado en S_1^2 y S_2^2 , que servirá para comparar σ_1^2 y σ_2^2 , con distribución F de Snedecor. Se tiene que

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n-1, m-1},$$

puesto que $\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi_{n-1}^2$, $\frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi_{m-1}^2$, y son independientes, y entonces

$$\frac{\frac{1}{n-1} \frac{(n-1)S_1^2}{\sigma_1^2}}{\frac{1}{m-1} \frac{(m-1)S_2^2}{\sigma_2^2}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

tiene distribución $F_{n,m}$.

Obsérvese que $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2/\sigma_1^2}{S_2^2/S_1^2}$.

Distribución del coeficiente de correlación muestral con $\rho = 0$

- Problema 2.

Puesto que la distribución teórica es normal bidimensional, la independencia de las componentes de (X_i, Y_i) es equivalente a la incorrelación ($\rho = 0$).

Para estudiar si las componentes son independientes, consideramos el parámetro muestral correspondiente al coeficiente de correlación poblacional $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$. El coeficiente de correlación muestral es

$$R = \frac{s_{12}}{s_1 s_2},$$

$$\text{siendo } s_{12} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}$$

la covarianza muestral.

El estadístico R converge, cuando n tiende a infinito, a una v.a. degenerada en ρ . Si $\rho = 0$, entonces R se distribuye alrededor de $\rho = 0$.

Para estudiar la independencia entre X e Y , resultará útil conocer la distribución de R bajo el supuesto de que $\rho = 0$. Se tiene que esta distribución no depende de los parámetros, y viene dada por el siguiente resultado: si $\rho = 0$ se tiene que

$$\sqrt{n-2} \frac{R}{\sqrt{1-R^2}} \sim t_{n-2}.$$

1.4 Estadísticos de orden

Para x_1, x_2, \dots, x_n reales sean $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ los mismos valores ordenados de menor a mayor. $x_{(1)}$ es el mínimo, $x_{(2)}$ el valor más pequeño después del mínimo, ..., y $x_{(n)}$ el máximo. Por ejemplo, si $n = 4$ y se obtiene $x_1 = 5, x_2 = 2, x_3 = 6$ y $x_4 = 2$, entonces $x_{(1)} = 2, x_{(2)} = 2, x_{(3)} = 5$ y $x_{(4)} = 6$.

A las v.a. $X_{(i)}$ se las denomina *estadísticos de orden*.

Consideremos la v.a. n -dimensional

$$(X_{(1)}, \dots, X_{(n)}),$$

función de (X_1, \dots, X_n) , que toma el valor $(X_{(1)}, \dots, X_{(n)}) = (x_{(1)}, \dots, x_{(n)})$ si $(X_1, \dots, X_n) = (x_1, \dots, x_n)$.

Los cuantiles muestrales se pueden expresar en función de los estadísticos de orden. Por ejemplo, para la mediana, si n es impar se tiene que $c_{1/2} = X_{(\frac{n+1}{2})}$, y si n es par $c_{1/2} = \frac{1}{2} (X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)})$.

Tengase en cuenta que $X_{(i)}$ toma el valor de alguna X_j , pero para distintos valores de j según la muestra obtenida. Además, aunque X_1, \dots, X_n son v.a.i.d., $X_{(1)}, \dots, X_{(n)}$ no son ni independientes ni idénticamente distribuidas.

Si la distribución teórica (la de las X_i) es continua con función de densidad f , entonces la v.a. $(X_{(1)}, \dots, X_{(n)})$ es continua con función de densidad

$$g(y_1, \dots, y_n) = n! f(y_1) \cdots f(y_n) \text{ para } (y_1, \dots, y_n) \in \Delta \text{ con} \\ \Delta = \{(y_1, \dots, y_n) \in \mathbb{R}^n : y_1 < y_2 < \cdots < y_n\}.$$

La función de densidad (marginal) de $X_{(i)}$ es

$$g_i(y) = n! f(y) \frac{[F(y)]^{i-1} [1 - F(y)]^{n-i}}{(i-1)! (n-i)!}.$$

Tema 2

Principios de reducción de datos. Introducción. El principio de suficiencia: estadísticos suficientes, suficientes minimales, auxiliares (ancillary) y completos. El principio de verosimilitud: la función de verosimilitud y el principio formal de verosimilitud. El principio de equivarianza.

Los contenidos de este tema corresponden a un curso básico de Estadística Matemática, y pueden ser consultados y completados con los libros de la bibliografía por ejemplo. El libro de Rohatgi y Ehsanes es el más completo.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

2.1 Introducción

En este tema consideramos solo el caso paramétrico. La distribución teórica F , sobre la que queremos obtener información, sabemos que pertenece a una familia paramétrica $\{F_\theta : \theta \in \Theta\}$. Entonces, para obtener información sobre $F = F_\theta$ hay que obtener información para el parámetro θ , a partir de los datos observados de una m.a.s. X_1, \dots, X_n . Para ello, no se suele utilizar la muestra en bruto, el dato n -dimensional observado, sino estadísticos, funciones de ella.

Estudiaremos qué estadísticos asimilan adecuadamente la información que aporta la muestra sobre el parámetro. En el próximo tema trataremos el problema de la estimación puntual: el estudio de estadísticos cuyas realizaciones muestrales (estimaciones) tomen, frecuentemente, valores cercanos a los del verdadero valor del parámetro θ . Comprobaremos que en muchos casos los estimadores puntuales adecuados son funciones de estos estadísticos que asimilan bien esta información.

Hablamos aquí de información de un modo impreciso, para ayudar a la comprensión intuitiva de los conceptos, y no estamos midiéndola. Lo que sí podemos precisar es cuándo esa información sobre θ es nula, y con un cambio de enfoque ingenioso, cuándo es completa.

La idea básica acerca de si un estadístico T puede aportar información sobre θ o no es la siguiente: si la distribución de T es independiente de θ (en el sentido de que es la misma siempre, no es diferente para los diferentes valores que pueda tomar θ en Θ), entonces T no puede aportar información sobre θ (es un estadístico auxiliar). Por tanto, para que T pueda aportar información sobre θ , es necesario que su distribución no sea siempre la

misma para los distintos valores $\theta \in \Theta$, sino que dependa de θ . El valor observado de T ayudará a realizar inferencias para θ , a discernir cuál puede ser aproximadamente el verdadero valor de θ , de cuál de esas distribuciones F_θ puede provenir esa observación. Por ejemplo, \bar{X} se distribuye alrededor de μ , en el sentido de que $E[\bar{X}] = \mu$, y sirve para obtener información sobre μ .

Una aplicación de esta idea a cierta distribución condicionada permite definir de un modo preciso, y obtener, aquellos estadísticos que, sin pérdida de información, resumen toda la información que aporta la muestra sobre el parámetro desconocido (estadísticos suficientes).

Estudiaremos la noción de completitud. Un estadístico suficiente y completo es siempre independiente de un estadístico auxiliar.

Definiremos algunos principios, que son presupuestos básicos en el manejo de la información, y que puede ser conveniente seguir. Es opcional, puesto que no hay una razón matemática que obligue a seguir unos u otros en este asunto práctico (aunque con herramientas matemáticas) de la inferencia estadística, aunque las propiedades matemáticas de uno u otro pueden aconsejarlos.

2.2 Estadísticos suficientes

A partir de los datos obtenidos como resultados (realizaciones) de un experimento aleatorio, queremos obtener información sobre el verdadero valor del parámetro θ , fijo pero desconocido. Sería útil poder resumir la información que aporta la muestra, que es un estadístico n -dimensional, mediante otro estadístico más simple, pero sin pérdida de información. Un estadístico suficiente cumple esta función.

Un estadístico puede aportar información sobre el valor del parámetro desconocido θ cuando su distribución depende de θ . Aplicamos esta idea al caso en que el estadístico es la muestra, pero en vez de su propia distribución consideramos su distribución condicionada con el valor de otro estadístico.

Si la distribución de la muestra, condicionada a que un estadístico T ha tomado un valor t , es independiente de θ para todo t en el soporte de T , entonces, sabiendo que $T = t$, la muestra no puede aportar más información sobre θ . La conclusión es que con T se agota la información que aporta la muestra sobre θ . La muestra no dispone de más información que la que aporta a través de T , y entonces, si conocemos qué valor t ha tomado T , la muestra concreta con la que se ha obtenido ese valor t es irrelevante para obtener información sobre θ (y también funciones de la muestra).

Definición 14. Un estadístico $T = T(X_1, \dots, X_n)$ se denomina estadístico suficiente para la familia paramétrica $\{F_\theta : \theta \in \Theta\}$ si la distribución de la muestra condicionada por el estadístico es independiente de θ .

Para hacer inferencias sobre θ se puede utilizar simplemente el valor $T(x_1, \dots, x_n)$ de un estadístico suficiente, sin utilizar la muestra completa x_1, \dots, x_n . Por ejemplo, para hacer inferencias sobre el parámetro de una distribución de Bernoulli no hace falta

disponer de la secuencia completa de ceros y unos obtenida, sino solamente del número de unos:

Ejemplo 15. Consideremos una muestra X_1, \dots, X_n de una distribución teórica $B(1, \theta)$ (Bernoulli), y sea $T = \sum_{i=1}^n X_i$. La distribución de la muestra (X_1, \dots, X_n) condicionada a $T = t$ no depende de θ , y por tanto T es un estadístico suficiente.

Para comprobarlo, calculamos la función de masa de la distribución de la muestra condicionada a $T = t$, con $t \in \{0, 1, \dots, n\}$. Para $x_1, \dots, x_n = 0, 1$ se tiene que:

$$P\{X_1 = x_1, \dots, X_n = x_n / T = t\} = \frac{P_\theta\{X_1 = x_1, \dots, X_n = x_n, T = t\}}{P_\theta\{T = t\}}, \text{ y}$$

$$P\{X_1 = x_1, \dots, X_n = x_n, T = t\} = \begin{cases} P_\theta\{X_1 = x_1, \dots, X_n = x_n\} & \text{si } \sum_{i=1}^n x_i = t \\ 0 & \text{si } \sum_{i=1}^n x_i \neq t \end{cases}$$

Teniendo en cuenta que $T \equiv B(n, \theta)$ (binomial), si $\sum_{i=1}^n x_i = t$ se tiene que

$$P\{X_1 = x_1, \dots, X_n = x_n / T = t\} = \frac{P_\theta\{X_1 = x_1, \dots, X_n = x_n\}}{P_\theta\{T = t\}}$$

$$= \frac{\theta^{x_1 + \dots + x_n} (1 - \theta)^{n - (x_1 + \dots + x_n)}}{\binom{n}{t} \theta^t (1 - \theta)^{n-t}} = \frac{1}{\binom{n}{t}},$$

y si $\sum_{i=1}^n x_i \neq t$ se tiene que $P\{X_1 = x_1, \dots, X_n = x_n / T = t\} = 0$. Obtenemos entonces que la distribución de (X_1, \dots, X_n) condicionada a $T = t$ es uniforme discreta en el conjunto $A_t = \{T = t\}$, que es $A_t = \{(x_1, \dots, x_n) \in \{0, 1\}^n : \sum_{i=1}^n x_i = t\}$, con cardinal $\binom{n}{t}$.

En particular, para todo t esta distribución condicionada no depende de θ , y por tanto T es un estadístico suficiente, que es lo que queríamos comprobar. Dado $T = t$ (número de unos), la información restante que proporciona la muestra es el orden de los ceros y unos, pero, conociendo el valor de t , esta información no es útil para hacer inferencias sobre θ .

Sin embargo, no es necesario realizar el cálculo de esta distribución condicionada para determinar si un estadístico es suficiente o no, lo que además es mas complejo en el caso continuo. El siguiente teorema permite obtener estadísticos suficientes a partir de la estructura de la función de masa o densidad de la muestra.

Teorema 16. (de factorización) Un estadístico $T(X_1, \dots, X_n)$ es suficiente si y solo si existen funciones de (x_1, \dots, x_n) reales positivas h y g_θ tales que la función de masa o densidad de la muestra se puede expresar como

$$f_\theta(x_1, \dots, x_n) = h(x_1, \dots, x_n) g_\theta(T(x_1, \dots, x_n)), \quad (2.1)$$

siendo g_θ una función que solo depende de la muestra a través de T y h una función que no depende de θ .

Ejemplo 17. Obtén estadísticos suficientes, mediante el teorema de factorización, para las distribuciones a) $B(1, \theta)$, b) $\mathcal{P}(\theta)$, c) $U(0, \theta)$.

Solución 18. a) Los argumentos x_1, \dots, x_n intervienen en la función de masa a través de $t = \sum_{i=1}^n x_i$, y esto lleva a obtener que $T = \sum_{i=1}^n X_i$ es un estadístico suficiente:

$$f_{\theta}(x_1, \dots, x_n) = \theta^{x_1 + \dots + x_n} (1 - \theta)^{n - (x_1 + \dots + x_n)} = \theta^t (1 - \theta)^{n-t},$$

y entonces $f_{\theta}(x_1, \dots, x_n)$ admite una expresión del tipo (2.1), con $g_{\theta}(t) = \theta^t (1 - \theta)^{n-t}$, siendo $t = T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$, y $h = 1$ (constante).

b) La función de masa es

$$f_{\theta}(x_1, \dots, x_n) = e^{-n\theta} \frac{\theta^{x_1 + \dots + x_n}}{x_1! \dots x_n!} = e^{-n\theta} \frac{\theta^t}{x_1! \dots x_n!}, \text{ con } t = \sum_{i=1}^n x_i,$$

y entonces $f_{\theta}(x_1, \dots, x_n)$ admite una expresión del tipo (2.1), con $g_{\theta}(t) = e^{-n\theta} \theta^t$, siendo $t = T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$, y $h(x_1, \dots, x_n) = \frac{1}{x_1! \dots x_n!}$. Por tanto $T = \sum_{i=1}^n X_i$ es un estadístico suficiente también en este caso.

c) La función de densidad es, si $x_1, \dots, x_n > 0$,

$$f_{\theta}(x_1, \dots, x_n) = \theta^{-n} I\{x_{(n)} < \theta\},$$

que solo depende de la muestra a través del máximo $x_{(n)}$. Entonces $T = X_{(n)}$ es un estadístico suficiente, puesto que podemos expresar $f_{\theta}(x_1, \dots, x_n)$ como en (2.1), con $g_{\theta}(t) = \theta^{-n} I\{t < \theta\}$, siendo $t = T(x_1, \dots, x_n) = X_{(n)}$, y $h = 1$.

El estadístico suficiente no es único. Dado un estadístico suficiente T , cualquier otro estadístico que aporte al menos la misma información que T es suficiente también. Esta información adicional no es sobre θ , porque ya T agota esta información.

Por ejemplo, la muestra completa, $T = (X_1, \dots, X_n)$, es siempre un estadístico suficiente. Para comprobarlo, simplemente hay que tener en cuenta que, obviamente, la muestra agota toda la información que posee la muestra sobre θ ; o que la distribución de (X_1, \dots, X_n) condicionada con $T = t$, con $t = (x_1, \dots, x_n)$, es degenerada en t , y entonces no depende de θ ; o, utilizando el teorema de factorización, que $f_{\theta}(x_1, \dots, x_n)$ admite una expresión del tipo (2.1), con $h = 1$ y $g_{\theta} = f_{\theta}$.

La muestra ordenada $T = (X_{(1)}, X_{(2)}, \dots, X_{(n)})$ también es un estadístico suficiente. Se tiene que f_{θ} es simétrica en sus argumentos, esto es, el valor de f_{θ} es el mismo en (x_1, \dots, x_n) que en una permutación de (x_1, \dots, x_n) , puesto que el orden de los factores no altera el producto en la expresión de $f_{\theta}(x_1, \dots, x_n) = f_{\theta}(x_1) \dots f_{\theta}(x_n)$. Entonces, f_{θ} no depende del orden de las observaciones x_1, \dots, x_n y admite una expresión del tipo (2.1):

$$f_{\theta}(x_1, \dots, x_n) = f_{\theta}(T(x_1, \dots, x_n)),$$

con $T(x_1, \dots, x_n) = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$, $g_{\theta} = f_{\theta}$, y $h = 1$. Por el teorema de factorización, T es un estadístico suficiente (siempre, para cualquier familia paramétrica).

Si T es suficiente y U es un estadístico cualquiera, entonces el estadístico (T, U) es suficiente también. Ésto es así porque (T, U) contiene al menos la misma información

sobre θ que T , que agota por si solo toda la información sobre θ . Además, siendo T suficiente, U no puede aportar información sobre θ , como componente del par (T, U) , y entonces conviene basar el estudio de θ directamente en T , y no en (T, U) .

La idea de que un estadístico aporte al menos la misma información que otro se puede explicar y formalizar de una manera sencilla. Y sirve para definir un tipo de estadísticos suficientes: los que resumen la información al máximo, incluyendo solo la que es relevante para el estudio de θ .

Cuando se procesan unos datos se obtienen otros que son resumen de ellos, quizá sin pérdida de información y solo transformados de un modo equivalente. Si los datos originales son el valor que ha tomado un estadístico T (la misma muestra, o el par de valores suma y suma de cuadrados por ejemplo), el algoritmo de procesamiento de los datos define una función φ que transforma los datos originales, del estadístico T , en los datos procesados, que será el valor de un estadístico U . De este modo, $U = \varphi(T)$.

De este modo, si T y U son dos estadísticos tales que existe una función φ con $U = \varphi(T)$, entonces T incluye al menos tanta información como U (sobre θ o no, como es la relativa al orden de los resultados en pruebas de Bernoulli, según se explicaba en el ejemplo 15).

Si además φ es biyectiva, existe la aplicación inversa φ^{-1} , y entonces también $T = \varphi^{-1}(U)$, y por tanto T y U incluyen la misma información. Este es el caso cuando de los datos procesados podemos recuperar los originales. Por ejemplo, $T = \sum_{i=1}^n X_i$ y \bar{X} incluyen la misma información, puesto que $\bar{x} = t/n$ y $t = n\bar{x}$, y de este modo a partir de uno podemos obtener el otro y viceversa.

Cada estadístico T define una partición $\mathcal{A} = \{A_t : t \in \text{sop}(T)\}$ del espacio muestral Ω con

$$A_t = \{T = t\} = \{(x_1, \dots, x_n) \in \Omega : T(x_1, \dots, x_n) = t\} .$$

Por ejemplo, si $n = 2$ y $T(x_1, x_2) = x_1 + x_2$, entonces A_t es la intersección de la recta $x_1 + x_2 = t$ del plano con Ω , para cada t en el soporte de T . Si $T = (X_1, \dots, X_n)$, entonces $\text{sop}(T) = \Omega$, cada A_t es un conjunto unitario $A_\omega = \{\omega\}$, y la partición es $\{\{\omega\} : \omega \in \Omega\}$.

Consideremos estadísticos T y U y sean \mathcal{A}_T y \mathcal{A}_U sus particiones asociadas. Si $T = \varphi(U)$, entonces $\mathcal{A}_T \subset \mathcal{A}_U$, esto es, \mathcal{A}_T incluye no mas conjuntos que \mathcal{A}_U , lo que expresamos diciendo que la partición \mathcal{A}_T es menos fina o igual que \mathcal{A}_U . Si φ es biyectiva, entonces $\mathcal{A}_T = \mathcal{A}_U$ (con $A_{T, \varphi(u)} = A_{U, u}$, puesto que $\{T = \varphi(u)\} = \{U = u\}$). Si φ no es biyectiva y $\varphi(t_1) = \varphi(t_2) = u$, entonces $A_{T, t_1} \cup A_{T, t_2} \subset A_{U, u}$ y por tanto $\mathcal{A}_U \neq \mathcal{A}_T$; en este caso, los valores t_1 y t_2 para el estadístico T , correspondientes a distintos grupos de muestras A_{T, t_1} y A_{T, t_2} , no quedan diferenciados por U , que pierde esta información. Además, $A_{U, u} = \bigcup_{t, \varphi(t)=u} A_{T, t}$.

Las particiones asociadas a estadísticos suficientes se denominan particiones suficientes. En este caso, la distribución de la muestra condicionada con $A_t = \{T = t\}$ no depende de θ , para todo t . Es deseable encontrar el estadístico suficiente con menor información, para eliminar información que sea irrelevante para el estudio de θ . Esto se consigue con

un estadístico suficiente cuya partición \mathcal{A} sea lo menos fina posible, según se define a continuación:

Definición 19. Un estadístico T es minimal suficiente si su partición asociada es suficiente y es lo menos fina posible entre todas las particiones suficientes, esto es, si para todo estadístico suficiente U se tiene que existe φ con $T = \varphi(U)$.

El estadístico minimal suficiente (abreviado, estadístico ms o ems) es único salvo transformaciones biyectivas (una transformación biyectiva preserva la información). Una manera sencilla de averiguar si existe una transformación biyectiva que transforma un estadístico T_1 en otro T_2 , es estudiar si T_2 se puede expresar en función de T_1 , y en caso afirmativo, comprobar si se puede despejar T_1 con unicidad en esta expresión.

Habitualmente, es posible identificar el estadístico ms a partir de la factorización del teorema de factorización. Sin embargo, es conveniente disponer de un procedimiento general para su obtención:

Teorema 20. Se tiene que un estadístico $T(X_1, \dots, X_n)$ es minimal suficiente si y solo si el cociente

$$\frac{f_{\theta}(x_1, \dots, x_n)}{f_{\theta}(x'_1, \dots, x'_n)} \quad (2.2)$$

no depende de θ cuando $T(x_1, \dots, x_n) = T(x'_1, \dots, x'_n)$ y sí depende de θ en caso contrario, cuando $T(x_1, \dots, x_n) \neq T(x'_1, \dots, x'_n)$.

La demostración está basada en la relación de equivalencia en Ω para la que dos puntos (x_1, \dots, x_n) y (x'_1, \dots, x'_n) son equivalentes si el cociente (2.2) no depende de θ . Se obtiene que la partición formada por las clases de equivalencia es minimal suficiente. Un estadístico T con las condiciones del enunciado tiene la misma partición asociada, y entonces debe ser ms.

Se entiende aquí y a continuación que (x_1, \dots, x_n) y (x'_1, \dots, x'_n) están en Ω , donde el denominador de (2.2) es positivo (casi seguro para F_{θ}).

Ejemplo 21. Determina el estadístico minimal suficiente para las siguientes distribuciones teóricas: a) $B(1, \theta)$, b) $\mathcal{P}(\theta)$, c) $Exp(\theta)$, d) $U(0, \theta)$.

Solución 22. Denotamos por C_{θ} al cociente $\frac{f_{\theta}(x_1, \dots, x_n)}{f_{\theta}(x'_1, \dots, x'_n)}$. Escribimos $n\bar{x}$ en vez de $\sum_{i=1}^n x_i$, puesto que es mas breve. La clave de la obtención del estadístico ms es identificar en C_{θ} los factores en los que no se puede separar θ de la muestra (x_1, \dots, x_n) , de modo que estos factores no se pueden expresar a su vez como un producto de una función de θ y una función de la muestra. Los estadísticos a través de los cuales se expresan estos factores, en los que no se puede separar θ de la muestra, son los estadísticos ms buscados. Estos factores, son potencias en los apartados a, b y c, y es una función indicatriz en d.

a) Se tiene que

$$C_{\theta} = \frac{\theta^{n\bar{x}} (1 - \theta)^{n - n\bar{x}}}{\theta^{n\bar{x}'} (1 - \theta)^{n - n\bar{x}'}} = \theta^{n(\bar{x} - \bar{x}')} (1 - \theta)^{-n(\bar{x} - \bar{x}')} = \left(\frac{\theta}{1 - \theta} \right)^{n(\bar{x} - \bar{x}')},$$

que no depende de θ cuando $\bar{x} = \bar{x}'$ (vale 1, puesto que $0 < \theta < 1$, y entonces $\frac{\theta}{1-\theta} > 0$), y sí depende de θ cuando $\bar{x} \neq \bar{x}'$. Por tanto, \bar{X} es ms. Aplicando a \bar{X} la función biyectiva $\varphi(t) = nt$ se obtiene que $\sum_{i=1}^n X_i$ es también ms. Ya sabíamos por los ejemplos 15 y 17a que T es suficiente, y ahora hemos comprobado que es además ms.

b) Se tiene que

$$C_\theta = \frac{e^{-n\theta} \frac{\theta^{n\bar{x}}}{x_1! \cdots x_n!}}{e^{-n\theta} \frac{\theta^{n\bar{x}'}}{x_1! \cdots x_n!}} = \theta^{n(\bar{x}-\bar{x}')} \frac{x_1! \cdots x_n!}{x_1! \cdots x_n!},$$

que no depende de θ cuando $\bar{x} = \bar{x}'$ ($\theta^0 = 1$), y sí depende de θ cuando $\bar{x} \neq \bar{x}'$. Por tanto, \bar{X} , ó $\sum_{i=1}^n X_i$, es ms.

c) Se tiene que

$$C_\theta = \frac{\theta^n e^{-\theta n\bar{x}}}{\theta^n e^{-\theta n\bar{x}'}} = e^{-n\theta(\bar{x}-\bar{x}')},$$

que no depende de θ cuando $\bar{x} = \bar{x}'$, y sí depende de θ cuando $\bar{x} \neq \bar{x}'$. Por tanto, \bar{X} , ó $\sum_{i=1}^n X_i$, es ms.

d) Se tiene que

$$C_\theta = \frac{\theta^{-n} I \{x_{(n)} < \theta\}}{\theta^{-n} I \{x'_{(n)} < \theta\}} = \frac{I \{x_{(n)} < \theta\}}{I \{x'_{(n)} < \theta\}}.$$

Se produce aquí una dificultad por depender de θ el soporte $[0, \theta]$, que se puede solventar considerando $\frac{0}{0} = 1$. De todos modos, obtendremos el resultado mas adelante de una manera alternativa. Cuando $x'_{(n)} = x_{(n)}$ se tiene que $C_\theta = 1$, y por tanto no depende de θ . Cuando $x'_{(n)} \neq x_{(n)}$, C_θ sí depende de θ . Entonces $X_{(n)}$ es ms.

Ejemplo 23. Determina el estadístico ms para una $N(\mu, \sigma)$ en los siguientes casos:

- $\theta = \mu$ (con $\sigma = \sigma_0$ conocida),
- $\theta = \sigma$ (con $\mu = \mu_0$ conocida),
- $\theta = (\mu, \sigma)$.

Solución 24. La función de densidad de la muestra es

$$f_\theta(x_1, \dots, x_n) = (2\pi)^{-n/2} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

a) Desarrollando el cuadrado y simplificando, se obtiene

$$f_\mu(x_1, \dots, x_n) = (2\pi)^{-n/2} \sigma_0^{-n} \exp \left\{ -\frac{1}{2\sigma_0^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right) \right\},$$

y de aquí

$$\begin{aligned} C_\mu &= \frac{(2\pi)^{-n/2} \sigma_0^{-n} \exp \left\{ -\frac{1}{2\sigma_0^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right) \right\}}{(2\pi)^{-n/2} \sigma_0^{-n} \exp \left\{ -\frac{1}{2\sigma_0^2} \left(\sum_{i=1}^n x_i'^2 - 2\mu \sum_{i=1}^n x_i' + n\mu^2 \right) \right\}} \\ &= \exp \left\{ -\frac{1}{2\sigma_0^2} \left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i'^2 - 2n\mu(\bar{x} - \bar{x}') \right) \right\}, \end{aligned}$$

que no depende de μ cuando $\bar{x} = \bar{x}'$, y sí depende de μ cuando $\bar{x} \neq \bar{x}'$. Por tanto, \bar{X} , ó $\sum_{i=1}^n X_i$, es ms.

b) Se tiene que

$$\begin{aligned} C_\sigma &= \frac{(2\pi)^{-n/2} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right\}}{(2\pi)^{-n/2} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x'_i - \mu_0)^2 \right\}} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x'_i - \mu_0)^2 \right) \right\}. \end{aligned}$$

Téngase en cuenta que μ_0 es una cantidad conocida. Consideremos el estadístico $T(x_1, \dots, x_n) = \sum_{i=1}^n (x_i - \mu_0)^2$. Se tiene que C_σ no depende del parámetro desconocido σ cuando $T(x_1, \dots, x_n) = T(x'_1, \dots, x'_n)$, y sí depende de σ cuando $T(x_1, \dots, x_n) \neq T(x'_1, \dots, x'_n)$. Por tanto, $T = T(X_1, \dots, X_n)$ es ms. Entonces, $\frac{1}{n}T = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2$ también es un estadístico ms.

c) Se tiene que

$$\begin{aligned} C_{\mu, \sigma} &= \frac{(2\pi)^{-n/2} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right) \right\}}{(2\pi)^{-n/2} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i'^2 - 2\mu \sum_{i=1}^n x_i' + n\mu^2 \right) \right\}} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i'^2 - 2\mu \left(\sum_{i=1}^n x_i - \sum_{i=1}^n x_i' \right) \right) \right\}. \end{aligned}$$

Sea $T(x_1, \dots, x_n) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right)$. Se tiene que el cociente $C_{\mu, \sigma}$ no depende del parámetro $\theta = (\mu, \sigma)$ cuando $T(x_1, \dots, x_n) = T(x'_1, \dots, x'_n)$, y sí depende de θ cuando $T(x_1, \dots, x_n) \neq T(x'_1, \dots, x'_n)$. Por tanto,

$$T(X_1, \dots, X_n) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)$$

es ms. Se tiene que el estadístico

$$(\bar{X}, s^2)$$

también es ms (y también lo es (\bar{X}, S^2)), puesto que se transforma en

$T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right) = (T_1, T_2)$ biyectivamente: se tiene que

$$\bar{X} = \frac{T_1}{n} \quad \text{y} \quad s^2 = \frac{T_2}{n} - \left(\frac{T_1}{n} \right)^2,$$

y despejando se obtiene $T_1 = n\bar{X}$ y $T_2 = n \left(s^2 + \bar{X}^2 \right)$. Puesto que se ha despejado con unicidad, entonces, efectivamente, la transformación es biyectiva y (\bar{X}, s^2) es ms.

2.3 Estadísticos auxiliares y completos

2.3.1 Estadísticos auxiliares

Definición 25. Consideremos una m.a.s. de una distribución teórica paramétrica F_θ . Un estadístico $U = U(X_1, \dots, X_n)$ se denomina *estadístico auxiliar para θ* si su distribución es independiente de θ .

Ejemplo 26. Consideremos una m.a.s. de tamaño $n \geq 2$ de una población $N(\theta, \sigma_0)$, con desviación típica $\sigma = \sigma_0$ conocida. El estadístico $U(X_1, \dots, X_n) = X_2 - X_1$ tiene distribución $N(0, \sqrt{2}\sigma_0)$, que no depende del parámetro desconocido θ , y entonces U es un estadístico auxiliar. El estadístico S^2 también es auxiliar en este caso: por el teorema de Fisher $(n-1) \frac{S^2}{\sigma_0^2}$, y entonces la distribución de S^2 es independiente del parámetro desconocido θ .

Un estadístico auxiliar no puede aportar información sobre θ como estadístico individual. Pero hay que matizar que en algunos casos sí si se considera conjuntamente con otro estadístico (si este último no es suficiente, porque en este caso no quedaría información que aportar). En el siguiente ejemplo se presenta un estadístico suficiente bidimensional, una de cuyas componentes es un estadístico no suficiente y la otra un estadístico auxiliar.

Ejemplo 27. Consideremos una m.a.s. de una distribución teórica $U(0, \theta)$. El estadístico $T = \left(X_{(1)}, \frac{X_{(n)}}{X_{(1)}} \right)$ es suficiente, puesto que $\varphi(T) = X_{(n)}$ es suficiente (ms además), con $\varphi(t_1, t_2) = t_1 t_2$.

Si consideramos las componentes individuales de T , se observa que $X_{(1)}$ no es suficiente, ya que no hay ninguna función de $X_{(1)}$ que lo aplique en el ems $X_{(n)}$, y, como comentaremos mas adelante, se tiene que $\frac{X_{(n)}}{X_{(1)}}$ es auxiliar. La conclusión es que la información sobre θ incluida en $X_{(1)}$, que no es toda la incluida en la muestra por no ser $X_{(1)}$ suficiente, queda complementada con la que aporta el estadístico auxiliar, hasta tener toda la información sobre θ . Puesto que la distribución marginal de $\frac{X_{(n)}}{X_{(1)}}$ no depende de θ , la información que aporta para θ en el par T vendrá dada a partir del condicionamiento entre las dos componentes, aunque en la operación que hemos hecho con φ esa circunstancia está solo implícita.

Para obtener la distribución de $\frac{X_{(n)}}{X_{(1)}}$, se puede obtener su función de distribución mediante condicionamiento con $X_{(1)}$ (o con $X_{(n)}$). O se puede utilizar el teorema de cambio de variable aleatoria para obtener la función de densidad de (U, V) con $U = X_{(1)}$ y $V = \frac{X_{(n)}}{X_{(1)}}$ a partir de la densidad conjunta de $(X_{(1)}, X_{(n)})$, y calcular después la función de densidad marginal de V . Se obtiene

$$f_V(v) = (n-1)v^{-2}(1-v^{-1})^{n-2} \text{ para } v \geq 1,$$

(y 0 en el resto $v < 1$, se sobreentiende). Puesto que esta distribución no depende de θ , se tiene que $V = \frac{X_{(n)}}{X_{(1)}}$ es un estadístico auxiliar.

2.3.2 Estadísticos completos

Si un estadístico suficiente satisface además otra propiedad, la completitud, entonces es independiente de cualquier estadístico auxiliar.

Definición 28. Una familia paramétrica $\{F_\theta : \theta \in \Theta\}$ se dice que es completa si cualquier función g que satisfaga

$$E_\theta[g(X)] = 0 \text{ para todo } \theta \in \Theta$$

debe ser necesariamente la función idénticamente nula casi seguro (respecto F_θ), esto es,

$$P_\theta \{g(X) = 0\} = 1, \text{ para todo } \theta \in \Theta.$$

Entonces, una familia es completa si ninguna función de $X \sim F_\theta$ no idénticamente nula tiene esperanza nula, para todo θ .

Definición 29. Un estadístico $T = T(X_1, \dots, X_n)$ es completo si la familia de sus distribuciones $\{F_{T,\theta} : \theta \in \Theta\}$ es completa, esto es, si

$$E_\theta [g(T)] = 0 \text{ para todo } \theta \in \Theta$$

implica que g debe ser necesariamente la función idénticamente nula casi seguro.

Entonces, si T es completo, ninguna función de T no idénticamente nula tiene esperanza nula, para todo θ (casi seguro respecto de F_θ y de $F_{T,\theta}$).

Ejemplo 30. Consideremos una m.a.s. X_1, \dots, X_n de una distribución $B(1, \theta)$. Se tiene que $T = \sum_{i=1}^n X_i$ es un estadístico completo, como comprobamos a continuación. Puesto que $T \sim B(n, \theta)$, se tiene que

$$\begin{aligned} E_\theta [g(T)] &= \sum_{t=0}^n g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} \\ &= (1-\theta)^n \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t. \end{aligned}$$

El sumatorio es un polinomio en $\frac{\theta}{1-\theta}$, y para que sea idénticamente nulo para todo θ necesariamente los coeficientes $g(t) \binom{n}{t}$ del polinomio deben ser nulos, y entonces $g(t) = 0$ para $t = 0, 1, \dots, n$.

Ejemplo 31. Consideremos una distribución teórica normal $N(0, \theta)$ (por ejemplo, el error de medida de un instrumento bien calibrado), con $\theta > 0$.

Esta familia paramétrica no es completa o, equivalentemente, el estadístico $X \sim N(0, \theta)$ no es completo ($X = X_1$). Tomando $g(x) = x$, no idénticamente nula, se tiene que $E_\theta [g(X)] = E_\theta [X] = 0$ para todo $\theta \in \Theta$.

Sin embargo, el estadístico X^2 sí es completo para esta familia paramétrica. Se tiene que $\left(\frac{X}{\theta}\right)^2 \sim \chi_1^2$. Expresando a partir de ello $E_\theta [g(X^2)]$ como una integral, en términos de la función de densidad χ_1^2 , se obtiene una expresión en términos de una transformada de Laplace. Si $E_\theta [g(X^2)] = 0$ para todo $\theta \in \Theta$ esta transformada es idénticamente nula, y entonces debe ser transformada de la función idénticamente nula, lo que permite establecer que g debe ser idénticamente nula.

Teorema 32. Un estadístico suficiente y completo es minimal suficiente.

Teorema 33. (Basu) Un estadístico suficiente y completo es independiente de cualquier estadístico auxiliar.

Entonces, un estadístico suficiente y completo es minimal suficiente e independiente de cualquier estadístico auxiliar. Pero hay casos en los que un estadístico minimal suficiente no es completo.

2.3.3 Familias paramétricas de tipo exponencial

La función de masa o densidad de muchas de las distribuciones de probabilidad mas utilizadas tiene una estructura similar: se expresa como productos de potencias de funciones de la variable x y del parámetro. Por ejemplo, la distribución de Bernoulli, la Poisson, la exponencial, la normal, la gamma y la beta son de este tipo. No así las distribuciones uniformes. Se pueden obtener resultados generales que se aplican a cualquier familia paramétrica de este tipo, como por ejemplo una fórmula para la obtención del estadístico ms, que es además completo en este tipo de familias.

Definición 34. Se dice que una familia k -paramétrica de distribuciones, esto es, con $\Theta \subset \mathbb{R}^k$, es una familia de tipo exponencial k -paramétrico, si existen funciones reales de variable real c, h, q_1, \dots, q_k y V_1, \dots, V_k tales que la función de masa o densidad admite una expresión de la forma

$$f_{\theta}(x) = c(\theta) h(x) \exp \left\{ \sum_{j=1}^k q_j(\theta) V_j(x) \right\}.$$

Suponemos que no existen relaciones lineales entre las $q_j(\theta)$, ni entre las $V_j(x)$. En este caso podríamos reducir el número de sumandos en el sumatorio, tomando factor común, lo que resulta en que de los k parámetros algunos son redundantes y pueden ser eliminados. El nuevo espacio paramétrico Θ , con dimensión menor que k , contiene un conjunto abierto, de modo que no puede ser expresado mediante un número menor de parámetros.

El término "exponencial" en el nombre de este tipo de distribuciones puede generar confusión en una primera lectura. La familia de las distribuciones exponenciales (uniparamétrica, con $k = 1$) es una familia de tipo exponencial, pero no es la única de este tipo.

Ejemplo 35. Es sencillo obtener en los siguientes casos que la familia paramétrica que se indica es de tipo exponencial uniparamétrico, y determinar las funciones $c(\theta)$, $h(x)$, $q_1(\theta)$ y $V_1(x)$.

$Exp(\theta)$: con $c(\theta) = \theta$, $h(x) = 1$, $q_1(\theta) = -\theta$ y $V_1(x) = x$.

$B(1, \theta)$: con $c(\theta) = 1 - \theta$, $h(x) = 1$, $q_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ y $V_1(x) = x$.

$\mathcal{P}(\theta)$: con $c(\theta) = e^{-\theta}$, $h(x) = \frac{1}{x!}$, $q_1(\theta) = \log \theta$ y $V_1(x) = x$.

$N(\mu, \sigma_0)$ ($\theta = \mu$, y $\sigma = \sigma_0$ conocido): con $c(\mu) = \frac{1}{\sqrt{2\pi\sigma_0}} \exp\left\{-\frac{1}{2\sigma_0^2}\mu^2\right\}$, $h(x) = \exp\left\{-\frac{1}{2\sigma_0^2}x^2\right\}$,

$q_1(\mu) = \frac{1}{\sigma_0^2}\mu$ y $V_1(x) = x$ (desarrollando el cuadrado y simplificando adecuadamente).

$N(\mu_0, \sigma)$ ($\theta = \sigma$, y $\mu = \mu_0$ conocido): con $c(\sigma) = \frac{1}{\sqrt{2\pi\sigma}}$, $h(x) = 1$, $q(\sigma) = -\frac{1}{2\sigma^2}$ y $V(x) = (x - \mu_0)^2$.

Ejemplo 36. Es sencillo comprobar que la familia $N(\mu, \sigma)$, con $\theta = (\mu, \sigma)$, es de tipo exponencial biparamétrico, con: $c(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}$, $h(x) = 1$, $q_1(\mu, \sigma) = \frac{\mu}{\sigma^2}$, $V_1(x) = x$, $q_2(\mu, \sigma) = -\frac{1}{2\sigma^2}$ y $V_2(x) = x^2$. También lo son la familia de las distribuciones gamma y la de las beta.

Teorema 37. Consideremos una m.a.s. de un familia de distribuciones de tipo exponencial k -paramétrico. Se tiene que el estadístico

$$T = (T_1, \dots, T_k) = \left(\sum_{i=1}^n V_1(X_i), \dots, \sum_{i=1}^n V_k(X_i) \right)$$

es suficiente y completo.

Por el teorema 32 se tiene que este estadístico es además minimal suficiente. Esto permite obtener, para familias de tipo exponencial, estadísticos minimales suficientes de un modo alternativo a la aplicación directa del teorema 20, que además son completos.

Demostración. Demostramos solo que T es suficiente, y además para $k = 1$. Y en vez de utilizar el teorema de factorización para ello, utilizamos el teorema 20 para obtener directamente el ems. Se tiene que

$$f_{\theta}(x_1, \dots, x_n) = \prod_{i=1}^n f_{\theta}(x_i) = c(\theta)^n \left(\prod_{i=1}^n h(x_i) \right) \exp \left\{ q(\theta) \sum_{i=1}^n V(x_i) \right\},$$

y de aquí se obtiene

$$\frac{f_{\theta}(x_1, \dots, x_n)}{f_{\theta}(x'_1, \dots, x'_n)} = \frac{\prod_{i=1}^n h(x_i)}{\prod_{i=1}^n h(x'_i)} \exp \left\{ q(\theta) \left(\sum_{i=1}^n V(x_i) - \sum_{i=1}^n V(x'_i) \right) \right\},$$

que no depende de θ si y solo si $\sum_{i=1}^n V(x_i) = \sum_{i=1}^n V(x'_i)$. Entonces, el estadístico ms es

$$T = \sum_{i=1}^n V(X_i).$$

■

En el ejemplo 35 se obtiene $V(x) = x$ para las distribuciones $Exp(\theta)$, $B(1, \theta)$, $\mathcal{P}(\theta)$ y $N(\mu, \sigma_0)$, con $\theta = \mu$, y $\sigma = \sigma_0$ conocida. Entonces, en estos cuatro casos el estadístico ms es

$$T = \sum_{i=1}^n V(X_i) = \sum_{i=1}^n X_i, \quad (2.3)$$

como ya sabíamos, y además es completo. Para la distribución $N(\mu_0, \sigma)$, $\theta = \sigma$ y $\mu = \mu_0$ conocido, se obtiene $V(x) = (x - \mu_0)^2$, y entonces

$$T = \sum_{i=1}^n (X_i - \mu_0)^2$$

es ms y completo. Para la distribución $N(\mu, \sigma)$, con $\theta = (\mu, \sigma)$, biparamétrica, se ha obtenido $V_1(x) = x$ y $V_2(x) = x^2$. Entonces, $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ es ms, como ya sabíamos, y además es completo.

Ejemplo 38. Para la familia paramétrica $N(\theta, 1)$ se tiene que \bar{X} es suficiente y completo (según acabamos de ver $\sum_{i=1}^n X_i$ es suficiente y completo, y entonces \bar{X} también lo es). Comprobaremos que S^2 es un estadístico auxiliar, y por tanto es independiente de \bar{X} por el teorema de Basu (resultado que ya conocíamos por el teorema de Fisher). Se tiene que $(n-1)S^2 \sim \chi_{n-1}^2$ (por el teorema de Fisher), distribución que no depende de θ , y entonces la distribución de S^2 no depende tampoco de θ , y por tanto S^2 es un estadístico auxiliar.

2.4 El principio de verosimilitud

Dada una realización muestral (x_1, \dots, x_n) , denominamos *función de verosimilitud* a

$$L(\theta) = L(\theta; x_1, \dots, x_n) = f_{\theta}(x_1, \dots, x_n),$$

que es la función de masa o densidad $f_{\theta}(x_1, \dots, x_n)$, pero con (x_1, \dots, x_n) fijo y considerada como función de θ . El argumento de $L(\theta; x_1, \dots, x_n)$ es θ , pero a veces se añade la muestra x_1, \dots, x_n en la notación para incidir en que $L(\theta)$ se define a partir de x_1, \dots, x_n , que denotamos vectorialmente como \mathbf{x} : $\mathbf{x} = (x_1, \dots, x_n)$.

Hay una función $L(\theta)$ para cada muestra \mathbf{x} . Para cada θ , la verosimilitud $L(\theta; \mathbf{x})$ de θ para la muestra \mathbf{x} es la masa o densidad de probabilidad que tiene la muestra \mathbf{x} para ese valor de θ . La elección del valor de θ mas verosimil da lugar a un procedimiento de estimación que se estudia en el próximo tema.

El principio de verosimilitud se basa en la función de verosimilitud. Se aplica a resultados de experimentos aleatorios, como pueden ser realizaciones de una m.a.s. Utilizaremos la expresión “evidencia que aporta sobre θ una realización muestral \mathbf{x} ” en el sentido de información, discernimiento sobre el parámetro cuando se utiliza \mathbf{x} para hacer inferencias.

Se dice que se sigue el *principio de verosimilitud* cuando se considera que, dos realizaciones \mathbf{x} , y aportan la misma evidencia, si existe una constante $c(\mathbf{x}, \mathbf{y})$ que no depende de θ con

$$L(\theta; \mathbf{x}) = c(\mathbf{x}, \mathbf{y}) L(\theta; \mathbf{y}).$$

Pero el principio se aplica a casos mas generales, en que las dos muestras pueden ser realizaciones de experimentos aleatorios diferentes, y en vez de muestras se consideran estadísticos (una muestra de tamaño 1 para el estadístico, con su distribución):

Ejemplo 39. Sea θ la probabilidad de éxito en pruebas de Bernoulli independientes. Consideramos dos resultados de dos experimentos aleatorios alternativos para estudiar θ .

Por una parte, si realizamos n pruebas de Bernoulli y consideramos el estadístico

número de éxitos, su distribución es $B(n, \theta)$. Si se obtienen t éxitos, la función de verosimilitud para esta observación es

$$L(\theta; t) = \binom{n}{t} \theta^t (1 - \theta)^{n-t}.$$

Por otra parte, si realizamos repeticiones de las pruebas de Bernoulli hasta obtener t éxitos, y consideramos el estadístico número de repeticiones realizadas, su distribución es binomial negativa con parámetro (t, θ) . Si se requieren n repeticiones, la función de verosimilitud para esta observación es

$$L(\theta; n) = \binom{n-1}{t-1} \theta^t (1 - \theta)^{n-t}.$$

Se tiene que $L(\theta; t) = c(t, n) L(\theta; n)$ con $c(t, n) = n/t$, que no depende de θ . Si se sigue el principio de verosimilitud, hay que considerar que estas dos realizaciones aportan la misma evidencia, aunque correspondan a experimentos aleatorios diferentes.

Si x e y son dos muestras del mismo experimento aleatorio, si se sigue el principio de verosimilitud, por el teorema 20, se tiene que x , y aportan la misma evidencia cuando están en la misma clase de la partición minimal suficiente. Esto es, cuando cualquier estadístico minimal suficiente T cumple que $T(x) = T(y)$.

El principio de verosimilitud está relacionado con otros principios.

El principio de suficiencia acepta que dos puntos muestrales x , y de un mismo experimento aportan la misma evidencia cuando están en la misma clase de la partición minimal suficiente. Pero téngase en cuenta que el principio de verosimilitud no está limitado al mismo experimento.

El principio de condicionalidad acepta que ningún mecanismo aleatorio que no dependa de θ puede aportar evidencia. Se puede formular en términos de un experimento compuesto, en el que un mecanismo aleatorio que no depende de θ selecciona uno entre varios posibles experimentos, del que se obtiene una realización x . Según el principio de condicionalidad, la única evidencia la aporta x , y no el mecanismo aleatorio de selección.

Por el teorema de Birnbaum, se tiene que el principio de verosimilitud es equivalente a los principios de suficiencia y condicionalidad considerados conjuntamente.

2.5 El principio de equivarianza

Las cuestiones que se desarrollan aquí se aplican principalmente a estadísticos que sean estimadores, que se definen en el próximo tema. Es suficiente saber de momento que los estimadores son estadísticos con los que se pretenden obtener realizaciones cercanas al verdadero valor de θ , estimaciones de θ , y que sirven por tanto como aproximaciones para θ .

Supongamos que se realizan mediciones repetidas x_1, \dots, x_n sobre un mismo material, en micras, con un aparato para medir longitudes. Se supone que los errores de medida

son normales con esperanza 0, de modo que la distribución teórica es $N(\mu, \sigma)$, siendo μ el verdadero valor de la longitud del material. Se considera un estimador T de μ , y se obtiene la estimación $T = t$.

Si se descubre que sistemáticamente las mediciones del aparato son erróneas, con un exceso de tres micras, habrá que rehacer los cálculos para t . Se podrían corregir los datos x_i considerando $x'_i = x_i - 3$, y calcular el nuevo valor t' de t a partir de ellos. O se podría corregir el valor de t a $t - 3$. Para que el estadístico T produzca resultados coherentes, parece razonable imponerle que el resultado sea el mismo por uno y otro procedimiento, esto es,

$$t' = t - 3.$$

Se dice en este caso que el estimador T es equivariante (o invariante) para traslaciones (o cambios de posición). Es sencillo comprobar que la media muestral, la mediana muestral, y $\frac{X_{(1)} + X_{(n)}}{2}$ cumplen esta propiedad. En cambio, la media geométrica y la media armónica no la cumplen.

También, si hacemos los cálculos para una estimación t de una medida de posición (la esperanza μ por ejemplo), a partir de los datos muestrales x_i de una medida expresada en metros, y volvemos a realizar los cálculos con la medida expresada en centímetros, con $x'_i = 100x_i$, la nueva estimación t' debería cumplir $t' = 100t$. Se dice en este caso que el estimador T del que hemos obtenido la realización t' es equivariante para cambios de escala. Es sencillo comprobar que la media muestral, la mediana muestral, $\frac{X_{(1)} + X_{(n)}}{2}$, y la media armónica cumplen esta propiedad. En cambio la media geométrica no la cumple.

En ambos ejemplos hemos considerado transformaciones $x'_i = g(x_i)$ sobre las x_i , siendo g la traslación $g(x) = x - a$ con $a = 3$ en un caso, y el cambio de escala $g(x) = bx$ con $b = 100$ en el otro.

La noción general de equivarianza para un estimador en una familia paramétrica se formula en término de un grupo de transformaciones (funciones), siendo la composición de funciones la operación interna considerada. Para cualquiera de estos grupos, el elemento neutro es la función g dada por $g(x) = x$. Cualquier función g del grupo será invertible (biyectiva), y su función inversa g^{-1} coincidirá con el elemento inverso en el grupo. Por ejemplo, para el grupo de las traslaciones, el elemento inverso de $g_1(x) = x + a$ es $g_2(x) = x - a$ ($g_2 = g_1^{-1}$), con $g_2(g_1(x)) = x$, de modo que g_2g_1 es el elemento neutro, y también $g_1g_2(x) = x$. Para el grupo de los cambios de escala, el elemento inverso de $g_1(x) = bx$ es $g_2(x) = x/b$, con $g_2(g_1(x)) = x$.

En primer lugar, hay que asegurarse de que la transformación que suponen las correcciones $x'_i = g(x_i)$ sobre las x_i de los ejemplos dejen la distribución de las X'_i en la misma familia paramétrica. Definiremos el concepto de familia de distribuciones invariante.

Se puede considerar que la muestra $\mathbf{X} = (X_1, \dots, X_n)$ es m.a.s. o que tiene otra distribución, pero nos limitamos al caso en que \mathbf{X} es una m.a.s. De este modo, la distribución de \mathbf{X} se obtiene a partir de la distribución F_θ de las X_i , independientes, que es la distribución de la muestra que venimos considerando. En los ejemplos hemos considerado

una misma función g sobre cada componente x_i de la muestra, y a partir de ello queda definida una función $g : \Omega \rightarrow \Omega$ que aplica la misma operación a cada componente. Es habitual considerar este tipo de funciones g , pero lo que se explicará se puede aplicar también a funciones de otro tipo, en un grupo de transformaciones.

Consideraremos transformaciones $g : \Omega \rightarrow \Omega$ en general, y el anterior será un caso particular. Además, por sencillez en la notación, escribimos g sin negrita, g . Y la transformación sobre T que se considerará para obtener T' estará relacionada con g , pero no necesariamente será la misma función g que se aplica sobre cada componente, como ocurre en los ejemplos.

Consideramos un familia paramétrica $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Podríamos especificar las distribuciones de la familia mediante F_θ , como venimos haciendo, pero lo expresamos mediante P_θ porque quizá resulta mas claro cuando hablemos de transformaciones de P_θ . Además, cometemos abuso de notación denotando también a la distribución de la muestra por P_θ (como hacíamos con F y f). De este modo, en la expresión

$$\mathbf{X} \sim P_\theta$$

se sobreentiende que se debe sustituir P_θ por la medida producto usual para la muestra.

Definición 40. Consideremos un grupo \mathcal{G} de transformaciones g sobre Ω , $g : \Omega \rightarrow \Omega$, y una familia de distribuciones $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Se dice que la familia \mathcal{P} es invariante por el grupo \mathcal{G} si para todo $g \in \mathcal{G}$ y $\theta \in \Theta$ se tiene que existe un único $\theta' \in \Theta$ tal que

$$g(\mathbf{X}) \sim P_{\theta'} .$$

Para una familia invariante, sea

$$\bar{g} : \Theta \rightarrow \Theta$$

la función que aplica a cada θ ese valor único θ' , con $\bar{g}(\theta) = \theta'$. Entonces, la distribución de $g(\mathbf{X})$ es:

$$g(\mathbf{X}) \sim P_{\bar{g}(\theta)} .$$

Bajo una transformación de \mathbf{X} por una aplicación g del grupo, la distribución de $g(\mathbf{X})$ sigue en la misma familia paramétrica, pero con un nuevo valor del parámetro $\theta' = \bar{g}(\theta)$.

Ejemplo 41. Consideremos una m.a.s. $\mathbf{X} = (X_1, \dots, X_n)$ de una población $N(\mu, \sigma)$, con parámetro $\theta = (\mu, \sigma)$, con $\mu \in \mathbb{R}$, $\sigma > 0$. Consideramos el grupo de transformaciones dado por

$$g(\mathbf{X}) = (a + bX_1, \dots, a + bX_n) , \text{ con } -\infty < a < \infty \text{ y } b > 0$$

(cambios de posición y escala o transformaciones afines).

Por ejemplo, supongamos que tenemos unos datos de temperatura en grados Celsius que sabemos que siguen una distribución normal, y pasamos a grados Fahrenheit.

Vamos a comprobar que la distribución normal (la familia) es invariante para este grupo, y a determinar $\bar{g}(\theta) = \bar{g}(\mu, \sigma)$.

Puesto que \mathbf{X} tiene distribución normal n -dimensional, y g es una transformación lineal de \mathbb{R}^n en \mathbb{R}^n no singular (no se aplica en un subespacio con dimensión inferior a n), se tiene por las propiedades de esta distribución que $g(\mathbf{X})$ es normal también. Las componentes de $g(\mathbf{X})$ son *vaiid*, puesto que son función de *vaiid*, y con la misma distribución, con esperanza y varianza:

$$\begin{aligned} E[a + bX_i] &= a + b\mu \in \mathbb{R}, \\ V[a + bX_i] &= b^2V[X_i] = b^2\sigma^2 > 0. \end{aligned}$$

La desviación típica de $a + bX_i$ es $b\sigma$. Se tiene entonces que esta familia es invariante para este grupo, con

$$\bar{g}(\mu, \sigma) = (a + b\mu, b\sigma).$$

Definición 42. Consideremos un grupo \mathcal{G} que deja invariante una familia paramétrica $\{P_\theta : \theta \in \Theta\}$. Un estimador $T(\mathbf{X})$ de θ se dice que es equivariante respecto de \mathcal{G} si

$$T(g(\mathbf{X})) = \bar{g}(T(\mathbf{X})) \quad \text{para todo } g \in \mathcal{G}.$$

También se utiliza el término invariante: estimador invariante.

Ejemplo 43. Consideramos una m.a.s. $N(\mu, \sigma)$, con $\theta = (\mu, \sigma)$, y sea $S = \sqrt{S^2}$, la raíz cuadrada de la cuasivarianza muestral. Comprobamos que el estimador $T(\mathbf{X}) = (\bar{X}, S)$ es equivariante para el grupo de cambios de posición y escala. En el ejemplo anterior se comprobó que esta familia es invariante para este grupo, con $\bar{g}(\mu, \sigma) = (a + b\mu, b\sigma)$. Comprobamos que $T(g(\mathbf{X})) = \bar{g}(T(\mathbf{X}))$:

$$\begin{aligned} T(g(\mathbf{X})) &= T(a + bX_1, \dots, a + bX_n) = (a + b\bar{X}, bS) \quad \text{y} \\ \bar{g}(T(\mathbf{X})) &= \bar{g}(\bar{X}, S) = (a + b\bar{X}, bS). \end{aligned}$$

Si se considera $\theta = (\mu, \sigma^2)$, se obtiene que $T(\mathbf{X}) = (\bar{X}, S^2)$ es equivariante para el grupo de cambios de posición y escala, y también lo es (\bar{X}, s^2) .

El principio de equivarianza aconseja que se consideren solo estimadores equivariantes.

Tema 3

Estimación puntual I. Introducción. Métodos para encontrar estimadores: método de los momentos, estimadores máximo-verosímiles, estimadores de Bayes, estimadores invariantes.

Los contenidos de este tema corresponden a un curso básico de Estadística Matemática, y pueden ser consultados y completados con los libros de la bibliografía por ejemplo. El libro de Rohatgi y Ehsanes es el más completo.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

3.1 Introducción

Suponemos que la distribución teórica sobre la que queremos obtener información pertenece a una familia paramétrica $\{F_\theta : \theta \in \Theta\}$. Consideramos una m.a.s. X_1, \dots, X_n , para la que observamos los datos x_1, \dots, x_n (realización muestral). Para obtener información sobre θ consideramos estadísticos, funciones de la muestra.

En este tema y el próximo estudiaremos un tipo de estadísticos cuyas realizaciones pretendemos que sean cercanas al valor del parámetro. Denominaremos estimadores a estos estadísticos, y estimaciones a sus realizaciones. De este modo, una estimación debería proporcionar un valor aproximado para el parámetro desconocido.

En este tema estudiamos métodos de obtención de estimadores. Son métodos generales que se aplican a distintas familias paramétricas. Estudiaremos el método de los momentos, basado en estimar parámetros expresándolos en términos de momentos poblacionales, y estimando estos mediante los correspondientes momentos muestrales. El método de máxima verosimilitud consiste en utilizar como estimación el valor del parámetro que maximiza la función de verosimilitud. Los estimadores de Bayes están basados en un enfoque para la inferencia estadística en el que el parámetro no es fijo, sino que hay una distribución de probabilidad sobre el espacio paramétrico, que actualizamos con la muestra obtenida. En el próximo tema estudiaremos métodos para evaluar el comportamiento de los estimadores, en relación al objetivo de obtener con las estimaciones buenas aproximaciones para el valor del parámetro.

Recordamos que el espacio muestral Ω es el conjunto de posibles muestras (x_1, \dots, x_n) , que coincide con probabilidad 1 con el soporte de la variable aleatoria (X_1, \dots, X_n) .

3.2 Estimadores puntuales

Definición 44. Un estimador puntual $T = T(X_1, \dots, X_n)$ de θ es un estadístico cuyo recorrido está contenido en el espacio paramétrico:

$$T : \Omega \rightarrow \Theta .$$

Al valor $t = T(x_1, \dots, x_n)$ del estimador obtenido con una muestra concreta (x_1, \dots, x_n) se le denomina *estimación puntual* de θ . Es una realización de la variable aleatoria T . En este contexto de estimación puntual se suele omitir el término “puntual”, y se habla de estimador y estimación.

Por ejemplo, si consideramos el estimador $T = \bar{X}$ y obtenemos la muestra, de tamaño $n = 3$, $(x_1, x_2, x_3) = (5.2, 2.8, 4.3)$, entonces obtenemos la estimación $T = \bar{x} = 4.1$. El hecho de que el recorrido de T esté contenido en Θ hace que tenga sentido utilizar una estimación como valor aproximado para θ . Por ejemplo, si $\Theta = [0, 1]$, una estimación para θ de $t = 1.2$ no corresponde a posible un valor del parámetro, y entonces la mencionada estimación realmente no lo es, porque no corresponde a una realización de un estimador.

El problema de la estimación puntual consiste en seleccionar un estimador adecuado, de modo que cuando obtengamos la realización muestral, el valor de la estimación sea “cercano” al de θ . En este tema se presentarán procedimientos para obtener estimadores de θ , y en el próximo tema se estudiarán criterios para valorar la conveniencia de uno u otro estimador.

Ejemplo 45. Consideremos una muestra X_1, X_2 de una distribución teórica $B(1, \theta)$, con $\Theta = [0, 1]$ (cerrado aquí). Es una muestra de un tamaño demasiado pequeño como para que pueda aportar mucha información sobre θ , pero sirve para manejar algunas nociones básicas.

Los estadísticos $T_1 = \bar{X}$ y $T_2 = \frac{1}{3}(X_1 + 2X_2)$ son estimadores de θ , puesto que toman valores en Θ . Tengase en cuenta que T_1 y T_2 son variables aleatorias y toman distintos valores según la muestra obtenida, según se presenta en la siguiente tabla:

$(x_1, x_2) =$	$(0, 0)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
Probabilidad	$(1 - \theta)^2$	$\theta(1 - \theta)$	$\theta(1 - \theta)$	θ^2
T_1	0	1/2	1/2	1
T_2	0	2/3	1/3	1

Mediante un cálculo directo a partir de la distribución de T_1 y T_2 , dadas en la tabla, o utilizando las propiedades de la esperanza y la varianza y conociendo los momentos de la distribución $B(1, \theta)$ (lo hacemos así), se obtiene

$$E[T_1] = E[\bar{X}] = \mu = \theta \text{ y}$$

$$E[T_2] = \frac{1}{3}(E[X_1] + 2E[X_2]) = \theta$$

La esperanza de ambos estimadores coincide con el valor θ del parámetro, que queremos estimar, sea cual sea este. Un criterio que se considerará en el próximo tema para

comparar estimadores de este tipo está basado en el valor de la varianza del estimador. La varianza es la esperanza (un promedio) de las desviaciones al cuadrado de la distribución del estimador respecto de su esperanza. Se tiene que

$$V [T_1] = V [\bar{X}] = \frac{\sigma^2}{2} \text{ y}$$

$$V [T_2] = \left(\frac{1}{3}\right)^2 (V [X_1] + 2^2 V [X_2]) = \frac{5}{9} \sigma^2 ,$$

con $\sigma^2 = \theta(1 - \theta)$, teniendo en cuenta que X_1 y X_2 son independientes, y entonces su covarianza es nula. Se obtiene que $V [T_1] < V [T_2]$ para todo θ . Esta circunstancia podría aconsejar utilizar T_1 mejor que T_2 , aunque esto no quiere decir que siempre obtengamos mejores estimaciones con T_1 que con T_2 . Por ejemplo, si fuera $\theta = 2/3$ y obtuviéramos $(x_1, x_2) = (0, 1)$, entonces la estimación $t_1 = 1/2$ es obviamente peor que la estimación $t_2 = 2/3 = \theta$. Sin embargo, cualquiera que sea θ (también con $\theta = 2/3$), en promedio nos alejaremos más de θ (medido con la varianza) con T_2 que con T_1 .

Es frecuente denotar a un estimador de θ (y también a la estimación) por $\hat{\theta}$. Por otra parte, algunas veces el objetivo es estimar no ya directamente θ sino una función $g(\theta)$. Por ejemplo, para una distribución $Exp(\theta)$ se tiene que la esperanza es $\mu = 1/\theta$, y podríamos estar interesados en estimar $\mu = g(\theta) = 1/\theta$ en vez de θ . El estimador se puede denotar por $\hat{\mu} = \widehat{g(\theta)}$. Si queremos estimar θ , entonces $g(\theta) = \theta$.

Ejemplo 46. La media muestral \bar{X} es un estimador de la esperanza μ , lo que podemos denotar como $\hat{\mu} = \bar{X}$. Esto es válido en todos los casos habituales, aunque se pueden construir casos con espacio paramétrico no conexo para los que el recorrido de \bar{X} (conjunto de valores que puede tomar) no está contenido en Θ .

Consideremos una muestra de una $B(1, \theta)$. La proporción de éxitos es \bar{X} , que es un estimador de $\mu = \theta$, lo que podemos denotar como $\hat{\theta} = \bar{X}$.

Consideremos una muestra de una $\gamma(p, a)$. En este caso $\theta = (p, a)$ y $\mu = g(p, a)$ con $g(p, a) = p/a$, y entonces $\hat{\mu} = \bar{X}$ es un estimador de p/a .

3.3 Método de los momentos para la obtención de estimadores

Los momentos poblacionales son funciones del parámetro $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$. El *método de los momentos* (mm) de estimación puntual consiste en despejar $\theta_1, \dots, \theta_k$ de estas funciones, para expresarlos en términos de los momentos, y en estimar estos momentos poblacionales mediante los correspondientes momentos muestrales.

Se plantean k ecuaciones con los k primeros momentos. Si el sistema no tiene solución única se añaden momentos posteriores a los k primeros hasta que la solución sea única. Se suelen utilizar momentos respecto del origen. Utilizando momentos centrales el resultado suele ser el mismo que utilizando momentos respecto del origen, pero no siempre es así. Además, el momento de orden 1 debe ser tomado respecto del origen, puesto que el central es nulo.

Este procedimiento no está asociado a propiedades de optimalidad, y puede dar lugar a buenos o malos estimadores (según los criterios que se presentarán en el tema 4).

Ejemplo 47. Determina el estimador de θ por el mm para las siguientes distribuciones teóricas:

a) $B(1, \theta)$, b) $Exp(\theta)$, c) $U(0, \theta)$, d) $U(-\theta, \theta)$, e) $\gamma(p, a)$, con $\theta = (p, a)$.

a) Se estima $E[X] = \mu = \theta$ mediante la media muestral, y entonces $\hat{\theta} = \bar{X}$.

b) Se despeja θ de la ecuación $E[X] = \mu = 1/\theta$, obteniéndose $\theta = 1/\mu$, y se estima μ mediante la media muestral, obteniéndose el estimador $\hat{\theta} = 1/\bar{X}$.

c) Se despeja θ de la ecuación $E[X] = \mu = \theta/2$, obteniéndose $\theta = 2\mu$, y se estima μ mediante la media muestral, obteniéndose el estimador $\hat{\theta} = 2\bar{X}$. Obsérvese que $\hat{\theta}$ no es función del estadístico sm $X_{(n)}$. No es un buen estimador como comprobaremos en el próximo tema. Además, el estimador $\hat{\theta} = 2\bar{X}$ puede dar lugar a malas estimaciones. Por ejemplo, si obtenemos la muestra $(x_1, x_2, x_3) = (1.3, 8.9, 2.1)$, con $n = 3$, se tiene que $\bar{x} = 4.1$, y la estimación vale $\hat{\theta} = 2 \cdot 4.1 = 8.2$. Puesto que $X_2 \sim U(0, \theta)$ y $x_2 = 8.9$, tiene que ser $\theta \geq 8.9$, y entonces la estimación $\hat{\theta} = 8.2$ es claramente inadecuada.

d) Se tiene que $E[X] = \mu = \frac{\theta + (-\theta)}{2} = 0$, que no proporciona ninguna ecuación para θ , y entonces debemos considerar el momento de orden 2. En algunas ocasiones, como en este ejemplo, los estimadores son diferentes si se consideran momentos respecto del origen o centrales. El momento poblacional central de orden 2 es la varianza

$$\beta_2 = V[X] = \sigma^2 = \frac{(\theta - (-\theta))^2}{12} = \frac{\theta^2}{3},$$

y se estima mediante el momento muestral central de orden 2 (la varianza muestral $s^2 = b_2$); despejando θ se obtiene $\theta = \sigma\sqrt{3}$, y el estimador es $\hat{\theta} = s\sqrt{3}$, siendo $s = \sqrt{s^2}$. El momento poblacional de orden 2 respecto del origen es

$$\alpha_2 = E[X^2] = V[X] + E[X]^2 = \frac{\theta^2}{3} + 0^2 = \frac{\theta^2}{3},$$

y se estima mediante el momento muestral de orden 2 respecto del origen $a_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$; despejando θ se obtiene $\theta = \sqrt{3\alpha_2}$, y el estimador es $\tilde{\theta} = \sqrt{3a_2} = \sqrt{\frac{3}{n} \sum_{i=1}^n X_i^2}$.

e) Puesto que hay dos parámetros desconocidos, planteamos dos ecuaciones. Además de la esperanza $E[X] = \mu = p/a$ (momento de orden 1), consideramos la varianza $V[X] = \sigma^2 = p/a^2$ (momento central de orden 2). Despejando se obtiene $p = \frac{\mu^2}{\sigma^2}$ y $a = \frac{\mu}{\sigma^2}$, y entonces el estimador mm de $\theta = (p, a)$ es

$$\hat{\theta} = (\hat{p}, \hat{a}) = \left(\frac{\bar{X}^2}{s^2}, \frac{\bar{X}}{s^2} \right).$$

Este estimador no es función del estadístico minimal suficiente, que es $(\prod_{i=1}^n X_i, \sum_{i=1}^n X_i)$, o $(\sum \log X_i, \bar{X})$.

Ejemplo 48. Para una distribución teórica $N(\mu, \sigma)$, el parámetro μ coincide con el momento de orden 1, que entonces no hay que despejar, obteniéndose el estimador mm $\hat{\mu} = \bar{X}$. Para estimar σ , si consideramos el momento de orden 2 respecto del origen se obtiene

$$\alpha_2 = E[X^2] = V[X] + E[X]^2 = \sigma^2 + \mu^2 = \sigma^2 + \alpha_1^2,$$

y de aquí $\sigma = \sqrt{\alpha_2 - \alpha_1^2}$, obteniéndose el estimador mm

$$\hat{\sigma} = \sqrt{a_2 - a_1^2} = s ,$$

con $s = \sqrt{s^2}$. Entonces, el estimador mm de (μ, σ) es (\bar{X}, s) . Si consideramos el momento central de orden 2 para estimar σ se obtiene el mismo resultado. El estimador mm para (μ, σ^2) es (\bar{X}, s^2) .

3.3.1 Método de máxima verosimilitud

Dada una realización muestral (x_1, \dots, x_n) , denominamos *función de verosimilitud* a

$$L(\theta) = L(\theta; x_1, \dots, x_n) = f_\theta(x_1, \dots, x_n) \quad \text{para } \theta \in \Theta,$$

que es la función de masa o densidad $f_\theta(x_1, \dots, x_n)$, pero con (x_1, \dots, x_n) fijo y considerada como función de θ .

Definición 49. La estimación de máxima verosimilitud de θ , $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$, asociada a la muestra (x_1, \dots, x_n) , es el valor de θ que maximiza $L(\theta)$:

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta) . \quad (3.1)$$

La correspondiente v.a. $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ es el *estimador de máxima verosimilitud (emv)* de θ .

Ejemplo 50. Consideremos dos urnas: $U_1(2b, 8n)$ y $U_2(9b, 1n)$. Alguién elige una urna y extrae una bola al azar de esa urna. No nos informa de cual ha sido la urna elegida, pero nos dice que la bola extraída ha resultado ser blanca. ¿Cual ha sido la urna de la que se ha extraído la bola blanca?

Podría ser cualquiera de las dos, puesto que en cada una de ellas hay alguna bola blanca. Pero si tuvieramos que decantarnos, parece más plausible que la urna haya sido U_2 , puesto que la probabilidad de que al extraer una bola ésta sea blanca, es mayor para U_2 que para U_1 . Con este modo de decidir se está realizando la estimación de máxima verosimilitud dada por (3.1), como comprobamos a continuación.

Consideremos la v.a. $X = X_1$ (muestra de tamaño 1) que vale 1 si se obtiene bola blanca y vale 0 si se obtiene negra (Bernoulli). Sea $\theta_1 = 0.2$ y $\theta_2 = 0.9$, que son las probabilidades de obtener bola blanca en la primera y segunda urna respectivamente. Con estos elementos, podemos formular la situación del enunciado del siguiente modo: tenemos una muestra de tamaño 1 de una distribución teórica de Bernoulli con espacio paramétrico $\Theta = \{\theta_1, \theta_2\}$. Obtenemos la estimación de máxima verosimilitud (mv) asociada al dato $x = 1$ (bola blanca).

La función de verosimilitud es

$$L(\theta) = L(\theta; x = 1) = f_\theta(1) = \theta^1(1 - \theta)^{1-1} = \theta .$$

Entonces, $L(\theta_1) = \theta_1 = 0.2$ y $L(\theta_2) = \theta_2 = 0.9$, y por tanto

$$\max_{\theta \in \Theta} L(\theta) = \max_{\theta \in \{\theta_1, \theta_2\}} L(\theta) = \max\{0.2, 0.9\} = 0.9 = L(\theta_2) .$$

Entonces, $L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta) = L(\theta_2)$, y por tanto $\hat{\theta} = \theta_2$ es la estimación mv de θ , correspondiente a la urna U_2 .

Si $L(\theta)$ es continua y derivable, entonces podemos encontrar el punto de máximo $\hat{\theta}$ mediante el método habitual: determinando los máximos locales entre los puntos críticos y buscando entre ellos (la frontera del soporte incluida) el máximo global $\hat{\theta}$. En los casos más habituales con $L(\theta)$ derivable, $L(\theta)$ tiene un único máximo local que es el máximo global.

Además, disponemos del siguiente resultado que simplifica los cálculos: el valor de θ para el cual una función $L(\theta)$ se maximiza, es el mismo para el cual la función $\log L(\theta)$ se maximiza. La razón de este resultado es que el logaritmo es una función creciente y, por tanto, cuanto mayor sea $L(\theta)$ mayor es $\log L(\theta)$. También se puede comprobar de un modo simple (para L derivable) observando que para $L(\theta) > 0$ se cumple $\frac{d}{d\theta} \log L(\theta) = 0$ si y solo si $L'(\theta) = 0$, puesto que $\frac{d}{d\theta} \log L(\theta) = \frac{L'(\theta)}{L(\theta)}$. Si $L(\theta)$ es un producto de funciones exponenciales, es más rápido derivar $\log L(\theta)$ (que es suma de productos) que $L(\theta)$. Ésto ocurre por ejemplo con las familias de tipo exponencial.

Ejemplo 51. Determina el emv de θ para las siguientes distribuciones teóricas:

a) $B(1, \theta)$, b) $\mathcal{P}(\theta)$, c) $Exp(\theta)$, d) $U(0, \theta)$.

Omitimos la comprobación de que el punto crítico obtenido es de máximo absoluto, que queda como ejercicio.

a) Se tiene que $L(\theta) = f_{\theta}(x_1, \dots, x_n) = \theta^{n\bar{x}}(1 - \theta)^{n(1 - \bar{x})}$, y de aquí:

$$\begin{aligned} \log L(\theta) &= n\bar{x} \log \theta + n(1 - \bar{x}) \log(1 - \theta), \text{ y} \\ \frac{d}{d\theta} \log L(\theta) &= \frac{n\bar{x}}{\theta} - \frac{n(1 - \bar{x})}{1 - \theta}. \end{aligned}$$

Resolviendo la ecuación $\frac{d}{d\theta} \log L(\theta) = 0$ se obtiene la solución única $\theta = \bar{x}$. La solución de la ecuación es el estimador de θ que buscamos, y entonces el emv es $\hat{\theta} = \bar{X}$.

b) Se tiene que $\log L(\theta) = -n\theta + n\bar{x} \log \theta - \log(x_1! \cdots x_n!)$, y

$$\frac{d}{d\theta} \log L(\theta) = -n + \frac{n\bar{x}}{\theta} = 0,$$

que tiene solución $\theta = \bar{x}$. Entonces, $\hat{\theta} = \bar{X}$ es el emv de θ .

c) $\log L(\theta) = n \log \theta - n\theta\bar{x}$, y $\frac{d}{d\theta} \log L(\theta) = \frac{n}{\theta} - n\bar{x} = 0$, que tiene solución $\theta = 1/\bar{x}$. Entonces, $\hat{\theta} = 1/\bar{X}$ es el emv de θ .

d) En este caso la solución no corresponde a un valor θ con $L'(\theta) = 0$. Por simplicidad, conviene modificar la función de densidad de la muestra de la $U(0, \theta)$, $f_{\theta}(x_1, \dots, x_n) = \theta^{-n} I\{x_{(n)} < \theta\}$, utilizando una desigualdad no estricta en la función $I: I\{x_{(n)} \leq \theta\}$. Haciendo ésto, modificamos el soporte en un conjunto de probabilidad 0, y no afecta a la distribución. Con ello, la función de verosimilitud es $L(\theta) = \theta^{-n} I\{x_{(n)} \leq \theta\}$, y se tiene que

$$L(\theta) = \theta^{-n} I\{x_{(n)} \leq \theta\} = \begin{cases} 0 & \text{si } 0 < \theta < x_{(n)} \\ \theta^{-n} & \text{si } \theta \geq x_{(n)} \end{cases}.$$

Se tiene que $\frac{d}{d\theta} \theta^{-n} = -n\theta^{-n-1} < 0$ para $\theta > 0$, y entonces θ^{-n} es decreciente para $\theta \geq x_{(n)}$. Por tanto, $L(\theta)$ se maximiza con $\theta = x_{(n)}$, y entonces $\hat{\theta} = X_{(n)}$ es el emv de θ .

Ejemplo 52. Determina el emv de θ para la distribución teórica $N(\mu, \sigma)$ en los siguientes casos: a) $\theta = \mu$, con $\sigma = \sigma_0$ conocida, b) $\theta = \sigma$, con $\mu = \mu_0$ conocida, c) $\theta = (\mu, \sigma)$. Se tiene que $\log L(\theta) = -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$.

a) $\frac{d}{d\mu} \log L(\mu) = -\frac{1}{2\sigma_0^2} \sum_{i=1}^n 2(x_i - \mu)^{2-1} (-1) = \frac{1}{\sigma_0^2} (\sum_{i=1}^n x_i - n\mu) = 0$ tiene solución $\mu = \bar{x}$. Entonces, $\hat{\mu} = \bar{X}$ es el emv de μ .

b) La ecuación $\frac{d}{d\sigma} \log L(\sigma) = -\frac{n}{\sigma} - \frac{1}{2} (-2) \sigma^{-3} \sum_{i=1}^n (x_i - \mu_0)^2 = 0$ se puede expresar como $\frac{1}{\sigma} (-n + \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2) = 0$, que tiene solución

$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2}$. Entonces, $\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2}$ es el emv de σ .

c) En este caso, para encontrar el máximo hay que resolver el sistema

$$\left. \begin{aligned} \frac{d}{d\mu} \log L(\mu, \sigma) &= \frac{1}{\sigma^2} (\sum_{i=1}^n x_i - n\mu) = 0 \\ \frac{d}{d\sigma} \log L(\mu, \sigma) &= \frac{1}{\sigma} (-n + \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2) = 0 \end{aligned} \right\}$$

que tiene solución $\mu = \bar{x}$ y $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = s$. Mediante la matriz hessiana de derivadas segundas se comprueba que este punto crítico es punto de máximo. Entonces, $\hat{\theta} = (\hat{\mu}, \hat{\sigma}) = (\bar{X}, s)$ es el emv de $\theta = (\mu, \sigma)$.

Hay casos en los que el emv no existe. Si existe, el emv puede no ser único, aunque frecuentemente lo es. Por ejemplo, para una distribución teórica uniforme $U(\theta, \theta + 1)$, se puede comprobar que para cualquier α con $0 \leq \alpha \leq 1$ se tiene que $\alpha(X_{(n)} - 1) + (1 - \alpha)X_{(1)}$ es un emv, teniendo en cuenta que $x_{(1)} \geq \theta$ y $x_{(n)} \leq \theta + 1$ si y solo si $x_{(n)} - 1 \leq \theta \leq x_{(1)}$.

Teorema 53. Sea T un estadístico suficiente. Si el emv existe y es único, es función de T .

Demostración. Se tiene que existen funciones h y g_θ con $L(\theta) = f_\theta(\mathbf{x}) = h(\mathbf{x}) g_\theta(T(\mathbf{x}))$, siendo $\mathbf{x} = (x_1, \dots, x_n)$. El valor de θ que maximiza $L(\theta)$ es el mismo que el que maximiza $g_\theta(T(\mathbf{x}))$. Este valor será función de $T(\mathbf{x})$, puesto que dos muestras \mathbf{x} y \mathbf{x}' con $T(\mathbf{x}) = T(\mathbf{x}') = t$ dan lugar a la misma función $g_\theta(t)$, con el mismo máximo. ■

Aunque con frecuencia el mismo emv es un estadístico suficiente, no siempre lo es.

El emv tiene una útil propiedad de invarianza para transformaciones del parámetro. Por ejemplo, si $\hat{\sigma}$ es el emv de σ , entonces $\hat{\sigma}^2$ es el emv de σ^2 . Para obtener el emv de σ^2 habría que expresar la función de masa o densidad teórica en términos de $\lambda = \sigma^2$ (reparametrizar), y maximizar la función de verosimilitud para λ . Es sencillo obtener que el máximo para λ es justamente $\hat{\sigma}^2$. Hemos considerado la transformación $g(\sigma) = \sigma^2$.

Esto se cumple en general para cualquier aplicación g definida sobre Θ : si $\hat{\theta}$ es el emv de θ , entonces $\hat{\lambda} = g(\hat{\theta})$ es el emv de $\lambda = g(\theta)$.

Ejemplo 54. El número de fallos en bobinas de cable de 1 km sigue una distribución de Poisson. Se puede comprobar si una bobina tiene algún fallo conectando el principio y el final de la bobina a un circuito eléctrico y comprobando si circula la electricidad. De

1.000 bobinas se comprobó que 78 de ellas tenían algún fallo. Determina el emv para el número medio (el número esperado) de fallos por bobina.

Solución 55. De los datos, lo único que se puede estimar directamente es la probabilidad, que llamamos θ , de que una bobina tenga algún fallo (mediante la proporción 0.078 de bobinas con fallos). Sin embargo, la utilización del supuesto paramétrico de que el número de fallos es de Poisson, y el uso de la propiedad anterior de los emv, permite estimar el número medio de fallos.

Disponemos de una muestra de una distribución teórica de Bernoulli (hemos observado si hay fallo o no). Para $i = 1, \dots, n$, con $n = 1.000$, sea X_i la v.a. que vale 1 si la bobina i -ésima tiene algún fallo y vale 0 si no tiene fallo. Se tiene que X_1, \dots, X_n es una m.a.s. de una $B(1, \theta)$. Se ha observado una realización x_1, \dots, x_n de esta muestra, con $\sum_{i=1}^n x_i = 78$.

Sea Z la v.a. "número de fallos en una bobina". No hemos observado Z , pero sabemos que $Z \sim \mathcal{P}(\lambda)$ para algún valor de λ . El número esperado de fallos es la esperanza de la $\mathcal{P}(\lambda)$, que es igual a λ . Queremos entonces estimar λ .

La probabilidad de que una bobina tenga algún fallo, se expresa en términos de X y de Z como $P\{X = 1\}$ y $P\{Z > 0\}$, respectivamente. Ésto permite relacionar θ con λ , y estimar λ mediante el emv de θ . Se tiene que:

$$\theta = P\{X = 1\} = P\{Z > 0\} = 1 - P\{Z = 0\} = 1 - e^{-\lambda} \frac{\lambda^0}{0!} = 1 - e^{-\lambda}.$$

Despejando, se obtiene $\lambda = g(\theta) = -\log(1 - \theta)$. El emv de θ (el parámetro de la $B(1, \theta)$) es $\hat{\theta} = \bar{X}$, y entonces se tiene que el emv de $\lambda = g(\theta)$ es

$$\hat{\lambda} = g(\hat{\theta}) = -\log(1 - \hat{\theta}) = -\log(1 - \bar{X}).$$

Con este estimador se obtiene la siguiente estimación:

$$\hat{\lambda} = -\log(1 - \bar{x}) = -\log(1 - 0.078) = 0.0812.$$

Ejemplo 56. Obtén el emv para el séptimo decil de una distribución teórica $N(\mu, \sigma)$.

Solución 57. Suponemos que ninguno de los parámetros es conocido. El séptimo decil es el cuantil $d = x_{0.7}$ tal que $P\{N(\mu, \sigma) \leq d\} = 0.7$. Lo expresamos en función de μ y σ , y obtenemos el estimador estimando μ y σ en la expresión mediante su emv, ya obtenido en un ejemplo anterior. Se tiene que

$$P\{N(\mu, \sigma) \leq d\} = P\{N(0, 1) \leq \frac{d-\mu}{\sigma}\} = 0.7,$$

y entonces $P\{N(0, 1) > \frac{d-\mu}{\sigma}\} = 0.3$. Puesto que $P\{N(0, 1) > 0.52\} = 0.3015$, se tiene que $\frac{d-\mu}{\sigma} = 0.52$, y despejando se obtiene que el séptimo decil es

$$d = \mu + 0.52\sigma = g(\mu, \sigma).$$

Se tiene que el emv de (μ, σ) es $(\hat{\mu}, \hat{\sigma}) = (\bar{X}, s)$, y entonces

$$\hat{d} = g(\hat{\mu}, \hat{\sigma}) = \hat{\mu} + 0.52\hat{\sigma} = \bar{X} + 0.52s$$

es el emv de $d = g(\mu, \sigma) = x_{0.7}$.

También podemos estimar $x_{0.7}$ mediante el correspondiente cuantil muestral $c_{0.7}$, por ejemplo. Obsérvese que en este caso el estimador no es función del estadístico minimal suficiente (\bar{X}, s) , como sí ocurre con el emv.

En el siguiente teorema se presenta otra propiedad muy útil del emv. Bajo ciertas condiciones que se cumplen con frecuencia, la distribución del emv es asintóticamente normal.

Teorema 58. Supongamos que $\Theta \subset \mathbb{R}$ es abierto, y que $f_\theta(x)$ satisface las siguientes condiciones:

1.- Existe la derivada $\frac{d^3}{d\theta^3} \log f_\theta(x)$, y su valor absoluto está acotado por una función $K(x)$ tal que $E_\theta[K(X)] \leq k$ (para algún $k > 0$) para todo θ .

2.- Se tiene que $E_\theta \left[\frac{d}{d\theta} \log f_\theta(X) \right] = 0$ y $E_\theta \left[\frac{1}{f_\theta(X)} \frac{d^2}{d\theta^2} f_\theta(X) \right] = 0$, al menos para un $\theta = \theta_0$.

3.- Se tiene que $i(\theta) > 0$, con

$$i(\theta) = E_\theta \left[\left(\frac{d}{d\theta} \log f_\theta(X) \right)^2 \right].$$

Bajo estas condiciones, si $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ es una sucesión de estimadores de máxima verosimilitud que converge en probabilidad a θ , se tiene que

$$\hat{\theta}_n \sim AN \left(\theta, \frac{1}{\sqrt{n \cdot i(\theta)}} \right) \quad (3.2)$$

para todo $\theta \in \Theta$ (asintóticamente normal).

Por otra parte, si el emv es único, entonces la sucesión $\hat{\theta}_n$ correspondiente a los sucesivos tamaños de muestra converge en probabilidad a θ .

Recuérdese que la expresión $\hat{\theta}_n \sim AN \left(\theta, \frac{1}{\sqrt{n \cdot i(\theta)}} \right)$ significa que $\frac{\hat{\theta}_n - \theta}{1/\sqrt{n \cdot i(\theta)}} \sim AN(0, 1)$,

esto es, que la v.a. $(\hat{\theta}_n - \theta) \sqrt{n \cdot i(\theta)}$ converge en distribución a una $N(0, 1)$. A la función $i(\theta)$ de la condición 3 se la denomina *información de Fisher*. Obsérvese que la desviación típica en (3.2) es menor cuanto mayor sea $i(\theta)$, y por ello puede ser considerada como una medida de la información que aporta una muestra de tamaño 1 sobre θ . Una expresión alternativa para $i(\theta)$, cuando se cumple la condición 2, es la siguiente:

$$i(\theta) = -E_\theta \left[\frac{d^2}{d\theta^2} \log f_\theta(X) \right]. \quad (3.3)$$

En algunas ocasiones es más sencillo calcular $i(\theta)$ de esta manera.

Las condiciones 1 y 2 son suficientes pero no necesarias para que se cumpla el resultado. Pueden ser difíciles de comprobar, pero se cumplen con frecuencia. Para comprobar

que se cumple 2 es útil tener en cuenta que

$$\frac{1}{f_{\theta}(X)} \frac{d^2}{d\theta^2} f_{\theta}(X) = \frac{d^2}{d\theta^2} \log f_{\theta}(X) + \left(\frac{d}{d\theta} \log f_{\theta}(X) \right)^2,$$

que se obtiene expresando $\frac{d^2}{d\theta^2} \log f_{\theta}(X)$ en términos de las derivadas de $f_{\theta}(X)$ y despejando. En algunas ocasiones puede ser más sencillo comprobar (3.2) de un modo directo que comprobar que se cumplen las condiciones 1 y 2. En los casos de los siguientes ejemplos se cumplen 1 y 2; se obtiene después de cálculos sencillos pero algo laboriosos, con alguna matización para la condición 1. Nos limitamos aquí a calcular la información de Fisher $i(\theta)$, que es positiva en todos los casos.

Ejemplo 59. Determinamos la distribución asintótica del emv $\hat{\theta}$ para el parámetro θ de las siguientes distribuciones teóricas: a) $B(1, \theta)$, b) $\mathcal{P}(\theta)$, c) $Exp(\theta)$.

En el ejemplo 51 se obtienen los emv $\hat{\theta}$ para estos casos.

a) Puesto que $f_{\theta}(x) = \theta^x (1 - \theta)^{1-x}$, se tiene que $f_{\theta}(X) = \theta^X (1 - \theta)^{1-X}$ (es una v.a. transformada de X), y $\log f_{\theta}(X) = X \log \theta + (1 - X) \log(1 - \theta)$. Derivando y operando se obtiene, para $0 < \theta < 1$:

$$\begin{aligned} i(\theta) &= E_{\theta} \left[\left(\frac{X - \theta}{\theta(1 - \theta)} \right)^2 \right] = \frac{1}{\theta^2(1 - \theta)^2} E_{\theta} [(X - \theta)^2] \\ &= \frac{1}{\theta^2(1 - \theta)^2} V_{\theta}[X] = \frac{\theta(1 - \theta)}{\theta^2(1 - \theta)^2} = \frac{1}{\theta(1 - \theta)} > 0. \end{aligned}$$

Se tiene que $\hat{\theta} = \bar{X}$ y, por el teorema 58, se tiene que

$$\bar{X} \sim N \left(\theta, \frac{\sqrt{\theta(1 - \theta)}}{\sqrt{n}} \right).$$

Obsérvese que también se puede obtener este resultado de un modo inmediato a partir del teorema central del límite (TCL), puesto que $\theta = \mu$ y $\theta(1 - \theta) = \sigma^2$.

b) Para $\theta > 0$ se tiene que $\log f_{\theta}(X) = -\theta + X \log \theta - \log(X!)$, y de aquí:

$$\begin{aligned} i(\theta) &= E_{\theta} \left[\left(\frac{X - \theta}{\theta} \right)^2 \right] = \frac{1}{\theta^2} E_{\theta} [(X - \theta)^2] \\ &= \frac{1}{\theta^2} V_{\theta}[X] = \frac{1}{\theta^2} \theta = \frac{1}{\theta} > 0 \end{aligned}$$

Se tiene que $\hat{\theta} = \bar{X}$ y, por el teorema 58, se tiene que

$$\bar{X} \sim N \left(\theta, \frac{\sqrt{\theta}}{\sqrt{n}} \right).$$

(También es inmediato a partir del TCL).

c) Para $\theta > 0$ se tiene que $\log f_{\theta}(X) = \log \theta - \theta X$, y $\frac{d}{d\theta} \log f_{\theta}(X) = \frac{1}{\theta} - X$, y de aquí se

obtiene

$$i(\theta) = E_{\theta} \left[\left(\frac{1}{\theta} - X \right)^2 \right] = V_{\theta} [X] = \frac{1}{\theta^2} > 0$$

Se tiene que $\hat{\theta} = 1/\bar{X}$ y, por el teorema 58, se tiene que

$$\frac{1}{\bar{X}} \sim N \left(\theta, \frac{\theta}{\sqrt{n}} \right).$$

Aunque ya es innecesario, también obtenemos $i(\theta)$ utilizando la expresión (3.3). Se tiene que

$$\begin{aligned} \frac{d^2}{d\theta^2} \log f_{\theta}(X) &= \frac{d}{d\theta} \left(\frac{d}{d\theta} \log f_{\theta}(X) \right) = \frac{d}{d\theta} \left(\frac{1}{\theta} - X \right) = -\frac{1}{\theta^2}, \text{ y} \\ i(\theta) &= -E_{\theta} [-1/\theta^2] = 1/\theta^2. \end{aligned}$$

Si se desea obtener la distribución asintótica del emv $\hat{\lambda} = g(\hat{\theta})$ de una función del parámetro $\lambda = g(\theta)$, siendo $\hat{\theta}$ el emv de θ , el siguiente resultado es útil.

Proposición 60. Sea $Y_n, n \in \mathbb{N}$, una sucesión de v.a. con $Y_n \sim AN(\mu, \sigma_n)$, siendo μ un número real dado y σ_n una sucesión de números que tiende a 0. Si g una función real derivable en μ , con $g'(\mu) \neq 0$. Entonces,

$$g(Y_n) \sim AN(g(\mu), |g'(\mu)| \sigma_n).$$

Si aplicamos el resultado a $\hat{\theta}_n \sim AN\left(\theta, \frac{1}{\sqrt{n \cdot i(\theta)}}\right)$ se obtiene

$$g(\hat{\theta}_n) \sim AN\left(g(\theta), \frac{|g'(\theta)|}{\sqrt{n \cdot i(\theta)}}\right)$$

Ejemplo 61. En el ejemplo 54 considerábamos la distribución teórica $B(1, \theta)$ y estimábamos $\lambda = g(\theta) = -\log(1 - \theta)$, obteniendo el emv $\hat{\lambda} = -\log(1 - \bar{X})$. Puesto que $g'(\theta) = \frac{1}{1-\theta} > 0$ para todo $\theta \in (0, 1)$, se obtiene que, sea cual sea el verdadero valor θ del parámetro, $-\log(1 - \bar{X}) \sim AN\left(-\log(1 - \theta), \sqrt{\frac{\theta}{1-\theta}}/\sqrt{n}\right)$, o

$$\hat{\lambda} \sim AN\left(\lambda, \sqrt{e^{\lambda} - 1}/\sqrt{n}\right).$$

Si dispusiéramos de los datos del número de roturas en cada cable en el ejemplo 54, podríamos estimar λ mediante el emv de la distribución de Poisson, que es la media del número de roturas, este estimador $\tilde{\lambda}$ cumple: $\tilde{\lambda} \sim AN\left(\lambda, \sqrt{\lambda}/\sqrt{n}\right)$. La varianza de $\tilde{\lambda}$ es menor que la de $\hat{\lambda}$ para todo n y $\lambda > 0$, aunque el cociente de varianzas es cercano a 1 para λ pequeño (y tiende a 1 cuando λ tiende a 0).

3.4 Estimadores de Bayes

3.4.1 Introducción a la inferencia bayesiana

Inferencia clásica y bayesiana

Otro enfoque en inferencia estadística es el *planteamiento bayesiano*. En el denominado enfoque clásico, que es el que venimos siguiendo, se considera que la distribución teórica F , desconocida, es fija, característica del experimento aleatorio en cuestión.

En el planteamiento bayesiano no se considera que exista una única distribución teórica F , fija, que es la verdadera distribución que genera los datos del experimento aleatorio. En lugar de ello, se considera una distribución de probabilidad sobre el conjunto de posibles distribuciones teóricas, que se denomina “distribución a priori”. Si la distribución teórica es paramétrica, la distribución a priori viene dada mediante una distribución de probabilidad sobre el espacio paramétrico. Este nuevo elemento complementa los elementos del planteamiento clásico.

En algunas ocasiones un problema se ajusta a un esquema bayesiano. Por ejemplo, supongamos que se dispone de varias urnas con distintas proporciones de bolas blancas. Se selecciona una urna mediante un procedimiento aleatorio, con una distribución de probabilidad específica; por ejemplo, con equiprobabilidad. Se extrae una bola al azar de la urna seleccionada, y se anota un 1 si resulta ser bola blanca y un 0 en otro caso. La manera adecuada de formular esta situación es mediante un enfoque bayesiano. Se está obteniendo una observación de una v.a. de Bernoulli. El espacio paramétrico es el conjunto de los valores de la proporción de bolas blancas en las distintas urnas. El valor del parámetro solo es fijo una vez seleccionada la urna, pero antes de ello se considera una distribución de probabilidad sobre los posibles valores del parámetro, de la que se obtiene una realización que lo determina.

En otras ocasiones, la distribución a priori es un artificio que permite añadir, a la información proporcionada por los datos, otra información disponible sobre la distribución teórica. Por ejemplo, queremos estudiar la longitud media μ de los tornillos producidos por una máquina. Por limitaciones físicas de la máquina, sabemos que el valor de μ está comprendido entre 7.2 y 7.5 mm. Con un enfoque clásico, esta información no es utilizada para realizar inferencias, y se pierde. Podemos utilizar esta información con un enfoque bayesiano, considerando una distribución a priori para μ con soporte en el intervalo $[7.2, 7.5]$. Si no tenemos más información, deberíamos utilizar la distribución uniforme $U(7.2, 7.5)$ como distribución a priori: $\pi \sim U(7.2, 7.5)$.

Una vez especificada la distribución a priori y la distribución de la muestra, el primer objetivo es la obtención de la “distribución a posteriori”, que es la distribución del parámetro condicionada a los valores observados en la muestra. Para ello se utiliza la fórmula de Bayes, que da nombre a este planteamiento. Las inferencias se realizan a partir de la distribución a posteriori.

Distribución a priori y distribución a posteriori

Nos limitamos a estudiar el caso paramétrico. La *distribución a priori* está definida sobre el espacio paramétrico y se denota por π , y entonces toma valores $\pi(\theta)$ para $\theta \in \Theta$. Puede ser una función de masa o una función de densidad. También se denomina *distribución prior*.

El hecho de que θ sea una variable aleatoria, determina que la función de masa o densidad de la muestra $f_{\theta}(\mathbf{x}) = f_{\theta}(x_1, \dots, x_n)$ sea una función de masa o densidad condicionada al valor θ del parámetro, y entonces se denota por $f(\mathbf{x}/\theta) = f(x_1, \dots, x_n/\theta)$, con $\mathbf{x} = (x_1, \dots, x_n)$. Es la misma función para los casos clásico y bayesiano, pero ahora, además de la notación, la interpretación es diferente, es una distribución condicionada.

La función de masa o densidad marginal de la muestra se denomina *distribución predictiva*. Se denota por m y viene dada por

$$m(\mathbf{x}) = \int_{\Theta} \pi(\theta) f(\mathbf{x}/\theta) d\theta.$$

Frecuentemente $\pi(\theta)$ es una función de densidad. Si Θ es finito o numerable, entonces $\pi(\theta)$ es una función de masa, y hay que reemplazar la integral por el correspondiente sumatorio en la expresión anterior.

La información disponible inicialmente sobre θ viene dada por la distribución a priori $\pi(\theta)$. Una vez observada la muestra \mathbf{x} , podemos incorporar la información proporcionada por estos datos utilizando la *distribución a posteriori* $\pi(\theta/\mathbf{x})$, o *distribución posterior*. Es la distribución del parámetro θ condicionada a la muestra observada $X_1 = x_1, \dots, X_n = x_n$, y viene dada por la fórmula de Bayes:

$$\pi(\theta/\mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x}/\theta)}{m(\mathbf{x})}. \quad (3.4)$$

Obsérvese que el denominador no depende de θ . Es el valor por el que hay que dividir la función de θ del numerador, $\pi(\theta) f(\mathbf{x}/\theta)$, que hace que $\pi(\theta/\mathbf{x})$ sea una función de densidad:

$$\begin{aligned} \int_{\Theta} \pi(\theta/\mathbf{x}) d\theta &= \int_{\Theta} \frac{\pi(\theta) f(\mathbf{x}/\theta)}{m(\mathbf{x})} d\theta \\ &= \frac{1}{m(\mathbf{x})} \int_{\Theta} \pi(\theta) f(\mathbf{x}/\theta) d\theta \\ &= \frac{1}{m(\mathbf{x})} m(\mathbf{x}) = 1. \end{aligned}$$

Si $\pi(\theta/\mathbf{x})$ es una función de densidad conocida, puede ser identificada a partir del numerador $\pi(\theta) f(\mathbf{x}/\theta)$. Si este numerador multiplicado por una constante (que no depende de θ) es una función de densidad conocida, entonces $\pi(\theta/\mathbf{x})$ es justamente esta densidad conocida.

Ejemplo 62. Se desea estudiar θ para una distribución de Bernoulli, con $\Theta = (0, 1)$. No se dispone de ninguna información sobre θ , y entonces es conveniente utilizar la

distribución uniforme sobre Θ , la $U(0, 1)$, como distribución a priori. Se ha observado la muestra $\mathbf{x} = (x_1, \dots, x_n)$. Determina la distribución a posteriori.

Se tiene que $\pi \sim U(0, 1)$, y entonces

$$\pi(\theta) = 1 \quad \text{si } 0 < \theta < 1 \quad (\text{y } \pi(\theta) = 0 \text{ en otro caso}).$$

Se tiene que

$$f(x_1, \dots, x_n/\theta) = \theta^t (1 - \theta)^{n-t}$$

con $x_1, \dots, x_n = 0, 1$, y $t = \sum_{i=1}^n x_i$,

y entonces

$$\pi(\theta/\mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x}/\theta)}{m(\mathbf{x})} = \frac{\theta^t (1 - \theta)^{n-t}}{m(\mathbf{x})} \quad \text{si } 0 < \theta < 1.$$

La función de densidad de la $Beta(\alpha, \beta)$ es $g_{\alpha, \beta}(z) = \frac{z^{\alpha-1}(1-z)^{\beta-1}}{Be(\alpha, \beta)}$ si $0 < z < 1$, con $Be(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$. Obsérvese que $g_{t+1, n-t+1}(\theta) = \frac{\theta^t(1-\theta)^{n-t}}{Be(\alpha, \beta)}$. La única posible diferencia entre las funciones de densidad $\pi(\theta/\mathbf{x})$ y $g_{t+1, n-t+1}(\theta)$ son los denominadores, que no dependen de θ , y entonces deben ser iguales, ya que ambas son funciones de densidad, que deben ser por tanto la misma. Entonces, $\pi(\theta/\mathbf{x}) = g_{t+1, n-t+1}(\theta)$, que es la función de densidad de una $Beta(t + 1, n - t + 1)$, y por tanto la distribución a posteriori es

$$\pi(\theta/\mathbf{x}) \sim Beta(t + 1, n - t + 1).$$

La distribución $U(0, 1)$ es una $Beta(1, 1)$. Si se considera, también para una distribución teórica de Bernoulli, como distribución a priori la $Beta(\alpha, \beta)$ para cualesquiera valores $\alpha, \beta > 0$, se obtiene de un modo similar:

$$\pi(\theta/\mathbf{x}) \sim Beta(\alpha + t, \beta + n - t),$$

(o $\pi(\theta/\mathbf{x}) \sim Beta(\alpha + t, \beta + n - t)$).

Si $\pi(\theta)$ es constante (distribución a priori uniforme), entonces la moda a posteriori coincide con el estimador de máxima verosimilitud, como explicamos a continuación. La moda a posteriori (la moda de la distribución a posteriori) es el valor de θ que maximiza $\pi(\theta/\mathbf{x})$ (véase 3.4). Puesto que ni $m(\mathbf{x})$ depende de θ ni $\pi(\theta)$ tampoco en este caso uniforme, entonces el valor de θ que maximiza $\pi(\theta/\mathbf{x})$ es el valor de θ que maximiza $f(\mathbf{x}/\theta)$. Este valor de θ es justamente el estimador de máxima verosimilitud.

Las inferencias sobre θ se realizan en función de la distribución a posteriori $\pi(\theta/\mathbf{x})$. Por ejemplo, cualquier medida de posición, como la esperanza, la moda o la mediana de $\pi(\theta/\mathbf{x})$, proporciona una estimación puntual para θ .

Los factores que intervienen en la expresión de $\pi(\theta/\mathbf{x})$ que no dependen de θ , forman parte de la constante multiplicativa que hacen que sea una función de masa o densidad. Es habitual omitir estas constantes en los cálculos para obtener $\pi(\theta/\mathbf{x})$, y para ello se utiliza la notación " \propto ", que indica que dos funciones son iguales salvo factores que no

dependen del argumento de las funciones, o lo que es lo mismo, que el cociente de las funciones no depende del argumento. De este modo, para obtener el resultado en el caso general del ejemplo anterior, podemos expresar la función de densidad posterior como:

$$\pi(\theta/\mathbf{x}) = \frac{\theta^{\alpha+t-1} (1-\theta)^{\beta+n-t-1}}{Be(\alpha, \beta) m(\mathbf{x})} \propto \theta^{\alpha+t-1} (1-\theta)^{\beta+n-t-1}. \quad (3.5)$$

Esto permite identificar la función de densidad $\pi(\theta/\mathbf{x})$, con argumento θ , como una *Beta* $(\alpha + t, \beta + n - t)$, puesto que la expresión de la derecha en (3.5) es la función de densidad de esa beta salvo constantes multiplicativas.

Ejemplo 63. Se desea estudiar θ para una distribución normal $N(\theta, \sigma)$ con σ conocida, con $\Theta = \mathbb{R}$. Se considera para θ la distribución a priori $N(\mu_0, \sigma_0)$. Se ha observado la muestra \mathbf{x} . Determina la distribución a posteriori.

Se tiene que

$$\begin{aligned} \pi(\theta) &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right\} \quad \text{si } -\infty < \theta < \infty, \text{ y} \\ f(\mathbf{x}/\theta) &= (2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right\}, \end{aligned}$$

y entonces, simplificando, se obtiene:

$$\begin{aligned} \pi(\theta/\mathbf{x}) &= \frac{\pi(\theta) f(\mathbf{x}/\theta)}{m(\mathbf{x})} \\ &= \frac{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right\} (2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right\}}{m(\mathbf{x})} \\ &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma_0^2}\mu_0^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta^2 - 2\theta\mu_0)\right\} \\ &\cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right\} \exp\left\{-\frac{1}{2\sigma^2}(n\theta^2 - 2n\theta\bar{x})\right\}, \end{aligned}$$

y entonces

$$\begin{aligned} \pi(\theta/\mathbf{x}) &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\theta^2 - 2\theta\mu_0)\right\} \exp\left\{-\frac{1}{2\sigma^2}(n\theta^2 - 2n\theta\bar{x})\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)\theta^2 + \left(\frac{\mu_0}{\sigma_0^2} + n\frac{\bar{x}}{\sigma^2}\right)\theta\right\}. \quad (3.6) \end{aligned}$$

Multiplicando por una constante adecuada para completar el cuadrado, a continuación, se obtiene que $\pi(\theta/\mathbf{x})$ es normal. Para todo a, b se tiene que $a(\theta - b)^2 = a\theta^2 - 2ab\theta + ab^2$. Igualando el exponente en (3.6) a los términos en θ^2 y θ en $a(\theta - b)^2$, se obtiene

$$\begin{aligned} -\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) &= a, \\ \frac{\mu_0}{\sigma_0^2} + n\frac{\bar{x}}{\sigma^2} &= -2ab. \end{aligned}$$

Despejando, se obtiene

$$b = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\bar{x}.$$

Entonces, con estos valores de a y b se obtiene, teniendo en cuenta que $\exp\{ab^2\}$ no depende de θ ,

$$\begin{aligned}\pi(\theta/\mathbf{x}) &\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)\theta^2 + \left(\frac{\mu_0}{\sigma_0^2} + n\frac{\bar{x}}{\sigma^2}\right)\theta\right\} \\ &= \exp\{a\theta^2 - 2ab\theta\} \\ &\propto \exp\{a\theta^2 - 2ab\theta\} \exp\{ab^2\} \\ &= \exp\{a\theta^2 - 2ab\theta + ab^2\} = \exp\{a(\theta - b)^2\} \\ &= \exp\left\{-\frac{1}{2\sigma_1^2}(\theta - \mu_1)^2\right\},\end{aligned}$$

con $\mu_1 = b$ y $\sigma_1^2 = -\frac{1}{2a} > 0$. Esta función exponencial es la función de densidad de una $N(\mu_1, \sigma_1)$, salvo la constante multiplicativa $(\frac{1}{\sqrt{2\pi}\sigma_1})$, que no tenemos que obtener ya), con

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\bar{x} \quad \text{y} \quad (3.7)$$

$$\sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}. \quad (3.8)$$

Por tanto, $\pi(\theta/\mathbf{x})$ es la función de densidad de una $N(\mu_1, \sigma_1)$, obteniéndose la distribución a posteriori:

$$\pi(\theta/\mathbf{x}) \sim N(\mu_1, \sigma_1),$$

con μ_1 y σ_1 dados en (3.7) y (3.8).

Familias conjugadas de distribuciones

Cuando las distribuciones prior y posterior pertenecen a la misma familia paramétrica, se dice que esta familia es *conjugada* de la distribución teórica que se esté considerando. De este modo, por el ejemplo 86 se tiene que la familia Beta es conjugada de la familia de Bernoulli, y por el ejemplo 63 se tiene que la Normal es conjugada de la Normal considerando $\theta = \mu$ y σ conocida.

Para una distribución teórica dada, el hecho de que una distribución a priori sea conjugada de ella no es un motivo suficiente para seleccionarla. Sin embargo, en algunos casos, como los de los ejemplos anteriores, una distribución conjugada puede ser adecuada.

Por ejemplo, la distribución beta puede ser adecuada como distribución a priori para el parámetro de una Bernoulli. Con una elección conveniente de los parámetros, se consigue concentrar la probabilidad a priori donde se quiera dentro del soporte, y con una dispersión también a elección.

Por ejemplo, si creemos que una moneda debería estar casi equilibrada, y tomamos como distribución a priori para θ la *Beta* (α, β) , como en el ejemplo 86, podemos elegir

los parámetros de modo que sea $\mu_0 = 1/2$. Ésto se consigue tomando $\alpha = \beta$. Además, podemos elegir este valor común de modo que σ_0 tome un valor especificado. Si pensamos que la verdadera probabilidad de éxito θ no debería alejarse mucho de $1/2$, tomaremos σ_0 pequeño, y si no creemos que sea así tomaremos σ_0 más grande. Por ejemplo, para que además de $\mu_0 = 1/2$ sea $\sigma_0 = 0.1$, se obtiene $\alpha = \beta = 12$ despejando α y β de las ecuaciones

$$\mu_0 = \frac{\alpha}{\alpha + \beta} = \frac{1}{2} \quad \text{y} \quad \sigma_0^2 = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} = 0.1^2.$$

Ejemplo 64. Consideremos una distribución teórica con función de densidad

$$f(x/\theta) = \theta(1+x)^{-\theta-1} \quad \text{para } x > 0,$$

siendo θ un parámetro positivo. Comprobamos que la distribución gamma es conjugada de esta distribución teórica, y determinamos la distribución a posteriori para una distribución a priori $Exp(\lambda)$.

Se tiene que $\Theta = (0, \infty)$. La función de densidad de la m.a.s. y la función de densidad a priori de $\theta \sim \gamma(p, a)$ son

$$f(\mathbf{x}/\theta) = \theta^n \left(\prod_{i=1}^n (1+x_i) \right)^{-\theta-1} \quad \text{para } x_1, \dots, x_n > 0, \text{ y}$$

$$\pi(\theta) = \frac{a^p}{\Gamma(p)} \theta^{p-1} e^{-a\theta} \quad \text{para } \theta > 0.$$

Para el cálculo de $\pi(\theta/\mathbf{x})$ identificamos los factores que dependen de θ en el producto $\pi(\theta) f(\mathbf{x}/\theta)$. En primer lugar, obsérvese que

$$\begin{aligned} \left(\prod_{i=1}^n (1+x_i) \right)^{-\theta-1} &= \left(\prod_{i=1}^n (1+x_i) \right)^{-\theta} \left(\prod_{i=1}^n (1+x_i) \right)^{-1} \\ &\propto \left(\prod_{i=1}^n (1+x_i) \right)^{-\theta} = \exp \left\{ -\theta \sum_{i=1}^n \log(1+x_i) \right\}. \end{aligned}$$

Se tiene que

$$\begin{aligned} \pi(\theta/\mathbf{x}) &\propto \pi(\theta) f(\mathbf{x}/\theta) \propto \theta^{p-1} e^{-a\theta} \theta^n \left(\prod_{i=1}^n (1+x_i) \right)^{-\theta-1} \\ &\propto \theta^{p-1} e^{-a\theta} \theta^n \exp \left\{ -\theta \sum_{i=1}^n \log(1+x_i) \right\} \\ &= \theta^{p+n-1} \exp \left\{ -\theta \left[a + \sum_{i=1}^n \log(1+x_i) \right] \right\}. \end{aligned}$$

Obsérvese que ésta es la función de densidad de una $\gamma(p+n, a+t)$, con

$$t = \sum_{i=1}^n \log(1+x_i),$$

salvo constantes multiplicativas, y entonces la distribución a posteriori es

$$\pi(\theta/\mathbf{x}) \sim \gamma(p+n, a+t).$$

Por tanto, la distribución gamma es conjugada de la distribución teórica considerada. Puesto que la $Exp(\lambda)$ es una $\gamma(1, \lambda)$, si $\pi(\theta) \sim Exp(\lambda)$ se tiene que

$$\pi(\theta/\mathbf{x}) \sim \gamma(n+1, t+\lambda).$$

Cuando la distribución prior no es conjugada de la distribución teórica, puede ocurrir que la distribución posterior sea poco manejable. En este caso, el uso del ordenador es muy útil para realizar cálculos numéricos para la distribución posterior.

Distribuciones a priori no informativas

La información que se incluye en el problema mediante la distribución a priori suele ser subjetiva. Se puede utilizar la experiencia de una persona en el campo de trabajo donde se esté realizando el estudio estadístico. Se puede incluir su opinión a través de la distribución a priori.

En algunos casos, también se puede considerar una distribución a priori que no contenga información. Si Θ es acotado, entonces la distribución uniforme en Θ es *no informativa*. La función de densidad es constante, y entonces la probabilidad de un intervalo en Θ es proporcional a la amplitud del intervalo. No hay zonas con mayor densidad de probabilidad que otras. En el ejemplo 86 considerábamos una distribución a priori uniforme, en $\Theta = (0, 1)$.

Si Θ es un intervalo no acotado, entonces no existe la distribución uniforme en Θ : para cualquier constante positiva c , la función constante $\pi(\theta) = c$ tiene integral sobre Θ que vale ∞ , y entonces π no sería una función de densidad. En un intento de acercarnos a una distribución que no concentre mayor probabilidad en unas zonas que en otras, podemos repartir la probabilidad en un rango amplio de valores tomando una varianza a priori grande. En el ejemplo 63 se estudia el parámetro θ para una distribución teórica $N(\theta, \sigma)$ con σ conocida, con distribución a priori $N(\mu_0, \sigma_0)$. Cuanto mayor es σ_0 , menos informativa es la $N(\mu_0, \sigma_0)$. Comprobamos cual es el efecto en la distribución a posteriori. Se obtuvo que $\pi(\theta/\mathbf{x}) \sim N(\mu_1, \sigma_1)$, con

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\bar{x} \quad \text{y} \quad \sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

Calculando el límite cuando $\sigma_0 \rightarrow \infty$, se obtiene que para σ_0 grande $\mu_1 \simeq \bar{x}$ y $\sigma_1^2 \simeq \sigma^2/n$, y entonces estos momentos a posteriori tienden a no depender de μ_0 ni de σ_0 para σ_0 grande.

En ocasiones, se utiliza el artificio de utilizar formalmente una función no negativa $\pi(\theta)$ con $\int_{\Theta} \pi(\theta) d\theta = \infty$, como si fuera una función de densidad a priori. A una tal función $\pi(\theta)$ se la denomina *función de densidad impropia*. En particular, ésto permite considerar lo que sería una "distribución uniforme" sobre Θ cuando es no acotado; por ejemplo, para el caso normal que acabamos de comentar. En el cálculo de $\pi(\theta/\mathbf{x})$, el elemento $\pi(\theta) = c$ se cancela en el numerador con el denominador:

$$\pi(\theta/\mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x}/\theta)}{\int_{\Theta} \pi(\theta) f(\mathbf{x}/\theta) d\theta} = \frac{c \cdot f(\mathbf{x}/\theta)}{c \cdot \int_{\Theta} f(\mathbf{x}/\theta) d\theta} = \frac{f(\mathbf{x}/\theta)}{\int_{\Theta} f(\mathbf{x}/\theta) d\theta}. \quad (3.9)$$

Obsérvese que aunque $\pi(\theta)$ no sea una función de densidad, $\pi(\theta/\mathbf{x})$ sí lo es:

$$\int_{\Theta} \pi(\theta/\mathbf{x}) d\theta = \int_{\Theta} \frac{f(\mathbf{x}/\theta)}{\int_{\Theta} f(\mathbf{x}/\theta) d\theta} = \frac{1}{\int_{\Theta} f(\mathbf{x}/\theta) d\theta} \int_{\Theta} f(\mathbf{x}/\theta) d\theta = 1.$$

La utilización de tal $\pi(\theta)$ es formalmente válida porque, aunque no sea función de densidad, se puede operar con ella para el cálculo de $\pi(\theta/\mathbf{x})$ obteniéndose una función de densidad.

Ejemplo 65. Se desea estudiar θ para una distribución normal $N(\theta, \sigma)$ con σ conocida, con $\Theta = \mathbb{R}$. Se considera a priori para θ la función de densidad impropia $\pi(\theta) = 1$. Se ha observado la muestra \mathbf{x} . Determinamos la distribución a posteriori.

Por (3.9) se tiene que $\pi(\theta/\mathbf{x}) \propto f(\mathbf{x}/\theta)$. Operando, se obtiene

$$\begin{aligned} \pi(\theta/\mathbf{x}) &\propto f(\mathbf{x}/\theta) \propto \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2} (\sum_{i=1}^n x_i^2 + n\theta^2 - 2n\theta\bar{x})\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2\right\} \exp\left\{-\frac{1}{2\sigma^2} (n\theta^2 - 2n\theta\bar{x})\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2} (n\theta^2 - 2n\theta\bar{x})\right\} \\ &= \exp\left\{-\frac{1}{2} \frac{n}{\sigma^2} \theta^2 + n \frac{\bar{x}}{\sigma^2} \theta\right\}. \end{aligned}$$

Entonces, la función de densidad $\pi(\theta/\mathbf{x})$ tiene una estructura como la del ejemplo 63: depende del argumento θ a través de la exponencial de un polinomio de segundo grado en θ . Razonamos como en el ejemplo 63 para obtener que la distribución a posteriori es normal, calculando también los parámetros. Para completar el cuadrado, consideramos la igualdad $a(\theta - b)^2 = a\theta^2 - 2ab\theta + ab^2$, planteamos las ecuaciones

$$-\frac{1}{2} \frac{n}{\sigma^2} = a, \quad n \frac{\bar{x}}{\sigma^2} = -2ab,$$

y despejamos, obteniendo $b = \bar{x}$. Entonces, con estos valores de a y b se obtiene,

$$\begin{aligned} \pi(\theta/\mathbf{x}) &\propto \exp\left\{-\frac{1}{2} \frac{n}{\sigma^2} \theta^2 + n \frac{\bar{x}}{\sigma^2} \theta\right\} = \exp\{a\theta^2 - 2ab\theta\} \\ &\propto \exp\{a\theta^2 - 2ab\theta\} \exp\{ab^2\} = \exp\{a\theta^2 - 2ab\theta + ab^2\} \\ &= \exp\{a(\theta - b)^2\} = \exp\left\{-\frac{1}{2} \frac{n}{\sigma^2} (\theta - \bar{x})^2\right\} \\ &= \exp\left\{-\frac{1}{2 \cdot (\sigma^2/n)} (\theta - \bar{x})^2\right\} \quad \text{para } -\infty < \theta < \infty. \end{aligned}$$

Recuérdese que aquí el argumento es θ , y el resto de elementos son constantes: σ^2 es conocida y \bar{x} es función de la realización muestral \mathbf{x} . La densidad normal $N(\mu_1, \sigma_1)$ tiene esta estructura, con $\mu_1 = \bar{x}$ y $\sigma_1^2 = \sigma^2/n$, y entonces

$$\pi(\theta/\mathbf{x}) \sim N\left(\bar{x}, \frac{\sigma}{\sqrt{n}}\right). \quad (3.10)$$

Como vemos, el hecho de poder ignorar las constantes multiplicativas simplifica las operaciones, puesto que no hay que calcular integrales. Además, simplifica también las expresiones al omitir estas constantes.

La noción de suficiencia se expresa del siguiente modo en un contexto bayesiano: un estadístico $T = T(\mathbf{x})$ es suficiente si se tiene que $\pi(\theta/\mathbf{x}) = \pi(\theta/t)$, con $t = T(\mathbf{x})$, para cualquier distribución a priori $\pi(\theta)$. De este modo, si T toma el mismo valor para dos muestras \mathbf{x} y \mathbf{x}' , $T(\mathbf{x}) = T(\mathbf{x}') = t$, entonces $\pi(\theta/\mathbf{x})$ coincide con $\pi(\theta/\mathbf{x}')$, que por este motivo se denota por $\pi(\theta/t)$. Se tiene que los estadísticos suficientes bayesianos para una familia paramétrica son los mismos que los estadísticos suficientes clásicos, estudiados en el tema 2.

3.4.2 Estimación puntual bayesiana

La información disponible sobre θ una vez observada la muestra \mathbf{x} viene dada por la distribución a posteriori $\pi(\theta/\mathbf{x})$. Se puede considerar como estimador puntual de θ a alguna característica de posición de $\pi(\theta/\mathbf{x})$, como la esperanza, la mediana o la moda.

Se denomina *estimador (puntual) de Bayes* a la esperanza de la distribución a posteriori:

$$\hat{\theta} = E[\theta/\mathbf{x}] = \int_{\Theta} \theta \pi(\theta/\mathbf{x}) d\theta.$$

En el ejemplo 86 se consideró una distribución teórica $B(1, \theta)$. Con una distribución a priori $Beta(\alpha, \beta)$, para cualesquiera valores $\alpha, \beta > 0$, la distribución a posteriori es:

$$\pi(\theta/\mathbf{x}) \sim Beta(\alpha + t, \beta + n - t),$$

con $t = \sum_{i=1}^n x_i$, con esperanza $\mu_1 = \frac{\alpha+t}{\alpha+\beta+n}$ (igual a $\frac{\alpha/n+\bar{x}}{\alpha/n+\beta/n+1}$ si $n \geq 1$). Se obtiene entonces la siguiente estimación de Bayes:

$$\hat{\theta} = \frac{\alpha + t}{\alpha + \beta + n}$$

Obsérvese que si no hay muestra, $n = 0$ (y por tanto $t = 0$), se tiene que $\mu_1 = \mu_0$, puesto que en este caso $\pi(\theta/\mathbf{x}) = \pi(\theta)$. En sentido contrario, la influencia de la distribución a priori en la distribución a posteriori tiende a desaparecer cuando el tamaño de muestra crece. En este caso, la información proporcionada por la muestra se impone sobre la información proporcionada por la distribución a priori: se obtiene que $\hat{\theta} = \hat{\theta}_n$ tiende a \bar{x} cuando n tiende a infinito, que no depende de α y β .

Puesto que la $U(0, 1)$ es la $Beta(1, 1)$, se tiene con esta distribución a priori no informativa que el estimador de Bayes es

$$\hat{\theta} = \frac{t + 1}{n + 2} = \frac{\bar{x} + 1/n}{1 + 2/n},$$

que es distinto del estimador de máxima verosimilitud clásico.

Ejemplo 66. La proporción de enfermos θ en una determinada colectividad puede ser modelizada mediante una variable aleatoria con distribución beta. De estudios previos, puede suponerse que $\theta \sim \text{Beta}(5, 10)$ (con $\mu_0 = 1/3$ y $\sigma_0 = 0.118$). En una muestra de 100 individuos hay 20 enfermos. Se obtiene la estimación de Bayes

$$\hat{\theta} = E[\theta/\mathbf{x}] = \frac{5}{23}.$$

En el ejemplo 63 se consideró una distribución teórica $N(\theta, \sigma)$ con σ conocida, con $\Theta = \mathbb{R}$. Para la distribución prior $N(\mu_0, \sigma_0)$ para θ , se obtuvo que la distribución posterior es $N(\mu_1, \sigma_1)$, con $\mu_1 = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\bar{x}$, y entonces el estimador de Bayes es

$$\hat{\theta} = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\bar{x}.$$

Obsérvese que $\hat{\theta}$ es una combinación lineal convexa (con pesos no negativos que suman 1) de μ_0 , la esperanza prior y \bar{x} , la media muestral. Para comparar los pesos consideramos el cociente, que vale $n\sigma_0^2/\sigma^2$. Por ejemplo, los pesos son iguales, valen 1/2, si $n\sigma_0^2/\sigma^2 = 1$, esto es, si $n = \sigma^2/\sigma_0^2$.

Si $n = 0$ se tiene que $\mu_1 = \mu_0$ y $\sigma_1 = \sigma_0$. Se obtiene que $\hat{\theta} = \hat{\theta}_n$ tiende a \bar{x} cuando n tiende a infinito.

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{\sigma_0^2}{\sigma^2/n + \sigma_0^2}\bar{x} \simeq \bar{x}.$$

Ejemplo 67. Consideremos una muestra x_1, \dots, x_n de una distribución teórica $N(\theta, \sigma)$ con σ conocida, y una distribución a priori no informativa. Determinamos el estimador de Bayes para θ . Por el ejemplo 65, página 19, se tiene que

$$\pi(\theta/\mathbf{x}) \sim N\left(\bar{x}, \frac{\sigma}{\sqrt{n}}\right),$$

y entonces, la estimación de Bayes es:

$$\hat{\theta} = E[\theta/\mathbf{x}] = \bar{x}.$$

El estimador $\hat{\theta} = \bar{X}$ es el mismo que el obtenido por el método de los momentos y el de máxima verosimilitud. Con el enfoque clásico, la distribución de este estimador cumple lo siguiente, siendo θ fijo el verdadero valor del parámetro:

$$\bar{X} - \theta \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right).$$

En el enfoque bayesiano se considera una muestra fija \mathbf{x} , con media \bar{x} , y θ es la variable aleatoria, cuya distribución posterior cumple:

$$\pi(\theta - \bar{x}/\mathbf{x}) \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right).$$

Tema 4

Estimación puntual II. Métodos para evaluar estimadores: error cuadrático medio, estimadores insesgados óptimos, suficiencia e insesgadez, optimalidad de la función de pérdida.

Los contenidos de este tema corresponden a un curso básico de Estadística Matemática, y pueden ser consultados y completados con los libros de la bibliografía por ejemplo. El libro de Rohatgi y Ehsanes es el más completo.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

4.1 Introducción

Suponemos que la distribución teórica sobre la que queremos obtener información pertenece a una familia paramétrica $\{F_\theta : \theta \in \Theta\}$. Consideramos una m.a.s. X_1, \dots, X_n , para la que observamos los datos x_1, \dots, x_n (realización muestral). Para obtener información sobre θ consideramos estimadores puntuales, con el propósito de utilizar sus estimaciones como aproximaciones para el valor de θ .

En el tema anterior hemos estudiado métodos para la obtención de estimadores, y en este tema estudiamos procedimientos para valorar las propiedades de los estimadores y para compararlos.

4.2 Estimadores consistentes e insesgados

La muestra tiende a dar información completa sobre la distribución teórica cuando crece su tamaño, según el teorema de Glivenko-Cantelli. En consonancia con ello, se estudiaba en el tema 1 que, cuando n crece, parámetros muestrales como momentos y cuantiles tienden a concentrar su probabilidad cada vez más cerca de los correspondientes parámetros poblacionales. Los unos sirven como estimadores de los otros (propiedad que el método de los momentos utiliza).

A la propiedad de convergencia mencionada se la denomina consistencia. En el tema 1 se obtenía convergencia casi segura, y la propiedad de consistencia puede ser considerada según distintos tipos de convergencia. Es habitual considerar convergencia en probabilidad, que asegura (es equivalente a) la convergencia en ley (o distribución), que quiere decir que, cuando n crece, el estimador tiende a concentrar la mayor parte de su probabilidad cada vez más cerca del parámetro poblacional que está estimando.

Entonces, para n suficientemente grande, con alta probabilidad, una estimación debería ser muy cercana al verdadero valor del parámetro.

Definición 68. Una sucesión de estimadores T_1, T_2, \dots , asociada a los sucesivos tamaños muestrales n , se denomina *consistente* para estimar una función $g(\theta)$ del parámetro θ si

$$T_n \xrightarrow{p} g(\theta) \quad \text{para todo } \theta \in \Theta. \quad (4.1)$$

En esta definición se utiliza convergencia en probabilidad a una distribución degenerada en $g(\theta)$.

Una sucesión consistente para $g(\theta) = \mu$ es la media muestral $\bar{X} = \bar{X}_n$ para sucesivos valores de n . Por la ley débil de los grandes números se tiene que $\bar{X} \xrightarrow{p} \mu$ sea cual sea el verdadero valor de μ , y entonces la sucesión de medias muestrales $\bar{X}_1, \bar{X}_2, \dots$ es consistente para la media poblacional μ . No se hará referencia explícita a la sucesión y se dirá simplemente que la media muestral es un estimador consistente para la media poblacional.

La consistencia de la media es válida para cualquier distribución teórica (con μ finita). Por ejemplo, para una distribución teórica $\gamma(p, a)$, con $\theta = (p, a)$ y $\mu = p/a$, se tiene que \bar{X} es consistente para $g(p, a) = p/a$.

Recuérdese que θ es fijo pero desconocido, y entonces la condición en (4.1) de que la convergencia sea para todo $\theta \in \Theta$ asegura la convergencia a $g(\theta)$ cualquiera que sea el verdadero valor de θ . Por ejemplo, si fuera $\mu = 2'1$, entonces $\bar{X} \xrightarrow{p} 2'1$, y si fuera $\mu = -5'7$, entonces $\bar{X} \xrightarrow{p} -5'7$.

La siguiente propiedad establece condiciones para la esperanza y la varianza del estimador que aseguran la consistencia (aunque no son necesarias): si

$$\lim_{n \rightarrow \infty} E_\theta[T_n] = g(\theta) \quad \text{y} \quad \lim_{n \rightarrow \infty} V_\theta[T_n] = 0 \quad \text{para todo } \theta \in \Theta, \quad (4.2)$$

entonces la sucesión T_1, T_2, \dots es consistente para $g(\theta)$.

Un modo alternativo para obtener la consistencia de $\hat{\mu} = \bar{X}$ consiste en tener en cuenta que $E[\bar{X}] = \mu$ y $V[\bar{X}] = \frac{\sigma^2}{n}$ y observar que se cumple 4.2.

Aunque en la práctica los estimadores que se utilizan suelen ser consistentes, puede ser útil en alguna ocasión un estimador no consistente.

Ejemplo 69. Supongamos que se desea estudiar la duración de un tipo de led que no envejece, de modo que su tendencia a fallar es constante en el tiempo, y no aumenta por envejecimiento del material. Se obtiene que la distribución del tiempo hasta el fallo debe ser exponencial $Exp(\theta)$. Consideramos una m.a.s. X_1, \dots, X_n correspondiente a la duración de n leds, conectados en funcionamiento continuo hasta el fallo. El emv de $\mu = 1/\theta = g(\theta)$, el tiempo esperado hasta el fallo para un led, es $\hat{\mu} = \bar{X}$, que es un estimador consistente.

Para obtener los datos, y una realización \bar{x} de \bar{X} , hay que esperar a que fallen los n leds.

El siguiente estimador no es consistente, pero su valor puede ser determinado en cuanto falla el primer led de los n , y puede servir para realizar una estimación provisional. Sea $T = nX_{(1)}$. Se obtiene la función de distribución

$$P \{T \leq t\} = 1 - e^{-\theta t} \text{ para } t > 0,$$

que es la función de distribución de una $Exp(\theta)$ (igual que la teórica, aunque es casual). Por tanto $T \sim Exp(\theta)$, con

$$E[T] = \frac{1}{\theta} = \mu \text{ y } V[T] = \frac{1}{\theta^2}.$$

La distribución de T tiene la particularidad de que la misma para todo n , con $E[T] = \mu$. Además no es degenerada, y entonces T no es consistente.

Todos los momentos y cuantiles muestrales son consistentes para los correspondientes momentos y cuantiles poblacionales.

La consistencia hace referencia al comportamiento cuando n tiende a infinito, pero no al comportamiento del estimador para un tamaño de muestra dado. Sin embargo, asegura que es posible conseguir un estimador con una distribución tan concentrada cerca de $g(\theta)$ como se quiera, tomando n suficientemente grande.

Definición 70. Se denomina *sesgo* del estimador T para estimar $g(\theta)$ a la cantidad

$$b_T(g(\theta)) = E_\theta[T] - g(\theta).$$

Se dice que T es *insesgado*, o *centrado*, para $g(\theta)$ si el sesgo es nulo, esto es, si $E_\theta[T] = g(\theta)$.

Consideraremos algunos estimadores que son insesgados y otros que no lo son, pero en este último caso el sesgo siempre tenderá a 0 cuando $n \rightarrow \infty$ (estimador asintóticamente insesgado).

Ejemplo 71. La media muestral y la cuasivarianza muestral son estimadores insesgados para la media y la varianza poblacional, puesto que

$$E[\bar{X}] = \mu \quad \text{y} \quad E[S^2] = \sigma^2.$$

El sesgo de la varianza muestral como estimador de la varianza poblacional es:

$$b_{s^2}(\sigma^2) = E[s^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2,$$

que tiende a 0 cuando $n \rightarrow \infty$. Por tanto, la varianza muestral es asintóticamente insesgada para la varianza poblacional.

Aunque S^2 es insesgado para σ^2 , S no es insesgado para σ en general. Si $\hat{\theta}$ es insesgado para θ , $g(\hat{\theta})$ es insesgado para $g(\theta)$ en general solo si g es una función lineal.

Habitualmente el estimador de máxima verosimilitud es consistente. En algunas ocasiones es insesgado y en otras no.

4.3 Error cuadrático medio como medida de la bondad de un estimador

Para medir la *bondad de un estimador* (cómo es de bueno) tendremos que medir de algún modo si los valores que toma se alejan mucho del verdadero valor del parámetro (o de la función del parámetro $g(\theta)$ que se considere). Puesto que el estimador es una v.a., y distintas realizaciones (a partir de distintas muestras que se obtuvieran) tomarían distintos valores, algunos más lejanos y otros más cercanos, debemos considerar un valor promedio (una esperanza).

Consideremos un estimador T de θ . Si obtenemos la estimación $T = t$, entonces la lejanía al verdadero valor se puede medir simplemente como la distancia $|\theta - t|$ entre la estimación y el valor que estamos estimando. Pero también podemos considerar otras nociones de distancia que pueden contemplar una penalización por el error cometido al tomar t como si fuera el verdadero valor de θ ; por ejemplo, los costes económicos consecuencia de la toma de esa decisión.

Esta penalización se expresa mediante la *función de pérdida* $L(\theta, t)$. Algunos ejemplos de función de pérdida son los siguientes.

$$\begin{aligned} L(\theta, t) &= |\theta - t| \quad \text{es el error absoluto de la estimación.} \\ L(\theta, t) &= (\theta - t)^2 \quad \text{es el error cuadrático de la estimación.} \\ L(\theta, t) &= \frac{|\theta - t|}{\theta} \quad \text{es el error relativo de la estimación.} \end{aligned}$$

O la función de pérdida con $L(\theta, t) = c$ si $|\theta - t| > \varepsilon$ y 0 en caso contrario, que penaliza con un coste c los errores mayores que ε .

A la pérdida esperada $E_{\theta}[L(\theta, T)]$, que es función de θ , se la denomina *función de riesgo*, y se la denota por $R_T(\theta)$.

Una función de pérdida muy utilizada es el error cuadrático. En este caso de pérdida cuadrática la función de riesgo es $R_T(\theta) = E_{\theta}[(\theta - T)^2]$, y se denomina *error cuadrático medio* de T como estimador de θ . Se suele denotar por $ECM_T(\theta)$ ó $ECM(\theta)$:

$$ECM(\theta) = E_{\theta}[(\theta - T)^2] .$$

El error cuadrático medio de T como estimador de $g(\theta)$ es

$$ECM(g(\theta)) = E_{\theta}[(T - g(\theta))^2] .$$

Se verifica que

$$ECM(g(\theta)) = V_{\theta}[T] + b_T(g(\theta))^2 ,$$

como comprobamos a continuación:

$$\begin{aligned} ECM(g(\theta)) &= E_{\theta}[(T - E_{\theta}[T] + E_{\theta}[T] - g(\theta))^2] \\ &= E_{\theta}[(T - E_{\theta}[T])^2] + E_{\theta}[(E_{\theta}[T] - g(\theta))^2] \\ &\quad + 2E_{\theta}[(T - E_{\theta}[T])(E_{\theta}[T] - g(\theta))] , \end{aligned}$$

y, puesto que

$$\begin{aligned} E_{\theta} [(T - E_{\theta} [T]) (E_{\theta} [T] - g(\theta))] &= (E_{\theta} [T] - g(\theta)) E_{\theta} [(T - E_{\theta} [T])] \\ &= (E_{\theta} [T] - g(\theta)) (E_{\theta} [T] - E_{\theta} [T]) = 0, \end{aligned}$$

se tiene que

$$ECM(g(\theta)) = V_{\theta} [T] + (E_{\theta} [T] - g(\theta))^2 + 0 = V_{\theta} [T] + b_T(g(\theta))^2.$$

De aquí se obtiene que, si T es insesgado, entonces su ECM coincide con su varian-za.

Para dos estimadores T_1 y T_2 en particular puede ocurrir que $R_{T_1}(\theta) < R_{T_2}(\theta)$ para algunos valores de θ , y $R_{T_1}(\theta) > R_{T_2}(\theta)$ para otros.

Si T_1 y T_2 son dos estimadores para $g(\theta)$, se dice que T_1 es *preferible* a T_2 si

$$\begin{aligned} R_{T_1}(\theta) &\leq R_{T_2}(\theta) \quad \text{para todo } \theta \in \Theta \quad \text{y} \\ R_{T_1}(\theta) &< R_{T_2}(\theta) \quad \text{para algún } \theta \in \Theta. \end{aligned}$$

Sería deseable encontrar un estimador que fuera preferible a todos los demás, pero en muchas ocasiones no existe. Puede ocurrir que $R_{T_1}(\theta) < R_{T_2}(\theta)$ para algunos θ y $R_{T_1}(\theta) > R_{T_2}(\theta)$ para otros, y en este caso ni es T_1 preferible a T_2 ni T_2 preferible a T_1 .

Sin embargo, si nos restringimos a estimadores que verifican ciertas condiciones puede ocurrir que sí exista uno preferible a todos los demás. En concreto, más adelante estudiaremos un método que permite, en algunas ocasiones, obtener un estimador centrado que es preferible a cualquier otro estimador centrado. Para ello, y en general en toda la inferencia paramétrica, el concepto de estadístico suficiente es fundamental.

Ejemplo 72. Consideremos una m.a.s. de tamaño n de una distribución $N(\mu, \sigma)$. Vamos a considerar la media y la mediana muestrales, \bar{X} y M , como estimadores de μ , y a comprobar que la media es preferible a la mediana, al menos para n suficientemente grande.

En el tema 1 se obtuvo $M \sim AN\left(\mu, \frac{\sigma\sqrt{\pi}}{\sqrt{2n}}\right)$ (asintóticamente normal). Se tiene que $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. La media es insesgada y la mediana asintóticamente insesgada. Para n suficientemente grande, el cociente de errores cuadráticos medios es:

$$\frac{ECM_{\bar{X}}(\mu)}{ECM_M(\mu)} \simeq \sqrt{\frac{2}{\pi}} \simeq 0.8 < 1 \quad \text{para todo } (\mu, \sigma) \in \Theta.$$

Sin embargo, la mediana tiene propiedades que pueden aconsejar su uso en algunas ocasiones. Si en la muestra normal hay datos anómalos, excesivamente grandes o pequeños, que pueden hacer sospechar de un error en los registros, hay que decidir si excluir o no esos datos por ser erróneos. La media \bar{X} es sensible a esta circunstancia,

pudiendo tomar valores muy diferentes dependiendo de que se eliminen esos datos extremos o no, y esa decisión tiene elementos subjetivos. Sin embargo la mediana es menos sensible. Hágase la prueba con la siguiente muestra de tamaño 4: 6.3, 5.2, 42.0, 6.5. El estimador M es más “robusto” en ese sentido.

En los siguientes ejemplos, con distribución teórica uniforme $U(0, \theta)$, se obtiene un estimador de la forma $cX_{(n)}$ que es insesgado y consistente para θ . Se obtiene que dentro de la clase de los estimadores de la forma $cX_{(n)}$, con $c = c(n) > 0$, hay uno que es preferible a todos los demás de esa forma, y que no es el insesgado. Y se comparan estos estimadores con el estimador por el método de los momentos.

Ejemplo 73. Consideremos una distribución teórica $U(0, \theta)$, con $\theta > 0$. Determinamos c tal que $T = cX_{(n)}$ es un estimador insesgado para θ . (Para ello hay que obtener la función de densidad de $X_{(n)}$, a partir de aquí su esperanza, y después hay que utilizar la linealidad de la esperanza para simplificar y obtener una ecuación para c).

Las funciones de distribución y densidad de la $U(0, \theta)$ son: $F(x) = x/\theta$ para $0 < x < \theta$ y $f(x) = 1/\theta$ si $0 < x < \theta$. Por tanto, las funciones de distribución y densidad de $Z = X_{(n)} = \max\{X_1, \dots, X_n\}$ son

$$G(z) = G_n(z) = [F(z)]^n = \left(\frac{z}{\theta}\right)^n \quad \text{si } 0 \leq z \leq \theta.$$

$$g(z) = n[F(z)]^{n-1} f(z) = n \left(\frac{z}{\theta}\right)^{n-1} \frac{1}{\theta} = \frac{nz^{n-1}}{\theta^n} \quad \text{si } 0 < z < \theta.$$

Se obtiene:

$$E[X_{(n)}] = \int_0^\infty zg(z)dz = \frac{n}{n+1}\theta,$$

y de aquí se obtiene que $ET = E[cX_{(n)}] = cE[X_{(n)}] = c\frac{n}{n+1}\theta$. Si $ET = \theta$, entonces debe ser $c\frac{n}{n+1} = 1$, y de aquí se obtiene que $c = \frac{n+1}{n}$. Por tanto,

$$T = \frac{n+1}{n}X_{(n)}$$

es centrado para θ .

Para $z < \theta$ se tiene que $\frac{z}{\theta} < 1$, y entonces $\lim_{n \rightarrow \infty} G_n(z) = \lim_{n \rightarrow \infty} \left(\frac{z}{\theta}\right)^n = 0$. Para $z = \theta$ se tiene que $G_n(z) = 1$, y entonces $\lim_{n \rightarrow \infty} G_n(z) = 1$. Por tanto, $\lim_{n \rightarrow \infty} G_n(z)$ es la función de distribución de una degenerada en θ , y entonces $X_{(n)} \xrightarrow{d} \theta$. Puesto que $\lim_{n \rightarrow \infty} \frac{n+1}{n} = 1$, se tiene que $T \xrightarrow{d} \theta$, y entonces $T \xrightarrow{p} \theta$. Por tanto, $T = \frac{n+1}{n}X_{(n)}$ es consistente para θ .

Ejemplo 74. Consideremos una distribución teórica $U(0, \theta)$, con $\theta > 0$. Existe una función de n , $c_0 = c_0(n) > 0$, tal que el estimador $T = c_0X_{(n)}$, de θ , es preferible a cualquier otro estimador de θ de la forma $cX_{(n)}$ con $c \neq c_0$. Determinamos c_0 .

Un estimador $c_0X_{(n)}$ que sea preferible a otro estimador $cX_{(n)}$ debe tener menor ECM para todo θ . A priori, no tiene porqué existir un $c_0X_{(n)}$ preferible a cualquier otro estimador $cX_{(n)}$ con $c \neq c_0$. Con los siguientes cálculos, comprobamos que sí existe y lo obtenemos.

En el ejercicio anterior se obtuvo $E[X_{(n)}] = \frac{n}{n+1}\theta$, y de un modo similar se obtiene $E[X_{(n)}^2] = \frac{n}{n+2}\theta^2$. El sesgo y la varianza de $cX_{(n)}$, con $c = c(n) > 0$ arbitrario, son:

$$b_{cX_{(n)}}(\theta) = E_{\theta}[cX_{(n)}] - \theta = c\frac{n}{n+1}\theta - \theta = \left(c\frac{n}{n+1} - 1\right)\theta \quad y$$

$$V[cX_{(n)}] = \left(E[X_{(n)}^2] - E[X_{(n)}]^2\right)c^2 = \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)\theta^2c^2.$$

Simplificando, se obtiene que el error cuadrático medio de $cX_{(n)}$ como estimador de θ es:

$$\begin{aligned} ECM_{cX_{(n)}}(\theta) &= V[cX_{(n)}] + \left(b_{cX_{(n)}}(\theta)\right)^2 \\ &= \left[\frac{n}{n+2}c^2 + 1 - 2\frac{n}{n+1}c\right]\theta^2 \\ &= h(c)\theta^2, \end{aligned}$$

con $h(c) = \frac{n}{n+2}c^2 - 2\frac{n}{n+1}c + 1$, que es un polinomio de segundo grado en c , con un único punto de mínimo, que es

$$c = \frac{n+2}{n+1}.$$

Entonces, tomando $c_0 = \frac{n+2}{n+1}$, se tiene que $h(c) > h(c_0)$ si $c \neq c_0$, y por tanto $h(c)\theta^2 > h(c_0)\theta^2$ para todo θ . Por tanto, si $c \neq c_0$, se tiene que $ECM_{cX_{(n)}}(\theta) > ECM_{c_0X_{(n)}}(\theta)$ para todo θ , y entonces

$$c_0X_{(n)} = \frac{n+2}{n+1}X_{(n)}$$

es preferible a $cX_{(n)}$ para cualquier $c \neq \frac{n+2}{n+1}$.

Ejemplo 75. Consideremos una distribución teórica $U(0, \theta)$, con $\theta > 0$. Consideremos los siguientes estimadores para $\mu = \frac{\theta}{2}$: $T_1 = \frac{c}{2}X_{(n)}$, con $c = \frac{n+2}{n+1}$, y $T_2 = \bar{X}$. ¿Es preferible alguno de ellos al otro?

Es sencillo comprobar que

$$ECM_{\frac{c}{2}X_{(n)}}(\mu) = ECM_{\frac{c}{2}X_{(n)}}(\theta/2) = \left(\frac{1}{2}\right)^2 ECM_{cX_{(n)}}(\theta).$$

Utilizando los cálculos del ejercicio anterior, obtenemos que el ECM del estimador $\frac{c}{2}X_{(n)} = \frac{1}{2}\frac{n+2}{n+1}X_{(n)}$ de $\mu = \theta/2$ es:

$$\begin{aligned} ECM_{\frac{c}{2}X_{(n)}}(\mu) &= \left(\frac{1}{2}\right)^2 ECM_{cX_{(n)}}(\theta) = \frac{1}{4}h\left(\frac{n+2}{n+1}\right)\theta^2 \\ &= \frac{1}{4}\left[\frac{n}{n+2}\left(\frac{n+2}{n+1}\right)^2 - 2\frac{n}{n+1}\left(\frac{n+2}{n+1}\right) + 1\right]\theta^2 \\ &= \frac{-n(n+2) + (n+1)^2}{4(n+1)^2}\theta^2 = \frac{1}{4(n+1)^2}\theta^2. \end{aligned}$$

Por otra parte, \bar{X} siempre es un estimador centrado de μ , con varianza

$$V[\bar{X}] = \frac{\sigma^2}{n} = \frac{\theta^2}{12n},$$

que coincide con el ECM en este caso insesgado. Comprobamos que $4(n+1)^2 > 12n$, y de aquí se obtiene que

$$ECM_{\frac{c}{2}X_{(n)}}(\mu) < ECM_{\bar{X}}(\mu)$$

para todo θ (y para todo n), y por tanto $\frac{c}{2}X_{(n)}$ es preferible a \bar{X} para estimar $\mu = \theta/2$. Se tiene que

$$4(n+1)^2 - 12n = 4n^2 - 4n + 4 = 4[n(n-1) + 1],$$

que es positivo para todo n , y de aquí se obtiene el resultado.

Además, se observa que $ECM_{\frac{c}{2}X_{(n)}}(\mu)$ converge a 0 de un modo más rápido del que lo hace $ECM_{\bar{X}}(\mu)$.

4.4 Estimador centrado de uniformemente mínima varianza

Un estimador T , centrado para $g(\theta)$ y con varianza finita, se dice que es un *estimador centrado de uniformemente mínima varianza (ECUMV)* para $g(\theta)$ si, para cualquier otro estimador T' centrado para $g(\theta)$ y con varianza finita, se tiene que

$$V_{\theta}[T] \leq V_{\theta}[T'] \text{ para todo } \theta \in \Theta.$$

El error cuadrático medio coincide con la varianza para un estimador centrado, y entonces un ECUMV es un estimador centrado preferible a cualquier otro estimador centrado.

Si existe el ECUMV, es único. Aunque no siempre existe, sí existe en muchos casos. El siguiente teorema es útil para su obtención.

Teorema 76. (de Rao-Blackwell) Sea S un estadístico suficiente para una familia paramétrica $\{F_{\theta} : \theta \in \Theta\}$ (por ejemplo, el ems). Para cualquier estimador T_1 centrado para $g(\theta)$ se tiene que

$$T_2 = E[T_1/S]$$

es un estimador de $g(\theta)$ que satisface las siguientes propiedades:

- En la expresión de T_2 no interviene θ .
- T_2 es centrado para $g(\theta)$.
- $V_{\theta}[T_2] \leq V_{\theta}[T_1]$ para todo $\theta \in \Theta$, con igualdad si y solo si $T_2 = T_1$.

Demostración. a) Puesto que la distribución de la muestra condicionada por el valor de S no depende de θ , entonces así ocurre con cualquier función de la muestra, y en particular con T_1 . Entonces, la esperanza condicionada basada en esa distribución condicionada, $E_{\theta}[T_1/S]$, no depende de θ . Por tanto, T_2 es un estimador.

b) $E_{\theta}[T_2] = E_{\theta}[E[T_1/S]] = E_{\theta}[T_1] = g(\theta)$.

c) $V_{\theta}[T_1] = V_{\theta}[E[T_1/S]] + E_{\theta}[V[T_1/S]] \geq V_{\theta}[E[T_1/S]] = V_{\theta}[T_2]$. Además,

$$\begin{aligned} E_{\theta}[V[T_1/S]] &= E_{\theta}[E[(T_1 - E[T_1/S])^2/S]] \\ &= E_{\theta}[E[(T_1 - T_2)^2/S]] = E_{\theta}[(T_1 - T_2)^2], \end{aligned}$$

que vale 0 si y solo si $T_2 = T_1$ con probabilidad 1. ■

En este apartado 4.4 estamos considerando solo estimadores centrados, y entonces el *ECM* coincide con la varianza. El teorema 76 proporciona un método para mejorar un estimador centrado T_1 , obteniéndose un estimador T_2 que es, o preferible o igual a T_1 .

La esperanza condicionada $T_2 = E [T_1/S]$ es función del estadístico *ms* S , y entonces el *ECUMV*, si existe, debe ser función de S : si un estimador insesgado T_1 no es función de S , entonces $T_2 = E [T_1/S]$ es distinto de T_1 (uno es función de S y el otro no), y preferible a T_1 (por el teorema 76), y por tanto T_1 no puede ser el *ECUMV*.

Obsérvese que si aplicamos el teorema ahora a T_2 en vez de a T_1 , obtenemos de nuevo T_2 , por ser T_2 función de S :

$$E [T_2/S] = E [T_2 \cdot 1/S] = T_2 E [1/S] = T_2 .$$

Frecuentemente existe una única función h del estadístico suficiente minimal S que es centrada para $g(\theta)$. En este caso, por el teorema 76, $h(S)$ debe ser el *ECUMV*, y además $h(S) = E [T_1/S]$ para cualquier estimador insesgado T_1 .

Teorema 77. (Lehmann-Scheffé) Si S es un estadístico suficiente y completo, y existe un estimador insesgado T de θ , entonces existe un único *ECUMV* para θ , y viene dado por

$$E [T/S]$$

En todos los ejemplos de este apartado 4.4, el estadístico *sm* es completo, y de aquí se obtiene que el *ECUMV* es el único estimador centrado función del estadístico *sm*. En algunos casos, esta función centrada se obtiene de un modo sencillo, y en otros hay que calcular $E [T/S]$ directamente.

Ejemplo 78. Para las siguientes familias paramétricas y funciones g del parámetro, el *ECUMV* para $g(\theta)$ es la única función centrada del estadístico *sm*, que llamamos *SM* (reservamos el símbolo S para la cuasidesviación típica muestral). Determinamos el *ECUMV* para $g(\theta)$. a) $B(1, \theta)$, $g(\theta) = \theta$, b) $\mathcal{P}(\theta)$, $g(\theta) = \theta$, c) $N(\mu, \sigma)$ ($\theta = (\mu, \sigma)$), $g(\mu, \sigma) = \mu$, d) $N(\mu, \sigma)$ ($\theta = (\mu, \sigma)$), $g(\mu, \sigma) = \sigma^2$.

a) Se tiene que $SM = \bar{X}$ es *sm* (y función del *sm* por tanto) y es insesgado para $\mu = \theta$. Por tanto, \bar{X} es el *ECUMV* para θ .

b) Se tiene que $SM = \bar{X}$ es *sm* (y función del *sm* por tanto) y es insesgado para $\mu = \theta$. Por tanto, \bar{X} es el *ECUMV* para θ .

c) Se tiene que $SM = (\bar{X}, S^2)$ es *sm* y \bar{X} es función de SM . Además, \bar{X} es centrado para μ , y por tanto \bar{X} es el *ECUMV* para μ .

d) Se tiene que $SM = (\bar{X}, S^2)$ es *sm* y S^2 es función de SM . Además, S^2 es centrado para σ^2 , y por tanto S^2 es el *ECUMV* para σ^2 .

Ejemplo 79. Consideremos una m.a.s. de una distribución teórica $B(1, \theta)$. Se tiene que el *ECUMV* para $g(\theta) = \theta(1 - \theta) = \sigma^2$ es la única función centrada del estadístico *sm* SM . Consideramos el estimador centrado T que vale 1 si $X_1 = 1$ y $X_2 = 0$, y 0 en caso contrario, y calculamos $E [T/SM]$.

Se tiene que $SM = \bar{X}$. No parece sencillo encontrar de un modo directo una función centrada del estadístico $sm \bar{X}$ (si no se ha trabajado con varianzas muestrales para variables de Bernoulli). Utilizamos el teorema 77 para encontrar la función centrada, que es única y es por tanto el ECUMV. Para ello, podemos partir de cualquier estimador centrado y, de hecho, cuanto más simple sea más sencillos son los cálculos. Partimos de un estimador que no es bueno pero que es simple.

Sea T_1 con $T_1 = 1$ si $X_1 = 1, X_2 = 0$ y 0 en caso contrario. Se tiene que T_1 es centrado, puesto que

$$\begin{aligned} E_{\theta} [T_1] &= 1 \cdot P \{T_1 = 1\} + 0 \cdot P \{T_1 = 0\} = 1 \cdot P \{X_1 = 1, X_2 = 0\} \\ &= P \{X_1 = 1\} P \{X_2 = 0\} = \theta (1 - \theta) = g(\theta) . \end{aligned}$$

Resulta ahora más cómodo considerar el estadístico sm en la forma $SM = \sum_{i=1}^n X_i$, que tiene distribución $B(n, \theta)$. Se tiene que

$$\begin{aligned} E [T_1 / SM = s] &= 1 \cdot P \{T_1 = 1 / SM = s\} + 0 \cdot P \{T_1 = 0 / SM = s\} \\ &= P \{X_1 = 1, X_2 = 0 / SM = s\} \\ &= \frac{P \{X_1 = 1, X_2 = 0, \sum_{i=1}^n X_i = s\}}{P \{SM = s\}} \\ &= \frac{P \{X_1 = 1, X_2 = 0, \sum_{i=3}^n X_i = s - 1\}}{P \{SM = s\}} \\ &= \frac{P \{X_1 = 1\} P \{X_2 = 0\} P \{\sum_{i=3}^n X_i = s - 1\}}{P \{SM = s\}} \\ &= \frac{P \{B(1, \theta) = 1\} P \{B(1, \theta) = 0\} P \{B(n - 2, \theta) = s - 1\}}{P \{B(n, \theta) = s\}} \\ &= \frac{\theta (1 - \theta) \binom{n-2}{s-1} \theta^{s-1} (1 - \theta)^{n-s-1}}{\binom{n}{s} \theta^s (1 - \theta)^{n-s}} = \frac{\binom{n-2}{s-1}}{\binom{n}{s}} = \frac{s(n-s)}{n(n-1)} . \end{aligned}$$

Por tanto, el ECUMV para $g(\theta) = \theta(1 - \theta) = \sigma^2$ es

$$E [T_1 / SM] = \frac{SM(n - SM)}{n(n - 1)} = \frac{n}{n - 1} \bar{X} (1 - \bar{X}) .$$

Ejemplo 80. El ECUMV para la función del parámetro $g(\theta) = P_{\theta} \{X > 0\} = 1 - e^{-\theta}$ de una distribución teórica $\mathcal{P}(\theta)$ es único.

No es sencillo encontrar de un modo directo una función centrada del estadístico sm $SM = \sum_{i=1}^n X_i$. Utilizamos el teorema 77 para encontrar la función centrada, que es única (es el ECUMV), por ser el ems completo. Partimos de un mal estimador pero que es simple.

Sea T_1 con $T_1 = 1$ si $X_1 > 0$ y $T_1 = 0$ si $X_1 = 0$. Se tiene que T_1 es centrado, puesto que

$$E_{\theta} T_1 = 1 \cdot P \{T_1 = 1\} = P \{X_1 > 0\} = 1 - e^{-\theta} = g(\theta) .$$

Se tiene que (la distribución de Poisson es reproductiva)

$$\begin{aligned}
 E[T_1/SM = s] &= P\{T_1 = 1/SM = s\} \\
 &= 1 - P\{X_1 = 0/SM = s\} \\
 &= 1 - \frac{P\{X_1 = 0, \sum_{i=1}^n X_i = s\}}{P\{SM = s\}} \\
 &= 1 - \frac{P\{X_1 = 0, \sum_{i=2}^n X_i = s\}}{P\{SM = s\}} \\
 &= 1 - \frac{P\{X_1 = 0\} P\{\sum_{i=2}^n X_i = s\}}{P\{SM = s\}} \\
 &= 1 - \frac{P\{\mathcal{P}(\theta) = 0\} P\{\mathcal{P}((n-1)\theta) = s\}}{P\{\mathcal{P}(n\theta) = s\}} \\
 &= 1 - \frac{e^{-\theta} e^{-(n-1)\theta} \frac{[(n-1)\theta]^s}{s!}}{e^{-n\theta} \frac{[n\theta]^s}{s!}} = 1 - \frac{(n-1)^s}{n^s}.
 \end{aligned}$$

Por tanto, el ECUMV para $g(\theta) = P\{X > 0\}$ es

$$E[T_1/SM] = 1 - \frac{(n-1)^{SM}}{n^{SM}} = 1 - \left(1 - \frac{1}{n}\right)^{n\bar{X}}. \quad (4.3)$$

Puesto que el estimador de máxima verosimilitud de θ es $\hat{\theta} = \bar{X}$, se tiene que el emv para la función del parámetro $g(\theta) = 1 - e^{-\theta}$ es $1 - e^{-\hat{\theta}} = 1 - e^{-\bar{X}}$. Para n grande se tiene que este estimador es muy cercano al ECUMV, en (4.3), puesto que $\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1}$, y entonces

$$E[T_1/SM] = 1 - \left[\left(1 - \frac{1}{n}\right)^n\right]^{\bar{X}} \simeq 1 - e^{-\bar{X}} = g(\hat{\theta}).$$

Si T_1 es el ECUMV para $g_1(\theta)$ y T_2 es el ECUMV para $g_2(\theta)$, se tiene que $T = aT_1 + bT_2$ es el ECUMV para $g(\theta) = ag_1(\theta) + bg_2(\theta)$.

Ejemplo 81. Consideremos una m.a.s. de un distribución teórica $B(1, \theta)$. Obtenemos el ECUMV para $g(\theta) = \theta^2$.

Se tiene que $g(\theta) = g_1(\theta) - g_2(\theta)$, con $g_1(\theta) = \theta$ y $g_2(\theta) = \theta(1 - \theta)$. En ejemplos anteriores obtuvimos el ECUMV para $g_1(\theta)$, que es $T_1 = \bar{X}$, y el ECUMV para $g_2(\theta)$, que es $T_2 = \frac{n}{n-1} \bar{X} (1 - \bar{X})$. Entonces, el ECUMV para $g(\theta) = g_1(\theta) - g_2(\theta) = \theta^2$ es

$$T_1 - T_2 = \bar{X} - \frac{n}{n-1} \bar{X} (1 - \bar{X}) = \frac{1}{n-1} (n\bar{X}^2 - \bar{X}) = \frac{n\bar{X} - 1}{n-1} \bar{X}.$$

4.5 Cota de Frechet-Cramer-Rao para la varianza de un estimador

El valor más pequeño para la varianza de un estimador centrado viene dado por la varianza del ECUMV (si existe). Aun sin conocer el ECUMV es posible en muchos

casos obtener funciones de θ que acotan inferiormente la varianza de los estimadores centrados.

En algunas ocasiones, estas acotaciones permiten obtener el ECUMV: dado un estimador centrado, si su varianza coincide con la cota entonces es el ECUMV.

Una de estas acotaciones es la desigualdad de Frechet-Cramer-Rao, que se presenta a continuación.

Sea T centrado para $g(\theta)$. Bajo ciertas condiciones de regularidad de la distribución teórica, una de ellas que su soporte no dependa de θ y el resto poco restrictivas, se verifica que

$$V_{\theta}[T] \geq \frac{g'(\theta)^2}{n \cdot i(\theta)} \text{ para todo } \theta \in \Theta. \quad (4.4)$$

Un estimador centrado para el que se dé la igualdad en (4.4) se denomina *estimador eficiente*.

Si un estimador es eficiente entonces es el ECUMV, y esto sirve como método de cálculo del ECUMV en algunas ocasiones. La relación opuesta no se verifica y de este modo el ECUMV puede no ser eficiente.

Si el ECUMV no es eficiente, entonces no existe ningún estimador eficiente, puesto que el ECUMV es preferible a cualquier otro estimador centrado.

Denominamos *eficiencia* de T , centrado para $g(\theta)$, al cociente entre su cota y su varianza:

$$e_T(\theta) = \frac{g'(\theta)^2 / (n \cdot i(\theta))}{V_{\theta}[T]}.$$

Se verifica que $e_T(\theta) \leq 1$. El estimador T es eficiente cuando $e_T(\theta) = 1$ para todo θ . Se verifica que un estimador solo puede ser eficiente cuando la distribución teórica es una familia de tipo exponencial.

4.6 Perdida final esperada como medida de la bondad de un estimador bayesiano

Consideremos un estimador puntual T de θ , y sea t la estimación puntual con la muestra \mathbf{x} , $t = T(\mathbf{x})$. En un contexto bayesiano se considera también una función de perdida $L(t, \theta)$. Pero, a diferencia del enfoque clásico, ahora la variable aleatoria es θ , y el valor t es fijo. La función de riesgo es la esperanza de $L(t, \theta)$. En lugar de considerar la esperanza para la distribución de T en el muestreo, como en el enfoque clásico, con θ fijo, se considera ahora la esperanza de $L(t, \theta)$ para la distribución posterior de θ (a posteriori, final) $\pi(\theta/\mathbf{x})$. A esta función de riesgo bayesiana se la denomina también perdida final esperada.

Definición 82. La perdida final esperada PFE es la esperanza de $L(t, \theta)$ para la distribución a posteriori:

$$PFE(T) = E[L(t, \theta) / \mathbf{x}] = \int_{\theta \in \Theta} L(t, \theta) \pi(\theta/\mathbf{x}) d\theta.$$

Sería conveniente utilizar estimadores que minimicen la PFE . Considerando la función de pérdida cuadrática $L(t, \theta) = (\theta - t)^2$, se obtiene que el estimador con pérdida final esperada mínima es la esperanza de $\pi(\theta/\mathbf{x})$: $T(\mathbf{x}) = E[\theta/\mathbf{x}]$. Es sencillo comprobarlo aplicando el siguiente resultado a esta distribución condicionada. Para una variable aleatoria Z , desarrollando el cuadrado $(Z - t)^2 = (Z - EZ + EZ - t)^2$, tomando esperanzas y simplificando, se obtiene

$$E[(Z - t)^2] = V[Z] + (EZ - t)^2,$$

que obviamente se minimiza con $t = EZ$, y el valor mínimo de esta esperanza es entonces $E[(Z - EZ)^2] = V[Z]$. Aplicando este resultado a la distribución condicionada $\pi(\theta/\mathbf{x})$, se obtiene, con pérdida cuadrática, que

$$PFE(T) = E[(\theta - t)^2/\mathbf{x}]$$

se minimiza tomando $t = E[\theta/\mathbf{x}]$. El estimador obtenido es entonces

$$T(\mathbf{x}) = E[\theta/\mathbf{x}],$$

que es el estimador Bayes que estudiábamos en el tema 3. El valor mínimo de esta esperanza es:

$$PFE(T) = E[(\theta - E[\theta/\mathbf{x}])^2/\mathbf{x}] = V[\theta/\mathbf{x}],$$

que es la varianza a posteriori.

Considerando función de pérdida absoluta $L(t, \theta) = |\theta - t|$, se obtiene que el estimador con pérdida final esperada mínima es la mediana de $\pi(\theta/\mathbf{x})$.

Ejemplo 83. La proporción de enfermos θ en una determinada colectividad puede ser modelizada mediante una variable aleatoria con distribución beta. De estudios previos, puede suponerse que $\theta \sim \text{Beta}(5, 10)$ (con $\mu_0 = 1/3$ y $\sigma_0 = 0.118$). En una muestra de 100 individuos hay 20 enfermos. En el tema 3 se obtuvo para este caso $\pi(\theta/\mathbf{x}) \sim \text{Beta}(25, 90)$. La estimación de Bayes es $\hat{\theta} = E[\theta/\mathbf{x}] = \frac{5}{23}$. La pérdida final esperada es:

$$PFE = V[\theta/\mathbf{x}] = V[\text{Beta}(25, 90)] = 0.00146.$$

Esta varianza a posteriori $\sigma_1^2 = V[\theta/\mathbf{x}]$ es menor que la varianza a priori $\sigma_0^2 = 0.0139$. La aportación de información muestral (datos) hace que se reduzca la incertidumbre sobre θ , disminuyendo la varianza de su distribución.

Una vez que se han obtenido los datos muestrales y se ha determinado la distribución posterior, la información de la que se dispone sobre θ está en esta distribución posterior. De este modo, la distribución posterior pasa a ser la nueva distribución prior, que ha sido actualizada con los datos muestrales.

Ejemplo 84. Consideremos una muestra $\mathbf{x} = (x_1, \dots, x_n)$ de una distribución teórica $B(1, \theta)$, y una distribución a priori $\theta \sim \text{Beta}(\alpha, \beta)$. En el tema 3 se obtuvo $\pi(\theta/\mathbf{x}) \sim \text{Beta}(\alpha + t, \beta + n - t)$, con $t = \sum_{i=1}^n x_i$. El estimador de Bayes es $\hat{\theta} = \frac{\alpha + t}{\alpha + \beta + n}$. La pérdida final esperada es:

$$PFE = V[\theta/\mathbf{x}] = \frac{(\alpha + t)(\beta + n - t)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}.$$

Para $n = 0$ la *PFE* (varianza de θ a posteriori) es la misma varianza a priori, puesto que la distribución posterior coincide con la prior. Para n grande se obtiene

$$\mu_1 \simeq \bar{x} \quad \text{y} \quad \sigma_1^2 \simeq \bar{x}(1 - \bar{x})/n,$$

con $\mu_1 = E[\theta/\mathbf{x}] = \hat{\theta}$ y $\sigma_1^2 = V[\theta/\mathbf{x}] = PFE$. Entonces, la *PFE* tiende a 0, coherentemente con el hecho de que la muestra tiende a dar información completa sobre el parámetro cuando n crece.

Ejemplo 85. Consideremos una muestra $\mathbf{x} = (x_1, \dots, x_n)$ de una distribución teórica $N(\theta, \sigma)$ con σ conocida, y una distribución a priori $\theta \sim N(\mu_0, \sigma_0)$. En el tema 3 se obtuvo $(\theta/\mathbf{x}) \sim N(\mu_1, \sigma_1)$, con

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\bar{x} \quad \text{y} \quad \sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

El estimador de Bayes es $\hat{\theta} = \mu_1$. La pérdida final esperada es:

$$PFE = V[\theta/\mathbf{x}] = \sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

Para n grande se tiene que

$$\begin{aligned} \mu_1 &= \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{\sigma_0^2}{\sigma^2/n + \sigma_0^2}\bar{x} \simeq \bar{x}, \\ \sigma_1^2 &= \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} = \frac{\sigma^2/n}{\frac{\sigma^2}{n\sigma_0^2} + 1} \simeq \frac{\sigma^2}{n}, \end{aligned}$$

de modo que la aportación de la distribución prior (con sus parámetros μ_0 y σ_0) sobre la distribución posterior tiende a desaparecer, diluida por la información aportada por una muestra cada vez mas grande.

Tema 5

Tests de hipótesis. Introducción. Métodos para encontrar tests: tests de ratios de verosimilitudes, tests bayesianos, tests unión-intersección e intersección-unión. Métodos para evaluar tests: probabilidades de error y función de potencia, tests más potentes, tamaños de tests unión-intersección e intersección-unión, optimalidad de la función de pérdida.

Los contenidos de este tema corresponden a un curso básico de Estadística Matemática, y pueden ser consultados y completados con los libros de la bibliografía por ejemplo. El libro de Rohatgi y Ehsanes es el más completo.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

5.1 Introducción

Un contraste de hipótesis, o test de hipótesis, es un procedimiento de inferencia estadística que sirve para valorar la credibilidad de una determinada hipótesis, o suposición, sobre la distribución teórica. Se puede utilizar por ejemplo para decidir si aceptamos que un determinado tratamiento es útil para la curación de una enfermedad, o que un dado está equilibrado.

La hipótesis se formula en términos de una clasificación (partición) de las posibles distribuciones teóricas. Por un lado aquellas en la que la hipótesis es cierta, y por otro aquellas en la que no. Por ejemplo, si la distribución teórica es de Bernoulli, la hipótesis para contrastar puede ser "el parámetro θ vale $1/2$ ".

El procedimiento consiste en una partición del espacio muestral en un conjunto de rechazo y otro de aceptación, de acuerdo con ciertos requerimientos. Si la muestra obtenida pertenece al primero se rechaza la hipótesis formulada, y si pertenece al segundo no se rechaza. Pero como veremos, este no rechazo no es exactamente una aceptación sin más. La decisión estará sujeta a un posible error, pero se puede medir la probabilidad de error.

Por ejemplo, se puede considerar que un determinado medicamento cura si un determinado parámetro toma ciertos valores; se tomará una muestra de n pacientes y, basándose en los datos obtenidos, habrá que decidir si el parámetro toma los valores que indican que el medicamento cura o no. O por ejemplo, ¿podemos aceptar que cierta moneda está equilibrada, esto es, que la probabilidad de obtener cara es $\theta = 1/2$?; se

lanza $n = 100$ veces y se obtienen 64 caras; si se estudia la distribución del número de caras, se comprobará que obtener observaciones tan lejanas como $|64 - 50| = 14$ o más del número esperado de caras en una moneda equilibrada, 50, es poco probable; el test habitual que se utiliza para este estudio, coherentemente con ello, rechaza con bastante rotundidad (según el “ p -valor” del test, pequeño) que la moneda esté equilibrada.

Esta formulación de las hipótesis y el procedimiento de contraste no presuponen un contexto paramétrico, aunque en el ejemplo de la Bernoulli sea así. Presentaremos las nociones generales y nos ocuparemos del caso paramétrico. En el caso paramétrico se considerará un subconjunto Θ_0 del espacio paramétrico Θ , y se estudiarán procedimientos para decidir sobre si el verdadero valor de θ está en Θ_0 o no.

5.2 Conceptos básicos

5.2.1 Hipótesis nula y alternativa, simple o compuesta

La hipótesis que queremos contrastar se denomina *hipótesis nula*, y se denota por H_0 . La negación de la hipótesis nula se denomina *hipótesis alternativa*, y se denota por H_1 . Diremos que estamos contrastando H_0 frente a H_1 , o simplemente, que estamos contrastando H_0 .

Una hipótesis que especifica con unicidad la distribución teórica se dice que es una *hipótesis simple*. En caso contrario se dice que es una *hipótesis compuesta*.

En el caso paramétrico, la hipótesis nula queda especificada mediante un subconjunto $\Theta_0 \subset \Theta$ adecuado, y se expresa cómo “ $\theta \in \Theta_0$ ”. Sea Θ_1 el complementario de Θ_0 , esto es, $\Theta_0 \cup \Theta_1 = \Theta$ y $\Theta_0 \cap \Theta_1 = \emptyset$. La hipótesis alternativa se expresa como “ $\theta \notin \Theta_0$ ” o “ $\theta \in \Theta_1$ ”. Diremos que estamos contrastando

$$H_0 : \theta \in \Theta_0 \quad \text{frente a} \quad H_1 : \theta \in \Theta_1,$$

o simplemente, que estamos contrastando $H_0 : \theta \in \Theta_0$.

En el ejemplo de la moneda del apartado anterior (con distribución teórica de Bernoulli), es $\Theta_0 = \{1/2\}$, y H_0 es una hipótesis simple. El contraste se expresa del siguiente modo: $H_0 : \theta = 1/2$ frente a $H_1 : \theta \neq 1/2$, o simplemente $H_0 : \theta = 1/2$.

5.2.2 Definición de contraste de hipótesis. Región crítica y región de aceptación.

Un *contraste* o *test de hipótesis* es una partición del espacio muestral en una *región crítica* C , en la cual se rechaza H_0 aceptándose H_1 , y una *región de aceptación* C^c , en la cual no se rechaza H_0 (se acepta). Una vez observada la muestra x_1, \dots, x_n , la decisión de rechazar o no H_0 se hace según la pertenencia o no de la muestra a C . Entonces, un test es un procedimiento de decisión para el rechazo o no rechazo de la hipótesis nula.

Según se ha definido, cualquier subconjunto del espacio muestral es una región crítica, y esta define un test de hipótesis. Sin embargo, unos serán mejores que otros. El criterio para la elección de una determinada región crítica C se detalla un poco más adelante,

y se basa en limitar la probabilidad de error. La teoría de los contrastes de hipótesis consiste en el estudio de métodos adecuados para la elección de C .

La región crítica C se expresa en términos de un estadístico, denominado *estadístico del contraste*. Por ejemplo, si contrastamos $H_0 : \theta = 1/2$ para una $B(1, \theta)$, es razonable rechazar H_0 si \bar{x} se aleja de $1/2$, puesto que \bar{x} es un estimador de θ . Entonces, rechazando simétricamente respecto de $1/2$, se puede considerar una región crítica de la forma

$$C = \left\{ \left| \bar{x} - \frac{1}{2} \right| > c \right\} = \left\{ (x_1, \dots, x_n) \in \Omega : \left| \bar{x} - \frac{1}{2} \right| > c \right\},$$

y habrá que elegir un valor adecuado para c . El estadístico del contraste es \bar{x} . La elección de c se hace en función de la probabilidad de error que se considere.

En ocasiones abreviamos “región crítica” como rc .

5.2.3 Errores de tipo I y II

Tenemos dos disyuntivas. Por una parte la verdad o falsedad de la hipótesis nula, cuestión que pretendemos estudiar, y por otra la aceptación o rechazo de esta hipótesis nula. Estaremos cometiendo un error si H_0 es cierta y la rechazamos, denominado *error de tipo I*, o si H_0 es falsa y la aceptamos, denominado *error de tipo II*. Expresamos estas posibilidades y esta terminología en forma de tabla:

	H_0 es cierta	H_0 es falsa
Rechazar H_0	<i>error de tipo I</i>	decisión correcta
Aceptar H_0	decisión correcta	<i>error de tipo II</i>

La decisión de rechazar o aceptar H_0 se hace según la pertenencia o no de la muestra a C . Puesto que la muestra es aleatoria, unas veces se cometerá error y otras no, y esto ocurre conforme a la distribución de la muestra. Lo ideal sería poder minimizar la probabilidad de ambos tipos de error, pero habitualmente la reducción de la probabilidad de uno se hace a costa del aumento en la probabilidad del otro.

Utilizamos la abreviatura “pet I” para la probabilidad de error de tipo I, y “pet II” para la probabilidad de error de tipo II.

Supongamos que H_0 es una hipótesis simple. Sea $P_0 = P_{H_0}$ la probabilidad bajo H_0 (esto es, suponiendo H_0 cierta). Puesto que el rechazo de H_0 ocurre cuando la muestra observada pertenece a la región crítica, entonces la pet I es la probabilidad $P_0(C)$.

5.2.4 Nivel de significación

El criterio comunmente usado para la elección de C consiste en fijar una cota α para la pet I, denominada *nivel de significación*, y elegir, a ser posible, entre los test con pet I menor o igual que α , el test con menor pet II. Con esto se consigue controlar el error de tipo I, pero sin embargo solo en algunos casos puede ser elegido el test con menor pet II.

Si tomamos α pequeño, por ejemplo 0.1, 0.05 ó 0.01, y la muestra observada pertenece a C y lleva por tanto al rechazo de H_0 , entonces podemos estar razonablemente seguros de que H_0 es falsa. Pudiera ser que H_0 fuera cierta, pero en este caso estaríamos cometiendo un error de tipo I; puesto que la *pet I* es pequeña ($\leq \alpha$), estaría ocurriendo un suceso con probabilidad pequeña, y es más razonable concluir que H_0 no es cierta. Por tanto, con este criterio, si la muestra observada lleva al rechazo de H_0 podemos estar razonablemente seguros de que la decisión es correcta. Sin embargo, la aceptación de H_0 solo indica que los datos no aportan evidencia en contra de H_0 .

Por ejemplo, si contrastamos $H_0 : \theta = 1/2$ para una $B(1, \theta)$, y consideramos la región crítica $C = \{|\bar{x} - \frac{1}{2}| > c\}$, el valor de c se elige en función del nivel de significación. En este caso, la hipótesis nula es simple, y se tiene que la *pet I* vale $P_0(C)$. Eligiendo c tal que $P_0(C) = \alpha$, el contraste con región crítica C tiene nivel de significación α . Obtenemos el valor de c mediante aproximación normal.

Por el teorema central del límite se tiene que $\frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)}}\sqrt{n} \sim AN(0, 1)$ (asintóticamente normal $N(0, 1)$), siendo θ el verdadero valor del parámetro. Si H_0 es cierta se tiene que $\theta = 1/2$, y entonces

$$\frac{\bar{X} - \frac{1}{2}}{\sqrt{\frac{1}{2}(1 - \frac{1}{2})}}\sqrt{n} \stackrel{H_0}{\sim} AN(0, 1), \quad (5.1)$$

denotando con " $\stackrel{H_0}{\sim}$ " la distribución bajo H_0 , suponiendo H_0 cierta, esto es, con $\theta = 1/2$. Se tiene que

$$\begin{aligned} \alpha &= P_0(C) = P_0\left\{\left|\bar{X} - \frac{1}{2}\right| > c\right\} = P_0\left\{\left|\frac{\bar{X} - \frac{1}{2}}{\sqrt{\frac{1}{2}(1 - \frac{1}{2})}}\sqrt{n}\right| > \frac{c\sqrt{n}}{\sqrt{\frac{1}{2}(1 - \frac{1}{2})}}\right\} \\ &\simeq P\{|N(0, 1)| > 2c\sqrt{n}\} = 1 - P\{|N(0, 1)| < 2c\sqrt{n}\} = \alpha, \end{aligned}$$

y entonces $P\{|N(0, 1)| < 2c\sqrt{n}\} = 1 - \alpha$. Por tanto debe ser $2c\sqrt{n} = z_{\alpha/2}$. De aquí se obtiene $c = \frac{1}{2\sqrt{n}}z_{\alpha/2}$, y por tanto

$$C = \left\{\left|\bar{x} - \frac{1}{2}\right| > \frac{1}{2\sqrt{n}}z_{\alpha/2}\right\} \quad (5.2)$$

es una región crítica para el contraste con nivel de significación (aproximado) α .

Ejemplo 86. Consideremos una distribución teórica $B(1, \theta)$. Consideremos una región crítica de la forma $C = \{|\bar{x} - \frac{1}{2}| > c\}$ para contrastar $H_0 : \theta = 1/2$. En una muestra de tamaño $n = 100$ se obtuvieron 61 éxitos. ¿Se debe rechazar H_0 , con nivel de significación $\alpha = 0.05$? ¿y con $\alpha = 0.01$?

Según se ha obtenido en (5.2), la región crítica concreta para un nivel de significación α (aproximado) es $C = \left\{|\bar{x} - \frac{1}{2}| > \frac{1}{2\sqrt{n}}z_{\alpha/2}\right\}$.

Con $\alpha = 0.05$ se obtiene $C = \left\{|\bar{x} - \frac{1}{2}| > 0.098\right\}$. Puesto que $|0.61 - \frac{1}{2}| = 0.11$ y $0.11 > 0.098$, la muestra está en la región crítica, y entonces rechazamos H_0 .

Con $\alpha = 0.01$ se obtiene $C = \left\{|\bar{x} - \frac{1}{2}| > 0.129\right\}$. Puesto que $|0.61 - \frac{1}{2}| = 0.11$ y $0.11 \not> 0.129$, la muestra no está en la región crítica sino que está en la región de aceptación, y entonces aceptamos H_0 .

5.2.5 Nivel crítico o p -valor

Habitualmente se consideran contrastes con una región crítica que depende de α , como en (5.2). Una vez fijado el nivel de significación α , se obtiene la región crítica concreta y, una vez observada la muestra x_1, \dots, x_n , se toma la decisión de aceptar o rechazar H_0 según la muestra pertenezca o no a la región crítica.

Hay un modo de manejar todos los niveles de significación α simultáneamente. Ésto se hace considerando un número, el nivel crítico, que es función de la muestra, de modo que con α menor que el nivel crítico se acepta H_0 y con α mayor o igual que el nivel crítico se rechaza H_0 .

El *nivel crítico* o *p-valor* del test, $\alpha(x_1, \dots, x_n)$, es el nivel de significación más pequeño para el que la muestra observada x_1, \dots, x_n lleva a rechazar la hipótesis nula.

Con un nivel de significación $\alpha = 0$ se acepta H_0 siempre, según crece α la región crítica se hace más grande, y con $\alpha = 1$ se rechaza H_0 siempre. Dada la muestra, el nivel crítico es el valor de α para el cual se pasa de aceptar a rechazar H_0 .

El p -valor se calcula del siguiente modo, una vez elegido un estadístico para el contraste, la forma de la región crítica, y obtenida la muestra. Se considera la región crítica concreta para la cual la observación está en el borde. Su nivel de significación es el p -valor.

Por ejemplo, si la región crítica es de la forma $C = \{T > c\}$, y después de obtener la muestra se calcula T y vale $t = 6.1$, entonces el p -valor es

$$\alpha(x_1, \dots, x_n) = P_0 \{T > 6.1\} ,$$

que se calcula a partir de la distribución nula (esto es, bajo H_0) de T .

Ejemplo 87. Determinamos el p -valor del test en la situación del ejemplo 86.

La región crítica concreta para la cual el valor observado $\bar{x} = \frac{61}{100} = 0.61$ está en el borde es $C = \{|\bar{x} - \frac{1}{2}| > 0.11\}$, puesto que $|0.61 - \frac{1}{2}| = 0.11$. El p -valor es el nivel de significación correspondiente a esta región crítica, que es la probabilidad de error de tipo I: puesto que $2(\bar{X} - \frac{1}{2})\sqrt{n} \stackrel{H_0}{\sim} AN(0, 1)$ (véase (5.1)), se tiene que

$$\begin{aligned} \alpha(x_1, \dots, x_n) &= P_0(C) = P_0\left\{|\bar{X} - \frac{1}{2}| > 0.11\right\} \\ &= P_0\left\{|2(\bar{X} - \frac{1}{2})\sqrt{n}| > 2 \cdot 0.11\sqrt{n}\right\} \\ &\simeq P\{|N(0, 1)| > 2.2\} = 2 \cdot 0.0139 = 0.0278 . \end{aligned}$$

(Otra manera de obtener el p -valor hubiera sido mediante la resolución de la ecuación $\frac{1}{2\sqrt{n}}z_{\alpha/2} = 0.11$, puesto que $C = \left\{|\bar{x} - \frac{1}{2}| > \frac{1}{2\sqrt{n}}z_{\alpha/2}\right\}$ es una región crítica para el contraste con nivel de significación α ; el p -valor es el valor de α obtenido como solución). El cálculo del p -valor permite resolver el ejemplo 86 de un modo alternativo, sin tener que calcular las regiones críticas concretas para los casos $\alpha = 0.05$ y 0.01 : puesto que $0.05 > 0.0278$, se rechaza H_0 con nivel de significación $\alpha = 0.05$, y puesto que $0.01 < 0.0278$, se acepta H_0 con $\alpha = 0.01$.

Resolvemos dos ejemplos más con hipótesis nula simple.

Ejemplo 88. Se supone que la longitud del ala de los individuos adultos de una especie de mosca tiene distribución $N(\mu, \sigma)$ con $\sigma = 0.4$ (en mm). Se obtuvo una muestra de 49 individuos y se calculó la media de las longitudes del ala, con el resultado $\bar{x} = 7.2$. Contrastamos la hipótesis $H_0 : \mu = 7$ utilizando una región crítica de la forma $C = \{|\bar{x} - 7| > c\}$: obtenemos el p -valor y determinamos la decisión que hay que tomar con nivel de significación $\alpha = 0.01$.

Calculamos el p -valor. La región crítica concreta para la cual el valor observado $\bar{x} = 7.2$ está en el borde es $C = \{|\bar{x} - 7| > 0.2\}$, puesto que $|7.2 - 7| = 0.2$. Se tiene que $\frac{\bar{X}-7}{\sigma}\sqrt{n} \stackrel{H_0}{\sim} N(0, 1)$ (exactamente $N(0, 1)$, no solo asintóticamente), y entonces

$$\begin{aligned}\alpha(x_1, \dots, x_n) &= P_0 \{|\bar{X} - 7| > 0.2\} = P_0 \left\{ \left| \frac{\bar{X} - 7}{\sigma} \sqrt{n} \right| > \frac{0.2}{\sigma} \sqrt{n} \right\} \\ &= P \{|N(0, 1)| > 3.5\} = 2 \cdot 0.0002 = 0.0004.\end{aligned}$$

Es un p -valor tan pequeño que aporta una evidencia muy clara en contra de H_0 . Puesto que $0.01 > 0.0004$, se rechaza H_0 con $\alpha = 0.01$.

Aunque ya no es necesario, calculamos la región crítica concreta para un nivel de significación α , y la particularizamos para $\alpha = 0.01$. Imponiendo la condición $P_0(C) = \alpha$, se tiene que

$$\begin{aligned}\alpha &= P_0 \{|\bar{x} - 7| > c\} = P_0 \left\{ \left| \frac{\bar{X} - 7}{\sigma} \sqrt{n} \right| > \frac{c}{\sigma} \sqrt{n} \right\} \\ &= P \left\{ |N(0, 1)| > \frac{c}{\sigma} \sqrt{n} \right\} = 1 - P \left\{ |N(0, 1)| < \frac{c}{\sigma} \sqrt{n} \right\} = \alpha,\end{aligned}$$

y entonces $P \left\{ |N(0, 1)| < \frac{c}{\sigma} \sqrt{n} \right\} = 1 - \alpha$. Por tanto debe ser $\frac{c}{\sigma} \sqrt{n} = z_{\alpha/2}$. De aquí se obtiene $c = \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$, y por tanto

$$C = \left\{ \left| \bar{x} - \frac{1}{2} \right| > \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\}$$

es una región crítica para el contraste con nivel de significación α . Con $\alpha = 0.01$ se obtiene la región crítica $C = \{|\bar{x} - \frac{1}{2}| > 0.147\}$. Puesto que $|7.2 - \frac{1}{2}| = 0.2 > 0.147$, entonces, si se ha observado $\bar{x} = 7.2$, se rechaza H_0 con $\alpha = 0.01$ (como ya sabíamos).

Ejemplo 89. Consideremos una distribución teórica $N(\mu, \sigma)$. Consideremos la hipótesis $H_0 : \sigma^2 = 8.4$. Determina una región crítica que rechaze H_0 cuando el estimador S^2 de σ^2 tome valores demasiado alejados de 8.4 (demasiado pequeños o demasiado grandes). Se obtuvo una muestra de tamaño 21 de la que resultó la estimación $S^2 = 5.7$; contrasta H_0 con $\alpha = 0.2$ y p -valor.

Solución 90. Consideramos una región crítica de la forma $C = \{S^2 < c_1\} \cup \{S^2 > c_2\}$, siendo $c_1 < c_2$ constantes adecuadas. Determinamos c_1 y c_2 tales que $P_0(C) = \alpha$. Sea

$\sigma_0^2 = 8.4$. Puesto que $\frac{(n-1)S^2}{\sigma_0^2} \stackrel{H_0}{\sim} \chi_{n-1}^2$ (teorema de Fisher), se tiene que

$$\begin{aligned} \alpha &= P_0 \left(\{S^2 < c_1\} \cup \{S^2 > c_2\} \right) = P_0 \{S^2 < c_1\} + P_0 \{S^2 > c_2\} \\ &= P_0 \left\{ \frac{(n-1)S^2}{\sigma_0^2} < \frac{(n-1)c_1}{\sigma_0^2} \right\} + P_0 \left\{ \frac{(n-1)S^2}{\sigma_0^2} > \frac{(n-1)c_2}{\sigma_0^2} \right\} \\ &= P_0 \left\{ \chi_{n-1}^2 < \frac{(n-1)c_1}{\sigma_0^2} \right\} + P_0 \left\{ \chi_{n-1}^2 > \frac{(n-1)c_2}{\sigma_0^2} \right\}. \end{aligned}$$

Elegimos c_1 y c_2 de modo que cada sumando valga $\alpha/2$ (como se hace habitualmente). En este caso, debe ser

$$\frac{(n-1)c_1}{\sigma_0^2} = \chi_{n-1,1-\alpha/2}^2 \quad \text{y} \quad \frac{(n-1)c_2}{\sigma_0^2} = \chi_{n-1,\alpha/2}^2,$$

y de aquí se obtiene $c_1 = \frac{\chi_{n-1,1-\alpha/2}^2 \sigma_0^2}{n-1}$ y $c_2 = \frac{\chi_{n-1,\alpha/2}^2 \sigma_0^2}{n-1}$. Entonces,

$$C = \left\{ S^2 < \frac{\chi_{n-1,1-\alpha/2}^2 \sigma_0^2}{n-1} \right\} \cup \left\{ S^2 > \frac{\chi_{n-1,\alpha/2}^2 \sigma_0^2}{n-1} \right\} \quad (5.3)$$

es una región crítica para contrastar $H_0 : \sigma^2 = \sigma_0^2$ con nivel de significación α .

Una vez más, obtenemos la región crítica para $\alpha = 0.2$ además del p -valor, aunque sea innecesario. Con $\alpha = 0.2$ se obtiene $\chi_{20,0.9}^2 = 12.44$, $\chi_{20,0.1}^2 = 28.41$ y la región crítica

$$C = \{S^2 < 5.22\} \cup \{S^2 > 11.93\}.$$

Puesto que se ha obtenido $S^2 = 5.7$, se tiene que $5.22 < S^2 < 11.93$, y por tanto $S^2 \notin 5.22$ y $S^2 \notin 11.93$. Entonces, la muestra no está en la región crítica, y aceptamos H_0 con $\alpha = 0.2$. Para obtener la región crítica en la cual la observación $S^2 = 5.7$ está en el borde, primero hay que determinar cual de los dos bordes. Vamos a ver que en este caso, en que $S^2 = 5.7 < 8.4 = \sigma_0^2$, hay que considerar el borde c_1 de la rc $C = \{S^2 < c_1\} \cup \{S^2 > c_2\}$ en (5.3). Con valores de α no muy grandes (con $\alpha < 0.8$ para $n > 4$) se tiene que

$$\frac{\chi_{n-1,1-\alpha/2}^2}{n-1} < 1 < \frac{\chi_{n-1,\alpha/2}^2}{n-1}.$$

(Además $\frac{\chi_{n-1,1-\alpha/2}^2}{n-1}$ es creciente en n y converge a 1, y $\frac{\chi_{n-1,\alpha/2}^2}{n-1}$ es decreciente en n y converge a 1). De aquí se obtiene que

$$c_1 < \sigma_0^2 < c_2.$$

Puesto que $S^2 = 5.7 < 8.4 = \sigma_0^2$, entonces si la observación $S^2 = 5.7$ está en el borde de la rc, lo está en el borde correspondiente a c_1 , y no en el borde correspondiente a c_2 , ya que $c_2 > \sigma_0^2 = 8.4$, y la ecuación $c_2 = \frac{\chi_{n-1,\alpha/2}^2 \sigma_0^2}{n-1} = 5.7$ no tiene solución. De aquí se obtiene que el p -valor es el valor α solución de la ecuación $c_1 = \frac{\chi_{n-1,1-\alpha/2}^2 \sigma_0^2}{n-1} = 5.7$. (Ésto corresponde al modo de cálculo del p -valor explicado entre paréntesis en el ejemplo 2). Se obtiene $\chi_{20,1-\alpha/2}^2 = \frac{n-1}{\sigma_0^2} 5.7 = 13.57$. Buscando en las tablas se obtiene $0.7 < 1 - \frac{\alpha}{2} < 0.9$, y de aquí $0.2 < \alpha < 0.6$. Entonces,

$$0.2 < \alpha(x_1, \dots, x_n) < 0.6.$$

Puesto que $\alpha(x_1, \dots, x_n) > 0.2$, aceptamos H_0 con $\alpha = 0.2$ (como ya sabíamos).

Los elementos básicos que se han introducido hasta ahora no suponen que la distribución teórica sea paramétrica, y son válidos en todos los casos. En el resto del capítulo suponemos que la distribución teórica pertenece a cierta familia paramétrica $\{F_\theta : \theta \in \Theta\}$, con $\Theta \subset \mathbb{R}^k$. Consideramos una m.a.s. X_1, \dots, X_n , para la que observamos los datos x_1, \dots, x_n (realización muestral).

5.3 Contrastes de hipótesis paramétricos

En la estimación puntual el objetivo es estudiar el valor del parámetro, fijo pero desconocido (en el enfoque clásico). En un contraste de hipótesis paramétrico se estudia no el valor concreto del parámetro, sino que se toma una decisión razonable sobre si éste pertenece o no a un cierto subconjunto Θ_0 del espacio paramétrico. En la misma línea que con la estimación puntual, los mejores contrastes estarán basados en estadísticos suficientes. En el resto de este capítulo estudiaremos los contrastes de hipótesis paramétricos.

5.3.1 Contrastes unilaterales y bilaterales

Los *contrastos unilaterales* son de la forma

$$\begin{aligned} H_0 : \theta \leq \theta_0 \quad \text{frente a} \quad H_1 : \theta > \theta_0, \text{ o bien} \\ H_0 : \theta \geq \theta_0 \quad \text{frente a} \quad H_1 : \theta < \theta_0. \end{aligned}$$

Los *contrastos bilaterales* son de la forma

$$\begin{aligned} H_0 : \theta = \theta_0 \quad \text{frente a} \quad H_1 : \theta \neq \theta_0, \text{ o bien} \\ H_0 : \theta \in [\theta_1, \theta_2] \quad \text{frente a} \quad H_1 : \theta \notin [\theta_1, \theta_2]. \end{aligned}$$

Los contrastes de todos los ejemplos anteriores, con H_0 simple, eran bilaterales.

5.3.2 Función de potencia

En un contraste paramétrico se define la *función de potencia* $\beta : \Theta \rightarrow [0, 1]$ como

$$\beta(\theta) = P_\theta(C),$$

de modo que, para cada θ , $\beta(\theta)$ es la probabilidad de rechazar H_0 si el verdadero valor del parámetro es θ . Entonces, un contraste paramétrico tiene nivel de significación α si

$$\beta(\theta) \leq \alpha \quad \text{para cada } \theta \in \Theta_0,$$

esto es, si la *pet I* es menor o igual que α sea cual sea el valor del parámetro en Θ_0 (bajo H_0).

Se denomina *tamaño del test* al valor

$$\sup_{\theta \in \Theta_0} \beta(\theta),$$

que es la máxima probabilidad de rechazar H_0 entre todos los valores del parámetro para los que es cierta (cometiendo error de tipo I). Entonces, si H_0 es cierta, sea cual sea el verdadero valor del parámetro en Θ_0 la *pet I* siempre será menor o igual que el tamaño del test.

Una *pet II* pequeña supone que $P_\theta(C^c) = 1 - \beta(\theta)$ sea pequeña bajo H_1 , o sea, que $\beta(\theta)$ sea grande para cualquier valor del parámetro en Θ_1 .

Dadas dos regiones críticas C y C' (dos tests) para contrastar $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta_1$, con nivel de significación α ambas, se dice que la primera es *uniformemente más potente* que la segunda si

$$\beta_C(\theta) \geq \beta_{C'}(\theta) \quad \text{para todo } \theta \in \Theta_1.$$

Se dice que un test es *uniformemente de máxima potencia (ump)*, dentro de una familia de tests, si es uniformemente más potente que cualquier otro con el mismo nivel de significación.

Ejemplo 91. Consideremos una distribución teórica $N(\mu, \sigma)$ con $\sigma = 0.4$. Consideremos la hipótesis $H_0 : \mu \geq \mu_0$, con $\mu_0 = 7$ y una rc de la forma $C = \{\bar{x} < c\}$. Determinamos la función de potencia, y a partir de ella una rc concreta para un test con nivel de significación α .

Es un contraste unilateral (la hipótesis alternativa es $H_1 : \mu < \mu_0$, la negación de H_0). Se tiene que

$$\begin{aligned} \beta(\mu) &= P_\mu(C) = P_\mu\{\bar{X} < c\} \\ &= P_\mu\left\{\frac{\bar{X} - \mu}{\sigma}\sqrt{n} < \frac{c - \mu}{\sigma}\sqrt{n}\right\} = P\left\{N(0, 1) < \frac{c - \mu}{\sigma}\sqrt{n}\right\}. \end{aligned}$$

Cuando μ crece se tiene que $\frac{c - \mu}{\sigma}\sqrt{n}$ decrece, y entonces $\beta(\mu)$ decrece. Por tanto, $\beta(\mu)$ es una función decreciente. Se tiene que $\lim_{\mu \rightarrow -\infty} \beta(\mu) = 1$ y $\lim_{\mu \rightarrow \infty} \beta(\mu) = 0$. Por tanto, tomando c tal que $\beta(\mu_0) = \alpha$ se consigue que sea $\beta(\mu) \leq \alpha$ para cada $\mu \in \Theta_0 = [\mu_0, \infty)$, esto es, se consigue que el test tenga nivel de significación α . Calculamos este valor de c . Imponiendo la condición $\beta(\mu_0) = \alpha$ se obtiene

$$\alpha = \beta(\mu_0) = P\left\{N(0, 1) < \frac{c - \mu_0}{\sigma}\sqrt{n}\right\} = 1 - P\left\{N(0, 1) > \frac{c - \mu_0}{\sigma}\sqrt{n}\right\},$$

y por tanto $P\left\{N(0, 1) > \frac{c - \mu_0}{\sigma}\sqrt{n}\right\} = 1 - \alpha$. Entonces, $\frac{c - \mu_0}{\sigma}\sqrt{n} = z_{1-\alpha} = -z_\alpha$, y de aquí se obtiene $c = \mu_0 - \frac{\sigma}{\sqrt{n}}z_\alpha$. Por tanto,

$$C = \left\{\bar{x} < \mu_0 - \frac{\sigma}{\sqrt{n}}z_\alpha\right\}$$

es una región crítica para contrastar $H_0 : \mu \geq \mu_0$ con nivel de significación α .

5.3.3 Contraste de hipótesis simple frente a simple

En el caso de hipótesis simple frente a simple el lema de Neyman-Pearson proporciona el test de (uniformemente) máxima potencia:

Lema 92. Si $C \subset \Omega$ es tal que

$$\begin{aligned} \{(x_1, \dots, x_n) \in \Omega : f_{\theta_1} > k f_{\theta_0}\} &\subset C \text{ y} \\ C &\subset \{(x_1, \dots, x_n) \in \Omega : f_{\theta_1} \geq k f_{\theta_0}\} \end{aligned}$$

para alguna constante $k > 0$, y se verifica que $P_{\theta_0}(C) = \alpha$, entonces C es la región crítica de un test de máxima potencia, con nivel de significación α , para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$. Además, C se expresa en función del estadístico ms (minimal suficiente).

Demostración. Solo lo demostramos en el caso continuo. El símbolo f_{θ_0} es una expresión abreviada de $f_{\theta_0}(x_1, \dots, x_n)$, y lo mismo para f_{θ_1} . La condición, válida en los casos discreto y continuo, se puede expresar de la siguiente manera, mas sencilla, en el caso continuo:

$$\begin{aligned} C &= \{f_{\theta_1} > k f_{\theta_0}\} \\ &= \{(x_1, \dots, x_n) \in \Omega : f_{\theta_1}(x_1, \dots, x_n) > k f_{\theta_0}(x_1, \dots, x_n)\} , \end{aligned}$$

puesto que la probabilidad de que sea $f_{\theta_1} = k f_{\theta_0}$ vale 0. Se verifica que la desigualdad $f_{\theta_1} > k f_{\theta_0}$ se expresa en términos del estadístico ms. Supongamos que C tiene nivel de significación α , esto es, $P_{\theta_0}(C) = \alpha$. Sea C' otra rc (cualquiera) con $P_{\theta_0}(C') = \alpha$. Demostraremos que $P_{\theta_1}(C^c) \leq P_{\theta_1}(C'^c)$ (o sea, que hay menor o igual pet II con C que con C'), con lo que queda demostrado el lema. Para cualquier $A \subset C$ se tiene que $P_{\theta_1}(A) > k P_{\theta_0}(A)$:

$$\begin{aligned} P_{\theta_1}(A) &= \int_A f_{\theta_1}(x_1, \dots, x_n) dx_1 \cdots dx_n > \int_A k f_{\theta_0}(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &= k P_{\theta_0}(A) . \end{aligned}$$

Del mismo modo se obtiene que $P_{\theta_1}(A) \leq k P_{\theta_0}(A)$ si $A \subset C^c$. Puesto que

$$P_{\theta}(C') = P_{\theta}(C' \cap C) + P_{\theta}(C' \cap C^c) \text{ para } \theta \in \{\theta_1, \theta_2\} ,$$

se tiene que

$$\begin{aligned} &P_{\theta_1}(C) - P_{\theta_1}(C') - k(P_{\theta_0}(C) - P_{\theta_0}(C')) \\ &= P_{\theta_1}(C) - k P_{\theta_0}(C) - (P_{\theta_1}(C') - k P_{\theta_0}(C')) \\ &= P_{\theta_1}(C) - k P_{\theta_0}(C) - (P_{\theta_1}(C' \cap C) - k P_{\theta_0}(C' \cap C)) \\ &\quad - (P_{\theta_1}(C' \cap C^c) - k P_{\theta_0}(C' \cap C^c)) \\ &= P_{\theta_1}(C) - P_{\theta_1}(C' \cap C) - k(P_{\theta_0}(C) - P_{\theta_0}(C' \cap C)) \\ &\quad + k P_{\theta_0}(C' \cap C^c) - P_{\theta_1}(C' \cap C^c) \\ &\geq 0 + 0 + 0 = 0 . \end{aligned}$$

De ello se obtiene que $P_{\theta_1}(C) - P_{\theta_1}(C') \geq k(P_{\theta_0}(C) - P_{\theta_0}(C')) = k(\alpha - \alpha) = 0$. Entonces, $P_{\theta_1}(C) \geq P_{\theta_1}(C')$, y de aquí $P_{\theta_1}(C^c) \leq P_{\theta_1}(C'^c)$. ■

Ejemplo 93. Consideremos una distribución teórica $\mathcal{P}(\theta)$ y sea $\theta_0 < \theta_1$. Determinamos la región crítica del test de máxima potencia para contrastar

$$H_0 : \theta = \theta_0 \text{ frente a } H_1 : \theta = \theta_1 .$$

Sea $t = \sum_{i=1}^n x_i$. Se tiene que:

$$\begin{aligned} \{f_{\theta_1} > k f_{\theta_0}\} &= \left\{ e^{-n\theta_1} \frac{\theta_1^t}{x_1! \cdots x_n!} > k e^{-n\theta_0} \frac{\theta_0^t}{x_1! \cdots x_n!} \right\} = \left\{ (\theta_1/\theta_0)^t > k e^{n(\theta_1-\theta_0)} \right\} \\ &= \{t \log(\theta_1/\theta_0) > \log k + n(\theta_1 - \theta_0)\} = \{\bar{x} > c\} , \end{aligned}$$

con $c = ((1/n) \log k + \theta_1 - \theta_0) / \log(\theta_1/\theta_0)$. Entonces, la rc óptima del test ump es de la forma

$$C = \{\bar{x} > c\} .$$

Calculamos la rc concreta para el nivel de significación α mediante aproximación normal (calculamos c).

$$\begin{aligned} \alpha &= P_{\theta_0}(C) = P_{\theta_0}\{\bar{X} > c\} = P_{\theta_0}\left\{\frac{\bar{X} - \theta_0}{\sqrt{\theta_0}}\sqrt{n} > \frac{c - \theta_0}{\sqrt{\theta_0}}\sqrt{n}\right\} \\ &\simeq P\left\{N(0, 1) > \frac{c - \theta_0}{\sqrt{\theta_0}}\sqrt{n}\right\} = \alpha . \end{aligned}$$

Por tanto, $\frac{c - \theta_0}{\sqrt{\theta_0}}\sqrt{n} = z_\alpha$, y de aquí se obtiene $c = \theta_0 + \frac{\sqrt{\theta_0}}{\sqrt{n}}z_\alpha$. Entonces, la región crítica concreta del test ump para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$ con nivel de significación α (aproximado) es

$$C = \left\{ \bar{x} > \theta_0 + \frac{\sqrt{\theta_0}}{\sqrt{n}}z_\alpha \right\} . \quad (5.4)$$

Obsérvese que C no depende de θ_1 . Lo que sí depende de θ_1 es la pet II:

$$\begin{aligned} P_{\theta_1}(C^c) &= P_{\theta_1}\{\bar{X} \leq c\} = P_{\theta_1}\left\{\frac{\bar{X} - \theta_1}{\sqrt{\theta_1}}\sqrt{n} \leq \frac{c - \theta_1}{\sqrt{\theta_1}}\sqrt{n}\right\} \\ &\simeq P\left\{N(0, 1) < \frac{c - \theta_1}{\sqrt{\theta_1}}\sqrt{n}\right\} = P\left\{N(0, 1) < \sqrt{\frac{\theta_0}{\theta_1}}z_\alpha + \frac{\theta_0 - \theta_1}{\sqrt{\theta_1}}\sqrt{n}\right\} . \end{aligned}$$

Ejemplo 94. Se obtuvo una muestra de tamaño 25 de una distribución teórica $\mathcal{P}(\theta)$, con $\bar{x} = 10.3$. Contrastamos la hipótesis $H_0 : \theta = 9$ frente a $H_1 : \theta = 12$, con $\alpha = 0.01$ (calculamos también la pet II).

Utilizamos el test ump dado en (5.4). Con $\alpha = 0.01$ se obtiene $C = \left\{ \bar{x} > 9 + \frac{\sqrt{9}}{\sqrt{25}}2.33 \right\} = \{\bar{x} > 10.40\}$. La pet II vale

$$\begin{aligned} P_{\theta=12}(C^c) &\simeq P\left\{N(0, 1) < \frac{10.40 - 12}{\sqrt{12}}\sqrt{25}\right\} \\ &= P\{N(0, 1) < -2.31\} = 0.0104 \end{aligned}$$

Puesto que $10.3 \not\approx 10.40$, se acepta H_0 con $\alpha = 0.01$. En este caso, a diferencia de lo que ocurre en la mayor parte de las ocasiones, podemos estar razonablemente seguros de que la aceptación es correcta (no solamente de que no hay evidencia en contra de H_0): si la aceptación no fuera correcta, es que estaríamos cometiendo un error de tipo II, que tiene probabilidad $P_{\theta=12}(C^c) \simeq 0.0104$, pequeña.

5.4 Teoría de contrastes paramétricos óptimos

El caso de hipótesis simple frente a simple, que tiene solución óptima según se vió en el apartado anterior, raramente se presenta en la práctica. Para contrastes paramétricos unilaterales y bilaterales no siempre hay un test uniformemente de máxima potencia, aunque sí en muchos casos.

Suponemos que el parámetro es unidimensional. Los resultados que se presentan para contrastes unilaterales y bilaterales son consecuencia del teorema de Karlin-Rubin y del de Lehmann. Por simplicidad, no se presenta el resultado completo de estos teoremas, aunque sí aspectos principales.

5.4.1 Contrastes unilaterales

Se dice que una familia paramétrica tiene *razón de verosimilitud monótona* (en T) si existe un estadístico unidimensional T tal que, para cada θ, θ' con $\theta < \theta'$,

$$\frac{f_{\theta'}(x_1, \dots, x_n)}{f_{\theta}(x_1, \dots, x_n)} \quad (5.5)$$

es una función creciente de $T(x_1, \dots, x_n)$.

Si una familia es de tipo exponencial uniparamétrico y $q(\theta)$ es monótona, entonces la familia tiene razón de verosimilitud monótona en el estadístico ms: se tiene que

$$\begin{aligned} \frac{f_{\theta'}(x_1, \dots, x_n)}{f_{\theta}(x_1, \dots, x_n)} &= \frac{c(\theta')^n \left(\prod_{i=1}^n h(x_i)\right) \exp\{q(\theta') \sum_{i=1}^n V(x_i)\}}{c(\theta)^n \left(\prod_{i=1}^n h(x_i)\right) \exp\{q(\theta) \sum_{i=1}^n V(x_i)\}} \\ &= \frac{c(\theta')^n}{c(\theta)^n} \exp\left\{(q(\theta') - q(\theta)) \sum_{i=1}^n V(x_i)\right\}, \end{aligned}$$

que es función creciente de $T = \sum_{i=1}^n V(x_i)$ si $q(\theta)$ es creciente, y es función creciente de $T = -\sum_{i=1}^n V(x_i)$ si $q(\theta)$ es decreciente.

Ejemplo 95. Determina el estadístico T para el que las siguientes distribuciones tienen razón de verosimilitud monótona: a) $B(1, \theta)$, b) $\mathcal{P}(\theta)$, c) $Exp(\theta)$, d) $N(\mu, \sigma)$, con $\theta = \mu$ y σ conocida, e) $N(\mu, \sigma)$, con $\theta = \sigma$ y μ conocida.

Solución 96. a) Puesto que $q(\theta) = \log \frac{\theta}{1-\theta} = \log \frac{1}{\frac{1}{\theta}-1}$ es creciente para $\theta \in \Theta = (0, 1)$ y $V(x) = x$, se tiene que la familia de las distribuciones de Bernoulli tiene razón de verosimilitud monótona en $T = \sum_{i=1}^n V(X_i) = \sum_{i=1}^n X_i$.

También podemos obtener T de un modo directo, sin utilizar que la distribución de Bernoulli es de tipo exponencial uniparamétrico. Sea $t = \sum_{i=1}^n x_i$. Se tiene que

$$\frac{f_{\theta'}(x_1, \dots, x_n)}{f_{\theta}(x_1, \dots, x_n)} = \frac{\theta'^t (1 - \theta')^{n-t}}{\theta^t (1 - \theta)^{n-t}} = \frac{(1 - \theta')^n}{(1 - \theta)^n} \left(\frac{\theta'}{1 - \theta'} \right)^t.$$

Puesto que $\frac{\theta}{1-\theta}$ es una función creciente de θ , se tiene que $\frac{\theta}{1-\theta} < \frac{\theta'}{1-\theta'}$, y de aquí se obtiene que el cociente es función creciente de t . Por tanto, $T = \sum_{i=1}^n X_i$, como ya sabíamos.

b) Puesto que $q(\theta) = \log \theta$ es creciente para $\theta \in \Theta = (0, \infty)$ y $V(x) = x$, se tiene que la familia de las distribuciones de Poisson tiene razón de verosimilitud monótona en $T = \sum_{i=1}^n V(X_i) = \sum_{i=1}^n X_i$.

c) Puesto que $q(\theta) = -\theta$ es decreciente y $V(x) = x$, se tiene que la familia de las distribuciones exponenciales tiene razón de verosimilitud monótona en el estadístico $T = -\sum_{i=1}^n V(X_i) = -\sum_{i=1}^n X_i$.

d) Puesto que $q(\mu) = \frac{\mu}{\sigma^2}$ es creciente y $V(x) = x$, se tiene que la familia considerada tiene razón de verosimilitud monótona en $T = \sum_{i=1}^n V(X_i) = \sum_{i=1}^n X_i$.

e) Puesto que $q(\sigma) = -\frac{1}{2\sigma^2}$ es creciente y $V(x) = (x - \mu)^2$, se tiene que la familia considerada tiene razón de verosimilitud monótona en el estadístico ms

$$T = \sum_{i=1}^n V(X_i) = \sum_{i=1}^n (X_i - \mu)^2.$$

Ejemplo 97. Consideremos la función de densidad teórica

$$f_{\theta}(x) = e^{\theta-x} I\{x > \theta\}$$

($X \sim \theta + Exp(1)$). No es una familia de tipo exponencial, y entonces debemos manipular el cociente (5.5) directamente. La función de densidad de la muestra es

$$f_{\theta}(x_1, \dots, x_n) = \exp \left\{ n\theta - \sum_{i=1}^n x_i \right\} I\{x_{(1)} > \theta\},$$

y se tiene que

$$\frac{f_{\theta'}(x_1, \dots, x_n)}{f_{\theta}(x_1, \dots, x_n)} = \frac{\exp \{n\theta' - \sum_{i=1}^n x_i\} I\{x_{(1)} > \theta'\}}{\exp \{n\theta - \sum_{i=1}^n x_i\} I\{x_{(1)} > \theta\}} = e^{n(\theta' - \theta)} \frac{I\{x_{(1)} > \theta'\}}{I\{x_{(1)} > \theta\}},$$

que solo está definida para $x_{(1)} > \theta$. Es función creciente de $x_{(1)}$ (no decreciente, para ser más precisos), puesto que vale 0 si $\theta < x_{(1)} \leq \theta'$ y vale $e^{n(\theta' - \theta)} > 0$ si $x_{(1)} > \theta'$. Por tanto, la familia tiene razón de verosimilitud monótona en $T = X_{(1)}$.

Si la distribución teórica tiene razón de verosimilitud monótona en el estadístico T , entonces, para contrastar

$$H_0 : \theta \leq \theta_0 \text{ frente a } H_1 : \theta > \theta_0,$$

el test de uniformemente máxima potencia tiene región crítica de la forma

$$C = \{T > c\}.$$

La función de potencia es creciente, y por tanto, si el nivel de significación es α debe ser

$$P_{\theta_0}(C) = \alpha.$$

Para contrastar $H_0 : \theta \geq \theta_0$, la región crítica del test de uniformemente máxima potencia es de la forma $C = \{T < c\}$. La función de potencia es decreciente y $P_{\theta_0}(C) = \alpha$.

Ejemplo 98. Determina la forma de la región crítica de uniformemente máxima potencia para contrastar $H_0 : \theta \leq \theta_0$ en los siguientes casos:

- a) $B(1, \theta)$, b) $\mathcal{P}(\theta)$, c) $Exp(\theta)$, d) $N(\mu, \sigma)$, con $\theta = \mu$ y σ conocida,
e) $N(\mu, \sigma)$, con $\theta = \sigma$ y μ conocida, f) $f_{\theta}(x) = e^{\theta-x} I\{x > \theta\}$.

Solución 99. a) $C = \{T > c\} = \{\sum_{i=1}^n X_i > c\} = \{\bar{X} > c'\}$ (con $c' = c/n$).

b) $C = \{T > c\} = \{\bar{X} > c'\}$.

c) $C = \{T > c\} = \{-\sum_{i=1}^n X_i > c\} = \{-\bar{X} > c'\} = \{\bar{X} < c''\}$.

d) $C = \{T > c\} = \{\bar{X} > c'\}$.

e) $C = \{T > c\} = \{\sum_{i=1}^n (X_i - \mu)^2 > c\} = \{\hat{\sigma}^2 > c'\}$, con

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

f) $C = \{T > c\} = \{X_{(1)} > c\}$.

5.4.2 Contrastes bilaterales

Si la distribución teórica es de tipo exponencial uniparamétrico, entonces, para contrastar

$$H_0 : \theta = \theta_0 \text{ frente a } H_1 : \theta \neq \theta_0,$$

el mejor test (no es exactamente de uniformemente máxima potencia) tiene región crítica de la forma

$$C = \{T < c_1\} \cup \{T > c_2\},$$

siendo $T = \sum_{i=1}^n V(X_i)$ el estadístico ms.

Observese que si el nivel de significación es α debe ser $P_{\theta_0}(C) = \alpha$.

El mejor test para contrastar $H_0 : \theta \in [\theta_1, \theta_2]$ también tiene región crítica de la forma $C = \{T < c_1\} \cup \{T > c_2\}$, con $P_{\theta_1}(C) = \alpha$ y $P_{\theta_2}(C) = \alpha$.

5.5 Contraste de razón de verosimilitudes

Sería conveniente disponer de algún método general que dé lugar a contrastes que, aunque no sean tests de uniformemente de máxima potencia, sean razonablemente buenos. Del mismo modo que, en estimación puntual, el método de máxima verosimilitud es un método general que se puede aplicar en cualquier situación paramétrica, y da lugar a buenos estimadores en la mayoría de ocasiones, así ocurre con el método de la razón de verosimilitudes para contrastes paramétricos.

El método de la razón de verosimilitudes está basado en estimadores de máxima verosimilitud, y permite obtener una región crítica razonablemente buena para cualquier contraste paramétrico. En muchas ocasiones coincide con el contraste de máxima potencia. Por ejemplo, en el caso de un test de hipótesis simple frente a simple, coincide con el contraste obtenido por el lema de Neyman-Pearson. El método es de una gran generalidad, puesto que proporciona un test para contrastar cualquier hipótesis, con cualquier dimensión del espacio paramétrico. También para el caso en que hay parámetros desconocidos a los que no se refieren las hipótesis explícitamente (por ejemplo, $H_0 : \mu = \mu_0$ para una $N(\mu, \sigma)$ con (μ, σ) desconocido).

Consideramos $\Theta \subset \mathbb{R}^k$, siendo k arbitrario. Sea $\Theta_0 \subset \Theta$ arbitrario y $\Theta_1 = \Theta_0^c$ su complementario en Θ . Vamos a presentar un procedimiento general para contrastar

$$H_0 : \theta \in \Theta_0 \text{ frente a } H_1 : \theta \in \Theta_1 .$$

Se denomina *razón de verosimilitudes* (rv) al estadístico

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} f_{\theta}(\mathbf{x})}{\sup_{\theta \in \Theta} f_{\theta}(\mathbf{x})} = \frac{f_{\hat{\theta}_0}(\mathbf{x})}{f_{\hat{\theta}}(\mathbf{x})} ,$$

siendo $\hat{\theta}$ la estimación de máxima verosimilitud de θ asociada a la muestra $\mathbf{x} = (x_1, \dots, x_n)$, y $\hat{\theta}_0$ la estimación de máxima verosimilitud en el espacio paramétrico Θ_0 (o sea, bajo H_0). Obsérvese que $\Lambda \leq 1$. El *test de razón de verosimilitudes* (trv) es el que tiene región crítica de la forma

$$C = \{\Lambda(\mathbf{x}) < c\}$$

para algún $c < 1$. Se rechaza la hipótesis nula cuando el supremo de la función de verosimilitud en Θ_0 (bajo H_0) es suficientemente más pequeña que el supremo en Θ , el espacio paramétrico completo.

Con frecuencia, la desigualdad $\Lambda(\mathbf{x}) < c$ puede expresarse en términos de un estadístico más simple que Λ , y cuya distribución es conocida. Si no es así, la principal dificultad en la construcción del test de razón de verosimilitudes consiste en obtener la distribución de Λ ; ésta es necesaria para obtener la región crítica concreta, esto es, el valor de c dado α .

Cuando la región crítica $C = \{\Lambda < c\}$ no puede expresarse en términos de un estadístico más simple que Λ , y tampoco se dispone de la distribución de Λ , se puede utilizar el siguiente resultado asintótico. Bajo las mismas condiciones que aseguran la distribución asintótica normal del emv (tema 3), se verifica que

$$-2 \log \Lambda(X_1, \dots, X_n) \stackrel{H_0}{\sim} \chi_{k-q}^2 , \quad (5.6)$$

siendo k el número de parámetros y q el número de parámetros bajo H_0 . De este modo, el número de grados de libertad $k - q$ coincide con la reducción en el número de parámetros del problema efectuada por la hipótesis nula. Con la aproximación dada por el resultado (5.6) obtenemos la siguiente región crítica para un nivel de significación α (aproximadamente):

$$C = \{-2 \log \Lambda > \chi_{k-q, \alpha}^2\} .$$

Con este método se puede obtener una región crítica para cualquier contraste paramétrico, con cualquier número de parámetros. Los desarrollos son algo laboriosos, pero están basados en operaciones elementales.

Ejemplo 100. Comprobamos que el trv para un contraste de hipótesis simple frente a simple coincide con el proporcionado por el lema de Neyman-Pearson.

Queremos contrastar

$$H_0 : \theta = \theta_0 \text{ frente a } H_1 : \theta = \theta_1 .$$

Obviamente, $\hat{\theta}_0 = \theta_0$, y entonces $f_{\hat{\theta}_0}(\mathbf{x}) = f_{\theta_0}(\mathbf{x})$, que denotamos por f_{θ_0} . El espacio paramétrico es $\Theta = \Theta_0 \cup \Theta_1 = \{\theta_0\} \cup \{\theta_1\} = \{\theta_0, \theta_1\}$, y entonces

$$\sup_{\theta \in \Theta} f_{\theta}(\mathbf{x}) = \max \{f_{\theta_0}(\mathbf{x}), f_{\theta_1}(\mathbf{x})\} = \max \{f_{\theta_0}, f_{\theta_1}\} .$$

De aquí se obtiene

$$\begin{aligned} \Lambda(\mathbf{x}) &= \frac{f_{\theta_0}}{\max \{f_{\theta_0}, f_{\theta_1}\}} , \text{ con} \\ \Lambda(\mathbf{x}) &= 1 \text{ si } f_{\theta_0}(\mathbf{x}) \geq f_{\theta_1}(\mathbf{x}) \text{ y} \\ \Lambda(\mathbf{x}) &= \frac{f_{\theta_0}}{f_{\theta_1}} \text{ si } f_{\theta_0}(\mathbf{x}) < f_{\theta_1}(\mathbf{x}) \end{aligned}$$

y entonces la rc del trv es

$$C = \{\Lambda < c\} = \left\{ \frac{f_{\theta_0}}{f_{\theta_1}} < c \right\} = \{f_{\theta_1} > c^{-1} f_{\theta_0}\} ,$$

que es justamente la rc $C = \{f_{\theta_1} > k f_{\theta_0}\}$ del lema de Neyman-Pearson, con $k = c^{-1}$.

Ejemplo 101. Consideremos una distribución teórica $N(\mu, \sigma)$ con $\sigma = \sigma_0$, conocida. Obtenemos el trv para contrastar $H_0 : \mu = \mu_0$.

La hipótesis nula es simple. La emv es $\hat{\mu} = \bar{x}$. Obviamente, se tiene que $\hat{\mu}_0 = \mu_0$. La función de densidad de una muestra de una $N(\mu, \sigma_0)$ es

$$f_{\mu}(\mathbf{x}) = (2\pi)^{-n/2} \sigma_0^{-n} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} ,$$

y por tanto

$$\begin{aligned} f_{\hat{\mu}}(\mathbf{x}) &= (2\pi)^{-n/2} \sigma_0^{-n} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right\} \\ &= (2\pi)^{-n/2} \sigma_0^{-n} \exp \left\{ -\frac{1}{2\sigma_0^2} n s^2 \right\} , \text{ y} \\ f_{\hat{\mu}_0}(\mathbf{x}) &= (2\pi)^{-n/2} \sigma_0^{-n} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right\} . \end{aligned}$$

De aquí se obtiene

$$\begin{aligned}\Lambda &= \frac{f_{\hat{\mu}_0}(\mathbf{x})}{f_{\hat{\mu}}(\mathbf{x})} = \frac{(2\pi)^{-n/2} \sigma_0^{-n} \exp\left\{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_0)^2\right\}}{(2\pi)^{-n/2} \sigma_0^{-n} \exp\left\{-\frac{1}{2\sigma_0^2} ns^2\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma_0^2} \left(\sum_{i=1}^n (x_i - \mu_0)^2 - ns^2\right)\right\} \\ &= \exp\left\{-\frac{1}{2\sigma_0^2} (n\mu_0^2 - 2n\mu_0\bar{x} - n\bar{x}^2)\right\} = \exp\left\{-\frac{n}{2\sigma_0^2} (\bar{x} - \mu_0)^2\right\}.\end{aligned}$$

El estadístico Λ es función del estadístico ms \bar{X} . Despejando \bar{x} en la desigualdad $\Lambda < c$ se obtiene:

$$\begin{aligned}C &= \{\Lambda < c\} = \left\{\exp\left(-\frac{n}{2\sigma_0^2} (\bar{x} - \mu_0)^2\right) < c\right\} = \left\{-\frac{n}{2\sigma_0^2} (\bar{x} - \mu_0)^2 < \log c\right\} \\ &= \left\{(\bar{x} - \mu_0)^2 > -\frac{2\sigma_0^2}{n} \log c\right\} = \left\{|\bar{x} - \mu_0| > \sqrt{-\frac{2\sigma_0^2}{n} \log c}\right\}.\end{aligned}$$

Por tanto, la rc es de la forma $C = \{|\bar{x} - \mu_0| > c'\}$, con $c' = \sqrt{-\frac{2\sigma_0^2}{n} \log c}$. Determinamos c' para un nivel de significación α , teniendo en cuenta que la distribución nula de \bar{X} es $N(0, 1)$. Por comodidad, llamamos c a c' .

$$\begin{aligned}\alpha &= P_0\{|\bar{X} - \mu_0| > c\} = P_0\left\{\left|\frac{\bar{X} - \mu_0}{\sigma_0} \sqrt{n}\right| > \frac{c}{\sigma_0} \sqrt{n}\right\} \\ &= P\left\{|N(0, 1)| > \frac{c}{\sigma_0} \sqrt{n}\right\},\end{aligned}$$

y entonces $\frac{c}{\sigma_0} \sqrt{n} = z_{\alpha/2}$. Despejando, $c = \frac{\sigma_0}{\sqrt{n}} z_{\alpha/2}$, obteniéndose la rc

$$C = \left\{|\bar{x} - \mu_0| > \frac{\sigma_0}{\sqrt{n}} z_{\alpha/2}\right\}.$$

Ejemplo 102. Consideremos una distribución teórica $N(\mu, \sigma)$. Obtenemos el trv para contrastar $H_0: \mu = \mu_0$.

Cuando no se mencione explícitamente que un parámetro es conocido, como ocurre aquí con σ , supondremos que es desconocido. Hay por tanto dos parámetros desconocidos, μ y σ , y la hipótesis nula es compuesta en este caso. Las emv son $\hat{\mu} = \bar{x}$ y $\hat{\sigma} = s$ (véase el tema 3). Obviamente, se tiene que $\hat{\mu}_0 = \mu_0$. La estimación mv de σ bajo H_0 , $\hat{\sigma}_0$, es la estimación mv de σ para una $N(\mu, \sigma)$ con $\mu = \mu_0$ conocida, que vale

$$\hat{\sigma}_0 = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2}.$$

La función de densidad de una muestra de una $N(\mu, \sigma)$ es

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

y por tanto

$$\begin{aligned} f_{\hat{\mu}, \hat{\sigma}}(\mathbf{x}) &= (2\pi)^{-n/2} s^{-n} \exp \left\{ -\frac{1}{2s^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right\} \\ &= (2\pi)^{-n/2} s^{-n} \exp \left\{ -\frac{1}{2s^2} ns^2 \right\} = (2\pi)^{-n/2} s^{-n} \exp \left\{ -\frac{n}{2} \right\}, \text{ y} \\ f_{\hat{\mu}_0, \hat{\sigma}_0}(\mathbf{x}) &= (2\pi)^{-n/2} \hat{\sigma}_0^{-n} \exp \left\{ -\frac{1}{2\hat{\sigma}_0^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right\} \\ &= (2\pi)^{-n/2} \hat{\sigma}_0^{-n} \exp \left\{ -\frac{n}{2} \right\}. \end{aligned}$$

De aquí se obtiene

$$\begin{aligned} \Lambda &= \frac{f_{\hat{\mu}_0, \hat{\sigma}_0}(\mathbf{x})}{f_{\hat{\mu}, \hat{\sigma}}(\mathbf{x})} = \frac{(2\pi)^{-n/2} \hat{\sigma}_0^{-n} \exp \left\{ -\frac{n}{2} \right\}}{(2\pi)^{-n/2} s^{-n} \exp \left\{ -\frac{n}{2} \right\}} = \frac{\hat{\sigma}_0^{-n}}{s^{-n}} = \left(\frac{\hat{\sigma}_0^2}{s^2} \right)^{-n/2} \\ &= \left(\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^{-n/2} = \left(\frac{\sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^{-n/2} \\ &= \left(\frac{ns^2 + n(\bar{x} - \mu_0)^2 + 0}{ns^2} \right)^{-n/2} = \left(1 + \frac{(\bar{x} - \mu_0)^2}{s^2} \right)^{-n/2} \end{aligned}$$

El estadístico Λ es función del estadístico ms (\bar{X}, s^2) . Despejando en la desigualdad $\Lambda < c$ se obtiene una expresión para la rc en función de un estadístico con distribución nula t de Student:

$$\begin{aligned} C = \{\Lambda < c\} &= \left\{ \left(1 + \frac{(\bar{x} - \mu_0)^2}{s^2} \right)^{-n/2} < c \right\} = \left\{ 1 + \frac{(\bar{x} - \mu_0)^2}{s^2} > c^{-2/n} \right\} \\ &= \left\{ \left| \frac{\bar{x} - \mu_0}{s} \right| > \sqrt{c^{-2/n} - 1} \right\} = \left\{ \left| \sqrt{n-1} \frac{\bar{x} - \mu_0}{s} \right| > \sqrt{n-1} \sqrt{c^{-2/n} - 1} \right\}. \end{aligned}$$

Por tanto $C = \left\{ \left| \sqrt{n-1} \frac{\bar{x} - \mu_0}{s} \right| > c' \right\}$. Esta expresión permite obtener la rc concreta de un modo muy sencillo. Para un nivel de significación α tomamos

$$\alpha = P_0 \left\{ \left| \sqrt{n-1} \frac{\bar{X} - \mu_0}{s} \right| > c \right\} = P_0 \{ |t_{n-1}| > c \},$$

y entonces $c = t_{n-1, \alpha/2}$, obteniéndose la rc

$$C = \left\{ \left| \sqrt{n-1} \frac{\bar{x} - \mu_0}{s} \right| > t_{n-1, \alpha/2} \right\}.$$

Ejemplo 103. Consideremos una distribución teórica $N(\mu, \sigma)$. Obtenemos el trv para contrastar $H_0 : \sigma = \sigma_0$.

Hay dos parámetros desconocidos, μ y σ , y la hipótesis nula es compuesta en este caso.

Las emv son $\hat{\mu} = \bar{x}$ y $\hat{\sigma} = s$. Obviamente, se tiene que $\hat{\sigma}_0 = \sigma_0$. La estimación mv de μ bajo H_0 es $\hat{\mu}_0 = \bar{x}$. De aquí se obtiene

$$\begin{aligned} f_{\hat{\mu}, \hat{\sigma}}(\mathbf{x}) &= (2\pi)^{-n/2} s^{-n} \exp\{-n/2\} \quad (\text{como en el ejemplo anterior}), \\ f_{\hat{\mu}_0, \hat{\sigma}_0}(\mathbf{x}) &= (2\pi)^{-n/2} \sigma_0^{-n} \exp\left\{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2\right\} \\ &= (2\pi)^{-n/2} \sigma_0^{-n} \exp\left\{-ns^2/(2\sigma_0^2)\right\}, \text{ y} \\ \Lambda &= \frac{f_{\hat{\mu}_0, \hat{\sigma}_0}(\mathbf{x})}{f_{\hat{\mu}, \hat{\sigma}}(\mathbf{x})} = \frac{(2\pi)^{-n/2} \sigma_0^{-n} \exp\{-ns^2/(2\sigma_0^2)\}}{(2\pi)^{-n/2} s^{-n} \exp\{-n/2\}} \\ &= \frac{\sigma_0^{-n} \exp\{-ns^2/(2\sigma_0^2)\}}{s^{-n} \exp\{-n/2\}} = \left(\frac{s}{\sigma_0}\right)^n \exp\left\{-\frac{n}{2} \left(\frac{s^2}{\sigma_0^2} - 1\right)\right\} = h(s/\sigma_0), \end{aligned}$$

con $h(t) = t^n \exp\left\{-\frac{n}{2}(t^2 - 1)\right\}$. Comprobamos que h alcanza el máximo en $t = 1$, y es creciente a la izquierda del máximo y decreciente a la derecha. Esto determina que Λ rechaza H_0 cuando s/σ_0 se aleja de 1 (o sea, cuando s se aleja de σ_0). El valor t que maximiza $h(t)$ es el mismo que el que maximiza $\log h(t)$. Trabajamos con el logaritmo. Se tiene que

$$\log h(t) = n \log t - \frac{n}{2}(t^2 - 1) \quad \text{y} \quad \frac{d}{dt} \log h(t) = \frac{n}{t} - nt = n \frac{1 - t^2}{t}.$$

Puesto que $t = s/\sigma_0 > 0$, se tiene que el signo de $\frac{d}{dt} \log h(t)$ es el mismo que el de $1 - t^2$, que es positivo cuando $0 < t < 1$, nulo con $t = 1$, y negativo cuando $t > 1$, con lo que queda comprobado el comportamiento de h descrito.

Por tanto, la rc es de la forma

$$C = \{\Lambda < c\} = \left\{ \frac{s}{\sigma_0} < c_1 \right\} \cup \left\{ \frac{s}{\sigma_0} > c_2 \right\},$$

con $c_1 < 1$ y $c_2 > 1$. No se suelen utilizar los valores exactos de c_1 y c_2 , soluciones de la ecuación $\Lambda = c$, sino que se toman de modo que $P_0\left\{\frac{s}{\sigma_0} < c_1\right\} = P_0\left\{\frac{s}{\sigma_0} > c_2\right\} = \frac{\alpha}{2}$. El teorema de Fisher proporciona la distribución nula de una función de $\frac{s}{\sigma_0}$, lo que permite obtener c_1 y c_2 :

$$\frac{\alpha}{2} = P_0\left\{\frac{s}{\sigma_0} < c_1\right\} = P_0\left\{\frac{ns^2}{\sigma_0^2} < nc_1^2\right\} = P_0\left\{\chi_{n-1}^2 < nc_1^2\right\},$$

y entonces $nc_1^2 = \chi_{n-1, 1-\alpha/2}^2$. De aquí se obtiene $c_1 = \sqrt{\frac{1}{n} \chi_{n-1, 1-\alpha/2}^2}$ y

$$\left\{ \frac{s}{\sigma_0} < c_1 \right\} = \left\{ \frac{s}{\sigma_0} < \sqrt{\frac{1}{n} \chi_{n-1, 1-\alpha/2}^2} \right\} = \left\{ s^2 < \frac{\sigma_0^2}{n} \chi_{n-1, 1-\alpha/2}^2 \right\}.$$

Igualmente se obtiene $c_2 = \sqrt{\frac{1}{n} \chi_{n-1, \alpha/2}^2}$, y de ello la rc con nivel de significación α

$$C = \left\{ s^2 < \frac{\sigma_0^2}{n} \chi_{n-1, 1-\alpha/2}^2 \right\} \cup \left\{ s^2 > \frac{\sigma_0^2}{n} \chi_{n-1, \alpha/2}^2 \right\}.$$

Ejemplo 104. Se lanza un dado 600 veces, obteniéndose los siguientes resultados:

Resultado	1	2	3	4	5	6
Frecuencia	83	119	94	81	101	122

Se pretende estudiar si estos datos son compatibles con un dado equilibrado o, por el contrario, si debemos inclinarnos a pensar que el dado no está equilibrado. Obtenemos el trv y determinamos el p -valor para los datos observados.

Los parámetros son las probabilidades de los distintos resultados, p_i , para $i = 1, \dots, 6$. Sin embargo, los parámetros desconocidos son 5, puesto que suman 1 y uno de ellos se puede obtener a partir de los demás. Si sabemos cuanto valen 5 de ellos conocemos también el restante. Ésto es relevante para establecer la distribución nula del estadístico del contraste y, a partir de ello, para obtener la región crítica concreta.

La hipótesis nula es que el dado está equilibrado, esto es,

$$H_0 : \mathbf{p} = \mathbf{p}_0 ,$$

siendo

$$\mathbf{p} = (p_1, \dots, p_6)$$

el vector de parámetros, fijos pero desconocidos, y $\mathbf{p}_0 = (\frac{1}{6}, \dots, \frac{1}{6})$.

Calculamos el emv de \mathbf{p} . Obtendremos el resultado habitual, estimando probabilidades mediante proporciones observadas. Para $i = 1, \dots, 6$, sea n_i el número de observaciones iguales a i (los datos del enunciado). Se tiene que

$$(N_1, \dots, N_6) \sim \text{mult}(n; p_1, \dots, p_6) ,$$

con $n = 600$, y por tanto la función de verosimilitud es

$$L(\mathbf{p}) = f_{\mathbf{p}}(n_1, \dots, n_6) = P_{\mathbf{p}}\{N_1 = n_1, \dots, N_6 = n_6\} = \frac{n!}{n_1! \dots n_6!} p_1^{n_1} \dots p_6^{n_6} .$$

El vector \mathbf{p} que maximiza $L(\mathbf{p})$ es el mismo que el que maximiza

$$\log L(\mathbf{p}) = \log \left(\frac{n!}{n_1! \dots n_6!} \right) + \sum_{i=1}^6 n_i \log p_i .$$

El máximo está restringido a que $\sum_{i=1}^6 p_i = 1$, y el método más adecuado para obtenerlo es el de los multiplicadores de Lagrange. Consideramos la función

$$\varphi(\mathbf{p}) = \log L(\mathbf{p}) + \lambda \left(\sum_{i=1}^6 p_i - 1 \right) .$$

El máximo se obtiene resolviendo el sistema

$$\begin{aligned} \frac{d}{dp_i} \varphi(\mathbf{p}) &= \frac{n_i}{p_i} + \lambda = 0 , \text{ para } i = 1, \dots, 6, \\ \frac{d}{d\lambda} \varphi(\mathbf{p}) &= \sum_{i=1}^6 p_i - 1 = 0 . \end{aligned}$$

Se obtiene $p_i = \frac{-1}{\lambda} n_i$ y $1 = \sum_{i=1}^6 p_i = \frac{-1}{\lambda} \sum_{i=1}^6 n_i = \frac{-1}{\lambda} n$. Por tanto, $\frac{-1}{\lambda} = \frac{1}{n}$, y entonces la estimación de máxima verosimilitud es

$$\hat{p}_i = \frac{n_i}{n}, \text{ para } i = 1, \dots, 6.$$

Entonces, $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_6)$. Se tiene que $\hat{\mathbf{p}}_0 = \mathbf{p}_0 = (\frac{1}{6}, \dots, \frac{1}{6})$, y el estadístico rv es

$$\begin{aligned} \Lambda &= \frac{f_{\hat{\mathbf{p}}_0}(n_1, \dots, n_6)}{f_{\hat{\mathbf{p}}}(n_1, \dots, n_6)} = \frac{\frac{n!}{n_1! \dots n_6!} \left(\frac{1}{6}\right)^{n_1} \dots \left(\frac{1}{6}\right)^{n_6}}{\frac{n!}{n_1! \dots n_6!} \left(\frac{n_1}{n}\right)^{n_1} \dots \left(\frac{n_6}{n}\right)^{n_6}} \\ &= \frac{\left(\frac{1}{6}\right)^{n_1} \dots \left(\frac{1}{6}\right)^{n_6}}{\left(\frac{n_1}{n}\right)^{n_1} \dots \left(\frac{n_6}{n}\right)^{n_6}} = \frac{1}{6^{n_1 + \dots + n_6}} = \left(\frac{n}{6}\right)^n \frac{1}{n_1^{n_1} \dots n_6^{n_6}}. \end{aligned}$$

El estadístico Λ es función de $\prod_{i=1}^6 n_i^{n_i}$ ó, tomando el logaritmo, de $\sum_{i=1}^6 n_i \log n_i$. El tamaño de muestra es grande y podemos utilizar la distribución asintótica. Se tiene que

$$-2 \log \Lambda = -2n \log(n/6) + 2 \sum_{i=1}^6 n_i \log n_i.$$

Hay $k = 5$ parámetros, y $q = 0$ parámetros bajo H_0 , puesto que es una hipótesis simple. Entonces, la rc $C = \{-2 \log \Lambda > \chi_{k-q, \alpha}^2\}$ es

$$C = \left\{ -2n \log(n/6) + 2 \sum_{i=1}^6 n_i \log n_i > \chi_{5, \alpha}^2 \right\}.$$

Con los datos obtenidos se tiene que $-2 \log \Lambda = 15.23$, y entonces el p -valor es

$$\alpha(x_1, \dots, x_n) = P_0 \{-2 \log \Lambda > 15.23\} \simeq P_0 \{\chi_5^2 > 15.23\} < 0.01.$$

Ejemplo 105. Se lanza un dado 600 veces, obteniéndose los siguientes resultados:

Resultado	1	2	3	4	5	6
Frecuencia	83	119	94	81	101	122

En el ejemplo 104 se ha realizado un contraste con estos datos, obteniéndose fuerte evidencia de que el dado no está equilibrado. Se piensa que ésto es debido a un defecto de construcción del dado, que hace que el material sea más denso cerca del 6. De este modo, el resultado 6 tendría mayor probabilidad que los demás resultados, y el resultado 1, en la cara opuesta al 6, tendría menor probabilidad que los demás resultados. Contrastamos la hipótesis $H_0 : p_1 = p_6$.

Los parámetros son las probabilidades de los distintos resultados, p_i , para $i = 1, \dots, 6$, y $\mathbf{p} = (p_1, \dots, p_6)$ es el vector de parámetros, fijos pero desconocidos. En este caso se tiene que

$$\begin{aligned} \Theta &= \left\{ (p_1, \dots, p_6) : p_1, \dots, p_6 \geq 0, \sum_{i=1}^6 p_i = 1 \right\}, \text{ y} \\ \Theta_0 &= \left\{ (p_1, \dots, p_6) : p_1, \dots, p_6 \geq 0, \sum_{i=1}^6 p_i = 1, p_1 = p_6 \right\}. \end{aligned}$$

Hay $k = 5$ parámetros desconocidos (seis con una restricción). Hay $q = 4$ parámetros desconocidos bajo H_0 , que son p_1, p_2, p_3 y p_4 , con $p_6 = p_1$ y $p_5 = 1 - 2p_1 - p_2 - p_3 - p_4$ (seis con dos restricciones). El emv de \mathbf{p} es (véase el ejemplo 104) $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_6) = (\frac{n_1}{n}, \dots, \frac{n_6}{n})$. Calculamos $\hat{\mathbf{p}}_0$. Se tiene que

$$L(\mathbf{p}) = \frac{n!}{n_1! \dots n_6!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} p_5^{n_5} p_6^{n_6} \stackrel{H_0}{=} \frac{n!}{n_1! \dots n_6!} p_1^{n_1+n_6} p_2^{n_2} p_3^{n_3} p_4^{n_4} p_5^{n_5}.$$

El vector \mathbf{p} que maximiza $L(\mathbf{p})$ bajo H_0 es el mismo que el que maximiza

$$\log L(\mathbf{p}) \stackrel{H_0}{=} \log \left(\frac{n!}{n_1! \dots n_6!} \right) + \sum_{i=1}^5 n_i \log p_i + n_6 \log p_1.$$

El máximo está restringido a que $p_1 + \sum_{i=1}^5 p_i = 1$; utilizamos el método de los multiplicadores de Lagrange. Consideramos la función

$$\varphi(\mathbf{p}) = \log \left(\frac{n!}{n_1! \dots n_6!} \right) + \sum_{i=1}^5 n_i \log p_i + n_6 \log p_1 + \lambda \left(p_1 + \sum_{i=1}^5 p_i - 1 \right).$$

El máximo se obtiene resolviendo el sistema

$$\frac{d}{dp_i} \varphi(\mathbf{p}) = 0 \text{ para } i = 1, \dots, 5, \text{ y } \frac{d}{d\lambda} \varphi(\mathbf{p}) = p_1 + \sum_{i=1}^5 p_i - 1 = 0,$$

que es:

$$\begin{aligned} \frac{n_i}{p_i} + \lambda &= 0, \text{ para } i = 2, 3, 4, 5, \\ \frac{n_1}{p_1} + \frac{n_6}{p_1} + 2\lambda &= 0, \text{ y} \\ p_1 + \sum_{i=1}^5 p_i &= 1. \end{aligned}$$

Se obtiene $p_i = \frac{-1}{\lambda} n_i$ para $i = 2, 3, 4, 5$, $p_1 = \frac{-1}{\lambda} \frac{n_1+n_6}{2}$ y

$$1 = p_1 + \sum_{i=1}^5 p_i = \frac{-1}{\lambda} \sum_{i=1}^6 n_i = \frac{-1}{\lambda} n.$$

Por tanto, $\frac{-1}{\lambda} = \frac{1}{n}$, y entonces la estimación de máxima verosimilitud bajo H_0 es

$$\hat{p}_i = \frac{n_i}{n}, \text{ para } i = 2, 3, 4, 5, \text{ y } \hat{p}_1 = \frac{n_1 + n_6}{2n}.$$

Entonces, $\hat{\mathbf{p}}_0 = \left(\frac{n_1+n_6}{2n}, \frac{n_2}{n}, \frac{n_3}{n}, \frac{n_4}{n}, \frac{n_5}{n}, \frac{n_1+n_6}{2n} \right)$, y el estadístico rv es

$$\begin{aligned} \Lambda &= \frac{f_{\hat{\mathbf{p}}_0}(n_1, \dots, n_6)}{f_{\hat{\mathbf{p}}}(n_1, \dots, n_6)} \\ &= \frac{\frac{n!}{n_1! \dots n_6!} \left(\frac{n_1+n_6}{2n} \right)^{n_1} \left(\frac{n_2}{n} \right)^{n_2} \left(\frac{n_3}{n} \right)^{n_3} \left(\frac{n_4}{n} \right)^{n_4} \left(\frac{n_5}{n} \right)^{n_5} \left(\frac{n_1+n_6}{2n} \right)^{n_6}}{\frac{n!}{n_1! \dots n_6!} \left(\frac{n_1}{n} \right)^{n_1} \dots \left(\frac{n_6}{n} \right)^{n_6}} \\ &= \frac{\left(\frac{n_1+n_6}{2n} \right)^{n_1+n_6}}{\left(\frac{n_1}{n} \right)^{n_1} \left(\frac{n_6}{n} \right)^{n_6}} = 2^{-(n_1+n_6)} \frac{(n_1+n_6)^{n_1+n_6}}{n_1^{n_1} n_6^{n_6}} \\ &= 2^{-(n_1+n_6)} \left(\frac{n_1}{n_1+n_6} \right)^{-n_1} \left(\frac{n_6}{n_1+n_6} \right)^{-n_6} = 2^{-(n_1+n_6)} t^{-n_1} (1-t)^{-n_6}, \end{aligned}$$

con $t = \frac{n_1}{n_1+n_6}$. El tamaño de muestra es grande y podemos utilizar la distribución asintótica. Se tiene que

$$\begin{aligned} -2 \log \Lambda &= 2(n_1+n_6) \log 2 + 2n_1 \log t + 2n_6 \log(1-t) \\ &= 2(n_1+n_6) [\log 2 + t \log t + (1-t) \log(1-t)], \end{aligned}$$

y entonces, la rc $C = \{-2 \log \Lambda > \chi_{k-q, \alpha}^2\}$ es

$$C = \{2(n_1+n_6) [\log 2 + t \log t + (1-t) \log(1-t)] > \chi_{1, \alpha}^2\}.$$

Con los datos obtenidos se tiene que $-2 \log \Lambda = 7.465$, y entonces el p -valor es

$$\begin{aligned} \alpha(x_1, \dots, x_n) &= P_0 \{-2 \log \Lambda > 7.465\} \\ &\simeq P_0 \{\chi_1^2 > 7.465\} < 0.01. \end{aligned}$$

La región crítica en el ejemplo 104 se puede expresar en función de la entropía de Shannon de la distribución que asigna una probabilidad n_i/n al valor i , para $i = 1, \dots, 6$. La región crítica en el ejemplo 105 se puede expresar en función de la entropía de Shannon de la distribución $B(1, t)$.

Ejemplo 106. Consideremos sendas muestras X_1, \dots, X_n e Y_1, \dots, Y_m de dos poblaciones normales independientes. Determinamos el trv para contrastar la igualdad de varianzas poblacionales.

Hay cuatro parámetros desconocidos: μ_1, σ_1, μ_2 y σ_2 . Entonces,

$$\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2).$$

Queremos contrastar $\sigma_1^2 = \sigma_2^2$, que es equivalente a contrastar

$$H_0 : \sigma_1 = \sigma_2 \quad \text{frente a} \quad H_1 : \sigma_1 \neq \sigma_2.$$

La función de verosimilitud tiene que considerar las dos muestras. Puesto que son

independientes, se obtiene

$$\begin{aligned}
 L(\mu_1, \mu_2, \sigma_1, \sigma_2) &= f_{\mu_1, \sigma_1, \mu_2, \sigma_2}(x_1, \dots, x_n, y_1, \dots, y_m) \\
 &= f_{\mu_1, \sigma_1}(x_1, \dots, x_n) \cdot f_{\mu_2, \sigma_2}(y_1, \dots, y_m) \\
 &= (2\pi)^{-n/2} \sigma_1^{-n} \exp \left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^n (x_i - \mu_1)^2 \right\} \cdot (2\pi)^{-m/2} \sigma_2^{-m} \exp \left\{ -\frac{1}{2\sigma_2^2} \sum_{j=1}^m (y_j - \mu_2)^2 \right\} \\
 &= (2\pi)^{-(n+m)/2} \sigma_1^{-n} \sigma_2^{-m} \exp \left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^n (x_i - \mu_1)^2 - \frac{1}{2\sigma_2^2} \sum_{j=1}^m (y_j - \mu_2)^2 \right\}.
 \end{aligned}$$

Puesto que $L = f_{\mu_1, \sigma_1} f_{\mu_2, \sigma_2}$, se tiene que $\log L = \log f_{\mu_1, \sigma_1} + \log f_{\mu_2, \sigma_2}$, y entonces, al derivar $\log L$ respecto de μ_1 y de σ_1 solo resulta involucrada f_{μ_1, σ_1} (la derivada de $\log f_{\mu_2, \sigma_2}$ vale 0). Por tanto, los estimadores de mv de μ_1 y σ_1 son los ya conocidos para una única población normal, e igualmente para los estimadores de μ_2 y σ_2 : se tiene que $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$, con

$$\hat{\mu}_1 = \bar{x}, \hat{\sigma}_1 = s_1, \hat{\mu}_2 = \bar{y}, \hat{\sigma}_2 = s_2.$$

La función de verosimilitud bajo H_0 es

$$\begin{aligned}
 L(\mu_1, \mu_2, \sigma) &= f_{\mu_1, \sigma}(x_1, \dots, x_n) \cdot f_{\mu_2, \sigma}(y_1, \dots, y_m) \\
 &= (2\pi)^{-(n+m)/2} \sigma^{-(n+m)} \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \mu_1)^2 + \sum_{j=1}^m (y_j - \mu_2)^2 \right) \right\}.
 \end{aligned}$$

Obtenemos la emv $\hat{\theta}_0$, y para ello, resolvemos las ecuaciones que resultan de igualar a cero las derivadas del logaritmo de $L(\mu_1, \mu_2, \sigma)$, que son:

$$\frac{dL}{d\mu_1} = -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(x_i - \mu_1)^{2-1} \cdot (-1) = \frac{n}{\sigma^2} (\bar{x} - \mu_1) = 0, \quad (5.7)$$

$$\frac{dL}{d\mu_2} = \frac{m}{\sigma^2} (\bar{y} - \mu_2) = 0, \quad (5.8)$$

$$\begin{aligned}
 \frac{dL}{d\sigma} &= -\frac{n+m}{\sigma} - \frac{1}{2}(-2)\sigma^{-3} \left(\sum_{i=1}^n (x_i - \mu_1)^2 + \sum_{j=1}^m (y_j - \mu_2)^2 \right) \\
 &= \frac{1}{\sigma} \left(-(n+m) + \frac{1}{\sigma^2} \left(\sum_{i=1}^n (x_i - \mu_1)^2 + \sum_{j=1}^m (y_j - \mu_2)^2 \right) \right) = 0. \quad (5.9)
 \end{aligned}$$

De (5.7) y (5.8) se obtiene $\mu_1 = \bar{x}$ y $\mu_2 = \bar{y}$. Sustituyendo estos valores en (5.9) y despejando, se obtiene

$$\sigma^2 = \frac{ns_1^2 + ms_2^2}{n+m} = \frac{n}{n+m} s_1^2 + \frac{m}{n+m} s_2^2.$$

Entonces, $\hat{\sigma}_0 = \sqrt{\frac{ns_1^2 + ms_2^2}{n+m}}$ y $\hat{\theta}_0 = (\bar{x}, \bar{y}, \hat{\sigma}_0)$.

El estadístico rv es

$$\begin{aligned} \Lambda &= \frac{f_{\hat{\theta}_0}(x_1, \dots, x_n, y_1, \dots, y_m)}{f_{\hat{\theta}}(x_1, \dots, x_n, y_1, \dots, y_m)} \\ &= \frac{(2\pi)^{-(n+m)/2} \hat{\sigma}_0^{-(n+m)} \exp\left\{-\frac{1}{2\hat{\sigma}_0^2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{j=1}^m (y_j - \bar{y})^2\right)\right\}}{(2\pi)^{-(n+m)/2} s_1^{-n} s_2^{-m} \exp\left\{-\frac{1}{2s_1^2} \sum_{i=1}^n (x_i - \bar{x})^2 - \frac{1}{2s_2^2} \sum_{j=1}^m (y_j - \bar{y})^2\right\}} \\ &= \frac{\hat{\sigma}_0^{-(n+m)} \exp\left\{-(n+m)/2\right\}}{s_1^{-n} s_2^{-m} \exp\left\{-\frac{n}{2} - \frac{m}{2}\right\}} = \frac{\hat{\sigma}_0^{-(n+m)}}{s_1^{-n} s_2^{-m}} = \left(\frac{\hat{\sigma}_0^2}{s_1^2}\right)^{-n/2} \left(\frac{\hat{\sigma}_0^2}{s_2^2}\right)^{-m/2} \\ &= \left(\frac{n + ms_2^2/s_1^2}{n+m}\right)^{-n/2} \left(\frac{ns_1^2/s_2^2 + m}{n+m}\right)^{-m/2} = (n+m)^{(n+m)/2} h(s_1^2/s_2^2), \end{aligned}$$

con

$$\begin{aligned} h(t) &= \left(n + \frac{m}{t}\right)^{-n/2} (nt + m)^{-m/2} = \left(\frac{nt + m}{t}\right)^{-n/2} (nt + m)^{-m/2} \\ &= t^{n/2} (nt + m)^{-(n+m)/2}. \end{aligned}$$

Hemos expresado Λ en función de s_1^2/s_2^2 , que es por tanto el estadístico del contraste. El comportamiento de la función h determina para qué valores de s_1^2/s_2^2 se tiene que $\Lambda < c$. La función h es creciente para $0 < t < 1$ y decreciente para $t > 1$: derivando y simplificando se obtiene $\frac{d}{dt} \log h(t) = \frac{nm(1-t)}{2t(nt+m)}$, que es positiva para $0 < t < 1$ y negativa para $t > 1$. De aquí se obtiene que el test rechaza la igualdad de varianzas cuando s_1^2/s_2^2 se aleja del 1:

$$C = \left\{ \frac{s_1^2}{s_2^2} < c_1 \right\} \cup \left\{ \frac{s_1^2}{s_2^2} > c_2 \right\}, \quad (5.10)$$

con $c_1 < 1$ y $c_2 > 1$.

Dado un nivel de significación α , no se suelen obtener los valores exactos de c_1 y c_2 , resolviendo la ecuación $\Lambda = c$ (con $c = c(\alpha)$). Se suelen tomar de modo que la probabilidad nula de cada una de las dos partes en (5.10) vale $\alpha/2$. Éste es justamente el contraste que se presenta en las tablas de contrastes en poblaciones normales (dos poblaciones normales independientes, contrastes para las varianzas con esperanzas desconocidas).

Parece un esfuerzo excesivo el realizado para obtener finalmente este test. Podríamos haberlo obtenido simplemente considerando los estadísticos ms S_1^2 y S_2^2 , y teniendo en cuenta que la distribución nula del cociente es conocida, es F de Snedecor.

Sin embargo, lo que motiva el esfuerzo es que el procedimiento puede ser fácilmente generalizado al caso de más de dos poblaciones normales. El resultado, sin comprobación, se presenta en el siguiente ejemplo.

Ejemplo 107. Consideremos muestras de tamaños n_1, \dots, n_k de k poblaciones normales independientes, y sea $n = \sum_{i=1}^k n_i$. Determinamos el trv para contrastar la igualdad de

varianzas poblacionales.

Hay $2k$ parámetros desconocidos, las esperanzas y varianzas. Queremos contrastar

$$H_0 : \sigma_1 = \sigma_2 = \dots = \sigma_k .$$

Bajo H_0 hay $k + 1$ parámetros desconocidos, las esperanzas y el valor común de la varianza.

Se obtiene

$$\Lambda = \frac{s_1^{n_1} \cdots s_k^{n_k}}{\widehat{\sigma}_0^n} , \text{ con } \widehat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^k n_i s_i^2 .$$

Teniendo en cuenta que $-2 \log \Lambda \stackrel{H_0}{\sim} \chi_{2k-(k+1)}^2$ se obtiene la rc

$$C = \left\{ 2n \log \widehat{\sigma}_0 - 2 \sum_{i=1}^k n_i \log s_i > \chi_{k-1, \alpha}^2 \right\} ,$$

correspondiente a un nivel de significación α .

En los ejemplos hemos considerado hipótesis nulas expresadas en términos de igualdades. Los contrastes unilaterales, y en general los que expresan la hipótesis nula utilizando \leq ó \geq , requieren de un trabajo adicional para la obtención de $\widehat{\theta}_0$.

Todos los contrastes para poblaciones normales presentados en las tablas del siguiente apartado se pueden obtener por el método de la razón de verosimilitudes. Son óptimos, con la excepción de las probabilidades $\alpha/2$ y $1 - \alpha/2$ que se consideran en los extremos para la χ^2 y la F de Snedecor (en los contrastes bilaterales que las involucran, por ejemplo $H_0 : \sigma = \sigma_0$). Aunque esta elección no dé exactamente lugar al contraste óptimo, es la práctica habitual, pero puede ser mejorada mediante procedimientos informáticos, sobre todo para muestras pequeñas.

5.6 Contrastes para poblaciones normales

5.6.1 Distribución teórica $N(\mu, \sigma)$ y muestra de tamaño n

Contrastes para μ

	σ conocido	σ desconocido
H_0	Se rechaza H_0 , con nivel α , si	
$\mu \leq \mu_0$	$\bar{x} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$	$\bar{x} > \mu_0 + \frac{S}{\sqrt{n}} t_{n-1, \alpha}$
$\mu \geq \mu_0$	$\bar{x} < \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$	$\bar{x} < \mu_0 - \frac{S}{\sqrt{n}} t_{n-1, \alpha}$
$\mu = \mu_0$	$ \bar{x} - \mu_0 > \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$	$ \bar{x} - \mu_0 > \frac{S}{\sqrt{n}} t_{n-1, \alpha/2}$

Contrastes para σ

$$T = \sum_{i=1}^n (x_i - \mu)^2$$

	μ conocido	μ desconocido
H_0	Se rechaza H_0 , con nivel α , si	
$\sigma \leq \sigma_0$	$T > \sigma_0^2 \chi_{n,\alpha}^2$	$s^2 > \frac{\sigma_0^2}{n} \chi_{n-1,\alpha}^2$
$\sigma \geq \sigma_0$	$T < \sigma_0^2 \chi_{n,1-\alpha}^2$	$s^2 < \frac{\sigma_0^2}{n} \chi_{n-1,1-\alpha}^2$
$\sigma = \sigma_0$	$T < \sigma_0^2 \chi_{n,1-\alpha/2}^2$ ó $T > \sigma_0^2 \chi_{n,\alpha/2}^2$	$s^2 < \frac{\sigma_0^2}{n} \chi_{n-1,1-\alpha/2}^2$ ó $s^2 > \frac{\sigma_0^2}{n} \chi_{n-1,\alpha/2}^2$

5.6.2 Dos poblaciones normales independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ y muestras de tamaños n y m respectivamente

Consideramos

una m.a.s. X_1, \dots, X_n de una $N(\mu_1, \sigma_1)$ y
una m.a.s. Y_1, \dots, Y_m de una $N(\mu_2, \sigma_2)$,

siendo las X_i independientes de las Y_j . Por ejemplo, utilizando los fertilizantes A ó B la producción de cierto cereal es $N(\mu_A, \sigma_A)$ o $N(\mu_B, \sigma_B)$ respectivamente. Un estudio de $\mu_A - \mu_B$ basado en $n + m$ cultivos en un vivero, n con A y m con B, nos permite comparar el rendimiento de ambos fertilizantes.

Denotamos por \bar{X} , s_1^2 y S_1^2 a la media, varianza y cuasivarianza de las X_i , y por \bar{Y} , s_2^2 y S_2^2 a las de las Y_j .

Contrastes para $\mu_1 - \mu_2$

δ es una constante dada ($\delta = 0$ habitualmente) y $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$

	σ_1 y σ_2 conocidos	$\sigma_1 = \sigma_2$, desconocido
H_0	Se rechaza H_0 , con nivel α , si	
$\mu_1 - \mu_2 \leq \delta$	$\bar{x} - \bar{y} > \delta + z_\alpha \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$	$\bar{x} - \bar{y} > \delta + t_{n+m-2,\alpha} \cdot S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$
$\mu_1 - \mu_2 \geq \delta$	$\bar{x} - \bar{y} < \delta - z_\alpha \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$	$\bar{x} - \bar{y} < \delta - t_{n+m-2,\alpha} \cdot S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$
$\mu_1 - \mu_2 = \delta$	$ \bar{x} - \bar{y} - \delta > z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$	$ \bar{x} - \bar{y} - \delta > t_{n+m-2,\alpha/2} \cdot S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$

Contrastes para σ_1/σ_2 ($\leq 1, \geq 1$ e $= 1$)

$$T = \frac{\sum_{i=1}^n (x_i - \mu_1)^2}{\sum_{j=1}^m (y_j - \mu_2)^2}$$

	μ_1 y μ_2 conocidos	μ_1 y μ_2 desconocidos
H_0	Se rechaza H_0 , con nivel α , si	
$\sigma_1 \leq \sigma_2$	$\frac{m}{n}T > F_{n,m,\alpha}$	$\frac{S_1^2}{S_2^2} > F_{n-1,m-1,\alpha}$
$\sigma_1 \geq \sigma_2$	$\frac{m}{n}T < F_{n,m,1-\alpha}$	$\frac{S_1^2}{S_2^2} < F_{n-1,m-1,1-\alpha}$
$\sigma_1 = \sigma_2$	$\frac{m}{n}T < F_{n,m,1-\alpha/2}$ ó $> F_{n,m,\alpha/2}$	$\frac{S_1^2}{S_2^2} < F_{n-1,m-1,1-\alpha/2}$ ó $> F_{n-1,m-1,\alpha/2}$

5.6.3 Una muestra pareada de tamaño n

Consideramos una m.a.s. $(X_1, Y_1), \dots, (X_n, Y_n)$ de una $N(\mu, \Sigma)$, donde $\mu = (\mu_1, \mu_2)'$ es el vector esperanzas y $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ es la matriz de covarianzas, con $\sigma_{12} = \rho\sigma_1\sigma_2$.

Los datos son bidimensionales. Por ejemplo, el nivel de colesterol de n pacientes antes y después de un tratamiento médico. Si $\mu_1 - \mu_2 > 0$ el tratamiento ha reducido el nivel.

Contrastes para $\mu_1 - \mu_2$

Sea $Z_i = X_i - Y_i$ para $i = 1, \dots, n$. Se tiene que Z_1, \dots, Z_n son vaaid con distribución normal unidimensional, con

$$\mu_Z = E[Z_i] = \mu_1 - \mu_2, \quad \sigma_Z^2 = V[Z_i] = \sigma_1^2 + (-1)^2 \sigma_2^2 - 2\sigma_{12},$$

y se tiene que $\bar{Z} = \bar{X} - \bar{Y}$, también con distribución normal.

Los contrastes $H_0 : \mu_1 - \mu_2 \leq 0, \geq 0$ e $= 0$ (o δ en vez de 0), se reducen a los contrastes $H_0 : \mu_Z \leq 0, \geq 0$ e $= 0$, estudiados en el apartado 5.6.1. Para σ_Z^2 desconocida no hace falta estimar los elementos σ_1^2, σ_2^2 y σ_{12} individualmente.

Contraste de independencia

Si las componentes de (X_i, Y_i) son independientes entonces son incorreladas, esto es, se tiene que $\rho = 0$, siendo ρ el coeficiente de correlación poblacional,

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

El coeficiente de correlación muestral es

$$R = \frac{s_{12}}{s_1 s_2},$$

siendo

$$s_{12} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}$$

la covarianza muestral, y s_1, s_2 la raíz cuadrada de las varianzas muestrales. El estadístico R es un estimador de ρ . Por tanto, valores alejados de 0 para el coeficiente muestral aportan evidencia en contra de la independencia, obteniéndose así una región crítica de la forma $C = \{R < c_1\} \cup \{R > c_2\}$, con $c_1 < 0$ y $c_2 > 0$.

En general, la distribución de R depende de la distribución teórica, que es desconocida, lo que impide obtener la región crítica concreta (obtener c_1 y c_2) para un nivel de significación dado. Sin embargo, tanto ρ como R son invariantes por cambios de posición y cambios de escala en ambas componentes. Puesto que cualquier distribución normal se puede transformar en cualquier otra distribución normal mediante un cambio adecuado de posición y escala (transformación lineal), entonces, en el caso normal, si $\rho = 0$ (lo que implica la independencia) la distribución de R no depende de los parámetros de las dos componentes normales (que son independientes). De este modo, R sirve como estadístico para contrastar la independencia en este caso normal.

Si la distribución teórica es normal y $\rho = 0$ se verifica que

$$\sqrt{n-2} \frac{R}{\sqrt{1-R^2}} \sim t_{n-2} .$$

Consideramos entonces la región crítica con nivel de significación α

$$C = \left\{ \left| \sqrt{n-2} \frac{R}{\sqrt{1-R^2}} \right| > t_{n-2, \alpha/2} \right\} .$$

Operando, esta rc se puede expresar como

$$C = \left\{ |R| > \left(\frac{n-2}{t_{n-2, \alpha/2}^2} + 1 \right)^{-1/2} \right\} .$$

Por ejemplo, con $\alpha = 0.01$, se obtiene para $n = 10$ la rc $C = \{|R| > 0.76\}$, para $n = 20$ la rc $C = \{|R| > 0.56\}$, para $n = 50$ la rc $C = \{|R| > 0.36\}$, y para $n = 500$ la rc $C = \{|R| > 0.12\}$.

Puesto que en este caso normal la independencia es equivalente a la incorrelación, la aceptación de H_0 supone una aceptación de la independencia.

Ejemplo 108. Se obtuvo la siguiente muestra de tamaño 6 de una distribución teórica normal bidimensional:

$(7.3, 2.4)$, $(6.8, 2.1)$, $(7.8, 1.7)$, $(7.4, 2.3)$, $(6.5, 1.8)$, $(7.6, 2.2)$.

Contrastamos la independencia entre las componentes de la distribución teórica.

Se obtiene $\bar{x} = 7.23$, $\bar{y} = 2.08$, $s_1 = 0.450$ y $s_2 = 0.254$. Además, se obtiene $\sum_{i=1}^n x_i y_i = 90.5$, y de aquí,

$$s_{12} = \frac{1}{n} \sum_{i=1}^6 x_i y_i - \bar{x}\bar{y} = 1.70 \cdot 10^{-2} \quad \text{y} \quad R = \frac{s_{12}}{s_1 s_2} = 0.149 .$$

Se tiene que $\sqrt{n-2} \frac{R}{\sqrt{1-R^2}} = 0.300$, y entonces la rc en la cual nuestra observación está al borde es $C = \left\{ \left| \sqrt{6-2} \frac{R}{\sqrt{1-R^2}} \right| > 0.300 \right\}$, y el p -valor es

$$\begin{aligned} \alpha(x_i, y_i) &= P_0 \left\{ \left| \frac{2R}{\sqrt{1-R^2}} \right| > 0.300 \right\} = P \{|t_{6-2}| > 0.300\} \\ &= 2 \cdot P \{t_{6-2} > 0.300\} > 2 \cdot 0.3 = 0.6 . \end{aligned}$$

Puesto que $0.6 > 0.2$, se acepta la independencia.

BIBLIOGRAFÍA

Dudewicz, E.J. y Mishra, S.N., *Modern Mathematical Statistics*, Wiley, New York, 1988

Gómez Villegas, M.A., *Inferencia Estadística*, Díaz de Santos, Madrid, 2005

Rohatgi, V.K. y Ehsanes Saleh, A. K., *An Introduction to Probability and Statistics*, Wiley, New York, 2000

Vélez Ibarrola, R. y García Pérez, A., *Principios de Inferencia Estadística*, UNED, Madrid, 2012

Tema 6

Estimación por intervalos I. Introducción. Métodos para encontrar estimadores de intervalos: inversión del estadístico de un test, cantidades pivota-les, intervalos de confianza para estadísticos de orden, intervalos bayesianos.

6.1 Intervalos de confianza. Introducción

Un intervalo de confianza es una técnica de estimación paramétrica utilizada en inferencia estadística que permite delimitar un par de valores, dentro de los cuales se encontrará el parámetro desconocido con una confianza alta elegida por el investigador.

Definición 1. Sea X_1, X_2, \dots, X_n una muestra aleatoria simple (m.a.s.) de $X \sim f(x, \theta)$, siendo θ un parámetro desconocido. Llamaremos intervalo de confianza de nivel $1 - \alpha$ para θ a (U, V) tal que

$$p(U < \theta < V) = 1 - \alpha \quad (6.1)$$

Observaciones:

- U y V son estadísticos (dependen de la muestra y sólo de la muestra). En la práctica, son estimadores por defecto y por exceso de θ , de manera que U infraestima θ y V lo sobrevalora.
- El nivel de confianza lo fija el investigador. Por defecto es 0,95. Otros valores muy utilizados son 0,9 y 0,99. Al incrementar el nivel de confianza, el intervalo se hace más largo, perdiéndose así precisión.
- La interpretación del concepto de intervalo es frecuentista en el siguiente sentido. Por ejemplo, si se dispone de 100 realizaciones muestrales, habría 100 intervalos numéricos, $(U, V), (U^1, V^1), \dots, (U^{99}, V^{99})$.

Si $1 - \alpha = 0,95$, al menos 95 de esos 100 intervalos contendrán el verdadero valor de θ .

Se dispone de dos métodos para obtener intervalos de confianza:

1. Método de la función pivote (o de la cantidad pivotal).
2. Método de Neyman.
3. Método bayesiano.

6.2 Método de la función pivote o cantidad pivotal

Sea X_1, X_2, \dots, X_n una m.a.s. de $X \sim f(x, \theta)$, siendo θ un parámetro desconocido. Se desea obtener un intervalo de confianza para θ de nivel $1 - \alpha$. Supongamos que existe una función de la muestra y de θ cuya distribución no depende de θ . A esta función se le denomina **función pivote**, lo que da nombre al método. Sea $h(X_1, X_2, \dots, X_n, \theta)$ dicha función.

Existirán constantes K_1 y K_2 (que no dependerán de θ porque la distribución de $h(X_1, X_2, \dots, X_n, \theta)$ no depende de θ), tales que:

$$p(K_1 \leq h(X_1, X_2, \dots, X_n, \theta) \leq K_2) = 1 - \alpha$$

Despejando θ en dichas inecuaciones, se obtienen los límites de confianza para θ .

Este método es especialmente apropiado para obtener intervalos de confianza para parámetros desconocidos en poblaciones normales. Analicemos estas situaciones.

6.2.1 Intervalo de confianza para la media de una distribución normal con varianza conocida

Sea X_1, X_2, \dots, X_n una m.a.s. de $N(\mu, \sigma_0)$, siendo la media μ desconocida y la desviación típica σ_0 conocida. Se desea obtener un intervalo de confianza para μ de nivel $1 - \alpha$. Sabemos por el Teorema de Fisher que $\bar{X} \sim N\left(\mu, \frac{\sigma_0}{\sqrt{n}}\right)$. Por tanto, $\frac{\bar{X} - \mu}{\sigma_0} \sqrt{n}$ es una función pivote para μ , ya que depende de la muestra y de μ , pero su distribución no depende de μ .

Por tanto, existe $z_{\frac{\alpha}{2}}$ tal que

$$p\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma_0} \sqrt{n} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

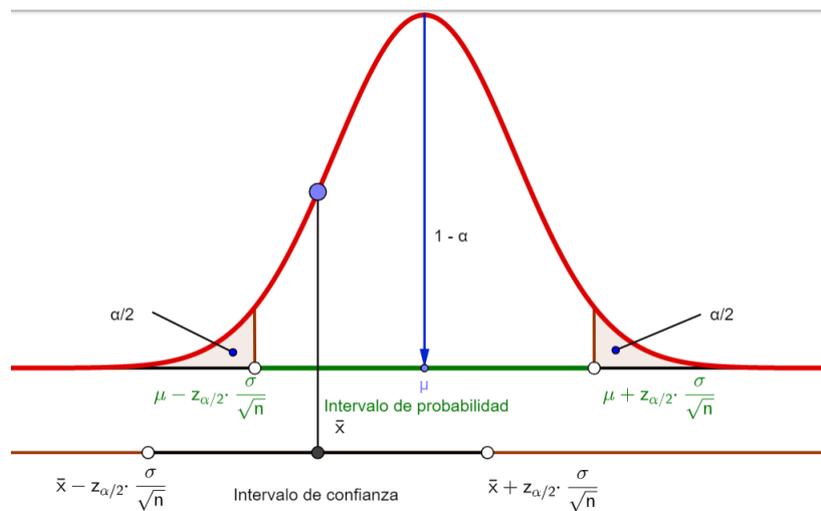
Entonces

$$-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma_0} \sqrt{n} \iff \bar{X} - \mu \geq -z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} \iff \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}$$

y

$$\frac{\bar{X} - \mu}{\sigma_0} \sqrt{n} \leq z_{\frac{\alpha}{2}} \iff \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} \leq \mu.$$

Por tanto, el intervalo de confianza, de nivel $1 - \alpha$, es $\left(\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}\right)$.

Figura 6.1: Intervalo de confianza $1 - \alpha$.

Como se puede observar, se han escogido los valores $-z_{\frac{\alpha}{2}}$ y $z_{\frac{\alpha}{2}}$ como valores críticos (simétricos) en la distribución $N(0, 1)$. Esta elección podría parecer inocente pero no lo es. Cualquier otra elección de valores críticos no simétricos, pero que dejen entre ambas colas de la $N(0, 1)$ una masa de α , garantizaría igualmente que el intervalo tenga nivel de confianza $1 - \alpha$. Cálculos sumamente engorrosos que tratamos de evitar aquí, demuestran que la elección de valores simétricos $z_{\frac{\alpha}{2}}$ y $-z_{\frac{\alpha}{2}}$ garantiza que la longitud del intervalo obtenido es mínima. Esta argumentación se extiende a todos los intervalos que se obtendrán a continuación a partir del método de la función pivote para la media de una normal o para la diferencia de medias en poblaciones normales. Para lectores que deseen conocer la metodología matemática según la cual se prueba que la longitud del intervalo es mínima, les remitimos a los ejemplos de intervalos de confianza obtenidos posteriormente a partir del método de Neyman, donde se ilustran estos cálculos, y al Tema 7.

Ejemplo 6.1:

Un proveedor de latas de sardinas afirma en el envase que el peso medio de las latas es de 100 gramos. Elegidas 36 latas al azar, el peso medio de la muestra ha resultado ser 96 gramos. Si se supone normalidad en el peso y varianza teórica (poblacional) igual a 9 gramos, obtener un intervalo del 95 % y otro del 99 % para la media poblacional, μ , con σ_0 conocido.

- Si $1 - \alpha = 0,95$, entonces $\alpha = 0,05$, con lo que $\frac{\alpha}{2} = 0,025$.

El intervalo de confianza numérico será $\left(96 \pm 1,96 \frac{3}{\sqrt{36}}\right) = (95,01, 96,98)$

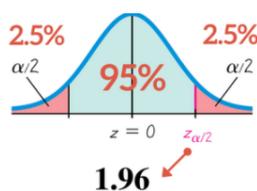


Figura 6.2: Intervalo al 95 % de confianza.

- Si $1 - \alpha = 0,99$, entonces $\alpha = 0,01$, con lo que $\frac{\alpha}{2} = 0,005$.

El intervalo de confianza numérico será $\left(96 \pm 2,57 \frac{3}{\sqrt{36}}\right) = (94,715, 97,528)$.

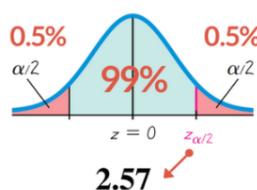


Figura 6.3: Intervalo al 99 % de confianza.

Se puede observar que a mayor confianza, menor es la precisión: el intervalo al 99 % de confianza es más ancho que el intervalo al 95 %.

6.2.2 Intervalo de confianza para la media de una normal con varianza desconocida

Sea X_1, X_2, \dots, X_n una m.a.s. de $N(\mu, \sigma)$, con μ y σ desconocidas. Se desea obtener un intervalo de nivel $1 - \alpha$ para μ .

Por el teorema de Fisher se tiene que:

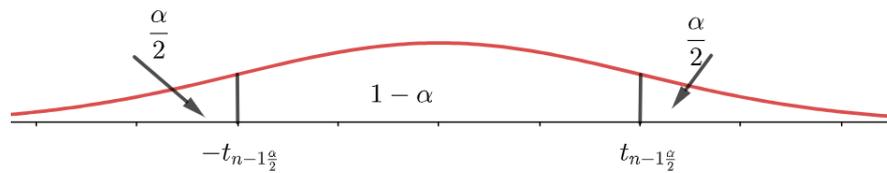
- $\frac{\bar{X} - \mu}{\sigma} \sqrt{n} \rightsquigarrow N(0, 1)$.
- $\sqrt{\frac{nS^2}{\sigma^2(n-1)}} \rightsquigarrow \sqrt{\frac{\chi_{n-1}^2}{n-2}}$.

Y entonces, como en la expresión siguiente el numerador y el denominador son independientes,

$$\frac{\frac{\bar{X} - \mu}{\sigma} \sqrt{n}}{\sqrt{\frac{nS^2}{\sigma^2(n-1)}}} = \frac{\bar{X} - \mu}{S} \sqrt{n-1} \rightsquigarrow t_{n-1} \quad (6.2)$$

Esta es la función pivote. Existe $t_{n-1, \frac{\alpha}{2}}$ tal que

$$P\left(-t_{n-1, \frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{S} \sqrt{n-1} \leq t_{n-1, \frac{\alpha}{2}}\right) = 1 - \alpha$$



Despejando, tendríamos

$$-t_{n-1, \frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{S} \sqrt{n-1} \iff \mu \leq \bar{X} + \frac{t_{n-1, \frac{\alpha}{2}} S}{\sqrt{n-1}}$$

$$\frac{\bar{X} - \mu}{S} \sqrt{n-1} \leq t_{n-1, \frac{\alpha}{2}} \iff \mu \geq \bar{X} - \frac{t_{n-1, \frac{\alpha}{2}} S}{\sqrt{n-1}}$$

Por lo tanto, el intervalo resultante es:

$$\left(\bar{X} \pm \frac{t_{n-1, \frac{\alpha}{2}} S}{\sqrt{n-1}} \right) \quad (6.3)$$

Si se escribe el intervalo en términos de \hat{S} , como

$$S^2 = \frac{n-1}{n} \hat{S}^2 \implies S = \frac{\sqrt{n-1}}{\sqrt{n}} \hat{S}$$

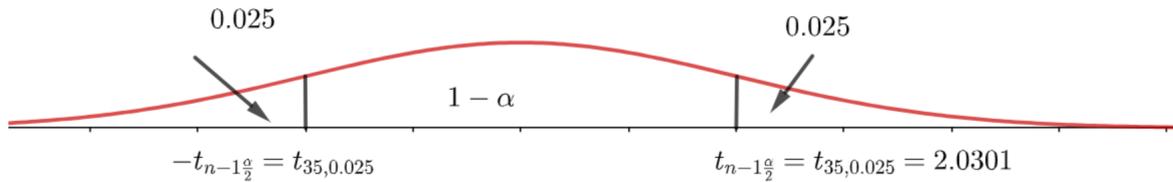
luego el intervalo es

$$\left(\bar{X} \pm t_{n-1, \frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \right). \quad (6.4)$$

Ejemplo 6.2:

Si en el ejemplo anterior se elimina el conocimiento sobre la varianza, el intervalo al 95% suponiendo que la cuasi-varianza del peso de las 36 latas es 16, se calcula como sigue.

Sabiendo que $\mu \in \left(\bar{X} \pm t_{n-1, \frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \right)$, y que $1 - \alpha = 0,95 \implies \alpha = 0,05 \implies \frac{\alpha}{2} = 0,025$, por lo que el intervalo al 95% será $\left(96 \pm 2,03 \frac{4}{\sqrt{36}} \right)$.

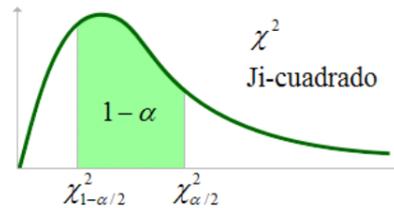


Si se quiere calcular el intervalo de confianza al 99%, se tendría que $t_{35, 0.005} = 2,7281$, con lo que el intervalo será $\left(96 \pm 2,7281 \frac{4}{\sqrt{36}}\right)$.

6.2.3 Intervalo de confianza para la varianza de una normal conocida la media

Supongamos que X_1, X_2, \dots, X_n es una m.a.s. de $N(\mu_0, \sigma)$, μ_0 conocida y σ desconocida. Se desea calcular un intervalo de nivel de confianza $1 - \alpha$ para σ^2 .

$$\sum_{i=1}^n \left(\frac{X_i - \mu_0}{\sigma} \right)^2 \sim \chi_n^2 \iff \sum_{i=1}^n \frac{(X_i - \mu_0)^2}{\sigma^2} \sim \chi_n^2 \quad (6.5)$$



Así se obtiene una función pivote para σ^2 , luego existen $\chi_{n, \frac{\alpha}{2}}^2$ y $\chi_{n, 1 - \frac{\alpha}{2}}^2$, tales que

$$P \left(\chi_{n, 1 - \frac{\alpha}{2}}^2 \leq \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma^2} \leq \chi_{n, \frac{\alpha}{2}}^2 \right) = 1 - \alpha.$$

Despejando,

$$\chi_{n, 1 - \frac{\alpha}{2}}^2 \leq \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma^2} \iff \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi_{n, 1 - \frac{\alpha}{2}}^2}$$

$$\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma^2} \leq \chi_{n, \frac{\alpha}{2}}^2 \iff \sigma^2 \geq \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi_{n, \frac{\alpha}{2}}^2}$$

luego el intervalo de nivel de confianza $1 - \alpha$ para σ^2 es

$$\left(\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi_{n, \frac{\alpha}{2}}^2}, \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi_{n, 1 - \frac{\alpha}{2}}^2} \right) \quad (6.6)$$

Lógicamente, para obtener un intervalo de nivel $1 - \alpha$ para σ , es suficiente con calcular la raíz cuadrada en los límites de confianza anteriores.

Ejemplo 6.3:

Supongamos que 1, 2, 3 es una realización muestral de una $N(\mu_0, \sigma)$, con $\mu_0 = 2$. Obtener un intervalo al 95 % para σ^2 .

$$\sum_{i=1}^n (X_i - \mu_0)^2 = (1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2 = 2.$$

Por tanto, siendo $\chi_{3, 0, 975}^2 = 0,2157$ y $\chi_{3, 0, 025}^2 = 9,3484$ el intervalo de confianza al 95 % es $\left(\frac{2}{9,3484}, \frac{2}{0,2157} \right)$.

6.2.4 Intervalo de confianza para la varianza de una normal. Media desconocida

Supongamos que X_1, X_2, \dots, X_n es una m.a.s. de $N(\mu, \sigma)$, con μ y σ desconocidas. Se desea un intervalo de nivel $1 - \alpha$ para σ^2 .

Por el teorema de Fisher, $\frac{nS^2}{\sigma^2} \sim \chi_{n-1}^2$, que es entonces función pivote para σ^2 . Por tanto, existen $\chi_{n-1, \frac{\alpha}{2}}^2$ y $\chi_{n-1, 1 - \frac{\alpha}{2}}^2$ tales que

$$P \left(\chi_{n-1, 1 - \frac{\alpha}{2}}^2 \leq \frac{nS^2}{\sigma^2} \leq \chi_{n-1, \frac{\alpha}{2}}^2 \right) = 1 - \alpha.$$

Despejando,

$$\chi_{n-1, 1 - \frac{\alpha}{2}}^2 \leq \frac{nS^2}{\sigma^2} \iff \sigma^2 \leq \frac{nS^2}{\chi_{n-1, 1 - \frac{\alpha}{2}}^2}$$

$$\frac{nS^2}{\sigma^2} \leq \chi_{n-1, \frac{\alpha}{2}}^2 \iff \sigma^2 \geq \frac{nS^2}{\chi_{n-1, \frac{\alpha}{2}}^2}.$$

Por tanto, el intervalo de nivel $1 - \alpha$ es

$$\left(\frac{nS^2}{\chi_{n-1, \frac{\alpha}{2}}^2}, \frac{nS^2}{\chi_{n-1, 1 - \frac{\alpha}{2}}^2} \right) \quad (6.7)$$

En realidad, la diferencia entre el caso anterior y éste es que si μ no es conocido, se estima con \bar{X} , con lo que se pierde un grado de libertad.

6.2.5 Intervalo de confianza para la diferencia de medias en poblaciones normales. Varianzas conocidas

Sea X_1, X_2, \dots, X_{n_1} una m.a.s. de $N(\mu_1, \sigma_1)$, con μ_1 desconocida y σ_1 conocida. A su vez, sea Y_1, Y_2, \dots, Y_{n_2} una m.a.s. de $N(\mu_2, \sigma_2)$, con μ_2 desconocida y σ_2 conocida. Además, se supone que ambas muestras son independientes. Se desea obtener un intervalo de confianza de nivel $1 - \alpha$ para $\mu_1 - \mu_2$. Sabemos que $\bar{X} \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$ y que $\bar{Y} \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$, y que ambas variables son independientes, puesto que las muestras correspondientes lo son. Entonces,

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \iff \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \quad (6.8)$$

Por tanto, existe $z_{\frac{\alpha}{2}}$ tal que

$$p\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha \quad (6.9)$$

Despejando en las inecuaciones:

$$-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \iff \mu_1 - \mu_2 \leq \bar{X} - \bar{Y} + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$z_{\frac{\alpha}{2}} \geq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \iff \mu_1 - \mu_2 \geq \bar{X} - \bar{Y} - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

luego el intervalo de nivel $1 - \alpha$ para $\mu_1 - \mu_2$ es

$$\left(\bar{X} - \bar{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \quad (6.10)$$

Ejemplo 6.4:

En el año 1980 la altura media de 1000 reclutas fue 1,7m, y en el año 1990 al altura media de 1021 reclutas fue de 1,73. Si se supone que las estaturas estaban normalmente distribuidas con varianzas respectivas de 15 y 17 cm^2 . Obtener un intervalo de confianza al 95% para $\mu_1 - \mu_2$.

Puesto que $1 - \alpha = 0,95$, $\alpha = 0,05$ y $\frac{\alpha}{2} = 0,025$, el intervalo será

$$\left(170 - 173 \pm 1,96 \sqrt{\frac{15}{1000} + \frac{17}{1021}}\right) = (-3 \pm 0,35) = (-3,35, -2,65)$$

Puesto que $\mu_1 - \mu_2 = 0$ no pertenece al intervalo al 95 %, existe evidencia estadística de que la estatura media se ha modificado. Como además el intervalo tiene ambos extremos negativos, podemos concluir que $\mu_1 - \mu_2 < 0$, y por tanto, la estatura media (poblacional) de los reclutas se ha incrementado en esos 10 años.

6.2.6 Intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas desconocidas pero iguales

Sea X_1, X_2, \dots, X_{n_1} una m.a.s. de $N(\mu_1, \sigma)$, e Y_1, Y_2, \dots, Y_{n_2} una m.a.s. de $N(\mu_2, \sigma)$, ambas muestras independientes, y μ_1, μ_2, σ desconocidos. Se desea obtener un intervalo de confianza para $\mu_1 - \mu_2$ de nivel $1 - \alpha$.

Sabemos que $\frac{n_1 S_X^2}{\sigma^2} \sim \chi_{n_1-1}^2$, y que $\frac{n_2 S_Y^2}{\sigma^2} \sim \chi_{n_2-1}^2$, ambas independientes. Entonces, $\frac{n_1 S_X^2}{\sigma^2} + \frac{n_2 S_Y^2}{\sigma^2} \sim \chi_{n_1-1}^2 + \chi_{n_2-1}^2 \equiv \Gamma\left(\frac{n_1-1}{2}, \frac{1}{2}\right) + \Gamma\left(\frac{n_2-1}{2}, \frac{1}{2}\right) \equiv \Gamma\left(\frac{n_1+n_2-2}{2}, \frac{1}{2}\right) \sim \chi_{n_1+n_2-2}^2$

La distribución χ^2 hereda la reproductividad de la Gamma, y si sumamos dos distribuciones χ^2 independientes obtendremos otra con grados de libertad la suma de los respectivos grados de ambas. Además, $\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}\right)$, y por tanto:

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

$$\left. \frac{\sqrt{\frac{n_1 S_X^2 + n_2 S_Y^2}{\sigma^2(n_1 + n_2 - 2)}}}{\sqrt{\frac{\chi_{n_1+n_2-2}^2}{n_1 + n_2 - 2}}} \right\} \sim \sqrt{\frac{\chi_{n_1+n_2-2}^2}{n_1 + n_2 - 2}}$$

ambos, numerador y denominador, independientes por el teorema de Fisher, es decir,

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{n_1 S_X^2 + n_2 S_Y^2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sqrt{n_1 + n_2 - 2}$$

sigue una $t_{n_1+n_2-2}$. Es la función pivote. Por tanto, existe $t_{n_1+n_2-2, \frac{\alpha}{2}}$ tal que

$$P\left(-t_{n_1+n_2-2, \frac{\alpha}{2}} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{n_1 S_X^2 + n_2 S_Y^2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{n_1+n_2-2, \frac{\alpha}{2}}\right) = 1 - \alpha$$

Despejando se obtendrán los límites y el intervalo de confianza de nivel $1 - \alpha$,

$$\left(\bar{X} - \bar{Y} \pm t_{n_1+n_2-2, \frac{\alpha}{2}} \frac{\sqrt{n_1 S_X^2 + n_2 S_Y^2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{n_1 + n_2 - 2}}\right). \quad (6.11)$$

Escrito en términos de las cuasi-varianzas, será igual a

$$\mu_1 - \mu_2 \in \left(\bar{X} - \bar{Y} \pm t_{n_1+n_2-2, \frac{\alpha}{2}} \frac{\sqrt{(n_1-1)\widehat{S}_X^2 + (n_2-1)\widehat{S}_Y^2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{n_1+n_2-2}} \right) \quad (6.12)$$

Ejemplo 6.5:

Una muestra de estatura de 10 niños arrojó una media de 97cms y una cuasi-desviación típica de 11cms. En una muestra de estatura de 20 niñas se obtuvo una media de 112cm y cuasi-desviación típica de 17cm. Calcular un intervalo de confianza para la diferencia de estaturas medias niños y niñas.

Se tiene que $1 - \alpha = 0,95$, por lo que $\frac{\alpha}{2} = 0,05$, de donde $t_{28,0,025} = 2,043$. Por tanto, el intervalo de confianza buscado es

$$\left(97 - 112 \pm 2,043 \frac{\sqrt{1089 + 5491} \sqrt{\frac{1}{10} + \frac{1}{20}}}{\sqrt{28}} \right) = (-15 \pm 12,13) = (-27,13, -2,87).$$

6.2.7 Intervalo de confianza para el cociente de varianzas en poblaciones normales

Sea X_1, X_2, \dots, X_{n_1} una m.a.s. de $N(\mu_1, \sigma_1)$ y Y_1, Y_2, \dots, Y_{n_2} una m.a.s. de $N(\mu_2, \sigma_2)$, ambas muestras independientes, $\mu_1, \mu_2, \sigma_1, \sigma_2$ desconocidas. Se desea obtener un intervalo de confianza de nivel $1 - \alpha$ para $\frac{\sigma_1^2}{\sigma_2^2}$.

Calculamos un intervalo para el cociente y no para la diferencia, porque existe una función pivote para cociente y no para diferencia. Sin embargo, si nuestro interés es analizar la igualdad de varianzas, ello equivale a que el cociente sea 1, y por tanto, con un intervalo para el cociente se podría abordar el problema. Sabemos que $\frac{n_1 S_X^2}{\sigma_1^2} \sim \chi_{n_1-1}^2$, y que $\frac{n_2 S_Y^2}{\sigma_2^2} \sim \chi_{n_2-1}^2$ por el teorema de Fisher, siendo ambas variables independientes. Entonces,

$$\frac{\frac{n_1 S_X^2}{\sigma_1^2 (n_1-1)}}{\frac{n_2 S_Y^2}{\sigma_2^2 (n_2-1)}} \sim \frac{\chi_{n_1-1}^2}{\chi_{n_2-1}^2}$$

es decir,

$$\frac{n_1(n_2-1)S_X^2\sigma_2^2}{n_2(n_1-1)S_Y^2\sigma_1^2} \sim F_{n_1-1, n_2-1}.$$

Existen por tanto $F_{n_1-1, n_2-1, \frac{\alpha}{2}}$ y $F_{n_1-1, n_2-1, 1-\frac{\alpha}{2}}$ tales que

$$P \left(F_{n_1-1, n_2-1, 1-\frac{\alpha}{2}} \leq \frac{n_1(n_2-1)S_X^2\sigma_2^2}{n_2(n_1-1)S_Y^2\sigma_1^2} \leq F_{n_1-1, n_2-1, \frac{\alpha}{2}} \right) = 1 - \alpha$$

Despejando, el intervalo será

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{n_1(n_2-1)S_X^2}{n_2(n_1-1)S_Y^2 F_{n_1-1, n_2-1, \frac{\alpha}{2}}}, \frac{n_1(n_2-1)S_X^2}{n_2(n_1-1)S_Y^2 F_{n_1-1, n_2-1, 1-\frac{\alpha}{2}}} \right).$$

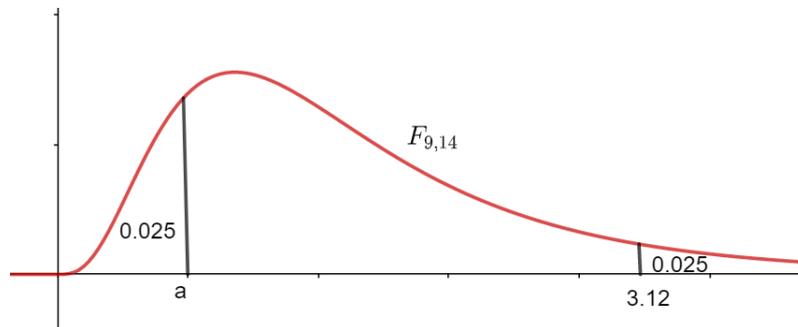
Si se desea escribir en términos de las cuasi-varianzas; como $\frac{n_1 S_X^2}{n_1-1} = \widehat{S}_X^2$ y $\frac{n_2 S_Y^2}{n_2-1} = \widehat{S}_Y^2$, entonces:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{\widehat{S}_X^2}{\widehat{S}_Y^2 F_{n_1-1, n_2-1, \frac{\alpha}{2}}}, \frac{\widehat{S}_X^2}{\widehat{S}_Y^2 F_{n_1-1, n_2-1, 1-\frac{\alpha}{2}}} \right).$$

Ejemplo 6.6:

La cuasi-varianza del peso de una muestra de 10 naranjas es de 81 gramos², y la cuasi-varianza en una muestra de 15 limones es de 75 gramos². Calcular un intervalo de confianza de nivel 0.95 para el cociente de varianzas poblacionales del peso de naranjas y limones. Se supone normalidad.

$$\left(\frac{81}{75 \times 3,12}, \frac{81}{75 \times a} \right)$$



$$0,025 = p(F_{9,14} < a) = p\left(F_{14,9} > \frac{1}{a}\right) \Rightarrow \frac{1}{a} = 3,798 \Rightarrow a = 0,26$$

El intervalo es (0,346, 4,154). Como 1 pertenece al intervalo al 95% de confianza, podemos asumir varianzas iguales.

6.3 Método general o de Neyman

Supongamos que se dispone de una m.a.s. X_1, X_2, \dots, X_n de una población $X \sim f(x, \theta)$, siendo θ un parámetro desconocido. Se desea obtener un intervalo de confianza para θ de nivel $1 - \alpha$. Si el método de la función pivote no fuera utilizable porque no existiera una función pivote,

o porque fuera desconocida, se utiliza el método general, también conocido como método de Neyman, que no es más que una generalización del método de la función pivote.

Cuando se aplica el método general, partimos del estimador de máxima verosimilitud (EMV) de θ , $\hat{\theta}$, y supongamos que su distribución depende de θ . Entonces, existirán dos cantidades, $g_1(\theta)$ y $g_2(\theta)$ que dependen de θ tales que $p(\hat{\theta} \leq g_1(\theta)) = \alpha_1$, y $p(\hat{\theta} > g_2(\theta)) = \alpha_2$, con $\alpha_1, \alpha_2 \geq 0$ y $\alpha_1 + \alpha_2 = \alpha$.

Despejando en estas expresiones, podemos obtener los límites de confianza para θ .

En realidad, con este procedimiento obtenemos una infinidad de intervalos para θ , uno por cada elección de α_1 y α_2 . Se requiere entonces un criterio adicional para elegir entre estos infinitos intervalos uno concreto. Tal criterio es determinar, entre todos los posibles intervalos, el más corto (lógicamente, a igualdad de nivel de confianza, preferimos el intervalo más corto posible, porque es más preciso).

Ejemplo 6.7:

Sea X_1, X_2, \dots, X_n una m.a.s. de $X \sim U(0, \theta)$. Obtener el intervalo de confianza más corto para θ , el máximo de la distribución, de nivel $1 - \alpha$.

Como no se conoce ninguna función pivote, se aplica el método de Neyman. Aquí, $\hat{\theta} = \max\{X_1, X_2, \dots, X_n\}$, con

$$f_{\hat{\theta}}(t) = \begin{cases} \frac{nt^{n-1}}{\theta^n}, & \text{si } t \in (0, \theta) \\ 0, & \text{en otro caso} \end{cases}$$

Fijamos $\alpha_1, \alpha_2 \in [0, 1]$, tales que $\alpha_1 + \alpha_2 = \alpha$.

Imponiendo la condición $\alpha_1 = p(\hat{\theta} \leq g_1(\theta)) = \int_0^{g_1(\theta)} \frac{nt^{n-1}}{\theta^n} dt = \frac{t^n}{\theta^n} \Big|_0^{g_1(\theta)} = \frac{g_1(\theta)^n}{\theta^n} \implies g_1(\theta) = \theta \sqrt[n]{\alpha_1}$.

Imponiendo la otra condición, $\alpha_2 = p(\hat{\theta} \geq g_2(\theta)) = \int_{g_2(\theta)}^{\theta} \frac{nt^{n-1}}{\theta^n} dt = \frac{t^n}{\theta^n} \Big|_{g_2(\theta)}^{\theta} = 1 - \frac{g_2(\theta)^n}{\theta^n} \implies \frac{g_2(\theta)^n}{\theta^n} = 1 - \alpha_2 \implies g_2(\theta) = \theta \sqrt[n]{1 - \alpha_2}$.

Por tanto,

$$p(g_1(\theta) \leq \hat{\theta} \leq g_2(\theta)) = 1 - \alpha \iff p(\theta \sqrt[n]{\alpha_1} \leq \hat{\theta} \leq \theta \sqrt[n]{1 - \alpha_2}) = 1 - \alpha.$$

Despejando,

$$\theta \sqrt[n]{\alpha_1} \leq \hat{\theta} \implies \theta \leq \frac{\hat{\theta}}{\sqrt[n]{\alpha_1}} \quad \text{y} \quad \hat{\theta} \geq \theta \sqrt[n]{1 - \alpha_2} \iff \theta \geq \frac{\hat{\theta}}{\sqrt[n]{1 - \alpha_2}}$$

luego un intervalo de confianza para θ de nivel $1 - \alpha$ será:

$$\left(\frac{\hat{\theta}}{\sqrt{1 - \alpha_2}}, \frac{\hat{\theta}}{\sqrt{\alpha_1}} \right) \quad (6.13)$$

En realidad, se han obtenido infinitos intervalos, uno para cada elección de α_1 y α_2 . Buscamos el más corto. La longitud es

$$\begin{aligned} L(\alpha_1) &= \hat{\theta} \left(\frac{1}{\sqrt[n]{\alpha_1}} - \frac{1}{\sqrt[n]{1 - \alpha - \alpha_1}} \right) \\ L'(\alpha_1) &= \hat{\theta} \left(-\frac{1}{n} \alpha_1^{\frac{-1}{n}-1} + \frac{1}{n} (1 - \alpha - \alpha_1)^{\frac{-1}{n}-1} \right) = 0 \implies \\ &\implies \frac{1}{n} (1 - \alpha - \alpha_1)^{\frac{-1}{n}-1} = \frac{1}{n} (\alpha_1)^{\frac{-1}{n}-1} \iff 1 - \alpha = 0 \quad \text{lo que es imposible} \end{aligned}$$

Es decir, no hay raíces y $L'(\alpha_1) \neq 0$ siempre. Así que, $L(\alpha_1)$ es siempre creciente o siempre decreciente, $L(\alpha_1)$ está definida para $\alpha_1 \in [0, \alpha]$.

Como $L(0) = \infty$, L decrece y la longitud mínima se obtiene cuando $\alpha_1 = \alpha$ y $\alpha_2 = 0$, y el intervalo de confianza resultante es:

$$\left(\hat{\theta} = \max\{X_1, \dots, X_n\}, \frac{\hat{\theta}}{\sqrt[n]{\alpha}} \right)$$

Si $x_1 = 0,3$, $x_2 = 0,5$, $x_3 = 0,1$, el intervalo para θ de nivel 0.95 y longitud menor, cuando $X \sim U(0, \theta)$ es $\left(0,5, \frac{0,5}{\sqrt[3]{0,05}} \right)$.

De manera similar, se puede obtener un intervalo para θ , el mínimo de la distribución, si $X \sim U(\theta, a)$, con a conocida. Este ejercicio se recomienda vivamente al lector.

Ejemplo 6.8:

Sea X_1, X_2, \dots, X_n una m.a.s. de $X \sim f(x, \theta)$, con

$$f(x, \theta) = \begin{cases} e^{-x} e^{\theta}, & x > \theta \\ 0, & \text{en otro caso.} \end{cases}$$

Se desea obtener un intervalo de nivel $(1 - \alpha)$ para θ , el mínimo de la distribución.

Entonces $\hat{\theta} = \min\{X_1, \dots, X_n\}$ y

$$f_{\hat{\theta}}(t) = \begin{cases} ne^{n\theta} e^{-nt} & \text{si } t > \theta \\ 0 & \text{en otro caso} \end{cases}$$

Fijando $\alpha_1, \alpha_2 \in [0, 1]$, tales que $\alpha_1 + \alpha_2 = \alpha$, e imponiendo la condición

$$\alpha_1 = P(\widehat{\theta} \leq g_1(\theta)) = \int_{\theta}^{g_1(\theta)} n e^{n\theta} e^{-nt} dt = e^{n\theta} | -e^{-nt} |_{\theta}^{g_1(\theta)} = e^{n\theta} [-e^{-ng_1(\theta)} + e^{-n\theta}] =$$

$$-e^{n(\theta - g_1(\theta))} + 1 \implies 1 - \alpha_1 = e^{n(\theta - g_1(\theta))} \implies \ln(1 - \alpha_1) = n(\theta - g_1(\theta)) \implies$$

$$g_1(\theta) = \theta - \frac{1}{n} \ln(1 - \alpha_1)$$

Por otra parte, imponiendo

$$\alpha_2 = P(\widehat{\theta} \geq g_2(\theta)) = \int_{g_2(\theta)}^{\infty} n e^{-nt} dr = e^{n\theta} [-e^{-nt}]_{g_2(\theta)}^{\infty} = e^{n\theta} e^{-ng_2(\theta)};$$

$$e^{n(\theta - g_2(\theta))} \implies \ln \alpha_2 = n(\theta - g_2(\theta)) \implies \frac{1}{n} \ln \alpha_2 = \theta - g_2(\theta) \implies$$

$$g_2(\theta) = \theta - \frac{1}{n} \ln \alpha_2.$$

$$\text{Así, } 1 - \alpha = P(g_1(\theta) \leq \widehat{\theta} \leq g_2(\theta)) = P\left(\theta - \frac{1}{n} \ln(1 - \alpha_1) \leq \widehat{\theta} \leq \theta - \frac{1}{n} \ln \alpha_2\right).$$

Despejando,

$$\theta - \frac{1}{n} \ln(1 - \alpha_2) \leq \widehat{\theta} \implies \theta \leq \widehat{\theta} + \frac{1}{n} \ln(1 - \alpha_1)$$

$$\widehat{\theta} \geq \theta - \frac{1}{n} \ln \alpha_2 \implies \theta \geq \widehat{\theta} + \frac{1}{n} \ln \alpha_2$$

Un intervalo de nivel $1 - \alpha$ es $\left(\widehat{\theta} + \frac{1}{n} \ln \alpha_2, \widehat{\theta} + \frac{1}{n} \ln(1 - \alpha_1)\right)$.

En realidad, para cada elección de α_1 y α_2 hay un intervalo distinto. Elegiremos los valores de α_1 y α_2 que nos garanticen que el intervalo es el más corto posible, para que así sea el más preciso.

$$L(\alpha_1) = \frac{1}{n} \ln(1 - \alpha_1) - \frac{1}{n} \ln \alpha_2 = \frac{1}{n} \ln(1 - \alpha_1) - \frac{1}{n} \ln(\alpha - \alpha_1)$$

$$L'(\alpha_1) = \frac{1}{n} \frac{(-1)}{1 - \alpha_1} - \frac{1}{n} \frac{(-1)}{\alpha - \alpha_1} = 0 \implies \frac{1}{1 - \alpha_1} - \frac{1}{\alpha - \alpha_1} = 0 \implies \alpha - 1 = 0,$$

lo que no es posible.

Por tanto, $L'(\alpha_1)$ no se anula, y $L(\alpha_1)$ es creciente o decreciente siempre en $[0, 1]$. Como $L(\alpha) = +\infty$, entonces L es creciente de 0 a α y la longitud mínima se obtiene para $\alpha_1 = 0$ y $\alpha_2 = \alpha$, o sea,

$$\left(\hat{\theta} + \frac{1}{n} \ln \alpha, \hat{\theta} \right). \quad (6.14)$$

Ejemplo 6.9:

Si se tiene

$$f(x, \theta) = \begin{cases} e^{-x} e^{\theta}, & x > \theta \\ 0, & \text{en otro caso} \end{cases}$$

y $\hat{\theta} = 2,3$ y tres observaciones ($n = 3$), el intervalo de nivel 0,95 más corto para θ será

$$\left(2,3 + \frac{1}{3} \ln 0,05, 2,3 \right).$$

6.4 Intervalos para parámetros de distribuciones discretas

Las funciones de distribución de variables discretas presentan saltos en todos los valores que toman las variables. Por tanto, para la mayoría de los valores $1 - \alpha$ no se puede obtener un intervalo de dicho nivel. En tales casos, se utilizan tamaños muestrales grandes para, haciendo uso del teorema central del límite, obtener intervalos de confianza aproximados para los parámetros. De nuevo, el punto de partida será el EMV $\hat{\theta}$ que, en general es consistente y asintóticamente normal (CAN).

6.4.1 Intervalo de confianza aproximado para p en una población $B(1, p)$ si n es grande ($n \geq 30$).

Sabemos que $\hat{p} = \bar{X}$ y que aproximadamente $\bar{X} \rightsquigarrow N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ si $n \geq 30$, luego $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow$ aproximadamente $N(0, 1)$.

Existe $z_{\frac{\alpha}{2}}$ tal que $p\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sqrt{n} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$.

Despejando,

$$\frac{\hat{p}-p}{\sqrt{p(1-p)}} \sqrt{n} \leq z_{\frac{\alpha}{2}} \implies p \geq \hat{p} - \frac{z_{\frac{\alpha}{2}} \sqrt{p(1-p)}}{\sqrt{n}}$$

$$\frac{\hat{p}-p}{\sqrt{p(1-p)}} \sqrt{n} \geq -z_{\frac{\alpha}{2}} \implies p \leq \hat{p} + \frac{z_{\frac{\alpha}{2}} \sqrt{p(1-p)}}{\sqrt{n}}$$

Como los límites del intervalo de confianza dependen de p (desconocido), no nos sirven, y estimamos p por \hat{p} para evitar la incertidumbre, de modo que el intervalo será:

$$\left(\hat{p} \pm z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right). \quad (6.15)$$

Ejemplo 6.10:

Se lanza una moneda 50 veces y se obtienen 20 caras. Obtener un intervalo aproximado al 95 % para la probabilidad de cara de la moneda.

$$z_{\frac{\alpha}{2}} = 1,96 \quad \hat{p} = \frac{20}{50} = 0,4$$

$$\left(0,4 - 1,96 \frac{\sqrt{0,24}}{\sqrt{50}}, 0,4 + 1,96 \frac{\sqrt{0,24}}{\sqrt{50}} \right)$$

6.4.2 Intervalo de confianza aproximado para λ en el modelo de Poisson

Sea X_1, X_2, \dots, X_n una m.a.s. de $X \sim P(\lambda)$. Se desea obtener un intervalo de confianza aproximado para λ de nivel $1 - \alpha$. Supondremos que $n \geq 30$. Entonces, por el teorema central del límite, \bar{X} sigue aproximadamente $N\left(\lambda, \frac{\sqrt{\lambda}}{\sqrt{n}}\right)$. Por tanto, existe $z_{\frac{\alpha}{2}}$ tal que $p\left(-z_{\frac{\alpha}{2}} \leq N(0, 1) \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$. En particular, $\frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \sim N(0, 1)$, luego $p\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$.

Despejando en las inecuaciones,

$$-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \iff \lambda \leq \bar{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\lambda}{n}}$$

$$z_{\frac{\alpha}{2}} \geq \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \iff \lambda \geq \bar{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\lambda}{n}}$$

Siendo n grande, estimamos λ en los límites de confianza, que dependen de λ , a través de \bar{X} (consistente) para obtener el intervalo aproximado de nivel $1 - \alpha$.

$$\left(\bar{X} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}}{n}} \right) \quad (6.16)$$

Ejemplo 6.11:

Se desea estimar el número medio de personas que entran en un centro comercial los lunes entre

las 2 y las 3 de la tarde. Examinamos los primeros 50 lunes de un año en esa franja horaria, y se obtiene que entraron un total de 550 personas. Obtener un intervalo de confianza aproximado para la media teórica de visitas, al 95 % de confianza.

Supongamos que las entradas en el centro comercial se distribuyen según Poisson. Como $1 - \alpha = 0,95$, $z_{\frac{\alpha}{2}} = 1,96$. Aplicando la fórmula,

$$\left(\frac{550}{50} \pm 1,96 \sqrt{\frac{550}{2500}} \right)$$

es un intervalo aproximado de nivel 95 % para λ .

6.4.3 Intervalo de confianza para la diferencia de proporciones

Sea X_1, X_2, \dots, X_{n_1} una m.a.s de $B(1, p_1)$, y sea Y_1, Y_2, \dots, Y_{n_2} una m.a.s de $B(1, p_2)$, ambas muestras independientes. Admitiremos que $n_1 \geq 30$, y $n_2 \geq 30$. Se desea obtener un intervalo de confianza para $p_1 - p_2$ de nivel $1 - \alpha$.

Por el teorema central del límite, se sabe que $\hat{p}_1 = \bar{X} \rightsquigarrow N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right)$ aproximadamente,

y que $\hat{p}_2 = \bar{Y} \rightsquigarrow N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$, aproximadamente.

Como las dos muestras son independientes, también lo van a ser las distribuciones de \hat{p}_1 y \hat{p}_2 . Entonces, de forma aproximada,

$$\hat{p}_1 - \hat{p}_2 \rightsquigarrow N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

o tipificado

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \rightsquigarrow N(0, 1), \quad \text{aproximadamente.}$$

Entonces, existe $z_{\frac{\alpha}{2}}$ tal que $p\left(-z_{\frac{\alpha}{2}} \leq N(0, 1) \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$. Por tanto,

$$p\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Despejando,

$$-z_{\frac{\alpha}{2}} \leq \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \iff p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{\frac{\alpha}{2}} \iff p_1 - p_2 \geq \hat{p}_1 - \hat{p}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Como los límites de confianza dependen de p_1 y p_2 , estimamos p_1 y p_2 a través de \hat{p}_1 y \hat{p}_2 , que son consistentes (cuanto mayor sea el tamaño de n_1 y n_2 , menor error se comete). El intervalo aproximado de nivel $1 - \alpha$ para $p_1 - p_2$ es:

$$\left(\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right). \quad (6.17)$$

Ejemplo 6.12:

Se desea obtener un intervalo de confianza aproximado de nivel 90% para la diferencia entre la proporción de aprobados en las asignaturas de A y B. De 44 alumnos consultados, aprobaron 33 el examen final de A, y de 32 alumnos consultados, aprobaron 18 el examen final de B. Supondremos que existe independencia entre las notas de ambos exámenes. Un intervalo del 90% para $p_1 - p_2$ siendo p_1 la proporción de aprobados en A y p_2 la proporción de aprobados en B es:

$$\left(\frac{33}{44} - \frac{18}{32} \pm 1,645 \sqrt{\frac{33 \times 11}{44 \times 44 \times 44} + \frac{18 \times 14}{32 \times 32 \times 32}} \right) = (-0,0157, 0,3907)$$

Al 95% de confianza, el intervalo obtenido sería

$$\left(\frac{33}{44} - \frac{18}{32} \pm 1,96 \sqrt{\frac{33 \times 11}{44 \times 44 \times 44} + \frac{18 \times 14}{32 \times 32 \times 32}} \right) = (-0,04, 0,42)$$

6.5 Inversión del estadístico de un test

En general, cuando se habla de invertir un test para obtener un intervalo de confianza, se trata de identificar el conjunto de valores del parámetro poblacional que se corresponderían con no rechazar la hipótesis nula.

A continuación se considera el método de la inversión de un test, y se exploran las relaciones entre el contraste de hipótesis de un parámetro θ y la confianza de un intervalo para θ . Se propone el siguiente ejemplo.

Ejemplo 6.13:

Sea X_1, \dots, X_n una m.a.s. de $X \sim N(\mu, \sigma_0^2)$, siendo σ_0^2 conocida. Se sabe que un intervalo para μ de nivel de confianza $1 - \alpha$ es

$$\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}} \sigma_0, \bar{X} + z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}} \sigma_0 \right).$$

Si se define un test φ que rechace el valor $\mu = \mu_0$ si y sólo si μ_0 está fuera de este intervalo, o lo que es lo mismo, si

$$\frac{\sqrt{n}|\bar{X} - \mu_0|}{\sigma_0} \geq z_{\frac{\alpha}{2}},$$

entonces

$$P_{\mu_0} \left(\sqrt{n} \frac{|\bar{X} - \mu_0|}{\sigma_0} \geq z_{\frac{\alpha}{2}} \right) = \alpha,$$

y el test φ es un test de tamaño α para $\mu = \mu_0$ en contra de la alternativa $\mu \neq \mu_0$.

En cambio, una familia de tests de nivel α para la hipótesis $\mu = \mu_0$, genera una familia de intervalos de confianza para μ , simplemente tomando, como posibles intervalos de confianza para μ_0 , el conjunto de intervalos para μ para los cuales no se puede rechazar $\mu = \mu_0$.

De manera similar, se puede generar una familia de tests de nivel α partiendo de una cota de confianza inferior (o superior) de nivel $(1 - \alpha)$. Por ejemplo, si se empieza con una cota inferior de nivel $(1 - \alpha)$ para μ , $\bar{X} - z_{\alpha} \left(\frac{\sigma_0}{\sqrt{n}} \right)$, entonces, definiendo el test $\varphi(X)$ que rechaza $\mu \leq \mu_0$ si y sólo si $\mu_0 < \bar{X} - z_{\alpha} \left(\frac{\sigma_0}{\sqrt{n}} \right)$, se obtiene un test de nivel α para la hipótesis $\mu \leq \mu_0$.

Este ejemplo es un caso especial del principio de dualidad que se presenta a continuación. Se escribirá $H_0(\theta_0)$ para referirnos a la hipótesis $H_0 : \theta = \theta_0$ y $H_1(\theta_0)$ para referirnos a la hipótesis alternativa, teniendo en cuenta que ambas podrían ser unilaterales o bilaterales.

Principio de dualidad:

Sea $A(\theta_0)$, con $\theta_0 \in \Theta$, la región de aceptación de un test de nivel α para $H_0(\theta_0)$. Para cada observación $x = (x_1, \dots, x_n)$, sea $S(x)$ el conjunto

$$S(x) = \{\theta | x \in A(\theta), \theta \in \Theta\}. \quad (6.18)$$

Entonces $S(x)$ es una familia de conjuntos de confianza para θ siendo $1 - \alpha$ el nivel de confianza. Además, si $A(\theta_0)$ es el contraste uniformemente más potente (UMP) para el problema $(\alpha, H_0(\theta_0), H_1(\theta_0))$, entonces $S(X_1, \dots, X_n)$ es el conjunto que minimiza

$$P_{\theta} (S(X_1, \dots, X_n) \ni \theta), \quad \forall \theta \in H_1(\theta') \quad (6.19)$$

de entre todas las familias de conjuntos de confianza de nivel $(1 - \alpha)$. Esto es, $S(X_1, \dots, X_n)$ es el contraste uniformemente más preciso, (*uniformly most accurate* UMA).

Demostración Se tiene que $\theta \in S(x)$ si y sólo si $x \in A(\theta)$. Por tanto,

$$P_{\theta} (S(X_1, \dots, X_n) \ni \theta) = P_{\theta} (X \in A(\theta)) \geq 1 - \alpha,$$

como se afirmó.

Si $S^*(X_1, \dots, X_n)$ es cualquier otra familia de conjuntos de nivel de confianza $1 - \alpha$, sea $A^*(\theta) = \{x | S^*(x) \ni \theta\}$. Entonces,

$$P_\theta((X_1, \dots, X_n) \in A^*(\theta)) = P_\theta(S^*(X_1, \dots, X_n) \ni \theta) \geq 1 - \alpha;$$

y puesto que $A(\theta_0)$ es UMP para $(\alpha, H_0(\theta_0), H_1(\theta_0))$, se cumple que

$$P_\theta((X_1, \dots, X_n) \in A^*(\theta_0)) \geq P_\theta(X \in A(\theta_0)), \quad \text{para cualquier } \theta \in H_1(\theta_0).$$

Por tanto,

$$P_\theta(S^*(X_1, \dots, X_n) \ni \theta_0) \geq P_\theta((X_1, \dots, X_n) \in A(\theta_0)) = P_\theta(S(X_1, \dots, X_n) \ni \theta_0)$$

para cada $\theta \in H_1(\theta_0)$. Esto completa la demostración.

6.6 Inferencia bayesiana. Intervalos bayesianos

En el contexto del modelo paramétrico, se dispone de una muestra aleatoria simple (m.a.s.), X_1, \dots, X_n de una población $X \sim f(x, \theta)$, siendo θ desconocido, $\theta \in \Theta$. A diferencia de la inferencia clásica, se supondrá que θ es una variable aleatoria, y que por tanto dispondrá de una función de probabilidad o densidad que se denotará por $p(\theta)$. En el contexto bayesiano, esta función $p(\theta)$ representa las creencias o información disponible “**a priori**”, es decir, antes de la experimentación por parte del investigador.

Una vez que el experimentador muestrea, y dispone de $X_1 = x, \dots, X_n = x_n$, se obtiene información adicional sobre θ que cambia la verosimilitud de θ , obteniéndose la llamada distribución “**a posteriori**” de θ , $h(\theta|x_1, \dots, x_n)$, que se calcula mediante el teorema de Bayes.

Entonces,

$$h(\theta|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n, \theta)p(\theta)}{\int_{\Theta} f(x_1, \dots, x_n, \theta)p(\theta)d\theta}$$

6.6.1 Familias conjugadas de distribuciones

Sería deseable que las creencias a priori del investigador, reflejadas en $p(\theta)$, fueran ratificadas por la muestra y, por tanto, que si $p(\theta)$ pertenece a una cierta familia de distribuciones $h(\theta|X_1, \dots, X_n)$ pertenezca a la misma familia. Cuando una familia de distribuciones satisface esta condición para un determinado parámetro, diremos que la familia es conjugada para dicho parámetro. Veremos a continuación diferentes tipos de familias conjugadas.

Proposición 1:

La familia de distribuciones Beta es conjugada para el parámetro θ de una distribución Bernoulli. Es decir, dada una m.a.s. X_1, X_2, \dots, X_n de una $B(1, \theta)$, si la distribución a priori de θ pertenece a la familia *Beta*, entonces la distribución a posteriori de θ también pertenece a la misma familia.

Demostración:

Supongamos que a priori $\theta \sim Beta(a, b)$, entonces, a posteriori, después de observar $X_1 = x_1, \dots, X_n = x_n$,

$$h(\theta|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n, \theta)p(\theta)}{\int_{\Theta} f(x_1, \dots, x_n, \theta)p(\theta)d\theta} = \frac{\prod_{i=1}^n f(x_i, \theta)p(\theta)}{\int_0^1 \prod_{i=1}^n f(x_i, \theta)p(\theta)d\theta} =$$

$$\frac{\prod_{i=1}^n [\theta^{x_i}(1-\theta)^{1-x_i}] \frac{1}{\beta(a,b)} \theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \prod_{i=1}^n [\theta^{x_i}(1-\theta)^{1-x_i}] \frac{1}{\beta(a,b)} \theta^{a-1}(1-\theta)^{b-1} d\theta}.$$

Puesto que $\beta(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt$, se tendrá que

$$\frac{\prod_{i=1}^n [\theta^{x_i}(1-\theta)^{1-x_i}] \frac{1}{\beta(a,b)} \theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \prod_{i=1}^n [\theta^{x_i}(1-\theta)^{1-x_i}] \frac{1}{\beta(a,b)} \theta^{a-1}(1-\theta)^{b-1} d\theta} = \frac{\theta^{\sum x_i + a - 1} (1-\theta)^{n - \sum x_i + b - 1}}{\beta(\sum x_i + a, n - \sum x_i + b)}$$

es decir, la densidad a posteriori es la *Beta* de parámetros $\sum x_i + a, n - \sum x_i + b$.

Ejemplo 6.14:

Un investigador tira una moneda cuya probabilidad de cara es desconocida, y no tiene información a priori sobre dicha probabilidad. Es decir, si θ es la probabilidad de cara, entonces $p(\theta) \sim U(0, 1) \equiv Beta(1, 1)$. Lanza la moneda 10 veces y salen 4 caras y 6 cruces, es decir, $\sum_{i=1}^{10} x_i = 4$. Entonces, la distribución a posteriori de θ es $Beta(4 + 1, 10 - 4 + 1) = Beta(5, 7)$.

Proposición 2.:

La familia de distribuciones Gamma es conjugada para el parámetro λ de la distribución de Poisson. Es decir, si X_1, X_2, \dots, X_n es una m.a.s. de $X \sim Poisson(\lambda)$, y a priori $\lambda \sim \Gamma(a, p)$, entonces la distribución a posteriori de λ también es de tipo Gamma.

Demostración:

A priori, $p(\lambda) = \frac{p^a}{\Gamma(a)} e^{-p\lambda} \lambda^{a-1}$, si $\lambda > 0$, y 0 en otro caso.

Si $X_1 = x_1, \dots, X_n = x_n$ es una realización muestral de $X \sim P(\lambda)$, entonces, la distribución a posteriori de λ es

$$h(\lambda|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n, \lambda)p(\lambda)d\lambda}{\int_0^{+\infty} f(x_1, \dots, x_n, \lambda)p(\lambda)d\lambda} = \frac{\prod_{i=1}^n \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) \frac{p^a}{\Gamma(a)} e^{-p\lambda} \lambda^{a-1}}{\int_0^{\infty} \prod_{i=1}^n \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) \frac{p^a}{\Gamma(a)} e^{-p\lambda} \lambda^{a-1} d\lambda} =$$

$$\frac{e^{-n\lambda} \lambda^{\sum x_i} e^{-p\lambda} \lambda^{a-1}}{\int_0^{\infty} e^{-n\lambda} \lambda^{\sum x_i} e^{-p\lambda} \lambda^{a-1} d\lambda} = \frac{e^{-(n+p)\lambda} \lambda^{\sum x_i + a - 1}}{\int_0^{\infty} e^{-(n+p)\lambda} \lambda^{\sum x_i + a - 1} d\lambda} = \frac{e^{-(n+p)\lambda} \lambda^{\sum x_i + a - 1}}{\Gamma(\sum X_i + a) / (n+p)^{\sum x_i + a}} =$$

$$\begin{cases} \frac{(n+p)\Sigma x_i + a}{\Gamma(\Sigma x_i + a)} e^{-(n+p)\lambda \Sigma x_i + a - 1} & \text{si } \lambda > 0 \\ 0 & \text{en otro caso} \end{cases} \rightsquigarrow \Gamma(\Sigma x_i + a, n + p)$$

Si la distribución a priori del parámetro λ en una distribución de Poisson es $\Gamma(a, p)$, entonces, a posteriori, $\lambda \rightsquigarrow \Gamma(\Sigma x_i + a, n + p)$.

Ejemplo 6.15:

Supongamos que el número de llamadas a un teléfono es, a priori, $P(\lambda)$. Se piensa que $\lambda \rightsquigarrow \Gamma(1, \frac{1}{4}) \equiv \text{exp}(\frac{1}{4})$. Se observan las llamadas en 10 días, obteniéndose la muestra 0, 3, 5, 1, 1, 6, 2, 8, 7, 1. Entonces, la distribución a posteriori de λ es

$$\Gamma\left(34 + 1, 10 + \frac{1}{4}\right) \equiv \Gamma\left(35, \frac{41}{4}\right)$$

$$p(\lambda) = \begin{cases} \frac{1}{4} e^{-\frac{1}{4}\lambda} & \text{si } \lambda > 0 \\ 0 & \text{en otro caso} \end{cases}$$

$$h(\lambda | x_1, \dots, x_n) = \begin{cases} \frac{(\frac{41}{4})^{35}}{34!} e^{-\frac{41}{4}\lambda} \lambda^{34} & \text{si } \lambda > 0 \\ 0 & \text{en otro caso} \end{cases}$$

En particular, a priori se esperan cuatro llamadas y, a posteriori, la muestra corrige este valor esperado a $\frac{35}{\frac{41}{4}} = 3,9$.

Proposición 3.:

La distribución Gamma es conjugada para el parámetro λ de la distribución exponencial, es decir, si X_1, \dots, X_n es una m.a.s. de $X \rightsquigarrow \text{exp}(\lambda)$, y, a priori $\lambda \rightsquigarrow \Gamma(a, p)$, entonces la distribución a posteriori de λ también es de tipo Gamma.

Demostración:

A priori,

$$p(\lambda) = \begin{cases} \frac{p^a}{\Gamma(a)} e^{-p\lambda} \lambda^{a-1} & \text{si } \lambda > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Entonces,

$$\begin{aligned} h(\lambda | x_1, \dots, x_n) &= \frac{f(x_1, \dots, x_n, \lambda) p(\lambda)}{\int_0^{+\infty} f(x_1, \dots, x_n, \lambda) p(\lambda) d\lambda} = \frac{\prod_{i=1}^n (\lambda e^{-\lambda x_i}) \frac{p^a}{\Gamma(a)} e^{-p\lambda} \lambda^{a-1}}{\int_0^{+\infty} \prod_{i=1}^n (\lambda e^{-\lambda x_i}) \frac{p^a}{\Gamma(a)} e^{-p\lambda} \lambda^{a-1} d\lambda} = \\ &= \frac{\lambda^n e^{-\lambda \Sigma x_i} e^{-p\lambda} \lambda^{a-1}}{\int_0^{+\infty} \lambda^n e^{-\lambda \Sigma x_i} e^{-p\lambda} \lambda^{a-1} d\lambda} = \frac{e^{-(\Sigma x_i + p)\lambda} \lambda^{n+a-1}}{\int_0^{+\infty} e^{-(\Sigma x_i + p)\lambda} \lambda^{n+a-1} d\lambda} = \frac{e^{-(\Sigma_{i=1}^n x_i + p)\lambda} \lambda^{n+a-1}}{\frac{\Gamma(n+a)}{(\Sigma x_i + p)^{n+a}}} = \end{aligned}$$

$$= \frac{(\sum x_i + p)^{n+a}}{\Gamma(n+a)} e^{-(\sum x_i + p)\lambda} \lambda^{n+a-1} \rightsquigarrow \Gamma(n+a, \sum x_i + p)$$

Si la distribución a priori del parámetro λ en una distribución $exp(\lambda)$ es $\Gamma(a, p)$, a posteriori $\lambda \rightsquigarrow \Gamma(n+a, \sum x_i + p)$.

Ejemplo 6.16:

Supongamos que los tiempos entre llegadas consecutivas a la cola de una oficina bancaria siguen una $exp(\lambda)$. A priori, se asume que $\lambda \rightsquigarrow \Gamma(2, 1)$. Observados los 4 valores 0,2, 0,4, 0,1, 0,3, la distribución a posteriori de λ es $\Gamma(4+2, 1+1) \equiv \Gamma(6, 2)$.

$$\text{A priori, } p(\lambda) = \begin{cases} \lambda e^{-\lambda} & \text{si } \lambda > 0 \\ 0 & \text{en otro caso} \end{cases}$$

$$\text{A posteriori, } h(\lambda|x_1, \dots, x_n) = \begin{cases} \frac{2^6}{120} e^{-2\lambda} \lambda^5 & \text{si } \lambda > 0 \\ 0 & \text{en otro caso.} \end{cases}$$

La demostración de la siguiente proposición es similar a las anteriores, -quizás un poco más tediosa-, y se deja como ejercicio para el lector.

Proposición 4.:

La familia de distribuciones normal es conjugada para el parámetro μ de la distribución normal. Es decir, si X_1, \dots, X_n es una m.a.s. de $X \rightsquigarrow N(\mu, \sigma_0)$, con σ_0 conocido y la distribución a priori de μ es $N(\mu_a, \sigma_a)$, entonces, la distribución a posteriori de μ es

$$N\left(\frac{\bar{X}\sigma_a^2 + \frac{\mu_a\sigma_0^2}{n}}{\sigma_a^2 + \frac{\sigma_0^2}{n}}, \frac{1}{\sqrt{\frac{n}{\sigma_0^2} + \frac{1}{\sigma_a^2}}}\right)$$

Ejemplo 6.17:

Sea $X \rightsquigarrow N(\mu, 1)$. a priori, $\mu \rightsquigarrow N(0, 1)$. Se observan 10 valores, obteniéndose 1.2, 0.7, 0.3, -0.3, 2, 1.9, 1.6, 0.5, -1, 1. Entonces, la distribución a posteriori de μ es

$$N\left(\frac{0,79 \times 1^2 + 0 \times \frac{1}{10}}{1^2 + \frac{1}{10}}, \frac{1}{\sqrt{\frac{10}{1} + \frac{1}{1}}}\right) \equiv N(0,71, 0,3015).$$

6.7 Intervalos de confianza bayesianos

Un tercer método para obtener intervalos de confianza está basado en el análisis bayesiano, en el que tenemos en cuenta el conocimiento *a priori* que el experimentador tiene sobre θ . Como se ha dicho, esto se refleja en la especificación de la distribución a priori $p(\theta)$ sobre Θ . En este escenario, la probabilidad de cobertura del parámetro no se basa en la distribución de

X (población), sino en la distribución condicional $h(\theta|x_1, \dots, x_n)$, distribución a posteriori de θ .

Definición:

Un intervalo (U, V) , donde U y V son estimadores de θ , tal que la $p(U \leq \theta \leq V) = 1 - \alpha$ recibe el nombre de intervalo bayesiano para θ . Notemos que en el escenario bayesiano podemos hablar con pleno sentido de la probabilidad de que $p(U \leq \theta \leq V) = 1 - \alpha$, ya que esa probabilidad se calcula utilizando la distribución a posteriori.

Para enfatizar esta distinción entre el análisis bayesiano y el clásico, algunos autores prefieren el término “conjuntos de credibilidad” para los intervalos de confianza bayesianos.

Ejemplo 6.18:

Sea X_1, \dots, X_n una m.a.s. de $X \sim N(\mu, 1)$ con $\mu \in \mathbb{R}$, y supongamos que la distribución a priori de μ es $N(0, 1)$. Entonces, haciendo uso de la familia conjugada normal, sabemos que la distribución a posteriori $h(\mu|x_1, \dots, x_n)$ es

$$N\left(\frac{\sum_{i=1}^n x_i}{n+1}, \frac{1}{\sqrt{n+1}}\right)$$

Entonces, el intervalo bayesiano de nivel $(1 - \alpha)$ es

$$\left(\frac{n\bar{x}}{n+1} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n+1}}, \frac{n\bar{x}}{n+1} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n+1}}\right),$$

ya que cubre con probabilidad $(1 - \alpha)$, (calculada con la distribución a posteriori), al verdadero valor de μ .

Ejemplo 6.19:

Sea X_1, \dots, X_n una m.a.s. de $X \sim B(1, p)$, y sea la distribución a priori sobre p una $U(0, 1)$. Sabemos, ya que la familia *Beta* es conjugada para el parámetro p , que la distribución a posteriori $h(p|x_1, \dots, x_n)$ es $Beta(\sum x_i + 1, n - \sum x_i + 1)$. Por tanto, si

$\beta(\sum x_i + 1, n - \sum x_i + 1)_{\frac{\alpha}{2}}$ y $\beta(\sum x_i + 1, n - \sum x_i + 1)_{1-\frac{\alpha}{2}}$ dejan a su derecha e izquierda, respectivamente, en la distribución a posteriori, una masa de $\frac{\alpha}{2}$, el intervalo bayesiano para p será

$$\left(\beta(\sum x_i + 1, n - \sum x_i + 1)_{1-\frac{\alpha}{2}}, \beta(\sum x_i + 1, n - \sum x_i + 1)_{\frac{\alpha}{2}}\right).$$

Bibliografía

1. Rohatgi, V.K. y Ehsanes Saleh, A.K. Md. (2001). *An Introduction to Probability and Statistics*. John Wiley and Sons.

2. Neyman, J. y Pearson, E. (1928). *On the use and interpretation of certain test criteria for purposes of statistical inference: part I*. *Biometrika*, 20(1/2):175-242.
3. Neyman, J. (1977). *Frequentist probability and frequentist statistics*. *Synthese*, 36:97-131.
4. Fisher, R. (1925). *Theory of Statistical Estimation*. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700-725.
5. Fisher, R. (1935). *The Logic of Inductive Inference*. *Journal of the Royal Statistical Society*, 39-82.
6. Butler, C. (1985). *Statistics in Linguistics*. Basil Blackwell.

Tema 7

Estimación por intervalos II. Métodos para evaluar estimadores de intervalos: tamaño y probabilidad de cobertura, optimalidad relacionada con los tests, optimalidad bayesiana, optimalidad de la función de pérdida

7.1 Introducción

Como se ha visto, uno de los problemas estadísticos clave desconocidos asociados a una variable o población. Al obtener un intervalo de confianza para un parámetro, se dispone de un rango de valores en el que existe una confianza alta, y predeterminada por el investigador, de encontrar el valor del parámetro desconocido. Se debe tener en cuenta que todo intervalo de confianza depende de la estimación hecha del parámetro de interés a partir de una realización muestral. Por tanto, la calidad y precisión de un intervalo de confianza están totalmente relacionadas con elementos como el tamaño muestral o el nivel de confianza requerido.

7.2 Métodos para evaluar estimadores de intervalos: tamaño y probabilidad de cobertura

Asumiremos, como hasta ahora, que se dispone de una muestra aleatoria simple (m.a.s.), X_1, \dots, X_n de una población $X \rightsquigarrow f(x, \theta)$, θ desconocido, $\theta \in \Theta$, siendo Θ el espacio paramétrico.

En la evaluación de los intervalos de confianza, hay dos conceptos esenciales que permiten determinar qué soluciones son las más apropiadas: la longitud o tamaño del intervalo, que indica su precisión, y la probabilidad de cobertura, que se define teóricamente como $p(U \leq \theta \leq V)$, donde U y V son dos estimadores, uno por defecto y otro por exceso de θ , que representan el límite superior y el inferior del intervalo, respectivamente.

Desafortunadamente, una de las leyes perversas que inunda la ciencia, hace que ambas, precisión y probabilidad de cobertura, no puedan ser controladas simultáneamente. Si se aumenta el nivel de confianza, entonces se reduce la precisión, ya que se incrementa la longitud del intervalo. Esto es claro si se llevan las cosas a los extremos. ¿Por qué no elegir como nivel de confianza la unidad? ¿Por qué conformarse con niveles del 95 % ó 99 %? Pues porque si el nivel se hace unitario, el intervalo de confianza se convierte en el espacio paramétrico. Para tener certeza total de encontrar el parámetro en un intervalo, acabaremos diciendo una obviedad, a saber, que el parámetro se encuentra en el espacio paramétrico. Pero ese es precisamente el punto de partida, y lo que se desea es reducir esa incertidumbre determinando un intervalo dentro del espacio paramétrico en el que exista una confianza alta (pero no total) de encontrar el parámetro.

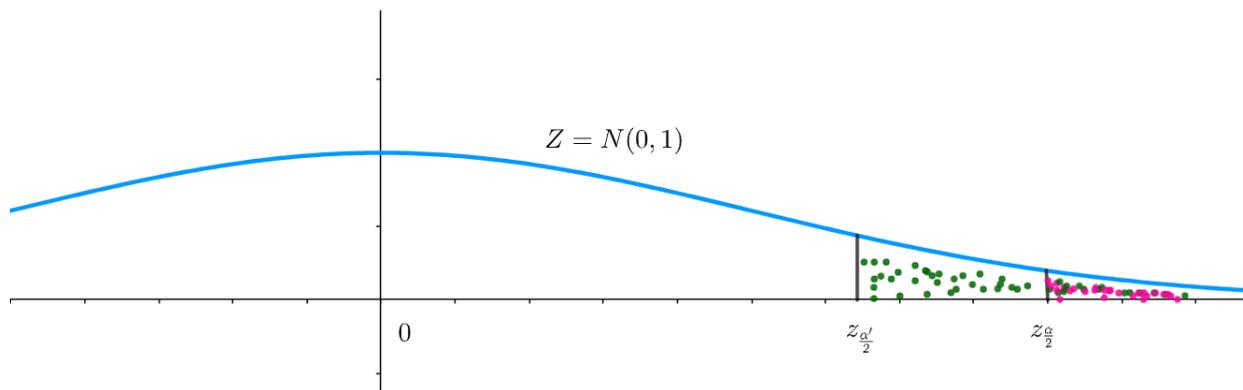
Ejemplo 7.1:

Como sabemos, si X_1, \dots, X_n es una muestra aleatoria simple de una $N(\mu, \sigma_0)$, μ desconocida y

σ_0 conocida, el intervalo de confianza para μ de nivel $1 - \alpha$ es

$$\left(\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} \right).$$

En dicha expresión, se hacen patentes los comentarios anteriores. Si $1 - \alpha > 1 - \alpha'$, ó $\alpha < \alpha'$, y entonces $\frac{\alpha}{2} < \frac{\alpha'}{2}$, se tiene que $z_{\frac{\alpha}{2}} > z_{\frac{\alpha'}{2}}$, y el intervalo de nivel $1 - \alpha$ tiene mayor longitud que el de nivel $1 - \alpha'$.



Como se puede sospechar, el incremento de información, - el aumento del tamaño muestral -, debe reducir la incertidumbre sobre el parámetro. Y esto es así, efectivamente. Al incrementar el tamaño muestral se reduce la longitud del intervalo. Una manera de compensar el incremento de longitud al aumentar el nivel de confianza es incorporar más observaciones, si ello es posible.

Ejemplo 7.2:

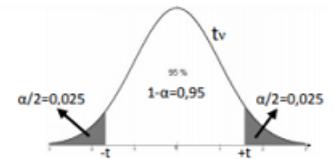
El intervalo de confianza para μ en una población X tal que $X \sim N(\mu, \sigma)$ a partir de una m.a.s. X_1, \dots, X_n , viene dado por

$$\left(\bar{X} \pm t_{n-1, \frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \right).$$

Si se incrementa el tamaño n , el denominador \sqrt{n} crece, y $t_{n-1, \frac{\alpha}{2}}$ disminuye.

Distribución t de Student

Contiene los valores de t tales que $\frac{\alpha}{2} = P(t_v \geq t)$, donde v son los Grados de Libertad



v grados de libertad	α/2												
	0,0005	0,001	0,005	0,01	0,025	0,05	0,1	0,2	0,25	0,3	0,4	0,45	0,475
1	636,619	318,309	63,657	31,821	12,706	6,314	3,078	1,376	1,000	0,727	0,325	0,158	0,079
2	31,599	22,327	9,925	6,965	4,303	2,920	1,886	1,061	0,816	0,617	0,289	0,142	0,071
3	12,924	10,215	5,841	4,541	3,182	2,353	1,638	0,978	0,765	0,584	0,277	0,137	0,068
4	8,610	7,173	4,604	3,747	2,776	2,132	1,533	0,941	0,741	0,569	0,271	0,134	0,067
5	6,869	5,893	4,032	3,365	2,571	2,015	1,476	0,920	0,727	0,559	0,267	0,132	0,066
6	5,959	5,208	3,707	3,143	2,447	1,943	1,440	0,906	0,718	0,553	0,265	0,131	0,065
7	5,408	4,785	3,499	2,991	2,365	1,895	1,415	0,896	0,711	0,549	0,263	0,130	0,065
8	5,041	4,501	3,355	2,891	2,306	1,860	1,397	0,889	0,706	0,546	0,262	0,130	0,065
9	4,781	4,297	3,250	2,821	2,262	1,833	1,383	0,883	0,703	0,543	0,261	0,129	0,064
10	4,587	4,144	3,169	2,761	2,228	1,812	1,372	0,879	0,700	0,542	0,260	0,129	0,064
11	4,437	4,025	3,106	2,713	2,201	1,796	1,363	0,876	0,697	0,540	0,260	0,129	0,064
12	4,318	3,930	3,055	2,661	2,179	1,782	1,356	0,873	0,695	0,539	0,259	0,128	0,064
13	4,221	3,852	3,012	2,610	2,160	1,771	1,350	0,870	0,694	0,538	0,259	0,128	0,064
14	4,140	3,787	2,977	2,564	2,145	1,761	1,345	0,868	0,692	0,537	0,258	0,128	0,064
15	4,073	3,733	2,947	2,522	2,131	1,753	1,341	0,866	0,691	0,536	0,258	0,128	0,064
16	4,015	3,686	2,921	2,533	2,120	1,746	1,337	0,865	0,690	0,535	0,258	0,128	0,064
17	3,965	3,646	2,898	2,547	2,110	1,740	1,333	0,863	0,689	0,534	0,257	0,128	0,064
18	3,922	3,610	2,878	2,542	2,101	1,734	1,330	0,862	0,688	0,534	0,257	0,127	0,064
19	3,883	3,579	2,861	2,539	2,093	1,729	1,328	0,861	0,688	0,533	0,257	0,127	0,064
20	3,850	3,552	2,845	2,538	2,086	1,725	1,325	0,860	0,687	0,533	0,257	0,127	0,063
21	3,819	3,527	2,831	2,538	2,080	1,721	1,323	0,859	0,686	0,532	0,257	0,127	0,063
22	3,792	3,505	2,819	2,538	2,074	1,717	1,321	0,858	0,686	0,532	0,256	0,127	0,063
23	3,768	3,485	2,807	2,530	2,069	1,714	1,319	0,858	0,685	0,532	0,256	0,127	0,063
24	3,745	3,467	2,797	2,492	2,064	1,711	1,318	0,857	0,685	0,531	0,256	0,127	0,063
25	3,725	3,450	2,787	2,465	2,060	1,708	1,316	0,856	0,684	0,531	0,256	0,127	0,063
26	3,707	3,435	2,779	2,470	2,056	1,706	1,315	0,856	0,684	0,531	0,256	0,127	0,063
27	3,690	3,421	2,771	2,471	2,052	1,703	1,314	0,855	0,684	0,531	0,256	0,127	0,063
28	3,674	3,408	2,763	2,461	2,048	1,701	1,313	0,855	0,683	0,530	0,256	0,127	0,063
29	3,659	3,396	2,756	2,462	2,045	1,699	1,311	0,854	0,683	0,530	0,256	0,127	0,063
30	3,646	3,385	2,750	2,457	2,042	1,697	1,310	0,854	0,683	0,530	0,256	0,127	0,063
31	3,633	3,375	2,744	2,453	2,040	1,696	1,309	0,853	0,682	0,530	0,256	0,127	0,063
32	3,622	3,365	2,738	2,449	2,037	1,694	1,309	0,853	0,682	0,530	0,256	0,127	0,063
33	3,611	3,356	2,733	2,445	2,035	1,692	1,308	0,853	0,682	0,530	0,255	0,127	0,063
34	3,601	3,348	2,728	2,441	2,032	1,691	1,307	0,852	0,682	0,529	0,255	0,127	0,063
35	3,591	3,340	2,724	2,438	2,030	1,690	1,306	0,852	0,682	0,529	0,255	0,127	0,063
α	0,001	0,002	0,01	0,02	0,05	0,1	0,2	0,4	0,5	0,6	0,8	0,9	0,95

v grados de libertad	α/2												
	0,0005	0,001	0,005	0,01	0,025	0,05	0,1	0,2	0,25	0,3	0,4	0,45	0,475
36	3,582	3,333	2,719	2,434	2,028	1,688	1,306	0,852	0,681	0,529	0,255	0,127	0,063
37	3,574	3,326	2,715	2,431	2,026	1,687	1,305	0,851	0,681	0,529	0,255	0,127	0,063
38	3,566	3,319	2,712	2,429	2,024	1,686	1,304	0,851	0,681	0,529	0,255	0,127	0,063
39	3,558	3,313	2,708	2,426	2,023	1,685	1,304	0,851	0,681	0,529	0,255	0,126	0,063
40	3,551	3,307	2,704	2,423	2,021	1,684	1,303	0,851	0,681	0,529	0,255	0,126	0,063
41	3,544	3,301	2,701	2,421	2,020	1,683	1,303	0,850	0,681	0,529	0,255	0,126	0,063
42	3,538	3,296	2,698	2,418	2,018	1,682	1,302	0,850	0,680	0,528	0,255	0,126	0,063
43	3,532	3,291	2,695	2,416	2,017	1,681	1,302	0,850	0,680	0,528	0,255	0,126	0,063
44	3,526	3,286	2,692	2,414	2,015	1,680	1,301	0,850	0,680	0,528	0,255	0,126	0,063
45	3,520	3,281	2,690	2,411	2,014	1,679	1,301	0,850	0,680	0,528	0,255	0,126	0,063
46	3,515	3,277	2,687	2,411	2,013	1,679	1,300	0,850	0,680	0,528	0,255	0,126	0,063
47	3,510	3,273	2,685	2,408	2,012	1,678	1,300	0,849	0,680	0,528	0,255	0,126	0,063
48	3,505	3,269	2,682	2,408	2,011	1,677	1,299	0,849	0,680	0,528	0,255	0,126	0,063
49	3,500	3,265	2,680	2,408	2,010	1,677	1,299	0,849	0,680	0,528	0,255	0,126	0,063
50	3,496	3,261	2,678	2,408	2,009	1,676	1,299	0,849	0,679	0,528	0,255	0,126	0,063
51	3,492	3,258	2,676	2,408	2,008	1,675	1,298	0,849	0,679	0,528	0,255	0,126	0,063
52	3,488	3,255	2,674	2,408	2,007	1,675	1,298	0,849	0,679	0,528	0,255	0,126	0,063
53	3,484	3,251	2,672	2,408	2,006	1,674	1,298	0,848	0,679	0,528	0,255	0,126	0,063
54	3,480	3,248	2,670	2,399	2,005	1,674	1,297	0,848	0,679	0,528	0,255	0,126	0,063
55	3,476	3,245	2,668	2,399	2,004	1,673	1,297	0,848	0,679	0,527	0,255	0,126	0,063
56	3,473	3,242	2,667	2,395	2,003	1,673	1,297	0,848	0,679	0,527	0,255	0,126	0,063
57	3,470	3,239	2,665	2,394	2,002	1,672	1,297	0,848	0,679	0,527	0,255	0,126	0,063
58	3,466	3,237	2,663	2,392	2,002	1,672	1,296	0,848	0,679	0,527	0,255	0,126	0,063
59	3,463	3,234	2,662	2,391	2,001	1,671	1,296	0,848	0,679	0,527	0,254	0,126	0,063
60	3,460	3,232	2,660	2,390	2,000	1,671	1,296	0,848	0,679	0,527	0,254	0,126	0,063
120	3,373	3,160	2,617	2,358	1,980	1,658	1,289	0,845	0,677	0,526	0,254	0,126	0,063
∞	3,300	3,098	2,581	2,330	1,967	1,646	1,282	0,842	0,675	0,525	0,253	0,126	0,063
α	0,001	0,002	0,01	0,02	0,05	0,1	0,2	0,4	0,5	0,6	0,8	0,9	0,95

En la tabla podemos observar cómo para cualquier cola de probabilidad (se ha resaltado la correspondiente a 0,025), el incremento del tamaño muestral que lleva asociado el aumento de los grados de libertad, supone una disminución del valor crítico. Para valores grandes de n ($n \geq 120$), la tabla recoge los valores críticos de la distribución normal. Es una consecuencia de que la distribución t_n de Student converge en ley a la distribución Normal cuando $n \rightarrow \infty$. En algunos textos se lee que el intervalo o el test de la t vale para muestras pequeñas. Lo que se quiere afirmar es que cuando el tamaño muestral es muy grande, entonces el valor crítico es muy aproximado al de la $N(0, 1)$.

7.3 Intervalo de confianza de longitud más corta

Ya se ha comentado que se puede incrementar el nivel de confianza simplemente considerando intervalos de confianza de mayor longitud. Además, el intervalo trivial $\theta \in \Theta$, que únicamente especifica que θ es cualquier valor posible, tiene nivel de confianza 1. En la práctica, uno querría establecer como nivel de confianza un número fijo $(1 - \alpha)$, siendo $0 < \alpha < 1$ y, si fuera posible, encontrar un intervalo lo más corto posible de entre todos los intervalos de confianza del mismo nivel, ya que dicho intervalo proporciona más información que el resto. En esta sección se estudia la posibilidad de construir intervalos de confianza de longitud mínima.

Sea X_1, \dots, X_n una m.a.s. de $X \sim f(x, \theta)$, y sea $T(X_1, \dots, X_n, \theta) = T_\theta$ un pivote para θ . Sea $\lambda_1 = \lambda_1(\alpha); \lambda_2 = \lambda_2(\alpha)$, escogidos de manera que

$$P(\lambda_1 < T_\theta < \lambda_2) = 1 - \alpha$$

lo que se puede reescribir como

$$P(\underline{\theta}(X_1, \dots, X_n) < \theta < \bar{\theta}(X_1, \dots, X_n)) = 1 - \alpha.$$

Para cada T_θ , λ_1 y λ_2 se pueden escoger de muchas formas distintas; sería deseable escogerlos de manera que $\bar{\theta} - \underline{\theta}$ toma un valor mínimo. Dicho intervalo es un intervalo de confianza $(1 - \alpha)$ de longitud mínima basado en T_θ . Sin embargo, sería posible encontrar otra variable aleatoria T_θ^* que pudiera producir un intervalo aún más corto. Por tanto, no se puede afirmar que este procedimiento proporcione un intervalo de confianza $(1 - \alpha)$ de longitud mínima de entre todos los intervalos de dicho nivel de confianza. Para T_θ , se utilizan las variables aleatorias más simples que sean una función de un estadístico suficiente para θ .

Nota:

Una alternativa para minimizar la longitud de los intervalos de confianza es minimizar la longitud esperada, $E_\theta(\bar{\theta}(X_1, \dots, X_n) - \underline{\theta}(X_1, \dots, X_n))$. Desafortunadamente, en general no existe un elemento de la clase de todos los intervalos de confianza de nivel $(1 - \alpha)$ que minimice $E_\theta(\bar{\theta}(X_1, \dots, X_n) - \underline{\theta}(X_1, \dots, X_n))$ para todo θ . El procedimiento aplicado para encontrar el intervalo de confianza más corto basado en un pivote, también se puede utilizar para encontrar intervalos que minimicen la longitud esperada. Es importante tener en cuenta que la restricción de insesgadez es natural si se quiere minimizar $E_\theta(\bar{\theta}(X_1, \dots, X_n) - \underline{\theta}(X_1, \dots, X_n))$.

Ejemplo 7.3:

Sea X_1, \dots, X_n una m.a.s. de $X \sim N(\mu, \sigma)$, siendo σ conocida. Entonces, \bar{X} es suficiente para μ , y

$$T_\mu = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

es un pivote para estimar μ .

Por tanto,

$$1 - \alpha = P\left(a < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < b\right) = P\left(\bar{X} - b \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - a \frac{\sigma}{\sqrt{n}}\right).$$

La longitud de este intervalo de confianza es $L = \frac{\sigma}{\sqrt{n}}(b - a)$, valor que se quiere minimizar de manera que:

$$\Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx = \int_a^b \varphi(x) dx = 1 - \alpha,$$

siendo φ y Φ las funciones de densidad y distribución de $N(0, 1)$, respectivamente. Por tanto,

$$\frac{dL}{da} = \frac{\sigma}{\sqrt{n}} \left(\frac{db}{da} - 1 \right)$$

y

$$\varphi(b) \frac{db}{da} - \varphi(a) = 0,$$

obteniéndose

$$\frac{dL}{da} = \frac{\sigma}{\sqrt{n}} \left[\frac{\varphi(a)}{\varphi(b)} - 1 \right].$$

El mínimo se obtiene cuando $\varphi(a) = \varphi(b)$, esto es, cuando $a = b$ ó $a = -b$. Como la igualdad $a = b$ no satisface

$$\int_a^b \varphi(t) dt = 1 - \alpha,$$

se escoge $a = -b$. Por tanto, el intervalo de confianza más corto basado en T_μ , es el intervalo de colas iguales,

$$\left(\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

ó

$$\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right).$$

La longitud de este intervalo es

$$2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

En este caso, se puede plantear el experimento para proporcionar un nivel de confianza prefijado, con un tamaño establecido. Para tener un nivel de confianza de $1 - \alpha$ y una longitud $\leq 2d$, se toma el valor más pequeño de n tal que

$$d \geq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

ó

$$n \geq z_{\frac{\alpha}{2}}^2 \frac{\sigma^2}{d^2}.$$

Otra interpretación que se puede dar es que, si se estima μ mediante \bar{X} , tomando una muestra de tamaño $n \geq z_{\frac{\alpha}{2}}^2 \frac{\sigma^2}{d^2}$, al $100(1 - \alpha)\%$ de confianza, el error de estimación será como mucho de d .

Ejemplo 7.4:

En el ejemplo anterior, supóngase que σ es desconocida. En este caso, se utiliza como pivote

$$T_{\mu} = \frac{\bar{X} - \mu}{\hat{S}} \sqrt{n}.$$

T_{μ} sigue una distribución t de Student con $n - 1$ grados de libertad. Por tanto,

$$1 - \alpha = P\left(a < \frac{\bar{X} - \mu}{\hat{S}} \sqrt{n} < b\right) = P\left(\bar{X} - b \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} - a \frac{\hat{S}}{\sqrt{n}}\right).$$

Se quiere minimizar

$$L = (b - a) \frac{\hat{S}}{\sqrt{n}},$$

sujeto a la condición

$$\int_a^b f_{n-1}(t) dt = 1 - \alpha,$$

donde f_{n-1} es la función de densidad de T_{μ} . Se tiene

$$\frac{dL}{da} = \left(\frac{db}{da} - 1\right) - \frac{\hat{S}}{\sqrt{n}}$$

y

$$f_{n-1}(b) \frac{db}{da} - f_{n-1}(a) = 0,$$

obteniéndose

$$\frac{dL}{da} = \left[\frac{f_{n-1}(a)}{f_{n-1}(b)} - 1 \right] \frac{\hat{S}}{\sqrt{n}}.$$

El mínimo se obtiene cuando $a = -b$ (la otra solución, $a = b$, no es admisible). El intervalo de confianza más corto basado en T_μ es el intervalo de colas iguales,

$$\left(\bar{X} - t_{n-1, \frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{n-1, \frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \right).$$

La longitud de este intervalo es $2t_{n-1, \frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}$, valor que podría ser arbitrariamente alto. El mismo intervalo de confianza minimiza la longitud esperada: $EL = (b - a)c_n \frac{\sigma}{\sqrt{n}}$, siendo c_n una constante determinada para que $ES = c_n \sigma$, y la longitud mínima esperada es $2t_{n-1, \frac{\alpha}{2}} c_n \frac{\sigma}{\sqrt{n}}$.

Ejemplo 7.5:

Sea X_1, \dots, X_n una m.a.s. de $X \sim N(\mu, \sigma)$. Se supone que μ es desconocida, y se quiere encontrar un intervalo de confianza para σ^2 . La elección obvia para un pivote T_{σ^2} viene dada por

$$T_{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2},$$

que tiene una distribución χ^2 con n grados de libertad. Ahora,

$$P\left(a < \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} < b\right) = 1 - \alpha,$$

de manera que

$$P\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{b} < \sigma^2 < \frac{\sum_{i=1}^n (X_i - \mu)^2}{a}\right) = 1 - \alpha.$$

Se quiere minimizar

$$L = \left(\frac{1}{a} - \frac{1}{b}\right) \sum_{i=1}^n (X_i - \mu)^2,$$

sujeto a

$$\int_a^b f_n(t) dt = 1 - \alpha,$$

siendo f_n la función de distribución de una χ^2 con n grados de libertad. Se tiene

$$\frac{dL}{da} = \left(\frac{1}{a^2} - \frac{1}{b^2} \frac{db}{da}\right) \sum_{i=1}^n (X_i - \mu)^2$$

y

$$\frac{db}{da} = \frac{f_n(a)}{f_n(b)},$$

de manera que

$$\frac{dL}{da} = \left[\frac{1}{a^2} - \frac{1}{b^2} \frac{f_n(a)}{f_n(b)}\right] \sum_{i=1}^n (X_i - \mu)^2,$$

que se anula si

$$\frac{1}{a^2} = \frac{1}{b^2} \frac{f_n(a)}{f_n(b)}.$$

Numéricamente, se han obtenido resultados para a y b con 4 decimales, como se puede consultar en la literatura. En la práctica, el intervalo de colas iguales, más simple,

$$\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n, \frac{\alpha}{2}}^2}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n, 1 - \frac{\alpha}{2}}^2} \right)$$

debería usarse, pero no es el más corto.

Si μ es desconocido, se utiliza

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

como pivote. T_{σ^2} tiene una distribución χ_{n-1}^2 . Se puede demostrar que el intervalo de confianza de longitud mínima basado en T_{σ^2} es

$$\left((n-1) \frac{\widehat{S}^2}{b}, (n-1) \frac{\widehat{S}^2}{a} \right),$$

donde a y b son las soluciones de

$$P(a < \chi_{n-1}^2 < b) = 1 - \alpha$$

y

$$a^2 f_{n-1}(a) = b^2 f_{n-1}(b),$$

siendo f_{n-1} la función de densidad de χ_{n-1}^2 . En la práctica, se utiliza el intervalo de colas iguales

$$\left(\frac{(n-1)\widehat{S}^2}{\chi_{n-1, \frac{\alpha}{2}}^2}, \frac{(n-1)\widehat{S}^2}{\chi_{n-1, 1 - \frac{\alpha}{2}}^2} \right).$$

Ejemplo 7.6:

Sea X_1, \dots, X_n una m.a.s. de $X \sim U(0, 1)$. Entonces, $X_{(n)}$ es suficiente para θ , con densidad

$$f_n(y) = n \frac{y^{n-1}}{\theta^n}, 0 < y < \theta.$$

La variable aleatoria $T_\theta = X_{(n)}/\theta$ tiene como función de densidad

$$h_n(t) = nt^{n-1}, 0 < t < 1.$$

Utilizando T_θ como pivote, se puede observar que el intervalo de confianza obtenido es

$$\left(\frac{X_{(n)}}{b}, \frac{X_{(n)}}{a} \right),$$

que tiene longitud

$$L = X_{(n)} \left(\frac{1}{a} - \frac{1}{b} \right).$$

Se minimiza L atendiendo a

$$\int_a^b nt^{n-1} dt = b^n - a^n = 1 - \alpha.$$

Ahora,

$$(1 - \alpha)^{\frac{1}{n}} < b \leq 1,$$

y

$$\frac{dL}{db} = X_{(n)} \left(-\frac{1}{a^2} \frac{da}{db} + \frac{1}{b^2} \right) = X_{(n)} \left(\frac{a^{n+1} - b^{n+1}}{b^2 a^{n+1}} \right) < 0,$$

de manera que el mínimo se alcanza cuando $b = 1$. El intervalo más corto es, por tanto,

$$\left(X_{(n)}, \frac{X_{(n)}}{\alpha^{\frac{1}{n}}} \right).$$

Nótese que

$$EL = \left(\frac{1}{a} - \frac{1}{b} \right) EX_{(n)} = \frac{n\theta}{n+1} \left(\frac{1}{a} - \frac{1}{b} \right),$$

expresión que se minimiza atendiendo a

$$b^n - a^n = 1 - \alpha,$$

donde $b = 1$ y $a = \alpha^{\frac{1}{n}}$. La longitud esperada del intervalo que minimiza EL es

$$\left(\frac{1}{\alpha^{\frac{1}{n}}} - 1 \right) \left(\frac{n\theta}{n+1} \right),$$

que es, así mismo, la longitud esperada del intervalo de confianza más corto basado en T_θ . Es importante tener en cuenta que la longitud del intervalo

$$\left(X_{(n)}, \alpha^{\frac{-1}{n}} X_{(n)} \right)$$

tiene a 0 cuando $n \rightarrow \infty$.

7.4 Intervalos de confianza insesgados y equivariantes

Los tests de inversión son uno de los métodos existentes para el cálculo de intervalos de confianza. En el tema anterior, se ha demostrado que los test UMP conducen a intervalos de confianza UMA. Sin embargo, los tests UMP generalmente no existen. En tales situaciones, se restringe la consideración a subclases más pequeñas de tests, debido al requerimiento de que las funciones de test tengan algunas propiedades deseables, o se restringe la clase de alternativas a aquellas que son próximas a parámetros de valor nulo. En esta sección se sigue una enfoque similar para la

construcción de intervalos de confianza. Supongamos que $X_1 = x_1, \dots, X_n = x_n$ es una realización muestral de una muestra X_1, \dots, X_n de una población $X \sim f(x, \theta)$, con θ desconocido.

Definición 1:

Una familia $S(x)$ de conjuntos de confianza para un parámetro θ se dice que es insesgada a nivel $1 - \alpha$ si se cumple:

- (1) $P_\theta(S(X_1, \dots, X_n) \ni \theta) \geq 1 - \alpha$.
- (2) $P_\theta(S(X_1, \dots, X_n) \ni \theta') \leq 1 - \alpha, \forall \theta, \theta' \in \Theta, \theta \neq \theta'$.

Si el intervalo $S(X_1, \dots, X_n)$ satisface los puntos (1) y (2) anteriores, se dice que es un intervalo de confianza insesgado de nivel $(1 - \alpha)$. Si una familia de conjuntos de confianza de nivel $(1 - \alpha)$ es UMA en la clase de todos los conjuntos de confianza insesgados de nivel $(1 - \alpha)$, se dice que es una familia de conjuntos de confianza de nivel $(1 - \alpha)$ UMA insesgada (UMAU). En otras palabras, si $S^*(x)$ satisface los puntos (1) y (2) y minimiza la expresión

$$P_\theta(S(X_1, \dots, X_n) \ni \theta'), \forall \theta, \theta' \in \Theta, \theta \neq \theta'$$

entre todas las familias de conjuntos de confianza $S(X_1, \dots, X_n)$ de nivel $(1 - \alpha)$, entonces $S^*(X_1, \dots, X_n)$ es una familia de conjuntos de confianza de nivel $(1 - \alpha)$ UMAU.

Nota:

La Definición 1 dice que la familia de conjuntos de confianza $S(X_1, \dots, X_n)$ para el parámetro θ es insesgada a nivel $(1 - \alpha)$ si la probabilidad de cobertura real es de al menos $(1 - \alpha)$, y la probabilidad de cobertura falsa es de $(1 - \alpha)$ a lo sumo. En otras palabras, $S(X_1, \dots, X_n)$ incluye un valor real del parámetro más a menudo de lo que incluye un valor falso del mismo.

Teorema 1:

Sea $A(\theta_0)$ la región de aceptación de un test UMP insesgado de tamaño α para $H_0 : \theta = \theta_0$ frente a $H_1 : \theta \neq \theta_0$, para cada θ_0 . Entonces, $S(x) = \{\theta | x \in A(\theta)\}$ es una familia de conjuntos de confianza insesgada de nivel $1 - \alpha$ (UMAU).

Ejemplo 7.7:

Sea X_1, \dots, X_n una m.a.s. de $X \sim N(\mu, \sigma)$, siendo μ y σ desconocidas. Para contrastar $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$, se sabe que el t -test

$$\varphi(x) = \begin{cases} 1, & \frac{|\sqrt{n}(\bar{x} - \mu_0)|}{\hat{S}} > c \\ 0, & \text{en otro caso.} \end{cases}$$

siendo $\bar{x} = \sum \frac{x_i}{n}$ y $\hat{S}^2 = (n - 1)^{-1} \sum (x_i - \bar{x})^2$ es UMP insesgado. Se escoge c a partir del requisito de tamaño

$$\alpha = P_{\mu=\mu_0} \left(\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\hat{S}} \right| > c \right),$$

de manera que $c = t_{n-1, \frac{\alpha}{2}}$. Por tanto,

$$A(\mu_0) = \left\{ x \text{ tal que } \left| \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \right| \leq t_{n-1, \frac{\alpha}{2}} \right\}$$

es la región de aceptación de un test insesgado UMP de tamaño α para $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$. Por el Teorema 1 se cumple que

$$S(x) = \{ \mu | x \in A(\mu) \} = \left(\bar{x} - \frac{\widehat{S}}{\sqrt{n}} t_{n-1, \frac{\alpha}{2}} \leq \mu \leq \bar{x} + \frac{\widehat{S}}{\sqrt{n}} t_{n-1, \frac{\alpha}{2}} \right)$$

es una familia de conjuntos de confianza UMA insesgados de nivel $1 - \alpha$.

Si la medida de precisión de un intervalo de confianza fuera su longitud esperada, de manera natural se puede pensar en la consideración de intervalos de confianza insesgados. Pratt (J.W. Pratt, 1931) demostró que la longitud esperada de un intervalo de confianza es la medida de las probabilidades de falsa cobertura.

Teorema 2:

Sea Θ un intervalo de la recta real, y sea $X \sim f(x, \theta)$. Sea $S(X_1, \dots, X_n)$ una familia de intervalos de confianza de nivel $(1 - \alpha)$ de longitud finita, esto es,

$$S(X_1, \dots, X_n) = (\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n)),$$

y supóngase que $\bar{\theta}(X_1, \dots, X_n) - \underline{\theta}(X_1, \dots, X_n)$ es un valor finito (aleatorio). Entonces,

$$\begin{aligned} \int_{\theta' \neq \theta} (\bar{\theta}(x_1, \dots, x_n) - \underline{\theta}(x_1, \dots, x_n)) f(x_1, \dots, x_n, \theta) dx &= \\ &= \int_{\theta' \neq \theta} P_{\theta'}(S(X_1, \dots, X_n) \ni \theta') d\theta', \quad \forall \theta \in \Theta \end{aligned} \quad (7.1)$$

Nota:

Si $S(X_1, \dots, X_n)$ es una familia de intervalos de confianza UMAU de nivel $(1 - \alpha)$, la longitud esperada de $S(X_1, \dots, X_n)$ es mínima. Esto se cumple por el Teorema 2 debido a que, como la parte izquierda de la ecuación (7.1) representa el tamaño esperado, si θ es el valor real de $S(X_1, \dots, X_n)$ y $P_{\theta'}(S(X_1, \dots, X_n) \ni \theta')$ es mínimo (por ser $S(X_1, \dots, X_n)$ UMAU) con respecto a todas las familias de intervalos de confianza insesgados de nivel $(1 - \alpha)$ uniformes en $\theta (\theta \neq \theta')$.

Un análisis más completo de los tests UMP insesgados queda fuera del marco de este tema; por ello, el procedimiento que se propone a continuación es una técnica verdaderamente útil en muchas situaciones para obtener intervalos de confianza insesgados.

Sea X_1, \dots, X_n una m.a.s. de una distribución absolutamente continua con $X \sim f(x, \theta)$. Se pretende encontrar un intervalo de confianza para θ . Supóngase que

$$T(X_1, \dots, X_n, \theta) = T(X, \theta) = T_\theta$$

es un pivote. Se supone que

$$P(\lambda_1(\alpha) < T_\theta < \lambda_2(\alpha)) = 1 - \alpha$$

se puede convertir a

$$P_\theta(\underline{\theta}(X_1, \dots, X_n) < \theta < \bar{\theta}(X_1, \dots, X_n)) = 1 - \alpha.$$

Para que $(\underline{\theta}, \bar{\theta})$ sea insesgado, se debe cumplir:

$$P(\theta, \theta') = P_\theta(\underline{\theta}(X_1, \dots, X_n) < \theta' < \bar{\theta}(X_1, \dots, X_n)) = 1 - \alpha, \text{ si } \theta' \neq \theta. \quad (7.2)$$

y

$$P(\theta, \theta') < 1 - \alpha, \text{ si } \theta' \neq \theta. \quad (7.3)$$

Si $P(\theta, \theta')$ sólo depende de una función γ de θ y θ' , se puede escribir

$$P(\gamma) \begin{cases} = (1 - \alpha), & \text{si } \theta' = \theta, \\ < (1 - \alpha), & \text{si } \theta' \neq \theta. \end{cases} \quad (7.4)$$

Se cumple que $P(\gamma)$ tiene un máximo en $\theta' = \theta$.

Ejemplo 7.8:

Sea X_1, \dots, X_n una m.a.s. de $X \sim N(\mu, \sigma)$. Se pretende encontrar un intervalo de confianza insesgado para σ . Entonces,

$$T_\sigma = \frac{(n-1)\widehat{S}^2}{\sigma^2}$$

tiene una distribución χ_{n-1}^2 , y se cumple

$$P\left(\lambda_1 < (n-1)\frac{\widehat{S}^2}{\sigma^2} < \lambda_2\right) = 1 - \alpha,$$

de manera que

$$P\left((n-1)\frac{\widehat{S}^2}{\lambda_2} < \sigma^2 < (n-1)\frac{\widehat{S}^2}{\lambda_1}\right) = 1 - \alpha.$$

Entonces,

$$P(\sigma^2, \sigma'^2) = P_{\sigma^2} \left((n-1) \frac{\widehat{S}^2}{\lambda_2} < \sigma'^2 < (n-1) \frac{\widehat{S}^2}{\lambda_1} \right) = P \left(\frac{T_\sigma}{\lambda_2} < \gamma < \frac{T_\sigma}{\lambda_1} \right),$$

siendo $\gamma = \frac{\sigma'^2}{\sigma^2}$ y $T_\sigma \sim \chi_{n-1}^2$. Por tanto,

$$P(\gamma) = P(\lambda_1 \gamma < T_\sigma < \lambda_2 \gamma).$$

Entonces,

$$P(1) = 1 - \alpha \text{ y } P(\gamma) < 1 - \alpha.$$

Por tanto, se necesitan λ_1 y λ_2 tales que

$$P(1) = 1 - \alpha$$

y

$$\left. \frac{dP(\gamma)}{d\gamma} \right|_{\gamma=1} = \lambda_2 f_{n-1}(\lambda_2) - \lambda_1 f_{n-1}(\lambda_1) = 0,$$

donde f_{n-1} es la función de densidad de T_σ . Algunos autores han resuelto numéricamente las expresiones previas, obteniendo valores concretos para λ_1 y λ_2 , de manera que se obtiene el siguiente intervalo de confianza insesgado de nivel $(1 - \alpha)$:

$$\left((n-1) \frac{\widehat{S}^2}{\lambda_2}, (n-1) \frac{\widehat{S}^2}{\lambda_1} \right).$$

Se debe tener en cuenta que en este caso, el intervalo más corto, basado en T_σ , es diferente a los intervalos de confianza clásicos con colas iguales.

7.5 Optimalidad bayesiana. Optimalidad de función de pérdida

Sea X_1, \dots, X_n una m.a.s. de $X \sim f(x, \theta)$, con $\theta \in \Theta$. En el problema de decisión general, una vez que el investigador observa $x = (x_1, \dots, x_n)$, obtiene un conjunto \mathcal{A} de opciones posibles. El problema es encontrar un función de decisión d que minimice el riesgo $R(\theta, \delta) = E_\theta L(\theta, \delta)$ en algún sentido. Entonces, una solución *minimax* requiere la minimización de $\max R(\theta, \delta)$, mientras que una solución de Bayes requiere la minimización de $R(p, \delta) = ER(\theta, \delta)$, siendo p la distribución a priori de Θ . En la teoría del contraste de hipótesis, el conjunto \mathcal{A} contiene dos puntos. a_0 y a_1 ; a_0 se corresponde con la aceptación de $H_0 : \theta \in \Theta_0$; y a_1 se corresponde con el rechazo de H_0 . Se supone la existencia de una función de pérdida definida como

$$\begin{cases} L(\theta, a_0) = a(\theta) & \text{si } \theta \in \Theta_1, \quad a(\theta) > 0 \\ L(\theta, a_1) = b(\theta) & \text{si } \theta \in \Theta_0, \quad b(\theta) > 0 \\ L(\theta, a_0) = 0 & \text{si } \theta \in \Theta_0 \\ L(\theta, a_1) = 0 & \text{si } \theta \in \Theta_1 \end{cases} \quad (7.5)$$

Teorema 3:

Sea X_1, \dots, X_n una m.a.s. de tipo discreto (continuo), con $X \sim f(x, \theta)$, siendo $\theta \in \Theta = \{\theta_0, \theta_1\}$. Sea $p(\theta_0) = p_0$, $p(\theta_1) = 1 - p_0 = p_1$ la probabilidad a priori de una función de densidad en Θ . Una solución de Bayes para el contraste $H_0 : X \sim f(x, \theta_0)$ frente a $H_1 : X \sim f(x, \theta_1)$, utilizando la función de pérdida definida en la ecuación (7.5) consiste en rechazar H_0 si

$$\frac{f(x_1, \dots, x_n, \theta_1)}{f(x_1, \dots, x_n, \theta_0)} \geq \frac{bp_0}{ap_1},$$

es decir, se rechazará H_0 si la verosimilitud de los resultados obtenidos bajo la hipótesis alternativa es suficientemente grande respecto a la verosimilitud respecto a la hipótesis nula.

Demostración:

Se trata de encontrar un δ que minimice la expresión

$$R(p, \delta) = bp_0P_{\theta_0}(\delta(X_1, \dots, X_n) = a_1) + ap_1P_{\theta_1}(\delta(X_1, \dots, X_n) = a_0).$$

Ahora,

$$R(p, \delta) = E_{\theta}R(\theta, \delta) = E[E_{\theta}(L(\theta, \delta)|X)],$$

con lo que es suficiente con minimizar el valor

$$E_{\theta}(L(\theta, \delta)|X),$$

ya que el operador esperanza es creciente.

La distribución a posteriori de θ viene dada por

$$\begin{aligned} h(\theta|x_1, \dots, x_n) &= \frac{p(\theta)f(x_1, \dots, x_n, \theta)}{\sum_{\theta} f(x_1, \dots, x_n, \theta)p(\theta)} = \\ &= \frac{p(\theta)f(x_1, \dots, x_n, \theta)}{p_0f(x_1, \dots, x_n, \theta_0) + p_1f(x_1, \dots, x_n, \theta_1)} = \\ &= \begin{cases} \frac{p_0f(x_1, \dots, x_n, \theta_0)}{p_0f(x_1, \dots, x_n, \theta_0) + p_1f(x_1, \dots, x_n, \theta_1)} & \text{si } \theta = \theta_0 \\ \frac{p_1f(x_1, \dots, x_n, \theta_1)}{p_0f(x_1, \dots, x_n, \theta_0) + p_1f(x_1, \dots, x_n, \theta_1)} & \text{si } \theta = \theta_1 \end{cases} \end{aligned}$$

Entonces,

$$E_{\theta}[L(\theta, \delta(X_1, \dots, X_n)|X_1 = x_1, \dots, X_n = x_n)] = \begin{cases} bh(\theta_0|x_1, \dots, x_n), & \theta = \theta_0, \delta(X_1, \dots, X_n) = a_1, \\ ah(\theta_1|x_1, \dots, x_n), & \theta = \theta_1, \delta(X_1, \dots, X_n) = a_0 \end{cases}$$

Se rechaza H_0 , esto es, $\delta(X_1, \dots, X_n) = a_1$ si

$$bh(\theta_0|x_1, \dots, x_n) \leq ah(\theta_1|x_1, \dots, x_n),$$

lo cual ocurre si y sólo si

$$bp_0f(x_1, \dots, x_n, \theta_0) \leq ap_1f(x_1, \dots, x_n, \theta_1),$$

lo que completa la demostración.

Nota:

La generalización del teorema anterior es sencilla para el caso en que haya múltiples decisiones. Sea X una variable aleatoria con función de densidad f_θ , siendo θ un parámetro que puede tomar k valores, $\theta_1, \theta_2, \dots, \theta_k$. El problema es observar x y decidir qué valor θ_i es el valor correcto de θ . Así, se define $H_i : \theta = \theta_i, \forall i = 1, \dots, k$, y se asume que $p(\theta_i) = p_i, \forall i = 1, \dots, k$ y $\sum_{i=1}^k p_i = 1$ es la distribución a priori de $\Theta = \{\theta_1, \dots, \theta_k\}$. Sea

$$L(\theta_i, \delta) = \begin{cases} 1 & \text{si } \delta \text{ escoge } \theta_j, \text{ con } i \neq j \\ 0 & \text{si } \delta \text{ escoge } \theta_i. \end{cases} \quad (7.6)$$

El problema es encontrar una regla δ que minimice $R(p, \delta)$. Se deja como ejercicio para el lector probar que la regla óptima de Bayes consiste en aceptar $H_i : \theta = \theta_i, i = 1, \dots, k$ cuando

$$p_i f_{\theta_i}(x) \geq p_j f_{\theta_j}(x), \quad \forall j \neq i, \quad j = 1, \dots, k,$$

donde cualquier punto puede caer en más de una de las regiones asignadas a cada uno de ellos.

Bibliografía

1. Rohatgi, V.K. y Ehsanes Saleh, A.K. Md. (2001). *An Introduction to Probability and Statistics*. John Wiley and Sons.
2. Cepeda-Cuervo, E.; Aguilar, W.; Cervantes, V.H.; Corrales, M.; Díaz, I. y Rodríguez, D. (2008). *Intervalos de confianza e intervalos de credibilidad para una proporción*. Revista Colombiana de Estadística 31(2).