Apuntes para el curso de
Estadística Matemática

Ernesto Barrios Zamudio

3 de enero de 2024

Versión 0.34

Índice

Prefacio

Las condiciones en que continuamos este año 2021 ha motivado el trabajo. La imposibilidad de compartir mis notas personales por su desorden me llevó a hacer manuscritos con la mayoría del material del temario de Cálculo de Probabilidades I. En paralelo, comencé a pasar las notas a una presentación más formal usando LaTeX. La mayoría de las gráficas están hechas con R y algunos de los dibujos con LaTeX mismo o *Mayura*. Este documento es el resultado.

Estas *apuntes* son precisamente eso, unos apuntes o notas para apoyar el curso Estadística Matemática que ofrezco anualmente en ITAM.

Las notas de apoyo para el curso de Cálculo de Probabilidades I y II las encuentra en Barrios (2024a) y Barrios (2024b).

Durante el curso es mi responsabilidad motivar y ligar los distintos temas y en este sentido las notas son de apoyo al desarrollo teórico y técnico de los mismos. No se pretende que los temas sean autocontenidos ni son una versión muy preliminar de algo más elaborado y formal. No es material para ser referenciado.

Cualquier error que identifique, comentario y/o sugerencia serán bienvenido. Diríjalo a Ernesto Barrios `<ebarrios at itam.mx>`.

Ciudad de México, 28 de julio de 2022

1. Introducción

¿Qué es la estadística?[1]

> Esta pregunta se planteó en fecha tan temprana como 1898 –refiriéndose a la Royal Statistical Society– y desde entonces se ha vuelto a plantear muchas veces. La persistencia de la pregunta y la variedad de respuestas que se le han dado a lo largo de los años son por sí mismas un fenómeno notable.Tomadas en conjunto, indican que la persistente perplejidad se debe a que la estadística no es una materia única. La estadística ha cambiado radicalmente desde sus primeros días hasta la actualidad, yendo de ser una profesión que reivindicaba una objetividad tan extrema que los estadísticos sólo reunirían datos –sin analizarlos–, hasta ser una profesión que busca asociarles con los científicos en todas las etapas de la investigación, desde la planeación hasta el análisis. Igualmente, la estadística presenta diferentes rostros de las diferentes ciencias: en algunas aplicaciones, aceptamos los modelos científicos como si provinieran de la teoría matemática; en otras, construimos un modelo que pueda adquirir luego un estatus tan sólido como cualquier construcción newtoniana. En algunas situaciones somos planificadores activos y analistas pasivos; en otras, somos lo opuesto. Con tantas caras, y con las consiguientes dificultades para mantener el equilibrio y evitar tropiezos, no debe sorprender que la pregunta sobre qué es la estadística haya surgido una y otra vez, siempre que se enfrenta un nuevo reto, sean ls estadísticas económicas de la década de 1830, sean las cuestiones biológicas de la de los 1930 o las preguntas imprecisamente planteadas sobre *big data* en los tiempos que corren.
>
> Dada la gran variedad de preguntas, aproximaciones e interpretaciones estadísticas, ¿acaso no existe un núcleo duro de la ciencia de la estadística? Si nos dedicamos de manera central a trabajar en tantas ciencias diferentes, desde el estudio de ls políticas públicas hasta la validación del descubrimiento del bosón de Higgs, y si a veces se nos considera como un simple personal técnico, ¿realmente podemos asumirnos, en un sentido razonable, como practicantes de una disciplina unificada, incluso una ciencia por mérito propio? . . . Intentaré formular siete principios, siete pilares que en el pasado han sostenido nuestra disciplina de diferentes maneras y que prometen hacerlo también en el futuro. Cada uno de ellos fue revolucionario cuando se presentó, y que cada uno se mantiene como un avance conceptual importante y profundo.
>
> I. Agregación. El valor de la reducción dirigida o la compresión de los datos.
> II. Medición de la Información. El decreciente valor de un creciente número de datos.
> III. Verosimilitud. Cómo poner una varita de medir probabilística a lo que hacemos.
> IV. Intercomparación. Cómo usar la variación interna en los datos para ayudar en el punto anterior.
> V. Regresión. Hacer preguntas desde distintas perspectivas puede conducir a respuestas reveladoramente diferentes.
> VI. Diseño. El papel esencial de la planeación de las observaciones.
> VII. Residuos. Cómo todas esas ideas pueden usarse para explorar y comparar explicaciones rivales en la ciencia.

[1] Cita tomada de Stigler (2016), Stigler (2017)

Estadística descriptiva

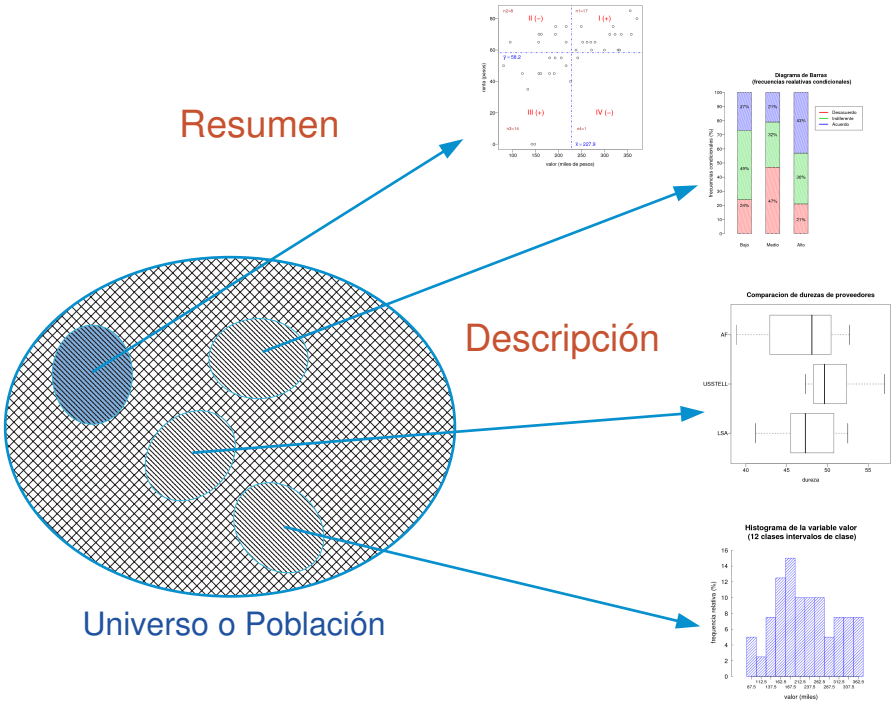


Figura 1: Estadística descriptiva

En el libro de Chihara and Hesterberg (2019) se señala la liga `https://www.bts.gov/topics/airlines-and-airports/quick-links-popular-air-carrier-statistics` donde encontrará estadísticas (información). ¿Cómo describir de manera eficiente la información incluida en ese sitio? ¿Gráficas, tablas, *dashboards*? Su respuesta es tema de investigación y desarrollo de la **estadística descriptiva** o el **análisis exploratorio** de datos. Vea por ejemplo Tukey (2020).

1.1. Deducción e Inferencia

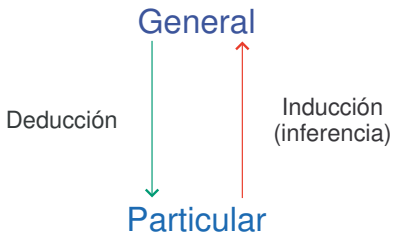


Figura 2: Deducción–Inducción.

Construcción del conocimiento (entre otras).

- **Deducción**. Pasar de lo general a lo particular. por ejemplo, la deducción lógica matemática.

  Ejemplo : A partir de principios (axiomas) concluir para casos particulares. Si $(\Omega, \mathcal{S}, \mathbb{P})$ es un espacio de probabilidad y $A \in \mathcal{S}$, con $\mathbb{P}(A) > 0$, se puede construir la

medida $\mathbb{P}_A$ definida por

$$\mathbb{P}_A(B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad \text{para todo} B \in \mathcal{S}$$

que es una medida de probabilidad.

- **Inferencia**. Pasar de lo particular a lo general. A partir de una muestra se pretende "aprender" sobre la población.

  Hay varias aproximaciones a la inferencia. En este curso nos interesa la **inferencia estadística**, y dentro de ésta hay dos enfoques principales: la inferencia *frecuentista* y la inferencia *bayesiana*. En este curso trabajaremos principalmente la **inferencia frecuentista** y muy brevemente se mencionarán algunos puntos de la **inferencia bayesiana**.

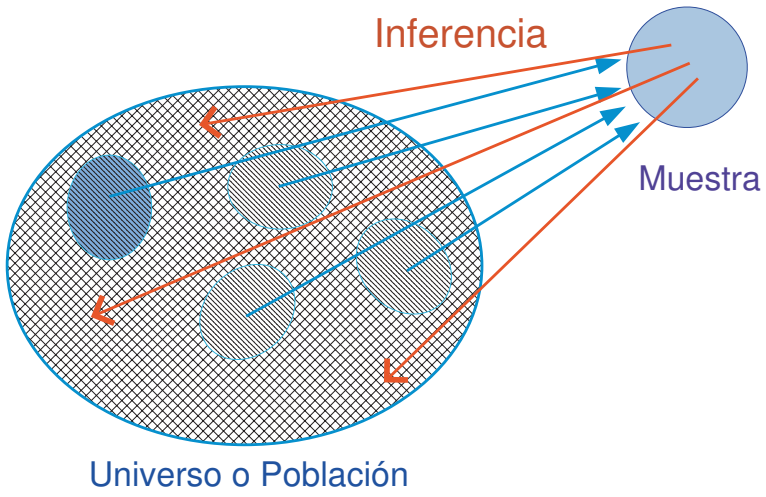


Figura 3: Inferencia estadística

Principios de la inferencia estadística[2]

La mayoría del trabajo estadístico tiene que ver con proveer e implementar métodos para el estudio del diseño, el análisis y la interpretación de los datos. La teoría estadística en principio tiene que ver con los conceptos generales basados en todos los aspectos de tal trabajo, y desde esta perspectiva la teoría formal de la inferencia estadística no es mas que una parte de esa teoría completa. De hecho, desde el punto de vista de las aplicaciones individuales, podría parecer una parte muy pequeña. Es probable que la preocupación sea más concentrada en ver si los modelos han sido formulados razonablemente para atender las preguntas más fructíferas o bien, si los datos están sujetos a errores no apreciados o contaminados y especialmente específico de la materia, en las interpretaciones del análisis y su relación con otros conocimientos en el área.

Y aún así la teoría formal es importante por varias razones. Sin una estructura sistemática los métodos estadísticos para el análisis de los datos se convierte en una colección de trucos que son difíciles de asimilar e interrelacionar entre ellos. El desarrollo de nuevos métodos apropiados para los nuevos problemas se

[2] Cox (2006)

convertiría en cosa de ingenio ad hoc. Por supuesto, dicho ingenio no debe ser subestimado y de hecho, uno de los papeles de la teoría es asimilar, generalizar y quizás modificar y mejorar los frutos de ese ingenio.

Mucho de la teoría tiene que ver con indicar la incertidumbre involucrada en las conclusiones del análisis estadístico y la valoración los méritos relativos a los diferentes métodos de análisis, lo que es importante aún en un nivel de meras aplicaciones para tener una comprensión de las fortalezas y limitaciones de tales discusiones. Esto tiene que ver con temas un poco más filosóficos relacionados con la naturaleza de la probabilidad. Una razón final, y muy buena, porque el estudio de la teoría es en sí interesante.

Inferencia estadística

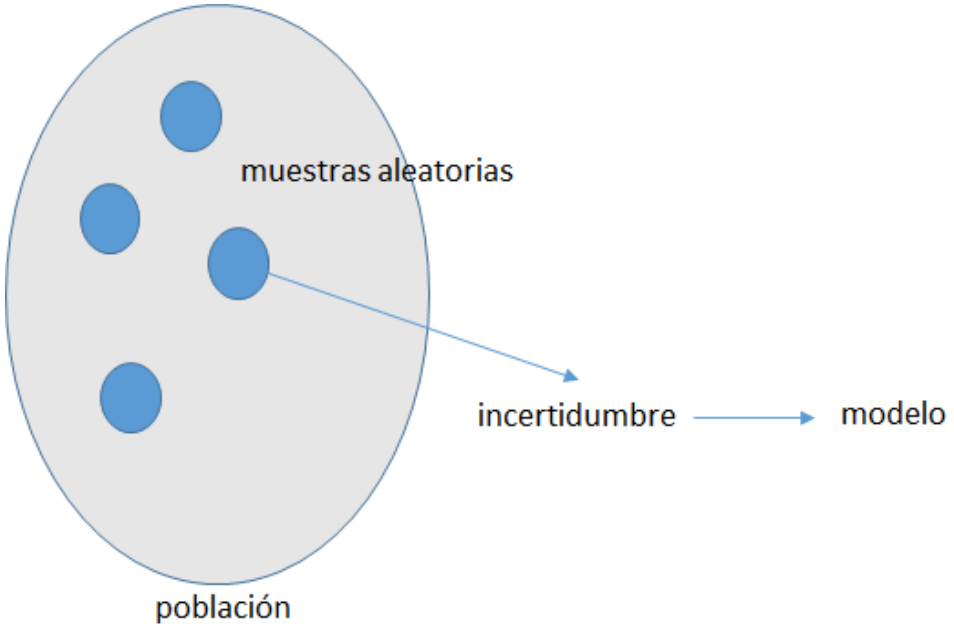


Figura 4: Modelación estadística. El modelo es representado por la variable aleatoria X y sus leyes de probabilidades.

Ejemplo : Considere una moneda y se pregunta por las posibilidades de que en un lanzamiento caiga *águila*.

$$X = \begin{cases} 1 & \text{águila} \\ 0 & \text{sol} \end{cases}$$

Supone que $\mathbb{P}(X = 1) = p$ y que $\mathbb{P}(X = 0) = q = 1 - p$. Entonces, se está asociando $X \sim \text{Ber}(p)$ ¿Qué se puede decir del valor de p? Para responderlo de manera objetiva podemos lanzar varias veces la moneda y *suponer* que cada lanzamiento es independiente de los otros en el sentido de que el resultado de un lanzamiento no influye ni se ve afectado por la salida de los otros. Luego, si además *suponemos* que la probabilidad no cambia entre lanzamientos. El procedimiento lo podríamos *modelar* como una sucesión de *ensayos Bernoulli*, $\{X_i\}$, donde $X_i =$ representa la salida del i-ésimo lanzamiento, entonces, se puede pensar a $\{X_i\}$ como una sucesión de variables aleatorias independientes e idénticamente distribuidas (*v.a.i.i.d.*).

Si se considera lanzar n veces la moneda, $X_1, \ldots, X_n$ son *v.a.i.i.d.* y en este caso representa una **muestra aleatoria (*m. a.*) de tamaño n**.

Si $(\Omega, \mathcal{S}, \mathbb{P})$ representa un espacio de probabilidad (EP) donde está definida la variable aleatoria X,

$$X : \Omega \longrightarrow \mathbb{R}$$
$$\omega \longmapsto X(\omega) = x$$

$X(\omega) = x$ se conoce com una **realización** de la *v. a.* X. En este caso, si $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ representa una *m. a.* de tamaño n de la población X, para $\omega \in \Omega$, $\boldsymbol{X}_n(\omega) = \big(X_1(\omega), \ldots, X_n(\omega)\big) = (x_1, \ldots, x_n)$, representa lo que se dice como la **muestra observada**.

En este ejemplo de la moneda, si el interés es "decir algo" sobre p, es posible *colectar* la información que se obtiene del lanzamiento n veces, $\boldsymbol{X}_n$ y por ejemplo *resumirla* con el número total de águilas, $S_n = \sum_{i=1}^n X_i$ es el número de águilas en n lanzamientos. Note que S_n es a su vez una variable aleatoria que depende de las n salidas individuales. S_n depende exclusivamente de la muestra $\boldsymbol{X}_n$. S_n se dice que es un **estadístico(a)** y $s_n = S_n(\omega) = \sum_{i=1}^n X_i(\omega) = \sum_{i=1}^n x_i$, se dice el **estadístico observado**.

Definición : Se dice que $S = S(\boldsymbol{X})$ es un **estadístico** si es función de la muestra $(\boldsymbol{X}_n)$ y que no depende de parámetros desconocidos.

Ejemplo : Dada la muestra $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, el promedio $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\tilde{X} = \text{mediana}\{X_1, \ldots, X_n\}$ y el rango $R = \text{máx}\{X_i\} - \text{mín}\{X_i\}$, son ejemplos de estadísticos. Por otro lado, $\sum_{i=1}^n (X_i - \mu)^2$ no es un estadístico pues depende del parámetro μ desconocido.

Si supone que $\boldsymbol{X}_n$ es una muestra aleatoria de $X_i \sim \text{Ber}(p)$, entonces $S_n = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$. La distribución del estadístico S_n se dice la **distribución muestral** de S_n, en este caso $\text{Bin}(n, p)$.

Recordar, si $X_i \sim \text{Ber}(p)$, entonces $\mathbb{E}[X_i] = p$ y $\text{var}(X_i) = p(1-p)$. Sea $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, entonces $\mathbb{E}[\bar{X}_n] = p$ y $\text{var}(\bar{X}_n) = p(1-p)/n$, por lo que sería razonable *aproximar* el verdadero (pero desconocido) valor de p por el observado $\bar{x}_n = \bar{X}_n(\omega) = \frac{1}{n}\sum_{i=1}^n x_i$. $\bar{X}$, es un estadístico pues es función de la muestra y además es un **estimador** de p y $\bar{x}_n$ es una **estimación** de p.

Note que siendo que $\text{var}(\bar{X}_n) = p(1-p)/n$, a mayor tamaño de muestra n, menor la *dispersión* (varianza) del estimador. De hecho, se tiene que por la *ley de los grandes números* (LGN), $\bar{X}_n \xrightarrow{P} p$, cuando $n \to \infty$. Luego, el promedio (*media muestral*) $\bar{X}_n$, es un excelente estimador del parámetro (*media poblacional*) μ.

Ejemplo : Considere la población de estudiantes y más en particular en el peso W de los estudiantes varones. Ω denota la población de estudiantes y $W(\omega)$ el peso en kilogramos del estudiante $\omega \in \Omega$. Se supone que $W \sim \text{N}(\mu, \sigma^2)$, es el modelo a utilizar y sea $\theta = (\mu, \sigma)$ el vector de parámetros de la distribución normal y $\theta \in \mathbb{R} \times \mathbb{R}^+ = \Theta$ es el **espacio de parámetros o parametral**.

Suponga que por experiencia se puede suponer que $\sigma = 5$ kg. y se desea *estimar* la media de la población (el parámetro) μ. Para esto, se toma una *m. a.* de tamaño n, $\boldsymbol{W}_n = (W_1, \ldots, W_n)$ de $W \sim \text{N}(\theta)$. Sean

$$S_n = \sum_{i=1}^n W_i \qquad \text{y} \qquad \bar{W}_n = \frac{1}{n} S_n$$

Entonces, $\mathbb{E}[\bar{W}_n] = \mu$ y $\text{var}(\bar{W}_n) = \sigma^2/n$. Luego, sería razonable estimar el parámetro μ mediante $\bar{W}_n$. Como en el ejemplo anterior, note que por LGN se tiene que $\bar{W}_n \xrightarrow{P} \mu$, cuando $n \to \infty$.

Por otro lado, siendo la distribución normal simétrica alrededor de la media, μ también localiza a la mediana de la distribución, por lo que un estimador alternativo del parámetros

sería el estadístico *mediana*, $\tilde{\boldsymbol{W}}_n = \text{mediana}\{W_1, \ldots, W_n\}$ y $\tilde{w}_n = \text{mediana}\{w_1, \ldots, w_n\}$ una estimación de μ.

Note que si la muestra seleccionada incluyese un valor atipico (*outlier*), digamos, w_i^*, más alto que $\mu + 5\sigma$, esto se reflejaría en un mayor valor de $\bar{w}_n$, *sobrestimando* el valor del parámetro μ. Por otro lado, la mediana muestral $\tilde{W}_n$ no se vería afectada por w_i^*. Que un estimador sea "poco sensible" se dice que es **robusto**. En este ejemplo, $\tilde{W}$ es un estimador de μ *robusto a valores atípicos*.

Un problema es que a diferencia de $\bar{W}$, no conocemos la distribución muestral de $\tilde{W}$. No conocemos su media ni su varianza. Luego, no se puede decir nada sobre su "puntería" (media) ni de su "precisión" (varianza). Sin embargo, podríamos ganar cierta idea por medio de ejercicios de simulación y *bootstrap* (**inferencia por computadora**).

Los ejemplos anteriores presentan el problema de **estimación de parámetros**, lo que da lugar a:

- distribuciones muestrales; métodos de estimación; propiedades de los estimadores; comparación de estimadores; estimación puntual; estimación por intervalos.

Ejemplo : ¿Quienes sacan mejor calificación en los cursos de Cálculo de Probabilidades, ellas o ellos?

Para abordar el problema se considera el siguiente modelo: Sean X y Y variables aleatorias independientes que representan las calificaciones de las mujeres y de los hombres respectivamente y sean μ_X y μ_Y las medias de las correspondientes distribuciones. Luego la pregunta se pudiera plantear con las siguientes **hipótesis**

$$\text{H}_0 : \mu_X = \mu_Y \qquad \textit{vs.} \qquad \text{H}_1 : \mu_X > \mu_Y$$

En palabras: la *hipótesis nula* H_0 supone que no hay diferencia entre ellas y ellos, mientras que la *hipótesis alternativa* H_1 supone que ellas obtienen calificaciones más altas que ellos.

Ahora bien, para intentar responder la pregunta de manera objetiva se tomarían muestras aleatorias de ambas poblaciones, $\boldsymbol{X}$ y $\boldsymbol{Y}$, y por ejemplo, se compararían los promedios $\bar{D} = \bar{X} - \bar{Y}$. ¿La diferencia es *importante*? ¿Qué tanto? Note que *siempre* corre el riesgo de concluir en **error**. Por ejemplo, declarar que aunque las muestras sugirieran que no hay diferencia entre los dos grupos, en realidad ellas tienen calificaciones más altas que ellos. O bien, concluir que ellas son mejores porque se observó que $\bar{D} > 0$, pero eso no es generalizado.

El tipo de preguntas que se presentan en el ejemplo anterior y bajo diferentes escenarios y distribuciones también son tema del la inferencia estadística y se abordarán en la segunda parte del curso.

1.2. Ejercicios

Refiérase a la Lista de Ejercicios 1.

Textos de apoyo

Chihara and Hesterberg (2019); Mood, Graybill, and Boes (1974); Wackerly, Mendenhall III, and Scheaffer (2008).

2. Elementos de Cálculo de Probabilidades I y II

2.1. Espacios de probabilidad

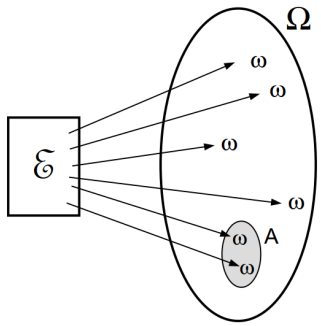


Figura 5: Experimento aleatorio $\mathcal{E}$ arroja salidas impredecibles ω. El espacio muestral Ω es el conjunto de todas las posibles resultados ω. El conjunto A de salidas de interés denota un evento.

- **Experimento aleatorio** $\mathcal{E}$, del que no es posible adelantar la salida.
- **Espacio muestral** Ω. Conjunto de posibles salidas del ω del experimento aleatorio.
- **Familia de eventos** $\mathcal{S}$. σ-álgebra de subconjuntos de Ω, tales que:

  i) Si $A \in \mathcal{S}$, entonces $A^C \in \mathcal{S}$.

  ii) Si $A_1, A_2, \cdots \in \mathcal{S}$, entonces $\cup_{i=1}^{\infty} A_i \in \mathcal{S}$.
- **Espacio medible** $(\Omega, \mathcal{S})$.
- **Conjuntos borelianos de** $\mathbb{R}$**,** $\mathcal{B}(\mathbb{R})$. σ-álgebra de subconjuntos de $\mathbb{R}$ generada por los intervalos $(a, b]$.
- **Medida de probabilidad** sobre el espacio medible $(\Omega, \mathcal{S})$, es una función $\mathbb{P} : \mathcal{S} \to \mathbb{R}$, tal que,

  K_1 : $\mathbb{P}(A) \geq 0$, para todo $A \in \mathcal{S}$.

  K_2 : $\mathbb{P}(\Omega) = 1$.

  K_3 : Si $A_1, A_2, \cdots \in \mathcal{S}$, tales que $A_i \cap A_j = \emptyset$, entonces $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.
- Las propiedades K_1, K_2, K_3 anteriores se dicen los **axiomas de probabilidad o de Kolmogorov**.
- A partir de los axiomas se derivan una serie de propiedades como:
  - $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$.
  - $\mathbb{P}(\emptyset) = 0$.
  - $\mathbb{P}(A \backslash B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$.
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Probabilidad condicional.

- Si $(\Omega, \mathcal{S}, \mathbb{P})$ es un espacio de probabilidad y $A \in \mathcal{S}$ tal que $\mathbb{P}(A) > 0$, se define

$$\mathbb{P}_A(B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \mathbb{P}(B|A), \quad \text{para todo } B \in \mathcal{S}$$

  $\mathbb{P}_A(B) = \mathbb{P}(B|A)$ se dice **probabilidad condicional** de B dado A y es una medida de probabilidad pues satisface los axiomas de Kolmogorov.

- **Regla de la multiplicación**. $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$.
- A y B se dicen **eventos independientes** si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
- $A_1, A_2, \dots$ se dicen eventos **mutuamente independientes** si para todo n se cumple que $\mathbb{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i)$.
- Sean $A_1, A_2, \dots \in \mathcal{S}$. $\{A_i\}$, se dice una **partición de** Ω si: *i)* $A_i \cap A_j = \emptyset$; *ii)* $\cup_i A_i = \Omega$.
- **Teorema de Probabilidad Total**. Sea $\{A_i\}$ una partición de Ω. Entonces, para todo evento B,

$$\mathbb{P}(B) = \sum_{j=1} \mathbb{P}(B|A_j)\mathbb{P}(A_j)$$

- **Regla de Bayes**. Sea $\{A_i\}$ una partición de Ω. Entonces, para todo evento B,

$$\mathbb{P}(A_k|B) = \frac{\mathbb{P}(B|A_k)\mathbb{P}(A_k)}{\sum_{j=1} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

2.2. Variables aleatorias

- Sea $(\Omega, \mathcal{S}, \mathbb{P})$ un *EP*. La función real X con dominio en Ω se dice **variable aleatoria** (*v. a.*) si para todo $x \in \mathbb{R}$, la preimagen de x es un evento. Esto es, si para todo $x \in \mathbb{R}$, $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{S}$.
- Sea X una *v. a.* definida en $(\Omega, \mathcal{S}, \mathbb{P})$. Se define su **función de probabilidad acumulada** o **función de distribución** por $F(x) = \mathbb{P}(X \leq x)$, para todo $x \in \mathbb{R}$.
- Sea X *v. a.* con función de distribución F. Entonces, F satisface:
  - $0 \leq F(t) \leq 1$, para todo $t \in \mathbb{R}$.
  - F es no decreciente.
  - $F(-\infty) = 0$, $F(+\infty) = 1$.
  - F es continua por la derecha.
  - Para $a < b$, $\mathbb{P}(a < X \leq b) = F(b) - F(a)$.
  - $\mathbb{P}(X = x) = F(x^+) - F(x^-)$.
- Sea X *v. a.* y $S_X = \{x_1, x_2, \dots\}$ tal que $\mathbb{P}(X = x_i) > 0$ y $\sum_{x_i \in S_X} \mathbb{P}(X = x_i) = 1$, X se dice que es una **variable aleatoria discreta** con **soporte** S_X.
- Si $A \in \mathcal{B}(\mathbb{R})$, $\mathbb{P}(X \in A) = \sum_{x_i \in A} \mathbb{P}(X = x_i)$.
- Sea X *v. a.* con *f. p. a.* F. Si existe una función h no negativa tal que para todo $x \in \mathbb{R}$, $\mathbb{P}(X \leq x) = \int_{-\infty}^x h(u)du$, X se dice que es una **variable aleatoria (absolutamente) continua** y h su **función de densidad de probabilidades**.

- Sea X *v. a.* con *f. d. p.* f. El conjunto $S_X = \{x \in \mathbb{R} : f(x) > 0\}$ se dice el **soporte de la distribución**.
- Sea X *v. a.* continua con *f. p. a.* F y *f. d. p.* f. Entonces, , para toda $t \in \mathbb{R}$,
  - $F(t) = \int_{-\infty}^{t} f(x)dx$.
  - $f(t) = \frac{dF(x)}{dx}\Big|_{x=t}$.

Cuantiles de una distribución.

- Sea X *v. a.* continua con función de distribución F y sea $0 < p < 1$. Se define x_p, el **p-ésimo cuantil** de la distribución como aquel tal que $p = F(x_p)$.

primer cuartil	$x_{.25} = q_1$	$p = 0.25$
mediana	$x_{.50} = x_{\text{med}}$	$p = 0.50$
tercer cuartil	$x_{.75} = q_3$	$p = 0.75$
Rango	$R = q_3 - q_1$	

2.3. Momentos de una variable aleatoria

- Sea X *v. a.* con función de probabilidades f y soporte S_X. Se define el **valor esperado** de X por
$$\mathbb{E}[X] = \begin{cases} \sum_{x_i \in S_x} x_i f(X_i) & \text{caso discreto} \\ \int_{\mathbb{R}} x f(x)dx & \text{caso continuo} \end{cases}$$
siempre que la suma o integral sean absolutamente finitas.
- **Teorema del estadístico inconsciente**(TEI). Sea X *v. a.* con función de probabilidad f y g una función tal que $g(X)$ es una *v. a.*. Entonces,
$$\mathbb{E}[g(X)] = \begin{cases} \sum_{x_i \in S_x} g(x_i) f(x_i) & \text{caso discreto} \\ \int_{\mathbb{R}} g(x) f(x)dx & \text{caso continuo} \end{cases}$$
siempre que la suma o integral sean absolutamente finitas.
- Se define el **r-ésimo momento** de X por $\mathbb{E}[X^r]$, siempre que este exista.
- Sea X *v. a.* con r-ésimo momento finito. Entonces X tiene k-ésimo momento finito, para todo $k \leq r$, pero no para $r + 1$ necesariamente.
- Sea X *v. a.* con media $\mu = \mathbb{E}[X]$. Se define la **varianza** de X por
$$\text{var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mathbb{E}^2[X]$$
- **Representación de los momentos**. Sea X con cuarto momento finito.

r	$\mathbb{E}[(X-\mu)^r]$	definición	interpretación
*1	μ	media	localización
2	σ^2	varianza	dispersión
3	η	sesgo	asimetría de la densidad
4	κ	curtosis	densidad plana o picuda

- Se define la **función generadora de momentos (*f. g. m.*)** de X por
$$m(t) = \mathbb{E}[e^{tX}]$$
para toda t siempre que la suma o integral exista en una vecindad de 0,

- Se define la **función característica** de X por
$$\varphi(t) = \mathbb{E}[e^{itX}]$$
con $i = \sqrt{-1}$. La función característica *siempre* existe.
- Si X tiene *f. g. m.* m y *f. c.* φ, entonces $\varphi(t) = m(it)$.
- Sea X *v. a.* con *f. g. m.* m diferenciable. Entonces, para $r = 1, 2, \ldots,$
$$\mathbb{E}[X^r] = \frac{d^r m(t)}{dt^r}\Big|_{t=0}$$
- **Teorema de unicidad de *f. g. m.***. Sean X y Y *v. a.*'s con *f. g. m.* m_X y m_Y, y funciones de distribución F_X y F_Y, respectivamente. Si $m_X(t) = m_Y(t)$ para toda t entonces, $F_X(w) = F_Y(w)$, para todo w punto de continuidad de F.
- Sea X *v. a.* con *f. g. m.* $m_X(t)$ y $a, b \in \mathbb{R}$, entonces
$$m_{a+bX}(t) = e^{at} m_X(bt)$$

2.4. Desigualdades

- **Markov:** Sea X una *v. a.* positiva. Entonces, para todo $k > 0$,
$$\mathbb{P}(X \geq k) \leq \frac{E[X]}{k}$$
- **Chabyshev:** Sea X una *v. a.* con media μ_X y varianza σ_X^2. Entonces, para todo $\epsilon > 0$,
$$\mathbb{P}(|X - \mu_X| \geq \epsilon) \leq \frac{\sigma_X^2}{\epsilon^2}$$
- **Jensen:** Sea X *v. a.* y h función convexa, entonces
$$\mathbb{E}[h((X)] \geq h\left(\mathbb{E}[X]\right)$$

2.5. Vectores aleatorios

- Sea $(\Omega, \mathcal{S}, \mathbb{P})$ un *EP.* $\boldsymbol{X} = (X_1, \ldots, X_n)$ es un **vector aleatorio** si para todo $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $\{\omega \in \Omega : \boldsymbol{X}(\omega) = \boldsymbol{x}\} \in \mathcal{S}$.
- Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$. Se define la ***f. p. a.* conjunta de $\boldsymbol{X}$** por $F(\boldsymbol{x}) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$, para todo $\boldsymbol{x} \in \mathbb{R}^n$. En este contexto, $F_i(x) = \mathbb{P}(X_i \leq x)$ se dice la *f. p. a.* **marginal** de X_i.
- Sea $\boldsymbol{X}$ *v. a.* con *f. p. a.* conjunta F. Si existe una función h no negativa tal que para todo $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$,
$$F(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} h(\boldsymbol{x}) dx_1 \cdots dx_n$$
h se dice una **función de densidad de probabilidad conjunta de $\boldsymbol{X}$**.
- Sea $\boldsymbol{X}$ *v. a.* con *f. p. a.* conjunta F y *f. d. p.* conjunta f. Entonces,
$$f(\boldsymbol{x}) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(\boldsymbol{x})$$
- Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ *v. a.* con *f. d. p.* conjunta f y f_i la correspondiente función marginal de X_i. Entonces,
$$f_i(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\ dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

Distribuciones condicionales.

- Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ con *f. p. a.* conjunta F. $\boldsymbol{X}$ se dice que es de **componentes independientes** $X_1, \ldots, X_n$, si y solo si, para todo $\boldsymbol{x} = (x_1, \ldots, x_n)$ y $\boldsymbol{t} = (t_1, \ldots, x_n)$,
  - $F(\boldsymbol{x}) = F_1(x_1) \cdots F_n(x_n)$.
  - $f(\boldsymbol{x}) = f_1(x_1) \cdots f_n(x_n)$.

  Basta que una se cumpla para que se cumpla la otra.
- Sea (X_1, X_2) un vector aleatorio con *f. d. p.* conjunta f y marginales f_i. Entonces,

$$\mathbb{P}(X_1 \leq x_1 | X_2 = x_2) = \int_{-\infty}^{x_1} h(u)du$$

  donde $h(u) = f(u|x_2) = \dfrac{f(u, x_2)}{f_2(x_2)}$. En esta definición, la función $f(\cdot|x_2)$ se dice la **función de densidad condicional de** X_1 **dado** $X_2 = x_2$.
- **Teorema de Probabilidad Total (**TPT**)** Sea (X, Y) un vector aleatorios con *f. d. p.* conjunta f, marginales f_X, f_Y y condicional $f(x|y)$. Entonces,

$$f_X(x) = \int_{\mathbb{R}} f(x|y) f_Y(y) dy$$

- **Regla de Bayes**

$$f(y|x) = \frac{f(x|y) f_Y(y)}{\int_{\mathbb{R}} f(x|v) f_Y(v) dv}$$

- **Teorema del estadístico inconsciente**. Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ un *v. a.* y sea $g : \mathbb{R}^n \to \mathbb{R}$, una función tal que $Y = g(\boldsymbol{X})$ es una variable aleatoria. Entonces,

$$\mathbb{E}[Y] = \mathbb{E}[g(\boldsymbol{X})] = \int_{\mathbb{R}^n} g(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

- **Esperanza condicional**. Sea (X, Y) un *v. a.*, entonces

$$\mathbb{E}[Y|X = x] = \int_{\mathbb{R}} y f(y|x) dy$$

- **Esperanza total**. Sea (X, Y) un *v. a.*, entonces
  - $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$
  - $\mathrm{var}(Y) = \mathrm{var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathrm{var}(Y|X)]$

Esperanza y covarianza de un vector aleatorio.

- Sea (X, Y) un *v. a.*. Se define la **covarianza** entre X y Y por

$$\sigma_{XY} = \mathrm{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

  y la **correlación lineal** entre X y Y por

$$\rho_{XY} = \mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Sea $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, se definen el **vector de medias** y la **matriz de covarianzas** por $\mu = (\mu_1, \ldots, \mu_n)^T$ y $\Sigma = (\sigma_{ij})$. Explícitamente,

$$\mathbb{E}[X] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \mu \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{pmatrix}$$

  donde $\mu_i = \mathbb{E}[X_i]$ y $\sigma_{ij} = \operatorname{cov}(X_i, X_j)$.

Función generadora de momentos.

- Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ un *v. a.*. Se define la **función generadora de momentos conjunta** por $m(\boldsymbol{t}) = \mathbb{E}[e^{\boldsymbol{t}^T \boldsymbol{X}}]$ y $m_i(t) = \mathbb{E}[e^{tX_i}]$ las correspondientes *f. g. m.* marginales.

- Sea $\boldsymbol{X} = (X_1, X_2)$ un *v. a.* con *f. g. m.* conjunta $m_{12}(t_1, t_2)$ y marginales $m_i(t_i)$. Entonces, $m_1(t) = m_{12}(t, 0)$.

- Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ con *f. p. a.* conjunta F, *f. d. p.* conjunta f y *f. g. m.* conjunta m. $\boldsymbol{X}$ se dice que es de **componentes independientes** $X_1, \ldots, X_n$, si y solo si, para todo $\boldsymbol{x} = (x_1, \ldots, x_n)$ y $\boldsymbol{t} = (t_1, \ldots, x_n)$,

  - $F(\boldsymbol{x}) = F_1(x_1) \cdots F_n(x_n)$.
  - $f(\boldsymbol{x}) = f_1(x_1) \cdots f_n(x_n)$.
  - $m(\boldsymbol{t}) = m_1(t_1) \cdots m_n(t_n)$.

  Basta que una se cumpla para que las otras dos también se cumplan.

- Sean X_1 y X_2 *v. a.*'s independientes con *f. g. m.* m_i, respectivamente. Entonces,

$$m_{X_1+X_2}(t) = m_1(t_1) \cdot m_2(t)$$

- **Teorema de unicidad:** Si X_1 y X_2 tienen *f. g. m.* y es la misma, entonces tienen la misma ley de probabilidades. Esto es, si $m_1(t) = m_2(t)$, para toda t, entonces $F_1(x) = F_2(x)$, para todo punto de continuidad x.

2.6. Transformaciones

- **Teorema de transformación**. Sea X *v. a.* con *f. d. p.* f_X y g una función con inversa $h = g^{-1}$ diferenciable y tal que $\frac{dh}{dy} \neq 0$. Si $Y = g(X)$ es una variable aleatoria. Entonces, f_Y, la *f. d. p.* de Y está dada por

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy} \right|$$

- **Teorema de transformación integral**.

  - Sea X *v. a.* continua con función de distribución F_X. Entonces $U = F_X(X) \sim \text{Unif}(0, 1)$.
  - Sea X con *f. p. a.* F_X invertible, $U \sim \text{Unif}(0, 1)$. Entonces, $Y = F_X^{-1}(U) \sim X$.

Distribuciones Discretas

Distribución	Función masa de probabilidad	Parámetros $p+q=1$	Media μ	Varianza σ^2	Función generadora de momentos $M(t)$
Bernoulli	$f(x)=p^x q^{1-x} I_{\{0,1\}}(x)$	$0<p<1$	p	pq	$q+pe^t$
Binomial	$f(x)=\binom{n}{x} p^x q^{n-x} I_{\{0,1,\ldots,n\}}(x)$	$0<p<1$ $n=1,2,\ldots$	np	npq	$(q+pe^t)^n$
Geométrica	$f(x)=pq^x I_{\{0,1,\ldots\}}(x)$	$0<p<1$	q/p	q/p^2	$p/(1-qe^t)$
Binomial Negativa	$f(x)=\binom{r+x-1}{x} p^r q^x I_{\{0,1,\ldots\}}(x)$	$0<p<1$ $r=1,2,\ldots$	rq/p	rq/p^2	$\left(\frac{p}{1-qe^t}\right)^r$
Poisson	$f(x)=\frac{\lambda^x}{x!} e^{-\lambda} I_{\{0,1,\ldots\}}(x)$	$\lambda>0$	λ	λ	$e^{\lambda(e^t-1)}$

Distribuciones Continuas

Distribución	Función de densidad de probabilidad	Parámetros	Media μ	Varianza σ^2	Función generadora de momentos $M(t)$
Uniforme	$f(x)=\frac{1}{b-a} I_{[a,b]}(x)$	$a,b\in\mathbb{R}$ $a<b$	$(a+b)/2$	$(b-a)^2/12$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Normal o Gaussiana	$f(x)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} I_{\mathbb{R}}(x)$	$\mu\in\mathbb{R}$ $\sigma\in\mathbb{R}^+$	μ	σ^2	$e^{\mu t+\frac{1}{2}\sigma^2 t^2}$
Gamma	$f(x)=\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)} I_{\mathbb{R}^+}(x)$	$\alpha\in\mathbb{R}^+$ $\beta\in\mathbb{R}^+$	$\alpha\beta$	$\alpha\beta^2$	$(1-\beta t)^{-\alpha}$ $t<1/\beta$
Ji-Cuadrada χ^2_n	$f(x)=\frac{x^{\frac{n}{2}-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)} I_{\mathbb{R}^+}(x)$	$n\in\mathbb{N}$	n	$2n$	$(1-2t)^{-n/2}$ $t<1/2$
t-Student	$f(x)=\frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1+x^2/n\right)^{-\frac{n+1}{2}} I_{\mathbb{R}}(x)$	$n\in\mathbb{N}$	0	$n/(n-2)$ $n>2$	no existe
F	$f(x)=\frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2}\frac{x^{(m-2)/2}}{[1+(m/n)x]^{(m+n)/2}} I_{\mathbb{R}^+}(x)$	$m,n\in\mathbb{N}$	$n/(n-2)$ $n>2$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ $n>4$	no existe
Weibull	$f(x)=\alpha\beta x^{\beta-1}e^{-\alpha x^\beta} I_{\mathbb{R}^+}(x)$	$\alpha>0$ $\beta>0$	$\alpha^{-1/\beta}\Gamma(1+1/\beta)$	$\alpha^{-2/\beta}\Gamma(1+2/\beta)$ $-\Gamma^2(1+1/\beta)$	$\mathbb{E}[X^r]=\alpha^{-r/\beta}\cdot\Gamma(1+r/\beta)$
Pareto	$f(x)=\frac{\theta\alpha^\theta}{x^{\theta+1}} I_{[\alpha,\infty)}(x)$	$\alpha>0$ $\theta>0$	$\frac{\alpha\cdot\theta}{\theta-1}$ $\theta>1$	$\frac{\alpha^2\theta}{(\theta-1)^2(\theta-2)}$ $\theta>2$	no existe

Figura 6: Funciones de densidad y generadora de momentos de algunas leyes de probabilidad discretas y continuas.

- **Teorema de transformación**. Sea $\boldsymbol{X}=(X_1,\ldots,X_n)^T$ con *f. d. p.* conjunta $f_{\boldsymbol{X}}$ y sea la función $\boldsymbol{g}:\mathbb{R}^n\to\mathbb{R}^n$ uno-a-uno tal que $\boldsymbol{h}=\boldsymbol{g}^{-1}$ con $|J\boldsymbol{h}|\neq 0$. Sea $\boldsymbol{Y}=\boldsymbol{g}(\boldsymbol{X})$. Entonces,

$$f_{\boldsymbol{Y}}(\boldsymbol{y})=f_{\boldsymbol{X}}(\boldsymbol{h}(\boldsymbol{y}))|J\boldsymbol{h}|$$

donde $|J\boldsymbol{h}|$ denota el valor absoluto del determinante de la matriz jacobiana $(\partial h_i/\partial y_j)$.

- **Ejemplos**

  - Sean $Z \sim \mathrm{N}(0,1)$ y $Y \sim \chi^2_n$ independientes. Defina $T = \sqrt{n}ZY^{-1/2}$ y $U = Y$.

    i) Defina $T = \sqrt{n}ZY^{-1/2}$, $U = Y$ y utilice el teorema de transformación para encontrar la *f. d. p.* conjunta de (T, U).

    iii) Encuentre la distribución marginal de T. Muestre que la *f. d. p.* de T está dada por

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + t^2/n)^{\frac{n+1}{2}}, \qquad \text{para todo } t \in \mathbb{R}$$

  - Sean $X_1, \ldots, X_n, \ldots$, *v.a.i.i.d.*'s con $X_i \sim \mathrm{Ber}(p)$. Entonces, $\sum_{i=1}^n X_i \sim \mathrm{Bin}(n, p)$.

    Solución: $X \sim \mathrm{Ber}(p)$, $m_X(t) = 1 + pe^t$. Entonces, $S_n = \sum_{i=1}^n X_i$ tiene un *f. g. m.* dada por $m_{S_n}(t) = \left[q + pe^t\right]^n$, que corresponde a la *f. g. m.* de una distribución Bernoulli parámetros n y p. Se concluye del teorema de unicidad que $S_n \sim \mathrm{Ber}(n, p)$.

  - Sean $X_1, \ldots, X_n, \ldots$, *v.* a.'s independientes con $X_i \sim \mathrm{Po}(\lambda_i)$. Entonces, $\sum_{i=1}^n X_i \sim \mathrm{Po}(\sum_{i=1}^n \lambda_i)$.

    Solución: $X \sim \mathrm{Po}(\lambda_i)$, $m_i(t) = e^{\lambda(e^t-1)}$. Entonces, $S_n = \sum_{i=1}^n X_i$ tiene un *f. g. m.* dada por $m_{S_n}(t) = e^{(e^t-1)\sum_{i=1}^n \lambda_i}$, que corresponde a la *f. g. m.* de una distribución Poisson parámetro $\sum_{i=1}^n \lambda_i$. Se concluye del teorema de unicidad que $S_n \sim \mathrm{Po}(\sum_{i=1}^n \lambda_i)$.

  - Sean $X_1, \ldots, X_n, \ldots$, *v.* a.'s independientes con $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. Entonces, $\sum_{i=1}^n X_i \sim \mathrm{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

    Solución: $X \sim \mathrm{N}(\mu_i, \sigma_i^2)$, $m_i(t) = \exp\{\mu_i t + \frac{1}{2}\sigma_i^2 t^2\}$. Entonces, $S_n = \sum_{i=1}^n X_i$ tiene un *f. g. m.* dada por $m_{S_n}(t) = \exp\{t\sum_{i=1}^n \mu_i, \frac{1}{2}t^2\sum_{i=1}^n \sigma_i^2\}$, que corresponde a la *f. g. m.* de una distribución normal con media $\sum \mu_i$ y varianza $\sum \sigma_i^2$. Se concluye del teorema de unicidad que $S_n \sim \mathrm{N}(\sum \mu_i, \sum \sigma_i^2)$.

  - Recuerde que $Y \sim \mathrm{Gamma}(\alpha, \beta)$, con **parámetro de forma** α y **parámetro de escala** β, entonces $\mathbb{E}[Y] = \alpha\beta$, $\mathrm{var}(Y) = \alpha\beta^2$ y *f. g. m.* $m_Y(t) = (1 - \beta t)^{-\alpha}$, para $t < 1/\beta$.

    $Y \sim \mathrm{Gamma}(\frac{n}{2}, 2)$ se dice que sigue una distribución **ji-cuadrada con** n **grados de libertad** y se denota por $Y \sim \chi^2_n$.

  - Sean $Y_1, \ldots, Y_n, \ldots$, *v.* a.'s independientes que siguen una distribución Gamma con parámetros de forma α_i y parámetro de escala común β. Luego, $\mathbb{E}[Y_i] = \alpha_i\beta$ y $\mathrm{var}(Y_i) = \alpha_i\beta^2$. Entonces, $\sum_{i=1}^n Y_i \sim \mathrm{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.

  - Sean $Y_1, \ldots, Y_n, \ldots$, *v.* a.'s independientes con $Y_i \sim \chi^2_{\nu_i}$. Entonces, $\sum_{i=1}^n Y_i \sim \chi^2_{\sum \nu_i}$.

  - Sea $Z \sim \mathrm{N}(0,1)$, entonces $Y = Z^2 \sim \chi^2_1$.

    Solución:

$$m_Y(t) = \mathbb{E}[e^{tY}] \overset{TEI}{=} \int_{\mathbb{R}^+} e^{tz^2}\phi(z)dz = (1 - 2t)^{-1/2}$$

    donde ϕ es la *f. d. p.* de la distribución normal estándar. La función $(1 - 2t)^{-1/2}$ constituye la *f. g. m.* de $\mathrm{Gamma}(1/2, 2) \equiv \chi^2_1$. La conclusión se sigue del Teorema de Unicidad de las *f. g. m.*.

- **Ejercicios:** Lista 01, problemas 2–14.

2.7. Estadísticos de orden

Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria (*v.a.i.i.d.*'s) de una población $X \sim f$. Esto es, X_i tiene una *f. d. p.* f. Se define $Y_i = X_{(i)}$ el **i-ésimo estadístico de orden**. Entonces, $Y_1 \leq Y_2 \leq \cdots \leq Y_n$, y si $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ y $\boldsymbol{x} = (x_1, \ldots, x_n)$,

$$f_{\boldsymbol{Y}}(\boldsymbol{x}) = n! \prod_{i=1}^{n} f(x_i)$$

-

estadístico		*f. d. p.*	*f. p. a.*
mín	$X_{(1)}$	$f_1(x) = nf(x)[1-F(x)]^{n-1}$	$F_1(x) = 1 - \big[1-F(x)\big]^n$
máx	$X_{(n)}$	$f_n(x) = nf(x)\big[F(x)\big]^{n-1}$	$F_n(x) = \big[F(x)\big]^n$

- La *f. p. a.* y la *f. d. p.* conjuntas del $(X_{(1)}, X_{(n)})$ son

$$F_{1n}(x,y) = \big[F(y)\big]^n - \big[F(y) - F(x)\big]^n$$
$$f_{1n}(x,y) = n(n-1)f(x)f(y)\big[F(y) - F(x)\big]^{n-2}$$

- Sea $R = \text{máx}\{X_i\} - \text{mín}\{X_i\}$ el rango de la muestra, entonces

$$f_R(r) = n(n-1)\int_{-\infty}^{\infty} f(u)f(r+u)\big[F(r+u) - F(u)\big]^{n-2}\,du$$

Ejercicios :

- Sea $\boldsymbol{U} = (U_1, \ldots, U_n)$ una *m. a.* de $U \sim \text{Unif}(a,b)$. Encuentre la distribución de $U_{(n)} = \text{máx}\{U_1, \ldots, U_n\}$.
- Sea $\boldsymbol{T} = (T_1, \ldots, T_n)$ una *m. a.* de $T \sim \text{Exp}(\theta)$. Encuentre la distribución de $T_{(1)} = \text{mín}\{T_1, \ldots, T_n\}$.
- Sea $\boldsymbol{U} = (U_1, \ldots, U_n)$ una *m. a.* de $U \sim \text{Unif}(0,1)$. Muestre que sus estadísticos de orden siguen distribuciones beta. A saber

$$U_{(k)} \sim \text{Beta}(k, n+1-k)$$

2.8. Método Delta

Vea Dudewicz and Mishra (1988).

Proposición : Sea X una variable aleatoria y suponga que su función generadora de momentos $m_X(t)$ existe para todo $|t| < T$, para algún $T > 0$. Entonces, X tiene todos sus momentos finitos $\mu_k = \mathbb{E}[X^k]$ y $m_X(t)$ se puede expandir alrededor de $t = 0$. A saber,

$$m_X(t) = 1 + \frac{\mathbb{E}[X]}{1!}t + \frac{\mathbb{E}[X^2]}{2!}t^2 + \cdots + \frac{\mathbb{E}[X^k]}{k!}t^k + \mathcal{O}(t^k)$$

donde $\lim\limits_{t\to\infty} \dfrac{\mathcal{O}(t^k)}{t^k} = 0$.

Proposición : Sea X una variable aleatoria con k-ésimo momento finito, $k > 2$ y sea h una función k veces diferenciable, entonces

$$\mathbb{E}[h(X)] = h(\mu) + h^{(2)}(\mu)\frac{\sigma^2}{2} + h^{(3)}(\mu)\frac{\mu^{(3)}}{3!} + \cdots + h^{(k-1)}(\mu)\frac{\mu^{(k-1)}}{(k-1)!} + \frac{1}{k!}\mathbb{E}[h^{(k)}(\xi)(X-\mu)^k]$$

donde $\mu^{(r)} = \mathbb{E}[(X-\mu)^r]$, $r = 3, 4, \ldots$ y ξ se encuentra en una vecindad de μ.

Demostración: Puesto que h es k-veces diferenciable, por el Teorema de Taylor con residuo y expandiendo al rededor de μ,

$$h(x) = \sum_{r=0}^{k-1} \frac{h^{(r)}(\mu)}{r!}(x-\mu)^r + \frac{h^{(k)}}{k!}(\xi)(x-\mu)^k$$

donde ξ está entre x y μ.

Ahora, evaluando la expresión anterior en la *v. a.* X y tomando valores esperados

$$\begin{aligned}\mathbb{E}[h(X)] = h^{(0)}(\mu) + \frac{h^{(1)}}{1!}(\mu)\mathbb{E}[(X-\mu)] + \frac{h^{(2)}}{2!}(\mu)\mathbb{E}[(X-\mu)^2] + \cdots \\ + \frac{h^{(k-1)}}{(k-1)!}(\mu)\mathbb{E}[(X-\mu)^{k-1}] + \frac{1}{k!}\mathbb{E}[h^{(k)}(\xi)(X-\mu)^k]\end{aligned}$$

y se sigue el resultado.

Corolario : Sea X una variable aleatoria con media μ_X y varianza σ_X^2. Si la función h es al menos dos veces diferenciable,

$$E[h(X)] \approx h(\mu_X) + \frac{h^{(2)}(\mu_X)}{2}\sigma_X^2$$

Demostración: Tome $k = 2$ en la proposición anterior.

$$\mathbb{E}[h(X)] = h(\mu_X) + h''(\mu_X)\frac{\sigma_X^2}{2} + \text{Residuo}$$

Corolario : Sea X una variable aleatoria con media μ_X y varianza σ_X^2. Si la función h es al menos dos veces diferenciable,

$$\text{var}[h(X)] \approx \left(h'(\mu_X)\right)^2 \sigma_X^2$$

Demostración: Sea $H = h^2$, $H' = 2hh^{'}$, $H'' = 2hh'' + 2(h')^2$. Entonces, se sigue del corolario anterior que

$$\begin{aligned}\mathbb{E}[H(X)] &\approx H(\mu_X) + \left(H''(\mu_X)\right)\frac{\sigma_X^2}{2} \\ &= h^2(\mu_X) + \left[2h(\mu_X)h''(\mu_x) + 2\left(h'(\mu_X)\right)^2\right]\frac{\sigma_X^2}{2}\end{aligned}$$

$$\begin{aligned}\text{var}(h(X)) &= \mathbb{E}[h^2(X)] - \mathbb{E}^2[h(X)] \\ &\approx h^2(\mu_X) + \sigma^2 h(\mu_X)h''(\mu_X) + \sigma_X^2\left(h'(\mu_X)\right)^2 - \left[h(\mu_X) - h''(\mu_X)\frac{\sigma_X^2}{2}\right]^2 \\ &= \sigma_X^2\left(h''(\mu_X)\right)^2 + \underbrace{\frac{\sigma_X^4}{4}\left(h''(\mu_X)\right)^2}_{\mathcal{O}(\sigma_X^4)} \\ &\approx \left(h'(\mu_X)\right)^2\sigma_X^2\end{aligned}$$

Teorema : Método Delta Sea X una variable aleatoria con media μ_X, varianza σ_X^2 y h una función dos veces diferenciable. Entonces,

- $E[h(X)] \approx h(\mu_X) + \frac{h^{(2)}(\mu_X)}{2}\sigma_X^2$
- $\text{var}[h(X)] \approx \left(h'(\mu_X)\right)^2 \sigma_X^2$

Ejemplo : Sea $U \sim \text{Unif}(0,1)$, $h(x) = \sqrt{X}$. Considere $Y = h(X)$. Mediante el método delta aproxime μ_Y, σ_Y^2 y σ_Y.

Solución: $X \sim \text{Unif}(0,1)$, entonces $\mu_X = 0.5$, $\sigma_X^2 = 1/12 = 0.0833$ y $\sigma_X = 0.2887$.

$$\mathbb{E}[Y] \overset{TEI}{=} \int_0^1 \sqrt{x}dx = \frac{x^{3/2}}{3/2}\Big|_0^1 = \frac{2}{3} = 0.667$$

$$\mathbb{E}[Y^2] \overset{TEI}{=} \int_0^1 (\sqrt{x})^2 dx = \frac{x^{3/2}}{3/2}\Big|_0^1 = \frac{1}{2} = 0.5$$

$$\text{var}(Y) = \mathbb{E}[Y^2] = \mathbb{E}^[Y] = 0.50 - (0.667)^2 = 0.0556$$

$$\sigma_Y = 0.2357$$

Y mediante el método delta

$$h(x) = x^{1/2}; \quad h'(x) = \frac{1}{2}x^{-1/2}; \quad h''(x) = -\frac{1}{4}x^{-3/2}$$

$$\mu_Y \approx h(\mu_X) + h''(\mu_X)\frac{\sigma_X^2}{2} = \sqrt{0.5} + \left[-\frac{1}{4}\left(\frac{1}{2}\right)^{-3/2}\right]\frac{0.0833}{2} = 0.7022$$

$$\sigma_Y^2 \approx \left(h'(\mu_X)\right)^2 \sigma_X^2 = \left(\frac{1}{2}(0,5)^{-1/2}\right)^2 0.0833 = 0.0589$$

$$\sigma_Y \approx 0.2427$$

Ejercicio : Repita el ejemplo para $X \sim \text{Unif}(1,2)$.

2.9. Ejercicios

Refiérase a la Lista de Ejercicios 2, problemas 15-18.

Textos de apoyo

Blitzstein and Hwang (2014); Hoel, Port, and Stone (1971); Mood, Graybill, and Boes (1974).

3. Distribuciones Muestrales

Definición : Z se dice que sigue una distribución **normal estándar** y se denota por $Z \sim \mathrm{N}(0,1)$ si tiene *f. d. p.* dada por

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad \text{para todo } z \in \mathbb{R}$$

Definición : Sea $Z \sim \mathrm{N}(0,1)$, $\mu \in \mathbb{R}$ y $\sigma > 0$. Entonces $X = \mu + \sigma Z$, se dice que sigue una distribución **normal con media μ y varianza σ^2**, y se denota por $X \sim \mathrm{N}(\mu, \sigma^2)$. X tiene *f. d. p.* dada por

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \text{para todo } x \in \mathbb{R}$$

Definición : $Y \sim \text{Gamma}(\frac{n}{2}, 2)$ se dice que sigue una distribución **ji-cuadrada con n grados de libertad** y se denota por $Y \sim \chi^2_n$. Su *f. d. p.* está dada por

$$f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}, \quad \text{para } y > 0$$

Definición : Sean $Z \sim \mathrm{N}(0,1)$ y $Y \sim \chi^2_n$ independientes. Defina $T = \dfrac{Z}{\sqrt{Y/n}}$. Entonces, T tiene una *f. d. p.* dada por

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} (1 + t^2/n)^{-\frac{n+1}{2}}, \qquad \text{para todo } t \in \mathbb{R}$$

Se dice que T sigue la **distribución t-Student con n grados de libertad** y se denota $T \sim t_n$.

Sumas, productos y concientes de variables aleatorias

- Sea $\boldsymbol{X} = (X_1, X_2)$ *v. a.* con *f. d. p.* conjunta f_{12} y marginales f_i. Entonces,
  - $f_{X_1+X_2}(z) = \displaystyle\int_{-\infty}^{\infty} f_{12}(x, z-x)dx$
  - $f_{X_1 \cdot X_2}(z) = \displaystyle\int_{-\infty}^{\infty} \frac{1}{|x|} f_{12}(x, z/x)dx$
  - $f_{X_2/X_1}(z) = \displaystyle\int_{-\infty}^{\infty} |x| f_{12}(x, zx)dx$

  Si además las X_i's son independientes
  - $f_{X_1+X_2}(z) = \displaystyle\int_{-\infty}^{\infty} f_1(x) f_2(z-x)dx$
  - $f_{X_1 \cdot X_2}(z) = \displaystyle\int_{-\infty}^{\infty} \frac{1}{|x|} f_1(x) f_2(z/x)dx$
  - $f_{X_2/X_1}(z) = \displaystyle\int_{-\infty}^{\infty} |x| f_1(x) f_2(zx)dx$

Ejemplo : Sean X_1 y X_2 *v. a.*'s independientes con $X_i \sim \text{Gamma}(\alpha_i, \lambda)$, $i = 1, 2$. Utilice la *f. d. p.* de las suma de *v. a.*'s para mostrar que $X_1 + X_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \lambda)$.

Solución: Para esto, sea $Z = X + Y$.

$f_Z(z) = 0$ para $z < 0$. Sea pues, $z \geq 0$,

$$\begin{aligned}
f_Z(z) &= \int_0^z f_X(x) f_Y(z-x) dx \\
&= \int_0^z \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)} x^{\alpha_1 - 1} e^{-\lambda x} \frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)} (z-x)^{\alpha_2 - 1} e^{-\lambda(z-x)} dx \\
&= \frac{\lambda^{\alpha_1+\alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z x^{\alpha_1 - 1} (z-x)^{\alpha_2 - 1} dx \\
&= \frac{\lambda^{\alpha_1+\alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 (zu)^{\alpha_1 - 1} (z - zu)^{\alpha_2 - 1} z du \qquad ; u = x/z \\
&= \frac{\lambda^{\alpha_1+\alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1+\alpha_2-1} \int_0^1 u^{\alpha_1 - 1} (1-u)^{\alpha_2 - 1} du \\
&= \frac{\lambda^{\alpha_1+\alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1+\alpha_2-1} \cdot B(\alpha_1, \alpha_2) \\
&= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)} z^{\alpha_1+\alpha_2-1} e^{-\lambda z}
\end{aligned}$$

con el uso del cambio de variable $u = x/z$ y $B(\alpha_1, \alpha_2)$ denota la **función Beta**[3] La función f_Z corresponde a la *f. d. p.* de una distribución Gamma con parámetro de forma $(\alpha_1 + \alpha_2)$ y parámetro tasa λ.

Ejemplo : Sean X_1, X_2 *v. a.*'s independientes con $X_i \sim \text{Gamma}(\alpha_i, \beta)$. Determine la función de densidad de $Z = X_2/X_1$.

Solución: $Z = X_2/X_1$, entonces el soporte de Z es $\mathbb{R}^+$. Sea pues $z > 0$.

$$\begin{aligned}
f_Z(z) &= \int_0^\infty x f_1(x) f_2(zx) dx \\
&= \int_0^\infty x \frac{x^{\alpha_1 - 1} \mathrm{e}^{-x/\beta}}{\beta^{\alpha_1}\Gamma(\alpha_1)} \cdot \frac{(zx)^{\alpha_2 - 1} \mathrm{e}^{-zx/\beta}}{\beta^{\alpha_2}\Gamma(\alpha_2)} dx \\
&= \frac{z^{\alpha_2 - 1}}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{1}{K} \int_0^\infty K x^{\alpha_1+\alpha_2-1} \mathrm{e}^{-\frac{1+z}{\beta}x} dx \\
&= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{z^{\alpha_2-1}}{(1+z)^{\alpha_1+\alpha_2}}
\end{aligned}$$

donde $K = \dfrac{\left[(1+z)/\beta\right]^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)}$ es la constante normalizadora de la función de densidad de una distribución Gamma con parámetro de forma $\alpha_1+\alpha_2$ y $(1+z)/\beta$ como parámetro tasa. Por lo tanto,

$$f_{X_2/X_1}(z) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{z^{\alpha_2-1}}{(1+z)^{\alpha_1+\alpha_2}} \mathbb{1}_{(0,\infty)}(z)$$

Corolario : Sean $X_i \sim \chi^2_{n_i}$, *v. a.*'s independientes para $i = 1, 2$. Entonces, X_1/X_2 tiene *f. d. p.* dada por

$$f_{X_1/X_2}(w) = \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \frac{w^{\frac{n_1}{2}-1}}{(1+w)^{\frac{n_1+n_2}{2}}} \mathbb{1}_{(0,\infty)}(w)$$

[3] $B(\alpha_1, \alpha_2) = \int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_2-1} du = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$.

Demostración: Recuerde que si $X_i \sim \chi^2_{n_i}$, entonces $X_i \sim \text{Gamma}(n_i/2, 2)$ y aplique el ejemplo anterior.

Corolario : Sean $X_i \sim \chi^2_{n_i}$, *v. a.*'s independientes para $i = 1, 2$. Entonces, $W = \dfrac{X_1/n_1}{X_2/n_2}$ tiene *f. d. p.* dada por

$$f_W(w) = \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \left(\frac{n_1}{n_2}\right)^{n_1/2} \frac{w^{\frac{n_1}{2}-1}}{(1+\frac{n_1}{n_2}w)^{\frac{n_1+n_2}{2}}} \mathbb{1}_{(0,\infty)}(w)$$

Definición : Sean $X_i \sim \chi^2_{n_i}$ independientes y defina $W = \dfrac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$. Entonces F tiene un *f. d. p.* dada por el corolario anterior y se dice que sigue la **distribución F con n_1 y n_2 grados de libertad**. Se denota $W \sim F_{n_1,n_2} \equiv F(n_1, n_2)$.

Ejercicios : Lista 01, problemas 15–24.

Definición : Sea X *v. a.* con media μ y varianza σ^2. Entonces, $Z := \dfrac{X-\mu}{\sigma}$ tiene media cero y varianza 1. Es decir,

$$Z = \frac{X-\mu}{\sigma} \Longrightarrow \quad \mathbb{E}[Z] = 0 \text{ y } \text{var}(Z) = 1$$

La definición de Z es la **estandarización** de la *v. a.* X. Se dice que Z es una *v. a.* estandarizada. De ahí que $N(0,1)$ se diga *normal estándar*.

Definición : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, una muestra aleatoria de X con $\mu = \mathbb{E}[X]$ y $\sigma^2 = \text{var}(X)$. Sean $S_n = X_1 + \cdots + X_n$, $\overline{X}_n = \dfrac{1}{n}S_n$ y $S^2 = \dfrac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$, entonces

estadístico	notación	media	varianza
Suma:	S_n	$\mathbb{E}[S_n] = n\mu$	$\text{var}(S_n) = n\sigma^2$
Media muestral:	$\overline{X}$	$\mathbb{E}[\overline{X}_n] = \mu$	$\text{var}(\overline{X}_n) = \sigma^2/n$
Varianza muestral:	S^2	$\mathbb{E}[S^2] = \sigma^2$	

Note que $(n-1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$. Luego,

$$\begin{aligned}
\mathbb{E}\left[(n-1)S^2\right] &= \mathbb{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \left\{(X_i - \mu) + (\mu - \bar{X})\right\}^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n (\mu - \bar{X})^2 - 2(\bar{X} - \mu)\sum_{i=1}^n (X_i - \mu)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n (X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)\cdot n(\bar{X} - \mu)\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[(X_i - \mu)^2\right] - n\mathbb{E}\left[(\bar{X} - \mu)^2\right] \\
&= n\sigma^2 - n\sigma^2/n \\
&= (n-1)\sigma^2
\end{aligned}$$

Por lo tanto, $\mathbb{E}[S^2] = \sigma^2$.

Resultado : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, una muestra aleatoria de $X \sim \mathrm{N}(\mu, \sigma^2)$. Entonces,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2_n$$

Resultado : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, una *m. a.* de $X \sim \mathrm{N}(\mu, \sigma^2)$. Entonces, $\overline{X}$ y $\boldsymbol{D} = (X_1 - \overline{X}, \ldots, X_n - \overline{X})$ son independientes.

Demostración: Recordar que la *f. g. m.* conjunta de $\boldsymbol{X}$ esta dada por

$$m_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[e^{\boldsymbol{t}^T\boldsymbol{X}}\right] = \mathbb{E}\left[e^{\sum t_i X_i}\right]$$

Luego, la *f. g. m.* conjunta de $(\overline{X}, \boldsymbol{D})$ es

$$m_{(\overline{X},\boldsymbol{D})}(s, \boldsymbol{t}) = \mathbb{E}\left[\exp\left\{s\overline{X} + \boldsymbol{t}^T\boldsymbol{D}\right\}\right] = \mathbb{E}\left[\exp\left\{s\overline{X} + \sum t_i(X - \overline{X})\right\}\right]$$

pero

$$\begin{aligned}\boldsymbol{t}^T\boldsymbol{D} &= \sum t_i(X_i - \overline{X}) = \sum t_i X_i - \overline{X}\sum t_i \\ &= \sum t_i X_i - n\bar{t}\overline{X} = \sum t_i X_i - \bar{t}\sum X_i \\ &= \sum (t_i - \bar{t})X_i\end{aligned}$$

Así,

$$s\overline{X} + \sum t_i(X_i - \overline{X}) = \sum\left[\frac{s}{n} + (t_i - \bar{t})\right]X_i = \sum a_i X_i$$

donde $a_i = \frac{s}{n} + (t_i - \bar{t})$. Ahora, note que $\sum(t_i - \bar{t}) = \sum t_i - n\bar{t} = 0$, luego

$$\sum a_i = \sum\left[\frac{s}{n} + (t_i - \bar{t})\right] = s + \sum(t_i - \bar{t}) = s$$

$$\sum a_i^2 = \sum\left[\frac{s}{n} + (t_i - \bar{t})\right]^2 = \sum\frac{s^2}{n^2} + 2\frac{s}{n}\sum(t_i - \bar{t}) + \sum(t_i - \bar{t})^2 = \frac{s^2}{n} + \sum(t_i - \bar{t})^2$$

Además,

$$\begin{aligned}m_{(\overline{X},\boldsymbol{D})}(s, \boldsymbol{t}) &= m_{\boldsymbol{X}}(\boldsymbol{a}) = \mathbb{E}[e^{\boldsymbol{a}^T\boldsymbol{X}}] = \mathbb{E}[e^{\sum a_i X_i}] \\ &\overset{\text{ind } X_i's}{=} \prod \mathbb{E}[e^{a_i X_i}] = \prod e^{\mu a_i + \frac{1}{2}a_i^2\sigma^2} \\ &= \exp\left\{\mu\sum a_i + \frac{\sigma^2}{2}\sum a_i^2\right\} \\ &= \exp\left\{\mu s + \frac{\sigma^2}{2}\left[\frac{s^2}{n} + \sum(t_i - \bar{t})^2\right]\right\} \\ &= \exp\left\{\mu s + \frac{1}{2}\frac{\sigma^2}{n}s^2\right\} \cdot \exp\left\{\frac{\sigma^2}{2}\sum(t_i - \bar{t})^2\right\} \\ &= m_{\overline{X}}(s) \cdot m_{\boldsymbol{D}}(\boldsymbol{t})\end{aligned}$$

donde los factores dependen exclusivamente de s y $\boldsymbol{t}$ respectivamente, por lo que $\overline{X}$ y $\boldsymbol{D}$ son independientes.

Note que $\overline{X} = \frac{1}{n}\sum X_i \sim \mathrm{N}(\mu, \sigma^2/n)$, por lo que su *f. g. m.* es $m_{\overline{X}}(s) = e^{\mu s + \frac{1}{2}s^2\sigma^2/n}$. Por lo que $m_{\boldsymbol{D}}(\boldsymbol{t}) = e^{\frac{\sigma^2}{2}\sum(t_i - \bar{t})^2}$.

Corolario : $\boldsymbol{X} = (X_, \ldots, X_n)$, *m. a.* de $X \sim \mathrm{N}(\mu, \sigma^2)$. Entonces, $\overline{X}$ y S^2, media y varianza muestrales, son independientes.

Demostración: $S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$ es una función continua de $\boldsymbol{D}$, independiente de $\overline{X}$ por la proposición anterior. Luego, $\overline{X}$ y S^2 son independientes.

Resultado : $\boldsymbol{X} = (X_, \ldots, X_n)$, *m. a.* de $X \sim \mathrm{N}(\mu, \sigma^2)$. Entonces $\dfrac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Demostración:

$$\begin{aligned} W &= \sum\left(\frac{x_i - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum(X_i - \mu)^2 \\ &= \frac{1}{\sigma^2}\sum\Big[(X_i - \overline{X}) + (\overline{X} - \mu)\Big]^2 \\ &= \frac{1}{\sigma^2}\sum(X_i - \overline{X})^2 + \frac{n}{\sigma^2}(\overline{X} - \mu)^2 + \frac{2}{\sigma^2}(\overline{X} - \mu)\sum(X_i - \overline{X}) \\ &= \frac{n-1}{\sigma^2}S^2 + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \\ &=: U + V \end{aligned}$$

Entonces, por proposición anterior U y V son *v. a.*'s independientes. Entonces, $m_W(t) = m_U(t)m_V(t)$, por lo que

$$m_U(t) = \frac{m_W(t)}{m_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2}$$

que corresponde a la *f. g. m.* de una distribución Gamma con $\alpha = \frac{n-1}{2}$ y $\beta = 2$. Se sigue del teorema de unicidad que $\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Proposición : $\boldsymbol{X} = (X_, \ldots, X_n)$, *m. a.* de $X \sim \mathrm{N}(\mu, \sigma^2)$. Entonces, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

Demostración:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{n-1}{\sigma^2}S^2/(n-1)}} \equiv \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1}$$

ya que el numerador y el denominador son independientes como ya se mostró.

Proposición : Sean $\boldsymbol{X}_{n_1} = (X_1, \ldots, X_{n_1})$ *m. a.* $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ y $\boldsymbol{Y}_{n_2} = (Y_1, \ldots, Y_{n_2})$ *m. a.* de $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$ independientes. Entonces,

$$\frac{S_X^2}{S_Y^2} \sim \frac{\sigma_1^2}{\sigma_2^2}F(n_1 - 1, n_2 - 1)$$

Demostración: Sea $S_X^2 = S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2$, $S_Y^2 = S_2^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$. Luego, $\frac{n_i - 1}{\sigma_i^2}S_i^2 \sim \chi^2_{n_i - 1}$, $i = 1, 2$ independientes. Entonces,

$$\frac{\frac{n_1 - 1}{\sigma_1^2}S_1^2/(n_1 - 1)}{\frac{n_2 - 1}{\sigma_2^2}S_2^2/(n_2 - 1)} \sim F(n_1 - 1, n_2 - 1)$$

Por lo tanto, $\dfrac{S_1^2}{S_2^2} \sim \dfrac{\sigma_1^2}{\sigma_2^2}F(n_1 - 1, n_2 - 1)$.

3.1. Ejercicios

Refiérase a la Lista de Ejercicios 3, problemas 1–15.

Textos de apoyo

Blitzstein and Hwang (2014); Hoel, Port, and Stone (1971); Mood, Graybill, and Boes (1974); Wackerly, Mendenhall III, and Scheaffer (2008).

4. Resultados Límite

4.1. Modos de convergencia de variables aleatorias

Sea $(\Omega, \mathcal{S}, \mathbb{P})$ un espacio de probabilidad (EP) y sean X y $X_1, X_2, \ldots$ variables aleatorias definidas sobre ese EP. A continuación se presentan distintos modos de convergencia de sucesiones de va's. Las definiciones fueron tomadas de Roussas (1997).

Definición : Se dice que $\{X_n\}$ **converge casi seguramente** o que **converge con probabilidad 1** a X y se denota por $X_n \xrightarrow{\text{cs}} X$, o bien, $X_n \xrightarrow{\text{cp1}} X$, si para "casi todo" $\omega \in \Omega$ [4], $X_n(\omega) \longrightarrow X(\omega)$.

Esto es, $X_n \xrightarrow{\text{cs}} X$, si $\forall\ \epsilon > 0$ y $\forall\ \omega \in \Omega \backslash D, \exists\ N(\omega, \epsilon)$, tal que

$$\mid X_n(\omega) - X(\omega) \mid < \epsilon, \quad n > N$$

El tipo de convergencia casi segura se le conoce también como *convergencia fuerte*.

Definición : Se dice que $\{X_n\}$ **converge en probabilidad** a X y se denota por $X_n \xrightarrow{P} X$, si para todo $\epsilon > 0$, $\mathbb{P}\left(\mid X_n - X \mid > \epsilon\right) \longrightarrow 0$.

Esto es, si $\forall\ \epsilon > 0$ y $\delta > 0,\ \exists\ N(\epsilon, \delta)$, tal que

$$\mathbb{P}\left(\mid X_n - X \mid > \epsilon\right) < \delta, \quad n > N$$

Note que la convergencia en probabilidad implica que $\mathbb{P}\left(\mid X_n - X \mid \leq \epsilon\right) \longrightarrow 1$. El tipo de convergencia en probabilidad se le conoce también como *convergencia débil*.

Definición : Se dice que $\{X_n\}$ **converge en media cuadrática** a X y se denota por $X_n \xrightarrow{\text{mc}} X$, si $\mathbb{E}\left[\mid X_n - X \mid^2\right] \longrightarrow 0$.

Esto es, si $\forall\ \epsilon > 0\ \exists\ N(\epsilon)$, tal que

$$\mathbb{E}\left[\mid X_n - X \mid^2\right] < \epsilon, \quad n > N$$

Definición : Se dice que $\{X_n\}$ **converge en distribución** a X y se denota por $X_n \xrightarrow{D} X$, si $F_n(x) \longrightarrow F(x)$ para todo $x \in \mathcal{C}(F)$, puntos de continuidad de F, la *f. p. a.* de X y las F_n's las correspondientes de X_n.

Esto es, si $\forall\ \epsilon > 0$ y todo $x \in \mathcal{C}(F)$, $\exists\ N(\epsilon, x)$, tal que

$$\mid F_X(x) - F(x) \mid < \epsilon, \quad n > N$$

Note que convergencia en distribución no implica la convergencia de las correspondientes *f. d. p.* ó *f. m. p.*. Considere por ejemplo,

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < 1 - 1/n \\ 1/2 & \text{si} \quad 1 - 1/n \leq x < 1 + 1/n \\ 1 & \text{si} \quad x \geq 1 + 1/n \end{cases}$$

Entonces, $F_n(x) \longrightarrow F(x) = \mathbb{1}_{[1,\infty)}(x)$, que es la función de distribución que asigna toda la probabilidad al punto $x = 1$. Sin embargo, para todo $x \in \mathbb{R}$,

$$f_n(x) = \frac{1}{2}\mathbb{1}_{\{1-\frac{1}{n}, 1+\frac{1}{n}\}}(x) \longrightarrow f(x) = \frac{1}{2}\mathbb{1}_{\{1\}}(x)$$

que no es *f. d. p.*.

Proposición : Relación entre los modos de convergencia

[4] $X_n(\omega) \longrightarrow X(\omega)$ para *"casi todo"* $\omega \in \Omega$, si la convergencia se da en todo $\omega \in \Omega$, excepto quizá para aquellos $\omega \in D$ pero tal que $\mathbb{P}(D) = 0$.

a). Convergencia en probabilidad implica convergencia en distribución pero no viceversa. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$.

b). Convergencia casi segura implica convergencia en probabilidad pero no viceversa. $X_n \xrightarrow{\text{cs}} X \implies X_n \xrightarrow{P} X$.

c). Convergencia media cuadrática implica convergencia en probabilidad pero no viceversa. $X_n \xrightarrow{\text{mc}} X \implies X_n \xrightarrow{P} X$.

d). Convergencia casi segura *no* implica convergencia cuadrática media *ni* viceversa.

4.2. Otros resultados límite

Teorema de continuidad de Lévy. Sea $\{F_n\}$ una sucesión de funciones de distribución y F una *f. p. a.*. Sean $\{\varphi_n\}$ y φ las correspondientes funciones características. Entonces,

I. Si $F_n(x) \longrightarrow F(x)$, $x \in \mathcal{C}(F)$, puntos de continuidad de F, entonces, $\varphi_n(t) \longrightarrow \varphi(t)$, para todo $t \in \mathbb{R}$.

II. Si para todo $t \in \mathbb{R}$, $\{\varphi_n(t)\}$ converge a una función $g(t)$, continua en $t = 0$, entonces, g es una función característica, y si F es la correspondiente *f. p. a.* entonces, $F_n(x) \longrightarrow F(x)$, para todo $x \in \mathcal{C}(F)$,

Teorema de mapeo continuo. Sea $\{X_n\}$ y X *v. a.*'s y g una función continua. Entonces,

I. Si $X_n \xrightarrow{P} X$, entonces $g(X_n) \xrightarrow{P} g(X)$.

II. Si $X_n \xrightarrow{D} X$, entonces, $g(X_n) \xrightarrow{D} g(X)$.

Proposición : Sean $\{X_n\}$, $\{Y_n\}$, X y Y, *v. a.*'s. Entonces,

I. Si $X_n \xrightarrow{P} X$ y $Y_n \xrightarrow{P} Y$, entonces $X_n + Y_n \xrightarrow{P} X + Y$.

II. Si $X_n \xrightarrow{\text{cm}} X$ y $Y_n \xrightarrow{\text{mc}} Y$, entonces $X_n + Y_n \xrightarrow{\text{mc}} X + Y$.

III. Si $X_n \xrightarrow{P} X$ y $Y_n \xrightarrow{P} Y$, entonces $X_n Y_n \xrightarrow{P} XY$.

Teorema de Slutsky. Sean $\{X_n\}$, $\{Y_n\}$, X y Y, *v. a.*'s y c una constante. Entonces,

I. Si $X_n \xrightarrow{P} X$ y $Y_n \xrightarrow{P} c$, entonces $X_n + Y_n \xrightarrow{P} c + X$

II. Si $X_n \xrightarrow{D} X$ y $Y_n \xrightarrow{D} c$, entonces $X_n Y_n \xrightarrow{D} cX$

Nota: no necesariamente se tiene que si $X_n \xrightarrow{D} X$ y $Y_n \xrightarrow{D} Y$, entonces $X_n + Y_n \xrightarrow{D} X + Y$

4.3. Ley de los grandes números

Sean $X_1, X_2, \ldots$ *v.a.i.i.d.*'s, $S_n = X_1 + \cdots + X_n$ y $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} S_n$.

Ley de los grandes números (LGN).

I. **Versión fuerte (LFGN)**. Sean las X_n's con media finita μ. Entonces,

$$\bar{X}_n \xrightarrow{\text{cs}} \mu$$

Y viceversa, si $\bar{X_n} \xrightarrow{\text{cs}} c$, para alguna constante finita c, entonces, $\mathbb{E}[X]$ es finita y $\mathbb{E}[X] = c$.

II. **Versión débil (LDGN)**. Sean las X_n *v. a.*'s con media finita μ. Entonces,

$$\bar{X}_n \xrightarrow{P} \mu$$

Demostración:

I. De un nivel más elevado que el presente curso ...

II. La demostración se sigue del inciso anterior pues convergencia casi segura implica la convergencia en probabilidad. Sin embargo, si además supone que las *v. a.*'s tienen varianza σ^2 finita , su demostración sigue de la desigualdad de Chebyshev. A saber, sea $\epsilon > 0$,

$$\mathbb{P}\left(\mid \bar{X}_n - \mu \mid \geq \epsilon\right) \leq \frac{\sigma^2_{\bar{X}_n}}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} \longrightarrow 0$$

pues $\mathbb{E}[\bar{X}_n] = \mu$ y $\text{var}(\bar{X}_n) = \sigma^2/n$.

Notas:

1. La constante ϵ puede verse como la precisión deseada en la aproximación de $\bar{X}_n$ a μ. La aproximación es buena para *n suficientemente grande.*
2. $\mathbb{P}\left(\mid \bar{X}_n - \mu \mid \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$
3. Considere que $X_1, X_2, \ldots$ son *v.a.i.i.d.* con $X \sim \text{Ber}(p)$. Entonces, $\mathbb{E}[X] = p$ y $\text{var}(X) = p(1-p)$. Luego, $\mathbb{E}[\bar{X}_n] = p$ y $\text{var}(\bar{X}_n) = p(1-p)/n$, por lo que para toda $0 < p < 1$,
$$\mathbb{P}\left(\mid \bar{X}_n - p \mid \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$
pues la varianza es máxima cuando $p = 1/2$. Esto es, para toda p, $p(1-p) \leq \frac{1}{2}(1-\frac{1}{2}) = \frac{1}{4}$.
4. Considere dados $\epsilon, \delta > 0$ y suponga que se desea que $\mathbb{P}\left(\mid \bar{X}_n - \mu \mid \geq \epsilon\right) \leq \delta$. Se tiene entonces que
$$\mathbb{P}\left(\mid \bar{X}_n - \mu \mid \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2} \leq \delta \quad \Longrightarrow \quad n \geq \frac{p(1-p)}{\delta\epsilon^2}$$
esto es, la condición $\mathbb{P}\left(\mid \bar{X}_n - \mu \mid \geq \epsilon\right) \leq \delta$ se satisface para $n \geq \frac{p(1-p)}{\delta\epsilon^2}$.
En el caso de *máxima incertidumbre* $p = 1/2$. se tiene $n \geq 1/4\delta\epsilon^2$.
5. Si $\epsilon = 0.05$ y $\delta = 0.10$, entonces $\delta\epsilon^2 = 2.5 \times 10^{-4}$, por lo que
$$n \geq \frac{1}{4(2.5 \times 10^{-4})} = 1000$$
Si por alguna razón se espera que $p \approx 0.30$, $p(1-p) = 0.21$ y
$$n \geq \frac{0.2}{2.5 \times 10^{-4}} = 840$$
y si $p \approx 0.15$, entonces $n \geq 510$.

4.4. Teorema central del límite

Sean $X_1, X_2, \ldots$ *v.a.i.i.d.*'s con media común μ_X y varianza σ_X^2 finita.
Recordar,

$$S_n = \sum_{i=1}^{n} X_i \qquad \bar{X}_n = S_n/n$$
$$\mathbb{E}[S_n] = n\mu_X \qquad \mathbb{E}[\bar{X}_n] = \mu_X$$
$$\mathrm{var}(S_n) = n\sigma_X^2 \qquad \mathrm{var}(\bar{X}_n) = \sigma_X^2/n$$

Luego, cuando $n \longrightarrow \infty$, se sigue que

$$\mathbb{E}[S_n] \longrightarrow \pm\infty \quad ; \quad \mathbb{E}[\bar{X}_n] \longrightarrow \mu$$
$$\mathrm{var}(S_n) \longrightarrow +\infty \quad ; \quad \mathrm{var}(\bar{X}_n) \longrightarrow 0$$

Note al crecer n, en el caso de la suma S_n, ésta "explota", mientras que la *media muestral* $\bar{X}_n$, se "colapsa" en el sentido que converge a la constante μ.

La versión del teorema central del límite que aquí se presenta es conocida como de Lindberg-Lévy.

Teorema central de límite (TCL) Sean S_n y $\bar{X}_n$ como antes. Entonces, para las variables estandarizadas $Z_n = \frac{S_n - n\mu_X}{\sqrt{n\sigma_X^2}} = \frac{\bar{X}_n - \mu_X}{\sqrt{\sigma_X^2/n}}$, se tiene que

$$Z_n \xrightarrow{D} Z \sim \mathrm{N}(0,1)$$

Demostración: Sin pérdida de generalidad suponga que las propias X_i's ya son estandarizadas, $\mathbb{E}[X_i] = 0$ y $\mathrm{var}(X_i) = 1$ y suponga que X tiene función generadora de momentos m_X finita. Entonces,

i). $m_X(0) = \mathbb{E}[e^{0\cdot X}] = 1$.

ii). $m'_X(0) = \dfrac{d}{dt} m_X(t)|_{t=0} = \mathbb{E}[X] = 0$

iii). $m''_X(0) = \dfrac{d^2}{dt^2} m_X(t)|_{t=0} = \mathrm{var}(X) = 1$

Se tiene también que

i). $m_{S_n}(t) = \big[m_X(t)\big]^n$, por ser *v.a.i.i.d.*.

ii). $m_{\bar{X}_n}(t) = \big[m_X(t/n)\big]^n$.

iii). $m_{\sqrt{n}\bar{X}_n}(t) = \big[m_X(t/\sqrt{n})\big]^n$.

Sea ahora $L(t) = \log(m_X(t))$. Entonces,

i). $L(0) = 0$.

ii). $L'(0) = \dfrac{d}{dt}\log(m_X(0)) = \left.\dfrac{m'_X(t)}{m_X(t)}\right|_{t=0} = 0$

iii). $L''(0) = \dfrac{d^2}{dt^2}\log(m_X(0)) = \left.\dfrac{m_X(t)m''_X(t) - \big[m'_X(t)\big]^2}{m_X^2(t)}\right|_{t=0} = \dfrac{1(1) - 0^2}{1} = 1$

$$\log(m_{\sqrt{n}\bar{X}_n}(t)) = \log\left[m_X(t/\sqrt{n})\right]^n = n\log(m_X(t/\sqrt{n})) = nL(t/\sqrt{n}) = \frac{L(t/\sqrt{n})}{1/n}$$

$$\begin{aligned}
\lim_{n\to\infty} \frac{L(t/\sqrt{n})}{1/n} &= \lim_{n\to\infty} \frac{L'(t/\sqrt{n})(-\frac{1}{2}n^{-3/2})t}{-n^{-2}}, && \text{Regla de L'Hopital} \\
&= \lim_{n\to\infty} \frac{L'(t/\sqrt{n})t}{2n^{-1/2}} \\
&= \lim_{n\to\infty} \frac{L''(t/\sqrt{n})(-\frac{1}{2}n^{-3/2})t^2}{-n^{-3/2}}, && \text{Regla de L'Hopital} \\
&= \frac{t^2}{2} \lim_{n\to\infty} L''(t/\sqrt{n})
\end{aligned}$$

pues $L''(0) = 1$. Por lo tanto,

$$\log(m_{\sqrt{n}\bar{X}_n}(t)) = \frac{L(t/\sqrt{n})}{1/n} \longrightarrow \frac{t^2}{2}$$

y por continuidad

$$m_{\sqrt{n}\bar{X}_n}(t) \longrightarrow e^{t^2/2}$$

que corresponde a la *f. g. m.* de $Z \sim \mathrm{N}(0,1)$. Se sigue del teorema de continuidad de Lévy que

$$\sqrt{n}\bar{X}_n = \frac{\bar{X}}{1/\sqrt{n}} \xrightarrow{D} Z \sim \mathrm{N}(0,1)$$

El teorema se concluye considerando $W_i = \mu_X + \sigma_X X_i$.

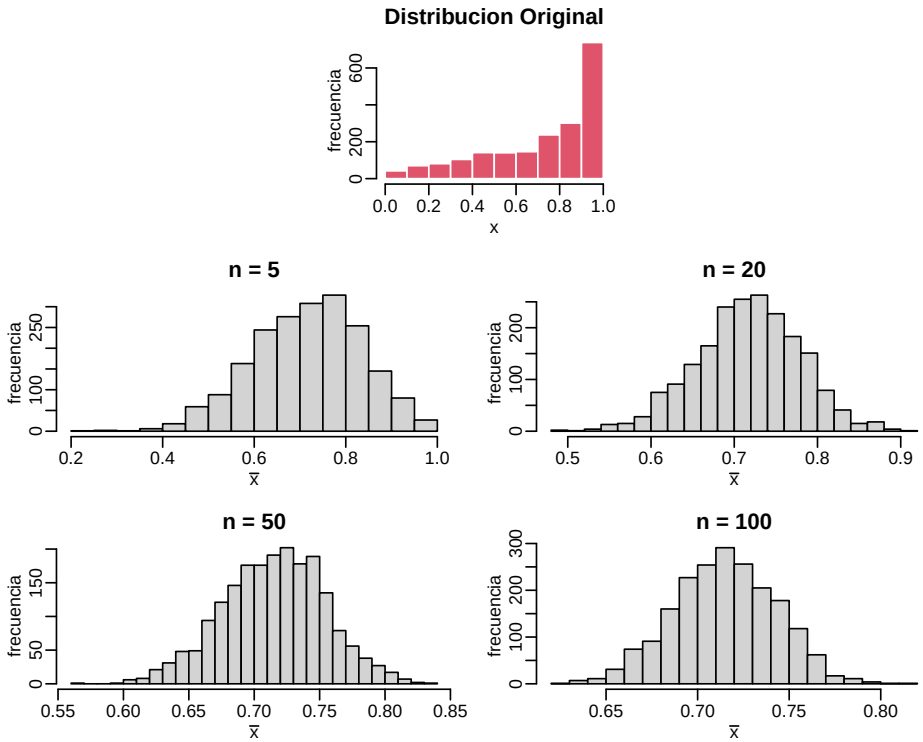


Figura 7: **Teorema central de límite**. Los valores de $\bar{X}_n$ se va distribuyendo a manera de "campana" al rededor de μ conforme crece el tamaño de muestra n. En el panel central se muestra el histograma de la distribución original Beta(1.3,0.5).

La figura 7 muestra el histograma de promedios al considerar muestras de tamaño $n = 5, 20, 50, 100$. Al aumentar el tamaño de la muestra el histograma se acerca más a una curva "normal". El histograma del panel superior representa la distribución original de las X_i's, en este caso, de una distribución Beta.

Nota: En la práctica, basta que el tamaño de la muestra n sea *"suficientemente grande"* para que la aproximación normal sea razonable. Esto es, si $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, y $\bar{X}_n = \frac{1}{n}(X + \cdots + X_n)$, para n suficientemente grande se tiene que

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \dot{\sim} \mathrm{N}(0,1), \quad \text{o bien,} \quad \bar{X}_n \dot{\sim} \mathrm{N}(\mu, \sigma^2/n)$$

¿Qué tan grande debe ser n para que la aproximación sea "razonable"? Dependerá de la distribución de los miembros de la muestra. Para distribuciones más o menos simétricas, una aproximación razonable se logra con $n \geq 30$. Distribuciones sesgadas requerirán de $n \geq 70$ a más. Los promedios de la distribución Beta(0.5,0,6) de la figura 8 parecen distribuidos normal aún para tamaños de muestra $n = 5$.

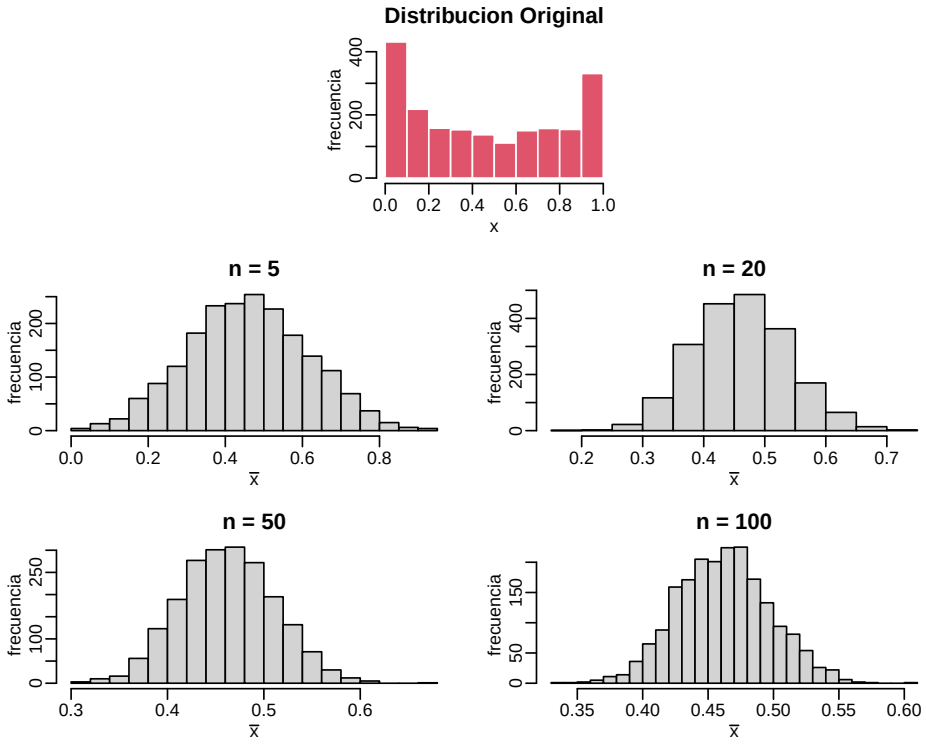


Figura 8: **Teorema central de límite**. En el panel central superior se muestra el histograma de la distribución original Beta(0.5, 0.6).

Ejemplo : El tiempo de vida de unas lámparas se distribuye exponencial con tiempo medio de vida de 10 días. Tan pronto fallan son reemplazadas por otras idénticas. ¿Cuál es la probabilidad de que se necesiten más de 50 lámparas en un año?

Solución: $X \sim \mathrm{Exp}(\theta = 10)$. $S_n = X_1 + \cdots + X_n$.

$$\mathbb{P}\left(S_{50} < 365\right) = \mathbb{P}\left(Z_n < \frac{365 - 50(10)}{10\sqrt{50}}\right) \overset{TCL}{\approx} \Phi(-1.91) = 0.028$$

Ejemplo : Sea $X_1, X_2, \ldots$ *v.a.i.i.d.*'s con media μ y varianza σ^2. Sean $S_n = X_1 + \cdots + X_n$, $\bar{X}_n = \frac{1}{n} S_n$. Aproxime $\mathbb{P}\left(\mid \bar{X}_n - \mu \mid \geq c\right)$.

Solución: Se sigue del teorema central del límite que para n *grande*,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \dot{\sim} \mathrm{N}(0,1)$$

Entonces,

$$\begin{aligned}
\mathbb{P}\left(\mid \bar{X}_n - \mu \mid \geq c\right) &= 1 - \mathbb{P}\left(-c \leq \bar{X}_n - \mu \leq c\right) \\
&= 1 - \mathbb{P}\left(\frac{-c}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{c}{\sigma/\sqrt{n}}\right) \\
&= 1 - \left[\mathbb{P}\left(Z_n \leq \frac{c}{\sigma/\sqrt{n}}\right) - \mathbb{P}\left(Z_n \leq \frac{-c}{\sigma/\sqrt{n}}\right)\right] \\
&\overset{\text{TCL}}{\approx} 1 - \left[\Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-c}{\sigma/\sqrt{n}}\right)\right] \\
&= 2\Phi\left(\frac{-c}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

Ejemplo : Se toma una muestra aleatoria de tamaño n para determinar el porcentaje que votará por el candidato X.

Sean $X_1, X_2, \ldots$ *v.a.i.i.d.*'s $X_i \sim \mathrm{Ber}(p)$. Luego, $\mathbb{E}[X] = p = \mu_X$ y $\mathrm{var}(X) = p(1-p) = \sigma_X^2 \leq 1/4$.

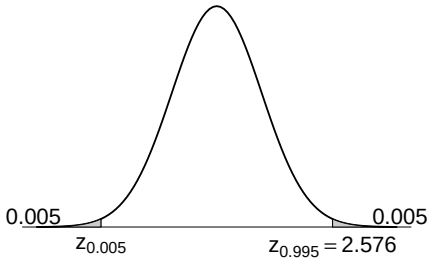
a). Si $n = 900$,

$$\mathbb{P}\left(\mid \bar{X}_n - p \mid \geq 0.025\right) \overset{\text{TCL}}{\approx} 2\Phi\left(\frac{-0.025}{\frac{1}{2}/\sqrt{900}}\right) = 2\Phi(-1.50) = 2(0.067) = 0.1336$$

b). $\mathbb{P}\left(\mid \bar{X}_n - p \mid \geq c\right) = 0.01$. Determine c si $n = 900$.

$$2\Phi\left(\frac{-c}{\sigma/\sqrt{n}}\right) = 0.01$$

$$\begin{aligned}
c &= -z_{.005}\sigma/\sqrt{n} \\
&= 2.576 \cdot \frac{1}{2}/\sqrt{900} \\
&= 0.043
\end{aligned}$$



Note que por simetría al rededor de 0, $z_{0.005} = -2.576 = -z_{0.995}$.

c). Determine n de manera que $\mathbb{P}\left(\mid \bar{X}_n - p \mid \geq 0.025\right) = 0.01$.

$\mathbb{P}\left(\frac{|\bar{X}_n - p|}{\sigma/\sqrt{n}} \geq \frac{0.025}{\sigma/\sqrt{n}}\right) = 0.01$.

Por otro lado, $z_{.995} = 2.576 = \frac{.025}{\frac{1}{2}/\sqrt{n}}$. Así,

$$\sqrt{n} = \frac{2.576}{.025} \cdot \frac{1}{2} = 51.517 \quad \text{y por lo tanto} \quad n \approx 2654$$

4.5. Ejercicios

Refiérase a la Lista de Ejercicios 4, problemas 1–19.

Textos de apoyo

Blitzstein and Hwang (2014); Hoel, Port, and Stone (1971); Mood, Graybill, and Boes (1974); Wackerly, Mendenhall III, and Scheaffer (2008).

5. Estimadores

5.1. Introducción

- En esta sección se presenta la manera de asociar *"leyes de probabilidad"* a observaciones.
- Muchas distribuciones dependen de pocos parámetros, que cuando se conocen sus valores las distribuciones quedan completamente determinadas. Por ejemplo, si $N \sim \text{Po}(\lambda)$ y se sabe que $\lambda = 3$, entonces la distribución está totalmente definida.
- El problema es ¿cómo se sabe que $\lambda = 3$? La más de las veces no se sabe y hay que *estimarlo de los datos*.
- Una vez estimados los parámetros se debe revisar qué tan bien la distribución estimada *ajusta los datos*. (*Pruebas de bondad de ajuste*.)
- En esta sección se verán ciertos métodos de estimación y algunas de sus propiedades.

Ejemplo: Ajuste de datos Poisson[5]

La tabla 1 muestra la emisión de partículas radioactivas en intervalos de 10 segundos. $N = 1207$ intervalos (independientes). 18 de ellos tuvieron 0, 1 o 2 emisiones; 28 de ellos tuvieron 3 emisiones; 56 tuvieron 4; ... 5 intervalos tuvieron 17 o más emisiones.

Tabla 1: Número de intervalos observados y estimados con un número n de partículas emitidas.

n	Observados	Esperados	χ^2
0−2	18	12.2	2.76
3	28	27.0	0.04
4	56	56.5	0.01
5	105	94.9	1.07
6	126	132.7	0.34
7	146	159.1	1.08
8	164	166.9	0.05
9	161	155.6	0.19
10	123	130.6	0.44
11	101	99.7	0.02
12	74	69.7	0.27
13	53	45.0	1.42
14	23	27.0	0.59
15	15	15.1	0.00
16	9	7.9	0.15
17+	5	7.1	0.62
	1207	1207	9.05

Si supone que la emisión de partículas sigue aproximadamente un distribución Poisson con λ, el número medio de emisiones por intervalo (de 10 segundos), entonces la probabilidad de tener k emisiones es

$$\pi_k = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0, 1, 2, \ldots$$

[5] Rice (2007).

El promedio *observado* de emisiones por intervalo de 10 segundos fue $\hat{\lambda} = \sum k_i n_i / N = 8.392$[6]

Note que las frecuencias observadas en los $N = 1207$ intervalos serán distintas si se volviese a observar otros tantos intervalos. Es la misma situación al lanzar 20 veces una moneda. Dos series de 20 volados no tienen porque tener la misma secuencia de águilas y soles. Una segunda sucesión de intervalos seguramente tendría un número promedio de emisiones distinto, digamos $\hat{\lambda}_2$. Luego, dependiendo de la serie de observaciones tendríamos distintos valores correspondientes a λ. Entonces, a su vez, λ es una variable aleatoria y su distribución se conoce como **distribución muestral**.

Ahora, la pregunta es ¿cómo evaluar la calidad del ajuste? Es decir, observados las $N = 1207$ intervalos, ¿su distribución se parece efectivamente a la de una distribución Poisson? Si la respuesta la creyésemos afirmativa, entonces para un valor de emisiones elegido k, en N ensayos se esperarían $e_k = N\pi_k$. Intuitivamente pareciera que una medida de la **bondad del ajuste** pudiera ser

$$S = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \tag{1}$$

donde el índice i se refiere al número de *"casilla"* donde se comparan los **valores esperados** e_i con los **valores observados** o_i.

Note que si la distribución del número de emisiones X fuese efectivamente Poisson con $\lambda = 8.392$, entonces, por ejemplo $p_4 = \mathbb{P}(X = 4) = 0.047$ y en $N = 1207$ realizaciones esperaríamos $e_4 = Np_4 = 1257(.047) = 56.6$ emisiones. Los conteos esperados y correspondientes sumandos de la expresión (1) se muestran en la tabla 1.

Se puede mostrar, no en este curso, que bajo ciertas condiciones, el **estadístico** S (función de la muestra que no depende de parámetros desconocidos) de la expresión (1) sigue *asintóticamente* una distribución χ^2_{n-1}. Si λ no se conoce y hubiera que estimarlo, entonces $S \sim \chi^2_{n-k-1}$, donde n es el número de casillas donde se compara lo esperado con lo estimado ($n = 16$) y k es el número de parámetros estimados de la muestra ($k = 1$). En nuestro ejemplo, $S = 9.047$, que, si suponemos que sigue una distribución χ^2_ν, con $\nu = 16 - 1 - 1 = 14$, $\mathbb{P}(\chi^2_{14} \geq 9.047) = 0.828$. Esta última probabilidad se interpreta como la probabilidad de ver el estadístico observado o algo más extremo. En nuestro ejemplo, con el **valor-p** de 0.82, concluimos que no es *extraño* lo observado y aceptaríamos el modelo de la distribución Poisson con $\lambda = 8.392$. Por otro lado, si por ejemplo, consideramos que $\lambda = 8.75$, el correspondiente estadístico $S = 30.186$ con un valor-p de 0.007, que se interpretaría como que la probabilidad de observar $S = 30.186$ o algo mayor es un evento de probabilidad de apenas 0.7 %, lo que haría sospechar sobre la validez del modelo propuesto, a saber, $X \sim \text{Po}(8.75)$. La figura 9 muestra las frecuencias observadas y esperadas para los dos casos mencionados anteriormente. La figura 10 presenta la función de densidad de la distribución χ^2_{14} y localiza los estadísticos observados y correspondientes valores-p.

[6] Resultado $\hat{\lambda} = 8.392$, tomado del libro de texto. Nuestros cálculos arrojan $\tilde{\lambda} = 8.370$.

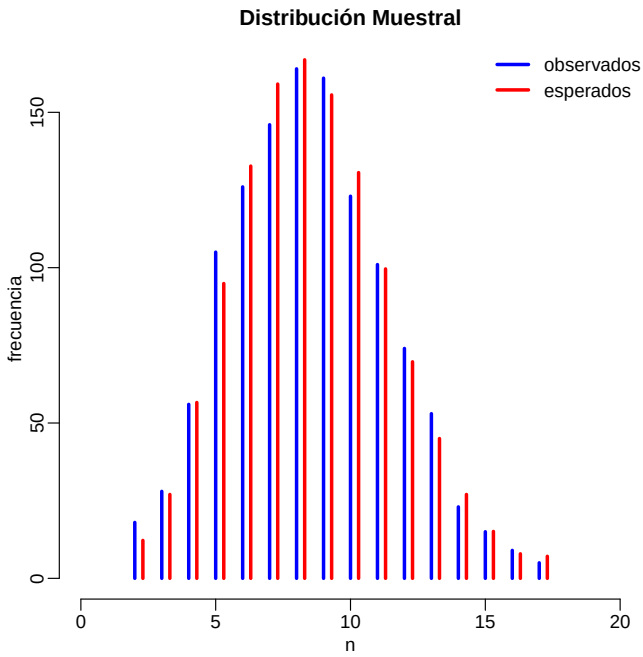


Figura 9: Observaciones.

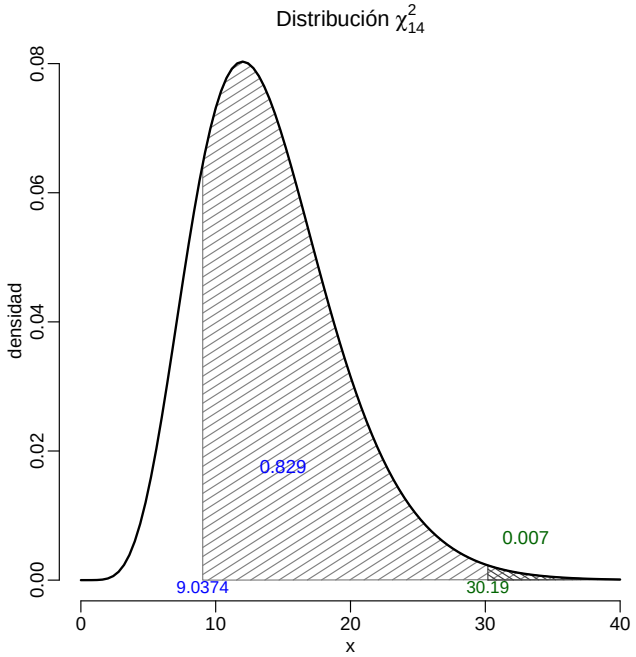


Figura 10: Densidad χ^2_{14} indicando los valores de los estadísticos observados y sus correspondientes valores-p, $S = 9.034(0.829)$ y $S = 8.75(0.007)$ correspondientes a los parámetros $\hat{\lambda} = 8.392$ y $\tilde{\lambda} = 8.750$, respectivamente.

5.2. Principios de estimación puntual

De manera muy somera se puede suponer que

- **Probabilidad**. Supone que el modelo es conocido y uno se pregunta por las probabilidades de ciertos eventos.

- **Estadística**. A partir de la salida de un experimento (observación, muestreo) se desea conocer los valores de los parámetros o el modelo (de distribución) mismo.

Modelos estadísticos.

$X \sim f$, con *f. d. p.*, *f. m. p.*, *f. p. a.* o función de distribución, representan el modelo (distribución) estadístico. Aunque se supone que X es observable, no siempre lo es pero aún así se puede hablar del modelo.

En muchos casos el modelo (distribución) depende de parámetros θ (univariado) ó $\boldsymbol{\theta}$ (vector de parámetros). Se supone por ejemplo que $\theta \in \Theta \subseteq \mathbb{R}$ ($\boldsymbol{\theta} \in \Theta \subseteq\in \mathbb{R}^k$), con Θ el *espacio de parámetros*.

Si $X \sim f(x;\theta)$, entonces

- $\mathbb{P}_\theta(A) = \int_A f(x;\theta)dx$;
- $E_\theta[X] = \int_{\mathbb{R}} xf(x;\theta)dx$;
- $\mathrm{var}_\theta(X) = \mathbb{E}_\theta[X^2] - \mathbb{E}^2_\theta[X]$.
- etcétera.

Los modelos pueden ser indexados por un vector de parámetros finito y en ese caso se dice que es un **modelo paramétrico**. Si no es posible se dice **modelo no paramétrico** con Θ de dimensión infinita.

Si para un modelo y parámetro θ dado corresponde una distribución F_θ, podría ser que $\theta_2 \neq \theta_1$ pero que $F_{\theta_1} = F_{\theta_2}$, por lo que podría tener un problema de estimación. Supondremos que si $F_{\theta_1} = F_{\theta_2} \Longrightarrow \theta_1 = \theta_2$. Esto es, el modelo es **identificable**. Parámetros de modelos identificables se dicen **estimables**.

Ejemplos.

- $\boldsymbol{X} = (X_1, \ldots, X_n)$ *m. a.* de $X \sim \mathrm{Po}(\lambda)$.

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda \in \Lambda = \mathbb{R}^+ \text{ espacio de parámetros}$$

- $\boldsymbol{X}$ *m. a.* de $X \sim F$, continua pero desconocida. En este caso, el espacio de parámetros sería el espacio de funciones continuas que no pueden ser indexadas por parámetros. Luego el modelo es no paramétrico.

- X una *m. a.* de $X \sim F$, con *f. d. p.* f tal que $f(x-\theta)$ desconocida pero que satisface $f(x) = f(-x)$. El problema es no paramétrico pero depende también de θ por lo que se dice **semi-paramétrico**.

- $\boldsymbol{X} = (X_1, \ldots, X_n)$ *v. a.*'s independientes normales con $\mathbb{E}[X_i] = \beta_0 + \beta_1 t_i + \beta_2 s_i$, $i = 1, \ldots, n$. s_i, t_i conocidas. $\mathrm{var}(X_i) = \sigma^2$. $\Theta = \{(\beta_0, \beta_1, \beta_2, \sigma^2) : \beta_i \in \mathbb{R}, \sigma^2 > 0\}$, el espacio de parámetros. El modelo (la parametrización) es identificable si y solo si los vectores $z_0 = \mathbf{1}$, $z_1 = \boldsymbol{t}$, $z_2 = \boldsymbol{s}$ son linealmente independientes.

5.3. Estimación de parámetros

Suponer que una población X puede ser modelada mediante una ley de probabilidades puede darse por distintas razones:

- **Razonamiento teórico**. La teoría dice que la variable observada X puede distribuirse de acuerdo a cierta ley. Por ejemplo, la distribución del error o variación modelada por la normal de acuerdo a Teorema Central del Límite.

- **Razonamiento empírico**. La teoría dice que la variable observada puede aproximarse razonablemente mediante cierta ley. Por ejemplo, la distribución Poisson en la emisión de partículas radioactivas.
- **Ajuste empírico** Las observaciones son aproximadas mediante una ley sin teoría que respalde tal aproximación. Por ejemplo, usar la distribución Gamma para aproximar la precipitación pluvial.

Recuerde, si $\boldsymbol{X} = (X_1, \ldots, X_n)$ es una muestra aleatoria de tamaño n de la población X, $S = S(\boldsymbol{X})$ es un **estadístico** si S es función de la muestra que no depende de parámetros desconocidos.

Se supone que la **población de interés** X es modelada mediante la distribución paramétrica $f(x, \theta)$, donde $\boldsymbol{\theta} = (\theta_1, \ldots \theta_k)$, es el vector de parámetros. Se ha visto que un **estimador** $\tilde{\boldsymbol{\theta}}$ es en general un estadístico que aproxima el valor del parámetro $\boldsymbol{\theta}$, usualmente desconocido.

Definición : Suponga que X sea modelada por la *f. d. p.* $f(x; \boldsymbol{\theta})$, donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, es un vector de parámetros. $\boldsymbol{\theta} \in \Theta = \{\theta : \theta \text{ es un valor factible}\}$, es el **espacio de parámetros**.

Ejemplo :

- $X \sim \mathrm{N}(\mu, \sigma^2)$. $\boldsymbol{\theta} = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+ = \Theta$.
- $Y \sim \mathrm{Gamma}(\alpha, \beta)$. $\boldsymbol{\theta} = (\alpha, \beta) \in \mathbb{R}^+ \times \mathbb{R}^+ = \Theta$.

Definición : Un **estimador** $\tilde{\boldsymbol{\theta}}$ es una regla, generalmente función de la muestra $\boldsymbol{X}$, que aproxima el valor real, pero desconocido, del vector de parámetros $\boldsymbol{\theta}$. Decimos que $\tilde{\boldsymbol{\theta}}(\boldsymbol{X})$ es un estimador del parámetro $\boldsymbol{\theta}$ y $\tilde{\boldsymbol{\theta}}(\boldsymbol{X})$ tiene una distribución nombrada como la **distribución muestral**. Dada la muestra observada $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\tilde{\boldsymbol{\theta}}(\boldsymbol{x})$ se dice que es una **estimación** de $\boldsymbol{\theta}$.

Definición : Sea $X \sim f(x; \theta)$ y suponga el parámetro θ univariado y $\tilde{\theta}(\boldsymbol{X})$ estimador de θ. Se define el **sesgo** del estimador $\tilde{\theta}$ por

$$\mathrm{sesgo}(\tilde{\theta}; \theta) = \mathrm{Bias}(\tilde{\theta}; \theta) = B(\tilde{\theta}; \theta) = \mathbb{E}[\tilde{\theta} - \theta]$$

Se dice que $\tilde{\theta}$ es un **estimador insesgado** si $B(\tilde{\theta}) = 0$, o bien, $\mathbb{E}\left[\tilde{\theta}(\boldsymbol{X})\right] = \theta$.

Definición : La desviación estándar de un estimador se conoce como **error estándar del estimador**.

$$\mathrm{ee}(\tilde{\theta}) = \sqrt{\mathrm{var}(\tilde{\theta})}$$

El error estándar de un estimador es una medida de lo preciso (variable) del estimador. La figura 11 muestra el caso de un estimador *preciso* pero sesgado (panel izquierdo); un estimador insesgado pero impreciso (panel central) y el caso de un estimador ideal: insesgado y preciso.

Definición : Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x; \theta)$. Sea $\tilde{\theta} = \tilde{\theta}(\boldsymbol{X})$ un estimador del parámetro θ. Se define el **error cuadrático medio** de $\tilde{\theta}$ por

$$\mathrm{ECM}(\tilde{\theta}) = \mathbb{E}\left[(\tilde{\theta} - \theta)^2\right]$$

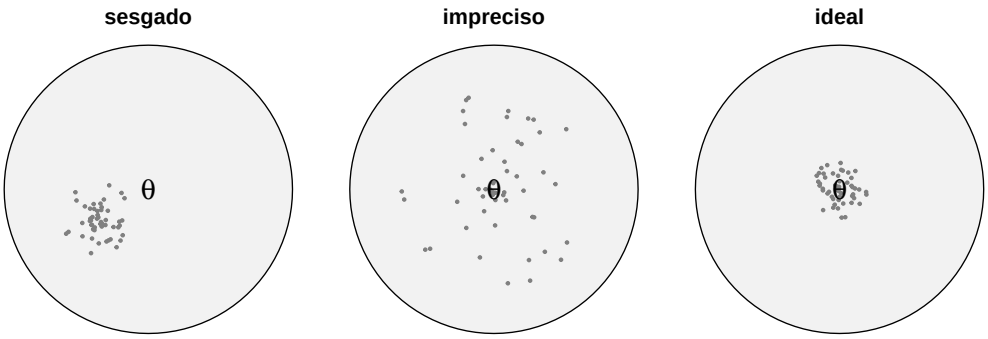


Figura 11: Estimadores. El panel de la derecha muestra estimaciones de un estimador del parámetro θ ideal: insesgado y preciso.

Proposición : Sea $\boldsymbol{X}$ una muestra aleatoria de $X \sim f(x;\theta)$. Para todo estimador $\tilde{\theta}$ de θ, se tiene que

$$\mathrm{ECM}(\tilde{\theta}) = \mathrm{var}(\tilde{\theta}) + \mathrm{sesgo}^2(\tilde{\theta})$$

Demostración:

$$\begin{aligned}
\mathrm{ECM}(\tilde{\theta}) &= \mathbb{E}[(\tilde{\theta}-\theta)^2] \\
&= \mathbb{E}\left[\left\{(\tilde{\theta}-\mathbb{E}\tilde{\theta}) + (\mathbb{E}\tilde{\theta}-\theta)\right)\right\}^2\Big] \\
&= \mathbb{E}[(\tilde{\theta}-\mathbb{E}\tilde{\theta})^2] + \mathbb{E}[(\mathbb{E}\tilde{\theta}-\theta)^2] + 2(\mathbb{E}\tilde{\theta}-\theta)\mathbb{E}[\tilde{\theta}-\mathbb{E}\tilde{\theta}] \\
&= \mathrm{var}(\tilde{\theta}) + \mathrm{sesgo}^2(\tilde{\theta})
\end{aligned}$$

pues el $\mathbb{E}[\tilde{\theta}-\mathbb{E}\tilde{\theta}] = 0$ y $\mathbb{E}[(\mathbb{E}\tilde{\theta}-\theta)^2] = \mathrm{sesgo}^2(\tilde{\theta})$.

Definición : Sea $\tilde{\theta}_1, \tilde{\theta}_2, \ldots$ una sucesión de estimadores de θ. Se dice que $\{\tilde{\theta}_n\}$ es un **estimador consistente** de θ si

$$\tilde{\theta}_n \xrightarrow{P} \theta$$

Esto es, si para todo $\epsilon > 0$, se tiene que $\mathbb{P}(|\tilde{\theta}_n - \theta| > \epsilon) \to 0$. En general, $\tilde{\theta}_n = \tilde{\theta}(\boldsymbol{X}_n)$, donde $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ es una *m. a.* de tamaño n.

Definición : $\{\tilde{\theta}_n\}$ se dice **consistente en error cuadrático medio** si

$$\mathrm{ECM}(\tilde{\theta}_n) = \mathbb{E}[(\tilde{\theta}-\theta)^2] \longrightarrow 0$$

Esto es, si para todo $\epsilon > 0$ $\mathrm{ECM}(\tilde{\theta}_n) < \epsilon$, para $n > N(\epsilon)$.

Proposición : Si $\{\tilde{\theta}_n\}$ es consistente en error cuadrático medio, entonces $\mathrm{var}(\tilde{\theta}_n) \to 0$ y $\mathrm{sesgo}(\tilde{\theta}) \to 0$.

Demostración: $\mathrm{ECM}(\tilde{\theta}_n) = \mathrm{var}(\tilde{\theta}_n) + \mathrm{sesgo}^2(\tilde{\theta}_n) \longrightarrow 0$, entonces cada uno de los sumandos se va a cero por ser ambas cantidades positivas.

Proposición : Si $\{\tilde{\theta}_n\}$ es consistente en error cuadrático medio, entonces es consistente. Lo contrario no es necesariamente cierto.

Demostración: Se sigue de la desigualdad de Chebyshev,

$$\mathbb{P}\left(|\tilde{\theta}_n - \theta| \geq \epsilon\right) \leq \frac{\mathrm{var}(\tilde{\theta}_n - \theta)}{\epsilon^2} \longrightarrow 0$$

Comparación de estimadores

Una manera de comparar estimadores es mediante el error cuadrático medio. Se prefiere $\tilde{\theta}_1$ sobre $\tilde{\theta}_2$ si $\text{ECM}(\tilde{\theta}_1) < \text{ECM}(\tilde{\theta}_2)$. Si ambos estimadores son insesgados, tal comparación se reduce a elegir aquel estimador con menor varianza.

Definición : Sean $\tilde{\theta}_1$ y $\tilde{\theta}_2$ dos estimadores insesgados de θ, Se define la **eficiencia** de $\tilde{\theta}_2$ relativa a $\tilde{\theta}_1$ por

$$\text{eff}(\tilde{\theta}_2, \tilde{\theta}_1) = \frac{\text{var}(\tilde{\theta}_1)}{\text{var}(\tilde{\theta}_2)}$$

Se dice que $\tilde{\theta}_2$ es más eficiente que $\tilde{\theta}_1$, si $\text{eff}(\tilde{\theta}_2, \tilde{\theta}_1) > 1$.

En ocasiones las varianzas son de la forma

$$\text{var}(\tilde{\theta}_1(\boldsymbol{X}_n)) = \frac{c_1}{n}; \qquad \text{var}(\tilde{\theta}_2(\boldsymbol{X}_n)) = \frac{c_2}{n}$$

En este caso, la eficiencia se define como el cociente entre los tamaños de muestra para hacer que ambas varianzas sean iguales.

Definición : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$, un estimador $T_n^* = T^*(\boldsymbol{X}_n)$ del parámetro $\tau(\theta)$ se dice que es un **estimador insesgado de varianza mínima** (UMVUE, *uniformly minimum-variance unbiased estimator*), si y solo si,

i) $\mathbb{E}_\theta[T_n^*] = \tau(\theta)$.

ii) $\text{var}_\theta(T_n^*) \leq \text{var}_\theta(T_n)$, para todo T_n, estimador insesgado de $\tau(\theta)$.

Cota inferior de Cramér-Rao.

Teorema (CICR): Sea θ un parámetro, τ una función real diferenciable y $\tau(\theta)$ un parámetro definido en términos de θ por $\tau = \tau(\theta)$. Sea $T_n = T(\boldsymbol{X}_n)$ un estimador insesgado de τ. Entonces, bajo *condiciones de regularidad*[7]

$$\text{var}_\theta(T_n) \geq \frac{\left[\tau'(\theta)\right]^2}{n\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]} \tag{2}$$

La igualdad se cumple si y solo si existe una función, digamos $K(\theta, n)$ tal que

$$\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}_n;\theta) = K(\theta;n)\left[T(\boldsymbol{X}_n) - \tau(\theta)\right] \tag{3}$$

La expresión (2) se conoce como la **desigualdad de Cramér-Rao** y el lado derecho de la desigualdad como la **cota inferior de Cramér-Rao** (CICR).

Ejemplo : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, una *m. a.* de $X \sim \text{Exp}(\lambda)$. Luego, $\mathbb{E}[X] = 1/\lambda = \tau$ y considere $T_n = T(\boldsymbol{X}_n) = \bar{X}_n$ estimador insesgado de τ. Por mostrar que T_n alcanza la cota inferior de Cramér-Rao.

- $f(x;\lambda) = \lambda e^{-\lambda x}$
- Sea $\ell(\lambda; x) = \log f(x;\lambda) = \log\lambda - \lambda x$.

[7] Condiciones apropiadas de suavidad, de acuerdo a Rice (2007).

- $\frac{\partial}{\partial\lambda}\ell(\lambda;x) = \frac{1}{\lambda} - x$.
- $\left(\frac{\partial}{\partial\lambda}\ell(\lambda;x)\right)^2 = \frac{1}{\lambda^2} - 2\frac{x}{\lambda} + x^2$.
- $\mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\ell(\lambda;X)\right)^2\right] = \frac{1}{\lambda^2} - 2\frac{1/\lambda}{\lambda} + (\frac{1}{\lambda^2} + \frac{1}{\lambda^2}) = \frac{1}{\lambda^2}$

Por otro lado, sea $\tau(\lambda) = 1/\lambda$. Entonces, $\tau'(\lambda) = -1/\lambda^2$ y $(\tau'(\lambda))^2 = 1/\lambda^4$, por lo que la desigualdad de Cramér-Rao queda

$$\text{var}(T_n) \geq \frac{1/\lambda^4}{n/\lambda^2} = \frac{1}{n\lambda^2}$$

Se tiene también que $\text{var}(T_n) = \text{var}(\bar{X}_n) = \text{var}(X)/n = \frac{1/\lambda^2}{n}$. Esto es, se alcanza la CICR.

Finalmente, note que

$$\frac{\partial}{\partial\lambda}\log f(\boldsymbol{X}_n;\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\log f(X_i;\theta) = n/\lambda - \sum_{i=1}^{n}X_i = -n\left(\bar{X}_n - \frac{1}{\lambda}\right)$$

con $K(\lambda,n) = -n$ y $(T(\boldsymbol{X}_n) - \tau(\lambda)) = (\bar{X}_n - 1/\lambda)$ de la expresión (3).

5.4. El método de momentos

Sea $X \sim f(x;\theta)$. Recuerde que el r-ésimo momento de X (o de la distribución de X) es $\mu_r = \mathbb{E}[X^r]$, siempre que el valor esperado exista.

Suponga $\boldsymbol{X}_n = (X_1,\ldots,X_n)$ una muestra aleatoria de la población X. Se define el **r-ésimo momento muestral** por $m_r = \dfrac{1}{n}\sum_{i=1}^{n}X_i^r$

En este caso, parecería intuitivo ver a m_r como un estimador de μ_r. Si, por ejemplo, $k = 2$ y se tiene el vector de parámetros $\boldsymbol{\theta} = (\theta_1,\theta_2)$ y $\theta_i = h_i(\mu_1,\mu_2)$, $i = 1,2$. Entonces, por el **método de momentos** se estimaría

$$\tilde{\theta}_1 = h_1(m_1,m_2) \quad \text{y} \quad \tilde{\theta}_2 = h_2(m_1,m_2)$$

Resumiendo: la estimación de parámetros por el método de momentos involucra el **principio de sustitución** de acuerdo a Knight (2000):

- Encuentre las expresiones que relacionen los momentos teóricos con los parámetros a estimar. Exprese los parámetros en términos de los momentos.
- Calcule los momentos muestrales de bajo orden que sean necesarios en el punto anterior.
- Sustituya los momentos muestrales en las expresiones del primer punto. El resultado son los **estimadores de momentos** (EMM).

Ejemplo : Sea $\boldsymbol{X}_n = (X_1,\ldots,X_n)$ una *m. a.* de $X \sim \text{Po}(\lambda)$.

$\mu_1 = \mathbb{E}[X] = \lambda$; $m_1 = \frac{1}{n}\sum_{i=1}^{n}X_i = \bar{X}_n$. Entonces, $\tilde{\lambda}_{\text{EMM}} = \bar{X}$.

Note que el estimador $\tilde{\lambda}$ es insesgado pues $\mathbb{E}[\tilde{\lambda}] = \mathbb{E}[\bar{X}] = \lambda$. Además, $\text{var}(\lambda) = \text{var}(\bar{X}) = \lambda/n$. Finalmente, su error estándar, $\text{ee}(\tilde{\lambda}) = \sqrt{\lambda/n}$ estimado a su vez por $\sqrt{\bar{X}/n}$.

Ejemplo : Sea $\boldsymbol{X}_n = (X_1,\ldots,X_n)$ una *m. a.* de $X \sim \text{N}(\mu,\sigma^2)$.

Entonces, $\mu_1 = \mu$ y $\mu_2 = \mu^2 + \sigma^2$. $m_1 = \frac{1}{n}\sum X_i$ y $m_2 = \frac{1}{n}\sum X_i^2$. Entonces,

- $\tilde{\mu} = m_1 = \bar{X}$.
- $\tilde{\sigma}^2 = m_2 - m_1^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$.

Note que $\tilde{\mu} = \bar{X}$ es un estimador insesgado de μ con un error estándar $\text{ee}(\bar{X}) = \sigma/\sqrt{n}$.

Por otro lado, si $\tilde{\sigma}^2 = \frac{n-1}{n}S^2$, con S^2 la varianza muestral, entonces $\mathbb{E}[\tilde{\sigma}^2] = \frac{n-1}{n}\mathbb{E}[S^2] = \frac{n-1}{n}\sigma^2$, por lo que $\tilde{\sigma}^2$ es un estimador sesgado de la varianza σ^2. Su error estándar

$$\text{ee}(\tilde{\sigma}^2) = \sqrt{\text{var}\left(\frac{n-1}{n}S^2\right)} = \frac{\sigma^2}{n}\sqrt{2(n-1)}$$

pues $\dfrac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$, por lo que $\text{var}(S^2) = \frac{\sigma^4}{(n-1)^2}2(n-1)$.

Ejemplo : Sea $\boldsymbol{U}_n = (U_1, \ldots, U_n)$, una *m. a.* de $U \sim \text{Unif}(0, \theta)$, con $\theta > 0$.

Entonces, $\mu_1 = \mathbb{E}[U] = \frac{\theta}{2}$. $m_1 = \frac{1}{n}\sum U_i = \bar{U}$. Y por lo tanto, $\tilde{\theta}_{\text{EMM}} = 2\bar{U}$, lo que puede dar lugar a estimaciones absurdas si, por ejemplo, $U_{(n)} > 2\bar{U} = \tilde{\theta}$.

Ejercicio[8]**:** El ángulo θ en que un electrón es emitido se puede modelar mediante un *f. d. p.* dada por

$$f(x; \alpha) = \frac{1 + \alpha x}{2}\mathbb{1}_{(-1,1)}(x)$$

para algún $|\alpha| \leq 1$ y donde $x = \cos\theta$. Consideraciones físicas sugieren que $|\alpha| < 1/3$.

$\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una *m. a.* de X. Muestre que $\tilde{\alpha}_{\text{EMM}} = 3\bar{X}$.

Proposición : Los estimadores por el método de momentos son estimadores consistentes. Sea $\tilde{\theta}_n = \tilde{\theta}(\boldsymbol{X}_n)$ el estimador por el método de momentos del parámetro θ. Entonces, bajo condiciones razonables

$$\tilde{\theta}_n \xrightarrow{P} \theta$$

Esto es, para todo $\epsilon > 0$, $\mathbb{P}(|\tilde{\theta}_n - \theta| > \epsilon) \to 0$.

Demostración: Se basa en la Ley de los Grandes Números y el teorema del mapeo continuo.

Ejemplo : Considere la *m. a.* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de $Y \sim \text{Ga}(\alpha, \beta)$, con α y β los parámetros de forma y escala respectivamente. Entonces,

- $m_1 \equiv \mu_1 = \mathbb{E}[Y] = \alpha\beta$
- $m_2 \equiv \mu_2 = \mathbb{E}[Y^2] = \alpha\beta^2 + \alpha^2\beta^2 = \alpha\beta^2(1 + \alpha)$

Resolviendo el sistema para α y β,

$$\tilde{\alpha}_{\text{EMM}} = \frac{m_1^2}{m_2 - m_1^2} = \frac{\bar{Y}^2}{\tilde{\sigma}^2} \quad \text{y} \quad \tilde{\beta}_{\text{EMM}} = \frac{m_2 - m_1^2}{m_1} = \frac{\tilde{\sigma}^2}{\bar{Y}}$$

El problema ahora sería determinar la distribución (muestral) de los estimadores $\tilde{\alpha}$ y $\tilde{\beta}$.

Una manera de evaluar (aproximar) a su vez el sesgo y el error estándar de los estimadores es mediante la **simulación Monte Carlo**.

[8] Rice (2007).

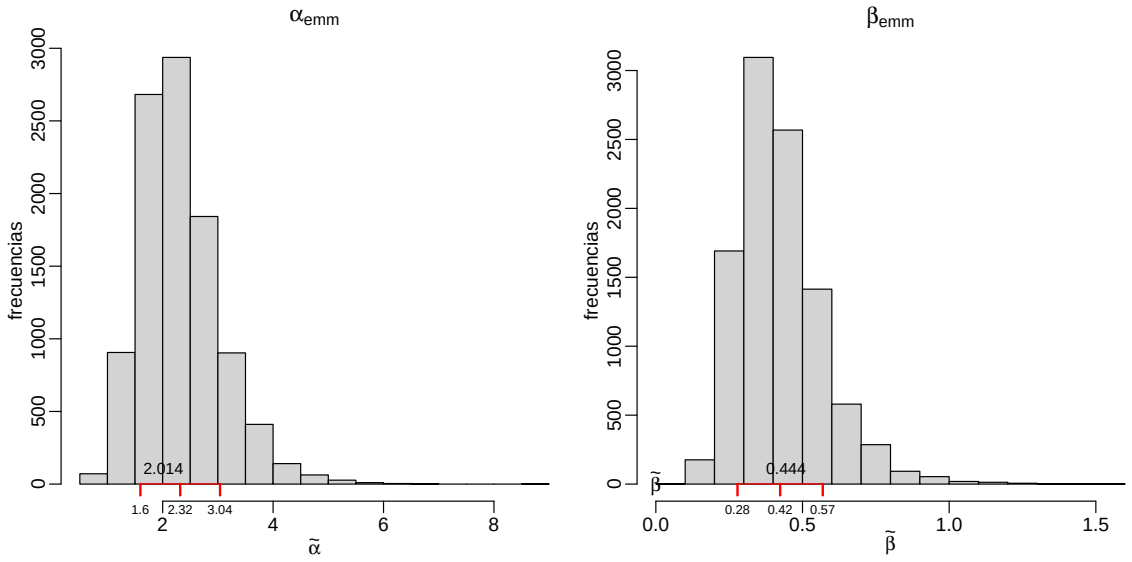


Figura 12: Histogramas de estimaciones de los parámetros de forma α y escala β, de una distribución gamma por el método de momentos. Los intervalos están centrados en el promedio de estimaciones y son de amplitud 2 desviaciones estándar.

A saber, en este ejemplo se simula una gran cantidad (N) de muestras aleatorias de una distribución Gamma del mismo tamaño n de la muestra original $\boldsymbol{Y}$ y de parámetros de forma $\tilde{\alpha}_{\text{EMM}}$ y de escala $\tilde{\beta}_{\text{EMM}}$.

Sean pues $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$, las N muestras simuladas. Por ejemplo, la j-ésima muestra sería $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn})$ Y con base en esta muestra se obtendrían las correspondientes estimaciones $\tilde{\alpha}_j$ y $\tilde{\beta}_j$. Finalmente, si definimos $\bar{\mu}$ y $\bar{\sigma}^2$ como la media y varianza de los estimadores. Así,

$$\bar{\mu}_{\tilde{\alpha}} = \frac{1}{N}\sum_{j=1}^{N}\tilde{\alpha}_j \qquad \bar{\mu}_{\tilde{\beta}} = \frac{1}{N}\sum_{j=1}^{N}\tilde{\beta}_j$$

$$\bar{\sigma}^2_{\tilde{\alpha}} = \frac{1}{N}\sum_{j=1}^{N}(\tilde{\alpha}_j - \bar{\mu}_{\tilde{\alpha}})^2 \qquad \bar{\sigma}^2_{\tilde{\beta}} = \frac{1}{N}\sum_{j=1}^{N}(\tilde{\beta}_j - \bar{\mu}_{\tilde{\beta}})^2$$

Esta manera de estimación por simulación (Monte Carlo) de muestras se conoce como **bootstrap**. Uno de varios *métodos de remuestreo*. La figura 12 muestra los histogramas de las estimaciones $\tilde{\alpha}$ y $\tilde{\beta}$ para una simulación de $N = 10,000$ muestras de tamaño $n = 30$ de una distribución $\text{Ga}(\alpha = 2.01, \beta = 0.44)$.

Nota: La consistencia de los estimadores $\tilde{\theta}$ se justifica por

$$S_{\tilde{\theta}} = \text{ee}(\tilde{\theta}) \xrightarrow{P} \sigma(\tilde{\theta}) = \text{stdev}(\tilde{\theta}) = h(\tilde{\theta})$$

Esto es, $\lim_{n\to\infty} \dfrac{S_{\tilde{\theta}}}{\text{ee}(\tilde{\theta})} = 1$ cuando $\sigma(\theta)$ es una función continua h de θ, pues $\tilde{\theta} \xrightarrow{P} \theta$, por LGN y $\sigma(\tilde{\theta}) \xrightarrow{P} \sigma(\theta)$ por el teorema del mapeo continuo.

5.5. El método de máxima verosimilitud

Ejemplo[9]**:** Suponga una urna con bolas blancas y negras en una proporción entre ellas de 3 a 1, pero no se sabe cuál es más numerosa. Para determinarlo se llevan a cabo n

[9] Mood, Graybill, and Boes (1974).

extracciones independientes con reemplazo. Si X denota el número de bolas negras en las n extracciones, $X \sim \text{Bin}(n,p)$ con $p = 1/4$ ó $p = 3/4$. Para $n = 3$ se construye la siguiente tabla de probabilidades:

p	0	1	2	3	
3/4	1/64	9/64	**27/64**	**27/64**	1
1/4	**27/64**	**27/64**	9/64	1/64	1

Por ejemplo, $\mathbb{P}(X = 1; p = 3/4) = 9/64$, mientras que $\mathbb{P}(X = 1; p = 1/4) = 27/64$.

Si en las tres extracciones salen 0 ó 1 bola negra uno se inclinaría por $\hat{p} = 1/4$ por tener asociadas las probabilidades más altas. Igualmente uno se inclinaría por $\hat{p} = 3/4$ si el número de negras fuera 2 ó 3.

Definición : Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ un vector aleatorio (muestra) con función de densidad conjunta $f(\boldsymbol{x}; \theta)$, con $\theta \in \Theta$. Se define la **función de verosimilitud** L, como la función de θ dado $\boldsymbol{x}$ igual a la función de densidad conjunta $f(\boldsymbol{x})$, dado θ . Esto es,

$$L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}; \theta), \qquad \theta \in \Theta$$

Nota: Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de una población $X \sim f(x; \theta)$, con $\theta \in \Theta$. Dada la realización $\boldsymbol{x} = (x_1, \ldots, x_n)$, por independencia queda la **función de verosimilitud** L como

$$L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Para cada realización $\boldsymbol{X} = \boldsymbol{x}$, $L(\theta; \boldsymbol{x})$ es una función real Con dominio en el espacio de parámetros Θ.

Definición : Suponga $\boldsymbol{x} = (x_1, \ldots, x_n)$, una muestra observada de $X \sim f(\boldsymbol{x}; \theta)$ se define $\hat{\theta}_{\text{EMV}}$, el **estimador de máxima verosimilitud** de θ por

$$\hat{\theta}_{\text{EMV}} = \arg \left\{ \sup_{\theta \in \Theta} L(\theta; \boldsymbol{x}) \right\}$$

Un **estimador de máxima verosimilitud** (EMV) de θ, es un estadístico $S(\boldsymbol{X}) = \hat{\theta}_{\text{EMV}}$, si sobre $\theta \in \Theta$, $\sup L(\theta, \boldsymbol{X}) = L\big(S(\boldsymbol{X}), \boldsymbol{X}\big)$.

Definición : Sea $\boldsymbol{x} = (x_1, \ldots, x_n)$ una muestra observada de $X \sim f(x; \theta)$. Se define la función **log de verosimilitud** por

$$\ell(\theta; \boldsymbol{x}) = \log L(\theta; \boldsymbol{x})$$

Si la muestra $\boldsymbol{x}$ es aleatoria,

$$\ell(\theta; \boldsymbol{x}) = \log L(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

Nota: Si $\hat{\theta}(\boldsymbol{x})$ maximiza la función de verosimilitud L, por monotonía creciente de la función logaritmo $\hat{\theta}$ también maximiza la función log de verosimilitud ℓ.

Ejemplo : Considere $\boldsymbol{X} = (X_1, \ldots, X_n)$ una *m. a.* de $X \sim \text{Po}(\lambda)$. Entonces $f(x;\lambda) = \lambda^x \dfrac{e^{-\lambda}}{x!}$, $x = 0, 1, 2, \ldots$ y

$$\begin{aligned} L(\lambda, \boldsymbol{x}) &= f(\boldsymbol{x};\lambda) \\ &= \prod f(x_i;\lambda) \\ &= \lambda^{\sum x_i} e^{-n\lambda} / \prod x_i! \\ \ell(\lambda;\boldsymbol{x}) &= \log\lambda \sum x_i - n\lambda - \sum \log x_i! \\ \frac{\partial \ell}{\partial \lambda} &= \frac{1}{\lambda}\sum x_i - n \end{aligned}$$

En este caso, se puede maximizar el log de la verosimilitud ℓ derivando (con respecto a λ) e igualando a cero para resolver con $\hat{\lambda}_{\text{EMV}} = \bar{X}$, que en este ejemplo coincide con el estimador del parámetro λ por el método de momentos $\tilde{\lambda}_{\text{EMM}}$. Luego, se tiene que

$$L(\bar{x};\boldsymbol{x}) \geq L(\lambda;\boldsymbol{x}), \quad \text{para todo } \lambda > 0$$

Notas:

1. No siempre se puede obtener el estimador de máxima verosimilitud por derivación, por ejemplo, cuando la función no es diferenciable o debido al espacio de parámetros.
2. Los estimadores de máxima verosimilitud no son necesariamente insesgados.
3. Los estimadores de máxima verosimilitud no son únicos necesariamente.

Ejemplo : Sea $\boldsymbol{T} = (T_1, \ldots, T_n)$ una *m. a.* de $T \sim \text{Exp}(\theta)$, donde $\theta = \mathbb{E}[T]$ para algún $\theta > 0$. Luego $f(t;\theta) = \frac{1}{\theta} e^{-t/\theta} \mathbb{1}_{\mathbb{R}^+}(t)$. Entonces,

$$\begin{aligned} L(\theta, \boldsymbol{t}) &= f(\boldsymbol{t};\theta) \\ &= \prod f(t_i;\theta) \\ &= \theta^{-n} e^{-\frac{1}{\theta}\sum t_i} \\ \ell(\theta;\boldsymbol{t}) &= -n\log\theta - \frac{1}{\theta}\sum t_i \\ \frac{\partial \ell}{\partial \theta} &= \frac{-n}{\theta} + \frac{1}{\theta^2}\sum t_i \end{aligned}$$

Nuevamente, igualando la derivada a cero y resolviendo para θ, se tiene que

$$\hat{\theta}_{\text{EMV}} = \frac{1}{n}\sum_{i=1}^{n} t_i = \bar{t}$$

Nota: En este caso, si $T \sim \text{Exp}(\lambda)$, con $\mathbb{E}[T] = \theta = 1/\lambda$. Entonces,

$$\hat{\lambda}_{\text{EMV}} = \frac{1}{\hat{\theta}_{\text{EMV}}}$$

por el **principio de invarianza** de los estimadores de máxima verosimilitud y que se verá más adelante.

Ejemplo : Sea $\boldsymbol{U} = (U_1, \ldots, U_n)$ una *m. a.* de $U \sim \text{Unif}(0, \theta)$, para algún $\theta > 0$. Luego, $f(u;\theta) = \frac{1}{\theta}\mathbb{1}_{(0,\theta)}(u)$ y se sigue que

$$L(\theta, \boldsymbol{u}) = \prod f(u_i;\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbb{1}_{(0,\theta)}(u_i)$$

Note que

$$L(\theta, \boldsymbol{u}) = \begin{cases} \theta^{-n} & \text{si, } 0 < u_i < \theta, \text{ para todo } i = 1, \ldots, n \\ 0 & \text{en caso contrario} \end{cases}$$

Luego, la función de verosimilitud L será mayor mientras más chico sea θ pero siempre y cuando sea mayor que todos los u_i's. Entonces, $\hat{\theta}_{\text{EMV}} = U_{(n)}$ pues

$$\text{máx}\{U_1, \ldots, U_n\} = \arg\left\{ \sup_{\theta > 0} L(\theta; \boldsymbol{u}) \right\}$$

En este caso, la maximización de la función de verosimilitud no fue por derivación y no ayudaría el uso del log de la verosimilitud. Por otro lado, note que la distribución muestral del estimador corresponde a la del n-ésimo estadístico de orden, el máximo de la muestra.

Ejemplo : Considere $\boldsymbol{X} = (X_1, \ldots, X_n)$ una *m. a.* de $X \sim \text{N}(\mu, \sigma^2)$. Sea $\boldsymbol{\theta} = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}^+$. Luego, la función de verosimilitud y log de la verosimilitud están dadas por

$$L(\theta, \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2}$$

$$\ell(\theta; \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

funciones suaves en $\boldsymbol{\theta} \in \Theta$ conjunto abierto por lo que derivando e igualando a cero permite la localización del máximo.

$$\frac{\partial \ell}{\partial \mu} = -\frac{2}{2\sigma^2}\sum(x_i - \mu)(-1) = 0 \Longrightarrow \hat{\mu} = \bar{x}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum(x_i - \hat{\mu})^2 = 0 \Longrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \hat{\mu})^2$$

Note que $\hat{\mu}_{\text{EMV}} = \bar{X}$ es un estimador insesgado de μ, pero $\hat{\sigma}^2_{\text{EMV}} = \frac{1}{n}\sum(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$ es sesgado pues $\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$. Además, la distribución muestral de $\hat{\mu}$ es $\text{N}(\mu, \sigma^2/n)$ mientras que $\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2_{n-1}$.

Ejemplo : Considere ahora $\boldsymbol{U} = (U_1, \ldots, U_n)$ una *m. a.* de $U \sim \text{Unif}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. La función de verosimilitud está dada por $L(\theta; \boldsymbol{U}) = \prod_{i=1}^{n} \mathbb{1}_{(\theta - \frac{1}{2}, \theta + \frac{1}{2})}(U_i)$. Luego,

$$L(\theta, \boldsymbol{u}) = \begin{cases} 1 & \text{si } \theta - \frac{1}{2} \leq u_i \leq \theta + \frac{1}{2}, \ i = 1, \ldots, n \\ 0 & \text{en caso contrario} \end{cases}$$

Entonces, la función es máxima si para todo $i = 1, \ldots, n$, $\theta - \frac{1}{2} \leq u_i \leq \theta + \frac{1}{2}$. Esto es, si y solo si $\theta \leq u_i + \frac{1}{2}$ y $u_i - \frac{1}{2} \leq \theta$, y esto ocurre si y solo si $u_{(n)} - \frac{1}{2} \leq \theta \leq u_{(1)} + \frac{1}{2}$. Luego, el estimador $\hat{\theta}_{\text{EMV}}$ será todo valor de θ en el intervalo $(U_{(n)} - \frac{1}{2}, U_{(1)} + \frac{1}{2})$. Se concluye que los estimadores de máxima verosimilitud no son necesariamente únicos.

Ejemplo : Considere $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ una muestra aleatoria de una población modelada mediante una distribución gamma con parámetro de forma α, β parámetro de escala y

$\lambda = 1/\beta$ parámetro tasa. Entonces, por ejemplo, el vector de parámetros es $\boldsymbol{\theta} = (\alpha, \beta)$ en el espacio de parametral $\Theta = \mathbb{R}^+ \times \mathbb{R}^+$.

Luego, $f(y; \boldsymbol{\theta}) = \dfrac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta} \mathbb{1}_{\mathbb{R}^+}(y)$ con $\mathbb{E}[Y] = \alpha\beta$ y $\text{var}(Y) = \alpha\beta^2$.

Sea $m_r = \frac{1}{n}\sum_{i=1}^n Y_i^r$, el r-ésimo momento muestral. Luego,

$$\begin{aligned} \mu &= \mu\beta = m_1 \\ \sigma^2 &= \alpha\beta^2 = m_2 - m_1^2 \end{aligned}$$

y resolviendo para α y β se tienen los estimadores por el método de momentos

$$\tilde{\alpha}_{\text{EMM}} = \frac{m_1^2}{m_2 - m_1^2} \quad \text{y} \quad \tilde{\beta}_{\text{EMM}} = \frac{m_2 - m_1^2}{m_1}$$

Ahora bien,

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{y}) &= \prod_{j=1}^n \frac{1}{\beta^\alpha \Gamma(\alpha)} y_j^{\alpha-1} e^{-y_j/\beta} \\ &= \beta^{-n\alpha} \Gamma^{-n}(\alpha) \left(\prod y_j\right)^{\alpha-1} e^{-\frac{1}{\beta}\sum y_j} \\ \ell(\alpha, \beta; \boldsymbol{y}) &= -n\alpha \log\beta - n\log\Gamma(\alpha) + (\alpha-1)\sum \log y_j - \frac{1}{\beta}\sum y_i \end{aligned}$$

Derivando la función log de la verosimilitud e igualando a cero se tiene el siguiente sistema

$$\frac{\partial \ell}{\partial \alpha} = -n\log\beta - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum \log y_j = 0 \tag{4}$$

$$\frac{\partial \ell}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum y_j = 0 \tag{5}$$

De la ecuación (5) se sigue que $\hat{\beta}_{\text{EMV}} = \dfrac{\bar{Y}}{\hat{\alpha}}$, mientras que de la ecuación (4) se tiene la siguiente ecuación

$$-n\log\bar{Y} + n\log\alpha - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum \log Y_j = 0$$

cuya solución arrojaría $\hat{\alpha}$, el estimador de máxima verosimilitud del parámetro de forma.

Note que la ecuación anterior no tiene una solución analítica, por lo que no hay una expresión cerrada para $\hat{\alpha}_{\text{EMV}}$, pero se puede resolver numéricamente y se dispondrían así de estimaciones de α y consecuentemente para β.

Ejercicio [10]**:**. Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim \text{Unif}(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma)$.

Encontrar los estimadores por el método de momentos y por el de máxima verosimilitud del vector de parámetros $\boldsymbol{\theta} = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}^+$.

Ejemplo : Suponga que $\boldsymbol{X} = (X_1, \ldots, X_k)$ es una muestra aleatoria de tamaño n de una población $X \sim \text{Multinomial}(n; p_1, \ldots, p_k)$. Entonces,

$$f(\boldsymbol{x}; \boldsymbol{p}) = \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k}$$

donde $n = \sum_{i=1}^k x_i$ y $1 = \sum_{i=1}^k p_i$. Entonces, la función log de la verosimilitud

$$\ell(\boldsymbol{p}; \boldsymbol{x}) = \log n! - \sum \log x_i! + \sum x_i \log p_i$$

[10] Mood, Graybill, and Boes (1974).

Note que maximizar $\ell(\boldsymbol{p};\boldsymbol{x})$ sobre el espacio de parámetros impone la restricción $p_i \geq 0$ y que $\sum p_i = 1$. Luego, se ha de resolver el problema

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \ell(\boldsymbol{p};\boldsymbol{x}) \\ \text{sujeto a} \quad & \sum p_i = 1 \end{aligned}$$

Para esto, se puede construir el *lagrangiano*

$$\mathcal{L}(\boldsymbol{p};\lambda) = \log n! - \sum \log x_i! + \sum x_i \log p_i + \lambda\left(\sum p_i - 1\right)$$

con λ como el *multiplicador de Lagrange* y que da lugar al sistema

$$\frac{\partial \mathcal{L}(\boldsymbol{p};\lambda)}{\partial p_i} = \frac{X_i}{p_i} + \lambda = 0, \qquad \frac{\partial \mathcal{L}(\boldsymbol{p};\lambda)}{\partial \lambda} = \sum p_i - 1 = 0$$

con solución

$$\hat{p}_i = \frac{X_i}{\sum X_j} = \frac{X_i}{n}, \quad i = 1, \ldots, k$$

Modelo de Equilibrio de Hardy-Weinberg[11]

De acuerdo con la *Ley de Hardy-Weinberg*, si las frecuencias de los genotipos están en equilibrio, los genotipos AA, Aa, y aa, ocurren con una frecuencia mostrada en la tabla

g	AA	Aa	aa
f	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2

En una muestra aleatoria (Hong Kong, población china, 1937) se encontró la siguiente distribución en el tipo de sangre:

g	M	MN	N	Total
f	342	500	187	1029

Interesa el parámetro θ. Para esto se tienen varias formas de estimarlo. Por ejemplo, $\tilde{\theta} = \sqrt{\frac{187}{1029}} = 0.4263$. Sin embargo, parecería que este procedimiento ignora la información de las otras celdas.

Alternativamente, si asumimos el modelo multinomial, $X \sim \text{Multinomial}(n, \boldsymbol{p})$.

$$\begin{aligned} \ell(\theta, \boldsymbol{x}) &= \log n! - \sum \log x_i! + \sum x_i \log p_i \\ &= \log n! - \sum \log x_i! + x_1 \log(1-\theta)^2 + x_2 \log 2\theta(1-\theta) + x_3 \log \theta^2 \\ &= K + (2x_1 + x_2)\log(1-\theta) + (x_2 + 2x_3)\log\theta + x_2 \log 2 \\ \frac{\partial \ell(\theta, \boldsymbol{x})}{\partial \theta} &= (2x_1 + x_2)\frac{-1}{1-\theta} + (x_2 + 2x_3)\frac{1}{\theta} \end{aligned}$$

Que igualando la derivada del log de la verosimilitud a cero y resolviendo para θ

$$\hat{\theta}_{\text{EMV}} = \frac{x_2 + 2x_3}{2x_1 + 2x_2 + 2x_3} = \frac{x_2 + 2x_3}{2n} = \frac{500 + 2(187)}{2(1029)} = 0.4247$$

En este caso, para determinar el error estándar del estimador se usaría *bootstrap*.

[11] Rice (2007).

Datos censurados[12]

Sea $\boldsymbol{T} = (T_1, \ldots, T_n)$ una muestra aleatoria de $T \sim \text{Exp}(\lambda)$. λ es el parámetro tasa de la distribución tal que $\mathbb{E}[T] = 1/\lambda$. Suponga por ejemplo que T denota el tiempo de procesamiento de un trabajo industrial con un tiempo medio de $1/\lambda$.

$$L(\lambda; \boldsymbol{t}) = \prod \lambda e^{-\lambda t_i}$$

$$\ell(\lambda; \boldsymbol{t}) = n \log \lambda - \lambda \sum t_i$$

$$\frac{\partial \ell(\lambda; \boldsymbol{t})}{\partial \lambda} = \frac{n}{\lambda} - \sum t_i$$

Nuevamente, igualando a cero al derivada del log de la verosimilitud y resolviendo se tiene

$$\hat{\lambda}_{\text{EMV}} = \frac{n}{\sum t_i} = 1/\bar{t}$$

Suponga ahora que los trabajos que se observaron hasta un tiempo τ fueron los primeros m y lo único que se sabe de los últimos $n - m$ es que duraron más de τ. En este caso, la función de verosimilitud queda

$$\begin{aligned} L(\lambda; \boldsymbol{t}) &= \prod_{i=1}^{m} f(t_i; \lambda) \prod_{j=m+1}^{n} \mathbb{P}(T_j > \tau) \\ &= \lambda^m e^{-\lambda \sum_{i=1}^{m} t_i} \left[1 - F(\tau)\right]^{n-m} \\ &= \lambda^m e^{-\lambda \sum_{i=1}^{m} t_i} \left[e^{-\lambda \tau}\right]^{n-m} \end{aligned}$$

$$\ell(\lambda; \boldsymbol{t}) = m \log \lambda - \lambda \sum_{i=1}^{m} t_i - \lambda \tau (n - m)$$

$$\frac{\partial \ell(\lambda : \boldsymbol{t})}{\partial \lambda} = \frac{m}{\lambda} - \left(\sum_{i=1}^{m} t_i + \tau(n - m) \right)$$

Igualando a cero y despejando para λ se tiene

$$\hat{\lambda}_{\text{EMV}} = \frac{m}{\sum_{i=1}^{m} T_i + \tau(n - m)}$$

[12] Garthwaite, Jolliffe, and Jones (2002).

5.6. Principio de Invarianza de EMV

Con el apoyo de Mood, Graybill, and Boes (1974) y Dudewicz and Mishra (1988).

Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ muestra aleatoria de $X \sim f(x;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}$ y sea $\tau = \tau(\theta) \in T \subseteq \mathbb{R}$ una transformación de θ.

Proposición : Sea $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ un estimador de máxima verosimilitud (EMV) del parámetro θ de $f(x;\theta)$ con $\theta \in \Theta \subseteq \mathbb{R}$. Si el parámetro $\tau = \tau(\theta)$ es una función invertible de θ, entonces el EMV de τ es $\hat{\tau} = \tau(\hat{\theta})$.

Demostración:

$$L(\hat{\tau}, \boldsymbol{x}) = \sup_{\{\tau \in T\}} L(\tau; \boldsymbol{x}) = \sup_{\{\theta : \tau(\theta) = \tau\}} L(\theta; \boldsymbol{x}) = L(\hat{\theta}; \boldsymbol{x})$$

La propiedad se puede extender a un vector de parámetros $\boldsymbol{\theta}$ y una función no necesariamente uno-a-uno. Por ejemplo, sea σ^2 la varianza de una distribución Bernoulli parámetro de éxito p. Entonces, $\sigma^2 = p(1-p)$ que no es función uno-a-uno de p, pero como se ha visto, $\hat{p} = \bar{X}$ y el EMV es $\hat{\sigma}^2 = \bar{x}(1 - \bar{x})$.

Considere ahora $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$ y sea $\boldsymbol{\tau}(\boldsymbol{\theta}) = (\tau_1(\boldsymbol{\theta}), \ldots, \tau_r(\boldsymbol{\theta})) \in T \subseteq \mathbb{R}^r$, $1 \leq r \leq k$. Sea $M(\boldsymbol{\tau}; \boldsymbol{x}) = \sup_{\{\boldsymbol{\theta} : \boldsymbol{\tau}(\theta) = \boldsymbol{\tau}\}} L(\boldsymbol{\theta}; \boldsymbol{x})$. El estimador de máxima verosimilitud $\hat{\boldsymbol{\tau}}$ maximiza la función de verosimilitud inducida por $\boldsymbol{\tau}$, $M(\boldsymbol{\tau}; \boldsymbol{x})$. Esto es, $\hat{\boldsymbol{\tau}}$ es tal que $M(\hat{\boldsymbol{\tau}}; \boldsymbol{x}) \geq M(\boldsymbol{\tau}; \boldsymbol{x})$, para todo $\boldsymbol{\tau} \in T$.

Proposición : Sea $\hat{\boldsymbol{\theta}} = (\hat{\theta_1}, \ldots, \hat{\theta_k})$, con $\hat{\theta}_j = \hat{\theta}_j(\boldsymbol{X})$ un estimador de máxima verosimilitud del vector de parámetros $\boldsymbol{\theta}$ de $f(x;\boldsymbol{\theta})$. Sea $\boldsymbol{\tau}(\boldsymbol{\theta}) = (\tau_1(\boldsymbol{\theta}), \ldots, \tau_r(\boldsymbol{\theta}))$. Entonces, el estimador de máxima verosimilitud de $\boldsymbol{\tau}(\boldsymbol{\theta})$ es $\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}(\hat{\boldsymbol{\theta}}) = (\tau_1(\hat{\boldsymbol{\theta}}), \ldots, \tau_r(\hat{\boldsymbol{\theta}}))$.

Demostración: Sea $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ un EMV de $\boldsymbol{\theta}$. Entonces,

$$\begin{aligned} M(\boldsymbol{\tau}, \boldsymbol{x}) &= \sup_{\{\boldsymbol{\theta} : \boldsymbol{\tau}(\boldsymbol{\theta}) = \boldsymbol{\tau})\}} L(\boldsymbol{\theta}; \boldsymbol{x}) \\ &\leq \sup_{\{\boldsymbol{\theta} \in \Theta\}} L(\boldsymbol{\theta}; \boldsymbol{x}) \\ &= L(\hat{\boldsymbol{\theta}}; \boldsymbol{x}) \\ &= \sup_{\{\boldsymbol{\theta} : \boldsymbol{\tau}(\boldsymbol{\theta}) = \boldsymbol{\tau}(\boldsymbol{\theta})\}} L(\boldsymbol{\theta}; \boldsymbol{x}) \\ &= M(\boldsymbol{\tau}(\hat{\boldsymbol{\theta}}); \boldsymbol{x}) \end{aligned}$$

Resumiendo: si $\hat{\boldsymbol{\theta}}$ es un EMV de $\boldsymbol{\theta}$ y $\boldsymbol{\tau} = \boldsymbol{\tau}(\boldsymbol{\theta})$, entonces $\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}(\hat{\boldsymbol{\theta}})$ es un EMV de $\boldsymbol{\tau}$.

Ejemplo : Sea $\boldsymbol{X}$ una *m. a.* de $X \sim \mathrm{N}(\mu, \sigma^2)$. $\theta = \mathbb{E}[X^2] = \mu^2 + \sigma^2$. Entonces, el EMV de θ es

$$\hat{\theta} = \bar{X} + \frac{1}{n} \sum (X_i - \bar{X})^2$$

5.7. Teoría de grandes muestras para EMV

Ejemplo : Considere $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una *m. a.* de $X \sim \mathrm{Geom}(\theta)$ con $\theta \in \Theta = (0, 1)$ y *f. m. p.* dada por

$$f(x;\theta) = \theta(1-\theta)^x \mathbb{1}_{\{0,1,\ldots,\}}(x)$$

con $\mathbb{E}[X] = (1-\theta)/\theta = \theta^{-1} - 1$ y $\text{var}(X) = (1-\theta)/\theta^2 = \theta^{-2}(1-\theta)$. Entonces,

$$L(\theta, \boldsymbol{x}) = \theta^n (1-\theta)^{\sum x_i}$$

$$\ell(\theta; \boldsymbol{x}) = n \log \theta + \sum x_i \log(1-\theta)$$

$$\frac{\partial \ell(\theta; \boldsymbol{x})}{\partial \theta} = \frac{n}{\theta} + \frac{1}{1-\theta} \sum x_i$$

De donde se puede ver que el estimador de máxima verosimilitud es $\hat{\theta} = \dfrac{1}{\bar{X}_n + 1}$.

Ahora bien, se sigue del teorema central del límite que

$$\left(\frac{\bar{X}_n - \frac{1-\theta}{\theta}}{\sqrt{\frac{1-\theta}{n\theta^2}}} \right) \xrightarrow{D} Z \sim \text{N}(0,1)$$

por lo que,

$$\sqrt{n} \left(\bar{X}_n - (\theta^{-1} - 1) \right) \dot{\sim} \text{N} \left(0, \theta^{-2}(1-\theta) \right)$$

Considere ahora $h(x) = 1/(1+x)$, luego $h'(x) = -1/(1+x)^2$. Se sigue del método delta que $\text{var}(h(X)) = (h'(\mu_X))^2 \sigma_X^2$, por lo que

$$\text{var}(h(\bar{X}_n)) = \frac{\sigma^2_{\bar{X}_n}}{(1+\mu_{\bar{X}_n})^4} = \frac{\frac{1-\theta}{\theta^2}}{n \left(1 + \frac{1-\theta}{\theta}\right)^4} = \theta^2(1-\theta)/n$$

Por lo tanto, puesto que $\hat{\theta}_n = h(\bar{X}_n)$ se sigue del teorema del mapeo continuo que para n *grande*

$$\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n} \left(h(\bar{X}_n) - h(\theta^{-1} - 1) \right) \dot{\sim} \text{N} \left(0, \theta^2(1-\theta) \right)$$

o bien, para n grande se tiene

$$\hat{\theta}_n \dot{\sim} \text{N}(\theta, \theta^2(1-\theta)/n)$$

En general, bajo ciertas condiciones razonables los estimadores de máxima verosimilitud son estimadores consistentes.

En esta sección se presentará también el concepto de varianza asintótica y se mostrará que para muestras grandes los estimadores de máxima verosimilitud se distribuyen aproximadamente de manera normal.

Las demostraciones formales de los resultados antes mencionados son muy técnicas. Aquí se presentarán versiones heurísticas y poco detalladas para los casos univariados y trabajando con muestras aleatorias (*v.a.i.i.d.*). Vea por ejemplo Rice (2007) y Knight (2000). Para mayor formalidad y detalle consulte Casella and Berger (2002), Dudewicz and Mishra (1988) o Bickel and Doksum (1977).

Sean,

- $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de tamaño n de una población $X \sim f(x; \theta)$ con $\theta \in \Theta$.
- $L(\theta; \boldsymbol{x}) = \prod f(x_i; \theta)$, la función de verosimilitud.
- $\ell(\theta; \boldsymbol{x}) = \log L(\theta, \boldsymbol{x}) = \sum \log f(x_i; \theta)$ la función log de la verosimilitud.

Sea θ_0 el verdadero (y desconocido) valor del parámetro θ. Se mostrará que $\hat{\theta}_n = \hat{\theta}(\boldsymbol{X}_n) = \hat{\theta}_{\text{EMV}}$ es un estimador consistente. Esto es,

$$\hat{\theta}_n \xrightarrow{\text{P}} \theta_0$$

Condiciones de Regularidad (CR)

Las **condiciones de regularidad** son en general condiciones de suavidad de las funciones relacionadas a la función de verosimilitud. Entre ellas está el permitir intercambiar el orden de integración y derivación. Dependiendo del resultado es la necesidad de apelar a una o varias de las condiciones o supuestos que a continuación se enuncian.

I) El espacio parametral Θ es un subconjuntos abierto de $\mathbb{R}$.

II) El soporte de la densidad $f(x;\theta)$ no depende del parámetro θ.

III) $\dfrac{\partial}{\partial\theta}\log f(x;\theta)$ existe para todo x y para todo θ.

IV) $\dfrac{\partial}{\partial\theta}\displaystyle\int_{\mathbb{R}^n}\prod_{i=1}^{n} f(x_i;\theta)\ d\boldsymbol{x} = \int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i;\theta)\ d\boldsymbol{x}$

V) $\dfrac{\partial}{\partial\theta}\displaystyle\int_{\mathbb{R}^n} t(\boldsymbol{x})\prod_{i=1}^{n} f(x_i;\theta)\ d\boldsymbol{x} = \int_{\mathbb{R}^n} t(\boldsymbol{x})\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i;\theta)\ d\boldsymbol{x}$

VI) $0 < \mathbb{E}_\theta\left[\left(\dfrac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] < \infty$, para todo $\theta\in\Theta$

Consistencia

Teorema : Bajo condiciones apropiadas de suavidad (CR) sobre f, los estimadores de máxima verosimilitud (EMV) a partir de una muestra aleatoria (*v.a.i.i.d.*'s) son consistentes.

Demostración: Considere maximizar

$$\frac{1}{n}\ell(\theta;X) = \frac{1}{n}\sum \log f(X;\theta)$$

Se sigue de la ley de los grandes números

$$\frac{1}{n}\sum \log f(X_i;\theta) \xrightarrow{P} \mathbb{E}\left[\log f(X;\theta_0)\right]$$

con $\mathbb{E}[\log f(X;\theta)] = \int_{\mathbb{R}}\log f(x;\theta)\cdot f(x;\theta_0)dx$. Luego, se puede pensar que para *"n grande"*, el valor de θ que maximiza $\ell(\theta;X)$, $\hat{\theta}_n$, esté cerca de θ_0, aquel que maximiza $\mathbb{E}[\log f(X:\theta)]$.

Ahora bien, para maximizar $\mathbb{E}[\log f(X;\theta)]$, se toma derivada

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathbb{E}\left[\log f(X;\theta_0)\right] &= \frac{\partial}{\partial\theta}\left[\int_{\mathbb{R}}\log f(x;\theta)f(x;\theta_0)dx\right]\\
&\overset{\text{CR}}{=} \int_{\mathbb{R}}\frac{\partial}{\partial\theta}\left(\log f(x;\theta)\right) f(x;\theta_0)dx\\
&= \int_{\mathbb{R}}\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta)\ f(x;\theta_0)dx\\
&= \int_{\mathbb{R}}\frac{\partial}{\partial\theta}f(x;\theta)dx\\
&\overset{\text{CR}}{=} \frac{\partial}{\partial\theta}\int_{\mathbb{R}}f(x;\theta)dx\\
&= 0
\end{aligned}$$

con la penúltima línea cuando $\theta=\theta_0$, lo que muestra que θ_0 es un punto estacionario de ℓ y posiblemente correspondiente a un máximo.

Definición : Se define la **función score** S por

$$S(\theta; x) = \frac{\partial}{\partial\theta} \log f(x;\theta) = \frac{\partial}{\partial\theta}\ell(\theta; x)$$

Proposición : Bajo condiciones de regularidad (CR) se tiene que

$$\mathbb{E}[S(\theta; X)] = \mathbb{E}\left[\frac{\partial}{\partial\theta} \log f(X;\theta)\right] = 0$$

Demostración:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial\theta} \log f(X:\theta)\right] &= \int_{\mathbb{R}} \frac{\partial}{\partial\theta} \log f(x;\theta) f(x;\theta) dx \\
&= \int_{\mathbb{R}} \frac{1}{f(x;\theta)} \ \frac{\partial}{\partial\theta} f(x;\theta) \ f(x;\theta) dx \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial\theta} f(x;\theta) dx \\
&= \frac{\partial}{\partial\theta} \int_{\mathbb{R}} f(x;\theta) dx \\
&= 0
\end{aligned}$$

Número de Información

Definición : Se define la varianza de la función *score* como el **número de información (de Fisher)**.

$$\mathrm{I}(\theta) = \mathrm{var}(S(\theta; X)) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f(X;\theta)\right)^2\right]$$

Lema : Sea $\mathrm{I}(\theta)$ el número de información de Fisher, entonces, bajo condiciones de regularidad se tiene

$$\mathrm{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X;\theta)\right]$$

Demostración: Recuerde que $\int_{\mathbb{R}} f(x;\theta)dx = 1$, luego $\dfrac{\partial}{\partial\theta}\displaystyle\int_{\mathbb{R}} f(x;\theta)dx = 0$. Note además que

$$\frac{\partial}{\partial\theta} f(x;\theta) = \left[\frac{\partial}{\partial\theta} \log f(x;\theta)\right] f(x;\theta) \tag{6}$$

pues $\dfrac{\partial}{\partial\theta} \log f(x;\theta) = \dfrac{1}{f(x;\theta)} \dfrac{\partial}{\partial\theta} f(x;\theta)$. Entonces,

$$0 = \frac{\partial}{\partial\theta} \int_{\mathbb{R}} f(x;\theta)dx \overset{\text{CR}}{=} \int_{\mathbb{R}} \left[\frac{\partial}{\partial\theta} \log f(x;\theta)\right] f(x;\theta)dx$$

Tomando segunda derivada y empleando (6) se tiene

$$0 = \frac{\partial}{\partial\theta} \int_{\mathbb{R}} \left[\frac{\partial}{\partial\theta} \log f(x;\theta)\right] f(x;\theta)dx = \int_{\mathbb{R}} \left[\frac{\partial^2}{\partial\theta^2} \log f(x;\theta)\right] f(x;\theta)dx + \int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2 f(x;\theta)dx$$

Así, se concluye que

$$\mathrm{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X:\theta)\right]$$

Distribución asintótica

Teorema : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, una muestra aleatoria de $X \sim f(x;\theta)$, con $\theta \in \Theta$, si $\hat{\theta}_n = \hat{\theta}(\boldsymbol{X}_n)$ es el EMV del parámetro θ_0. Entonces, bajo condiciones de regularidad se tiene que

$$\sqrt{n\mathrm{I}(\theta_0)}\,(\hat{\theta}_n - \theta_0) \xrightarrow{D} Z \sim \mathrm{N}(0,1)$$

Esto es, los estimadores de máxima verosimilitud se distribuyen asintóticamente normal. El estimador $\hat{\theta}_n$ se dice **asintóticamente insesgado** y se define la **varianza asintótica** del estimador $\hat{\theta}$ por $\dfrac{1}{n\mathrm{I}(\theta_0)}$.

Demostración: Considere la función log de verosimilitud $\ell(\theta;\boldsymbol{X}) = \log L(\theta;\boldsymbol{X})$. Entonces, puesto que $\hat{\theta}$ corresponde a un máximo de $\ell(\theta)$, se sigue que

$$0 = \ell'(\hat{\theta}) \approx \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0)$$

al expandir la función *score* ℓ' alrededor del verdadero valor θ_0. Se sigue que

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\frac{n^{-1/2}\,\ell'(\theta_0)}{n^{-1}\,\ell''(\theta_0)}$$

Considere el numerador:

$$\mathbb{E}\left[n^{-1/2}\ell'(\theta_0,\boldsymbol{X})\right] = \frac{1}{\sqrt{n}}\mathbb{E}\left[S(\theta,\boldsymbol{X})\right] = 0$$

$$\mathrm{var}(n^{-1/2}\ell'(\theta_0,\boldsymbol{X})) = \frac{1}{n}\mathrm{var}\left(S(\theta,\boldsymbol{X})\right) = \mathrm{I}(\theta_0)$$

Considere ahora el denominador, se sigue de la LGN,

$$\frac{1}{n}\ell''(\theta_0;,\boldsymbol{X}) = \frac{1}{n}\sum \frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta_0) \xrightarrow{P} \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] = -\mathrm{I}(\theta_0)$$

Entonces,

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \frac{n^{-1/2}\ell'(\theta_0)}{\mathrm{I}(\theta_0)}$$

De donde,

$$\mathbb{E}[\sqrt{n}(\hat{\theta} - \theta_0)] \approx 0$$

$$\mathrm{var}\left(\sqrt{n}(\hat{\theta} - \theta_0)\right) \approx \frac{1}{\mathrm{I}^2(\theta_0)}\mathrm{I}(\theta_0)$$

$$\mathrm{var}\left(\hat{\theta} - \theta_0\right) \approx \frac{1}{n\mathrm{I}(\theta_0)}$$

Finalmente, note que $\ell'(\theta_0;\boldsymbol{X}) = \sum S(\theta;X_i) = \sum \frac{\partial}{\partial\theta}\log f(X_i;\theta)$, por lo que se sigue del *TCL* que $\sqrt{n/\mathrm{I}(\theta_0)}\ell'(\theta;\boldsymbol{X}) \xrightarrow{D} Z \sim \mathrm{N}(0,1)$, y por el teorema de Slutsky,

$$\sqrt{n\mathrm{I}(\theta_0)}(\hat{\theta} - \theta_0) = -\sqrt{\mathrm{I}(\theta_0)}\frac{n^{-1/2}\ell'(\theta_0,\boldsymbol{X})}{n^{-1}\ell''(\theta_0;\boldsymbol{X})} \xrightarrow{D} Z \sim \mathrm{N}(0,1)$$

Corolario : $\hat{\theta}_n$, estimador de máxima verosimilitud del parámetro θ_0. Entonces,

$$\hat{\theta}_n \dot{\sim} \mathrm{N}\left(\theta_0, \frac{1}{n\mathrm{I}(\theta_0)}\right)$$

Definición : Se define la **varianza asintótica** de $\hat{\theta}$ por

$$\text{v.a.}(\hat{\theta}) = \frac{1}{n\mathrm{I}(\theta_0)} = \frac{1}{\mathbb{E}[\ell''(\theta_0, \boldsymbol{X})]}$$

Note, cuando $\mathbb{E}[\ell''(\theta_0)]$ es grande, en promedio $\ell(\theta)$ cambia rápidamente en vecindad de θ_0 y la varianza del estimador $\hat{\theta}$ es pequeña. O bien, *"La información sobre θ contenida en $L(\theta; \boldsymbol{X})$ es grande"*.

Existe el resultado equivalente para el caso multivariado. El vector $\hat{\boldsymbol{\theta}}_n$ es distribuido asintóticamente normal con $\mathbb{E}[\hat{\boldsymbol{\theta}}_n] \to \boldsymbol{\theta}_0$ y $\text{cov}(\boldsymbol{\theta}_n) \to \frac{1}{n}\mathrm{I}^{-1}(\boldsymbol{\theta}_0)$, donde

$$\big(\mathrm{I}(\boldsymbol{\theta}_0)\big)_{ij} = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}_i} \log f(\boldsymbol{X}; \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}_j} \log f(\boldsymbol{X}; \boldsymbol{\theta})\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(\boldsymbol{X}; \boldsymbol{\theta})\right]$$

5.8. Cota inferior de Cramér-Rao

En esta sección se revisa y demuestra la desigualdad de Cramér-Rao. Para esto se presentan los siguientes resultados.

Lema :

$$\int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} \prod f(x_i; \theta) d\boldsymbol{x} = 0$$

Demostración:

$$\int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} \prod f(x_i; \theta) d\boldsymbol{x} \overset{\text{CR}}{=} \frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} \prod f(x_i; \theta) d\boldsymbol{x} = \frac{\partial}{\partial \theta}[1] = 0$$

Lema :

$$\frac{\partial}{\partial \theta} \prod f(x_i; \theta) = \prod f(x_i; \theta) \frac{\partial}{\partial \theta} \log \prod f(x_j; \theta)$$

Demostración:

$$\begin{aligned}
\frac{\partial}{\partial \theta} \prod f(x_i; \theta) &= \frac{\partial}{\partial \theta} e^{\log \prod f(x_i; \theta)} \\
&= e^{\log \prod f(x_i; \theta)} \cdot \frac{\partial}{\partial \theta} \log \prod f(x_j; \theta) \\
&= \prod f(x_i; \theta) \frac{\partial}{\partial \theta} \log \prod f(x_j; \theta)
\end{aligned}$$

Teorema : Desigualdad de Cauchy-Schwarz Sea X y Y dos variables aleatorias con varianza finita. Entonces,

$$\mathbb{E}^2[XY] \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

cumpliéndose la igualdad si y solo si $\mathbb{P}(Y = 0) = 1$, o bien, $\mathbb{P}(X = cY) = 1$, para alguna constante c.

Teorema : Desigualdad de Cramér-Rao Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x; \theta)$. Sea $T_n = T(\boldsymbol{X}_n)$ un estimador insesgado de $\tau = \tau(\theta)$, τ, función real diferenciable de θ. Entonces, bajo ciertas condiciones de regularidad (CR),

$$\text{var}_\theta(T_n) \geq \frac{\big[\tau'(\theta)\big]^2}{n\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial \theta} \log f(X; \theta)\right)^2\right]}$$

La igualdad se cumple si y solo si existe una función, digamos $K(\theta; n)$ tal que

$$\sum \frac{\partial}{\partial \theta} \log f(x_i; \theta) = K(\theta; n) \left[T(\boldsymbol{x}_n) - \tau(\theta) \right]$$

Demostración:

$$\begin{aligned}
\tau'(\theta) &= \frac{\partial}{\partial \theta} \tau(\theta) \\
&= \frac{\partial}{\partial \theta} \mathbb{E}_\theta[T_n] \\
&= \frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} T(\boldsymbol{x}) f(\boldsymbol{x}; \theta) d\boldsymbol{x} + 0 \\
&\overset{\text{CR}}{=} \int_{\mathbb{R}^n} T(\boldsymbol{x}) \frac{\partial}{\partial \theta} \prod f(x_i; \theta)\, d\boldsymbol{x} - \tau(\theta) \frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} \prod f(x_i; \theta)\, d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} \left[T(\boldsymbol{x}) - \tau(\theta) \right] \frac{\partial}{\partial \theta} \prod f(x_i; \theta) d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} \left(T(\boldsymbol{x}) - \tau(\theta) \right) \frac{\partial}{\partial \theta} \log \prod f(x_j; \theta) \prod f(x_i; \theta) d\boldsymbol{x} \\
&= \mathbb{E}_\theta \left[\left(T(\boldsymbol{X}) - \tau(\theta) \right] \left\{ \frac{\partial}{\partial \theta} \log \prod f(X_i; \theta) \right\} \right]
\end{aligned}$$

Se sigue de la desigualdad de Cauchy-Schwarz que

$$\left[\tau'(\theta) \right]^2 \leq \mathbb{E}_\theta \left[\left(T(\boldsymbol{X}) - \tau(\theta) \right)^2 \right] \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log \prod f(X_j; \theta) \right)^2 \right]$$

O bien,

$$\operatorname{var}(T(\boldsymbol{X})) \geq \frac{\left[\tau'(\theta) \right]^2}{\mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log \prod f(X_j; \theta) \right)^2 \right]}$$

Pero, se tiene que,

$$\begin{aligned}
\mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log \prod f(X_j; \theta) \right)^2 \right] &= \mathbb{E}_\theta \left[\left(\sum \frac{\partial}{\partial \theta} \log f(X_j; \theta) \right)^2 \right] \\
&= \sum_{i,j} \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X_i; \theta) \right) \left(\frac{\partial}{\partial \theta} \log f(X_j; \theta) \right) \right] \\
&= \sum_{i} \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X_i; \theta) \right)^2 \right] \\
&= n \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X_i; \theta) \right)^2 \right]
\end{aligned}$$

pues X_i y X_j son independientes. Además,

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] &= \int_{\mathbb{R}}\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] f(x;\theta)dx\\ &= \int_{\mathbb{R}}\frac{\partial}{\partial\theta}f(x;\theta)dx\\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f(x;\theta)dx\\ &= 0\end{aligned}$$

En la desigualdad de Cauchy-Schwarz se da la igualdad si y solo si una de las *v. a.*'s es múltiplo de la otra con probabilidad 1. En este caso, se requiere que

$$\frac{\partial}{\partial\theta}\log\prod f(x_j;\theta) \propto \big(T(\boldsymbol{x}_n)-\tau(\theta)\big)$$

O bien, que exista una constante $K = K(\theta;n)$ tal que

$$\frac{\partial}{\partial\theta}\log\prod f(x_j;\theta) = K(\theta,n)\left[T(\boldsymbol{x})-\tau(\theta)\right]$$

De la desigualdad de Cramér-Rao se sigue:

- Si la varianza del estimador es cercana a la CICR es estimador insesgado es *"bueno"*.
- Si la varianza del estimador alcanza la CICR entonces el estimador es UMVUE.

Ejemplo[13]: Sea $\boldsymbol{X} = (X_1,\ldots,X_n)$ una *m. a.* de $X \sim \text{Po}(\lambda)$. Luego,

- $f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0,1,2,\ldots$.
- $\frac{\partial}{\partial\lambda}\log f(x;\lambda) = \frac{\partial}{\partial\lambda}(-\lambda + x\log\lambda - \log x!) = -1 + \frac{x}{\lambda}$

Entonces,

$$\begin{aligned}\mathbb{E}_\lambda\left[\left(\frac{\partial}{\partial\lambda}\log f(X;\lambda)\right)^2\right] &= \mathbb{E}_\lambda\left[1 - \frac{2X}{\lambda} + \frac{X^2}{\lambda^2}\right]\\ &= 1 - 2 + \frac{1}{\lambda^2}(\lambda+\lambda^2)\\ &= 1/\lambda\end{aligned}$$

por lo que el denominador de la cota de Cramer-Rao es n/λ.

Por otro lado, si $\tau = \tau(\lambda) = \mathbb{P}(X=0) = e^{-\lambda}$, y $T_n = T(\boldsymbol{X}_n)$, estimador insesgado de τ, se sigue de la desigualdad de Cramér-Rao,

$$\text{var}(T_n) \geq \frac{\left[e^{-\lambda}\right]^2}{n/\lambda} = \frac{1}{n}\lambda e^{-2\lambda}$$

Note, si $T_n = T(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{0\}}(X_i)$ entonces $\mathbb{E}[T_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{P}(X_i=0) = e^{-\lambda} = \tau$. Esto es, T_n es un estimador insesgado de τ, y

$$\text{var}(T_n) = \frac{1}{n}e^{-\lambda}(1-e^{-\lambda})$$

que ciertamente es mayor que la cota inferior de Cramér-Rao, $\lambda e^{-2\lambda}/n!$. Más adelante se encontrará un UMVUE para τ.

[13] Mood, Graybill, and Boes (1974)

5.9. Estadísticos suficientes

A principios del los años 1920's Ronald A. Fisher estudiaba la estimación de σ^2 de una población normal con base en una muestra aleatoria $\boldsymbol{X}$. Comparó la inferencia sobre σ basado en los estadísticos[14]

$$T_1(\boldsymbol{X}) = \sum |X_i - \bar{X}| \quad \text{y} \quad T_2(\boldsymbol{X}) = \sum (X_i - \bar{X})^2$$

Fisher mostró que la distribución de $T_1|T_2 = t$ no dependía de σ mientras que la de $T_2|T_1 = t$, sí dependía de σ. Concluyó que $T_2(\boldsymbol{X})$ capturaba toda la información sobre σ contenida en la muestra mientras que $T_1(\boldsymbol{X})$ no.

Suponga $\boldsymbol{X}$ una muestra aleatoria de $X \sim f(x;\theta)$, $\theta \in \Theta$ y sea $T(\boldsymbol{X})$ un estadístico. Para todo t sea $A_t = \{\boldsymbol{x} \in \boldsymbol{X} : T(\boldsymbol{x}) = t\}$. Los A_t forman una partición de $\boldsymbol{X}$. Considere ahora la distribución de $\boldsymbol{X}$ en el conjunto A_t (condicional de $\boldsymbol{X}$ dado $T = t$). Si tal distribución es independiente de θ, $\boldsymbol{X}$ no tiene información de θ en A_t, $\boldsymbol{X}$ es un estadístico auxiliar en A_t. Si lo anterior se cumple para todo A_t, $T(\boldsymbol{X})$ tiene la misma información sobre θ que $\boldsymbol{X}$. La misma idea se presenta en el siguiente principio.

Definición : Sea $\boldsymbol{X}$ una muestra aleatoria de $X \sim f(x;\theta)$. $T(\boldsymbol{X})$ se dice **estadístico auxiliar** (*ancilliary statistic*) de θ si su distribución es independiente de θ. Esto es, si para todo $\theta \in \Theta$, la distribución de T es la misma.

Ejercicio : Sea $\boldsymbol{U} = (U_1, \ldots, U_n)$ una muestra aleatoria de $U \sim \text{Unif}(\mu-\theta, \mu+\theta)$. Entonces, el rango muestral $R(\boldsymbol{U}) = U_{(n)} - U_{(1)}$, es un estadístico auxiliar para μ.

Definición : Sea $\boldsymbol{X}$ una muestra aleatoria de $X \sim f(x;\theta)$. Un estadístico $S(\boldsymbol{X})$ se dice **estadístico suficiente** para el parámetro θ (puede ser un vector) si y solo si la distribución condicional de $\boldsymbol{X}$ dado $S = s$ no depende de θ para cualquier valor de $S(\boldsymbol{x}) = s$.

Equivalentemente[15], $S(\boldsymbol{X})$ es un estadístico suficiente si y solo si, para todo $S(\boldsymbol{x}) = s$, fijo con *f. d. p.* f_S,

$$\frac{\prod f(x_i;\theta)}{f_S(s;\theta)} = h(\boldsymbol{x})$$

y donde $h(\boldsymbol{x})$ no depende de θ. Como consecuencia, si $T(\boldsymbol{X})$ es otro estadístico, dado $S(\boldsymbol{x})$, no hay manera que T pueda ser usado para hacer inferencia sobre θ, pues $h(\boldsymbol{x})$ no depende de ella.

Note que trivialmente la muestra misma $(X_1, \ldots, X_n)$ o la muestra ordenada $(X_{(1)}, \ldots, X_{(n)})$ son estadísticos suficientes.

Principio de Suficiencia[16]

> Si $T(\boldsymbol{X})$ es un estadístico suficiente de θ, entonces la inferencia sobre θ deberá depender de $\boldsymbol{X}$ solo a través de T. Esto es, si $\boldsymbol{x}$ y $\boldsymbol{y}$ son dos muestras tales que $T(\boldsymbol{x}) = T(\boldsymbol{y})$, entonces la inferencia sobre θ es indistinta si se basa en $\boldsymbol{x}$ o $\boldsymbol{y}$.

Ejemplo[17]: Considere $\boldsymbol{X}_3 = (X_1, X_2, X_3)$, *m. a.* de $X \sim \text{Ber}(p)$. Considere los estadísticos

[14] Knight (2000).
[15] Hogg and Craig (1978).
[16] Casella and Berger (2002).
[17] Mood, Graybill, and Boes (1974).

$S(\boldsymbol{X}) = X_1 + X_2 + X_3$ y $T(\boldsymbol{X}) = X_1X_2 + X_3$. Se tiene por ejemplo que

$$\begin{aligned} f_{\boldsymbol{X}|S}((0,1,0)|S=1) &= \mathbb{P}(X_1=0, X_2=1, X_3=0|S=1) \\ &= \frac{\mathbb{P}(X_1=0, X_2=1, X_3=0, S=1)}{\mathbb{P}(S=1)} \\ &= \frac{(1-p)p(1-p)}{\binom{3}{1}p(1-p)^2} \\ &= \frac{1}{3} \end{aligned}$$

mientras que

$$\begin{aligned} f_{\boldsymbol{X}|T}(0,1,0|T=0) &= \mathbb{P}(X_1=0, X_2=1, X_3=0|T=0) \\ &= \frac{\mathbb{P}(X_1=0, X_2=1, X_3=0, T=0)}{\mathbb{P}(T=0)} \\ &= \frac{p(1-p)^2}{(1-p)^3 + 2p(1-p)^2} \\ &= \frac{p}{1+p} \end{aligned}$$

De esta forma se obtiene la tabla

$\boldsymbol{x}$	$S=s$	$T=t$	$f(\boldsymbol{x}\|s)$	$f(\boldsymbol{x}\|t)$
$(0,0,0)$	0	0	1	$\frac{1-p}{1+p}$
$(0,0,1)$	1	1	$\frac{1}{3}$	$\frac{1-p}{1+2p}$
$(0,1,0)$	1	0	$\frac{1}{3}$	$\frac{p}{1+p}$
$(1,0,0)$	1	0	$\frac{1}{3}$	$\frac{p}{1+p}$
$(0,1,1)$	2	1	$\frac{1}{3}$	$\frac{p}{1+2p}$
$(1,0,1)$	2	1	$\frac{1}{3}$	$\frac{p}{1+2p}$
$(1,1,0)$	2	1	$\frac{1}{3}$	$\frac{p}{1+2p}$
$(1,1,1)$	3	2	1	1

En el ejemplo anterior note que la distribución de $\boldsymbol{X}$ dado $S = X_1+X_2+X_3$ no dependen del parámetro p mientras que la distribución condicionada en $T(\boldsymbol{X}) = X_1X_2+X_3$ sí depende de p. S es un estadístico suficiente para p, T no lo es.

Teorema Factorización Neyman–Fisher [18] Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra con función de densidad conjunta $f(\boldsymbol{x};\theta)$, $\theta \in \Theta$, Entonces, $S = S(\boldsymbol{X})$ es un estadístico suficiente para θ, si y solo si, se tienen g y h, funciones no negativas tales que

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = g(S(\boldsymbol{x});\theta)h(\boldsymbol{x}) \tag{7}$$

para todo $S(\boldsymbol{x}) = s$ fijo y la función h no depende de θ. S y θ pueden ambos ser vectores.
Demostración:

[18] Rice (2007), Hogg and Craig (1978).

- *Caso discreto:* Suponga que S es un estadístico suficiente. Entonces,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x};\theta) &= \mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}) \\ &= \sum_s \mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}, S=s) \\ &= \mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}, S=S(\boldsymbol{x})) \\ &= \mathbb{P}_\theta(S=S(\boldsymbol{x}))\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|S=S(\boldsymbol{x})) \\ &= g(S(\boldsymbol{x});\theta)h(\boldsymbol{x}) \end{aligned}$$

  pues S es un estadístico suficiente por lo que $\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|S=S(\boldsymbol{x}))$ es independiente de θ.

  Suponga ahora que $f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = g(S(\boldsymbol{x});\theta)h(x)$. Entonces, para $S(\boldsymbol{x})=s$,

$$\begin{aligned} \mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|S=s) &= \frac{\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}_\theta(S=s)} \\ &= \frac{g(S(\boldsymbol{x});\theta)h(\boldsymbol{x})}{\sum_{S(\boldsymbol{y})=s} g(S(\boldsymbol{y});\theta)h(\boldsymbol{y})} \\ &= \frac{h(\boldsymbol{x})}{\sum_{S(\boldsymbol{y})=s} h(\boldsymbol{y})} \end{aligned}$$

  que no depende de θ. Si $S(\boldsymbol{x}) \neq s$, $\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|S=s)=0$. En cualquier caso, $\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|S=s)$ es independiente de θ, por lo que S es suficiente.

- *Caso continuo:* Suponga X *v. a.* continua con *f. d. p.* $f(x;\theta)$, $\theta\in\Theta$ y tal que acepta la factorización (7) para todo $s=S(\boldsymbol{x})$ fijo. Defina $y_1 = s(\boldsymbol{x}) = G_1(\boldsymbol{x}), y_2 = G_2(\boldsymbol{x}), \ldots, y_n = G_n(\boldsymbol{x}))$, con la transformación inversa $\boldsymbol{H} = (H_1,\ldots,H_n)$, tal que $x_i = H_i(\boldsymbol{y})$. Se sigue del teorema de transformación que

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\theta) = f_{\boldsymbol{X}}\left(\boldsymbol{H}(\boldsymbol{y});\theta\right)\,|J\boldsymbol{H}(\boldsymbol{y})|$$

  Se sigue de (7) que la *f. d. p.* marginal de $S=Y_1$

$$\begin{aligned} f_S(s;\theta) &= \int_{\mathbb{R}^{n-1}} f_{\boldsymbol{X}}\left(\boldsymbol{H}(\boldsymbol{y});\theta\right)|J\boldsymbol{H}(\boldsymbol{y})|dy_2\cdots dy_n \\ &= g(s;\theta)\int_{\mathbb{R}^{n-1}} h\left(\boldsymbol{H}(\boldsymbol{y})\right)|J\boldsymbol{H}(\boldsymbol{y})|dy_2\cdots dy_n \end{aligned} \tag{8}$$

  Considere ahora la función h, para $s=y_1=G_1(\boldsymbol{x})$ fijo, h no depende de θ, tampoco $J\boldsymbol{H}$ ni los límites de integración, por lo que la integral en (8) depende solamente de s, digamos $I(s)$, por lo que

$$f_S(s;\theta) = g(s;\theta))I(s)$$

  Ahora, si $I(s)=0$, entonces $f_S(s;\theta)=0$, y si $I(s)>0$, $g(s;\theta) = g(s(\boldsymbol{x});\theta) = f_X(s(\boldsymbol{x});\theta)/I(s(\boldsymbol{x}))$, y la factorización queda

$$f(\boldsymbol{X};\theta) = f_S(s(\boldsymbol{x});\theta)\frac{h(\boldsymbol{x})}{I(s(\boldsymbol{x}))}$$

  y el cociente no depende de θ. Luego, $S(\boldsymbol{X})$ es un estadístico suficiente para θ de acuerdo con la definición alternativa.

  Y viceversa, si S es un estadístico suficiente para θ, la factorización se construye tomando a f_S como la función g en la expresión (7).

Ejemplo : Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una *m. a.* de $X \sim \text{Ber}(\theta)$. Luego

$$\begin{aligned} f(\boldsymbol{x};\theta) &= \prod f(x_i;\theta) \\ &= \prod \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \\ &= \left[\left(\frac{\theta}{1-\theta}\right)^{\sum x_i}(1-\theta)^n\right]\cdot 1 \\ &= g(T(\boldsymbol{x});\theta)\cdot h(\boldsymbol{x}) \end{aligned}$$

donde $T(\boldsymbol{x}) = \sum x_i$ y $h(\boldsymbol{x}) = 1$. Se sigue del teorema de factorización que $T(\boldsymbol{X}_n) = \sum_{i=1}^n X_i$ es un estadístico suficiente para θ.

Note: Puesto que $T = \sum X_i$ es un estadístico suficiente para θ, la maximización de $f(\boldsymbol{x};\theta) = g(T(\boldsymbol{x});\theta)h(\boldsymbol{x})$ es a través de la función $g(t,\theta)$. Luego,

$$\begin{aligned} g(t;\theta) &= \theta^t(1-\theta)^{n-t} \\ \log g(t,\theta) &= t\log\theta + (n-t)\log(1-\theta) \\ \frac{\partial}{\partial\theta}\log g &= \frac{t}{\theta} - \frac{n-t}{1-\theta} \end{aligned}$$

Igualando a cero la derivada y despejando para θ se concluye que

$$\hat{\theta}_{\text{EMV}} = \frac{t}{n} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

Ejemplo : Sea $X = (X_1, \ldots, X_n)$ una *m. a.* de $X \sim \text{N}(\mu, \sigma^2)$. Sea el vector de parámetros $\boldsymbol{\theta} = (\mu, \sigma) \in \Theta = \mathbb{R}\times\mathbb{R}^+$. Entonces,

$$\begin{aligned} f(\boldsymbol{x} : \boldsymbol{\theta}) &= \prod f(x_i;\boldsymbol{\theta}) \\ &= (2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right\} \\ &= (2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right]\right\} \\ &= \underbrace{(2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\left[S_1(\boldsymbol{x}) - 2\mu S_2(\boldsymbol{x}) + n\mu^2\right]\right\}}_{g(\boldsymbol{T}(\boldsymbol{x});\boldsymbol{\theta})}\cdot\underbrace{1}_{h(\boldsymbol{x})} \end{aligned}$$

Por lo tanto, $\boldsymbol{S}(\boldsymbol{X}) = \big(S_1(\boldsymbol{X}), S_2(\boldsymbol{X})\big) = \big(\sum X_i^2, \sum X_i\big)$ es un *estadístico suficiente conjunto* para $\boldsymbol{\theta} = (\mu, \sigma)$.

Definición : Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$. El estadístico $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), \ldots, T_r(\boldsymbol{X}))$ es un **estadístico suficiente conjunto** para $\boldsymbol{\theta}$ si y solo si

$$f(\boldsymbol{x};\boldsymbol{\theta}) = g(\boldsymbol{T}(\boldsymbol{x}),\boldsymbol{\theta})\,h(\boldsymbol{x})$$

donde h es una función no negativa que no depende de θ y g es una función no negativa de depende de $\boldsymbol{x}$ solamente a través de $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}$.

Ejemplo : Sea $\boldsymbol{U}_n = (U_1, \ldots, U_n)$ una *m. a.* de $U \sim \text{Unif}(\theta_1, \theta_2)$. Entonces, con $\boldsymbol{\theta} = (\theta_1, \theta_2)$, la *f. d. p.* es $f(u; \boldsymbol{\theta}) = \frac{1}{\theta_2 - \theta_1} \mathbb{1}_{(\theta_1, \theta_2)}(u)$ y la conjunta puede escribirse como

$$f(\boldsymbol{u}_n; \boldsymbol{\theta}) = (\theta_2 - \theta_1)^{-n} \mathbb{1}_{(\theta_1, y_n)}(y_1) \mathbb{1}_{(y_1, \theta_2)}(y_n)$$

donde $y_i = u_{(i)}$, $i = 1, \ldots, n$, son los estadísticos de orden. Luego, $\boldsymbol{T}(\boldsymbol{U}_n) = (U_{(1)}, U_{(n)})$, es un estadístico suficiente conjunto para $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

Note que si $\theta_1 = \theta$ y $\theta_2 = \theta + 1$, $\boldsymbol{T}(\boldsymbol{U}_n) = (U_{(1)}, U_{(n)})$, sigue siendo un estadístico suficiente conjunto. Sin embargo, si $\theta_1 = 0$ y $\theta_2 = \theta$, entonces, $T(\boldsymbol{U}_n) = U_{(n)}$ es un estadístico suficiente por él mismo.

El estudio de las distribuciones paramétricas con estadísticos suficientes $\boldsymbol{T}$ con la misma dimensión que el espacio parametral Θ independiente del tamaño de la muestra llevó a la construcción de la *familia exponencial*[19]. Muchas de las distribuciones más comunes pertenecen a esta familia, como las distribuciones Poisson, Normal, Gamma, etc.

Definición : $X \sim f(x; \theta)$, $\theta \in \Theta \subseteq \mathbb{R}$, f se dice miembro de la **familia o clase exponencial de un parámetro** si para todo $\theta \in \Theta$,

$$f(x; \theta) = \exp\left\{c(\theta)T(x) + d(\theta) + S(x)\right\} \mathbb{1}_{\mathcal{S}_X}(x) \tag{9}$$

para funciones c, d, T y S, apropiadas y donde $\mathcal{S}_X$ es el soporte de la distribución que no depende de θ.

Proposición : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x; \theta)$ miembro de la familia exponencial. Entonces,

$$\begin{aligned} f(\boldsymbol{x}_n; \theta) &= \exp\left\{c(\theta) \sum_{i=1}^n T(x_i) + nd(\theta) + \sum_{i=1}^n S(x_i)\right\} \prod_{i=1}^n \mathbb{1}_{\mathcal{S}_X}(x_i) \\ &= \underbrace{\exp\left\{c(\theta) \sum_i T(x_i) + nd(\theta)\right\}}_{g(T(\boldsymbol{x}_n); \theta)} \underbrace{\exp\left\{\sum_i S(x_i)\right\} \prod_i \mathbb{1}_{\mathcal{S}_X}(x_i)}_{h(\boldsymbol{x}_n)} \end{aligned}$$

por lo que es claro que $T_n = T(\boldsymbol{X}_n) = \sum T(X_i)$ es un estadístico suficiente para θ.

Ejemplo : $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ es una *m. a.* de $X \sim \text{Ber}(\theta)$. Entonces,

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x} \mathbb{1}_{\{0,1\}}(x) = \exp\left\{x \log \frac{\theta}{1 - \theta} + \log(1 - \theta)\right\} \mathbb{1}_{\{0,1\}}(x)$$

Por lo que comparando con (9), se tiene: $c(\theta) = \log \frac{\theta}{1-\theta}$, $T(x) = x$, $d(\theta) = \log(1 - \theta)$ y $S(x) = 0$. Luego, la distribución Bernoulli es miembro de la familia exponencial.

Ejercicio : Muestre que las distribuciones Poisson, geométrica y binomial negativa son miembros de la familia exponencial de un parámetro.

Definición : $X \sim f(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$, f se dice miembro de la **familia o clase exponencial de m-parámetros** si para todo $\boldsymbol{\theta} \in \Theta$,

$$f(x; \boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^m c_j(\boldsymbol{\theta}) T_j(x) + d(\boldsymbol{\theta}) + S(x)\right\} \mathbb{1}_{\mathcal{S}_X}(x) \tag{10}$$

[19] Rice (2007).

para funciones c_j, d, T_j y S, apropiadas y donde $\mathcal{S}_X$ es el soporte de la distribución que no depende del vector de parámetros $\boldsymbol{\theta}$.

Nota: m y k no tienen porque ser iguales, aunque en muchos casos lo es.

Los parámetros $\phi_j = c_j(\boldsymbol{\theta})$ se conocen como los **parámetros naturales o canónicos** de la distribución.

Proposición : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x; \boldsymbol{\theta})$ como en la definición anterior. Entonces,

$$\begin{aligned} f(\boldsymbol{x}_n; \boldsymbol{\theta}) &= \exp\left\{ \sum_{j=1}^{m} c_j(\boldsymbol{\theta}) \sum_{i=1}^{n} T_j(x_i) + nd(\boldsymbol{\theta}) + \sum_{i=1}^{n} S(x_i) \right\} \prod_{i=1}^{n} \mathbb{1}_{\mathcal{S}_X}(x_i) \\ &= \underbrace{\exp\left\{ \sum_{j=1}^{m} c_j(\boldsymbol{\theta}) \sum_{i=1}^{n} T_j(x_i) + nd(\boldsymbol{\theta}) \right\}}_{g(\boldsymbol{T}(\boldsymbol{x}); \boldsymbol{\theta})} \underbrace{\exp\left\{ \sum_{i=1}^{n} S(x_i) \right\} \prod_{i=1}^{n} \mathbb{1}_{\mathcal{S}_X}(x_i)}_{h(\boldsymbol{x})} \end{aligned}$$

Por lo que $\boldsymbol{T}_n = \boldsymbol{T}(\boldsymbol{X}_n) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_m(X_i) \right)$ es un estadístico suficiente conjunto para el vector de parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

Ejercicio : Muestre que las distribuciones $\mathrm{N}(\mu, \sigma^2)$ y $\mathrm{Ga}(\alpha, \beta)$ son miembros de la familia exponencial de 2 parámetros y encuentre correspondientes estadísticos suficientes.

Ejercicio : Considere $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, una *m. a.* de $X \sim \mathrm{N}(\theta, \theta^2)$, $\theta \in \Theta = \mathbb{R}^+$. Muestre que $f(x; \theta)$ es miembro de la familia exponencial de dos parámetros a pesar de que el espacio parametral Θ es unidimensional.

5.10. Estimadores insesgados uniformes de varianza mínima

[20] Si un estimador insesgado alcanza la cota de Cramér-Rao (CICR), éste es un estimador insesgado uniforme de varianza mínima (UMVUE). Existen sin embargo estimadores que no alcanzan la cota pero son UMVUE. El detalle es que la CICR puede ser muy restrictiva.

Proposición : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x; \theta)$, $\hat{\theta}_n = \hat{\theta}(\boldsymbol{x}_n)$, un estimador de máxima verosimilitud, solución de

$$\frac{\partial}{\partial \theta} \log L(\theta; \boldsymbol{x}_n) = \frac{\partial}{\partial \theta} \sum \log f(x_i; \theta) = 0$$

Si $T_n = T(\boldsymbol{X}_n)$ es un estimador insesgado de $\tau(\theta)$ cuya varianza alcanza la CICR, entonces $T_n = \tau(\hat{\theta}_n)$.

Demostración: Se sigue del Teorema de Cramér-Rao y la definición de estimadores de máxima verosimilitud que si T_n alcanza la cota,

$$0 = \frac{\partial}{\partial \theta} \log L(\theta; \boldsymbol{x}_n) = K(\theta; n) \left[T(\boldsymbol{x}_n) - \tau(\theta) \right]$$

Lo que dice que bajo condiciones de regularidad los EMV son UMVUE.

Nota: Si $T_n = T(\boldsymbol{X}_n)$ es un estimador insesgado de $\tau(\theta)$ cuya varianza alcanza la CICR entonces, $f(x; \theta)$ es un miembro de la familia exponencial ($\mathcal{F}_{\mathrm{Exp}}$).Y viceversa, si $f \in \mathcal{F}_{\mathrm{Exp}}$, entonces existe un estimador insesgado de $\tau(\theta)$, digamos, $T(\boldsymbol{X}_n)$ cuya varianza alcanza la CICR. De hecho, existe solamente una función (y transformaciones lineales de ella) para

[20] Mood, Graybill, and Boes (1974).

cuya varianza alcanza la cota. Por lo que la cota no es muy útil para encontrar UMVUE salvo en los casos de la familia exponencial de un solo parámetro. Se mostrará la ventaja de los estimadores insesgados función de estadísticos suficientes.

Teorema de Rao-Blackwell[21] Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$, $\theta \in \Theta$ y $T = T(\boldsymbol{X})$ un estimador de $\tau(\theta)$ con varianza finita $\left(\text{var}_\theta(T) < \infty\right)$ para todo θ. Suponga $S = S(\boldsymbol{X})$ un estadístico suficiente para θ y defina $\widehat{T} = \mathbb{E}[T|S]$. Entonces,

$$\mathbb{E}_\theta\left[(\widehat{T} - \tau(\theta))^2\right] \leq \mathbb{E}_\theta\left[(T - \tau(\theta))^2\right], \quad \text{para todo } \theta \in \Theta$$

Demostración: Recuerde que $\mathbb{E}[\widehat{T}] = \mathbb{E}\left[\mathbb{E}\left[T|S\right]\right] = \mathbb{E}[T]$, por lo que comparar ECMs de los estimadores se reduce a comparar sus varianzas.

$$\begin{aligned}
\text{var}(T) &= \text{var}(\mathbb{E}[T|S]) + \mathbb{E}[\text{var}(T|S)] \\
&= \text{var}(\widehat{T}) + \mathbb{E}[\text{var}(T|S)] \\
&\geq \text{var}(\widehat{T})
\end{aligned}$$

a menos que $\text{var}(T|S) = 0$, que sería el caso solamente si T es una función de S, lo que implicaría que $\widehat{T} = T$.

Notas:

- Puesto que $\mathbb{E}[T|S]$ es función del estadístico suficiente S, Rao–Blackwell justifica el uso de estimadores basados en estadísticos suficientes cuando existen. Si un estimador no es función de un estadístico suficiente, éste se puede mejorar.
- Si T estimador de τ es ya función del estadístico suficiente S, entonces $\widehat{T} = \mathbb{E}[T|S] = T$.
- En la práctica, se intenta condicionar en estadísticos *suficientes minimales*.
- El hecho de *"mejorar"* la precisión de un estadístico mediante el proceso de Rao–Blackwell, no lo hace necesariamente UMVUE.

Definición : Un estadístico suficiente se dice **minimal** si y solo si éste es función de cualquier otro estadístico suficiente.

Ejemplo[22]**:** Se tiene una sucesión de ensayos Bernoulli con parámetro de éxito p. Considere $n = 4$ ensayos y las siguientes tres particiones del espacio, inducidas por los correspondientes estadísticos:

T_1: Salida del primer ensayo. Partición muy burda, gruesa.

T_2: Número de éxitos: *Suficiente minimal* para p.

T_3: $T_3 = (T_1, T_2)$. Partición *suficiente* más fina de lo necesario.

La figura 13 ilustra el espacio muestral y las particiones T_i.

[21] Rice (2007).
[22] Garthwaite, Jolliffe, and Jones (2002).

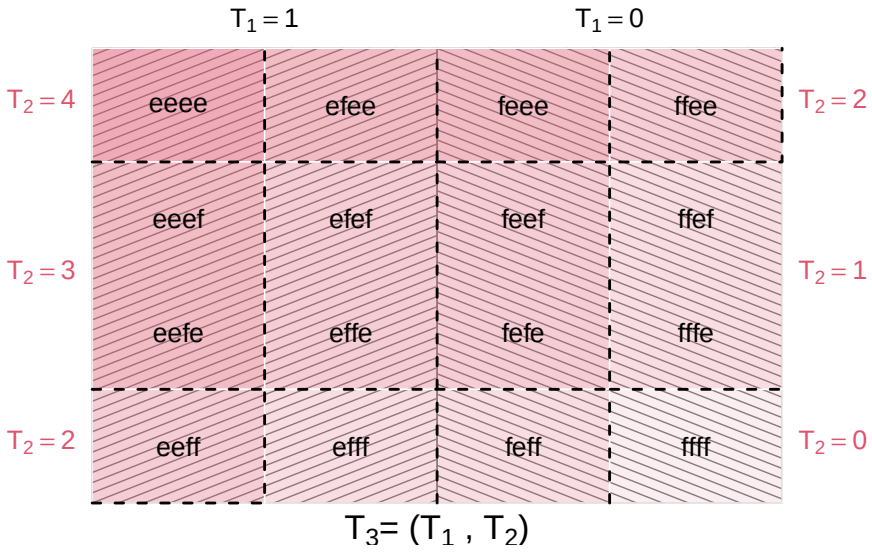


Figura 13: Particiones inducidas por los estadísticos T_1, T_2 y T_3.

Definición : Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$, con $\theta \in \Theta$ y sea $T = T(\boldsymbol{X})$ un estadístico. La familia de densidades $\{f_T(t;\theta) : \theta \in \Theta\}$ se dice **completa** si y solo si $\mathbb{E}_\theta[S(T)] = 0$, para todo $\theta \in \Theta$ implica que $\mathbb{P}_\theta(S(T) = 0) = 1$, para todo $\theta \in \Theta$ y todo estadístico $S(T)$. En este caso se dice que T es un **estadístico completo**.

Alternativamente, T es un estadístico completo si y solo si el único estimador insesgado del 0, función de T es S tal que $\mathbb{P}(S = 0) = 1$.

Ejemplo[23]**:** Sea $X_n = (X_1, \ldots, X_n)$ una *m. a.* de $X \sim \text{Unif}(0, \theta)$, $\theta \in \Theta = \mathbb{R}^+$. Sea $Y_n = X_{(n)} = \text{máx}\{X_i\}$. Entonces Y_n es un estadístico completo. Esto es, por mostrar que si $\mathbb{E}_\theta[T(Y_n)] \equiv 0$, para todo $\theta > 0$, entonces $\mathbb{P}_\theta(T(Y_n) = 0) = 1$, para todo $\theta > 0$.

En efecto, Sea f_n la *f. d. p.* de Y_n. Luego, si para todo $\theta > 0$

$$\mathbb{E}_\theta[T(Y_n)] = \int_{\mathbb{R}} T(y) f_n(y) dy = \int_0^\theta T(y)\theta^{-n} n y^{n-1} dy = \frac{n}{\theta^n} \int_0^\theta T(y) y^{n-1} dy \equiv 0$$

diferenciando ambos lados de la integral de la derecha con respecto a θ, se tiene que

$$T(\theta)\theta^{n-1} = 0$$

para todo θ. Entonces, se debe tener que $T(\theta) = 0$, para todo $\theta > 0$.

Teorema[24] Si T es un estadístico suficiente completo entonces es suficiente minimal pero no viceversa.

Determinar si un estadístico o familia es completa puede ser una tarea muy técnica y no fácil. Sin embargo hay casos donde se puede asegurar la completez. Tal es el caso de la familia exponencial.

Teorema[25] Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$, con $\theta \in \Theta$ un intervalo. Si $f \in \mathcal{F}_{\text{Exp}}$, esto es si

$$f(x;\theta) = \exp\{c(\theta)T(x) + d(\theta) + S(x)\}\mathbb{1}_{S_X}(x)$$

Entonces, $T(\underset{\sim}{X}) = \sum T(X_i)$ es un estadístico completo minimal.

[23] Mood, Graybill, and Boes (1974) Ejemplo VII.5-33.
[24] Knight (2000).
[25] Mood, Graybill, and Boes (1974).

Teorema de Lehmann–Scheffé Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$. Si $S(\boldsymbol{X})$ es un estadístico suficiente completo y $\widehat{T} = \widehat{T}(S)$ un estimador insesgado de $\tau(\theta)$, entonces es UMVUE de τ.

Demostración: Se sigue de la completez de S y del teorema de Rao–Blackwell. A saber, sea $\widetilde{T}(S)$ cualquier otro estadístico insesgado de τ función de S. Entonces, $\mathbb{E}_\theta[\widehat{T} - \widetilde{T}] \equiv 0$, para todo $\theta \in \Theta$. Por la completez de S, se sigue que $\mathbb{P}_\theta(\widehat{T} - \widetilde{T} = 0) = 1$, para todo $\theta \in \Theta$. Así, hay solamente un estimador insesgado de τ función de S. Sea T cualquier otro estimador insesgado de τ. Entonces, $\widehat{T}$ debe ser igual al $\mathbb{E}[T|S]$ puesto que $E[T|S]$ es un estimador insesgado de τ que depende de S. Se sigue del teorema de Rao–Blackwell que

$$\mathrm{var}_\theta(\widehat{T}) \leq \mathrm{var}_\theta(T), \text{ para todo } \theta \in \Theta$$

Se concluye que $\widehat{T}$ es UMVUE de $\tau(\theta)$.

5.11. Ejercicios

Refiérase a la Lista de Ejercicios 5.

Textos de apoyo.

Casella and Berger (2002); Knight (2000); Mood, Graybill, and Boes (1974); Rice (2007); Wackerly, Mendenhall III, and Scheaffer (2008).

6. Intervalos y Regiones de Confianza

6.1. Intervalos de confianza

Considere una población modelada por $\mathrm{N}(\mu, \sigma^2)$, suponiendo σ dada. Para conocer (*estimar*) el parámetro μ, se toma una muestra aleatoria $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ de $X \sim \mathrm{N}(\mu, \sigma^2)$. Se sabe que en este caso

$$\bar{X}_n = \frac{1}{n}\sum_{1=1}^{n} X_i \sim \mathrm{N}(\mu, \sigma^2/n)$$

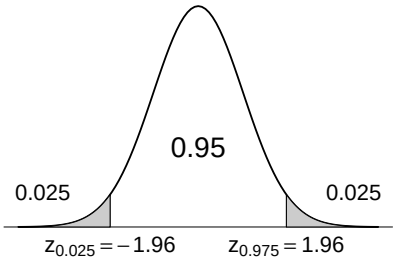
o bien, estandarizando

$$Zn = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

por lo que si z_p denota el p-ésimo cuantil de la distribución normal estándar, se tiene que para $0 < \alpha < 1$ pequeño,

$$P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

que para $\alpha = 0.05$, $1-\alpha = 0.95$, $z_{0.025} = -1.96$ y $z_{0.975} = 1.96$



Luego, se tiene que $\mathbb{P}(-2 \leq Z_n \leq 2) = 0.954$ y

$$\mathbb{P}\left(\bar{X}_n - 2\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + 2\sigma\right) = 0.945$$

Note que los límites del intervalo para μ arriba son: $\mathrm{LI}(\bar{X}_n) = \bar{X}_n - 2\sigma/\sqrt{n}$ y $\mathrm{LS}(\bar{X}_n) = \bar{X}_n + 2\sigma/\sqrt{n}$, límites inferior y superior respectivamente son ambos estadísticos, función de la muestra, y por lo tanto aleatorios. Luego, el intervalo $\big(\mathrm{LI}(\bar{X}_n), \mathrm{LS}(\bar{X}_n)\big)$ es un **intervalo aleatorio** que con 95 % de probabilidad incluye al parámetro μ.

Ahora bien, observada la muestra $\boldsymbol{x}_n = (x_1, \ldots, x_n)$, sea $\bar{x}_n = \frac{1}{n}\sum x_i$, y el intervalo $\big(\mathrm{LI}(\bar{x}_n), \mathrm{LS}(\bar{x}_n)\big)$ es un **intervalo determinístico** que contiene o no al parámetro μ, por lo que decir que tiene una probabilidad aproximada del 95 % es incorrecto. Si se repitiese el experimento varias veces, es decir, la toma de otras muestras del mismo tamaño, se esperaría que $100(1-\alpha)$ % de ellos incluya al parámetro μ. En este sentido, dada la muestra observada $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ se *"confía"* que el intervalo $\big(\mathrm{LI}(\boldsymbol{x}_n), \mathrm{LS}(\boldsymbol{x}_n)\big)$ contenga a μ con una *cobertura de probabilidad* o *nivel de confianza* $(1-\alpha)$ que corresponde a la probabilidad de que el intervalo aleatorio contenga a μ, el parámetro estimado.

Dada la muestra aleatoria $\boldsymbol{x}_n = (x_1, \ldots, x_n)$, $\big(\mathrm{LI}(\boldsymbol{x}_n), \mathrm{LS}(\boldsymbol{x}_n)\big)$ se dice **intervalo de confianza** con un **nivel de confianza** del $(1-\alpha)$, o bien, un **intervalo de $100(1-\alpha)$ % de confianza**.

Así, el proceso de inferencia es *intervalo de confianza*. A saber,

$$\left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{X} + z_{1-\alpha/2}\sigma/\sqrt{n}\right)$$

Note que en este caso, por simetría de distribución normal, el intervalo anterior se podría representar como $(\bar{X} \pm z_{1-\alpha/2}\sigma/\sqrt{n})$.

Si σ no es conocida los límites del intervalo serían desconocidos aunque la probabilidad de cobertura seguiría siendo válida.

Definición : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$, $\theta \in \Theta$ desconocida. Sean $T_1 = T_1(\boldsymbol{X}_n)$, $T_2 = T_2(\boldsymbol{X}_n)$ estadísticos tales que $T_1 < T_2$ para los cuales

$\mathbb{P}\left(T_1 \leq \tau(\theta) \leq T_2\right) = \gamma$, con $0 < \gamma < 1$ y que no depende de θ. El intervalo (T_1, T_2) se dice un intervalo de $100\gamma\,\%$ de confianza para $\tau(\theta)$. γ es el nivel de confianza, T_1 y T_2 los límites de confianza inferior y superior.

Si $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ es la muestra observada y $t_i = T_i(\boldsymbol{x}_n)$, (t_1, t_2) constituye un intervalo de $100\gamma\,\%$ de confianza para $\tau(\theta)$.

En ocasiones interesa nada más uno de los límites del intervalo, por ejemplo, quizás un intervalo de confianza para el parámetro desviación estándar, el límite inferior no resulta de interés por saber que éste no será menor que cero. Luego se podrían tener intervalos de la forma $\left(T_1 \leq \tau(\theta)\right)$, o bien, $(\tau(\theta) \leq T_2)$. En esos casos, T_1 y T_2 se dicen **límites de confianza inferior y superior**, respectivamente.

Nota: Si (T_1, T_2) constituye un intervalo del $100\gamma\,\%$ de confianza para $\tau(\theta)$ y g es una función estrictamente monótona, se puede construir un intervalo de confianza para $g(\tau)$, por ejemplo, si la función g es estrictamente creciente,

$$\mathbb{P}\left(g(T_1) \leq g(\tau) \leq g(T_2)\right) = \gamma$$

6.2. Cantidad pivotal

Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$. $\theta \in \Theta$. Sea $Q = Q(\boldsymbol{X}_n; \theta)$. Si Q sigue una distribución que no depende de θ, Q se dice **cantidad pivotal**.

Ejemplo : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una *m. a.* de $X \sim \mathrm{N}(\mu, 7)$. Luego, $(\bar{X}_n - \mu) \sim \mathrm{N}(0, 7/n)$ no depende de μ. Y

$$Q = \frac{\bar{X}_n - \mu}{\sqrt{7/n}} \sim \mathrm{N}(0,1)$$

Q también es una cantidad pivotal para μ.

Método de construcción de IC mediante cantidades pivotales

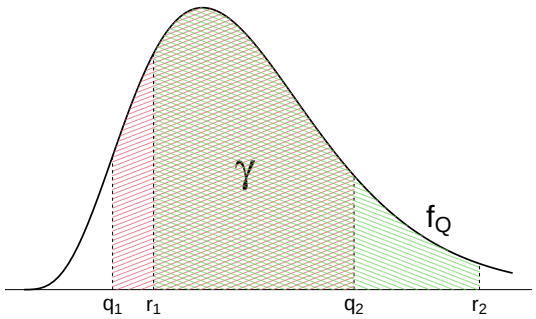
Si $Q = Q(\boldsymbol{X}_n; \theta)$ es una cantidad pivotal con *f. d. p.* f_Q, entonces para γ fijo, existen q_1 y q_2 que dependen de γ tales que $\mathbb{P}\left(q_1 \leq Q \leq q_2\right) = \gamma$.

Ahora, si para todo $\boldsymbol{X}_n$, $q_1 \leq Q(\boldsymbol{X}_n; \theta) \leq q_2$ si y solo si $T_1(\boldsymbol{X}_n) \leq \tau(\theta) \leq T_2(\boldsymbol{X}_n)$, para funciones T_i que no dependen de θ, (T_1, T_2) constituye un intervalo de confianza para $\tau(\theta)$.

Notas:

1. q_1 y q_2 no dependen de θ pues no lo hace f_Q, su *f. d. p.*.

2. Para γ fijo, podría haber un número infinito de parejas q_1, q_2 para los cuales $\mathbb{P}\left(q_1 \leq Q \leq q_2\right) = \gamma$. La siguiente figura ilustra dos casos (q_1, q_2) y (r_1, r_2), de tales parejas.



3. Distintos q_1 y q_2 producirán distintos t_1 y t_2.

4. Un criterio para elegir t_1 y t_2 es que la longitud (media) $t_2 - t_1$ sea mínima.

5. Lo importante en el uso de las cantidades pivotales es que puedan ser *"invertidas"*. Es decir, $\{q_1 \leq Q \leq q_2\}$ si y solo si $\{t_1 \leq \tau(\theta) \leq t_2\}$ para toda $\boldsymbol{x}_n$, de ahí el nombre de *cantidad pivotal*.

6.3. Muestreo de una población normal

6.3.1. Intervalo de confianza para la media conocida la varianza

Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim \mathrm{N}(\theta, 1)$.

$$Q = \frac{\bar{X}_n - \theta}{1/\sqrt{n}} \sim \mathrm{N}(0,1)$$

Q es una cantidad pivotal. Sea $0 < \gamma < 1$, q_1, q_2 tales que

$$\begin{aligned}\gamma &= \mathbb{P}\left(q_1 \leq Q \leq q_2\right)\\ \gamma &= \mathbb{P}\left(\bar{X}_n - q_2/\sqrt{n} \leq \theta \leq \bar{X}_n - q_1/\sqrt{n}\right)\end{aligned}$$

La longitud del intervalo es $\ell = q_2/\sqrt{n} - q_1/\sqrt{n} = \frac{1}{\sqrt{n}}(q_2 - q_1)$ y ℓ es mínimo si $q_2 = -q_1$ con $\gamma = \Phi(q_2) - \Phi(q_1)$.

Más formalmente, se buscan q_1 y q_2 de manera que

$$\min_{q_1, q_2} \quad \tfrac{1}{\sqrt{n}}\left(q_2 - q_1\right) \tag{11}$$

$$\text{s.a.} \quad \textstyle\int_{q_1}^{q_2} \phi(z)dz = \gamma \tag{12}$$

Se sigue de la restricción (12), derivando con respecto a q_1,

$$\phi(q_2)\frac{dq_2}{dq_1} - \phi(q_1) = 0$$

$$\begin{aligned}\frac{dL}{dq_1} &= \frac{1}{\sqrt{n}}\left(\frac{dq_2}{dq_1} - 1\right) = 0\\ \frac{1}{n}\left(\frac{\Phi(q_1)}{\Phi(q_2) - 1}\right) &= 0\end{aligned}$$

Por lo tanto $\Phi(q_1) = \Phi(q_2)$, por lo que $q_1 = q_2$, ó $q_1 = -q_2$.

Ejercicio : Sea $\boldsymbol{X}_n = (X_1, \ldots, X - n)$ una $m.$ a. de $X \sim \mathrm{N}(\theta, \sigma^2)$, σ conocida. Verifique que

$$Q = \frac{\bar{X}_n - \theta}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

es una cantidad pivotal tal que

$$\begin{aligned}\gamma &= \mathbb{P}\left(q_1 \leq Q \leq q_2\right)\\ &= \mathbb{P}\left(\bar{X}_n - q_2\sigma/\sqrt{n} \leq \theta \leq \bar{X}_n + q_2\sigma/\sqrt{n}\right)\end{aligned}$$

donde $q_1 = -q_2$ y $q_2 = z_{1-\alpha/2}$, con $\gamma = 1 - \alpha$. Así,

$$\left(\bar{X}_n - z_{1-\alpha/2}\sigma/\sqrt{n},\ \bar{X}_n + z_{1-\alpha/2}\sigma/\sqrt{n}\right)$$

El intervalo anterior constituye un intervalo de confianza para la media, conocida la varianza.

6.3.2. Intervalo de confianza para la media desconocida la varianza

Considere nuevamente $\boldsymbol{X}_n$ una muestra aleatoria de $X \sim \mathrm{N}(\theta, \sigma^2)$, pero ahora σ no es conocida. De cualquier modo se tiene que

$$\frac{\bar{X}_n - \theta}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

pero no sirve como cantidad pivotal pues depende de σ desconocida. Luego, $\left\{ q_1 \leq \dfrac{\bar{X}_n - \theta}{\sigma/\sqrt{n}} \leq q_2 \right\}$ no sería posible *"invertirse"*. Por otro lado, recuerde que

$$Q = \frac{\frac{\bar{X}_n - \theta}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2/(n-1)}} = \frac{\bar{X}_n - \theta}{S/\sqrt{n}} \sim t_{n-1}$$

donde $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ es la varianza muestral.

Entonces, $\gamma = \mathbb{P}(t_1 \leq Q \leq t_2)$. Nuevamente, por simetría de la densidad de la distribución t, el intervalo (t_1, t_2) es de longitud mínima si $t_1 = -t_2$. Luego,

$$\begin{aligned}
\gamma &= \mathbb{P}(t_1 \leq Q \leq t_2) \\
&= \mathbb{P}\left(\bar{X}_n - t_2 S/\sqrt{n} \leq \theta \leq \bar{X}_n - t_1 S/\sqrt{n}\right) \\
&= \mathbb{P}\left(\bar{X}_n - t(1-\alpha/2; n-1) S/\sqrt{n} \leq \theta \leq \bar{X}_n + t(1-\alpha/2; n-1) S/\sqrt{n}\right)
\end{aligned}$$

donde $t(p;\nu)$ representa el p-ésimo quantil de la distribución *t-Student* con ν grados de libertad. Por lo que

$$\left(\bar{X}_n - t_{1-\alpha/2;n-1} S/\sqrt{n},\ \bar{X}_n + t_{1-\alpha/2;n-1} S/\sqrt{n}\right)$$

constituye un intervalo de confianza para la media de la distribución normal, desconocida la varianza.

6.3.3. Intervalo de confianza para la varianza

Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim \mathrm{N}(\mu, \sigma^2)$. Sea $\boldsymbol{\theta} = (\mu, \sigma^2)$ desconocida. Considere $Q = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2$. Recordar $Q = \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$, donde $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ es la varianza muestral. Luego, Q es una cantidad pivotal y

$$\left\{ q_1 \leq \frac{n-1}{\sigma^2} S^2 \leq q_2 \right\} \iff \left\{ \frac{1}{q_2} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{q_1} \right\}$$

por lo que

$$\mathbb{P}\left(\frac{(n-1)S^2}{q_2} \leq \sigma^2 \leq \frac{(n-1)S^2}{q_1} \right) = \gamma$$

y donde q_i's representarían cuantiles apropiados de la distribución χ^2_{n-1}.

A diferencia de los casos anteriores, encontrar el intervalo de longitud mínima no tiene solución analítica y habría que encontrarlo por medio de algún método numérico.

Una manera empírica de proceder es elegir las colas del intervalo inicial para Q del mismo peso. Por ejemplo, si se desea un intervalo de $100\gamma\,\%$ de confianza. Las colas serían de probabilidad $(1-\gamma)/2$ cada una. Si $\gamma = 1 - \alpha$, la colas serían de probabilidad $\alpha/2$. Refierase a la figura 14

Entonces, el intervalo de nivel γ de confianza para la varianza σ^2 estaría dado por

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2;\ n-1}},\ \frac{(n-1)S^2}{\chi^2_{\alpha/2;\ n-1}} \right)$$

donde $\chi^2_{p;\nu}$ representa el p-ésimo cuantil de la distribución χ^2 con ν grados de libertad.

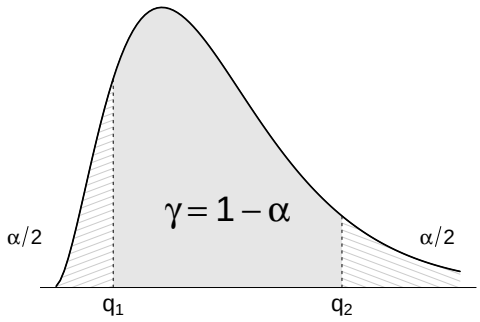


Figura 14: Intervalo (q_1, q_2) de $100\gamma\,\%$ de confianza para la cantidad pivotal Q con colas del mismo peso de probabilidad.

6.4. Comparación de poblaciones normales por intervalos de confianza

Suponga que se tienen dos poblaciones modeladas por $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ y $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ y se desea compararlas a partir de sus medias y varianzas. Se consideraran los casos cuando las poblaciones son o no independientes.

6.4.1. Comparación de medias, varianzas conocidas.

Considere X y Y poblaciones independientes y sean $\boldsymbol{X}_{n_1} = (X_1, \ldots, X_{n_1})$ una muestra aleatoria de $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ y $\boldsymbol{Y}_{n_2} = (Y_1, \ldots, Y_{n_2})$ una muestra aleatoria de $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$.

Sea $\theta = \mu_1 - \mu_2$. Se desea construir un intervalo de confianza para θ. Note que si el intervalo incluye al cero, éste sería un valor posible para θ, lo que implicaría en el proceso de inferencia que $\mu_1 = \mu_2$.

Entonces, $\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \sim \mathrm{N}(\mu_1, \sigma_1^2/n_1)$ y $\bar{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j \sim \mathrm{N}(\mu_2, \sigma_2^2/n_2)$ independientes por lo que si $\theta = \mu_1 - \mu_2$,

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

opera como cantidad pivotal y siguiendo el procedimiento de antes, un intervalo de $100\gamma\,\%$ de confianza para la diferencia de medias θ es

$$(\bar{X} - \bar{Y}) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

con $\gamma = 1 - \alpha$ y $z_{1-\alpha/2}$ el cuantil $1 - \alpha/2$ de la distribución normal estándar.

6.4.2. Comparación de medias, varianzas iguales desconocidas.

En este caso, $X \sim \mathrm{N}(\mu_1, \sigma^2)$ y $Y \sim \mathrm{N}(\mu_2, \sigma^2)$ independientes. Recuerde que $(n_i - 1)S_i^2/\sigma^2 \sim \chi^2_{n_i-1}$, donde S_1^2 y S_2^2 son las varianzas muestrales para X y Y, independientes. Luego,

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

es un estimador insesgado de la varianza común σ^2 con

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{1}{\sigma^2}\left[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right] \sim \chi^2_{n_1+n_2-2}$$

e independiente de $\bar{X}$ y de $\bar{Y}$. Entonces, si $\theta = \mu_1 - \mu_2$,

$$\frac{(\bar{X} - \bar{Y}) - \theta}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathrm{N}(0,1)$$

Se sigue de la independencia del numerador y denominador que

$$Q = \frac{(\bar{X} - \bar{Y}) - \theta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\frac{(\bar{X} - \bar{Y}) - \theta}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2}/(n_1 + n_2 - 2)}} \sim t_{n_1+n_2-2}$$

Por lo que Q es una cantidad pivotal y $p = 1 - \alpha/2$ y $\nu = n_1 + n_2 - 2$

$$(\bar{X} - \bar{Y}) \pm t_{p;\nu}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

constituye un intervalo de $100(1-\alpha)\,\%$ de confianza para $\theta = \mu_1 - \mu_2$. Nuevamente, si el cero está incluido en el intervalo anterior, se concluye que *no hay evidencia* en los datos de que las medias sean distintas.

Ejercicio : Considere X y Y poblaciones independientes. Sean $\boldsymbol{X} = (X_1, \ldots, X_m)$ y $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ muestras aleatorias de $X \sim \mathrm{N}(\mu_1, \sigma^2)$ y $Y \sim \mathrm{N}(\mu_2, k\sigma^2)$, respectivamente con la constante $k > 0$ conocida y σ desconocida. Construya un intervalo del $100(1-\alpha)\,\%$ de confianza para la diferencia de medias $\theta = \mu_1 - \mu_2$.

6.4.3. Comparación de medias con varianzas desconocidas

Considere nuevamente las poblaciones $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ y $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$ independientes y el problema de construir un intervalo de confianza para la diferencia de medias $\theta = \mu_1 - \mu_2$, pero con ambas varianzas desconocidas.

El problema, conocido como de **Behrens–Fisher** es que no se conoce el parámetro $\rho = \sigma_1/\sigma_2$ (vea el ejercicio anterior), por lo que se desconoce la distribución exacta de $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$.

Hay varias aproximaciones a la solución práctica del problema. A saber, por razonamientos fiduciarios, bayesianos y frecuentistas. En estas notas presentamos la solución frecuentista debida a Welch y Satterthwaite.

Sean $\boldsymbol{X}_{n_1} = (X_1, \ldots, X_{n1})$ y $\boldsymbol{Y}_{n_2} = (Y_1, \ldots, Y_{n_2})$ muestras aleatorias de $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ y $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$ independientes. Considere el estadístico $D = \bar{X} - \bar{Y} = \widehat{\mu}_D$. Entonces, $\mathbb{E}[D] = \theta = \mu_1 - \mu_2$ y $\mathrm{var}(D) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma_D^2$. Luego, $\frac{D - \mu_D}{\sigma_D} \sim \mathrm{N}(0,1)$, pero ¿cómo estimar σ_D?

i) Si n_1 y n_2 son suficientemente grandes, se sigue del Teorema Central de Límite y el Teorema de Slutsky que

$$\frac{D - \mu_D}{S_D} \dot{\sim} \mathrm{N}(0,1)$$

donde $S_D^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$.

- Si n_1 y n_2 son pequeños, se puede usar la **aproximación de Welch–Satterthwaite** basada en la corrección del método de momentos. A saber, si $Y_1, \ldots, Y_k$ son *v.* a.'s

independientes con $Y_i \sim \chi^2_{\nu_i}$ y $a_1, \ldots, a_k$ constantes conocidas, entonces

$$\sum a_i Y_i \,\dot\sim\, \frac{1}{\nu}\chi^2_\nu, \qquad \text{donde } \nu = \frac{\left(\sum a_i S_i^2\right)^2}{\sum \frac{a_i^2}{\nu_i} S_i^2}$$

Sea entonces

$$Q = \frac{D-\theta}{S_D} \,\dot\sim\, t_{\nu^*}$$

con

$$S_D^2 = S_1^2/n_1 + S_2^2/n_2 \quad \text{y} \quad \nu^* = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$$

Finalmente,

$$D \pm t(1-\alpha/2;\nu^*)S_D$$

constituye un intervalo de confianza (aproximada) de $100(1-\alpha)\,\%$ para $\theta = \mu_1 - \mu_2$.

6.4.4. Comparación de varianzas.

Sean $\boldsymbol{X}_{n_1} = (X_1, \ldots, X_{n_1})$ y $\boldsymbol{Y}_{n_2} = (Y_1, \ldots, Y_{n_2})$ muestras aleatorias de poblaciones $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ y $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$ independientes, respectivamente. Por construir un intervalo de confianza para $\theta = \sigma_1^2/\sigma_2^2$.

Sean $S_1^2 = \sum_{i=1}^{n_1}(X_i - \bar X)^2/(n_1-1)$ y $S_2^2 = \sum_{j=1}^{n_2}(Y_j - \bar Y)^2/(n_2-1)$. Luego, $\dfrac{n_1-1}{\sigma_1^2}S_1^2 \sim \chi^2_{n_1-1}$ y $\dfrac{n_2-1}{\sigma_2^2}S_2^2 \sim \chi^2_{n_2-1}$ independientes por lo que

$$Q = \frac{\frac{n_1-1}{\sigma_1^2}S_1^2/(n_1-1)}{\frac{n_2-1}{\sigma_2^2}S_2^2/(n_2-1)} = \frac{1}{\theta}\frac{S_1^2}{S_2^2} \sim F_{\nu_1;\nu_2}$$

con $\nu_1 = n_1 - 1$ y $\nu_2 = n_2 - 1$. Se sigue entonces que para q_1 y q_2 se tiene que

$$\gamma = \mathbb{P}\left(q_1 \le Q \le q_2\right) = \mathbb{P}\left(\frac{S_1^2/S_2^2}{q_2} \le \theta \le \frac{S_1^2/S_2^2}{q_1}\right) = 1-\alpha$$

Finalmente,

$$\left(\frac{S_1^2/S_2^2}{F(1-\alpha/2;n_1-1,n_2-1)},\ \frac{S_1^2/S_2^2}{F(\alpha/2;n_1-1,n_2-1)}\right)$$

constituye un intervalos de $100(1-\alpha)\,\%$ de confianza para $\theta = \sigma_1^2/\sigma_2^2$.

6.4.5. Observaciones pareadas.

Considere ahora $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ una muestra aleatoria de $\boldsymbol{X} \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con

$$\boldsymbol{X}_i = \begin{bmatrix} X_{1i} \\ X_{2i} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{y} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Se desea construir un intervalo de confianza para $\theta = \mu_1 - \mu_2$.

Para $i \neq j$, las observaciones $\boldsymbol{X}_i$ y $\boldsymbol{X}_j$ son independientes, pero sus componentes X_{1i} y X_{2i} no lo son salvo que $\sigma_{12} = 0$. Considere entonces $D_i = X_{1i} - X_{2i}$, $i = 1, \ldots, n$, la muestra aleatoria de $D \sim \mathrm{N}_1(\mu_D, \sigma_D^2)$, con $\mu_D = \mu_1 - \mu_2$ y $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$. Luego, se

puede construir un intervalo de confianza para $\theta = \mu_D$, considerando σ_D desconocida como se hizo anteriormente. Entonces,

$$\left(\bar{D} - t_{1-\alpha/2;n-1} S_D/\sqrt{n},\ \bar{D} + t_{1-\alpha/2;n-1} S_D/\sqrt{n}\right)$$

donde $\bar{D} = \frac{1}{n}\sum_{i=1}^n D_i$, $S_D^2 = \frac{1}{n-1}\sum_{i=1}^n (D_i - \bar{D})^2$

Note que si las componentes X_1 y X_2 están asociadas positivamente ($\sigma_{12} > 0$) la varianza de la diferencia $D = X_1 - X_2$ es menor que si las componentes fuesen independientes, lo que resultará en intervalos de confianza más estrechos, preferibles en la práctica. Pensar por ejemplo, el muestreo de manera de tener componentes asociadas positivamente es área del **diseño de experimentos estadístico**.

6.5. Poblaciones no normales

6.5.1. Intervalos de confianza para proporciones.

La aproximación a la distribución binomial por la distribución normal y que se muestra en el curso de Cálculo de Probabilidades I se justifica por la distribución de la suma de *v.a.i.i.d.*'s y el Teorema Central del Límite. De manera relacionada se construye un intervalo de confianza para proporciones o probabilidades de *éxito* de una distribución Bernoulli (binomial)

Considere $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de una población $X \sim \text{Ber}(p)$ y sea $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Por lo que, del teorema central del límite $\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{D} Z \sim \text{N}(0,1)$. Por otro lado, en la sección de teoremas límite se vio que $\bar{X}_n \xrightarrow{P} p$. Entonces, para n suficientemente grande se sigue del TCL y del teorema de Slutsky que $\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)/n}} \dot{\sim} N(0,1)$. Así,

$$\bar{X}_n \pm z_{1-\alpha/2}\sqrt{\bar{X}_n(1-\bar{X})_n)/n}$$

constituye un intervalo de confianza de nivel $(1-\alpha)$ para el parámetro p, la probabilidad de éxito o proporción de una población Bernoulli.

Considere ahora que se tiene una segunda población $Y \sim \text{Ber}(p_2)$, independiente de $X \sim \text{Ber}(p_1)$ y sea $\boldsymbol{Y}_m = (Y_1, \ldots, Y_m)$ una muestra aleatoria de ella. Se desea construir un intervalo de confianza para el parámetro $\theta = p_1 - p_2$, diferencia de proporciones. Entonces, como antes, $\frac{\bar{Y}_m - p_2}{\sqrt{\bar{Y}_m(1-\bar{Y}_m)/m}} \dot{\sim} N(0,1)$ y para m y n grandes se tiene

$$\frac{\left(\bar{X}_n - \bar{Y}_m\right) - (p_1 - p_2)}{\sqrt{\bar{X}_n(1-\bar{X}_n)/n + \bar{Y}_m(1-\bar{Y}_m)/m}} \dot{\sim} N(0,1)$$

por lo que

$$\left(\bar{X}_n - \bar{Y}_m\right) \pm z_{1-\alpha/2}\sqrt{\bar{X}_n(1-\bar{X}_n)/n + \bar{Y}_m(1-\bar{Y}_m)/m}$$

constituye un intervalo de confianza de nivel $100(1-\alpha)\,\%$ para la diferencia de proporciones $\theta = p_1 - p_2$.

6.5.2. Intervalos de confianza para la media

Considere $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de una población X con media $\mu = \mathbb{E}[X]$ y varianza $\sigma^2 = \text{var}(X)$. Sean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ y $S_n^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X}_n)^2$. Entonces, se ha visto ya que $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim \text{N}(0,1)$ y $\frac{S_n^2}{\sigma_2} \xrightarrow{P} 1$. Por lo que nuevamente por

el teorema de Slutsky se tiene que

$$\frac{\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S_n^2}{\sigma_2}}} = \frac{\bar{X}_n-\mu}{S_n/\sqrt{n}} \xrightarrow{D} \mathrm{N}(0,1)$$

Luego,

$$\bar{X}_n \pm z_{1-\alpha/2} S_n/\sqrt{n}$$

constituye un intervalo de confianza de nivel $(1-\alpha)$ aproximado para la media μ.

6.5.3. Intervalos de confianza por medio de EMV

Considere $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de una población X y θ parámetro de la distribución de X. Sea $\hat{\theta}_n = \hat{\theta}(\boldsymbol{X}_n)$ el estimador por máxima verosimilitud (EMV). Luego, por la distribución asintótica de $\hat{\theta}_n$

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\mathrm{var}(\hat{\theta}_n)}} = \frac{\hat{\theta}_n - \theta}{\sqrt{1/n\mathrm{I}(\theta)}} \dot{\sim} \mathrm{N}(0,1)$$

Por lo que un intervalo de nivel aproximado de confianza $(1-\alpha)$ estaría dado por

$$\hat{\theta} \pm z_{1-\alpha/2}/\sqrt{n\mathrm{I}(\theta)}$$

Ejemplo : Considere $\boldsymbol{X} = (X_1, \ldots, X_n)$ una *m. a.* de $X \sim \mathrm{Po}(\lambda)$. Entonces, $f(x;\lambda) = \lambda^x e^{-\lambda}/x!\, \mathbb{1}_{\mathbb{N}0}(x)$. Se ha visto que $\bar{X} = \hat{\lambda}_{\mathrm{EMV}}$ y

$$\begin{aligned} L(\lambda; x) &= \lambda^x e^{-\lambda}/x! \\ \ell(\lambda; x) &= x\log\lambda - \lambda - \log x! \\ \frac{\partial \ell}{\partial \lambda} &= -\frac{x}{\lambda} - 1 \\ \frac{\partial^2 \ell}{\partial \lambda^2} &= \frac{x}{\lambda^2} \end{aligned}$$

Por lo que

$$-\mathbb{E}\left[\frac{\partial^2 \ell(\lambda; X)}{\partial \lambda^2}\right] = -\mathbb{E}\left[-\frac{X}{\lambda^2}\right] = \frac{1}{\lambda} = \mathrm{I}(\lambda)$$

Así, un intervalo de confianza aproximado de $100(1-\alpha)\,\%$ para λ sería

$$\hat{\lambda} \pm z_{1-\alpha/2}/\sqrt{n/\hat{\lambda}}, \quad \text{o bien,} \quad \bar{X} \pm z_{1-\alpha/2}\sqrt{\bar{X}/n}$$

6.6. Regiones de confianza

6.6.1. Población normal

Con el apoyo de Mood, Graybill, and Boes (1974).

Considere $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de una población $X \sim \mathrm{N}(\mu, \sigma^2)$. Se han construido ya los intervalos marginales I_1 para μ

$$\left(\bar{X} - t_{1-\alpha_1/2;\, n-1} S/\sqrt{n},\ \bar{X} + t_{1-\alpha_1/2;\, n-1} S/\sqrt{n}\right) \equiv (\mu_{\mathrm{inf}},\ \mu_{\mathrm{sup}})$$

de $(1-\alpha_1)$ de confianza y I_2 para σ^2

$$\left((n-1)S^2/\chi^2_{1-\alpha_2/2;\, n-1},\ (n-1)S^2/\chi^2_{\alpha_2/2;\, n-1}\right) \equiv (\sigma^2_{\mathrm{inf}},\ \sigma^2_{\mathrm{sup}})$$

de $(1-\alpha_2)$ de confianza. Sin embargo, la región de confianza $\mathcal{R} = I_1 \times I_2$ para el vector de parámetros $\boldsymbol{\theta} = (\mu, \sigma^2)$ es menor a la confianza $\gamma = (1-\alpha_1)(1-\alpha_2)$ pues no son independientes. Vea el panel de la izquierda de la figura 15.

Por otro lado, si se consideran las cantidades pivotales

$$Q_1 = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{y} \quad Q_2 = \frac{(n-1)S^2}{\sigma^2}$$

que sí son independientes, se tiene que

$$1-\alpha_1 = \mathbb{P}\left(\ell_1 \leq Q_1 \leq u_1\right) \quad \text{y} \quad 1-\alpha_2 = \mathbb{P}\left(\ell_2 \leq Q_2 \leq u_2\right)$$

por lo que la región

$$\mathcal{R} = \left\{ \theta = (\mu, \sigma^2) \middle| \ell_1 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq u_1,\ \ell_2 \leq \frac{(n-1)S^2}{\sigma^2} \leq u_2 \right\}$$

que se muestra en el panel de la derecha de la figura 15 tiene una confianza conjunta de $\gamma = (1-\alpha_1)(1-\alpha_2)$.

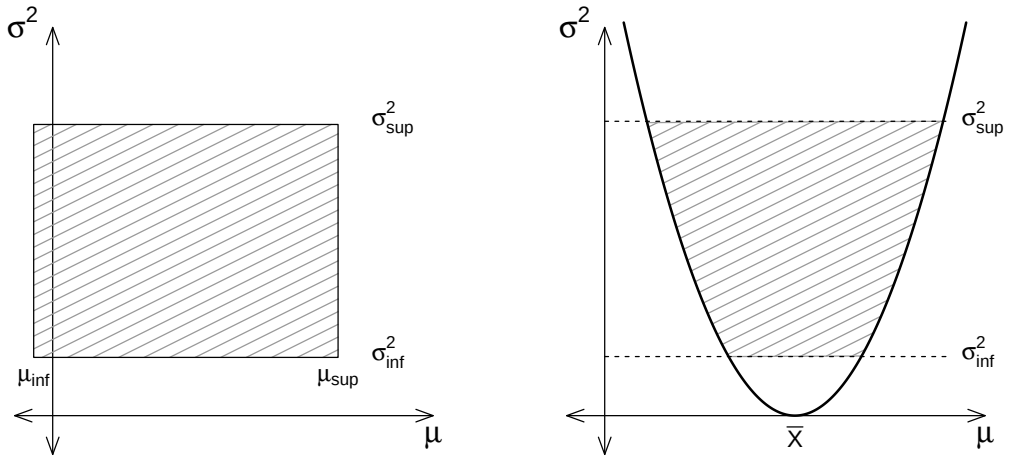


Figura 15: Regiones de confianza $\mathcal{R}$. El panel de la izquierda muestra el producto $I_1 \times I_2$ de los intervalos de confianza *marginales* de μ y σ^2. El panel de la derecha muestra la región de confianza *conjunta* para $\boldsymbol{\theta} = (\mu, \sigma^2)$.

6.6.2. Intervalos de Bonferroni

Los intervalos de Bonferroni ofrecen una manera práctica de construir intervalos de confianza marginales de manera que la confianza conjunta sea al menos la deseada.

Sea $\boldsymbol{\theta} = (\theta_1, \theta_2)$ un vector de parámetros y C_i un intervalo de $(1-\alpha_i)$ de confianza para θ_i, con $i = 1, 2$. Sean los eventos $E_i = \{\theta_i \in C_i\}$, luego $1-\alpha_1 = \mathbb{P}(E_i)$ y $E_i^C = \{\theta_i \notin C_i\}$. Así,

$$\begin{aligned}
\mathbb{P}(E_1 \cap E_2) &= \mathbb{P}\left(\left(E_1^C \cup E_2^C\right)^C\right) \\
&= 1 - \left[\mathbb{P}\left(E_1^C\right) + \mathbb{P}\left(E_2^C\right) - \mathbb{P}\left(E_1^C \cap E_2^C\right)\right] \\
&\geq 1 - (\alpha_1 + \alpha_2)
\end{aligned}$$

Entonces, si para cada uno de los θ_i se construye un intervalo de $(1-\alpha/2)$ de confianza, la confianza conjunta es de al menos

$$\mathbb{P}(E_1 \cap E_2) \geq 1 - (\alpha/2 + \alpha/2) = 1 - \alpha$$

Esta situación se ilustra en la figura 16.

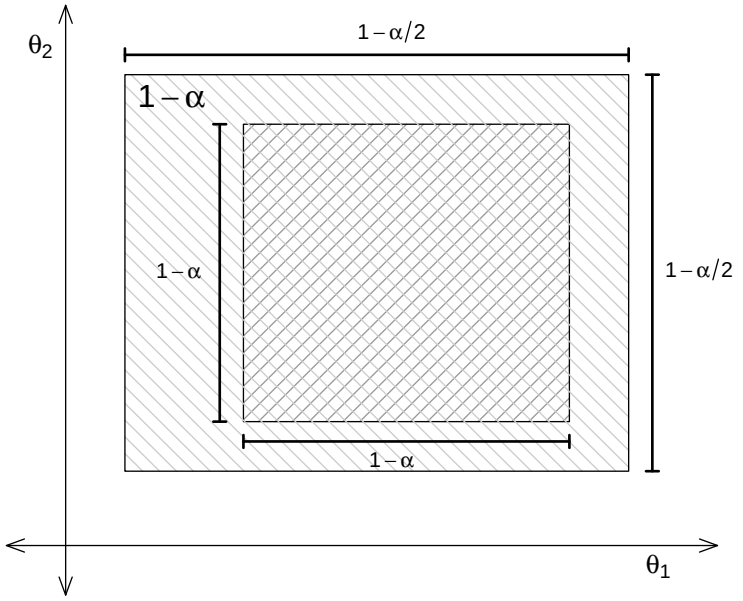


Figura 16: Regiones de confianza. El panel de la izquierda muestra el producto cruz de los intervalos de confianza *marginales* de μ y σ^2. El panel de la derecha muestra la región de confianza *conjunta* para (μ, σ^2).

6.7. Ejercicios

Refiérase a la Lista de Ejercicios 6.

Textos de apoyo.

Casella and Berger (2002); Knight (2000); Mood, Graybill, and Boes (1974); Rice (2007); Wackerly, Mendenhall III, and Scheaffer (2008).

7. Contraste de Hipótesis

7.1. Introducción

El siguiente ejemplo ha sido tomado de Rice (2007). Con él se introducen varios elementos de las pruebas o contraste de hipótesis.

Considere dos monedas m_0 y m_1 con correspondientes probabilidad de águila p_0 y p_1, respectivamente. Se elije una de las dos monedas y se lanza 10 veces. Con base en el número observado de águilas usted decide cuál de las monedas fue lanzado. Sea X el número observadas de águilas en los 10 lanzamientos. Entonces, $X \sim \text{Bin}(n, p)$ con $n = 10$. Suponga que $p_0 = 0.5$ y $p_1 = 0.7$.

Considere ahora x, el número observado de águilas. Entonces, la correspondiente verosimilitud de la moneda m_i es $L(p_i; x) = f(x; p_i)$, $i = 0, 1$ y el cociente de verosimilitudes (CV), también conocido como razón de verosimilitud (RV) se define como

$$\text{CV}(x) = \frac{L(p_0, x)}{L(p_1; x)} = \frac{f(x; p_0)}{f(x; p_1)}$$

Por ejemplo, si $\mathbb{P}_i(x) = \mathbb{P}(X = x; p = p_i) = \binom{10}{x} p_i^x (1 - p_i)^{10-x}$, entonces $\mathbb{P}_0(2) = 0.0439$, $\mathbb{P}_1(2) = 0.0014$ y $\text{CV}(2) = 30.2762$. Luego si $\text{CV}(2) = 30.38$ resulta razonable interpretarlo que favorece a la moneda m_0 mientras que $\text{CV}(7) = 0.44$ favorecería a m_1. La tabla 2 muestra las probabilidades $\mathbb{P}(X = x; p_i)$ y el correspondiente cociente de verosimilitudes. Éste último se presenta en la figura 17.

Tabla 2: Probabilidades del número de águilas en los 10 lanzamientos para las dos monedas y el correspondiente cociente de verosimilitudes CV.

x	0	1	2	3	4	5	6	7	8	9	10
m_0	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010
m_1	0.0000	0.0001	0.0014	0.0090	0.0368	0.1029	0.2001	0.2668	0.2335	0.1211	0.0282
CV	165.3817	70.8779	30.3762	13.0184	5.5793	2.3911	1.0248	0.4392	0.1882	0.0807	0.0346

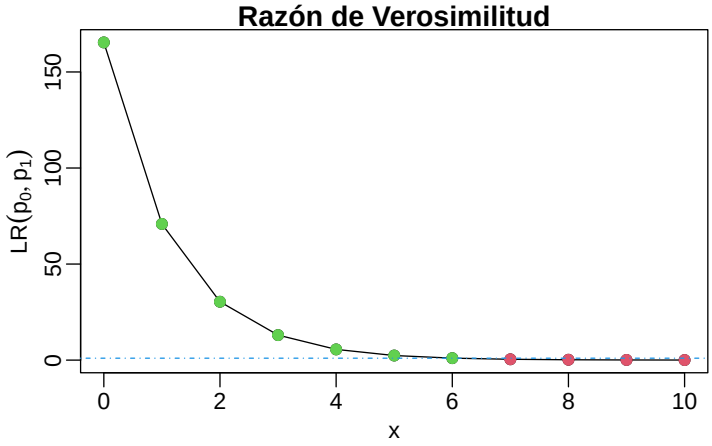


Figura 17: Ejemplo de las monedas. Cociente de verosimilitudes.

Sea H_i la hipótesis que supone que la moneda m_i fue lanzada. Para este ejemplo se ilustran elementos con un "enfoque bayesiano".

- Se tiene información previa (*a priori*) y no hay razón para suponer que las monedas

son distintas y deshonestas

$$\pi_0 = \mathbb{P}(\mathrm{H}_0) = \mathbb{P}(\mathrm{H}_1) = \pi_1 = \frac{1}{2}$$

- Después de observar el número de águilas X en los 10 lanzamientos, las probabilidades posteriores (*posteriori*)
$$\mathbb{P}(\mathrm{H}_i|X) = \frac{\mathbb{P}(X|\mathrm{H}_i)\mathbb{P}(\mathrm{H}_i)}{\mathbb{P}(X)}$$

- Luego, el cociente de verosimilitud CV,
$$\frac{\mathbb{P}(\mathrm{H}_0|X)}{\mathbb{P}(\mathrm{H}_1|X)} = \frac{\mathbb{P}(\mathrm{H}_0)}{\mathbb{P}(\mathrm{H}_1)}\,\frac{\mathbb{P}(X|\mathrm{H}_0)}{\mathbb{P}(X|\mathrm{H}_1)} = \frac{\pi_0}{\pi_1}\,\mathrm{CV}(X)$$
donde los π_i's son las probabilidades iniciales y CV denota el cociente de verosimilitudes.

- En el ejemplo, CV es una función monótona en X. Si las probabilidades iniciales son iguales $X \leq 6$ favorece a H_0 y $X \geq 7$ favorece a H_1.

- Si se ha de decidir entre H_0 y H_1 con base en lo observado, sería razonable preferir aquella hipótesis con probabilidad posterior $\mathbb{P}(\mathrm{H}_i|X)$ mayor. Es decir, favorece a H_0 si $\dfrac{\mathbb{P}(\mathrm{H}_0|X)}{\mathbb{P}(\mathrm{H}_1|X)} \geq c$, donde el valor crítico c depende de las probabilidades iniciales y el cociente de verosimilitudes.

- Suponga que $c = 1$, entonces sea acepta H_0 si $X \leq 6$ y se rechaza H_0 si $X \geq 7$. H_0 se dice la hipótesis nula y generalmente representa el *status quo*, en este ejemplo $\pi_0 = \pi_1 = 1/2$. H_1 se dice la hipótesis alternativa que muchas veces se interesa mostrar su validez a partir de los datos.

- En esta situación se pueden cometer errores de dos tipos:

  I : Rechazar H_0 cuando ésta es verdadera ($p = \pi_0$).

  II : Aceptar H_0 cuando ésta es falsa ($p = \pi_1$).

- Las correspondientes probabilidades de error:

$$\alpha = \mathbb{P}(\text{Error tipo I}) = \mathbb{P}(\text{Rechazar } \mathrm{H}_0|\mathrm{H}_0 \text{ verdadera}) = \mathbb{P}_0(X \geq 7) = 0.1719$$

$$\beta = \mathbb{P}(\text{Error tipo II}) = \mathbb{P}(\text{Aceptar } \mathrm{H}_0|\mathrm{H}_0 \text{ falsa}) = \mathbb{P}_0(X \leq 6) = 0.3504$$

- Los contrastes o pruebas de hipótesis estadísticas son procedimientos formales para distinguir entre distribuciones.

7.2. Definiciones

Ejemplo : Por ser una temporada de escasez de agua se quiere reducir el consumo en cierta comunidad y para este fin se dio una campaña de carteles a lo largo de 4 semanas. Se cree que con ella haya disminuido en promedio de 3000 a 2500 litros por mes por persona. Se desea verificar la efectividad de la campaña.

Suponga que el consumo mensual de agua por persona es razonablemente modelado por la distribución normal. Luego, sea X la variable aleatoria (*v. a.*) que denota el consumo

de agua por persona con $X \sim \mathrm{N}(\mu, \sigma^2)$, y en este ejemplo considere que la varianza σ^2 es conocida. Entonces, se desea contrastar las hipótesis

$$\mathrm{H}_0 : \mu = 3 \text{ m}^3/\text{mes} \quad vs. \quad \mathrm{H}_1 : \mu = 2.5 \text{ m}^3/\text{mes}$$

o bien: $\mathrm{H}_0 : \mu = \mu_0 = 3.0$ *vs.* $\mathrm{H}_1 : \mu = \mu_1 = 2.5$. Las hipótesis definen completamente las distribuciones involucradas, $X \sim \mathrm{N}(3.0, \sigma^2)$ y $X \sim \mathrm{N}(2.5, \sigma^2)$, pues en el ejemplo se supone σ conocida. Cuando las hipótesis definen completamente la distribución se dicen **hipótesis simples**, en caso contrario, por ejemplo, $\mathrm{H}_0 : \mu < 3$, $X \sim \mathrm{N}(\mu_i, \sigma^2)$, pero σ desconocida, etcétera, se dicen **hipótesis compuestas**.

Para decidir sobre la veracidad de las hipótesis de manera objetiva se toma una muestra aleatoria (*m. a.*) $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, para una n fija y, en este ejemplo, se decide rechazar la hipótesis H_0 si para cierta c, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \leq c$ lo que define la región de rechazo $\mathcal{R} = \{\boldsymbol{X}_n : \bar{X} \leq c\}$. Note que $\mathcal{R} = \{\boldsymbol{X}_n : \bar{X} \leq c\} \subseteq \{\boldsymbol{X}_n : \boldsymbol{X}_n = (X_1, \ldots, X_n)\} = \mathcal{X}$, el espacio muestral.

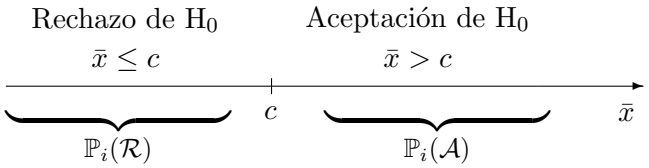
Para tener una idea de la sensibilidad del estadístico $\bar{X}$ para detectar desviaciones, la tabla 3 muestra las probabilidades $\mathbb{P}(\bar{X}_n \leq x)$, para distintos valores de x y tamaños de muestra n. Por ejemplo, si $X \sim \mathrm{N}(3.0, 1)$, aproximadamente 31 de cada 100 observaciones caerían a la izquierda de 2.5, pero del promedio de 10 observaciones, solo en 6 ocasiones caerían a la izquierda y si la muestra fuese de 25 observaciones, solamente 0.6 % de los casos carían a la izquierda.

Tabla 3: Probabilidades acumuladas por la izquierda de x del promedio $\bar{X}_n \sim \mathrm{N}(3.0, 1/\sqrt{n})$, para $n = 1, 5, 10, 25, 50$.

	$\mathbb{P}(\bar{X}_n \leq x)$				
x	1	5	10	25	50
3.0	0.500	0.50000	0.50000	0.50000	0.50000
2.8	0.421	0.32736	0.26354	0.15866	0.07865
2.5	**0.309**	**0.13178**	**0.05692**	**0.00621**	**0.00020**
2.3	0.242	0.05876	0.01343	0.00023	0.00000
2.0	0.159	0.01267	0.00078	0.00000	0.00000

En el resto de la nota se evitará el uso del subíndice n para la muestra $\boldsymbol{X}$ o la media $\bar{X}$, salvo que se quiera enfatizar la dependencia con el tamaño de muestra.

Así, en este ejemplo se tiene el **estadístico de prueba** $\bar{X}$ y el **valor crítico** c que define la **región de rechazo** $\mathcal{R} = \{\boldsymbol{X} \in \mathcal{X} : \bar{X} \leq c\}$ o **regla de decisión** para rechazar la **hipótesis nula** H_0 en favor de la **hipótesis alternativa** H_1. Las definiciones anteriores quedan ilustradas en la siguiente figura.



donde $\mathbb{P}_i(\cdot)$, denota la probabilidad de rechazar ó aceptar H_0 suponiendo que la hipótesis H_i es verdadera.

Aunque no es regla, se acostumbra establecer en la hipótesis nula lo común, la situación actual o el *status quo* para que sea la muestra, la evidencia en los datos lo que lleve a su rechazo en favor de la hipótesis alternativa H_1. Como el rechazo de H_0 depende del valor del estadístico de prueba que a su vez depende de la muestra observada, es posible, que aún bajo las mismas condiciones en ocasiones se rechace la hipótesis nula pero otras veces no, lo que no implica que se demuestre la validez de H_0. Es decir, que no se rechace la

hipótesis no implica haber demostrado su veracidad. La inferencia estadística no sigue las reglas de la lógica formal o matemática. Así, no es costumbre "aceptar la hipótesis nula", que se entendería como aceptar su validez, en todo caso "no se rechaza H_0".

Entonces, como la decisión de rechazar o no la hipótesis nula H_0 depende del estadístico de prueba y que a su vez depende de la muestra, es posible cometer errores: rechazar H_0 cuando ésta es verdadera, o no rechazarla cuando es falsa. Si se consideran las hipótesis simples $H_0 : \mu = \mu_0$ *vs.* $H_1 : \mu = \mu_1$, éstas y las decisiones que se pueden tomar pueden representarse en la siguiente tabla

	Condición de H_0	
	Verdadera	Falsa
Rechazar H_0	**Error Tipo I**	$\surd$
No rechazar H_0	$\surd$	**Error Tipo II**

Rechazar H_0 cuando ésta es falsa, o no es rechazada cuando la hipótesis es verdadera son la decisiones correctas. En caso contrario, rechazar H_0 cuando ésta es correcta o no rechazarla cuando es falsa son decisiones erróneas pero de naturaleza distinta por sus consecuencias potenciales.

Rechazar la hipótesis nula H_0 cuando ésta es válida se conoce como **error tipo I**, mientras que aceptar H_0 cuando es falsa se dice que se comete un **error tipo II**. La probabilidad de cometer el error tipo I y el tipo II se acostumbran denotar por α y β, respectivamente.

$$\alpha = \mathbb{P}(\text{Error tipo I}) = \mathbb{P}(\text{Rechazar } H_0 | H_0 \text{ verdadera})$$

$$\beta = \mathbb{P}(\text{Error tipo II}) = \mathbb{P}(\text{Aceptar } H_0 | H_0 \text{ falsa})$$

Para continuar con el ejemplo, el consumo mensual de agua por persona $X \sim \mathrm{N}(\mu, \sigma^2)$, con σ conocida. Se toma una muestra de tamaño n $\boldsymbol{X} = (X_1, \ldots, X_n)$.

$$\alpha = \mathbb{P}(\mathcal{R}|H_0 \text{ verdadera}) = \mathbb{P}_0(\bar{X} \leq c) = \int_{-\infty}^{c} f_{\bar{X}}(x)dx = \Phi\left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right)$$

donde $f_{\bar{X}}$ denota la función de densidad del estadístico de prueba $\bar{X} \sim \mathrm{N}(\mu_0, \sigma^2/n)$ bajo la suposición de que la hipótesis nula H_0 es verdadera. La distribución del estadístico de prueba suponiendo válida la hipótesis nula H_0 se conoce como la **distribución nula** del estadístico. Por otro lado, se tiene que

$$\beta = \mathbb{P}(\mathcal{R}^C|H_0 \text{ falsa}) = \mathbb{P}_1(\bar{X} > c) = \int_{c}^{\infty} f^*_{\bar{X}}(x)dx = 1 - \Phi\left(\frac{c-\mu_1}{\sigma/\sqrt{n}}\right)$$

donde $f^*_{\bar{X}}$ denota la *f. d. p.* de $\bar{X}$ bajo el supuesto de que H_1 es verdadera, esto es, $\bar{X} \sim \mathrm{N}(\mu_1, \sigma^2)$.

Suponga que en el problema de consumo mensual de agua la desviación estándar es de $\sigma = 1.0$ m^3 y que el tamaño de la muestra es de $n = 25$ observaciones. Suponga también que el valor crítico fuera $c = 2.7$. Luego, la regla de decisión queda definida como: rechazar $H_0 : \mu = 3.0$ si $\bar{X} \leq 2.75$. Luego, la región de rechazo es $\mathcal{R} = \{\boldsymbol{X} = (X_1, \ldots, X_{25}) : \frac{1}{n}\sum_{i=1}^{25} X_i \leq 2.75\}$. Entonces, las probabilidades de error serían,

$$\alpha = \Phi\left(\frac{2.7-3.0}{1.0/\sqrt{25}}\right) = \Phi(-1.5) = 0.067$$

$$\beta = 1 - \Phi\left(\frac{2.7-2.5}{1/5}\right) = 1 - \Phi(1.0) = 0.159$$

Note que las probabilidades de error tipo I y II no son complementarias, es decir, no tienen que sumar 1. La figura 18 ilustra los errores en el ejemplo del consumo mensual de agua.

La probabilidad α del error tipo I es la medida o tamaño de la región de rechazo $\mathcal{R}$ suponiendo válida la hipótesis nula H_0 y en tal sentido $\boldsymbol{\alpha}$ se conoce como el **tamaño de la prueba** o la **significancia de la prueba**.

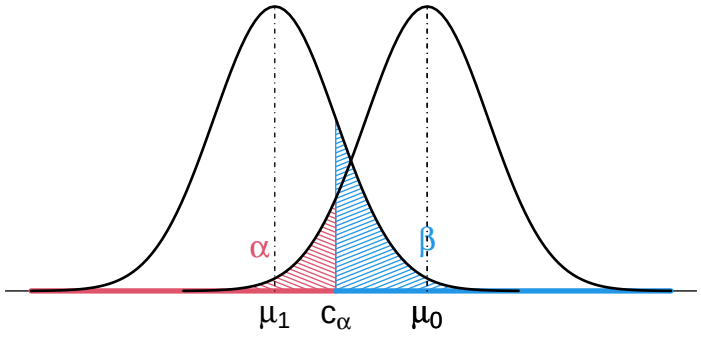


Figura 18: Contraste de hipótesis. Probabilidades de error tipo I y II.

En el ejemplo, la localización del valor crítico define la región de rechazo y en consecuencia las probabilidades de error, por lo que su determinación debe considerar las probabilidades de error consecuentes. En la figura 18 es claro que al disminuir un error necesariamente aumenta el otro. En general, en la práctica se predetermina la *significancia de la prueba* α o probabilidad de error tipo I o riesgos que se aceptan correr. Así, si determinara que en el ejemplo, se desea una prueba de significancia del 5 %, entonces, $\alpha = \Phi\left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right)$ implica que $z_{0.05} = \frac{c-3.0}{1/\sqrt{25}} = -1.644$, por lo que, si $\alpha = 0.05$, $c_\alpha = 2.671$. Esto es, la región de rechazo de tamaño α queda dada definida por $\mathcal{R}_\alpha = \{\bar{X} \leq c_\alpha\}$ de manera que $\alpha = \mathbb{P}_0(\mathcal{R}_\alpha)$ y en el ejemplo, $\mathbb{P}(\bar{X} \leq 2.671|\mu = 3.0) = 0.05$. Pero ahora, la correspondiente probabilidad del error tipo II es $\beta = 1 - \Phi\left(\frac{c_\alpha-\mu_1}{\sigma/\sqrt{n}}\right) = 0.196$.

Para disminuir simultáneamente las probabilidades de los errores la única manera es aumentando el tamaño de la muestra. Esto se ilustra en la figura 19 Si en el ejemplo, se

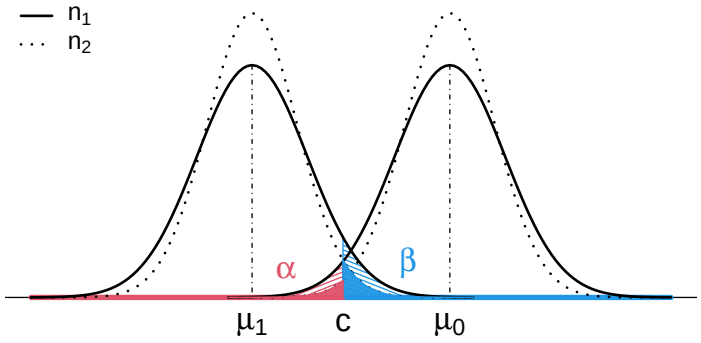


Figura 19: Disminución simultánea de las probabilidades de error tipo I y II aumentando el tamaño de la muestra, $n_1 < n_2$.

tomara en su lugar una muestra de tamaño $n = 40$, utilizando el mismo valor crítico de $c = 2.671$, los correspondientes errores serían $\alpha = \Phi\left(\frac{2.671-3.0}{1/\sqrt{40}}\right) = \Phi(-2.081) = 0.019$,

mientras que $\beta = 1 - \Phi\left(\frac{2.671-2.5}{1/\sqrt{40}}\right) = 1 - \Phi(1.081) = 0.140$. Compárense con los errores de 0.05 y 0.196, respectivamente cuando $n = 25$.

Otro concepto importante en la teoría de las pruebas de hipótesis es la **potencia de la prueba**, En estas notas se considerará la potencia de una prueba como la probabilidad de rechazar la hipótesis nula H_0 y se denotará por K. Así, uno hablaría de la potencia de prueba bajo H_0 o bien bajo H_1. En el ejemplo, la potencia de la prueba se define como $K(\mu_i) = \mathbb{P}_i(\mathcal{R})$. luego, idealmente se desearía $K(\mu_0) = 0$, y $K(\mu_1) = 1$. Pero la región de rechazo es $\mathcal{R} = \{\boldsymbol{x} : \bar{x} \leq c\}$, por lo que para $n = 25$ y $c = 2.671$, se tiene $K(3.0) = 0.05$ y $K(\mu_1) = 0.804$.

Más formalmente, se incluyen algunas de las definiciones antes presentadas para después incluir el lema de Neyman–Pearson, fundamental en el origen y el desarrollo de la Estadística Matemática. Varias de las siguientes definiciones son adaptadas de texto de Mood, Graybill, and Boes (1974).

Definición : En estas notas, una población de interés es definida por una característica de la misma, llamada **población estadística** que es representada o *modelada* mediante una distribución de probabilidad.

Así, el consumo de agua de cierta comunidad $\mathcal{X}$ quedó representada por una variable aleatoria X distribuida normalmente $\mathrm{N}(\mu, \sigma^2)$.

Definición : La inferencia sobre la población $\mathcal{X}$ usualmente se basa en la información colectada de una **muestra aleatroria** (*m. a.*) de X, entendiendo ésta como una colección de variables aleatorias independientes e idénticamente distribuidas a X, aquí denotada por $\boldsymbol{X} = (X_1, \ldots, X_n)$. Si se deseara hacer explícito el tamaño de la muestra, se usará el subíndice n, $\boldsymbol{X}_n$.

Definición : Una **hipótesis estadística** H es una aseveración o conjetura sobre la distribución de la variable o vector aleatorio X representando el modelo de la población.

Por ejemplo, suponer que la moneda lanzada es la moneda justa, con la probabilidad del águila $p = 1/2$. O bien, suponer que el consumo medio mensual de agua es de $\mu = 3$ m^3.

Definición : Se dice que una hipótesis estadística es **simple** si define completamente a distribución de X. En caso contrario se dice que es una **hipótesis compuesta**.

Por ejemplo, $\mathrm{H} : \mu = 3$ m^3 representa una hipótesis simple pues suponiendo *válido* H, se tendría que el consumo mensual de agua es $X \sim \mathrm{N}(3, 1)$, considerando $\sigma = 1$ conocida. Si no se conociera σ o se estableciera que $X \sim \mathrm{N}(\mu, \sigma^2)$, para $\mu \geq 3$, éstas serían hipótesis compuestas.

Definición : Una **prueba estadística** Υ es una regla o procedimiento empleado para decidir sobre la validez o no de la hipótesis H. Anteriormente, se distinguían la pruebas estadísticas como pruebas aleatorizadas o no-aleatorizadas.

Definición : Se dice que Υ es una **prueba aleatorizada** si para decidir sobre la hipótesis H, además del procedimiento resultado en la muestra aleatoria, la conclusión se basa también en el resultado de algún experimento aleatorio no relacionado con el procedimiento original.

Si, en el ejemplo inicial de decidir sobre qué moneda fue lanzada, suponga que la regla de decisión definida por la prueba Υ, fuese: aceptar $\mathrm{H}_0 : p = 0.5$ si $X < 6$; aceptar $\mathrm{H}_1 : p = 0.7$, si $X > 6$; y si $X = 6$, lanzar un dado *honesto* y favorecer H_0 si la cara muestra un número par y a H_1 si es impar.

Las pruebas aleatorizadas son rara vez empleadas, pues dos personas podría concluir sobre la hipótesis H de manera contraria a partir de la misma muestra aleatoria debido al

resultado aleatorio del experimento complementario. Más adelante se verá un ejemplo de tales pruebas.

Definición : Una **prueba no aleatorizada** Υ, es una regla de decisión sobre la hipótesis H definida a partir de un **estadístico (de prueba)** $T = T(\boldsymbol{X}_n)$ y una **región de rechazo** de manera que si la muestra aleatoria $\boldsymbol{X}$ es tal que si $T(\boldsymbol{X})$ satisface cierta condición $\mathcal{C}$ la hipótesis H es rechazada. Esto es, la región de rechazo es $\mathcal{R} = \{\boldsymbol{X} \in \mathcal{X} : T(\boldsymbol{X}) \in \mathcal{C} \Rightarrow$ rechazar H$\}$.

En el contraste de monedas, la hipótesis $H : p = 1/2$, el estadístico de prueba T es el número de águilas en los 10 lanzamientos y la regla de decisión es rechazar H si $T \geq 7$. O bien, en el ejemplo de consumo de agua, para decidir sobre la hipótesis $H : \mu = 3.0$, a partir de la muestra $\boldsymbol{X} = (X_1, \ldots, X_n)$, el estadístico de prueba es el promedio $T(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$ y la region de rechazo es $\mathcal{R} = \{\boldsymbol{X} : T(\boldsymbol{X}) < c\}$, para algún c, definido como **valor crítico**.

Definición : Sea Υ una prueba estadística para H_0. Si rechaza H_0 cuando ésta es correcta se dice que se comente el **error tipo I**. Por otro lado, Si no se rechaza H_0 cuando ésta es falsa se comete el **error tipo II**.

Definición : Sea Υ una prueba para la hipótesis H_0. Se define la **potencia de la prueba**, $K(\theta)$, como la probabilidad de rechazar H_0, cuando la distribución ha sido parametrizada por θ.

Esto es, si $X \sim f(x;\theta)$, con $\theta \in \Theta = \Theta_0 \cup \Theta_1$, donde Θ_0 denota los posibles valores de θ bajo la hipótesis H_0. Sea Υ la prueba definida por el estadístico de prueba $T(\boldsymbol{X})$ y la región de rechazo $\mathcal{R}$. Entonces, la potencia de la prueba Υ es

$$K(\theta) = \mathbb{P}(\mathcal{R}|\theta) = \int_{\mathcal{R}} f(\boldsymbol{x};\theta)d\boldsymbol{x}, \qquad \text{para } \theta \in \Theta$$

Si se considera nuevamente el ejemplo del consumo mensual de agua, $X \sim \mathrm{N}(\mu, \sigma^2)$, con $\sigma = 1$ y $n = 25$. El estadístico de prueba es $T(\boldsymbol{X}) = \bar{X}$ y $\mathcal{R} = \{\bar{X} < 2.671\}$. Entonces,

$$K(\mu) = \mathbb{P}(\bar{X} < c|\mu) = \Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(5(2.671-\mu)\right)$$

La función potencia K del ejemplo de consumo de agua se muestra en la figura 20.

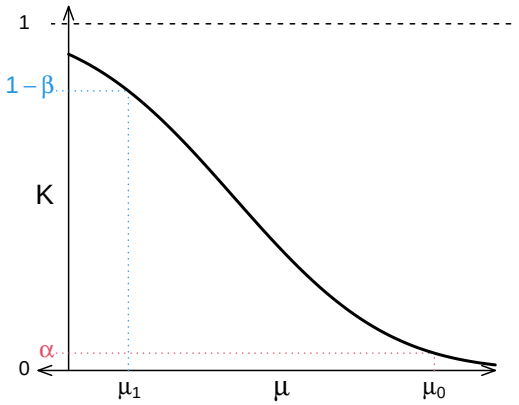


Figura 20: Función potencia K para el ejemplo del consumo mensual de agua. La prueba muestra una potencia de α y $1-\beta$ para los valores de la media $\mu_0 = 3.0$ y $\mu_1 = 2.5$, respectivamente.

La función potencia juega un papel muy relevante en la comparación y definición de pruebas estadísticas. Idealmente, se quisieran pruebas con potencia 0 bajo H_0 y 1 cuando ésta no se cumple. Así, se prefieren las pruebas con mayor potencia.
Notación: Sea $K(\theta)$ la potencia de la prueba Υ para la hipótesis H_0. La probabilidad de los errores tipo I y II se acostumbran a denotar por

$$\alpha(\theta) = \mathbb{P}(\text{Rechazar } H_0 | H_0 \text{ verdadera}) = K_0(\theta)$$
$$\beta(\theta) = \mathbb{P}(\text{No rechazar } H_0 | H_0 \text{ falsa}) = 1 - K_0(\theta)$$

La figura 20 muestra las probabilidades de errores $\alpha(\mu_0)$ y $\beta(\mu_1)$ para el ejemplo del consumo de agua.

Definición : Sea Υ una prueba de la hipótesis $H_0 : \theta \in \Theta_0$, donde $\Theta = \Theta_0 \cup \Theta_1$ es el espacio parametral ($\Theta_0 \cap \Theta_1 = \emptyset$). Sea $\mathcal{R}$ la correspondiente región de rechazo de H_0. Se define el **tamaño de la prueba** o **tamaño de la región crítica** por

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(\mathcal{R}|\theta)$$

Algunos texto se refieren al tamaño de la región crítica como la **significancia de la prueba** $\boldsymbol{\alpha}$.

Suponga que en el ejemplo sobre consumo de agua se considera la hipótesis $H_0 : \mu \geq 3.0$, $\big(\Theta_0 = \{\theta \geq 3.0\}\big)$ y la prueba Υ con el estadístico de prueba $\bar{X} = \frac{1}{n}\sum_{1i=1}^{n} X_i$ y la región de rechazo $\mathcal{R} = \{\bar{X} \leq c\}$. Entonces, como se puede ver en la figura 20,

$$K(\mu) = \mathbb{P}(\mathcal{R}|\mu) = \Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right)$$

que alcanza el supremo en el extremo $\mu = 3.0 \in \Theta_0$. Así, la significancia de la prueba Υ es α. Con $n = 25$, si $c = 2.70$, la significancia de la prueba es $\alpha = 0.0.067$. Si en su lugar, se considera el valor crítico de $c = 2.67$, la significancia sería de $\alpha = 0.05$.

Note que la significancia de una prueba es la probabilidad de cometer el error de tipo I. Otra manera de ver las significancia sería como el riesgo de cometer el error tipo I. En la práctica toda prueba tiene asociado tal riesgo. Luego, la manera formal de proceder en una prueba de hipótesis estadística es fijar de antemano el riesgo (significancia) que se acepta correr, determinar el estadístico de prueba y regla de decisión correspondiente y llevar a cabo el experimento (observación, procedimiento, etc.). Una vez que se determina el valor del estadístico concluir si se rechaza o no la hipótesis nula. En este contexto, no es válido (¿ético?) cambiar la decisión después de obtener el estadístico. En algunos casos, pruebas aleatorizadas y pruebas secuenciales, se pueden no tomar la decisión final sobre H_0 después de observar el estadístico de prueba, pero el procedimiento fue así establecido en un principio.

Por otro lado, como se comentó anteriormente, aumentando el tamaño de la muestra se reduce las probabilidades de los errores. Igualmente, la potencia de una prueba se puede aumentar con muestras de mayor tamaño. La figura 21 muestra la función potencia K para pruebas de significancia $\alpha = 0.05$ y distintos tamaños de muestra en el ejemplo del consumo mensual de agua.

Definición : Sea Υ una prueba de significancia para la hipótesis H_0. Se define la **distribución nula** del estadístico de prueba T a su distribución suponiendo la hipótesis nula verdadera.

En el ejemplo, para una muestra aleatoria $\boldsymbol{X}_n$, la distribución nula de $T = \bar{X}$ es $N(\mu_0, \sigma^2/n)$, que para $\sigma = 1$ y $n = 25$, $T \sim N(3.0, 1/25)$.

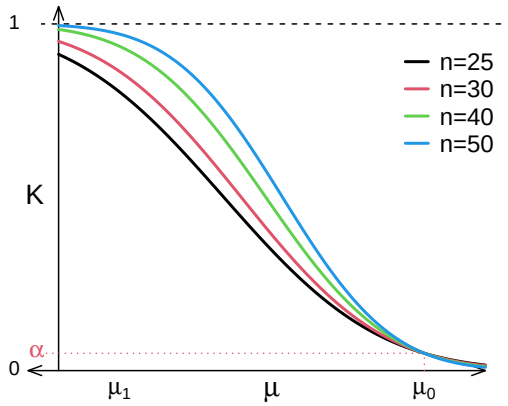


Figura 21: Función potencia K del ejemplo del consumo mensual de agua para distintos tamaños de muestra n y significancia común α.

Definición : Sea $\boldsymbol{x} = (x_1, \ldots, x_n)$ la muestra observada. Entonces, $t = T(\boldsymbol{x}) \in \mathbb{R}$ y se le dice el **estadístico observado**.

Dada $H_0 : \mu = 3.0$, y una muestra de tamaño 25, la prueba de significancia 0.05 tiene la región de rechazo $\mathcal{R} = \{\bar{X} \leq 2.671\}$. Luego, un estadístico observado $t = 2.73$ llevaría a no rechazar la H_0 con una significancia de 5 %. Si por el contrario, se observara $t = 2.63$, se concluiría que se rechaza la hipótesis con una significancia del 5 %. Pero igual se rechazaría con $t = 2.24$. Si bien los dos últimos estadísticos observados llevan a la misma conclusión: rechazar H_0 al 5 % de significancia, pareciera que la información en contra de la hipótesis es más evidente cuando $t = 2.24$ que $t = 2.63$. Una manera de valor al diferencia entre las dos observaciones en mediante el valor$-p$.

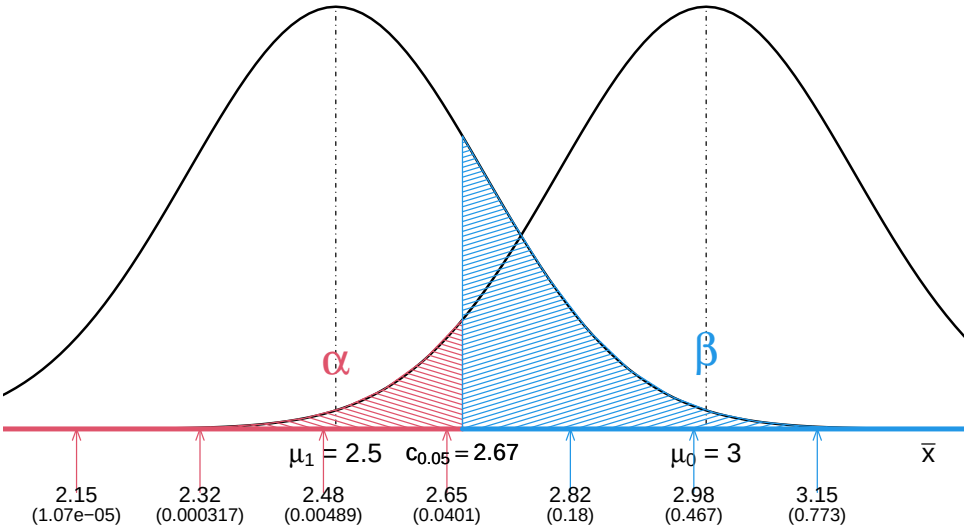


Figura 22: Estadísticos de prueba ($\bar{x}$) observados y sus correspondientes valores-p.

Definición : Se define el **valor$-p$** del estadístico observado $t = T(\boldsymbol{x})$ como la probabilidad de obtener t o algo más extremo en contra de la hipótesis H_0 suponiendo ésta verdadera.

En el ejemplo, $\bar{x} = 2.7$ tiene el valor$-p$ de 0.067 y si $\bar{x} = 2.671$, el valor$-p$ es del 5 %. La figura 22 muestra distintos ejemplos de estadísticos y debajo de ellos, entre paréntesis, los

correspondientes valores-p. Note que aquellos estadísticos $\bar{x}$ menores al valor crítico $\alpha_{0.05} = 2.671$, que llevarían a rechazar la hipótesis H_0 con una significancia del $5\,\%$, tiene valores-p menores 0.05. Es aparente que la evidencia en contra de H_0 de $\bar{x} = 2.15$ (1.07×10^{-5}) es más clara que la que presenta $\bar{x} = 2.65$ (0.040). Es común reportar en textos y artículos el valor$-p$ de los estadísticos observados en paréntesis inmediatamente después o debajo de la cantidad. Esto se ilustra en la figura 22.

Note que para determinar el valor$-p$ no se necesita conocer la hipótesis alternativa. De hecho, tampoco se requiere de un nivel de significancia de la prueba definido. En cierta forma, el valor$-p$ indica el respaldo de los datos a la hipótesis nula. Quien tome la decisión de rechazarla o no considerará el riesgo que desea correr.

En los últimos años ha habido mucha controversia sobre el uso, o mal uso, del valor$-p$. Personalmente creo que son unos indicadores, función de la distribución nula y la muestra observada, útiles por lo informativos, pero mal empleados y abusados. Su interpretación es la misma, independientemente de la hipótesis bajo consideración y estadístico de prueba empleado. Pero no dice nada de si, por ejemplo, lo apropiado de la hipótesis para propósito del estudio, si los datos son obtenidos por experimentación, cómo se llevó este último a cabo, si los supuestos fueron verificados y se satisfacen, etcétera. Hay mucho escrito al respecto que vale la pena revisar. En un inicio, puede acudir a Bastian (2013).

Finalmente, para terminar esta sección se presenta el siguiente teorema que refuerza la importancia de los estadísticos suficientes en la inferencia estadística.

Teorema : Sean $\boldsymbol{X}$ una muestra aleatoria de $X \sim f(x;\theta)$, con $\theta \in \Theta$ y $T(\boldsymbol{X})$ un estadístico suficiente para θ. Entonces, para cualquier prueba estadística Υ con potencia $K_\Upsilon(\theta)$ existe una prueba Υ^* con estadístico de prueba T con la misma potencia, esto es, $K_{\Upsilon^*}(\theta) = K_\Upsilon(\theta)$, para todo $\theta \in \Theta$.

Demostración: refiérase al teorema IX.1.1 p.408 de Mood, Graybill, and Boes (1974).

7.3. Lema de Neyman–Pearson

Con el apoyo de Mood, Graybill, and Boes (1974) y Rice (2007).

Teorema de Neyman-Pearson (1933) Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f$. Sean H_i hipótesis simples con $f_i(\boldsymbol{x})$, $i = 0, 1$, las correspondientes funciones de densidad conjunta completamente definidas. Sea Υ^* la prueba que define la región de rechazo de tamaño α, $\mathcal{R} = \{\boldsymbol{x} : f_0(\boldsymbol{x})/f_1(\boldsymbol{x}) \leq c\}$. Entonces, cualquier otra prueba Υ de significancia menor o igual a α tiene una potencia menor o igual a que aquella basada en el cociente de verosimilitudes $f_0(\boldsymbol{x})/f_1(\boldsymbol{x})$.

Demostración: Sea $\boldsymbol{X}$ una *m. a.* de $X \sim f$ y las hipótesis simples $\mathrm{H}_0 : f = f_0$ y $\mathrm{H}_1 : f = f_1$ que se desean contrastar. Sea d la regla o función de decisión tal que

$$d(\boldsymbol{X}) = \begin{cases} 0 & \text{si se acepta } \mathrm{H}_0 \\ 1 & \text{si se rechaza } \mathrm{H}_0 \end{cases}$$

Entonces, $d(\boldsymbol{X}) \sim \mathrm{Ber}(p)$, con $p = \mathbb{E}[d(\boldsymbol{X})] = \mathbb{P}(d(\boldsymbol{X}) = 1) = \mathbb{P}(\text{Rechazar } \mathrm{H}_0)$. Luego,

$$\alpha = \mathbb{P}(\mathcal{R}|\mathrm{H}_0) = \mathbb{P}_0(\mathcal{R}) = \mathbb{E}_0[d(\boldsymbol{X})] = K_0(d(\boldsymbol{X}))$$

y $K_1(d(\boldsymbol{X})) = \mathbb{P}_1(\mathcal{R}) = \mathbb{E}_1[d(\boldsymbol{X})]$.

Sea ahora $d^*(\boldsymbol{X})$ la regla de decisión definida por Υ^* del problema tal que, $d^*(\boldsymbol{X}) = 1$, si $f_0(\boldsymbol{X}) \leq c f_1(\boldsymbol{X})$ con $\alpha = \mathbb{E}_0[d^*(\boldsymbol{X})]$ y sea Υ cualquier otra prueba con regla de decisión $d(\boldsymbol{X})$ tal que $\mathbb{E}_0[d(\boldsymbol{X})] \leq \mathbb{E}_0[d^*(\boldsymbol{X})] = \alpha$. Por mostrar que $K_1(d(\boldsymbol{X})) \leq K_1(d^*(\boldsymbol{X}))$, esto es

$$K_1(d(\boldsymbol{X})) = \mathbb{E}_1[d(\boldsymbol{X})] \leq \mathbb{E}_1[d^*(\boldsymbol{X})] = K_1(d^*(\boldsymbol{X}))$$

para esto, note que

$$d(\boldsymbol{X}) \left[c f_1(\boldsymbol{X}) - f_0(\boldsymbol{X})\right] \leq d^*(\boldsymbol{X}) \left[c f_1(\boldsymbol{X}) - f_0(\boldsymbol{X})\right] \tag{13}$$

puesto que, si

$$\begin{aligned} &i) \quad \text{Si } d^*(\boldsymbol{X}) = 1, \text{ es porque } f_0(\boldsymbol{X}) \leq c f_1(\boldsymbol{X}) \\ &ii) \quad \text{Si } d^*(\boldsymbol{X}) = 0, \text{ es porque } f_0(\boldsymbol{X}) > c f_1(\boldsymbol{X}) \end{aligned}$$

Ahora, tomando integrales a ambos lados de la expresión (13), se sigue del teorema del estadístico inconsciente (TEI) se tiene

$$c\mathbb{E}_1[d(\boldsymbol{X})] - \mathbb{E}_0[d(\boldsymbol{X})] \leq c\mathbb{E}_1[d^*(\boldsymbol{X})] - \mathbb{E}_0[d^*(\boldsymbol{X})] \tag{14}$$

Luego,

$$0 \leq E_0[d^*(\boldsymbol{X})] - \mathbb{E}_0[d(\boldsymbol{X})] \leq c\left\{\mathbb{E}_1[d^*(\boldsymbol{X})] - \mathbb{E}_1[d(\boldsymbol{X})]\right\}$$

y por lo tanto,

$$K_1\big(d(\boldsymbol{X})\big) \leq K_1\big(d^*(\boldsymbol{X})\big)$$

En palabras: en el contraste de un par de hipótesis simples, la prueba de significancia α basada en el cociente de verosimilitudes es la de mayor potencia entre todas las pruebas de significancia menor o igual α.

Ejemplo : Considere el ejemplo del consumo mensual medio de agua, las hipótesis simples, $\mathrm{H}_0 : \mu = \mu_0$ *vs.* $\mathrm{H}_1 : \mu = \mu_1 (< \mu_0)$. Sea $\boldsymbol{X}_n$ una *m. a.* de $X \sim \mathrm{N}(\mu, \sigma^2)$, con σ conocida. Entonces, la *f. d. p.* conjunta del *v. a.* $\boldsymbol{X}$ es $f(\boldsymbol{x}; \mu, \sigma) = (2\pi)^{-n/2}\sigma^{-n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i -$

$\mu)^2\}$. Luego, de acuerdo al lema de Neyman-Pearson, el cociente de verosimilitudes $f_0/f_1 \leq c$ da lugar a

$$\begin{aligned}
\frac{f(\boldsymbol{x};\mu_0)}{f(\boldsymbol{x};\mu_1)} &\leq c \\
\frac{(2\pi)^{-n/2}\sigma^{-n}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu_0)^2\}}{(2\pi)^{-n/2}\sigma^{-n}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu_1)^2\}} &\leq c \\
-\frac{1}{2\sigma^2}\left\{-2(\mu_0-\mu_1)\sum x_i + n(\mu_o^2-\mu_1^2)\right\} &\leq \log c \\
n(\mu_0-\mu_1)\bar{x} - n(\mu_0^2-\mu_1^2) &\leq \sigma^2\log c \\
\bar{x} &\leq c^*
\end{aligned}$$

con $c^* = \frac{1}{n}\left(\sigma^2\log c - n(\mu_0^2-\mu_1^2)\right)$. Luego, la prueba con mayor potencia para contrastar las hipótesis $\mathrm{H}_0 : \mu = \mu_0$ *vs.* $\mathrm{H}_1 : \mu = \mu_1 (< \mu_0)$, induce una región de rechazo $\mathcal{R}^* = \{\boldsymbol{X} : \bar{X} \leq c^*\}$, como se había propuesto anteriormente. Note que $\bar{X}$ es un estadístico suficiente para μ. Finalmente, si la prueba se desea de tamaño α, se ha de elegir c_α tal que $\mathcal{R}_\alpha = \{\bar{X} \leq c_\alpha\}$ de manera que $\alpha = \mathbb{P}_0(\mathcal{R}_\alpha)$.

Una muestra observada que lleve a $\bar{x} = 2.4$ tiene un valor$-p$ de $\Phi\left(5 * (2.4 - 3.0)\right) = 0.0013$. Por lo que uno concluiría en rechazar la hipótesis $\mathrm{H}_0 : \mu = 3.0$, aún con una significancia del 0.5 %.

Ejercicio : Recupere el ejemplo inicial del lanzamiento de monedas. Muestre que $\mathcal{R} = \{X : X > c\}$ es la región de rechazo determinada por el lema de Neyman-Pearson. Si el valor crítico es $c = 6$, determine la significancia de la prueba α y la probabilidad de error tipo II β. Determine la significancia de la prueba. Calcule la potencia de la prueba $K(p)$, para $p = 0.5$ y $p = 0.7$.

Ejercicio : (Hogg and Craig (1978)) Se desea contrastar las hipótesis $H_0 : f = f_0$ *vs.* $H_1 : f = f_1$, donde

$$f_0(x) = \frac{e^{-1}}{x!}\mathbb{1}_{\{0,1,2,\dots\}}(x) \qquad \text{y} \qquad f_1(x) = \left(\frac{1}{2}\right)^{x+1}\mathbb{1}_{\{0,1,2,\dots\}}(x)$$

Sea $\boldsymbol{X} = (X_1, \dots, X_n)$ es una muestra aleatoria de $X \sim f$,

1. Muestre que la mejor región de rechazo para el contraste de las hipótesis es de la forma
$$\mathcal{R} = \left\{\boldsymbol{x} : \log 2 \cdot \sum_{i=1}^n x_i - \sum_{i=1}^n \log x_i! < c^*\right\}$$
con $c^* = \log\, c_0 - n\log(2e^{-1})$ y donde c_0 viene del cociente de verosimilitudes $\mathrm{CV}(\boldsymbol{x}) = \frac{f_0(\boldsymbol{x})}{f_1(\boldsymbol{x})} < c_0$.
2. Para $c_0 = 1$ y $n = 1$, determine $\mathcal{R}$ del inciso anterior.
3. Calcule la significancia de la prueba.
4. Calcule β, la probabilidad del error tipo II.
5. Grafique K, la función potencia de la prueba.

Ejemplo : Considere la población $Y \sim \mathrm{Exp}(\lambda)$, tal que $\mathbb{E}[Y] = 1/\lambda$. Se desea contrastar las hipótesis $\mathrm{H}_0 : \lambda = \lambda_0$ *vs.* $\mathrm{H}_1 : \lambda = \lambda_1 (< \lambda_0)$. Por encontrar la prueba significancia α más potente.

Sea $\boldsymbol{Y}_n$ una muestra aleatoria de $Y \sim \text{Exp}(\lambda)$ de tamaño n. Luego, se sigue del lema de Neyman-Pearson

$$\begin{aligned} \text{CV} = \frac{f_0(\boldsymbol{y})}{f_1(\boldsymbol{y})} &\leq c \\ \frac{\lambda_0^n e^{-\lambda_0 \sum y_i}}{\lambda_1^n e^{-\lambda_1 \sum y_i}} &\leq c \\ n\log(\lambda_0/\lambda_1) - (\lambda_0 - \lambda_1)\sum y_i &\leq \log c \\ \bar{y} &\geq -\frac{\log c - n\log(\lambda_0/\lambda_1)}{\lambda_0 - \lambda_1} \end{aligned}$$

Entonces la prueba más potente de significancia α tiene una región de rechazo de la forma $\mathcal{R}_\alpha = \{\boldsymbol{y} : \bar{y} > c_\alpha\}$, tal que $\alpha = \mathbb{P}(\mathcal{R}_\alpha|\text{H}_0) = \mathbb{P}_0(\bar{Y}_n > c_\alpha)$. Note que el estadístico de prueba $\bar{Y} = \frac{1}{n}\sum Y_i$ es suficiente para el parámetro λ y rechazará la hipótesis nula para valores grandes, lo que sugiere una media poblacional μ grande que corresponde a valores de λ pequeños.

Por otro lado verifique que Si $Y \sim \text{Exp}(\lambda)$ entonces $2\lambda\sum Y_i \sim \chi^2_{2n}$. Por lo que la región de rechazo de significancia α de más potencia tiene un valor crítico $c_\alpha = \dfrac{1}{2n\lambda_0}\chi^2(1-\alpha; 2n)$, siendo el último término el cuantil $1-\alpha$ de la distribución χ^2_{2n}.

Así, con $\lambda_1 < \lambda_0$, para el contraste de hipótesis

$$\text{H}_0 : \lambda = \lambda_0 \quad \textit{vs.} \quad \text{H}_1 : \lambda = \lambda_1$$

la mejor región de rechazo de significancia α es

$$\mathcal{R}_\alpha = \left\{ \boldsymbol{Y}_n : \bar{Y} > \frac{1}{2n\lambda_0}\chi^2(1-\alpha; 2n) \right\}$$

Si por ejemplo, $\lambda_0 = 2$ y $\lambda_1 = 1$, considerando muestras de tamaño $n = 20$, para una prueba de significancia $\alpha = 0.05$ el valor crítico es $c = 0.697$. Y si, el promedio observado fuese $\bar{y} = 0.65$ $(< c)$, se concluiría que no se rechaza la hipótesis $\text{H}_0 : \lambda = 2$, con una significancia del 5 %. Más aún, el valor-p de $\bar{y} = 0.65$, el estadístico observado, es $\mathbb{P}_0(\bar{Y} \geq 0.65) = \mathbb{P}\left(2n\lambda_0\bar{Y} > 80(0.65)\right) = 0.0968$. O bien, si el (promedio) estadístico observado fuese $\bar{y} = 0.76$, el correspondiente valor-p es $\mathbb{P}_0(\bar{Y} \geq 0.76) = \mathbb{P}\left(2n\lambda_0\bar{Y} > 80(0.75)\right) = 0.0185$, lo que llevaría a rechazar la H_0 con una significancia del 5 %. Finalmente, note que si la media observada fuese de 0.89, el correspondiente valor-p es de 0.0017. Ambos estadísticos observados, $\bar{y}_{20} = 0.76,\ 0.89$ llevan a rechazar H_0 con una significancia del 5 %, pero $\bar{y} = 0.89$, con un valor-p más pequeño, se concluiría que hay más evidencia en contra de la hipótesis nula.

Para graficar la potencia de la prueba recuerde que ésta es la probabilidad de rechazar la hipótesis nula H_0. Entonces,

$$K(\lambda) = \mathbb{P}_\lambda(\mathcal{R}) = \mathbb{P}_\lambda(\bar{Y} > c_\alpha) = \mathbb{P}(2n\lambda\bar{Y} > 2n\lambda c_\alpha) = \mathbb{P}(W > 2n\lambda c_\alpha)$$

donde $W \sim \chi^2_{2n}$. Luego, $K(\lambda_0) = \mathbb{P}(W > 2n\lambda_0 c_\alpha) = \mathbb{P}(W > 2(20)(2)0.697) = 0.05$ y $K(\lambda_1) = \mathbb{P}(W > 2n\lambda_1 c_\alpha) = \mathbb{P}(W > 2(20)(1)0.697) = 0.926$. Calculando $K(\lambda)$ para varios valores de λ en $[0.5, 2.5]$ se tiene la función potencia que se muestra en la figura 23.

Pruebas aleatorizadas

Considere el modelo $X \sim \text{Po}(\lambda)$. Se desean contrastar las hipótesis

$$\text{H}_0 : \lambda = \lambda_0 = 1 \quad \textit{vs.} \quad \text{H}_1 : \lambda = \lambda_1 = 5$$

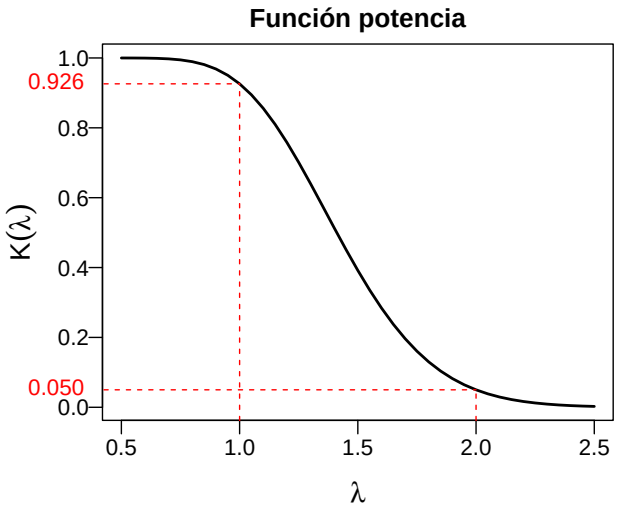


Figura 23: Función potencia de la prueba de hipótesis del ejemplo de la distribución exponencial

.

Para esto se toman muestras de tamaño $n = 10$ y se define la región de rechazo $\mathcal{R} = \{\boldsymbol{X} : T(\boldsymbol{X}) \geq 3\}$, con $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ el estadístico de prueba. Entonces, $\mathcal{R}$ es la mejor prueba para el contraste de las hipótesis.

En efecto, se sigue del lema de Neyman-Pearson que

$$\begin{aligned} \frac{f(\underline{x}; \lambda_0)}{f(\boldsymbol{x}; \lambda_1)} &\leq c_0 \\ \frac{e^{-\lambda_0} \lambda_0^{\sum x_i} / \prod x_i!}{e^{-\lambda_1} \lambda_1^{\sum x_i} / \prod x_i!} &\leq c_0 \\ -n(\lambda_0 - \lambda_1) + \log(\lambda_0/\lambda_1) \sum x_i &\leq c_1 \\ \sum x_i &\geq c_2 \end{aligned}$$

puesto que $\lambda_0/\lambda_1 < 1$.

La distribución nula del estadístico $T = \sum X_i \sim \text{Po}(n\lambda_0)$, por lo que la significancia de la prueba con región de rechazo $\mathcal{R} = \{\sum X_i \geq 3\}$ es

$$\alpha = \mathbb{P}_0(\mathcal{R}) = 1 - \mathbb{P}_0\left(\sum X_i \leq 2\right) = 0.080$$

mientras que si se considerase la región $\mathcal{R} = \{\sum X_i \geq 4\}$, la significancia sería de $\mathbb{P}_0(\sum X_i \geq 4) = 0.019$.

Si se deseara una prueba de significancia del 5 % exactamente se podría recurrir al siguiente procedimiento:

i) Construya la variable aleatoria $Y \sim \text{Ber}(p)$, con $p = \dfrac{0.05 - 0.019}{0.080 - 0.19} = 0.508.$, independiente de las X's.

ii) Rechace H_0 si $\sum X_i \geq 4$ ó $\sum X_i = 3$ y $Y = 1$.

iii) Acepte H_0 si $\sum X_i \leq 2$.

Esto es, se define la región de rechazo $\mathcal{R} = \{\sum X_i \geq 4\} \cup \{\sum X_i = 3, Y = 1\}$. Y en tal caso, la significancia de la prueba es

$$\begin{aligned}\mathbb{P}_0(\mathcal{R}) &= \mathbb{P}_0\left(\sum X_i \geq 4\right) + \mathbb{P}_0\left(\sum X_i = 3, Y = 1\right) \\ &= \mathbb{P}_0\left(\sum X_i \geq 4\right) + \mathbb{P}_0\left(\sum X_i = 3\right)\mathbb{P}_0\left(Y = 1\right) \\ &= 0.019 + 0.061(0.508) \\ &= 0.05\end{aligned}$$

El recurso anterior se conoce como una prueba aleatorizada. Note que mediante el procedimiento anterior se puede llegar a decisiones contrarias aún bajo los mismos datos (información), lo que sugiere el uso de los valores-p que no están ligados a la significancia de la prueba.

7.4. Hipótesis compuestas

7.4.1. Pruebas uniformemente más potentes

Recupere el ejemplo del consumo mensual de agua. Para contrastar las hipótesis $\mathrm{H}_0 : \mu = \mu_0$ *vs.* $\mathrm{H}_1 : \mu = \mu_1$ $(< \mu_0)$, se encontró Υ^*, la mejor prueba de significancia α, con una región de rechazo $\mathcal{R}_\alpha = \{\bar{X} < c_\alpha\}$. Se ha visto también que si la hipótesis nula fuese $\mathrm{H}_0 : \mu \geq \mu_0$, Υ^* seguiría siendo la mejor prueba de significancia α. Por otro lado, note que para todo $\mu_1^* < \mu_0$ fijo, la región de rechazo sigue siendo la misma y Υ^* sigue siendo la mejor prueba de significancia α, y es en este sentido que se dice que Υ^* es uniformemente $(\mu < \mu_0)$ más potente.

Más formalmente, sea $X \sim f(x;\theta)$, con $\theta \in \Theta = \Theta_0 \cup \Theta_1$, $\Theta_1 = \Theta \setminus \Theta_0$.

Definición : Una prueba Υ^* para el contraste de hipótesis $\mathrm{H}_0 : \theta \in \Theta_0$ *vs.* $\mathrm{H}_1 : \theta \in \Theta_1$ se dice que es una **prueba uniformemente más potente (PUMP)** de significancia α si y solo si:

i) $\alpha = \sup_{\theta\in\Theta_0} K_{\Upsilon^*}(\theta)$.

ii) $K_{\Upsilon^*}(\theta) \geq K_\Upsilon(\theta)$, para todo $\theta \in \Theta_1$ y para toda prueba Υ de significancia α.

Ejemplo : Como se comentó anteriormente, en el ejemplo de consumo de agua, si la hipótesis alternativa fuese $\mathrm{H}_1 : \mu < \mu_0$, la prueba Υ^*, con región de rechazo $\mathcal{R}_\alpha = \{\bar{X} < c_\alpha\}$, es PUMP de significancia α.

El caso anterior como el presentado para el modelo exponencial, ambos ejemplos son PUMP pero también resultan ser uniformemente más potentes al variar α.

Definición : Una prueba Υ para contrastar las hipótesis $\mathrm{H}_0 : \theta \in \Theta_0$ *vs.* $\mathrm{H}_1 : \theta \in \Theta_1$ se dice **insesgada** si

$$\sup_{\theta\in\Theta_0} K_\Upsilon(\theta) \leq \inf_{\theta\in\Theta_1} K_\Upsilon(\theta)$$

Esto es, *la probabilidad de rechazar H_0 cuando ésta es falsa es mayor o igual a la probabilidad de rechazarla cuando es verdadera.*

7.4.2. Pruebas dos colas

Suponga que en el ejemplo del consumo mensual de agua, la hipótesis nula es $\mathrm{H}_0 : \mu = \mu_0$ nuevamente pero la alternativa es ahora $\mathrm{H}_1 : \mu \neq \mu_0$. La región de rechazo $\{\bar{X} < c\}$ corresponde a la PUMP para el contraste con hipótesis alternativa $\mathrm{H}_1 : \mu < \mu_0$, mientras

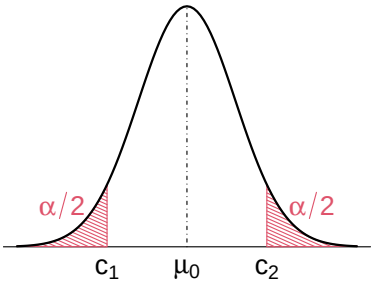


Figura 24: Ejemplo consumo de agua. Prueba de dos colas.

que la PUMP para la hipótesis $\mathrm{H}_1 : \mu > \mu_0$, la región es $\{\bar{X} > c\}$. Luego, no existe una prueba uniformemente más potente para cuando la hipótesis alternativa es $\mathrm{H}_1 : \mu \neq \mu_0$.

Considere entonces la región de rechazo $\mathcal{R} = \{\bar{X} \leq c_1\} \cup \{\bar{X} \geq c_2\}$. En tal caso, la significancia de la prueba es

$$\mathbb{P}_0(\mathcal{R}) = \mathbb{P}_0(\bar{X} \leq c_1) + \mathbb{P}_0(\bar{X} \geq c_2)$$

Si la hipótesis nula no se verifica y se supone que le media real puede igualmente estar a la izquierda o derecha de μ_0, una opción es asignar el mismo peso (probabilidades) a las colas. Luego, si la significancia es α, cada una de las colas seria con probabilidad $\alpha/2$. La figura 24 muestra la región de rechazo $\mathcal{R}$. Los valores críticos serían

$$c_1 = \mu_0 + z_{\alpha/2}\,\sigma/\sqrt{n} \quad \text{y} \quad c_2 = \mu_0 + z_{1-\alpha/2}\,\sigma/\sqrt{n}$$

Si la prueba es de significancia $\alpha = 0.05$, entonces $c_1 = 3.0 - 1.96(1)/\sqrt{25} = 2.671$ y $c_2 = 3.0 + 1.96(1)/\sqrt{25} = 3.329$.

7.4.3. Pruebas de hipótesis e intervalos de confianza

Basado en Rice (2007).

Ejemplo : Recupere el ejemplo del consumo de agua. Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim \mathrm{N}(\mu, \sigma^2)$, con σ conocida. Se desea contrastar las hipótesis

$$\mathrm{H}_0 : \mu = \mu_0 \quad v.\,a. \quad \mathrm{H}_1 : \mu \neq \mu_0$$

Como se vio anteriormente una "buena prueba" sería una prueba de dos colas con la región de rechazo $\mathcal{R} = \{\bar{X} > c_1\} \cup \{\bar{X} < c_2\}$, y si las colas son del mismo peso (probabilidad) en las colas, $\mathcal{R} = \left\{|\bar{X} - \mu_0| > z_{1-\alpha/2}\sigma/\sqrt{n}\right\}$ define una región de rechazo de significancia α. Sea,

$$\begin{aligned}\mathcal{A} = \mathcal{R}^C &= \left\{\boldsymbol{X} : |\bar{X} - \mu_0| \leq z_{1-\alpha/2}\sigma/\sqrt{n}\right\} \\ &= \left\{\boldsymbol{X} : \bar{X} - z_{1-\alpha/2}\sigma/\sqrt{n} \leq \mu_0 \leq \bar{X} + z_{1-\alpha/2}\sigma/\sqrt{n}\right\}\end{aligned}$$

por lo que fija $\bar{X}$, $\left\{\mu : \mu \in \left(\bar{X} \pm z_{1-\alpha/2}\sigma/\sqrt{n}\right)\right\}$ constituye una región de confianza $1-\alpha$ para μ. En palabras, *la región de confianza para μ la constituye todos los μ_0 para los cuales $\bar{X}$ aceptaría la hipótesis $H_0 : \mu = \mu_0$.*

Sea θ un parámetro de la familia de distribuciones de probabilidad. Se tiene $\theta \in \Theta$ espacio parametral sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$.

Teorema : Suponga que para cada $\theta_0 \in \Theta$, existe una prueba de significancia α para la hipótesis $\mathrm{H}_0 : \theta = \theta_0$. Sea $\mathcal{A}(\theta_0) = \{\boldsymbol{X} : \text{se acepta } \mathrm{H}_0 : \theta = \theta_0\}$. Entonces el conjunto $\mathcal{C}(\boldsymbol{X}) = \{\theta : \boldsymbol{X} \in \mathcal{A}(\theta)\}$ constituye una región de confianza $1-\alpha$ para θ.

Demostración: Como $\mathcal{A}$ es la región de aceptación de la prueba de nivel α, $\mathbb{P}_0\left(\boldsymbol{X} \in \mathcal{A}(\theta_0)\right) = 1-\alpha$. Luego,

$$\begin{aligned}\mathbb{P}_0\left(\theta_0 \in \mathcal{C}(\boldsymbol{X})\right\} &= \mathbb{P}_0\left\{\boldsymbol{X} : \theta_0 \in \mathcal{C}(\boldsymbol{X})\right\} \\ &= \mathbb{P}_0\left\{\boldsymbol{X} : \boldsymbol{X} \in \mathcal{C}(\theta)\right\} \\ &= 1-\alpha\end{aligned}$$

Entonces, la región $\mathcal{C}(\boldsymbol{X})$ es de nivel de confianza $1-\alpha$. En palabras, *una región de confianza $1-\alpha$ para θ consiste en todos aquellos θ_0 para los que $\boldsymbol{X}$ no rechazaría H_0 con una significancia α.*

Teorema : Suponga que $\mathcal{C}(\boldsymbol{X})$ es ua región de confianza de $1-\alpha$ para θ. Esto es, para todo θ_0, $\mathbb{P}_0\left\{\theta_0 \in \mathcal{C}(\boldsymbol{X})\right) = 1-\alpha$. Entonces, la región de aceptación $\mathrm{H}_0 : \theta = \theta_0$ de nivel α es $\mathcal{A}(\theta_0) = \left\{\boldsymbol{X} : \theta_0 \in \mathcal{C}(\boldsymbol{X})\right\}$.

Demostración: La prueba es de nivel α puesto que

$$\mathbb{P}_0\left(\boldsymbol{X} \in \mathcal{A}(\theta_0)\right) = \mathbb{P}_0\left(\theta_0 \in \mathcal{C}(\boldsymbol{X})\right) = 1-\alpha$$

Esto es, *la hipótesis $H_0 : \theta = \theta_0$ no se rechaza con significancia α. Si θ_0 está en la región de confianza $1-\alpha$.*

Regresando al ejemplo, el intervalo del 95 % de confianza para μ está dado por $(\bar{X} \pm 1.96/5)$, por lo que si el intervalo incluye $\mu = 3.0$ se acepta la hipótesis nula $\mathrm{H}_0 : \mu = 3.0$, con una significancia del 5 %. Si por ejemplo, el promedio observado fuese $\bar{x} = 3.51$, éste da lugar al intervalo del 95 % confianza $(3.12, 3.90)$, que por no contener $\mu_0 = 3.0$, se rechazaría la hipótesis H_0 con una significancia del 5 %.

7.5. Cociente de verosimilitud generalizado (CVG)

7.5.1. Introducción

Considere el modelo $X \sim \mathrm{N}(\mu, \sigma^2)$. Sea $\theta = (\mu, \sigma^2) \in \Theta = \big\{(\mu, \sigma^2) : \infty < \mu < \infty, \sigma > 0\big\}$. Se desea contrastar las hipótesis

$$\mathrm{H}_0 : \mu = 0, \sigma > 0 \quad \textit{vs.} \quad \mathrm{H}_1 : \mu \neq 0, \sigma > 0$$

Sea el **espacio nulo** $\Theta_0 = \big\{(0, \sigma^2) : \sigma > 0\big\}$. Si $\theta_0 = (0, \sigma^2)$, entonces las hipótesis se pueden presentar como

$$\mathrm{H}_0 : \theta \in \Theta_0 \quad \textit{vs.} \quad \mathrm{H}_1 : \theta \in \Theta_1$$

con $\Theta_1 = \Theta \setminus \Theta_0$. Para esto, sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de X. La función de verosimilitud

$$L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}; \theta) = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2\right\}$$

y bajo H_0, $L(\theta_0; \boldsymbol{x}) = (2\pi)^{-n/2}\sigma^{-n}\exp\{-\frac{1}{2\sigma^2}\sum x_i^2\}$. Pero, como σ no está especificada no es posible aplicar el lema de Neyman–Pearson al cociente de verosimilitudes.

Suponga sin embargo que en cada caso, numerador y denominador, se tienen *"buenas muestras"* viniendo de distribuciones con parámetros en Θ_0 y de Θ_1 respectivamente. Ésto ofrecería funciones de verosimilitud "altas". De hecho, se podrían pensar en

$$\frac{\sup_{\theta\in\Theta_0} L(\theta; \boldsymbol{X})}{\sup_{\theta_1\in\Theta_1} L(\theta; \boldsymbol{X})}$$

como cociente de verosimilitudes que podría tomar valores en todo $\mathbb{R}^+$. Por razones técnicas se extiende el denominador a todo Θ. Luego, se define el cociente

$$\Lambda(\boldsymbol{X}) = \frac{\sup_{\theta\in\Theta_0} L(\theta; \boldsymbol{X})}{\sup_{\theta\in\Theta} L(\theta; \boldsymbol{X})} = \frac{L(\hat{\theta}_0; \boldsymbol{X})}{L(\hat{\theta}; \boldsymbol{X})}$$

Notar que $0 \leq \Lambda(\boldsymbol{X}) \leq 1$, para todo $\boldsymbol{X} \in \mathcal{X}$.

Regresando al ejemplo, bajo H_0, en Θ_0, $\ell(\theta; \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum x_i^2$

$$\begin{aligned}
\ell(\theta; \boldsymbol{x}) &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum x_i^2 \\
\frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{2}{2\sigma^3}\sum x_i^2
\end{aligned}$$

que igualando a cero, se sigue que $\hat{\sigma}_0^2 = \frac{1}{n}\sum x_i^2$. Por lo que

$$L(\theta_0; \boldsymbol{x}) = (2\pi)^{-n/2}\left(\frac{1}{n}\sum x_i^2\right)^{-n/2}\exp\left\{-\frac{n}{2}\right\} = \left(\frac{ne^{-1}}{2\pi\sum x_i^2}\right)^{n/2}$$

En el caso general sobre Θ,

$$\begin{aligned}
&\frac{\partial \ell}{\partial \mu} \equiv 0 \quad \Rightarrow \quad \hat{\mu} = \bar{x} \\
&\frac{\partial \ell}{\partial \sigma} \equiv 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2
\end{aligned} \qquad \text{y} \qquad L(\hat{\theta}; \boldsymbol{x}) = \left(\frac{ne^{-1}}{2\pi\sum(x_i - \bar{x})^2}\right)^{n/2}$$

Luego,

$$\Lambda(\boldsymbol{X}) = \frac{L(\hat{\theta}_0; \boldsymbol{X})}{L(\hat{\theta}; \boldsymbol{X})} = \left(\frac{\sum(X_i - \bar{X})^2}{\sum X_i^2}\right)^{n/2} = \left(\frac{\sum(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2 + n\bar{X}^2}\right)^{n/2}$$

Entonces,

$$\Lambda(\boldsymbol{X}) = \left(\frac{1}{1 + \frac{n\bar{X}^2}{\sum(X_i - \bar{X})^2}} \right)^{n/2} < \lambda$$

Ahora bien, si $\mathrm{H}_0 : \mu = 0, \sigma > 0$ *vs.* $\mathrm{H}_1 : \mu \neq 0, \sigma > 0$, valores $\bar{X}$ cercanos a cero apoyarían H_0 y se observarían $\Lambda(\boldsymbol{X})$ cercanos a 1. Por otro lado, si $\mathrm{H}_1 : \mu \neq 0, \sigma > 0$, $\dfrac{n\bar{X}^2}{\sum(X_i - \bar{X})^2} \neq 0$ y la muestra se opone a H_0, con $\Lambda(\boldsymbol{X})$ pequeña. Así, con una región de rechazo de tamaño α, $\mathcal{R}_\alpha = \{\boldsymbol{X} : 0 \leq \Lambda(\boldsymbol{X}) \leq c_\alpha\}$. En este ejemplo se tiene

$$\begin{aligned} \Lambda(\boldsymbol{X}) &\leq \lambda \\ (\lambda^{-n/2} - 1) &\leq \frac{n\bar{X}^2}{\sum(X_i - \bar{X})^2} \\ \sqrt{\frac{n-1}{n}(\lambda^{-n/2} - 1)} &\leq \frac{|\bar{X}|\sqrt{n}}{\sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2}} \end{aligned}$$

Bajo H_0, el estadístico $T(\boldsymbol{X}) = \dfrac{\sqrt{n}\bar{X}}{\sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2}} \sim t_{n-1}$

Finalmente, la región de rechazo de tamaño α, queda

$$\mathcal{R}_\alpha = \left\{ \boldsymbol{X} : \frac{\sqrt{n}|\bar{X}|}{\sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2}} > t_{(1-\alpha/2;\ n-1)} \right\}$$

7.5.2. CVG

Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$ con $\theta \in \Theta = \Theta_0 \cup \Theta_1$. Considere el contraste de hipótesis

$$\mathrm{H}_0 : \theta \in \Theta_0 \quad \textit{vs.} \quad \mathrm{H}_1 : \theta \in \Theta_1$$

Sea $L(\theta; \boldsymbol{x}) = f(\boldsymbol{x};\theta)$, la función de verosimilitud. Se define el **cociente de verosimilitud generalizado (CVG)**

$$\Lambda(\boldsymbol{X}) = \frac{\sup\limits_{\theta \in \Theta_0} L(\theta; \boldsymbol{X})}{\sup\limits_{\theta \in \Theta} L(\theta; \boldsymbol{X})} = \frac{L(\hat{\theta}_0; \boldsymbol{X})}{L(\hat{\theta}; \boldsymbol{X})}$$

Notas:

1. $\Lambda(\boldsymbol{X})$ es el estadístico de prueba.
2. $0 \leq \Lambda(\boldsymbol{X}) \leq 1$.
3. La región de rechazo es $\mathcal{R} = \{\boldsymbol{X} : \Lambda(\boldsymbol{X}) \leq c_0\}$.
4. CVG no siempre es la mejor prueba pero en general será una buena prueba.
5. CVG no necesariamente coincide con la derivada del lema de Neyma-Pearson si las hipótesis son simples.
6. No siempre es fácil encontrar $\sup\limits_{\theta \in \Theta} L(\theta; \boldsymbol{x})$.
7. No siempre es fácil determinar la distribución nula de $\Lambda(\boldsymbol{X})$.

7.5.3. Ejemplo

Tomado de Mood, Graybill, and Boes (1974).

Considere el modelo $X \sim f(x;\theta) = \theta e^{-\theta x}\mathbb{1}_{\mathbb{R}^+}(x)$, con $\theta > 0$. Se desea contrastar las hipótesis

$$\mathrm{H}_0 : \theta \leq \theta_0 \quad vs. \quad \mathrm{H}_1 : \theta > \theta_0$$

Solución: Se considera el espacio parametral $\Theta = \{\theta : \theta > 0\}$ y el espacio nulo $\Theta_0 = \{\theta \leq \theta_0\} \subset \Theta$. Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de tamaño n de $X \sim f$. Luego,

$$L(\theta;\boldsymbol{x}) = \theta^n e^{-\theta\sum x_i}; \quad \ell(\theta;\boldsymbol{x}) = n\log\theta - n\theta\bar{x}; \quad \ell'(\theta;\boldsymbol{x}) = \frac{n}{\theta} - n\bar{x}$$

de donde, el EMV $\hat{\theta} = 1/\bar{X}$. Luego,

i) $L(\hat{\theta};\boldsymbol{x}) = (\bar{x}e)^{-n}$

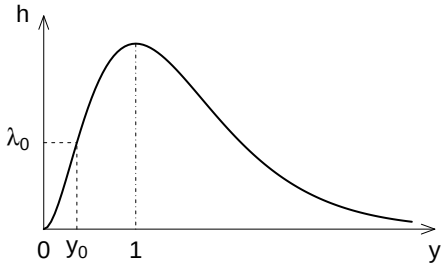
ii) $\sup_{\theta\in\Theta_0} L(\theta;\boldsymbol{x}) = \begin{cases} (\bar{x}e)^{-n} & \text{si } \ 1/\bar{x} \leq \theta_0 \\ \theta_0^n e^{-n\theta_0\bar{x}} & \text{si } \ 1/\bar{x} > \theta_0 \end{cases}$

Entonces,

$$\Lambda(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \ \bar{x} \geq 1/\theta_0 \\ \dfrac{\theta_0^n e^{-n\theta_0\bar{x}}}{(\bar{x}e)^{-n}} & \text{si } \ \bar{x} < 1/\theta_0 \end{cases}$$

Por lo que se rechaza H_0 si $\bar{x} < 1/\theta_0$ y $(\theta_0\bar{x})^n e^{-n(\theta_0\bar{x}-1)} < \lambda_0$. O bien, rechace H_0 si $\theta_0\bar{x} < 1$ y $\left(\theta_0\bar{x}/e^{\theta_0\bar{x}-1}\right)^n < \lambda_0$.

Sea ahora, $y = \theta_0\bar{x}$. Note ahora que $h(y) = y^n e^{-n(y-1)}$ tiene un máximo en $y = 1$, como se muestra en la figura de la derecha.



Entonces, la regla de decisión es: rechazar H_0 si $y < 1$ y $h(y) = y^n e^{-(y-1)} < \lambda_0$. Equivalentemente, si $y < y_0$ y $y < 1$, o bien, si $\bar{x} < 1/\theta_0$ y $\bar{x} < y_0/\theta_0$. Recuerde, si $X \sim \mathrm{Exp}(\theta)$, $\sum_{i=1}^n X_i \sim \frac{1}{2\theta}\mathrm{Ga}(\alpha = 2n/2, \beta = 2) \equiv \frac{1}{2\theta}\chi^2_{2n}$. Luego, si la prueba es de significancia α, $\alpha = \mathbb{P}_0(\bar{X} < y_0/\theta_0)$, por lo que $y_0 = \chi^2(\alpha; 2n)$. Por lo tanto, la región de rechazo para esta prueba $\mathrm{H}_0 : \theta < \theta_0$ *vs.* $\mathrm{H}_1 : \theta > \theta_0$ y de significancia α es,

$$\mathcal{R}_\alpha = \left\{ \boldsymbol{X}_n : \bar{X} < \frac{\chi^2(\alpha; 2n)}{2n\theta_0} \right\}$$

Si por ejemplo, $X \sim \mathrm{Exp}(\theta)$ y de desea contrastar $\mathrm{H}_0 : \theta \leq 2.0$ *v. a.* $\mathrm{H}_1 : \theta > 2.0$. Una prueba de significancia del 10 % considerando muestras de tamaño $n = 20$, tendría un valor crítico de $c_{0.10} = \frac{\chi^2(0.10;\ 2\cdot 20)}{2(20)(2)} = \frac{29.05}{80} = 0.363$. Por lo tanto,

$$\mathcal{R}_{0.10} = \left\{ \boldsymbol{X}_{20} : \bar{X} < 0.363 \right\}$$

Luego, en muestras de tamaño 20, promedios de 0.35 llevarían a rechazar $\mathrm{H}_0 : \theta \leq 2.0$ con una significancia del 10 %, mientras que si $\bar{x} = 0.42$ no se rechazaría la hipótesis.

Los correspondientes valores-p, son el resultado de resolver la ecuación $\bar{x} = \frac{\chi^2(p;2n)}{2n\theta_0}$. Por lo que, el $\mathbb{P}\left(\chi^2_{40} \leq 0.35(2)(20)(2)\right) = 0.077$ y $\mathbb{P}\left(\chi^2_{40} \leq 0.42(2)(20)(2)\right) = 0.248$, son los correspondientes valores-p. Esto es, 0.35 (0.077) y 0.42 (0.248), valores menor y mayor que el valor crítico $c = 0.363$, correspondiente a la significancia de 0.10. La siguiente tabla muestra distintos valores del estadístico de prueba $\bar{X}$, los valores-p respectivos y las respectivas decisiones considerando un nivel de significancia $\alpha = 0.10$.

$\bar{x}$	0.29	0.31	0.35	**0.363**	0.39	0.42
valor-p	(0.016)	(0.029)	(0.077)	**(0.10)**	(0.161)	(0.248)
Decisión	Rechazar H_0				Aceptar H_0	

Note que mientras más chicos los valores-p, mayor sería la evidencia asociada en contra de la hipótesis nula H_0.

Ahora bien, para calcular la potencia de la prueba para distintos valores de θ, recuerde

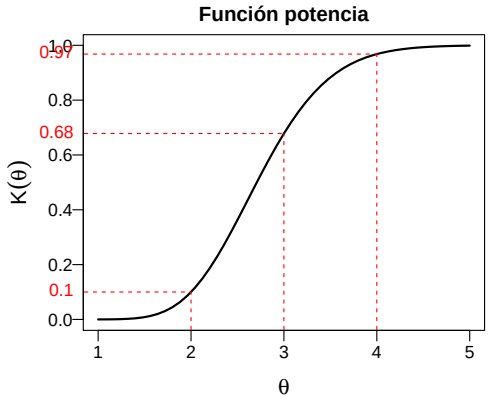
$$K(\theta) = \mathbb{P}_\theta(\mathcal{R}_\alpha) = \mathbb{P}_\theta(\bar{X} < c_\alpha)$$

y que $\sum_{i=1}^n X_i \sim \frac{1}{2\theta}\chi^2_{2n}$. Por lo que se tiene que

$$K(\theta) = \mathbb{P}_\theta\left(\sum X_i \leq nc_\alpha\right) = \mathbb{P}\left(2\theta \sum X_i \leq 2n\theta c_\alpha\right) = \mathbb{P}\left(\chi^2_{2n} \leq 2n\theta c_\alpha\right)$$

La siguiente tabla muestra la potencia de la prueba para varios valores del parámetro θ. La figura muestra la función potencia correspondiente.

θ	$K(\theta)$
1.0	0.0001
1.5	0.0084
2.0	0.1000
3.0	0.6782
4.0	0.9680
5.0	0.9988



Finalmente, note una vez más que encontrar la distribución nula de $\Lambda(\boldsymbol{X})$ puede ser difícil, sino es que imposible. Pero en ocasiones como en este ejemplo, es posible resolver el caso.

7.5.4. Distribución asintótica del cociente de verosimilitudes generalizado (CVG)

Con el apoyo de Casella and Berger (2002) y Knight (2000).

Teorema (H_0 simple) Considere el modelo $X \sim f(x;\theta)$, con $\theta \in \Theta$, subconjunto abierto de $\mathbb{R}$. Se desea contrastar las hipótesis $H_0 : \theta = \theta_0$ *vs.* $H_1 : \theta \neq \theta_0$.

Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim f(x;\theta)$. Sea $\hat{\theta}_n = \hat{\theta}(\boldsymbol{X}_n)$ el estimador máxima verosimilitud de θ. Suponga que f satisface las condiciones de regularidad (CR) enunciadas en la sección de estimación. Entonces, bajo H_0,

$$-2\log\Lambda(\boldsymbol{X}_n) \xrightarrow{\text{D}} \chi^2_1$$

Demostración: Note que la expansión por Taylor de $\ell(\theta;\boldsymbol{x}) = \log L(\theta;\boldsymbol{x})$ alrededor de $\hat{\theta}$ es

$$\ell(\theta;\boldsymbol{x}) = \ell(\hat{\theta};\boldsymbol{x}) + \ell'(\hat{\theta};\boldsymbol{x})(\theta - \hat{\theta}) + \frac{1}{2}\ell''(\theta^*;\boldsymbol{x})(\theta - \hat{\theta})^2$$

donde θ^* está entre θ y $\hat{\theta}$. Entonces,

$$\begin{aligned}
-2\log(\Lambda(\boldsymbol{X})) &= -2\ell(\theta_0;\boldsymbol{X}) + 2\ell(\hat{\theta};\boldsymbol{X}) \\
&= -2\left[\ell(\hat{\theta};\boldsymbol{X}) + \ell'(\hat{\theta};\boldsymbol{X})(\theta_0-\hat{\theta}) + \frac{1}{2}\ell''(\theta^*;\boldsymbol{X})(\theta_0-\hat{\theta})^2\right] + 2\ell(\hat{\theta};\boldsymbol{X}) \\
&= -\ell''(\theta^*;\boldsymbol{X})(\hat{\theta}-\theta_0)^2; \qquad \text{pues } \ell'(\hat{\theta};\boldsymbol{X}) = 0 \\
&= n(\hat{\theta}-\theta_0)^2\left[-\frac{1}{n}\sum_{i=1}^{n}\ell''(\theta^*;X_i)\right] \\
&\approx n(\hat{\theta}-\theta_0)^2\left(-\mathbb{E}_{\theta_0}[\ell''(\theta^*;X)]\right); \qquad \text{por LGN} \\
&= n(\hat{\theta}-\theta_0)^2\mathrm{I}(\theta_0) \\
&= \left[\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\sqrt{\mathrm{var}(\hat{\theta})}}\right]^2
\end{aligned}$$

Y como se mostró ya, $\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\sqrt{\mathrm{var}(\hat{\theta})}} \xrightarrow{D} \mathrm{N}(0,1)$, se sigue que

$$-2\log\Lambda(\boldsymbol{X}_n) \xrightarrow{\mathrm{D}} \chi^2_1$$

Teorema . Considere el modelo $X \sim f(x;\boldsymbol{\theta})$, con $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k) \in \Theta$, subconjunto abierto de $\mathbb{R}^k$. Se desea probar la hipótesis $\mathrm{H}_0 : \theta_1 = \theta_{10},\ldots,\theta_r = \theta_{r0}$, con $r \leq k$.

Suponga que se satisfacen las condiciones de regularidad (CR). Si $\hat{\boldsymbol{\theta}}_n$, el estimador máximo verosimilitud de $\boldsymbol{\theta}$, satisface que $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \dot{\sim} \mathrm{N}_k\left(\mathbf{0}, \mathrm{I}^{-1}(\boldsymbol{\theta})\right)$. Entonces, bajo H_0, el estadístico del cociente de verosimilitud generalizado, $\Lambda(\boldsymbol{X}_n)$ sigue una distribución χ^2_r. Esto es,

$$-2\log\Lambda(\boldsymbol{X}_n) \xrightarrow{D} \chi^2_r$$

En palabras, r es la diferencia en el número de *parámetros libres* de cada hipótesis.

Ejemplo : Considere los modelos independientes $X \sim \mathrm{N}(\mu_1,\sigma_1^2)$ y $Y \sim \mathrm{N}(\mu_2,\sigma_2^2)$. Se desea contrastar las hipótesis $\mathrm{H}_0 : \mu_1 = \mu_2$ *v. a.* $\mathrm{H}_1 : \mu_1 \neq \mu_2$, pero no se conocen las varianzas. Para efectos del ejemplo, se define el vector de parámetros $\boldsymbol{\theta} = (\mu_1-\mu_2, \mu_2, \sigma_1, \sigma_2)$. Considere entonces el espacio parametral $\Theta = \{\boldsymbol{\theta} = (\mu_1-\mu_2, \mu_2, \sigma_2, \sigma_2) : \mu_i \in \mathbb{R}, \sigma_i \in \mathbb{R}^+, i = 1,2\}$, $\Theta_0 = \{\boldsymbol{\theta} = (0, \mu_2, \sigma_1, \sigma_2)\} \subset \Theta$ y $\Theta_1 = \Theta \setminus \Theta_0$. Luego, las hipótesis pueden plantearse

$$\mathrm{H}_0 : \theta \in \Theta_0 \quad \textit{vs.} \quad \mathrm{H}_1 : \theta \in \Theta_1$$

Solución: Como se discutió anteriormente, desconociendo las varianzas el problema, conocido como Beherens–Fisher, no tiene una solución al desconocerse la distribución exacta. Una solución sería construir un intervalo de confianza para θ_1, utilizando la aproximación de Welch o Satterwhite, como se comentó en la sección anterior.

Alternativamente, se puede emplear la prueba del cociente de verosimilitud generalizado. Para esto, se consideran las muestras aleatorias $\boldsymbol{X} = (X_1,\ldots,X_{n_1})$ de X y $\boldsymbol{Y} = (Y_1,\ldots,Y_{n_2})$ de Y. La función de verosimilitud

$$\begin{aligned}
L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) &= f(\boldsymbol{x};\mu_1,\sigma_1)\cdot f(\boldsymbol{y};\mu_2,\sigma_2) \\
&= (2\pi)^{(n_1+n_2)/2}\sigma_1^{-n_1}\sigma_2^{-n_2}\exp\left\{-\frac{1}{2\sigma_1^2}\sum(x_i-\mu_1)^2 - \frac{1}{2\sigma_2^2}\sum(y_j-\mu_2)^2\right\}
\end{aligned}$$

El estadístico de prueba es

$$\Lambda(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sup\limits_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Y})}{\sup\limits_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Y})} = \frac{L(\hat{\boldsymbol{\theta}}_0;\boldsymbol{X},\boldsymbol{Y})}{L(\hat{\boldsymbol{\theta}};\boldsymbol{X},\boldsymbol{Y})}$$

Para determinar maximizar $L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y})$ en Θ_0 se requerirá de un método numérico por la restricción impuesta por la hipótesis H_0.

De acuerdo a los teoremas presentados anteriormente,

$$-2\log\Lambda(\boldsymbol{X},\boldsymbol{Y}) \xrightarrow{D} \chi^2_\nu$$

En este caso, $\nu = 4-3 = 1$, luego $-2\log\Lambda(\boldsymbol{X},\boldsymbol{Y}) \dot{\sim} \chi^2_1$. Luego, dada las muestras $\boldsymbol{x}$ y $\boldsymbol{y}$, se calcula $-2\log\Lambda(\boldsymbol{x},\boldsymbol{y})$ y determina su valor$-p$ de acuerdo a la distribución χ^2_1 y se toma la decisión sobre H_0.

7.5.5. Prueba Ji-cuadrada para bondad de ajuste

Con el apoyo de Rice (2007).

Sean $\{\ell_i\}$ sujetos que son asignados a una y solo una de k categorías $C_1,\ldots,C_k$, exclusivas y exhaustivas, de manera que todo ℓ queda clasificado en una solo una C_j. Sea $p_j = \mathbb{P}(\ell \in C_j) \geq 0$, para todo $j = 1,\ldots,k$ y por lo que $1 = p_1 \cdots + p_k$.

Sea $\Theta \subset \mathbb{R}^r$ el espacio de parámetros tales que $p_j = p_j(\theta)$, para $\theta \in \Theta$. Sean $\Theta_0 \subset \Theta$ el espacio nulo y $\Theta_1 = \Theta \setminus \Theta_0$. Se desea contrastar las hipótesis

$$\mathrm{H}_0 : \theta \in \Theta_0 \quad vs. \quad \mathrm{H}_1 : \theta \in \Theta_1$$

Para esto, considere n sujetos ℓ's independientes y sea $\boldsymbol{X} = (X_1,\ldots,X_k)$ con X_j el número de sujetos que fueron clasificados C_j. Luego, $n = X_1 + \cdots + X_k$. $\boldsymbol{X}$ sigue una distribución multinomial de parámetros n y $p_1,\ldots,p_k$. Entonces, la función de verosimilitud de $\boldsymbol{p}$ es

$$L(\boldsymbol{p};\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{p}) = \binom{n}{x_1\cdots x_k} p_1^{x_1}\cdots p_k^{x_k}$$

Se sigue del principio de invarianza de los estimadores de máxima verosimilitud que

$$\sup_{\boldsymbol{p}\in\mathrm{H}_0} L(\boldsymbol{p};\boldsymbol{x}) = L(\boldsymbol{p}(\hat{\theta});\boldsymbol{x})$$

donde $\hat{\theta}$ es el EMV de $\theta \in \Theta_0$. Por otro lado, recuerde que sin restricciones sobre θ, los EMV de las p_j's es $\hat{p}_j = \dfrac{x_j}{n}$. Por lo que el cociente de verosimilitud generalizado (CVG) es

$$\Lambda(\boldsymbol{X}) = \frac{\binom{n}{x_1\cdots x_k} p_1(\hat{\theta})^{x_1}\cdots p_k(\hat{\theta})^{x_k}}{\binom{n}{x_1\cdots x_k}\hat{p}_1^{x_1}\cdots\hat{p}^{x_k}} = \prod_{j=1}^{k}\left(\frac{p_j(\hat{\theta})}{\hat{p}_j}\right)^{x_j}$$

y puesto que $x_j = n\hat{p}_k$, se tiene que

$$\begin{aligned} -2\log\Lambda(\boldsymbol{X}) &= -2\sum_{j=1}^{k}\log\left(\frac{p_j(\hat{\theta})}{\hat{p}_j}\right)^{n\hat{p}_j} \\ &= -2\sum n\hat{p}_j\log\left(\frac{n\hat{p}_j(\hat{\theta})}{n\hat{p}_j}\right) \\ &= 2\sum_{j=1}^{k} O_j\log\left(\frac{O_j}{E_j}\right) \end{aligned} \tag{15}$$

donde $O_j = n\hat{p}_j$ y $E_j = np_j(\hat{\theta})$.

Ahora, puesto que $1 = p_1 + \cdots + p_k$, se tienen $k-1$ parámetros libres. Si bajo H_0 las $p_j(\hat{\theta})$ dependen del espacio nulo Θ_0, de dimensión m, esto es, $\dim(\Theta_0) = m$, entonces, $-2\log\Lambda(\boldsymbol{X}) \sim \chi^2_{k-1-m}$.

Por otro lado, existe el **estadístico X^2 de Pearson** (Karl), con

$$\mathrm{X}^2 = \sum_{j=1}^{k} \frac{\Big(x_j - np_j(\hat{\theta})\Big)^2}{np_j(\hat{\theta})} \xrightarrow{D} \chi^2_{k-1-m} \tag{16}$$

bajo H_0. Entonces, bajo la hipótesis H_0, el cociente de verosimilitudes generalizado y el estadístico X^2 de Pearson son asintóticamente equivalentes.

En efecto, si H_0 es verdadero y n es "grande", $\hat{p}_j \approx p_j(\hat{\theta})$ y expandiendo por Taylor $-2\log\Lambda(\boldsymbol{X})$ al rededor de $p_j(\hat{\theta})$,

$$\begin{aligned} -2\log\Lambda(\boldsymbol{X}) &= 2n\sum(\hat{p}_j - p_j(\hat{\theta})) + n\sum \frac{(\hat{p}_j - p_j(\hat{\theta}))^2}{p_j(\hat{\theta})} + \cdots \\ &\approx \sum_{i=1}^{k} \frac{\Big(x_j - np_j(\hat{\theta})\Big)^2}{np_j(\hat{\theta})} \end{aligned}$$

pues la primera suma es aproximadamente cero, $x_j = n\hat{p}_j$ e ignorando términos de mayor grado.

En la práctica es más común el uso del estadístico X^2 de Pearson pues su fácil de calcular y pocas veces requiere de computadora.

Ejemplo : *Modelo de equilibrio de Hardy-Weinberg*
Recupere el ejemplo de la Ley Hardy-Weinberg visto en la sección de estimadores. De acuerdo al modelo, si las frecuencias de los genotipos están en equilibrio, los genotipos AA, Aa, y aa, ocurren con una frecuencia mostrada en la tabla

g	AA	Aa	aa
f	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2

En la siguiente tabla se muestra la distribución en el tipo de sangre de 1029 personas, probabilidades estimadas, y la cantidad esperada de acuerdo al modelo

genotipo	M	MN	N	Total
O_j	342	500	187	1029
$p_j(\hat{\theta})$	0.331	0.489	0.180	1.0
E_j	340.6	502.8	185.6	1029

Como fue ya discutido, a partir de la muestra, el EMV $\hat{\theta} = 0.4247$. Con ella, se estiman las probabilidades $p_j(\hat{\theta})$ y los correspondientes observaciones esperadas $np_j(\hat{\theta})$. Por ejemplo, para el genoma AA, $p_1(\hat{\theta}) = (1-0.4247)^2 = 0.331$ y $E_1 = 1029(0.331) = 340.6$.

Se considera la hipótesis nula H_0 de que el modelo multinomial propuesto es adecuado, mientras que la alternativa H_1 es que las probabilidades siguen alguna otra relación. Se desean contrastar las hipótesis con una prueba de significancia α. El estadístico X^2 de Pearson sigue aproximadamente una distribución χ^2 con 1 (= 3 categorías -1 -estimación de θ) grado de libertad. Luego, la correspondiente región de rechazo será $\mathcal{R}_\alpha = \{\boldsymbol{X} : \mathrm{X}^2 > \chi^2(1-\alpha;1)\}$. Que si la significancia es $\alpha = 0.05$, entonces $\mathcal{R}_{0.05} = \{\mathrm{X}^2 \geq 3.841\}$. Ahora bien, de (16),

$$\mathrm{X}^2 = \frac{(342-340.6)^2}{340.6} + \frac{(500-502.8)^2}{502.8} + \frac{(187-185.6)^2}{185.6} = 0.0325 \ (0.857)$$

Por lo que no se rechaza la hipótesis H_0 con una significancia del 5 %. Pero note, el estadístico observado 0.032 es bastante menor que el valor crítico 3.84. Esto se ve reflejado en su correspondiente valor-p de 0.857.

Por otro lado, si se calcula el estadístico de prueba (15) derivado del CVG

$$-2\log\Lambda(\boldsymbol{X}) = -2\left[342\log\left(\frac{342}{340.6}\right)+\cdots\right] = 0.0325 \ (0.857)$$

Finalmente, note que el cociente de verosimilitudes es $\Lambda(\boldsymbol{X}) = 0.9839$, apoyando claramente la hipótesis nula.

Ejemplo : Recupere el ejemplo de *ajuste de datos Poisson* visto al principio de la sección de Estimadores. Ahí se presenta otro ejemplo de la prueba χ^2 de bondad de ajuste y aunque no es nombrado, el estadístico X^2 de Pearson y el valor-p.

7.6. Bondad de Ajuste

7.6.1. Gráficas de Probabilidad

Con el apoyo de Rice (2007).

Las gráficas de probabilidades son usadas para pruebas visuales de bondad del ajuste por las distribuciones de probabilidad. Refiérase nuevamente al ejemplo de los datos Poisson de emisión de partículas del principio de la sección de Estimadores. ¿Se pueden considerar los datos generados por una distribución Poisson? O bien, suponga que los siguientes datos (simulados) representan una muestra del consumo mensual de agua de 50 alumnos. ¿Qué distribución siguen éstos datos? Suponer que la distribución normal aproxima razonablemente a las observaciones, ¿se cumple? Este es el problema de *qué tan bien el modelo (distribución) ajusta los datos.*

2.55	2.20	2.62	1.71	3.63	2.33	4.74	3.73	3.70	2.98	3.46	2.70
2.34	3.23	3.42	2.17	3.26	4.70	3.04	2.12	2.53	3.15	3.20	1.98
3.07	3.69	3.69	3.77	3.72	3.22	3.30	3.05	4.10	3.82	3.50	3.45
3.37	3.90	3.82	3.13	1.93	2.20	4.36	3.52	4.27	4.51	3.81	6.02
3.07	2.81										

Suponga $X_1,\ldots,X_n$ una muestra aleatoria de una distribución uniforme $[0,1]$. Sean $Y_j = X_{(j)}$, los correspondiente estadísticos de orden. Entonces, se sabe[26] que $\mathbb{E}[X_{(j)}] = \dfrac{j}{n+1}$, lo que sugiere graficar la muestra ordenada $\{X_{(i)}\}$ contra los valores esperados $\frac{1}{n+1},\cdots,\frac{n}{n+1}$. La gráfica seguirá mas o menos una línea recta como lo muestra el panel de la izquierda de la figura 25. En el panel de la derecha se muestra la gráfica de la muestra ordenada de una distribución triangular (suma de uniformes) contra los cuantiles de la distribución uniforme. El evidente la separación de los puntos a la línea recta, sugiriendo que la distribución uniforme no ajusta razonablemente los datos triangulares.

Ahora bien, el teorema de transformación integral[27] afirma que si la variable aleatoria X tiene una función de distribución F, continua estrictamente creciente, entonces $F(X) \sim$ Unif$[0,1]$ y se podría graficar $Y_j = F(X_{(j)})$ contra $\frac{j}{n+1}$, o bien, $X_{(j)}$ *vs.* $F^{-1}(\frac{j}{n+1})$. El panel izquierdo de la figura 26 muestra la gráfica de $X_{(j)}$ contra $F^{-1}(\frac{j}{n+1})$ de una distribución gamma. Nótese nuevamente que los puntos siguen una línea recta, que no en el panel de la derecha, puntos simulados también de una distribución gamma pero de distintos parámetros que la de la izquierda. Ambas gráficas tiene el mismo eje horizontal, $F^{-1}(\frac{j}{n+1})$, correspondiente a la distribución de la izquierda.

[26] Cálculo de Probabilidades II.
[27] Cálculo de Probabilidades II.

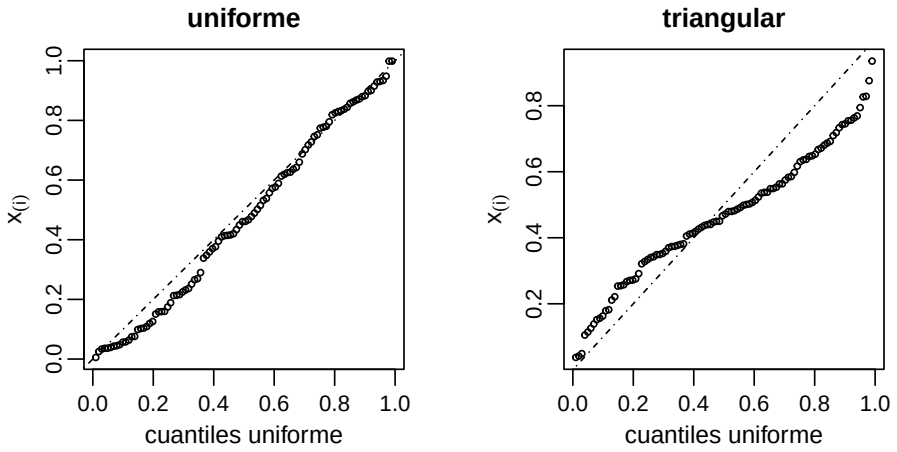


Figura 25: Datos simulados de: a) distribución uniforme; b) distribución triágular. Eje horizontal en cuantiles uniformes $[0, 1]$.

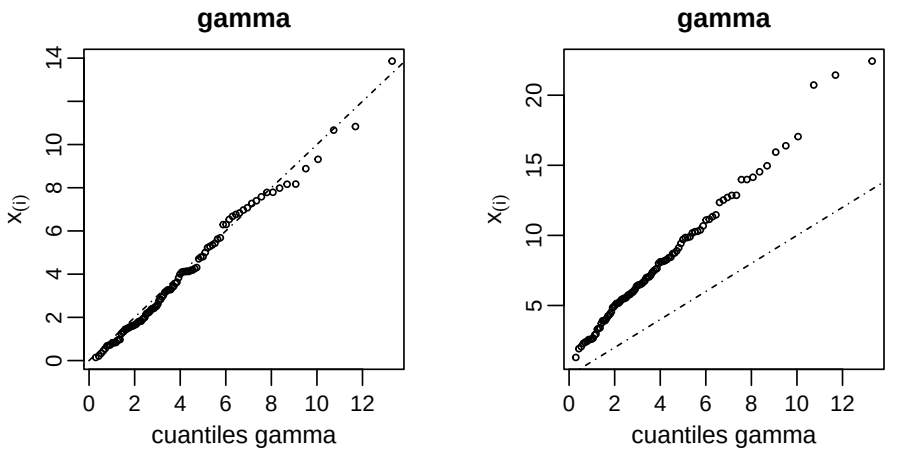


Figura 26: Datos simulados de: a) distribución gamma; b) distribución gamma distintos parámetros. Eje horizontal en cuantiles de la distribución con los parámetros del panel izquierdo.

En algunos casos, la distribución $F(x) = G\left(\frac{x-\mu}{\sigma}\right)$, con μ y σ llamados parámetros de localización y escala, como es el caso de la distribución normal. Si $X \sim \mathrm{N}(\mu, \sigma)$, $F_X = \Phi\left(\frac{x-\mu}{\sigma}\right)$, donde Φ es la función de distribución de la normal estándar. Luego, se puede graficar $\dfrac{X_{(j)} - \mu}{\sigma}$ *vs.* $G^{-1}\left(\dfrac{j}{n+1}\right)$, o bien, $X_{(j)}$ *vs.* $\mu + \sigma G^{-1}\left(\dfrac{j}{n+1}\right)$. El panel izquierdo de la gráfica 27 muestra datos normales ($\mathrm{N}(-3, 2^2)$) contra los correspondientes cuantiles. El panel de la derecha muestra los mismos datos graficados con la función `qqnorm` del lenguaje R. Note que la escala del eje horizontal corresponde a los cuantiles del la distribución normal estándar. Si los puntos siguen más o menos una línea recta se pueden suponer que la distribución normal describe razonablemente los datos.

La figura 28 muestra la gráfica cuantil-cuantil normal de los datos simulados del consumo mensual de agua, del principio de esta sección. Los puntos siguen más o menos una línea recta (exhibida) por lo que se podría suponer que siguen una distribución normal. Por otro lado, al mostrar los cuantiles de la distribución normal estándar en el eje horizontal, de la gráfica se puede estimar la media y la desviación estándar de la distribución origen de los

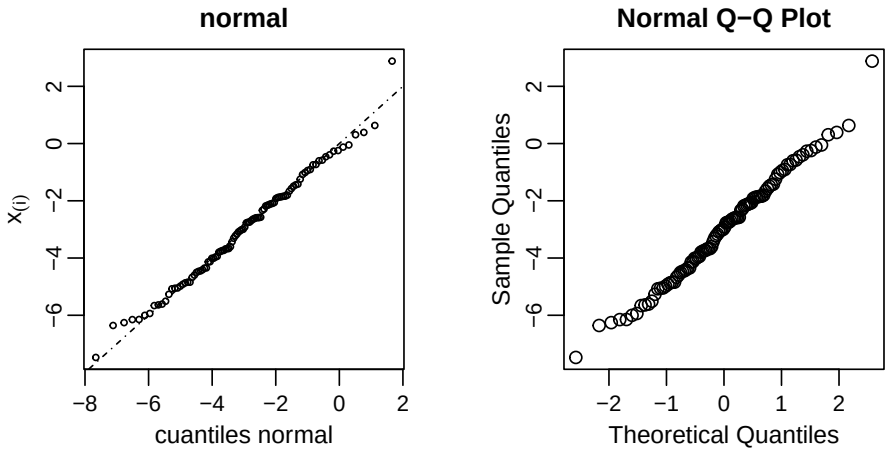


Figura 27: Datos simulados de distribución normal. Panel izquierdo es construido con los cuantiles correspondientes. Panel derecho obtenido con la función `qqnorm` de R.

datos.

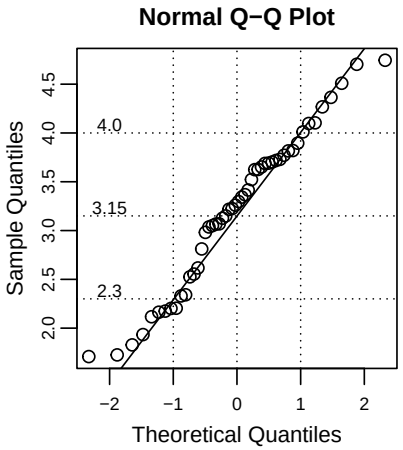


Figura 28: Gráfica cuantil-cuantil normal de los datos simulados del consumo mensual de agua.

A saber, localice las abscisas -1, 0 y 1 sobre la recta que ajusta los datos y proyéctelos sobre la escala vertical. Los valores son aproximadamente 2.3, 3.15 y 4.0. La media de la normal estándar es 0, la media estimada de la distribución será $\tilde{\mu} = 3.15$. Entre ± 1 se tiene 2 veces la desviación estándar de la distribución estándar, luego $\tilde{\sigma} = (4.0 - 2.3)/2 = 0.85$.

7.6.2. Función de distribución empírica

Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim F$ y sean $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ los correspondientes estadísticos de orden, $Y_j = X_{(j)}$, para $j = 1, \ldots, n$.

Definición : Se define la **función de distribución empírica** (*f.d.e.*; *e.c.d.f.* por *empirical cumulative distribution function*) por

$$F_n(x) = \frac{1}{n} \text{máx}\{i : Y_i \leq x\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, x)}(X_i)$$

Note que:

- Para $x \in \mathbb{R}$ fijo, $F_n(x)$ es un estadístico (función de la muestra que no depende de parámetros desconocidos).
- F_n se distribuye como el promedio de n ensayos Bernoulli (Binomial/n).

Teorema : $F_n(x)$ denota la función de distribución empírica de una muestra aleatoria de Bernoulli's tamaño n, entonces,

$$\mathbb{P}\left(F_n(x) = \frac{k}{n}\right) = \binom{n}{k} \left[F(x)\right]^k \left[1 - F(x)\right]^{n-k}, \qquad k = 0, \ldots, n$$

Demostración: Sean $Z_i = \mathbb{1}_{(-\infty,x]}(X_i)$. Entonces, $Z_i \sim \text{Ber}(F(x))$, por lo que $\sum Z_i \sim \text{Bin}(n, F(x))$ y $F_n(x) = \frac{1}{n}\sum Z_i$.

Corolario : $\boldsymbol{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim F$. $F_n(x) = \frac{1}{n}\sum \mathbb{1}_{(-\infty,x]}(X_i)$. Entonces,

i) $\mathbb{E}[F_n(x)] = F(x)$

ii) $\text{var}(F_n(X)) = \frac{1}{n}F(x)\left[1 - F(x)\right]$

De hecho, puesto F_n es una media muestral. se sigue del Teorema Central de Límite que

$$F_n(x) \dot{\sim} \text{N}\left(F(x), \frac{1}{n}F(x)(1 - F(x))\right)$$

Además, se sigue del corolario anterior que para x fijo, $F_n(x)$ es un estimador insesgado y consistente (ECM) de $F(x)$.

Las propiedades anteriores son para cualquier función de distribución F.

Para la estimación de F y no solo de $F(x)$, se ha de ver la distancia de $F_n(x)$ a $F(x)$ para todo $x \in \mathbb{R}$.

Teorema de Glivenko-Cantelli .

$$\mathbb{P}\left(\sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right| \longrightarrow 0\right) = 1$$

La convergencia *casi segura (cp1)* de $F_n(x)$ a $F(x)$ es uniforme en x.

Sea $D_n = \sup_{x\in\mathbb{R}} \left|F_n(x) - F(x)\right|$. Entonces, D_n es aleatorio y mide la distancia de F_n a F y el Teorema de Glivenko-Cantelli indica que $\mathbb{P}(D_n \to 0) = 1$. En particular, la función de probabilidad acumulada de D_n, F_{D_n} tiene toda su masa en cero. ¿Qué pasa con $F_{\sqrt{n}D_n}$?

El Teorema de Glivenko-Cantelli dice que la convergencia casi segura de $F_n(x)$ a $F(x)$ es uniforme en x. Si $x < y$, ¿cómo es la estimación de $F(y) - F(x) = \mathbb{P}(x < X \leq y)$?

Proposición :

$$\text{cov}(F_n(x), F_n(y)) = \frac{1}{n}F(x)[1 - F(y)], \quad x \leq y$$

Demostración: Sea $x < y$, luego

$$\begin{aligned}
\operatorname{cov}\big(F_n(x), F_n(y)\big) &= \operatorname{cov}\left(\frac{1}{n}\sum_i \mathbb{1}_{(\infty,x)}(X_i), \frac{1}{n}\sum_j \mathbb{1}_{(-\infty,y)}(X_j)\right) \\
&= \left(\frac{1}{n}\right)^2 \sum_i \sum_j \operatorname{cov}\Big(\mathbb{1}_{(-\infty,x)}(X_i), \mathbb{1}_{(-\infty,x)}(X_j)\Big) \\
&= \left(\frac{1}{n}\right)^2 \sum_i \operatorname{cov}\Big(\mathbb{1}_{(-\infty,x)}(X_i), \mathbb{1}_{(-\infty,y)}(X_j)\Big) \\
&= \frac{1}{n}\operatorname{cov}\Big(\mathbb{1}_{(-\infty,x)}(X_1), \mathbb{1}_{(-\infty,y)}(X_1)\Big) \\
&= \frac{1}{n}\Big[\mathbb{E}[\mathbb{1}_{(-\infty,x)}(X)\mathbb{1}_{(-\infty,y)}(X)] - \mathbb{E}[\mathbb{1}_{(-\infty,x)}(X)]\cdot\mathbb{E}[\mathbb{1}_{(-\infty,y)}(X)]\Big] \\
&= \frac{1}{n}\big[F_n(x) - F(x)F(y)\big] \\
&= \frac{1}{n}F(x)[1-F(y)]
\end{aligned}$$

Corolario :

$$\begin{aligned}
\operatorname{var}(F_n(y) - F_n(x)) &= \operatorname{var}(F_n(y)) + \operatorname{var}(F_n(x)) - 2\operatorname{cov}\big(F_n(y), F_n(x)\big) \\
&= \frac{1}{n}[F(y) - F(x)][1 - F(y) + F(x)]
\end{aligned}$$

Corolario : $\big(F_n(y) - F_n(x)\big)$ es un estimador consistente cuadrático medio de $(F(y) - F(x))$.

Proposición : Si $B \in \mathcal{B}(\mathbb{R})$, entonces, $\mathbb{P}_n(B) = \frac{1}{n}\mathbb{1}_B(X_i)$ es un estimador insesgado y consistente (cuadrático medio) de $\mathbb{P}(X \in B)$.

Demostración: Note que

$$\operatorname{var}(\mathbb{P}_n(X \in B)) = \operatorname{var}\left(\frac{1}{n}\sum \mathbb{1}_B(X_i)\right) = \frac{1}{n}\mathbb{P}(X \in B)[1 - \mathbb{P}(X \in B)]$$

7.6.3. Prueba Kolmogorov-Smirnov

Con base en Mood, Graybill, and Boes (1974).

Recordar que $F_n(x)$ sigue se distribuye asintóticamente normal.

$$\sqrt{n}\left[F_n(x) - F(x)\right] \xrightarrow{D} \mathrm{N}(0, F(x)(1-F(x)))$$

El siguiente teorema establece la distribución límite de $\sqrt{n}D_n$ definido anteriormente.

Teorema : Sea $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ una muestra aleatoria de $X \sim F$, continua. Sea

$$D_n = D(\boldsymbol{X}_n) = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

Entonces, la distribución asintótica de $\sqrt{n}D_n$ está dada por

$$\begin{aligned}
H(x) &:= \lim_{n\to\infty} F_{\sqrt{n}D_n}(x) \\
&= \lim_{x\to\infty} \mathbb{P}(\sqrt{n}D_n \le x) \\
&= \left[1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{2j^2x^2}\right]\mathbb{1}_{\mathbb{R}^+}(x)
\end{aligned}$$

Notas:

1. La distribución no depende de F, es decir, la distribución de $\sqrt{n}D_n$ es *libre de distribución*, ("no paramétrica"). Lo que hace a D_n muy práctica para las *pruebas de bondad de ajuste*.

2. EL resultado se debe a Andrey Kolmogorov (1933). Tablas de la distribución H se deben a Nicolai Smirnov (1948).

3. Su uso para pruebas de bondad de ajuste:

$$\mathrm{H}_0 : F = F_0 \quad vs. \quad \mathrm{H}_1 : F \neq F_0$$

Bajo H_0, sea $K_n = K(\boldsymbol{X}_n) = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \sim H$, la distribución de $\sqrt{n}D_n$. Si H_0 es falsa, F_n tendera a la verdadera F y no a F_0 y por lo tanto $\sup_{x\in\mathbb{R}} |Fn(x) - F_0(x)|$ tenderá a ser grande. Luego, una regla de decisión razonable (región de rechazo de tamaño α) sería

$$\mathcal{R}_\alpha = \left\{ \boldsymbol{X}_n : K_n := \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > c_\alpha \right\}$$

con el valor crítico $c_\alpha = H(1-\alpha; n)$, disponible en tablas.

La prueba anterior se conoce como la **prueba de bondad de ajuste de Kolmogorov-Smirnov (KS)**. KS valora que "tan bien" un conjunto de datos es ajustado por la distribución F_0. El ajuste "se mide" mediante el **estadístico de Kolmogorov**

$$D_n = D(\boldsymbol{X}) = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

El teorema indica la distribución asintótica de D_n. Hay tablas de probabilidad acumulada de H_n para distintos tamaños de muestra n[28].

Notas:

1. La distribución de H_n supone F_0 conocida, luego H_0 es simple. Si hay necesidad de estimar algún parámetro la distribución ya no es válida.

2. La adaptación de KS a $F_0 \equiv \mathrm{N}(\mu, \sigma^2)$ con $\boldsymbol{\theta} = (\mu, \sigma)$ desconocida se debe a H. Lilliefors (1967).

3. La figura 29 muestra un ejemplo de una distribución nula, la empírica y dónde se da la máxima D_n que define el estadístico K_n observado.

Ejemplo :

El siguiente arreglo presenta 40 observaciones que se dice representa una muestra aleatoria de una distribución uniforme en $[0, 1]$.

0.178	0.550	0.797	0.576	0.255	0.618	0.517	0.598	0.556	0.599
0.621	0.518	0.419	0.292	0.479	0.609	0.631	0.734	0.107	0.792
0.487	0.401	0.619	0.510	0.506	0.462	0.651	0.542	0.205	0.534
0.287	0.503	0.257	0.301	0.085	0.207	0.667	0.582	0.252	0.837

La gráfica 30 ilustra de la muestra ordenada contra los cuantiles teóricos (línea recta). La máxima distancia de los datos a la recta es 0.233, correspondiente a la observación 36. Luego, $K_n = \text{máx}_{1\leq k\leq 40}\{|x_{(k)} - k/n|\} = |0.667 - 0.90| = 0.233$.

La prueba de hipótesis con significancia α tiene como región de rechazo $\mathcal{R}_\alpha = \{\boldsymbol{X} : K_n(\boldsymbol{X}) > c(\alpha, n)\}$. Luego, para $\alpha = 0.05$ y $n = 40$, $c = .165$ menor que 0.23, por lo que se rechaza H_0 con una significancia del 5 %.

[28] **?** (**?**).

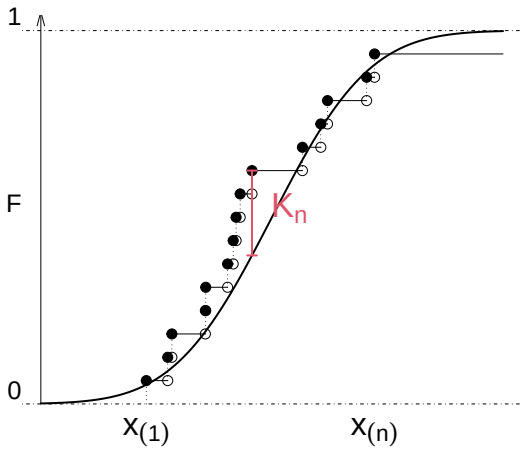


Figura 29: Funciones de distribución, teórica (continua) y empírica (escalonada). Se muestra el estadístico K_n observado.

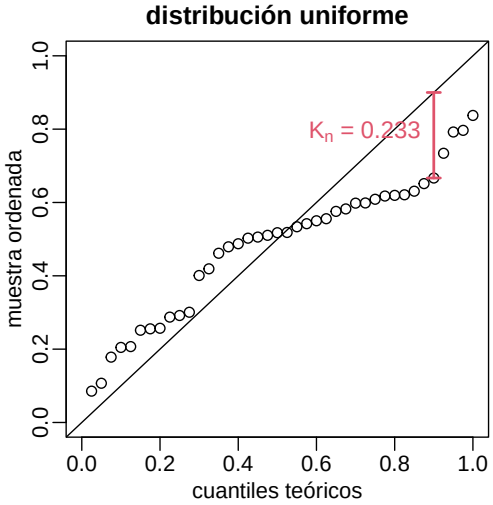


Figura 30: Funciones de distribución, teórica (continua) y empírica (escalonada). Se muestra el estadístico K_n observado.

7.7. Ejercicios

Refiérase a la Lista de Ejercicios 7.

Textos de apoyo.

Bickel and Doksum (1977); Casella and Berger (2002); Dudewicz and Mishra (1988); Knight (2000); Mood, Graybill, and Boes (1974); Rice (2007); Rincón (2019); Wackerly, Mendenhall III, and Scheaffer (2008).

Referencias

Barrios, E. (2024a). Apuntes para el curso de Cálculo de Probabilidades I. https://gente.itam.mx/ebarrios/docs/apuntes_CP1.pdf. (3 de enero de 2024).

Barrios, E. (2024b). Apuntes para el curso de Cálculo de Probabilidades II. https://gente.itam.mx/ebarrios/docs/apuntes_CP2.pdf. (3 de enero de 2024).

Bastian, H. (2013). Statistical Significance and Its Part in Science Downfalls. https://absolutelymaybe.plos.org/2013/11/11/statistical-significance-and-its-part-in-science-downfalls. (Consultado: 11/07/2022).

Bickel, P. and K. Doksum (1977). *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice Hall.

Blitzstein, J. K. and J. Hwang (2014). *Intorduction to Probability*. Boca Raton, FL: CRC Press.

Casella, G. and R. L. Berger (2002). *Statistical Inference* (2nd ed.). Pacific Gove, CA: Duxbury.

Chihara, L. and T. Hesterberg (2019). *Mathematical Statistics with R* (2 ed.). Hoboken, NJ: Wiley.

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

Dudewicz, E. J. and S. N. Mishra (1988). *Modern Mathematical Statistics*. New York, N.Y.: Wiley.

Garthwaite, P., I. Jolliffe, and B. Jones (2002). *Statistical Inference*. Oxford, UK: Oxford University Press.

Hoel, P. G., S. C. Port, and C. J. Stone (1971). *Introduction to Probability Theory*. Boston: Houghton Miffling Company.

Hogg, R. V. and A. T. Craig (1978). *Introduction to Mathematical Statistics* (4 ed.). New York: Macmillan Publishing Co., Inc.

Knight, K. (2000). *Mathematical Statistics*. Boca Raton, Florida: Chapman & Hall/CRC.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics* (3rd ed.). Singapore: McGraw-Hill.

Rice, J. S. (2007). *Mathematical Statistics and Data Analysis* (3rd ed.). Belmont, California: Brooks/Cole: Cengage Learning.

Rincón, L. (2019). *Una introducción a la estadística inferencial*. CDMX: Universidad Nacional Autónoma de México.

Roussas, G. G. (1997). *A Course in Mathematical Statistics* (2nd ed.). San Diego, CA.: Academic Press.

Stigler, S. (2016). *The Seven Pillars of Statistical Wisdom*. Cambridge, MA.: Harvard University Press.

Stigler, S. (2017). *Los siete pilares de la sabiduría estadística*. CDMX, México: Libros Grano de Sal. AME.

Tukey, J. (2020). *Exploratory Data Analysis* (1 ed.). Hoboken, N.J.: Pearson Education.

Wackerly, D. D., W. Mendenhall III, and R. L. Scheaffer (2008). *Mathematical Statistics with Applications* (7 ed.). Australia: Thomson.