Di Tella University - 2024

Lecture Notes - INFERENCIA ESTAD

Sesto Francisco

9 de septiembre de 2024

Índice

Inferencia estadística

Introducción

A partir de los datos x queremos inferir algún aspecto sobre la distribución $F(x)$ asumiendo que F pertenece a un modelo estadístico $\mathcal{F}$. La distribución F, o algún aspecto de la misma, es de interés. Conocerla nos provee información relevante para responder alguna pregunta concreta.

Los ingredientes de la inferencia estadística son:

1. Un **conjunto de mediciones**: $x_1, \ldots, x_n$ llamados **datos**.
2. Un **modelo estadístico** $\mathcal{F}$ con dos supuestos:
   a) $\underline{x} = (x_1, \ldots, x_n)$ es el valor observado (o muestra aleatoria) de un vector aleatorio $\underline{X} = (X_1, \ldots, X_n)$ con distribución acumulada F desconocida.
   b) F pertenece a un conjunto de distribuciones $\mathcal{F}$ específico.

En **inferencia estadística** inferimos a partir de lo particular algo general, **a partir de la muestra queremos encontrar a que familia de distribución pertenecen**. Por lo tanto, las conclusiones siempre son tentativas, nunca son definitivas.

Conceptos

El **parámetro del modelo** es el parámetro θ que indexa la colección de posibles distribuciones del modelo estadístico:

$$\mathcal{F} = \{F(\cdot, \theta) : \theta \in \Theta\}$$

Para datos $\underline{X}$ pensamos que hay un θ^* de manera que $F(\cdot, \theta^*)$ genera los datos. A θ^* se lo llama **el valor verdadero** de θ.

A la distribución $F = F(\cdot, \theta^*)$ del vector aleatorio $\underline{X} = (X_1, \ldots, X_n)$ del cual los datos $\underline{x}$ representan su valor observado, se la llama la **distribución verdadera**, o la distribución que generó los datos.

A la familia $\mathcal{F}$ la podemos también pensar como **una familia de densidades** (o funciones de probabilidad puntuales)

$$\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta\}$$

donde $f(\cdot, \theta)$ es la densidad o función puntual de probabilidad correspondiente a la distribución acumulada $F(\cdot, \theta)$.

Cuando es claro por el contexto del problema, que $X_i \overset{iid}{\sim}$, entonces $F(\cdot, \theta)$ se refiere a la distribución acumulada marginal de cada X_i. De lo contrario, se refiere a la distribución acumulada conjunta del vector $\underline{X} = (X_1, \ldots, X_n)$.

Un ejemplo de esto podría ser si $X_i \sim^{iid} Poi(\lambda)$, con $\lambda \in (0, +\infty)$ entonces $Sop(X_i) = \{0, 1, 2, ...\}$ y el modelo $\mathcal{F} = \{f(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}\}$ es una familia de soporte común pues es independiente de λ.

Otro ejemplo pero que no cumple es si $X_i \sim^{iid} U(0, \theta)$, con $\theta \in (3, +\infty)$ entonces $Sop(X_i) = (0, \theta]$ y el modelo $\mathcal{F} = \{f(x, \lambda) = \frac{1}{b-a}\}$ no es una familia de soporte común pues depende de θ.

Definición 1 (Soporte común)**.** Diremos que un *modelo estadístico* $\mathcal{F}$ tiene soporte común si $\text{Sop}(X)$ no depende de θ. Es decir, $f(x,\theta) > 0$ para los mismos valores de x para cualquier valor de θ.

Tipos de inferencia

1. **Paramétrica:** $\Theta \subseteq \mathbb{R}^k$ para algún k. [1]
2. **No paramétrica:** Θ es el conjunto irrestricto de todas las posibles distribuciones de X, o de las distribuciones que tienen densidades suaves (es decir m veces diferenciables, para algún $m \in \mathbb{N}$).
3. **Semiparamétrica:** Θ no es ni paramétrico ni irrestricto. [2]

[1] Ejemplos:

- $X_i \overset{iid}{\sim} \text{Be}(p), p \in [0,1] = \Theta$.
- $X_i \overset{iid}{\sim} \text{Be}(p), p \in \{0,0{,}25, 0{,}5, 0{,}75, 1\} = \Theta$.
- $X_i \overset{iid}{\sim} N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in (0, +\infty)$. Entonces $\Theta = \mathbb{R} \times \mathbb{R}^+$ y $k = 2$.

[2] Un ejemplo podría ser Θ que indexa todas las distribuciones de $X = (Y, Z)$ donde $\mathbb{E}(Y \mid Z) = \alpha_0 + \alpha_1 Z$ pero no se sabe nada más sobre Y ni Z.

Inferencia frecuentista o bayesiana (caso paramétrico)

1. **Inferencia clásica o frecuentista.**[3] Para los frecuentistas **la probabilidad cuantifica qué tan azaroso es un evento.**
   - θ^* es el número que la distribución verdadera que genera a los datos $F(\cdot, \theta^*) \in \mathcal{F}$ dentro del modelo estadístico.
   - La probabilidad del evento A se interpreta como la *frecuencia relativa* bajo infinitas repeticiones si los datos $\underline{x}$ son generados con $F(\cdot, \theta^*)$.
2. **Inferencia bayesiana.**[4] Para los bayesianos **la probabilidad cuantifica la incertidumbre que se tiene sobre un evento.**
   - θ^* es una v.a. que asumimos que tiene una distribución "a priori" $\pi(\theta)$.
   - Bajo este paradigma la evidencia en los datos $\underline{x}$ nos permite actualizar la creencia sobre la distribución de θ^*.
   - Dicha distribución queda reflejada en la distribución condicional $\pi^*(\theta \mid \underline{x})$ de θ^* dado $\underline{X} = \underline{x}$. Esta distribución se llama distribución "a posteriori".

[3] Bajo el paradigma frecuentista la probabilidad de un evento bajo las mismas circunstancias es o o 1. Y por ello creen que existe un número o vector θ verdadero que genera los datos.

[4] Bajo el paradigma bayesiano la probabilidad de un evento bajo las mismas circunstancias es algún número entre o y 1 pues se define en base a mi incertidumbre sobre ese evento. Y por ello dicen que no conocen θ y que se distribuye con densidad a priori $\pi(\theta)$.

Principios de reducción de datos

Estadísticos

Definición 2 (Estadístico)**.** Un estadístico $t(\underline{X})$ es cualquier función del vector aleatorio $\underline{X}$ es una forma de reducción de los datos aleatorios $\underline{X}$.

Cuando usamos los datos para hacer inferencia sobre θ lo hacemos usando algún estadístico $t(\underline{X})$. El problema es que al reducir los datos, pasando de $\underline{X}$ a $t(\underline{X})$, podríamos perder información importante acerca de θ. Es importante entonces usar estadísticos que capturen toda la información sobre θ en la muestra.

Estadístico suficiente

Un estadístico suficiente captura toda la información sobre θ en la muestra $\underline{X}$.

Definición 3 (Estadístico suficiente)**.** Un estadístico $T = t(\underline{X})$ es **suficiente** para θ si la distribución de $\underline{X}|_{T=t}$ es la misma para cada t y, en particular, no depende de θ. Es decir, cuando conocemos el valor de t no queda información adicional en $\underline{X}$ que dependa de θ:

$$f_{\underline{X}|_{T=t}}(x_1, \cdots, x_n|t) = f_{\underline{X}|_{T=t}}(y_1, \cdots, y_n|t)$$

Para hacer inferencia sobre θ si $t(\underline{x})$ y $t(\underline{y})$ coinciden entonces la muestra sobre θ en ambos casos es la misma.

Observación 4. Notar que $T = t(\underline{X})$ puede ser un vector de funciones de $\underline{X}$, no es necesariamente una función escalar. Por ejemplo, $t(\underline{X}) = \underline{X}$ es un estadístico suficiente, aunque no es uno muy interesante.

Si tenemos un estadístico $T = t(\underline{X})$ y queremos chequear si es o no un estadístico suficiente, debemos calcular $f_{\underline{X}|T}(\underline{x}|t)$. A menudo ése no es un cálculo fácil. Este teorema, facilita demostrar que un estimador $T = t(\underline{X})$ es suficiente para un parámetro θ.

Proposición 5 (Teorema de factorización de Neyman)**.** *Un estadístico $T = t(\underline{X})$* ***es suficiente*** *para el parámetro θ que indexa el modelo $\mathcal{F}$ para la distribución de $\underline{X}$* ***si y sólo si existen funciones*** *k_1* ***y*** *k_2* ***tales que***

$$f(\underline{x}, \theta) = k_1[t(\underline{x}), \theta] \cdot k_2(\underline{x})$$

Observación 6. **Las funciones k_1 y k_2 no son únicas.** De hecho por ejemplo, tomando $k_1 = f(\underline{x}, \theta)$ y $k_2 = 1$ se tiene que $t(\underline{X}) = \underline{X}$ es un estadístico suficiente para θ.

Es deseable que cualquier inferencia que se haga sobre θ esté basada en un estadístico suficiente. Si hacemos inferencia con un estadístico suficiente $t(\underline{X})$ no se descarta información relevante acerca del valor verdadero del parámetro θ. Por ende, no se descarta información relevante acerca de la distribución que generó los datos $\underline{X}$.

Proposición 7 (Principio de suficiencia)**.** *Si $T = t(\underline{X})$ es* ***un estadístico suficiente****, entonces* ***cualquier inferencia*** *sobre θ* ***debe depender de los datos de la muestra $\underline{x}$ sólo a través de*** *$t(\underline{x})$. Es decir, si $t(\underline{x}) = t(\underline{y})$ entonces las conclusiones que sacamos acerca de θ cuando observamos $t(\underline{x})$ deben ser idénticas a las que sacamos cuando observamos $t(\underline{y})$.*

Estadístico suficiente minimal

Si T es un estadístico suficiente, podríamos preguntarnos si es posible "resumir" T de manera que $U = g(T)$ sigue siendo suficiente. ¿Hasta cuándo se puede seguir repitiendo el proceso?

Queremos estadísticos que son suficientes y que reducen "al máximo" los datos. Es decir, son estadísticos suficientes T para los cuales no es posible encontrar otro estadístico suficiente $U = g(T)$ con g no inyectiva.

Definición 8 (Estadístico suficiente minimal)**.** Un estadístico suficiente $T = t(\underline{X})$ es minimal suficiente si para cualquier otro estadístico suficiente $U = u(\underline{X})$ existe una función k tal que $T = k(U)$.

Proposición 9 (Propiedades)**.** *Notar que cumplen determinadas propiedades*

- *Si un estadístico suficiente es escalar,*[5] *entonces es suficiente minimal.*

[5] Un estadístico suficiente escalar es aquel que tiene solo un valor o dimensión 1.

- *Los estadísticos minimales suficientes no son únicos: si T es suficiente minimal y g es biyectiva, entonces $U = g(T)$ también es suficiente minimal.*
- *Si dos estadísticos son suficientes minimales, entonces tienen necesariamente la misma dimensión.*
- *En algunos modelos, los estadísticos suficientes minimales pueden tener dimensión más grande que la dimensión del parámetro.*

Este teorema da condiciones suficientes para encontrar estadísticos suficientes minimales, no los caracteriza.

Teorema 10 (Lehmann-Scheffé)**.** *En una familia de distribuciones*

$$\mathcal{F} = \{f(\underline{x}, \theta) : \theta \in \Theta\}$$

un estadístico $t(\underline{X})$ ***es suficiente minimal*** *para θ si Es decir, un estadístico $t(x)$ es* ***suficiente minimal*** *para θ si*

$$\frac{f(\underline{x}, \theta)}{f(\underline{y}, \theta)} \text{ no depende de } \theta \Leftrightarrow t(\underline{x}) = t(\underline{y})$$

El siguiente teorema es equivalente al teorema de Lehman-Scheffé.

Teorema 11 (Young and Smith Thm 6.1)**.** *Una condición necesaria y suficiente para que el estadístico $t(\underline{X})$ sea suficiente minimal es que*

$$t(\underline{x}) = t(\underline{y}) \iff \frac{f(\underline{x}, \theta_1)}{f(\underline{x}, \theta_2)} = \frac{f(\underline{y}, \theta_1)}{f(\underline{y}, \theta_2)}$$

Estadístico ancillar

Definición 12 (Estadístico ancillar)**.** Un estadístico $U = u(\underline{X})$ se dice **ancillar para θ si la distribución de U no depende de θ**. O sea,

$$f_U(u, \theta) = f_U(u) \text{ ya que no depende de } \theta$$

Es decir, si la distribución de $u(\underline{X})$ es la misma cualquiera sea la distribución de X en la clase $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$.

Un ejemplo de estadístico ancillar es dado $X_i \sim^{iid} N(\mu, 1)$. Entonces $U = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}$. Luego U y por lo tanto la varianza muestral $S^2 = \frac{1}{n-1} \cdot U$ son estadísticos ancillares para μ.

Intuitivamente, los estadísticos ancillares constituyen la noción contrapuesta a la de los estadísticos suficientes porque si a los datos x los reducimos a $u(\underline{x})$ perdemos todo lo que los datos tienen para informarnos sobre θ.

Ancillaridad en modelos de locación y escala

Definición 13 (Modelo de locación)**.** Un modelo de locación, es un modelo en el que $X_i \overset{iid}{\sim}$ de manera que

$$X_i = \theta + W_i$$

donde W_i tiene una distribución G conocida, es decir, en particular G no puede depender de θ.

Definición 14 (Modelo de escala)**.** Un modelo de escala, es un modelo en el que $X_i \overset{iid}{\sim}$ de manera que

$$X_i = \theta \cdot W_i$$

donde W_i tiene una distribución G conocida, es decir, en particular G no puede depender de θ.

Ahora si los estadísticos en modelos de locación cumplen ciertas características vamos a poder afirmar que son ancillares

Definición 15 (Estadístico invariante por traslaciones)**.** Un estadístico $u(\underline{X})$ es **invariante por traslaciones** si verifica para todo c y para toda $(x_1, x_2, \ldots, x_n)$ vale que

$$u(x_1 - c, x_2 - c, \ldots, x_n - c) = u(x_1, x_2, \ldots, x_n)$$

Proposición 16. *En un modelo de locación, un estadístico $u(\underline{X})$ invariante por traslaciones es un estadístico ancillar para θ.*

Análogamente sucede lo mismo con los modelos de escala

Definición 17 (Estadístico invariante por escala)**.** Un estadístico $u(\underline{X})$ es *invariante por escala* si verifica para todo k y para toda $(x_1, x_2, \ldots, x_n)$ vale que

$$u(k \cdot x_1, k \cdot x_2, \ldots, k \cdot x_n) = u(x_1, x_2, \ldots, x_n)$$

Proposición 18. *En un modelo de escala, un estadístico $u(\underline{X})$ invariante por transformaciones de escala es un estadístico ancillar para θ.*

Principio de condicionalidad

Proposición 19 (Principio de condicionalidad)**.** *Toda inferencia acerca de θ debe hacerse con la distribución condicional de $\underline{X}$ dado U donde U es un estadístico ancillar para θ como si el único posible valor de U fuese su valor observado u.*

Por ejemplo, como los frecuentistas toman el tamaño n de muestra como dado, se considera que toda la inferencia realizada es condicionado al tamaño n.

Observación 20. El principio de condicionalidad peca de un problema serio: en muchos modelos, existen muchos estadísticos ancillares distintos. El principio nos dice que si U es ancillar, debemos conducir inferencias condicionando en U. Pero, ¿si hubiera más de un estadístico ancillar? ¿En cuál(es) condicionamos? Deberíamos condicionar en todos, técnicamente hablando y eso no es muy útil o práctico.

Estadísticos completos

Definición 21 (Estadístico completo)**.** Un estadístico $T = t(\underline{X})$ es completo para θ si para toda función que vale

$$E_\theta[g(T)] = 0 \ \forall \theta \in \Theta \Rightarrow P_\theta(g(T) = 0) = 1$$

E_θ quiere resaltar que estamos bajo el supuesto que el parámetro verdadero del modelo es θ.

En otras palabras, T es completo cuando si reducimos al estadístico T con una función g de manera que el estadístico $g(T)$ es mean-independent de θ entonces el estadístico $g(T) = 0$ con probabilidad igual a 1.

Proposición 22. *Si tenemos $T = t(\underline{X})$ un estadístico suficiente y completo entonces si condicionamos o no condicionamos a un estadístico ancillar nos va a devolver la misma distribución.*

Teorema 23 (Teorema de Bahadur)**.** *Si T es un estadístico completo y suficiente, entonces T es minimal.*

El teorema de Bahadur da condiciones suficientes para describir estadísticos minimales, pero no caracteriza a los estadísticos minimales. Es decir, pueden existir estadísticos suficientes minimales que no sean completos.

La idea es la siguiente: a un estadístico suficiente no le falta nada y a un estadístico completo no le sobra nada entonces si un estadístico las cumple a la vez entonces es minimal.

Teorema 24 (Teorema de Basu)**.** *Si T es un estadístico suficiente y completo de θ y U es un estadístico ancillar de θ, entonces T y U son independientes.*

Del teorema de Basu se obtienen consecuencias importantes

Proposición 25 (Completitud, suficiencia y principio de condicionalidad)**.** *Si T es suficiente y completo para θ entonces satisface el principio de condicionalidad porque la distribución $T|_{ancillares}$ es la misma que la distribución marginal de T.*

Observación 26 (Existencia de estadístico suficiente y completos). Para cualquier modelo estadístico no podemos asegurar que exista un estadístico T suficiente y

completo.

Familias exponenciales

Definición 27 (Familia exponencial)**.** El modelo estadístico $\mathcal{F}$ para $X_i \sim_{iid} F \in \mathcal{F}$ es una ***familia exponencial*** si $\mathcal{F} = \{f(\underline{x}, \theta) : \theta \in \Theta\}$ se puede escribir como:

$$f(\underline{x}, \theta) = h(\underline{x})c(\theta)e^{\sum_{j=1}^{k} \eta_j(\theta)t_j(\underline{x})}$$

Si además se cumplen las siguientes propiedades la familia $\mathcal{F}$ es ***exponencial de rango completo***:

1. $t_1(x), \ldots, t_k(x)$ no son linealmente dependientes [6]
2. $\eta_1(\theta), \ldots, \eta_k(\theta)$ no son linealmente dependientes
3. Θ contiene un abierto en $\mathbb{R}^k$.

[6] que no satifaga ninguna restrición lineal o sea que no sean linealmente dependientes significa que si tomamos $a_1, ..., a_k \in \mathbb{R} | a_1 t_1(\underline{x}) + ... + a_k t_k(\underline{x}) = 0 \implies a_i = 0 \forall i = 1, ..., k$

Teorema 28. *Si $\mathcal{F}$ es una **familia exponencial de rango completo**, $t(\underline{X}) = (t_1(\underline{X}), \ldots, t_k(\underline{X}))$ es un estadístico suficiente completo.*

Inferencia con grandes muestras

Estimadores

En un modelo estadístico $\mathcal{F} = \{f(\cdot, \theta)\}$ nos puede interesar un parámetro de interés $\beta(\theta)$.[7]

[7] **Ejemplos para** $\beta(\theta)$: $E_\theta(X)$, $Var_\theta(X)$, $F_\theta(x)$ o $F_\theta^{-1}(0{,}9)$.

Definición 29 (Estimador)**.** Para un parámetro $\beta(\theta)$ un estimador es una función $\widehat{\beta} = \delta(X)$. Siendo $\widehat{\beta} = \delta(X)$ **una estrategia de resumir los datos** X **de manera de poder hacer inferencia sobre** $\beta(\theta)$.

Los estimadores $\delta(X)$ **suelen ser funciones de los estadísticos** T. Es decir, $\delta(X) = g(t(X))$ donde T es un estadístico que idealmente es suficiente y/o minimal y/o completo.

Observación 30 (Estimación puntual)*.* Si tomamos una muestra particular x el valor $\delta(x)$ se llama **estimación puntual de** $\beta(\theta)$. Es el valor de la función δ evaluada en los datos observados x.

Estimadores y muestreo

Notar que usamos $\delta(\underline{X})$ para

- **Estimación puntual:** inferir el valor desconocido de $\beta(\theta^*)$.
- **Estimación por intervalos:** inferir un conjunto de valores para $\beta(\theta^*)$.
- **Contraste de hipótesis:** elegir entre dos conjuntos disjuntos para $\beta(\theta^*)$.

Función de perdida

Nos interesa poder cuantificar cuantos nos equivamos en nuestra estimación para ello definimos los siguentes conceptos

Definición 31 (Error de estimación)**.** Para una muestra particular $\underline{x}$, si los datos son generados a partir de $F(\cdot,\theta) \in F$, definimos **el error de estimación** como

$$\delta(\underline{x}) - \beta(\theta)$$

Ahora dado el error de estimación queremos medir la "perdida" cuando hay una diferencia tanto por subestimar o sobreestimar o sea $|\delta(\underline{x}) - \beta(\theta)| > 0$.

Definición 32 (Función de perdida)**.** Es una función L que **mide cuán serio es el error de decidir que el parámetro de interés es** $\delta(\underline{x})$ dado que verdadero valor del modelo es $\beta(\theta)$

$$L(\theta, \delta(\underline{x})) = g(|\delta(\underline{x}) - \beta(\theta)|)$$

Algunas funciones de perdida son
Pérdida cuadrática: $L(\theta,\delta) = (|\delta - \beta(\theta)|)^2$
Pérdida del arquero: $L(\theta,\delta) =$
$\begin{cases} 1 & |\delta - \beta(\theta)| > a \\ 0 & |\delta - \beta(\theta)| \leq a \end{cases}$

Función de riesgo

Ahora aunque tenemos como medir el error notemos que el error va a ser diferente dependiendo de cada valor puntual y nos gustaría tener un modo más general de medirlo por eso definimos la función de riesgo.

Definición 33 (Función de riesgo)**.** Una función de riesgo R para el estimador $\delta(\underline{x})$ es

$$R(\theta, \delta(\underline{x})) = E_\theta[L(\theta, \delta(\underline{x}))] = \int_{sop(\underline{X})} L(\delta(\underline{x}), \beta(\theta)) f(\underline{x}, \theta) d\underline{x}$$

El riesgo $R(\theta, \delta(\underline{x}))$ es el valor promedio que toma $L(\theta, \delta(\underline{x}))$ en infinitas hipotéticas repeticiones en las que usaremos el estimador $\delta(\underline{x})$ para estimar a $\beta(\theta)$ si los datos son generados con $F(\cdot,\theta) \in F$.

Definición 34 (Error cuadrático medio)**.** El error cuadrático medio es un caso particular de una función de riesgo para un $L(\theta, \delta(\underline{x})) = (\beta(\theta) - \delta(\underline{x}))^2$ y entonces

$$ECM_\delta(\theta) = E_\theta[(\beta(\theta) - \delta(\underline{x}))^2] = Var_\theta(\overbrace{\beta(\theta)}^{constante} - \delta(\underline{x})) + [E_\theta(\beta(\theta) - \delta(\underline{x}))]^2$$

$$ECM_\delta(\theta) = Var_\theta(\delta(\underline{x})) + [\underbrace{\beta(\theta) - E_\theta(\delta(\underline{x}))}_{=sesgo_\delta(\theta)}]^2$$

Estimador insesgado

Podemos también distinguir aquellos estimadores que su media no se desvía del verdadero valor.

Definición 35 (Estimador insesgado)**.** Un estimador δ de un parámetro $\beta(\theta)$ se dice insesgado si

$$sesgo_\delta(\theta) = 0 \, \forall \, \theta \in \Theta$$

Elección de estimador por minimización

Proposición 36 (Criterio de elección de estimador por minimización)**.** *Si tenemos una lista de estimadores $\delta_a(\overline{X})$ nos gustaría elegir a^* de manera que $\delta_{a^*}(\overline{X})$ tenga la menor función de riesgo posbile dentro de los $\delta_a(\overline{X})$. O sea queremos el a^* tal que*

$$R_\theta(\delta_{a^*}, \theta) \leq R_\theta(\delta_a, \theta) \;\; \forall a \neq a^* \;\; y \;\; \forall \theta \in \Theta$$

Inadmisibilidad de un estimador

Hay algunos estimadores que nunca convienen utilizar y los caracterizamos del siguiente modo

Definición 37 (Inadmisibilidad de un estimador)**.** Un estimador $\delta(\underline{X})$ de $\beta(\theta)$ es inadmisible con respecto a una función de riesgo $R(\theta, \mu)$ si existe otro estimador $\delta^*(\underline{X})$ tal que

$$R(\theta, \delta^*(\underline{X})) \leq R\left(\theta, \delta(\underline{X})\right) \text{ para todo } \theta \in \Theta \;\; y$$

$$R(\theta, \delta^*(\underline{X})) < R\left(\theta, \delta(\underline{X})\right) \text{ para algun } \theta \in \Theta$$

Limitación de la función de riesgo

Siguiendo el razonamiento anterior querríamos encontrar $\delta^*(\underline{X})$ que tenga el menor riesgo $R(\theta, \delta)$ entre todos los demás estimadores para todos los valores de $\theta \in \Theta$.

Observación 38 (Imposibilidad de minimizar uniformemente la función de riesgo). Es imposible encontrar δ^* tal que $R(\delta^*, \theta) \leq R(\delta, \theta) \forall \delta \; \forall \theta \in \Theta$.

Elección de estimador por minimax

Otro criterio de optimalidad para obtener un estimador $\delta^*(\underline{X})$ consiste en seleccionar (si es que existe) el estimador que minimiza el máximo valor posible de la función de riesgo

Básicamente busca el estimador cuyo máximo valor de la función de riesgo es menor que el máximo valor para los otros.

El procedimiento descarta estimadores $\delta(\underline{X})$ que tienen riesgo bajo en casi todo Θ excepto en alguna pequeña región de Θ para quedarse con estimadores $\delta^*(\underline{X})$ que tal vez tengan peor performance que $\delta(\underline{X})$ excepto en esa pequeña región.

Definición 39 (Criterio de elección de estimador minimax)**.** Un estimador $\delta^*(\underline{X})$ se dice minimax con respecto a una función de pérdida dada si para cualquier otro estimador δ se satisface

$$\max_{\theta\in\Theta} R(\theta,\delta^*) \leq \max_{\theta\in\Theta} R(\theta,\delta)$$

Observación 40. El criterio de minimaxidad es cuestionable, porque es demasiado conservador. Asume que el "oráculo" esperará a que elijamos nuestro estimador δ para luego recién generar los datos de la "peor" distribución $F(\cdot,\theta)$ posible para nuestra elección $\delta(\underline{X})$.

Observación 41. Hay δ^* minimax que son inadmisibles.

Elección de estimador por Bayes

Otro criterio, llamado el criterio de Bayes, selecciona como óptimo aquel estimador $\delta^*(\underline{X})$ que minimiza "un promedio ponderado" de la función de riesgo $R(\theta,\delta)$.

Definición 42 (Criterio de elección de estimador de Bayes)**.** Un estimador $\delta^*(\underline{X})$ se dice estiamdor de Bayes si miniza un promedio ponderado de $R(\theta,\delta)$ usando la distribución a priori $\pi(\theta)$

$$\delta^* = \arg\min_{\delta} B_\pi(\delta) = \arg\min_{\delta} \int_\Theta R(\theta,\delta)\pi(\theta)d\theta$$

Elección de estimador por UMVUE

Definición 43 (Criterio de elcción de estimador UMVUE)**.** Un estimador $\delta^*(\underline{X})$ es insesgado de minima varianza uniforme (UMVUE) de $\beta(\theta)$ si

- $\delta^*(\underline{X})$ es insesgado para $\beta(\theta)$
- para cualquier otro estimador $\delta(\underline{X})$ insesgado de $\beta(\theta)$ se verifica que

$$Var_\theta[\delta^*(\underline{X})] \leq Var_\theta[\delta(\underline{X})] \quad \forall\theta\in\Theta$$

Proposición 44. *Si un estimador es suficiente y completo es UMVUE*

Observación 45. EI criterio de ser UMVUE es cuestionable porque no garantiza admisibilidad

Eficiencia de estimadores

Esta definición se puede encontrar también con $\widehat{\theta}_n$ en vez de $\delta(\underline{X})$ y $\widetilde{\theta}_n$ en vez de $\widetilde{\delta}(\underline{X})$.

Definición 46 (Estimador eficiente)**.** Dados dos estimadores insesgados $\delta(\underline{X})$ y $\widetilde{\delta}(\underline{X})$ de θ decimos que $\delta(\underline{X})$ es eficiente si

$$Var_\theta(\delta(\underline{X})) \leq Var_\theta(\widetilde{\delta}(\underline{X}))$$

Definición 47 (Eficiencia relativa)**.** Definimos la eficiencia relativa entre dos estimadores insesgados $\delta(\underline{X})$ y $\widetilde{\delta}(\underline{X})$ como

$$ER(\delta(\underline{X}), \widetilde{\delta}(\underline{X})) = \frac{Var(\delta(\underline{X}))}{Var(\widetilde{\delta}(\underline{X}))}$$

Observación 48. Preferiremos $\delta(\underline{X})$ para estimar a θ si $ER(\delta(\underline{X}), \widetilde{\delta}(\underline{X})) < 1$.

Rao Blackwell

Sea $T = t(\underline{X})$ es un estadístico suficiente de θ. A partir de un estimador $\delta(\underline{X})$ de θ, podemos construir un nuevo estimador

$$\eta_\delta(T) = E[\delta(\underline{X}) \mid T]$$

Notemos que la suficiencia de $T = t(\underline{X})$ nos garantiza que $\eta_\delta(T)$ depende de los datos $\underline{X}$ y por lo tanto puede ser un estimador de θ.

Notemos que el teorema de Rao-Blackwell nos da un algoritmo para partir de un estadístico suficiente y a partir de allí obtener otro estadístico que tenga menor ECM.

Teorema 49 (Teorema de Rao-Blackwell)**.** *Sea $\underline{X}$ un vector con distribución conjunta que pertenece a la familia $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta\}$. Sea $\delta(\underline{X})$ un estimador de θ y sea $T = t(\underline{X})$ un estadístico suficiente para θ. Supongamos que para cada $\theta \in \Theta$, $L(\theta, \delta)$ es una función convexa en δ entonces*

$$R(\theta, \eta_\delta) \leq R(\theta, \delta) \quad \textit{para todo } \theta \in \Theta$$

Además, si $L(\theta, \delta)$ es estrictamente convexa en δ, entonces la desigualdad se satisface para todo $\theta \in \Theta$.

Propiedades asintóticas de estadísticos

Convergencia

Nos interesa analizar las propiedades asintóticas de los estadísticos

Definición 50 (Consistencia de sucesión de estimadores)**.** Considere una sucesión de estimadores $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \ldots, \widehat{\beta}_n$, de un parámetro $\beta(\theta)$ donde

$$\widehat{\beta}_n = \delta_n(X_1, \ldots, X_n) = \delta_n(\underline{X}_n)$$

La sucesión de estimadores $\{\widehat{\beta}_n\}_{n\in\mathbb{N}}$, es consistente para el parámetro $\beta(\theta)$ para las familias $\mathcal{F}_n = \{F_n(\underline{x}_n, \theta) : \theta \in \Theta\}$, $n = 1, 2, \ldots$ si cualquiera sea $\theta \in \Theta$, se verifica que

$$\lim_{n\to\infty} P_\theta\left(\left|\widehat{\beta}_n - \beta(\theta)\right| < \varepsilon\right) = 1 \quad \text{para todo } \varepsilon > 0$$

Es decir si $\lim_{n\to\infty} \widehat{\beta}_n \xrightarrow{P_\theta} \beta(\theta)$ donde $\xrightarrow{P_\theta}$ significa convergencia en probabilidad cuando los datos $\underline{X}_n = (X_1, \ldots, X_n)$ son generados bajo la distribución $F_n(\underline{x}_n, \theta)$.

Cabe recordar que si se cumple convergencia en media cuadrática también se cumple en probabilidad

Proposición 51 (Convergencia en media cuadrática implica consistencia)**.** *Si para todo $\theta \in \Theta$ vale que $\lim_{n\to\infty} \widehat{\beta_n} \xrightarrow{m.c._\theta} \beta(\theta)$ i.e.*

$$\lim_{n\to\infty} ECM_{\delta_n}(\theta) = \lim_{n\to\infty} E_\theta[(\beta(\theta) - \widehat{\beta}_n)^2] = \lim_{n\to\infty} sesgo^2_{\delta_n}(\theta) + Var_\theta[\delta_n(\underline{X}_n)] = 0$$

entonces $\widehat{\beta_n} = \delta_n(\underline{X}_n)$ es consistente para $\beta(\theta)$.

Velocidad

Sea $\widehat{\beta}_n$ consistente para $\beta(\theta)$. ¿Existe α de manera que $n^\alpha\left(\widehat{\beta}_n - \beta(\theta)\right)$ converja en distribución a una distribución no degenerada para todo θ?

- Si α es demasiado pequeño $n^\alpha\left(\widehat{\beta}_n - \beta(\theta)\right) \xrightarrow{P_\theta} 0$
- Si α es demasiado grande $\left|n^\alpha\left(\widehat{\beta}_n - \beta(\theta)\right)\right| \xrightarrow{P_\theta} \infty$
- Si $\exists\alpha : n^\alpha\left(\widehat{\beta}_n - \beta(\theta)\right) \xrightarrow{L(F_\theta)} G_\theta$, con G_θ es una distribución no-degenerada (o sea que no es la distribución de una constante) entonces la forma de la distribución acumulada de la distribuciín G_θ "ser´a indistinguible" de la distribución en el límite de $n^\alpha\left(\widehat{\beta}_n - \beta(\theta)\right)$

$L(F_\theta)$ es lo mismo que D_{F_θ} o sea converger en distribución o en ley

Definición 52 (Velocidad de convergencia)**.** Sea un modelo $\mathcal{F}_n = \{F_n(\cdot, \theta) : \theta \in \Theta\}$ para la distribución de $\underline{X}_n$ y $\beta(\theta) \in \mathbb{R}^k$ es algún parámetro. Si para algún $\alpha > 0$ se verifica que

$$n^{\alpha}\left(\widehat{\beta}_n - \beta(\theta)\right) \xrightarrow{L(F_\theta)} G_\theta$$

donde G_θ es una distribución no degenerada $\forall \theta \in \Theta$, entonces se dice que $\widehat{\beta}_n$ tiene velocidad de convergencia $n^{-\alpha}$. Se llama constante de normalización a n^{α}.

Estimadores asintóticamente normales

Definición 53 (Sucesión de estimadores asintóticamente normal)**.** Una sucesión de estimadores $\widehat{\beta}_n$ es **asintóticamente normal** bajo una familia $\mathcal{F}_n = (F_n(\cdot, \theta) : \theta \in \Theta)$ si para todo $\theta \in \Theta$ se verifica

$$\sqrt{n}\left(\hat{\beta}_n - \beta(\theta)\right) \xrightarrow{L(F_\theta)} N(0, V(\theta))$$

En este caso:

- la velocidad de convergencia es $n^{-1/2}$,
- la distribución límite G_θ es Normal (univariada si $\hat{\beta}_n$ es escalar y p-multivariada si $\hat{\beta}_n$ es de dimensión $p > 1$) con media o para alguna matriz positiva definida o escalar positivo, $V(\theta)$.

Observación 54. Se denomina $V(\theta)$ como varianza asintótica de $\hat{\beta}_n$. La denominación es imprecisa por dos motivos:

1. Porque $V(\theta)$ no necesariamente el límite de la sucesión de varianzas $\text{Var}_\theta\left[\sqrt{n}\left(\hat{\beta}_n - \beta(\theta)\right)\right]$.
2. Porque $V(\theta)$ es la varianza de la distribución asintótica de $\sqrt{n}\left(\hat{\beta}_n - \beta(\theta)\right)$ y no de la distribución asintótica de $\hat{\beta}_n$. Esta última tiene varianza 0.

Distribución normal k-variada

Definición 55 (Distribución normal k-variada)**.** Un vector aleatorio $\underline{U} = (U_1, \ldots, U_k)^T$ tiene distribución normal k-variada $N_k(\underline{\mu}, \Sigma)$ no-degenerada, si la densidad conjunta de $\underline{U}$ es

$$f(u_1, \ldots, u_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(\underline{u} - \underline{\mu})^T \Sigma^{-1} (\underline{u} - \underline{\mu})}$$

para algún $\underline{\mu} = (\mu_1, \ldots, \mu_p)^T \in \mathbb{R}^k$ y alguna matriz Σ definida positiva.

Teorema Central del límite multivariado

Teorema 56 (Teorema central del límite multivariado)**.** *Si* $\underline{X_i} = (X_1, \ldots, X_{ki})^T$ *con* $i = 1, \ldots, n$*, son vectores aleatorios k-variados independientes e igualmente distribuidos, tales que la varianza de cada componente del vector* $\underline{X_i}$ *es finita, entonces*

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\underline{X_i} - \mathbb{E}(\underline{X_i})\right] \xrightarrow{L_F} N_k(\underline{0}, \Sigma)$$

donde $\frac{1}{n}\sum_{i=1}^{n}\underline{X_i}$ denota el vector $\left(\frac{1}{n}\sum_{i=1}^{n}X_{i1}, \ldots, \frac{1}{n}\sum_{i=1}^{n}X_{ik}\right)^T$ y

$$\mathbb{E}(\underline{X_i}) = (\mathbb{E}(X_{1i}), \ldots, \mathbb{E}(X_{ki}))^T$$

siendo

- Σ es la matriz de dim $p \times p$ de varianza-covarianza del vector $\underline{X_j}$.
- Σ es la misma para todo i por la suposición de que todos los $\underline{X_i}$ son igualmente distribuidos.
- Es decir, la componente $(j, \ell) \in \{1, \ldots, k\}^2$ de la matriz Σ es

$$\Sigma_{j,k} = \mathrm{Cov}(X_{ji}, X_{\ell i})$$

.

Método Delta multivariado

Definición 57. Sea $\underline{Z_1}, \underline{Z_2}, \ldots, \underline{Z_n}, \ldots$ una sucesión infinita de vectores aleatorios ℓ-variados, donde $\underline{Z_n} = (Z_{1n}, Z_{2n}, \ldots, Z_{\ell n})^T$. Supongamos que existe un vector constante $\underline{a} \in \mathbb{R}^\ell$ y una matriz Σ de dimensión $\ell \times \ell$ tal que

$$\sqrt{n}\,(\underline{Z_n} - \underline{a}) \xrightarrow{L} N_\ell(\underline{0}, \Sigma)$$

donde $\underline{0} = (0, \ldots, 0)^T$. Supongamos que $g(\underline{z}) = (g_1(\underline{z}), \ldots, g_k(\underline{z}))^T : \mathbb{R}^\ell \to \mathbb{R}^k$ es una función continuamente diferenciable de $\underline{z} = (z_1, \ldots, z_\ell)^T$. Sea

$$\nabla(\underline{a}) = \begin{bmatrix} \left.\frac{\partial g_1(\underline{z})}{\partial z_1}\right|_{\underline{z}=\underline{a}} & \left.\frac{\partial g_1(\underline{z})}{\partial z_2}\right|_{\underline{z}=\underline{a}} & \cdots & \left.\frac{\partial g_1(\underline{z})}{\partial z_\ell}\right|_{\underline{z}=\underline{a}} \\ \left.\frac{\partial g_2(\underline{z})}{\partial z_1}\right|_{\underline{z}=\underline{a}} & \left.\frac{\partial g_2(\underline{z})}{\partial z_2}\right|_{\underline{z}=\underline{a}} & \cdots & \left.\frac{\partial g_2(\underline{z})}{\partial z_\ell}\right|_{\underline{z}=\underline{a}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial g_k(\underline{z})}{\partial z_1}\right|_{\underline{z}=\underline{a}} & \left.\frac{\partial g_k(\underline{z})}{\partial z_2}\right|_{\underline{z}=\underline{a}} & \cdots & \left.\frac{\partial g_k(\underline{z})}{\partial z_\ell}\right|_{\underline{z}=\underline{a}} \end{bmatrix}$$

Supongamos que $\nabla(\underline{a})^T \Sigma \nabla(\underline{a})$ es positiva definida. Entonces se verifica que

$$\sqrt{n}\,[g(\underline{Z_n}) - g(\underline{a})]^T \xrightarrow{L} N_k\left(\underline{0}, \nabla(\underline{a})^T \Sigma \nabla(\underline{a})\right)$$

Desigualdad de Hoeffding

Definición 58 (Desigualdad de Hoeffding)**.** Dadas variables aleatorias Y_i independientes tales que $\mathbb{E}(Y_i) = 0$ y $Y_i \in [a_i, b_i]$, la desigualdad de Hoeffding establece que para cualquier $\epsilon > 0$:

$$P\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Tomemos $X_i \overset{\text{i.i.d.}}{\sim} F \in \mathcal{F}$, donde $\mathcal{F}$ es un modelo no paramétrico. Sean $I_{(-\infty,x]}(X_i) \overset{\text{i.i.d.}}{\sim} \text{Ber}(F(x))$, entonces $\mathbb{E}\left(I_{(-\infty,x]}(X_i)\right) = F(x)$.

Sea $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$.

Definimos las variables $Y_i = \frac{1}{n}\left(I_{(-\infty,x]}(X_i) - F(x)\right)$, las cuales son independientes e identicamente distribuidas.

Notemos que $\mathbb{E}(Y_i) = 0$ y $Y_i \in \left[-\frac{1}{n}, \frac{1}{n}\right]$.

Aplicando Hoeffding a estas variables Y_i, se tiene que para todo $x \in \text{Soporte}(X)$:

$$P\left(|\hat{F}_n(x) - F(x)| > \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right)$$

Conjuntos e intervalos de confianza

Podemos distinguir a los intervalos de confianza según

1. **Exactos y Asíntoticos:** Un intervalo de confianza exacto garantiza cobertura precisa según la distribución. El asintótico es aproximado y se basa en muestras grandes.
2. **Basado en estadísticos o en desiguldades**: Un intervalo de confianza basado en estadísticos usa estimadores de la muestra y supuestos de distribución. El basado en desigualdades se construye usando límites teóricos sin asumir una distribución específica.
3. **Puntual y uniforme:** Un intervalo de confianza puntual es aquel en el cual la longitud del intervalo depende de datos muestrales o parámetros conocidos. Un intervalo de confianza uniforme es aquel en el cual la longitud del intervalo es constante para todas las distribuciones en un conjunto especificado.

Definición 59 (Intervalo de confianza exactos y asintóticos)**.** Dada una muestra aleatoria $X_1, \ldots, X_n \overset{iid}{\sim} \theta$ y θ un parámetro desconocido en la población; un **intervalo de confianza exacto (asintótico)** de nivel $1-\alpha$ (nivel aproximado $1-\alpha$) para el parámetro θ es un **conjunto aleatorio**

$$IC_{\theta}^{(1-\alpha)100\,\%} \quad o \quad IC_{\theta,as}^{(1-\alpha)100\,\%} = (A(\underline{X}), B(\underline{X}))$$

donde A y B son **funciones de** $X = (X_1, \ldots, X_n)$ y para todo θ se cumple que:

$$P_\theta\left(\theta \in IC_{\theta}^{(1-\alpha)100\,\%}\right) \geq 1-\alpha \quad \text{ó} \quad P_\theta\left(\theta \in IC_{\theta,as}^{(1-\alpha)100\,\%}\right) \overset{n\to+\infty}{\longrightarrow} 1-\alpha$$

Definición 60 (Intervalo de confianza puntual)**.** $C_n = c_n(\underline{X}_n),\ n = 1, 2, \ldots$ son intervalos o conjuntos de confianza **asintóticos puntuales de nivel aproximado** $1-\alpha$ para un parámetro $\beta(\theta)$ si

$$P_\theta\left(\beta(\theta) \in C_n\right) \overset{n\to+\infty}{\longrightarrow} 1-\alpha \quad \text{para todo } \theta \in \Theta$$

Definición 61 (Intervalo de confianza uniforme)**.** $C_n = c_n(\underline{X}_n),\ n = 1, 2, \ldots$ son intervalos o conjuntos de confianza **asintóticos uniforme de nivel aproximado** $1-\alpha$ para un parámetro $\beta(\theta)$ si

$$\inf_{\theta\in\Theta} P_\theta\left(\beta(\theta) \in C_n\right) \overset{n\to+\infty}{\longrightarrow} 1-\alpha$$

Algunas consideraciones a tomar en cuenta

- Algunos valores frecuentes para $1-\alpha$ son 0.95 y 0.99.
- Hay un *trade-off* entre la longitud del IC, L, y el nivel de confianza, $1-\alpha$.

Proposición 62 (Intervalo de confianza uniforme implica puntual)**.** *Notar que*

- *IC asintótico uniforme* $\Rightarrow$ *IC asintótico puntual.*
- *IC asintótico puntual* $\not\Rightarrow$ *IC asintótico uniforme.*

IC exacto para μ si $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ con σ^2 conocida

Supongamos una población normal con varianza σ^2 **conocida**, es decir, $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$. Con el **estadístico** Z construiremos un IC exacto para $E(X)$ de nivel $1-\alpha$

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

ya que tomando z_{α_1} y $z_{1-\alpha_2}$

$$P(z_{\alpha_1} \leq Z \leq z_{1-\alpha_2}) = 1-\alpha$$

Para que el intervalo tenga la **menor longitud posible** elegimos $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$

$$P_{\mu,\sigma^2}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right) = P_{\mu,\sigma^2}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Por lo tanto, el $IC_\mu^{(1-\alpha)100\,\%}$ está dado por

$$IC_\mu^{(1-\alpha)100\,\%} = \Big[\underbrace{\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{A_n}, \quad \underbrace{\bar{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{B_n}\Big]$$

Observación 63 (Propiedades). Para datos $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ con σ^2 conocida definimos

- **Margen de error del IC:** $ME = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
- **Longitud del IC:** $Longitud(IC_\mu^{(1-\alpha)100\,\%}) = 2ME$
- **Error de estimación:** $Error = |\overline{X}_n - \mu|$

Algunas propiedades son

- **El error es menor igual que el margen de error** $Error \leq ME$
- Como l**a longitud** es $Longitud(IC_\mu^{(1-\alpha)100\,\%}) = B_n - A_n = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ entonces vemos que
  - Disminuye si aumenta n
  - Aumenta si aumenta la confianza $1-\alpha$
  - Aumenta si aumenta el desvío σ (el cual no controlamos)[8]

[8] Con esto si conocemos o podemos acotar σ, podemos despejar un valor de n para que el error sea menor que algún valor dado, con un nivel de confianza $1-\alpha$

IC exacto basado en desigualdad de Tchebyshev para $E(X) = \mu$ si $X_i \overset{iid}{\sim}$

Podemos usar Tchebysehv para construir un intervalo para $E(X)$ si $X_i \overset{iid}{\sim}$. Por Tcheby para $\overline{X}_n$ y $\epsilon = \frac{\sigma}{\sqrt{n\alpha}}$ sabemos que

$$P_{\mu,\sigma}(|\overline{X}_n - E(X)| \geq \frac{\sigma}{\sqrt{n\sigma}}) \leq \alpha$$

Entonces

$$P_{\mu,\sigma}(\overline{X}_n - \frac{\sigma}{\sqrt{n\sigma}} \leq E(X) \leq \overline{X}_n + \frac{\sigma}{\sqrt{n\sigma}}) \geq 1 - \alpha$$

Por lo tanto

$$IC^{(1-\alpha)100\,\%}_{E(X),Tcheby} = [\overline{X}_n - \frac{\sigma}{\sqrt{n\sigma}}, \overline{X}_n + \frac{\sigma}{\sqrt{n\sigma}}]$$

es un intervalo de confianza de nivel $1 - \alpha$ para $E(X)$. Notemos que en este caso el IC para μ no está basado en un estadístico y su distribución (si bien $\overline{X}_n$ es un estimador de μ) sino en la desigualdad de Tchebychev.

IC exacto puntual para σ^2 si $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$

Para construir este intervalo de confianza usaremos que si $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, el estadístico $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ tiene distribución chi-cuadrado con $n-1$ grados de libertad. Entonces, denotando $\chi^2_{n-1,\beta}$ como el cuantil de la distribución chi-cuadrado con $n-1$ grados de libertad que cumple que: $P_{(\mu,\sigma)}\left(\chi^2 \leq \chi^2_{n-1,\beta}\right) = \beta$ se tiene que:

$$\begin{aligned}
1 - \alpha &= P_{(\mu,\sigma)}\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) \\
&= P_{(\mu,\sigma)}\left(\frac{1}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) \\
&= P_{(\mu,\sigma)}\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)
\end{aligned}$$

Por lo tanto, el intervalo de confianza $(1-\alpha)100\,\%$ para σ^2 está dado por:

$$IC^{(1-\alpha)100\,\%}_{\sigma^2} = \left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$$

IC para μ si $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ y σ^2 es desconocida

Si σ^2 era conocida para construir el $IC^{(1-\alpha)100\,\%}_{\mu}$ usamos el estadístico:

$$Z = \frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Si σ^2 es desconocida, para construir el $IC^{(1-\alpha)100\,\%}_{\mu}$ usaremos el estadístico:[9]

[9] Recordemos que la varianza muestral se define como

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}$$

$$W = \frac{\overline{X}_n - \mu}{\sqrt{\frac{S^2}{n}}} \sim T$$

No sabemos la distribución T en este caso, pero tenemos dos opciones:

- **Opción 1:** buscamos la distribución exacta $W \sim t_{n-1}$.
- **Opción 2:** buscamos una aproximación asintótica $W \xrightarrow{D} N(0,1)$.

Siguiendo la opción 1 del caso anterior entonces si $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ y σ es desconocida, y si utilizamos el estadístico W, tiene distribución exacta t-Student con $n-1$ grados de libertad:

$$W = \frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Entonces, el intervalo de confianza de $(1-\alpha)100\,\%$ para μ está dado por:

$$IC_{\mu}^{(1-\alpha)100\,\%} = \left[\overline{X}_n - t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \overline{X}_n + t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right]$$

donde $t_{1-\frac{\alpha}{2},n-1}$ es el cuantil $1-\frac{\alpha}{2}$ de la distribución t de Student con $n-1$ grados de libertad.

IC exacto para p si $X_i \overset{iid}{\sim}$ Be(p) usando Hoeffding

Si $X_i \overset{iid}{\sim} \text{Be}(p)$ con $p \in \Theta = [0,1]$ y $\epsilon > 0$, entonces la desigualdad de Hoeffding dice que:

$$P\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Entonces, tomando $\epsilon = \sqrt{\frac{\ln(2)-\ln(\alpha)}{2n}} = \sqrt{\frac{\ln(2/\alpha)}{2n}}$, vale que $\alpha = 2e^{-2n\epsilon^2}$.

$$P_p\left(|\overline{X}_n - p| > \sqrt{\frac{\ln(2/\alpha)}{2n}}\right) \leq \alpha$$

Por lo tanto:

$$P\left(\overline{X}_n - \sqrt{\frac{\ln(2/\alpha)}{2n}} \leq p \leq \overline{X}_n + \sqrt{\frac{\ln(2/\alpha)}{2n}}\right) \geq 1-\alpha$$

Luego:

$$IC_{p,\text{Hoeff}}^{(1-\alpha)100\,\%} = \left[\overline{X}_n - \sqrt{\frac{\ln(2/\alpha)}{2n}}, \overline{X}_n + \sqrt{\frac{\ln(2/\alpha)}{2n}}\right]$$

es un intervalo de confianza uniforme exacto para p de nivel $(1-\alpha)100\,\%$. Notemos que este intervalo vale para cualquier valor de $p \in \Theta$.

IC asintóticos para μ si $X_i \overset{iid}{\sim}$ y, en particular, p si $X_i \overset{iid}{\sim}$ Be(p)

1) Si $X_i \overset{iid}{\sim}$ con σ^2 desconocida, un $IC^{(1-\alpha)100\,\%}_{\mu,as}$ está dado por:

$$IC^{(1-\alpha)100\,\%}_{\mu,as} = \left[\overline{X}_n - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

porque $\sqrt{n}\left(\overline{X}_n - \mu\right)/S \overset{D}{\to} N(0,1)$ por: (a) TCL para $X_i \overset{iid}{\sim}$, (b) $\sigma/S \overset{P}{\to} 1$, (c) Slutsky aplicado a (a) y (b). Este es un caso general que permite dar la distribución asintótica del estadístico W.

2) Si $X_i \overset{iid}{\sim}$ Be(p), si $\hat{p} = \overline{X}_n$, un $IC^{p,as}_{(1-\alpha)100\,\%}$ está dado por:

$$IC^{(1-\alpha)100\,\%}_{p,as} = \left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

porque $\frac{\sqrt{n}\left(\overline{X}_n - p\right)}{\sqrt{\hat{p}(1-\hat{p})}} \overset{D}{\to} N(0,1)$ por (a) TCL para $X_i \overset{iid}{\sim}$, (b) $\frac{\sqrt{p(1-p)}}{\sqrt{\hat{p}(1-\hat{p})}} \overset{P}{\to} 1$, (c) Slutsky aplicado a (a) y (b).

Comentarios sobre $IC^{(1-\alpha)100\,\%}_{p,as}$

Si $p \cdot (1-p) \approx 0$, entonces es preferible utilizar el intervalo de confianza para p construido por la desigualdad de Hoeffding.

Por la fórmula del intervalo, podría potencialmente pasar que, para algunas muestras, el borde inferior del intervalo sea menor que cero o que el borde superior sea mayor que uno.

Por ejemplo, imagine que para ciertos datos muestrales x, se tiene que el intervalo de confianza da $[-0{,}1, 0{,}3]$. En ese caso, podemos decir que el intervalo de confianza es $[0, 0{,}3]$.

De manera análoga, imagine que para ciertos datos muestrales x, se tiene que el intervalo de confianza da $[0{,}8, 1{,}1]$. En ese caso, podemos decir que el intervalo de confianza es $[0{,}8, 1]$.

IC asintóticos de Wald (basados en estimadores asintóticamente normales)

Supongamos ahora más generalmente que $\hat{\theta}_n$ es un estimador de θ que es asintóticamente normal, es decir que cumple:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \overset{D}{\to} N(0, V(\theta))$$

donde $V(\theta)$ es la varianza de la distribución asintótica del estimador $\hat{\theta}_n$. Supongamos que tenemos un estimador consistente de $V(\theta)$. Por ejemplo, consideremos el estimador plug-in $V(\hat{\theta}_n)$ que es consistente para $V(\theta)$. Entonces, por Slutsky:

$$\frac{\sqrt{n}\left(\hat{\theta}_n - \theta\right)}{\sqrt{V(\hat{\theta}_n)}} \overset{D}{\to} N(0,1)$$

Luego, dado un $\alpha > 0$, vale que, cuando n tiende a infinito:

$$P_\theta\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{V(\hat{\theta}_n)}} \leq z_{1-\frac{\alpha}{2}}\right) \overset{n\to+\infty}{\to} 1-\alpha$$

Por lo tanto, un intervalo de confianza de nivel aproximado $(1-\alpha)100\,\%$ para θ basado en el estimador asintóticamente normal $\hat{\theta}_n$ es **conocido como el intervalo de Wald**:

$$IC^{(1-\alpha)100\,\%}_{\theta,as} = \left[\hat{\theta}_n - z_{1-\frac{\alpha}{2}}\frac{\sqrt{V(\hat{\theta}_n)}}{\sqrt{n}}, \hat{\theta}_n + z_{1-\frac{\alpha}{2}}\frac{\sqrt{V(\hat{\theta}_n)}}{\sqrt{n}}\right]$$

IC asintótico en modelo no paramétrico (basado en estimador asint normal)

Sean $X_i \overset{iid}{\sim}$, supongamos que queremos construir un $IC^{95\,\%}_{P(X\leq x_0),as}$. En primer lugar, un estimador del parámetro $F(x_0) = P(X \leq x_0)$ es:

$$\hat{F}_n(x_0) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_0)$$

Notemos que $I(X_i \leq x_0) \overset{iid}{\sim} \text{Ber}(F(x_0))$. $\hat{F}_n(x)$ es simplemente la media muestral de la muestra aleatoria $I(X_1 \leq x), \ldots, I(X_n \leq x)$. Entonces:

$$E(\hat{F}_n(x)) = F(x), \quad \text{Var}(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}, \quad \hat{F}_n(x) \xrightarrow{P} F(x)$$

Por lo tanto:

$$IC^{95\,\%}_{F(x_0),as} = \left[\hat{F}_n(x_0) - 1{,}96\sqrt{\frac{\hat{F}_n(x_0)(1-\hat{F}_n(x_0))}{n}}, \hat{F}_n(x_0) + 1{,}96\sqrt{\frac{\hat{F}_n(x_0)(1-\hat{F}_n(x_0))}{n}}\right]$$

es un intervalo de confianza de nivel asintótico 95 % para $F(x_0)$.

IC puntuales vs IC uniformes

Los IC puntuales tienen un margen de error que depende de algún parámetro conocido o cuyo margen de error es aleatorio.

Los IC uniformes tienen un margen de error que no depende de ningún parámetro conocido ni de la distribución de los datos y que para cualquier valor del parámetro θ el margen de error es el mismo.

Por ejemplo supongamos que $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(\theta)$ con $\theta \in \Theta = (0,1)$. Usando que:

$$W_n = \frac{\sqrt{n}\left(\hat{\theta}_n - \theta\right)}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow{L(F_\theta)} N(0,1)$$

Entonces, C_n es un intervalo de confianza asintótico de nivel aproximado $1-\alpha$:

$$IC^{(1-\alpha)100\,\%}_{\theta,\text{as., punt.}} = \left[\hat{\theta}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}, \hat{\theta}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}\right]$$

Decimos que este intervalo es un **IC puntual** porque la longitud del intervalo depende de $\hat{\theta}_n$. En cambio, si tomáramos el siguiente intervalo, éste es un **IC uniforme**:

$$IC^{(1-\alpha)100\,\%}_{\theta,\text{as., unif.}} = \left[\hat{\theta}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}}, \hat{\theta}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}}\right]$$

Intervalo de Predicción si $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$

Dadas $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, queremos predecir X_{n+1}. Consideremos el estimador $\bar{X}_n$. El **error de predicción (aleatorio) es** $\bar{X}_n - X_{n+1}$. La esperanza del **error de predicción** es:

$$\mathbb{E}\left(\bar{X}_n - X_{n+1}\right) = \mathbb{E}\left(\bar{X}_n\right) - \mathbb{E}\left(X_{n+1}\right) = 0$$

Como X_{n+1} es independiente de X_i para $1 \leq i \leq n$, también es independiente de $\bar{X}_n$. Luego la varianza del **error de predicción** es:

$$\text{Var}\left(\bar{X}_n - X_{n+1}\right) = \text{Var}\left(\bar{X}_n\right) + \text{Var}\left(X_{n+1}\right) = \left(\frac{1}{n} + 1\right)\sigma^2$$

Notemos que $\bar{X}_n - X_{n+1} \sim N\left(0, \left(\frac{1}{n}+1\right)\sigma^2\right)$, entonces

$$Z = \frac{\bar{X}_n - X_{n+1}}{\sigma\sqrt{\frac{1}{n}+1}} \sim N(0,1)$$

Por lo tanto,

$$\frac{\bar{X}_n - X_{n+1}}{S\sqrt{\frac{1}{n}+1}} \sim t_{n-1}$$

De esta forma, el siguiente intervalo es un intervalo de predicción de nivel $(1-\alpha)100\,\%$:

$$\text{IPred}_{X_{n+1}}^{(1-\alpha)100\,\%} = \left[\bar{X}_n - t_{1-\frac{\alpha}{2},n-1}S\sqrt{1+\frac{1}{n}}, \bar{X}_n + t_{1-\frac{\alpha}{2},n-1}S\sqrt{1+\frac{1}{n}}\right]$$

¿Por qué en el intervalo de predicción aparece un factor $\sqrt{1+\frac{1}{n}} > 1$ que no aparece en el intervalo de confianza si $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$?

El error de predicción es una resta de dos variables aleatorias $\bar{X}_n - X_{n+1}$, mientras que el error de estimación es una resta entre una variable aleatoria y un valor fijo $\bar{X}_n - \mu$.

El intervalo de predicción es más ancho que el intervalo de confianza porque hay más variabilidad en el error de predicción (debido a X_{n+1}) que en el error de estimación. De hecho, cuando $n \to \infty$, el intervalo de confianza se reduce a μ, pero el intervalo de predicción se convierte en $\mu \pm z_{1-\frac{\alpha}{2}}\sigma$.

Estadística bayesiana

Paradigmas de la probabilidad

En el **paradigma bayesiano la probabilidad se entiende como una medida de incertidumbre** acerca de la veracidad de cualquier proposición.

Sean n variables aleatorias cada una con distribución Bernoulli, originadas por un "ensayo" incierto repetitivo y queremos conocer la proporción de personas empleadas (1 = empleo).

- **Si soy frecuentista,** plantearé que $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, donde θ es un valor fijo, aunque desconocido, que mide la proporción de éxitos que obtendría si pudiera repetir el ensayo infinitas veces. Asímismo estimaré a θ con, por ejemplo, $\delta(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ y calcularé el desvío estándar estimado:$\hat{SE} = \frac{1}{n}X_n(1-X_n)$ que cuantifica el comportamiento de $\delta(X)$ en repetidos experimentos o "muestras".

- **Si soy bayesiano**, será muy poco razonable que asuma que $X_i \overset{\text{indep}}{\sim}$ ya que mi creencia antes del experimento debería verse afectada luego de que se me hayan revelado los valores realizados por $X_1, \ldots, X_{i-1}$.

Ambos paradigmas asumen: x es la realización de $X \sim f(\cdot;\theta)$ con $\theta \in \Theta$. No obstante, en cada paradigma la interpretación de $f(\cdot;\theta)$ es diferente.

Proceso del bayesiano

Si soy **bayesiano** mi inferencia seguirá estos pasos asumiendo que:

(a) Existe una variable aleatoria θ con soporte $[0,1]$, $\theta \sim \pi(\theta)$, llamada **distribución a priori**.

(b) Dado θ, las variables aleatorias X_i son intercambiables, es decir, $X_i|\theta \overset{\text{iid}}{\sim} f_X(x_i|\theta)$.

$$f_X(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta).$$

(c) Usando $\pi(\theta)$, calcularé la **distribución a posterior**i de θ, que cuantifica mi probabilidad acerca de θ luego de conocer el resultado de los n ensayos:

$$\pi^*(\theta|X) = \frac{f_X(x_1, \ldots, x_n|\theta)\pi(\theta)}{\int_\Theta f_X(x_1, \ldots, x_n|\theta)\pi(\theta)d\theta}$$

Teoremas de Finetti y Hewitt-Savage

Teorema 64 (Finetti)**.** *Una sucesión $\{X_i\}_{i=1}^{\infty}$ de variables aleatorias Bernoulli es intercambiable si y sólo si*

1. *Con probabilidad 1, existe la v.a. L cuya densidad designamos con $\pi(\cdot)$.*

$$L = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i$$

2. *Dada $L = \theta$, las variables aleatorias $\{X_i\}_{i=1}^{\infty}$ son iid Ber(θ).*

Teorema 65 (Hewitt-Savage)**.** *Una sucesión $\{X_i\}_{i=1}^{\infty}$ de variables aleatorias cualesquiera es intercambiable si y solo si:*

1. *Con probabilidad 1, existe la función de distribución acumulada aleatoria:*

$$\mathbf{F}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,t]}(X_i)$$

2. *Dado un valor de $\mathbf{F}(\cdot) = F(\cdot)$, las variables aleatorias $\{X_i\}_{i=1}^{\infty}$ son iid, cada una con distribución acumulada $F(\cdot)$.*

Inferencia bayesiana para variables aleatorias intercambiables

Asumimos que $X_1, \ldots, X_n$ son los primeros n elementos de una hipotética infinita sucesión $\{X_i\}_{i=1}^{\infty}$ de variables aleatorias intercambiables.

De acuerdo a los teoremas de De Finetti y Hewitt-Savage, existe una distribución acumulada aleatoria $\mathbf{F}(\cdot)$ tal que, dado $\mathbf{F}(\cdot) = F(\cdot)$, $\{X_i\}_{i=1}^{\infty}$ son iid, cada una con distribución F.

- **Si las X_i son binarias**, $\mathbf{F}(\cdot)$ queda enteramente determinada por la frecuencia relativa $\mathbf{L} = \mathbf{1} - \mathbf{F}(\mathbf{0})$ de ocurrencia de éxito en el límite. El bayesiano especifica la distribución "a priori" $\pi(\theta)$ de $\mathbf{L}$, la cual codifica su incertidumbre acerca del límite $\mathbf{L}$.
- **Si las X_i son continuas**, el bayesiano debe especificar su distribución a priori (codificando su incertidumbre) acerca de la función límite $\mathbf{F}(t)$, $t \in \mathbb{R}$.

En estadística bayesiana paramétrica, el bayesiano típicamente elimina de llano, es decir, le asigna a priori probabilidad o a todas las distribuciones $\mathbf{F}(\cdot)$, excepto aquellas que pertenecen a una familia:

$$\mathcal{F} = \{F(\cdot;\theta) : \theta \in \Theta\}, \quad \Theta \subseteq \mathbb{R}^k \text{ para algún } k \in \mathbb{N}$$

Distribuciones a priori y a posteriori

Para variables aleatorias X_i intercambiables, para hacer inferencia sobre θ necesitaremos:

- **Distribución a priori** de θ: $\pi(\theta)$
- **Distribución a posteriori** (ignoramos la constante de integración que no depende de θ):

$$\pi^*(\theta|x) = \frac{\pi(\theta) \cdot f(\underline{x}|\theta)}{\int_\Theta \pi(u) \cdot f(x|u)\, du} \propto \pi(\theta) \cdot f(\underline{x}|\theta)$$

La **verosimilitud** es la función de distribución de $\underline{X}$ dado que $\theta^* = \theta$, cuya variable es θ para $\underline{x}$ dada. Nótese que $f(\underline{x}|\theta)$ coincide con la $f(\underline{x};\theta)$ del frecuentista.

$$L_n(\theta) = f(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Luego, la distribución a posteriori es proporcional al producto de la distribución a priori por la verosimilitud:

$$\pi^*(\theta|\underline{x}) \propto \pi(\theta) \cdot L_n(\theta)$$

Principio de verosimilitud

Típicamente, el bayesiano determina la clase $\mathcal{F}$ en base a consideraciones acerca del problema en cuestión que esté analizando. Como $F(\cdot)$ pertenece a $\mathcal{F}$, entonces debe existir un $\theta^* \in \Theta$ tal que $\mathbf{F} = F(\cdot;\theta^*)$. Ese θ^* es aleatorio porque $\mathbf{F}$ es aleatorio.

Habiendo otorgado probabilidad a priori 1 a la proposición $\mathbf{F}(\cdot)$ pertenece a $\mathcal{F}$, el bayesiano completa su especificación de la distribución a priori de $\mathbf{F}$, simplemente especificando su distribución a priori $\pi(\theta)$ de θ^*.

Habiendo observado los resultados $x_1, \ldots, x_n$ de las primeras n variables aleatorias $X_1, \ldots, X_n$, de la sucesión infinita $\{X_i\}_{i=1}^{\infty}$, la inferencia bayesiana simplemente consiste en actualizar la distribución sobre θ^* empleando la regla de Bayes.

En la inferencia bayesiana, la inferencia se basa en la distribución a posteriori: $\pi^*(\theta|\underline{x}) \propto \pi(\theta) \cdot L_n(\theta)$ es decir, los datos $\underline{x}$ intervienen solo a través de la verosimilitud, y si **dos muestras resultan en la misma distribución a posteriori, la inferencia será la misma. A esto se lo conoce como el principio de verosimilitud**.

Verosimilitud

¿Qué NO es la verosimilitud? No es una densidad sobre θ. Entonces podemos definir la verosimilitud para dos casos

- Si X es discreta, la verosimilitud es: $L_n(\theta) = P(X = x|\theta)$ Es decir, es la probabilidad de observar $X = x$ si el valor desconocido del parámetro θ^* fuera θ.
- Si X es continua y $\underline{\epsilon} = \epsilon \cdot 1$ con $\epsilon \approx 0$, entonces: $L_n(\theta) \cdot \epsilon^n \approx P(\underline{x} \leq \underline{X} \leq \underline{x} + \underline{\epsilon}|\theta)$
  Luego, la razón de verosimilitudes: $\lambda(\theta, \theta') = \frac{L_n(\theta)}{L_n(\theta')}$ cuantifica cuánto más o menos probable sería observar $\underline{X} = \underline{x}$ si θ^* fuera θ que si fuera θ'.

Ejemplo 66 (Verosimilitud)**.** Supongamos que $X_1, \ldots, X_4|\theta \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, y observamos $x = (1, 0, 1, 1)$. Entonces, la verosimilitud es: $L_n(\theta) = \theta^3(1 - \theta)$ Si, en cambio, observamos $x = (1, 0, 0, 1)$, la verosimilitud es: $L_n(\theta) = \theta^2(1 - \theta)^2$

Mejor predictor de θ^* según la función de pérdida

Supongamos que queremos predecir el valor de una variable aleatoria θ^* con un número fijo δ. Si asignamos una pérdida $L(\theta^*, \delta)$ a predecir θ^* con δ, entonces la mejor predicción es aquel valor δ^* que minimiza la esperanza de $L(\theta^*, \delta)$:

$$\delta^* = \arg\min_{\delta} \mathbb{E}[L(\theta^*, \delta)] = \arg\min_{\delta} R(\theta^*, \delta)$$

Tenemos dos casos

- **Para variables aleatorias θ^* continuas**, una función de pérdida usada es $L(\theta^*, \delta) = (\theta^* - \delta)^2$. Dado que:

  $$\mathbb{E}[(\theta^* - \delta)^2] \geq \mathbb{E}[(\theta^* - \mathbb{E}(\theta^*))^2]$$

  la mejor predicción de θ^* es la esperanza de θ^* o sea $\delta^* = \mathbb{E}(\theta^*)$
- **Para variables aleatorias θ^* discretas**, una función de pérdida usada es $L(\theta^*, \delta) = 1 - I_{\{\theta^*\}}(\delta)$. En este caso, tenemos:

  $$\mathbb{E}[L(\theta^*, \delta)] = 1 - P(\theta^* = \delta)$$

  Para esta función de pérdida, la mejor predicción de θ^* es la moda de θ^* o sea, $\delta^* = \arg\max_{\delta} P(\theta^* = \delta)$

Mejor predicción de θ^* según verosimilitud

Habiendo observado los datos $\underline{x}$ de n variables $X_1, \ldots, X_n$ intercambiables, un bayesiano desea predecir el valor de la variable aleatoria θ^* que se corresponde con la distribución límite $F(\cdot)$ del Teorema de Hewitt-Savage.

Habiendo observado x, el bayesiano considera que $\theta^* \sim \pi^*(\theta|\underline{x})$.

Dada una función de pérdida $L(\theta^*, \delta)$, la mejor predicción es el valor $\delta(\underline{x})$ que minimiza la esperanza condicional de $L(\theta^*, \delta)$ dado $\underline{X} = \underline{x}$:

$$\delta(\underline{x}) = \arg\min_{\delta} \mathbb{E}[L(\theta^*, \delta(\underline{x}))|\underline{X} = \underline{x}]$$

Entonces tenemos dos casos

- Si, dado $\underline{x}$, θ^* toma valores en un **conjunto continuo** y usamos $L(\theta^*, \delta) = (\theta^* - \delta)^2$, entonces la mejor predicción es la esperanza de θ^*:

$$\delta(\underline{x}) = \mathbb{E}(\theta^*|X = x)$$

- Si, dado $\underline{x}$, θ^* toma valores en un **conjunto finito** o numerable, y usamos $L(\theta^*, \delta) = 1 - I_{\{\delta\}}(\theta^*)$, entonces la mejor predicción es la moda de la distribución de θ^* dada $\underline{X} = \underline{x}$:

$$\delta(\underline{x}) = \text{moda}(\theta^*|\underline{X} = \underline{x})$$

Notemos que, en realidad, el bayesiano no estima sino que predice.

Teorema de Bernstein-von Mises

Teorema 67 (Bernstein-von Mises)**.** *Si el tamaño de muestra $n \to \infty$:*

1. *La distribución a posteriori se concentra alrededor de $\hat{\theta} = \arg\max_\theta L_n(\theta)$.*
2. *Además, la distribución a posteriori se aproxima más y más a una distribución normal con media $\hat{\theta}$ y varianza $n^{-1}V$, donde:*

$$V = -n^{-1} \frac{\partial^2}{\partial \theta^2} \ln L_n(\theta) \Big|_{\theta=\hat{\theta}}$$

Aquí, $\hat{\theta} = \arg\max_\theta L_n(\theta)$ es el estimador de máxima verosimilitud de la estadística clásica

Intervalos de credibilidad

Definición 68 (Intervalo de credibilidad)**.** Para $\theta \in \mathbb{R}$, un intervalo de credibilidad $1 - \alpha$ es cualquier intervalo (θ_L, θ_U) tal que:

$$P(\theta_L < \theta^* < \theta_U|\underline{X} = \underline{x}) = 1 - \alpha$$

donde θ_L y θ_U dependen de $\underline{x}$.

Definición 69 (Intervalo de credibilidad de mínima longitud)**.** El intervalo (θ_L, θ_U) es un intervalo de credibilidad $1-\alpha$ de mínima longitud si cualquier otro intervalo (θ'_L, θ'_U) de credibilidad $1-\alpha$ satisface:

$$\theta'_L - \theta'_U \geq \theta_L - \theta_U$$

Observación 70. Si la distribución a posteriori $\pi^*(\theta|x)$ es simétrica respecto de su media y unimodal, el intervalo de credibilidad $1-\alpha$ de mínima longitud es el intervalo de probabilidad igual en los extremos (*equal tail probability interval*) que usa los percentiles $100 \cdot \frac{\alpha}{2}$ y $100 \cdot \left(1 - \frac{\alpha}{2}\right)$.

Cuando la distribución a posteriori no es simétrica, un bayesiano suele reportar los *equal tail probability intervals*, a pesar de no ser de mínima longitud.

Los intervalos de credibilidad para θ se construyen usando la distribución a posteriori.

Diferencias entre intervalos de credibilidad y intervalos de confianza

- En un intervalo de confianza (I.C.) frecuentista, el nivel $1-\alpha$ es una propiedad del comportamiento en hipotéticas infinitas muestras repetidas del procedimiento que se usó para calcular el intervalo, y no sobre el intervalo específico calculado con la muestra observada x.
- En un intervalo de credibilidad (I.Cred.) bayesiano, el nivel $1-\alpha$ se refiere a la probabilidad de cobertura que tiene el intervalo calculado con la muestra x.

Ejemplo 71. La distribución a posteriori es $\theta^*|\underline{X} = \underline{x} \sim \Gamma(n\bar{x}+\alpha, n+\beta)$. Usando que, si $X \sim \Gamma(k,\theta)$, entonces para $c > 0$, se tiene que $cX \sim \Gamma(k, \theta/c)$. Tomando $c = 2(n+\beta)$, se obtiene que:

$$2(n+\beta)\theta^* \sim \Gamma(n\bar{x}+\alpha, 0{,}5)$$

Si, por ejemplo, α fuera un número natural, entonces:

$$2(n+\beta)\theta^*|X = x \sim \Gamma(n\bar{x}+\alpha, 0{,}5) = \chi^2_m \quad \text{donde} \quad m = 2(n\bar{x}+\alpha)$$

Por lo tanto, un intervalo de credibilidad 95 % (que no es el de mínima longitud) para θ^* es:

$$IC_{95\,\%\theta^*} = \left(\frac{\chi^2_{m,0,025}}{2(n+\beta)}, \frac{\chi^2_{m,0,975}}{2(n+\beta)} \right)$$

donde $m = 2(n\bar{x}+\alpha)$. Si $n \gg 0$, se tiene que: $\chi^2_m \approx N(m, 2m), \quad m = 2(n\bar{x}+\alpha)$, $\frac{n\bar{x}+\alpha}{n+\beta} = \frac{\bar{x}+\frac{\alpha}{n}}{1+\frac{\beta}{n}} \approx \bar{x}$ y $\frac{n\bar{x}+\alpha}{(n+\beta)^2} \approx \frac{1}{n\bar{x}}$.

Un intervalo de credibilidad aproximado 95 % es:

$$IC^{95\,\%}_{\text{aprox},\theta^*} \approx \left[\bar{x} - 1{,}96\sqrt{\frac{\bar{x}}{n}}, \bar{x} + 1{,}96\sqrt{\frac{\bar{x}}{n}} \right]$$

Probabilidad predictiva

Habiendo observado $X_1 = x_1, \ldots, X_n = x_n$, queremos hacer una predicción sobre el valor de X_{n+1}. Por lo tanto, el bayesiano calcula:

$$f(x_{n+1}|x_1, \ldots, x_n) = \int_\Theta f(x_{n+1}|x_1, \ldots, x_n, \theta)\pi^*(\theta|x_1, \ldots, x_n)\, d\theta$$

Esta integral se conoce como la distribución predictiva (**predictive distribution**).

Si las variables aleatorias $X_1, X_2, \ldots$ son intercambiables, la distribución predictiva se reduce a:

$$f(x_{n+1}|x_1, \ldots, x_n) = \int_\Theta f(x_{n+1}|\theta)\pi^*(\theta|x_1, \ldots, x_n)\, d\theta$$

Ejemplo 72 (Distribución predicitiva)**.** Calculemos la distribución predictiva usando que $\pi^*(\theta|x) = \text{Beta}(\alpha_n, \beta_n)$:

$$p(x_{n+1}|x_n) = \frac{1}{B(\alpha_n, \beta_n)} \int_0^1 \theta^{x_{n+1}+\alpha_n-1}(1-\theta)^{1-x_{n+1}+\beta_n-1}\, d\theta$$

Sabemos que:

$$p(x_{n+1}|x_n) = \frac{B(x_{n+1}+\alpha_n, 1-x_{n+1}+\beta_n)}{B(\alpha_n, \beta_n)}$$

Finalmente, obtenemos la probabilidad predictiva como:

$$p(x_{n+1}|x_n) = \frac{\alpha_n^{x_{n+1}}\beta_n^{1-x_{n+1}}}{\alpha_n + \beta_n}$$

Distribución a priori conjugada

Definición 73. La distribución a priori es **conjugada** para el modelo $f(\underline{x}|\theta)$ si la distribución a priori y la distribución a posteriori pertenecen a la misma familia.

Ejemplos 74. Podemos verlo en los siguientes casos

1. $\pi(\theta) = \text{Beta}(\alpha, \beta)$ y $\pi^*(\theta|x) = \text{Beta}(\alpha + n\bar{x}, \beta + n(1-\bar{x}))$.
2. $\pi(\theta) = \Gamma(\alpha, \beta)$ y $\pi^*(\theta|x) = \Gamma(n\bar{x} + \alpha, n + \beta)$.

Distribuciones no informativas

Una de las principales críticas al enfoque bayesiano es la elección de la distribución a priori. En particular, los bayesianos objetivos buscan **distribuciones a priori que influyan lo menos posible en la distribución a posteriori**, conocidas como **distribuciones no informativas**.

Máxima Verosimilitud

Estimador de máxima verosimilitud de θ

Definición 75. Para una muestra aleatoria $\underline{X}$ se denomina estimador de máxima verosimilitud de θ^* y abreviamos $\widehat{\theta}_{MV}$ a

$$\widehat{\theta}_{MV} = \arg\max_{\theta\in\Theta} f(\underline{X}; \theta)$$

Note que en la definición de $\widehat{\theta}_{MV}$, la función f está evaluada en el vector aleatorio $\underline{X}$ y no en el valor observado $\underline{x}$, porque $\widehat{\theta}_{MV}$ es un estimador de θ^*, y no el valor estimado de θ^*.

- La función $\mathcal{L}_n(\theta) = f(\underline{X};\theta)$ es la función de verosimilitud para una muestra aleatoria X.
- La inferencia bayesiana considera $L_n(\theta)$ porque toma una muestra particular.

En la inferencia frecuentista, cuando pensamos al proceso de resumir datos para calcular el $\widehat{\theta}_{MV}$, la función $\mathcal{L}_n(\theta)$ es aleatoria porque depende del vector aleatorio $\underline{X}$. Notemos también que si las $X_1, \ldots, X_n \overset{iid}{\sim} f(\cdot;\theta)$ entonces

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

Log-verosimilitud

A partir de la función de verosimilitud $\mathcal{L}_n = f(\underline{X};\theta)$ definimos la log-verosimilitud $\ell_n(\theta) = \ln \mathcal{L}_n$, donde ln es el logaritmo natural (en base e). Notemos que como el $\ln(u)$ es una función estrictamente creciente, entonces

$$\widehat{\theta}_{MV} = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta) = \arg\max_{\theta\in\Theta} \ell_n(\theta)$$

Por lo tanto, el método de máxima verosimilitud consiste en, para una muestra aleatoria $\underline{X}$, obtener el estimador de θ que maximiza $L_n(\theta)$ (y en el caso que sea posible, y porque es más cómodo de calcular, el estimador que maximice $\ell_n(\theta)$).

Propiedades del estimador de máxima verosimilitud

Proposición 76 (Propiedad de invarianza en el método de MV)**.** *: Para una función $\beta : \Theta \to \mathbb{R}^k$, el estimador de máxima verosimilitud de $\beta(\theta)$ se define como el estimador plug-in:*

$$\widehat{\beta(\theta)}_{MV} = \beta(\widehat{\theta}_{MV})$$

Proposición 77 (Propiedad de consistencia en el método de MV)**.** *Bajo ciertas condiciones generales el estimador de máxima verosimilitud basado en n v.a iid es consistente, es decir:*

$$\widehat{\theta}_{MV} \xrightarrow{P} \theta$$

Métodos numéricos

Cuando no podemos encontrar analíticamente $\arg\max \ell_n(\theta)$, debemos recurrir a métodos numéricos iterativos que, en realidad, no dan la solución exacta sino que la aproximan tanto como se desee. Un tal método numérico es el llamado **método de Newton-Raphson**. El método en realidad encuentra una solución de la ecuación de primer orden $\ell_n'(\theta) = 0$, siendo $\ell_n'(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta)$.

Partiendo de un valor inicial cualquiera, digamos θ_0, en el paso $k+1$, se calcula el valor de θ que resuelve la ecuación:

$$0 = \ell_n'(\theta_k) + \ell_n''(\theta_k)^\top \cdot (\theta - \theta_k)$$

siendo $\ell_n''(\theta) = \frac{\partial^2}{\partial\theta\partial\theta^\top}\ell_n(\theta)$ la matriz de segundas derivadas de $\ell_n(\theta)$. La fórmula en la parte derecha es la fórmula del plano (o recta si θ es escalar) tangente al gráfico de la función $\ell_n(\theta)$ en el punto θ_k. El valor de θ que resuelve la ecuación es

$$\theta_{k+1} = \theta_k - \left[\ell_n''(\theta_k)\right]^{-1} \ell_n'(\theta_k)$$

El algoritmo se detiene cuando se alcanza alguna condición preestablecida de convergencia, por ejemplo $\|\theta_{k+1} - \theta_k\| < \epsilon$ para un ϵ dado.

Función score

Definición 78 (Función score)**.** Sea $\underline{X} = (X_1, \ldots, X_n)$ el vector compuesto por las observaciones de la muestra, se define a la función **score** de θ que escribimos como $s(\underline{X}, \theta)$ como:

$$s(\underline{X};\theta) = \begin{pmatrix} s_1(\underline{X};\theta) \\ s_2(\underline{X};\theta) \\ \vdots \\ s_r(\underline{X};\theta) \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \ln(f(\underline{X};\theta))}{\partial\theta_1} \\ \dfrac{\partial \ln(f(\underline{X};\theta))}{\partial\theta_2} \\ \vdots \\ \dfrac{\partial \ln(f(\underline{X};\theta))}{\partial\theta_r} \end{pmatrix}$$

Si el modelo es uniparamétrico, entonces $s(\underline{X}, \theta) = \dfrac{\partial \ln f(\underline{x};\theta)}{\partial\theta} = \ell_n'(\theta)$.

Lema 79. *Supongamos que:*

1. Θ *es abierto.*
2. $f(\underline{x};\theta)$ *tiene el mismo soporte cualquiera sea* $\theta \in \Theta$.
3. $f(\underline{x};\theta)$ *es diferenciable con respecto a* θ*, para cada* $\underline{x}$.
4. *Vale que podemos intercambiar derivadas e integrales. Es decir, para todo* $\theta^* \in \Theta$ $\frac{\partial}{\partial\theta_j}\left[\int f(\underline{x};\theta)d\underline{x}\right]\big|_{\theta=\theta^*} = \int \frac{\partial}{\partial\theta_j} f(\underline{x};\theta)\Big|_{\theta=\theta^*} d\underline{x}$

Entonces:

$$\mathbb{E}_\theta\left[s(\underline{X};\theta)\right] = 0$$

Si el modelo es uniparamétrico, esta propiedad nos dice que $\mathbb{E}_\theta\left[\ell_n'(\theta)\right] = 0$.

Matriz de información

Definición 80 (Matriz de información)**.** Se define a la **matriz de información** como:

$$I_n(\theta) = \mathbb{E}_\theta \left[s(\underline{X};\theta)\ s(\underline{X};\theta)^T \right]$$

Definición 81 (Igualdad de información)**.** Definimos a la **igualdad de información** como:

$$\mathbb{E}_\theta \left[s(\underline{X};\theta)\ s(\underline{X};\theta)^T \right] = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln f(\underline{X};\theta) \right]$$

Cuando el modelo es uniparamétrico (θ es un escalar), la **información** se reduce a:

$$I_n(\theta) = \mathbb{E}_\theta \left[s(\underline{X};\theta)^2 \right] = \mathbb{E}_\theta \left[\ell_n'(\theta)^2 \right]$$

Y la **igualdad de información** a:

$$\mathbb{E}_\theta \left[s(\underline{X};\theta)^2 \right] = -\mathbb{E}_\theta \left[\frac{\partial^2 \ln f(\underline{X};\theta)}{\partial\theta^2} \right]$$
$$\mathbb{E}_\theta \left[\ell_n'(\theta)^2 \right] = -\mathbb{E}_\theta \left[\ell_n''(\theta) \right]$$

Observación 82. $I_n(\theta)$ es la información (o matriz de información) sobre θ basada en n observaciones (por eso el subíndice). A menudo la llamamos matriz de información basada en toda la muestra. Por el contrario, cuando escribimos $I_1(\theta)$ o simplemente $I(\theta)$, nos referimos a la información sobre θ **por unidad de muestra.**

Corolario 83. *Cuando $X_1, \ldots, X_n$ son iid. obtenemos que:*

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

De manera que cuando las observaciones de la muestra son iid, entonces la información basada en n observaciones es igual a n veces la información por unidad.

Normalidad asintótica de el estimador de máxima verosimilitud

Teorema 84 (Normalidad asintótica)**.** *Supongamos que* $X_i \overset{iid}{\sim} f(x) \in \mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^r\}$*. Bajo condiciones de regularidad sobre las distribuciones en la familia* $\mathcal{F}$ *se verifica que:*

$$\sqrt{n}\left(\widehat{\theta}_{MV} - \theta\right) \xrightarrow{D} N_r\left(0, I_1(\theta)^{-1}\right)$$

donde $N_r(0, I_1(\theta)^{-1})$ *es la distribución normal multivariada de dimensión* r *con vector de medias* 0 *y matriz de varianzas y covarianzas* $I_1(\theta)^{-1}$*.*

Análogamente se puede expresar esta propiedad como:

$$\sqrt{nI_1(\theta)}\left(\widehat{\theta}_{MV} - \theta\right) \xrightarrow{D} N_r(0,1)$$

ó

$$\sqrt{I_n(\theta)}\left(\widehat{\theta}_{MV} - \theta\right) \xrightarrow{D} N_r(0,1)$$

Corolario 85. *Si* $\beta(.) : \Theta \to \mathbb{R}^p$ *es una función diferenciable de* θ *en cada* $\theta \in \Theta$*, entonces* $\widehat{\beta}_{MV} = \beta(\widehat{\theta}_{MV})$ *verifica:*

$$\sqrt{n}\left[\widehat{\beta}_{MV} - \beta(\theta)\right] \xrightarrow{D} N_p(0, V(\theta))$$

donde:

$$V(\theta) = \frac{\partial \beta(\theta)}{\partial \theta'} I_1(\theta)^{-1} \frac{\partial \beta(\theta)}{\partial \theta}$$

Si $\beta(.) : \Theta \to \mathbb{R}$*, entonces:*

$$\sqrt{n}\left[\widehat{\beta}_{MV} - \beta(\theta)\right] \xrightarrow{D} N_p\left(0, \frac{(\beta'(\theta))^2}{I_1(\theta)}\right)$$

Modelo mal especificado

Definición 86 (Divergencia de Kullback-Leibler)**.** La divergencica de Kullback-Leibler que mide la distancia entre dos funciones de densidad, f_0 y $f(\cdot,\theta)$ y se define como

$$D(f_0|f(\cdot,\theta)) = -\int_{\text{Sop}(X)} \ln\left(\frac{f(x,\theta)}{f_0(x)}\right) f_0(x)dx = -E_0\left[\ln\left(\frac{f(X,\theta)}{f_0(X)}\right)\right]$$

Notemos que si f_0 fuera igual a $f(\cdot,\theta)$, entonces $D(f_0\|f(\cdot,\theta)) = 0$. Además, por la desigualdad de Jensen, como $g(x) = -\ln(x)$ es convexa, sabemos que

$$-E_0\left[\ln\left(\frac{f(X,\theta)}{f_0(X)}\right)\right] \geq -\ln\left[E_0\left(\frac{f(X,\theta)}{f_0(X)}\right)\right]$$
$$-\ln\left[E_0\left(\frac{f(X,\theta)}{f_0(X)}\right)\right] = 0$$

Observación 87. Si $f_0 \notin \mathcal{F}$, o sea, si el modelo está mal especificado se prueba que $\widehat{\theta}_{MV}$ minimiza la divergencia de Kullback-Leibler encontrando θ^* de manera que

$$D(f_0|f(\cdot,\theta^*)) \leq D(f_0\|f(\cdot,\theta)) \quad \forall\theta \in \Theta$$

Eficiencia asintótica del estimador de máxima verosimilitud

Desigualdad de Cramer-Rao

Teorema 88 (Desigualdad de Cramer-Rao)**.** *Supongamos que $\widehat{\theta} = \delta(X_1, \ldots, X_n)$ es un estimador **insesgado** de θ. Sea $I_n(\theta)$ la información de θ basada en las n observaciones. Entonces:*

$$Var_\theta\left(\widehat{\theta}\right) \geq \frac{1}{I_n(\theta)}$$

Si $\theta \in \mathbb{R}^r$, la desigualdad se expresa como:

$$Var_\theta(\widehat{\theta}) - I_n(\theta)^{-1} \geq 0$$

*Lo que significa que la matriz $Var_\theta(\widehat{\theta}) - I_n(\theta)^{-1}$ es **semi-definida positiva**.*

La desigualdad de Cramer-Rao nos pone una cota **inferior** a la varianza de cualquier estimador insesgado.

Eficiencia asintótica

Definición 89 (Estimador asintóticamente normal)**.** $\widehat{\beta}$ se dice que es un estimador asintóticamente normal de un parámetro $\beta(\theta)$ si satisface:

$$\sqrt{n}(\widehat{\beta} - \beta(\theta)) \xrightarrow{L(F_\theta)} N(0, W(\theta))$$

para algún $W(\theta)$.

Con muestras grandes donde $n \gg 0$, preferimos al estimador que tenga el menor $W(\theta)$.

Definición 90 (Eficiencia asintótica)**.** Supongamos que dos estimadores asintóticamente normales $\widehat{\beta}$ y $\widetilde{\beta}$ son tales que:

$$\sqrt{n}\left(\widehat{\beta} - \beta(\theta)\right) \xrightarrow{D} N(0, V_1(\theta))$$
$$\sqrt{n}\left(\widetilde{\beta} - \beta(\theta)\right) \xrightarrow{D} N(0, V_2(\theta))$$

Entonces diremos que $\widehat{\beta}$ es **asintóticamente más eficiente** que $\widetilde{\beta}$ si:

$$V_1(\theta) < V_2(\theta) \forall \theta \in \Theta$$

Definición 91 (Eficiencia asintótica relativa)**.** Dados dos estimadores $\widehat{\beta}$ y $\widetilde{\beta}$ de un parámetro escalar $\beta(\theta)$, que satisfacen (1) y (2) respectivamente, la cantidad

$$\tau(\theta) = \frac{V_2(\theta)}{V_1(\theta)}$$

se denomina la **eficiencia asintótica relativa** de $\widehat{\beta}$ con respecto a $\widetilde{\beta}$. Si $\widehat{\beta} = \widehat{\beta}_{MV}$, se suele poner $V_1(\theta) = I_1(\theta)^{-1}$ en el denominador.

Interpretamos a $\tau(\theta)$ como un indicador de cuánto más grande o más pequeña debe ser la muestra que debemos tener cuando usamos $\widetilde{\beta}$ para obtener la misma precisión que si hubiésemos usado $\widehat{\beta}$.

Por ejemplo, si $\tau(\theta) = 2$, entonces deberemos usar una muestra el doble de grande si usamos $\widetilde{\beta}$ que si usamos $\widehat{\beta}$ para obtener la misma precisión en la estimación.

Relación entre tamaños de muestra Esto se puede ver de la siguiente manera. Si n_1 es el tamaño de la muestra con la que calculamos $\widehat{\beta}$ y n_2 es el tamaño de la muestra con la que calculamos $\widetilde{\beta}$, si queremos obtener la misma precisión, los tamaños muestrales deben cumplir:

$$\frac{V_1(\theta)}{n_1} = \frac{V_2(\theta)}{n_2},$$

o equivalentemente,

$$\frac{n_2}{n_1} = \frac{V_2(\theta)}{V_1(\theta)} = \tau(\theta).$$

De modo que

$$n_2 = \tau(\theta) n_1.$$

Cuanto mayor sea $\tau(\theta)$, más eficiente (asintóticamente) será $\widehat{\beta}$ con respecto a $\widetilde{\beta}$.

Eficiencia del estimador de máxima verosimilitud

Definición 92 (Estimador de máxima verosimilitud asintóticamente eficiente)**.** Como **el estimador de máxima verosimilitud** de $\beta(\theta)$ es asintóticamente normal bajo condiciones de regularidad y su varianza asintótica es: $\frac{\partial \beta(\theta)}{\partial \theta^\top} I(\theta)^{-1} \frac{\partial \beta(\theta)}{\partial \theta}$, entonces **es asintóticamente eficiente**.

Observación 93. Si bien es posible que dado un modelo F y un parámetro $\beta(\theta)$, se dé alguna de las siguientes situaciones:

1. No exista ningún estimador insesgado de $\beta(\theta)$.
2. Existan estimadores insesgados de $\beta(\theta)$, pero ninguno tenga varianza igual a la Cota de Cramer-Rao.

La normalidad asintótica y la eficiencia del estimador de máxima verosimilitud bajo una gran cantidad de modelos F implican que bajo esos modelos, con muestras grandes ($n \gg 0$), es posible obtener un estimador casi insesgado de $\beta(\theta)$ cuya varianza es casi igual a la cota de Cramer-Rao. Este estimador es precisamente el estimador de máxima verosimilitud de $\beta(\theta)$.

Información

La eficiencia asintótica de $\widehat{\beta}(\theta)_{\mathrm{MV}}$ puede interpretarse como indicando que este estimador es el que, con muestras donde $n \gg 0$, extrae toda la información disponible en los datos sobre el parámetro de interés $\beta(\theta)$.

A raíz de esto, a la inversa de la varianza de la distribución asintótica de $\widehat{\beta}(\theta)_{\mathrm{MV}}$, es decir, a la matriz (o al escalar, si $\beta(\theta)$ es escalar):

$$I_{\beta(\theta)} = \left[\frac{\partial \beta(\theta)}{\partial \theta^\top} I_1(\theta)^{-1} \frac{\partial \beta(\theta)}{\partial \theta} \right]^{-1}$$

se la denomina información acerca de $\beta(\theta)$ bajo el modelo F en θ.

En particular, si $\beta(\theta) = \theta$, la información $I_\theta(\theta)$ acerca de θ bajo el modelo F en θ coincide precisamente con $I_1(\theta)$.

Pérdida de información por usar estimadores ineficientes

Cuando usamos un estimador asintóticamente normal de $\beta(\theta)$ pero ineficiente, con varianza de su distribución asintótica, digamos igual a $W(\theta)$, es posible calcular cuánto hemos perdido en información por no haber usado el estimador de máxima verosimilitud.

La cantidad:

$$\frac{I_{\beta(\theta)} - W(\theta)^{-1}}{I_{\beta(\theta)}}$$

nos indica la fracción de información disponible que fue perdida por usar el estimador ineficiente.

Test de Hipótesis

Contraste de hipotesis

Pasos en el proceso del contraste de hipótesis

1. Se postulan dos opciones acerca de la distribución que generó los datos (llamada la distribución poblacional).
   a) La primera opción es una afirmación acerca de la distribución poblacional que **representa el "status quo"**. Es aquella que **de ser verdadera, llevará a la decisión de no modificar el curso de acción corriente**. Los datos se recogen precisamente para desafiar esta afirmación. A esta afirmación **se la denomina hipótesis nula**, y se la simboliza con H_0.
   b) La segunda opción es la que **cubre todas las alternativas posibles cuando la primera opción es falsa**. A esta afirmación **se la denomina hipótesis alternativa**, y se la simboliza con H_1.
2. Cualquier regla de decisión que, en función de los datos observados, permita optar entre H_0 y H_1 está sujeta a dos tipos de errores de tipo I y de tipo II.
3. Se **plantea un máximo nivel de tolerancia,** denotado por α, para la probabilidad de error de tipo I.
4. Se **formula una regla de decisión cuya probabilidad de error de tipo I no supere** α. A esta regla de decisión se la llama test.

Tipos de errores

Ninguna regla de decisión es infalible. Cualquier regla está sujeta a la posibilidad de cometer alguno de los siguientes errores:

1. **Error de tipo I:** Rechazar la hipótesis nula H_0, es decir, optar por la hipótesis alternativa H_1, cuando en realidad H_0 es verdadera.[10]
2. **Error de tipo II:** No rechazar la hipótesis nula H_0, es decir, optar por H_0, cuando en realidad H_1 es verdadera. [11]

[10] Cometeremos un error de tipo I si, al aplicar la regla de decisión, obtenemos un resultado que lleva a rechazar H_0 cuando H_0 es, de hecho, verdadera.
[11] Cometeremos un error de tipo II si, al aplicar la regla de decisión, no logramos rechazar H_0 cuando H_1 es, en realidad, verdadera.

Formalización general

En el problema de contraste o testeo de hipótesis, se asume que los datos $\underline{X} \sim f(\underline{x}) \in F = \{f(\underline{x};\theta) : \theta \in \Theta\}$ y el problema es averiguar si $\theta \in \Theta_0$ o $\theta \in \Theta_1$, siendo Θ_0 y Θ_1 una partición del conjunto Θ de posibles valores de θ, es decir:

$$\Theta = \Theta_0 \cup \Theta_1 \quad \text{y} \quad \Theta_0 \cap \Theta_1 = \emptyset$$

La afirmación $H_0 : \theta \in \Theta_0$ se llama hipótesis nula y la afirmación $H_1 : \theta \in \Theta_1$ se llama hipótesis alternativa.

Un test de hipótesis (no aleatorizado) para $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ es una función binaria $\phi(\underline{x})$ tal que:

- Se rechaza H_0 cuando $\phi(\underline{x}) = 1$
- No se rechaza H_0 cuando $\phi(\underline{x}) = 0$

Definición 94 (Región de rechazo)**.** El conjunto

$$\mathcal{R} = \{\underline{x} : \phi(\underline{x}) = 1\}$$

se denomina región de rechazo del test ϕ.

Hipótesis simples o compuestas

Definición 95 (Definición de Hipótesis Simple y Compuesta)**.** Se dice que una hipótesis H acerca de algún parámetro θ es **simple** cuando sólo un valor de θ verifica H. En caso contrario, se dice que H es **compuesta**.

Más generalmente:

- $H_0 : \theta \in \Theta_0$ se llama hipótesis nula simple si Θ_0 contiene un solo elemento; de lo contrario, se llama compuesta.
- $H_1 : \theta \in \Theta_1$ se llama hipótesis alternativa simple si Θ_1 contiene un solo elemento; de lo contrario, se llama compuesta.

Por ejemplo la hipotesis

$$H_0 : \mu = 2 \quad vs \quad H_1 : \mu \neq 2$$

Hipótesis unilaterales o bilaterales

Definición 96 (Hipótesis alternativa unilateral)**.** Se dice que la hipótesis alternativa H_1 acerca de algún parámetro escalar θ es **unilateral** cuando es de la forma

$$H_1 : \theta > \theta_0$$

o de la forma

$$H_1 : \theta < \theta_0$$

para algún θ_0 especificado.

Definición 97 (Hipótesis alternativa bilateral)**.** Se dice que la hipótesis alternativa H_1 acerca de algún parámetro escalar θ es **bilateral** cuando es de la forma

$$H_1 : \theta \neq \theta_0$$

para algún θ_0 especificado.

Función de potencia de un test

Definición 98 (Función de potencia)**.** La función de potencia de un test dado φ de hipótesis H_0 vs H_1 acerca de un parámetro dado θ es la función de θ

$$\pi_\varphi(\theta) \equiv P_\theta(\text{rechazar } H_0) \equiv P_\theta(\varphi(\underline{X}) = 1)$$

Aclaración notacional: cuando el contexto deja claro cuál es el test φ, obviaremos el subíndice φ y escribiremos $\pi(\theta)$ en vez de $\pi_\varphi(\theta)$.

Observe también que la notación β, que es ambigua, debería escribirse como $\beta(\mu)$. En cambio, en estas notas nosotros:

- Escribimos $P_\mu(\text{rechazar } H_0)$ en vez de $P(\text{rechazar } H_0 \mid \mu)$, pues el símbolo $\mid$ lo reservamos para referirnos a probabilidad condicional. Como estamos conduciendo inferencia frecuentista, no corresponde condicionar, ya que μ no es una variable aleatoria bajo el paradigma frecuentista.
- Usamos la letra griega $\pi = 1 - \beta$ para denominar a la función de potencia.
- Definimos la función de potencia sobre todo el conjunto de parámetros, no solo los que verifican la hipótesis alternativa H_1.

Notar que

- Si θ verifica H_0, entonces

$$\pi_\varphi(\theta) = \text{Probabilidad de error de tipo I del test } \varphi \text{ bajo } \theta$$

- Si θ verifica H_1, entonces

$$\pi_\varphi(\theta) = 1 - \text{Probabilidad de error de tipo II del test } \varphi \text{ bajo } \theta$$

Observación 99. Hemos definido la función de potencia de un test dado acerca de la media poblacional μ

$$\pi(\mu) = P_\mu(\text{rechazar } H_0)$$

A menudo se define la potencia como la probabilidad de rechazar una hipótesis nula que es falsa y se la escribe como

$$\text{Potencia}(\mu) = P_\mu(\text{rechazar } H_0 \mid \mu \in \Theta_1)$$

$$= 1 - \text{Probabilidad de error de tipo II}$$
$$= 1 - \beta$$

El símbolo "$\mid \mu \in \Theta_1$" y el subíndice μ significan "bajo la media μ para μ verificando H_1", de modo que se piensa en la potencia solo calculada en los μ que verifican H_1.

Observe que, aun definiendo la potencia de este modo, no hay un único valor de la potencia ya que $P_\mu(\text{rechazar } H_0 \mid \mu \in \Theta_1)$ cambia con μ.

Nivel de significación

El nivel de significación de un test es la peor (la más alta) de las probabilidades de error de tipo I que uno puede cometer cuando se usa ese test para contrastar una hipótesis H_0 con otra H_1.

Definición 100 (Nivel de significación)**.** El nivel de significación de un test es el máximo de las probabilidades de error de tipo I bajo todos los θ que verifican H_0:

$$\alpha = \text{nivel de significación} = \max_{\theta:\theta \text{ verifica } H_0} P_\theta(\text{el test } \varphi \text{ rechace } H_0)$$

equivalentemente,

$$\alpha = \text{nivel de significación} = \max_{\theta:\theta \text{ verifica } H_0} P_\theta(\varphi(\underline{X}) = 1)$$

Observación 101. A menudo se dice que el nivel de significación α es la probabilidad de rechazar una hipótesis nula que es verdadera.

Esta definición es precisa y coincide con la nuestra solo si existe un único valor de θ que verifique H_0, es decir, solo si H_0 es simple. En caso contrario, la definición es ambigua porque no existe una única probabilidad de error de tipo I.

Nuestra definición, en cambio, es precisa ya que α está unívocamente determinado como el máximo de las probabilidades de error de tipo I.

Contraste de hipótesis para media de una dist normal y varianza conocida

Formalización general

Supongamos que $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ donde σ^2 es conocida pero μ es desconocida. Las hipótesis a contrastar son:

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

El test propuesto es de la forma:

$$\text{rechazo } H_0 \text{ si y solo si } \overline{X}_n \geq k$$

Calcularemos k de modo que, para un nivel de significación α dado,

$$\alpha = \max_{\mu:\mu\leq\mu_0} P_\mu(\overline{X}_n \geq k)$$

(en nuestro ejemplo, $\alpha = 0{,}05$). Al valor k se lo llama valor crítico unilateral. Vamos a demostrar que el valor crítico que verifica el requisito es:

$$k = \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

donde, de ahora en más, para simplificar, usaremos la notación $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$.

Demostración de que $k = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$

Consideremos el test propuesto que rechaza H_0 si y solo si $\overline{X}_n \geq k$. La probabilidad de rechazo bajo H_0 es:

$$P_\mu(\overline{X}_n \geq k) = P_\mu\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{k-\mu}{\sigma/\sqrt{n}}\right)$$

donde $\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Esto nos lleva a:

$$P_\mu(\overline{X}_n \geq k) = 1 - \Phi\left(\frac{k-\mu}{\sigma/\sqrt{n}}\right)$$

Como $1 - \Phi\left(\frac{k-\mu}{\sigma/\sqrt{n}}\right)$ es una función creciente de μ, el k deseado debe verificar:

$$\alpha = \max_{\mu:\mu\leq\mu_0} P_\mu(\overline{X}_n \geq k) = \max_{\mu:\mu\leq\mu_0} 1 - \Phi\left(\frac{k-\mu}{\sigma/\sqrt{n}}\right)$$

$$\alpha = 1 - \Phi\left(\frac{k-\mu_0}{\sigma/\sqrt{n}}\right)$$

Equivalentemente,

$$1 - \alpha = \Phi\left(\frac{k-\mu_0}{\sigma/\sqrt{n}}\right)$$

Por lo tanto,

$$z_{1-\alpha} = \frac{k-\mu_0}{\sigma/\sqrt{n}}$$

Finalmente, despejando k, obtenemos:

$$k = \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

Test unilateral para la media de una población normal con varianza conocida

Hemos arribado a la siguiente conclusión: el test de nivel α de

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

rechaza H_0 cuando

$$\overline{X}_n \geq \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

o equivalentemente, cuando

$$\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$$

Al cociente

$$\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

se lo llama *estadístico Z* para el contraste de la media de una distribución normal.

Cálculo de la función de potencia

Para el test de

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

que rechaza H_0 cuando

$$\overline{X}_n \geq \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

su función de potencia es

$$\pi(\mu) = P_\mu \left(\overline{X}_n \geq \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

esto es,

$$\pi(\mu) = P_\mu \left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha} \right)$$

donde $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Por lo tanto,

$$\pi(\mu) = P \left(Z \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha} \right)$$

Finalmente, la función de potencia se expresa como:

$$\pi(\mu) = 1 - \Phi \left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha} \right)$$

La función de potencia

- aumenta a medida que μ aumenta
- aumenta a medida que σ disminuye (cuando $\mu > \mu_0$)
- aumenta a medida que n aumenta (cuando $\mu > \mu_0$)
- aumenta a medida que α aumenta

p-valor

Definición 102 (p-valor)**.** Para un test dado de una hipótesis nula H_0 dada, el p-valor es el mínimo nivel de significación tal que, con los datos obtenidos, H_0 es rechazada.

Notar que el p-valor depende de los datos obtenidos, luego antes de recoger los datos, el p-valor es una variable aleatoria.

Definición 103 (p-valor)**.** El p-valor para testear la hipótesis nula unilateral $H_0 : \mu \leq \mu_0$ cuando la población es normal y la varianza es conocida es igual a la máxima probabilidad de observar un test estadístico Z tan o más extremo que el valor observado del test estadístico Z cuando H_0 es cierta

Recordemos que el test de nivel α de $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ que hemos desarrollado rechaza H_0 cuando

$$\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$$

Ahora, como $z_{1-\alpha}$ aumenta a medida que α disminuye, entonces si la media muestral observada es $\overline{x}_n$, el p-valor debe satisfacer la igualdad

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} = z_{1-\text{p-valor}}$$

Luego, como

$$\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

cuando $\mu = \mu_0$, entonces

$$\begin{aligned}\text{p-valor} = P_{\mu_0}\left(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\right) &= \max_{\mu:\mu\in\Theta_0} P_\mu\left(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\right)\end{aligned}$$

Test uniformemente más potente

Definición 104 (Test más potente)**.** Un test de nivel α de $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ se dice más potente de nivel α para detectar a un $\theta_1 \in \Theta_1$ si

$$\pi_\phi(\theta_1) \geq \pi_{\phi^\top}(\theta_1)$$

para todo otro test $\phi^\top$ de nivel α.

Definición 105 (Test uniformemente más potente)**.** Un test de nivel α de $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ se dice uniformemente más potente de nivel α si verifica

$$\pi_{\phi}(\theta) \geq \pi_{\phi^\top}(\theta)$$

para todo $\theta \in \Theta_1$ y para todo otro test $\phi^\top$ de nivel α.

Teorema 106 (Neyman-Pearson)**.** *Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria, donde $n \in \mathbb{N}$, con función de densidad $f(x;\theta)$. Entonces, la verosimilitud de $X_1, X_2, \ldots, X_n$ es*

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i;\theta), \quad \textit{para } x' = (x_1, \ldots, x_n).$$

Sean θ' y θ'' distintos valores de θ, de manera que $\Omega = \{\theta', \theta''\}$, y sea $k > 0$. Definimos $\mathcal{C} \subseteq Sop(X)$ de manera que:

- *(a)* $\mathcal{C} = \left\{ x \in Sop(X) : \frac{L(\theta';x)}{L(\theta'';x)} \leq k \right\}$.
- *(b)* $\mathcal{C}^c = \left\{ x \in Sop(X) : \frac{L(\theta';x)}{L(\theta'';x)} \geq k \right\}$.
- *(c) Se define k de manera que se cumpla que* $\alpha = P_{H_0}(X \in \mathcal{C})$.

Entonces, $\mathcal{C}$ es la mejor región crítica de nivel α para testear la hipótesis simple $H_0 : \theta = \theta'$ vs la hipótesis alternativa simple $H_1 : \theta = \theta''$.

Test con máxima potencia uniforme (UMP Test)

Definición 107 (Test UMP de nivel α)**.** La región crítica $\mathcal{C}$ es una región crítica con máxima potencia uniforme (UMP) de nivel α para testear la hipótesis simple $H_0 : \theta = \theta'$ vs la hipótesis alternativa H_1 si el conjunto $\mathcal{C}$ es el mejor conjunto crítico de nivel α para testear H_0 contra cada una de las hipótesis simples bajo H_1. **Un test definido por una región crítica $\mathcal{C}$ se conoce como un test UMP de nivel α.**

Definición 108 (Mejor región crítica de nivel α)**.** Sea $C \subseteq \text{Sop}(\underline{X})$. Decimos que una región $\mathcal{C}$ es la mejor región crítica de nivel α para testear $H_0 : \theta = \theta'$ vs $H_1 : \theta = \theta''$ si:

- (a) $P_{\theta'}(\underline{X} \in \mathcal{C}) = \alpha$.
- (b) Para cualquier subconjunto $\mathcal{A} \subseteq \text{Sop}(X)$ se cumple que $P_{\theta'}(\underline{X} \in \mathcal{A}) = \alpha \implies P_{\theta''}(\underline{X} \in \mathcal{C}) \geq P_{\theta''}(\underline{X} \in \mathcal{A})$.

Definición 109 (Cociente de verosimilitud monótono)**.** Decimos que la verosimilitud $L(\theta, x)$ tiene la propiedad de cociente de verosimilitud monótono (MLR - monotone likelihood ratio) en el estadístico $y = u(\underline{x})$ si, para $\theta_1 < \theta_2$, el cociente

$$\frac{L(\theta_1, \underline{x})}{L(\theta_2, \underline{x})}$$

es una función monótona (puede ser creciente o decreciente) de $y = u(\underline{x})$

Proposición 110 (Test UMP con propiedad MLR)**.** *Supongamos que la verosimilitud $L(\theta, \underline{x})$ tiene la propiedad MLR decreciente en el estadístico $Y = u(\underline{X})$. Entonces, el cociente*

$$\frac{L(\theta_1, \underline{X})}{L(\theta_2, \underline{X})} = g(Y)$$

donde g es una función decreciente.
Sea α el nivel de significación. Entonces el test UMP de nivel α para las hipótesis $H_0 : \theta = \theta'$ versus $H_1 : \theta > \theta'$ es:

$$\textit{Rechazar } H_0 \textit{ si } Y \geq c_Y,$$

donde c_Y se determina por $\alpha = P_{\theta'}(Y \geq c_Y)$.

Test con Máxima Potencia Local si θ es Escalar

Definición 111 (Test con máxima potencia local)**.** Sea $F = \{f(\underline{x}; \theta) : \theta \in \Theta \subseteq \mathbb{R}\}$ y supongamos que Θ contiene un intervalo alrededor de θ_0. Supongamos que para cualquier test $\varphi(\underline{x})$ de

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0,$$

la función de potencia $\pi_\varphi(\theta)$ es dos veces diferenciable en $\theta = \theta_0$.
Diremos que **un test $\varphi^*(\underline{x})$ de H_0 vs H_1 es de máxima potencia local de nivel α si:**

- φ^* es de nivel α,
- $\frac{\partial}{\partial\theta}\pi_{\varphi^*}(\theta)\Big|_{\theta=\theta_0} = 0$
- $\frac{\partial^2}{\partial\theta^2}\pi_{\varphi}(\theta)\Big|_{\theta=\theta_0} \leq \frac{\partial^2}{\partial\theta^2}\pi_{\varphi^*}(\theta)\Big|_{\theta=\theta_0}$ para cualquier otro test $\varphi(\underline{x})$ de nivel α tal que $\frac{\partial}{\partial\theta}\pi_{\varphi}(\theta)\Big|_{\theta=\theta_0} = 0$.

La idea de esta definición es que la función de potencia $\pi_{\varphi^*}(\theta)$ del test φ^* se aleja mucho más rápido de $\pi_{\varphi^*}(\theta_0) = \alpha$ que la función de potencia $\pi_\varphi(\theta)$ de cualquier otro test φ de nivel α.

De modo que el test de máxima potencia local es el que mejor detecta aquellos θ que son más difíciles de detectar (los que son muy cercanos a θ_0).

Vamos a estudiar tres test con máxima potencia local:

1. Test *score*. [12]

[12] No es necesariamente un test asintótico

2. Test de Wald.

3. Test del cociente de verosimilitud.

Test score o multiplicador de Lagrange

Es posible probar que cuando $\underline{X} = (X_1, \ldots, X_n)$ con $X_i \overset{\text{iid}}{\sim}$, entonces para n grande la región de rechazo $\mathcal{C}$ de un test de máxima potencia local de nivel α de:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

es muy parecida a la región:

$$\left[\frac{1}{\sqrt{n}} s(\underline{X}; \theta_0)\right]^2 \geq k$$

donde k es una constante tal que:

$$P_{\theta=\theta_0}\left(\left[\frac{1}{\sqrt{n}} s(\underline{X}; \theta_0)\right]^2 \geq k\right) = \alpha$$

Teorema 112. *Supongamos que $X_i \overset{iid}{\sim} f(x;\theta)$ y supongamos que se cumplen todas las condiciones de regularidad de Máxima Verosimilitud. Sea $s(\underline{X};\theta_0)$ el score basado en toda la muestra $\underline{X}$ evaluado en $\theta = \theta_0$ y sea $I(\theta_0)$ la información basada en **una** unidad de la muestra. Entonces:*

$$\frac{\frac{1}{n} s(\underline{X};\theta_0)^2}{I(\theta_0)} \xrightarrow{D} \chi^2_{(1)}$$

La consecuencia del teorema es que el test rechaza cuando

$$\frac{\frac{1}{n} s(\underline{X};\theta_0)^2}{I(\theta_0)} > \chi^2_{(1),1-\alpha}$$

donde $\chi^2_{(1),1-\alpha}$ es el percentil $1-\alpha$ de la distribución chi-cuadrado con 1 grado de libertad.

Test score usando $i(\theta_0)$ en vez de $I(\theta_0)$: Si consideramos:

$$i(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \ln f(X_i;\theta)$$

Entonces vale que:

$$\frac{\frac{1}{n} s(\underline{X};\theta_0)^2}{i(\theta_0)} \xrightarrow{D} \chi^2_{(1)}$$

Este resultado implica que el test que rechaza cuando:

$$\frac{1}{n} s(X;\theta_0)^2 i(\theta_0) > \chi^2_{1,\alpha}$$

es un test con nivel aproximadamente igual a α cuando $n \gg 0$ para:

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0.$$

A la variable aleatoria:

$$S_2 = \frac{1}{n} s(X;\theta_0)^2 i(\theta_0),$$

también se la llama comúnmente el test estadístico score o test estadístico del multiplicador de Lagrange.

Test de Wald

Supongamos que $X_i \overset{\text{iid}}{\sim} f(x;\theta)$. Para contratastar:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

existen otros tests muy usados en la práctica, conocidos como test de Wald.

Un test de Wald es un test cuya región de rechazo de H_0 es de la forma:

$$W_j \geq k$$

para $j = 1, 2, 3$ ó 4, donde:

$$W_1 = I(\theta_0)\, n \left(\hat{\theta}_{MV} - \theta_0\right)^2 \qquad W_2 = I(\hat{\theta}_{MV})\, n \left(\hat{\theta}_{MV} - \theta_0\right)^2$$
$$W_3 = i(\theta_0)\, n \left(\hat{\theta}_{MV} - \theta_0\right)^2 \qquad W_4 = i(\hat{\theta}_{MV})\, n \left(\hat{\theta}_{MV} - \theta_0\right)^2$$

donde $\hat{\theta}_{MV}$ es el estimador de máxima verosimilitud de θ e:

$$I(\theta) = -E_\theta \left(\frac{\partial^2}{\partial \theta^2} \ln f(X_i;\theta) \right) \quad \text{e} \quad i(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(X_i;\theta).$$

Teorema 113. *Bajo condiciones de regularidad, vale que para $j = 1, 2, 3$ ó 4,*

$$P_{\theta=\theta_0} \left(W_j \geq \chi^2_{(1),\alpha} \right) \underset{n\to\infty}{\to} \alpha$$

De modo que el test rechaza cuando se verifica $W_j \geq k = \chi^2_{(1),\alpha}$.

Test del cociente de verosimilitud

Supongamos que $X_i \overset{\text{iid}}{\sim} f(x;\theta)$. Para contrastar:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

El test del cociente de verosimilitud tiene región de rechazo de la forma:

$$\lambda(\underline{x}) = \frac{f(\underline{X}; \hat{\theta}_{MV})}{f(\underline{X}; \theta_0)} > k$$

para alguna constante k donde $\hat{\theta}_{MV}$ es el estimador de máxima verosimilitud de θ.

Teorema 114. *Bajo condiciones de regularidad vale que:*

$$P_{\theta=\theta_0}\left(2\ln\left(\lambda(\underline{X})\right) \geq \chi^2_{(1),1-\alpha}\right) \underset{n\to\infty}{\to} \alpha$$

De modo que el test rechaza cuando $2\ln\left(\lambda(\underline{X})\right) \geq \chi^2_{(1),1-\alpha}$.

Potencia de la Trinidad de Tests

Sea W_1 el test de Wald de $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

Teorema 115. *Sea* $W_1 = n(\hat{\theta}_{MV} - \theta_0)^2 \cdot I(\theta_0)$ *y sea* $\theta_1 \neq \theta_0$. *Bajo las condiciones de regularidad:*

$$P_{\theta_1}(W_1 > k^2) \to 1 \quad \textit{cuando} \quad n \to \infty.$$

Potencia Local Asintótica

El Teorema anterior establece que, dado $\delta > 0$, existe un $n(\delta)$ tal que:

$$1 - P_{\theta_1}(W_1 > k) < \delta \quad \text{si} \quad n > n(\delta).$$

Sin embargo, $n(\delta)$ no solo depende de δ, sino también de θ_1. Para un n dado, no es posible garantizar que la potencia del test de Wald sea alta para detectar todos los $\theta \neq \theta_0$.

A medida que $n \to \infty$, los valores de θ en la alternativa que se encuentran a una distancia de θ_0 de orden:

- Igual que $\frac{1}{\sqrt{n}}$ se pueden detectar con probabilidad mayor que el nivel del test.
- Menor que $\frac{1}{\sqrt{n}}$ se pueden detectar con probabilidad igual al nivel del test.
- Mayor que $\frac{1}{\sqrt{n}}$ se pueden detectar con certeza.

A una sucesión de distribuciones $f(x;\theta_n)$ tal que $\sqrt{n}(\theta_n - \theta_0) \to c$ con $c \in (-\infty, \infty)$ se la llama **sucesión contigua** a la distribución $f(x;\theta_0)$.

Intervalos de confianza construidos por medio de inversión de tests

Supongamos que para cada $\theta_0 \in \Theta$, $\phi\left(X;\theta_0\right)$ es un test no aleatorizado de

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0 \text{ de nivel } \alpha$$

Notemos que incluyo θ_0 en la notación de la función $\phi\,(X;\theta_0)$ porque la regla de decisión depende de θ_0. Consideremos el siguiente subconjunto de Θ que es aleatorio

$$C(X) = \{\theta : \phi(X;\theta) = 0\}$$

Este subconjunto aleatorio satisface para todo $\theta_0 \in \Theta$

$$P_{\theta_0}\{\theta_0 \in C(X)\} = P_{\theta_0}[\phi(X;\theta_0) = 0]$$

$$= 1 - P_{\theta_0}[\phi(X;\theta_0) = 1] \geq 1 - \alpha$$

luego es una región de confianza de nivel $1 - \alpha$ para θ.

Se **conoce a** $C(X)$ **como región de confianza obtenida a partir de invertir el test** $\phi(X)$**.** En particular, si $C(X)$ es un intervalo, entonces se dice que es un intervalo de confianza obtenido a partir de la inversión del test.

Parametrización para el mismo modelo

Sean dos personas que cuentan con los mismos datos $\underline{X}$ y desean conducir un test con H_0 de que la distribución $f(\underline{x})$ que generó los datos es $f_0(\underline{x})$.

Supongamos que la primer persona plantea el modelo

$$\mathcal{F} = \{f(\underline{x};\theta) : \theta \in \Theta\}$$

La segunda persona plantea el mismo modelo con **otra parametrización**,

$$\mathcal{F} = \{\tilde{f}(\underline{x};\tau) : \tau \in \Lambda\}$$

Sean θ_0 y τ_0 los valores de θ y τ que corresponden a la densidad $f_0(\underline{x})$.

$$f(\underline{x}) = f_0(\underline{x}) \iff H_0 : \theta = \theta_0 \iff H_0 : \tau = \tau_0$$

Claramente si ambas personas observan los mismos datos lo lógico sería que si aplican el mismo "test", la primera para testear $H_0\,y$ la segunda para testear H_0, ambas arriben a la misma conclusión.

Lamentablemente s**i ambas personas aplican el test de Wald** (digamos basado en W_1) **no necesariamente arribarán a la misma conclusión,** porque el valor del estadístico W_1 depende de la parametrización del modelo (lo mismo ocurre con W_2, W_3 y W_4).

Apéndice I: Distribución de variables aleatorias discretas

Distribución Bernoulli

Una variable aleatoria con distribución Bernoulli es una variable aleatoria binaria que toma dos valores posibles, 0 y 1, con probabilidades $1-p$ y p respectivamente. Su función de probabilidad puntual es:

$$p_X(1) = p, \quad p_X(0) = 1-p, \quad pX(x) = 0\,(si\,x \neq 0\,y\,x \neq 1)$$

La función de probabilidad puntual se puede reescribir como:

$$P(X = x) = p_X(x) = \begin{cases} p^x(1-p)^{1-x} & \text{si } x = 1 \quad \text{ó} \quad x = 0 \\ 0 & \text{en otro caso} \end{cases}$$

Se denota a su distribución como $Be(p)$.

Las variables de Bernoulli se pueden escribir como variables indicadoras: Si A es un evento, la variable indicadora I_A toma el valor 1 si A ocurre y 0 si A no ocurre. De este modo, $I_A(\omega) = 1$ si $\omega \in A$ y $I_A(\omega) = 0$ si $\omega \notin A$. Es decir que I_A es una variable Bernoulli con parámetro $p = p(A)$.

Ejemplo: De un envío grande de mercadería, elegimos un lote al azar, y nos interesa averiguar si en el lote hay alguna mercadería fallada. Entonces $X = 1$ si el lote tiene una mercadería fallada y $X = 0$ en caso contrario.

Notemos que la varianza de una variable $X \sim Be(p)$ depende de p. Notemos que $\text{Var}(X) = p(1-p) \leq 0{,}25$ para cualquier $p \in [0, 1]$.

Distribución Binomial

Una variable aleatoria se llama binomial de parámetros n y p si es igual al número total de éxitos en n ensayos de Bernoulli independientes. Si X denota una variable aleatoria con distribución $Bin(n, p)$, su función de probabilidad de masa es:

$$p_X(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & \text{si } x = 0, 1, 2, ..., n. \\ 0 & \text{en otro caso} \end{cases}$$

Donde $\binom{n}{x}$ es el coeficiente binomial, que se define como:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Esta fórmula representa la probabilidad de obtener x éxitos en n ensayos, donde cada ensayo tiene probabilidad de éxito p y probabilidad de fracaso $1-p$.

La distribución binomial se denota como $Bin(n, p)$.

Distribución Geométrica

Una variable aleatoria geométrica $Ge(p)$ cuenta la cantidad total de ensayos de Bernoulli, con probabilidad de éxito p, que se realizan para obtener el primer éxito. Su función de probabilidad de masa es:

$$P(X = x) = p_X(x) = \begin{cases} p(1-p)^{x-1} & \text{si } x \in \{1, 2, 3, \ldots\} \\ 0 & \text{en otro caso} \end{cases}$$

Esta función representa la probabilidad de que se necesiten x ensayos para obtener el primer éxito en una secuencia de ensayos independientes, donde cada ensayo tiene probabilidad de éxito p y probabilidad de fracaso $1-p$.

La distribución geométrica se denota como $Ge(p)$.

Distribución Binomial Negativa

Una variable aleatoria binomial negativa $BN(p,r)$ cuenta la cantidad total de ensayos de Bernoulli, con probabilidad de éxito p, necesarios para obtener r éxitos. Su función de probabilidad de masa es:

$$p_X(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r}, & \text{si } x \in \{r, r+1, r+2, \ldots\} \\ 0, & \text{en otro caso} \end{cases}$$

Esta función representa la probabilidad de que se necesiten x ensayos para obtener r éxitos en una secuencia de ensayos independientes, donde cada ensayo tiene probabilidad de éxito p y probabilidad de fracaso $1-p$.

Es importante notar cómo se define una variable con distribución negativa ya que difiere según la bibliografía que se consulte. Es una generalización de la variable geométrica, ya que $BN(p,1) = Ge(p)$.

Notemos que podríamos escribir que si $X \sim BN(p,r)$ y $W_i \sim Ge(p)$ independientes, con $i = 1, 2, \ldots, r$, entonces:

$$X = W_1 + W_2 + \ldots + W_r$$

Distribución Hipergeométrica

Una variable aleatoria hipergeométrica $H(n,r,m)$ cuenta la cantidad de bolillas azules x que se obtienen de una urna cuando se extraen m bolillas al azar sin reemplazo. En la urna hay un total de n bolillas, r de las cuales son de color azul y $n-r$ son de color rojo. Su función de probabilidad de masa es:

$$P_X(x) = \begin{cases} \frac{\binom{r}{x}\binom{n-r}{m-x}}{\binom{n}{m}} & \text{si } x \in \{0, 1, 2, \ldots, r\} \\ 0 & \text{en otro caso} \end{cases}$$

Esta función representa la probabilidad de obtener exactamente x bolillas azules en una muestra de m bolillas extraídas de la urna sin reemplazo.

Donde $\binom{a}{b}$ representa el coeficiente binomial, que se define como:

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

Aproximación de la distribución binomial a la distribución hipergeométrica: Si la cantidad total de bolillas n cumple que $r \gg m$, se puede aproximar la probabilidad de que $X \sim H(n,r,m)$ tome el valor x como:

$$p_X(x) \approx \binom{m}{x} \left(\frac{r}{n}\right)^x \left(1 - \frac{r}{n}\right)^{m-x}$$

Esta aproximación se basa en la idea de que la probabilidad de extraer una bolilla azul es constante en cada ensayo y que las extracciones son independientes.

Distribución de Poisson

Una variable aleatoria se llama Poisson de parámetro λ, $Pois(\lambda)$, si puede tomar cualquier valor en el conjunto $\{0, 1, 2, 3, \ldots\}$ y su función de masa de probabilidad es:

$$p_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{si } x \in \{0, 1, 2, 3, \ldots\} \\ 0 & \text{en otro caso} \end{cases}$$

Donde λ es un número real positivo.

Las variables de Poisson modelan el número de veces que ocurre algún evento en una franja de tiempo (o espacio) bajo algunos supuestos. Por ejemplo, puede representar el número de errores de tipeo en una página, el número de llamados que llegan a un call-center, o el número de autos que pasan por una calle en un minuto.

Aproximación de Poisson por una binomial: Si $n \gg 0$ y $p := \frac{\lambda}{n}$ se verifica que

$$p_X(x) = e^{-\lambda}\frac{\lambda^x}{x!} \approx \binom{n}{x} p^x (1-p)^{n-x}$$

O sea, si $n \geq 100$ y $p \leq \frac{1}{100}$, $\lambda = np$ vale que $\text{Pois}(\lambda) \approx \text{Bin}(n, \lambda/n)$

Resumen de distribuciones discretas

Resumen de momentos de variables aleatorias

Distribución	Esperanza	Varianza
Be(p)	$E(X) = p$	$\text{Var}(X) = p(1-p)$
Bin(n, p)	$E(X) = np$	$\text{Var}(X) = np(1-p)$
Ge(p)	$E(X) = \frac{1}{p}$	$\text{Var}(X) = \frac{1-p}{p^2}$
BN(n, r)	$E(X) = \frac{r}{p}$	$\text{Var}(X) = \frac{r(1-p)}{p^2}$
H(n, r, m)	$E(X) = \frac{mr}{n}$	$\text{Var}(X) = \frac{(m-1)(r-1)nr+1-mn}{n^2(n-1)}$
Poi(λ)	$E(X) = \lambda$	$\text{Var}(X) = \lambda$

Resumen de funciones de probabilidad puntual (PMF) y FGM

Distribución	F. de masa de prob.	F. generatriz de momentos
Be(p)	$P(x) = p(1-p)^{1-x}$	$M_X(t) = pe^t + 1 - p$
Bin(n, p)	$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$	$M_X(t) = (pe^t + 1 - p)^n$
Ge(p)	$P(x) = p(1-p)^{x-1}$	$M_X(t) = \frac{pe^t}{1-(1-p)e^t}$
BN(p, r)	$P(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$	$M_X(t) = (\frac{pe^t}{1-(1-p)e^t})^r$
H(n, r, m)	$P(x) = \frac{\binom{r}{x}\binom{n-r}{m-x}}{\binom{n}{m}}$	$M_X(t) = \sum_{x=0}^{r} \frac{\binom{r}{x}\binom{n-r}{m-x}}{\binom{n}{m}} e^{tx}$
Poi(λ)	$P(x) = e^{-\lambda}\frac{\lambda^x}{x!}$	$M_X(t) = e^{\lambda(e^t-1)}$

Apéndice II: Distribución de variables aleatorias continuas

Distribución Uniforme

Una variable aleatoria continua X tiene distribución uniforme sobre un intervalo $E = [a, b]$, $X \sim \text{Unif}[a, b]$, cuando su función de densidad de probabilidad es:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{de otro modo} \end{cases}$$

La función de distribución acumulada es:

$$F_X(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b \end{cases}$$

Si $a = 0$ y $b = 1$, $X \sim U[0,1]$, es un caso particular de la distribución Beta cuando $\alpha = 1$ y $\beta = 1$. Es decir $U[0,1] \equiv Beta(1,1)$.

La distribución uniforme está caracterizada por dos parámetros a y b. Notemos que se puede escribir la PDF como $f(x) = \frac{1}{b-a} I_{[a,b]}(x)$

Es importante destacar que si $X \sim U(a,b)$ o $X \sim U(a,b]$ o $X \sim U[a,b)$, solo afecta al soporte de X, pero no a la expresión de la función de densidad.

Distribución exponencial

Una variable aleatoria continua X tiene distribución exponencial en el intervalo $(0,\infty)$, $X \sim \text{Exp}(\lambda)$, cuando su función de densidad de probabilidad es:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

Donde λ es el parámetro de escala de la distribución, que representa la tasa promedio de ocurrencia de eventos. A medida que λ es mayor, la función de densidad cae más abruptamente.

La función de distribución acumulada es:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

En la práctica, esta distribución se utiliza en modelos relacionados a tiempos de espera o vida útil bajo ciertos supuestos. La aplicación clásica de esta distribución es la de modelar el tiempo de vida de algún componente electrónico.

En economía, una aplicación podemos, bajo ciertos supuestos, modelizar el tiempo de duración del desempleo como una variable exponencial.

Una característica particular de la distribución exponencial es que representa procesos que no tienen memoria. Eso quiere decir que:

$$P(T \geq t + s | T \geq s) = P(T \geq t) = e^{-\lambda t}$$

Relación entre exponencial y poisson

La relación entre la variable aleatoria Poisson y la exponencial se puede ver en el siguiente ejemplo: imagina que N mide la cantidad de personas que llegan (de manera independiente) a una parada de colectivo a una tasa de λ personas por hora. La cantidad promedio de personas que llegan en t horas es $\lambda \cdot t$. La variable aleatoria T que mide el tiempo entre la llegada de dos personas consecutivas es una variable aleatoria que depende de $N \sim Poi(\lambda)$ la variable aleatoria Poisson. Entonces $T \sim Exp(\lambda)$ La función de densidad de probabilidad de T es $\lambda e^{-\lambda t}$.

Distribución Normal

Una variable aleatoria continua X tiene distribución normal o Gaussiana en el intervalo $(-\infty, +\infty)$, $X \sim N(\mu, \sigma^2)$, cuando su función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

La función de distribución acumulada es:

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(u-\mu)^2} du$$

Las características de $f(x)$ son:

- Toma valores positivos en toda la recta real (pero asigna valores muy pequeños a valores de x que difieren de μ en más de $\pm 3\sigma$).
- Tiene un solo máximo global en $x = \mu$ y es simétrica alrededor de μ.
- El valor de σ controla la dispersión de la curva alrededor de $x = \mu$. Tiene puntos de inflexión (cambio de concavidad) en $x = \mu - \sigma$ y $x = \mu + \sigma$.
- Cuando $\mu = 0$ y $\sigma = 1$, X tiene una **distribución normal estándar** $N(0,1)$.

Para una variable aleatoria con distribución $N(\mu, \sigma^2)$, la integral

$$\int_a^b f_X(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

no se puede calcular "a mano". Se calcula numéricamente.

Distribución Gamma

Una variable aleatoria continua X tiene distribución Gamma en el intervalo $(0, +\infty)$, $X \sim \Gamma(\alpha, \lambda)$, cuando su función de densidad de probabilidad es:

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

donde $\Gamma(\alpha)$ es la función Gamma que está definida como:

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

En particular, $\Gamma(n) = (n-1)!$ si n es un número entero positivo.

Una propiedad importante de la función Gamma es que:

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

La distribución Gamma generaliza a la distribución exponencial, con $Exp(\lambda) \equiv \Gamma(1, \lambda)$. También, la distribución chi-cuadrado con $\chi^2(\nu)$ es un caso particular con $\Gamma(\nu/2, 1/2)$.

Distribución Beta

Una variable aleatoria continua X tiene distribución Beta en el intervalo $[0, 1]$, $X \sim \text{Beta}(\alpha, \beta)$, con parámetros $\alpha > 0$ y $\beta > 0$, cuando su función de densidad de probabilidad es:

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & \text{si } 0 \leq x \leq 1 \\ 0 & \text{de otro modo} \end{cases}$$

donde $B(\alpha, \beta)$ es la función Beta, definida como:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

y $\Gamma(\cdot)$ es la función Gamma.

La distribución Beta es flexible para modelar variables aleatorias en el intervalo [0,1] y tiene varias aplicaciones, como en pruebas Bayesianas o modelado de proporciones.

Resumen de distribuciones continuas

Resumen de momentos de variables aleatorias continuas

Distribución	Esperanza	Varianza
Unif(a, b)	$E(X) = \frac{a+b}{2}$	$\text{Var}(X) = \frac{(b-a)^2}{12}$
Exp(λ)	$E(X) = \frac{1}{\lambda}$	$\text{Var}(X) = \frac{1}{\lambda^2}$
N(μ, σ^2)	$E(X) = \mu$	$\text{Var}(X) = \sigma^2$
Gamma(α, λ)	$E(X) = \frac{\alpha}{\lambda}$	$\text{Var}(X) = \frac{\alpha}{\lambda^2}$
Beta(α, β)	$E(X) = \frac{\alpha}{\alpha+\beta}$	$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Resumen de funciones de densidad de probabilidad (PDF) y FGM

Distribución	F. de densidad de prob. (PDF)	F. generatriz de momentos (MGF)
Unif(a,b)	$f(x)=\frac{1}{b-a}$	$M_X(t)=\frac{e^{tb}-e^{ta}}{t(b-a)}$
Exp(λ)	$f(x)=\lambda e^{-\lambda x}$	$M_X(t)=\frac{\lambda}{\lambda-t}, \quad t<\lambda$
N(μ,σ^2)	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$M_X(t)=e^{\mu t+\frac{\sigma^2 t^2}{2}}$
Gamma(α,λ)	$f(x)=\frac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}$	$M_X(t)=\left(\frac{\lambda}{\lambda-t}\right)^\alpha, \quad t<\lambda$
Beta(α,β)	$f(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	No tiene FGM cerrada

Distribución chi-cuadrado con n grados de libertad

La distribución chi-cuadrado con n grados de libertad es la suma de los cuadrados de n variables aleatorias normal estándar $N(0,1)$ independientes.

Sean $Z_1,\ldots,Z_n$ variables aleatorias independientes tales que $Z_i \sim N(0,1)$ para $i=1,2,\ldots,n$ entonces la variable aleatoria X definida por

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

tiene una distribución chi cuadrada con n grados de libertad y se denota χ^2_n.

Distribución t-Student

Para $X_1,\ldots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$:

(a) $\overline{X}_n$ y S_n^2 son variables aleatorias independientes por el teorema de Basu.

(b) $\overline{X}_n \sim N\left(\mu,\frac{\sigma^2}{n}\right)$ y $(n-1)\cdot\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$, donde χ^2_{n-1} es la distribución chi-cuadrado con $n-1$ grados de libertad.

Si $Z=\sqrt{n}\frac{X_n-\mu}{\sigma} \sim N(0,1)$ y $V=(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$ son variables aleatorias independientes, entonces la variable W:

$$W=\frac{Z}{\sqrt{\frac{V}{n-1}}}=\frac{\sqrt{n}(\overline{X}_n-\mu)}{S}\sim t_{n-1}$$

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