MA34B – Estadística
Estimación por Intervalo

Prof. Rodrigo Abt B.
rabt@dim.uchile.cl

Introducción (1)

- En un problema de estimación puntual, el objetivo es determinar una fórmula o receta que permita dar un valor para el(los) parámetro(s) desconocidos de una distribución dada.
- Pero, una vez hecha una estimación, no sabemos si en realidad estamos cerca o no del parámetro estimado.
- Sólo tenemos una idea de la precisión de un estimador a través de su varianza.

Introducción (2)

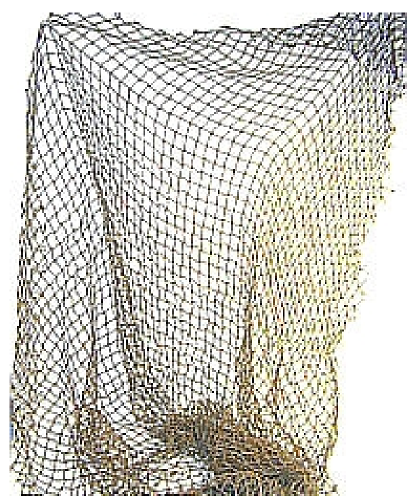
- En la práctica, la estimación puntual es rara, ya que si bien un estimador es bueno, la muestra obtenida para la estimación puede hacernos dudar de qué tan lejos o cerca está la estimación del verdadero valor de un parámetro.
- Por lo mismo, es más frecuente buscar un *intervalo* de valores entre los cuales pudiese estar el parámetro buscado con algún grado de confiabilidad.
- Por ejemplo, encontrar un intervalo de valores para la media μ de una Normal con un alto nivel de confiabilidad

Introducción (3)

- Para hacerse una idea, hacer una estimación puntual es como pescar con un anzuelo:

$$\longrightarrow \theta$$

- Mientras que hacer una estimación por intervalo es como pescar con una red.



$$\longrightarrow \theta$$

Introducción(4)

- Ahora bien, ¿de qué tamaño debería ser esta red para lograr "atrapar" un buen valor?, y ¿cómo nos aseguramos de que estamos "atrapando" lo correcto?
- Supongamos que estimamos la media μ de una Normal. Sabemos que μ puede tomar cualquier valor dentro de los reales, por ende, el único intervalo que nos garantiza un 100% de confiabilidad para encontrar al parámetro en él es…de –infinito a +infinito
- Sin embargo, dicho intervalo es de muy poca utilidad. De hecho no nos dice nada informativo.
- Sin embargo, si tuviésemos un intervalo como [2,3] sería BASTANTE más informativo.
- Desgraciadamente, para tener un intervalo debemos tolerar un margen de error, y "descontarlo" de la confiabilidad esperada. O sea, ¿podríamos encontrar valores "a" y "b" tales que la confiabilidad de encontrar μ en [a,b] sea de un 95% por ejemplo?

Planteamiento

- Formalmente, buscamos “a” y “b” tales que:

$$P(a \leq \mu \leq b) = 1 - \alpha$$

- En este caso el intervalo [a,b] se denomina “intervalo de confianza”, donde 1-α es el “nivel de confianza” o “confianza”.
- Ahora bien, dado que no conocemos la distribución asociada a μ, debemos utilizar los valores muestrales para determinar los valores de “a” y “b”, es decir, buscamos $a=a(X_1,X_2,\ldots,X_n)$ y $b=b(X_1,X_2,\ldots,X_n)$.
- Por ende “a” y “b” son variables aleatorias.

Resolución

- Para resolver la ecuación anterior, debemos encontrar una variable aleatoria que dependa de μ, pero cuya distibución NO dependa de μ. Es decir, debemos transformar la ecuación original.

$$P(a \leq \mu \leq b) = 1 - \alpha$$

$$\downarrow$$

$$P(L \leq Z \leq U) = 1 - \alpha$$

- En que Z es una v.a. con distribución conocida.

Caso 1: Media μ para Normal con varianza conocida

- Retomemos el caso de la media
  - Buscamos "a" y "b" tales que $P(a<\mu<b)=1-\alpha$
  - Sabemos por otro lado que:

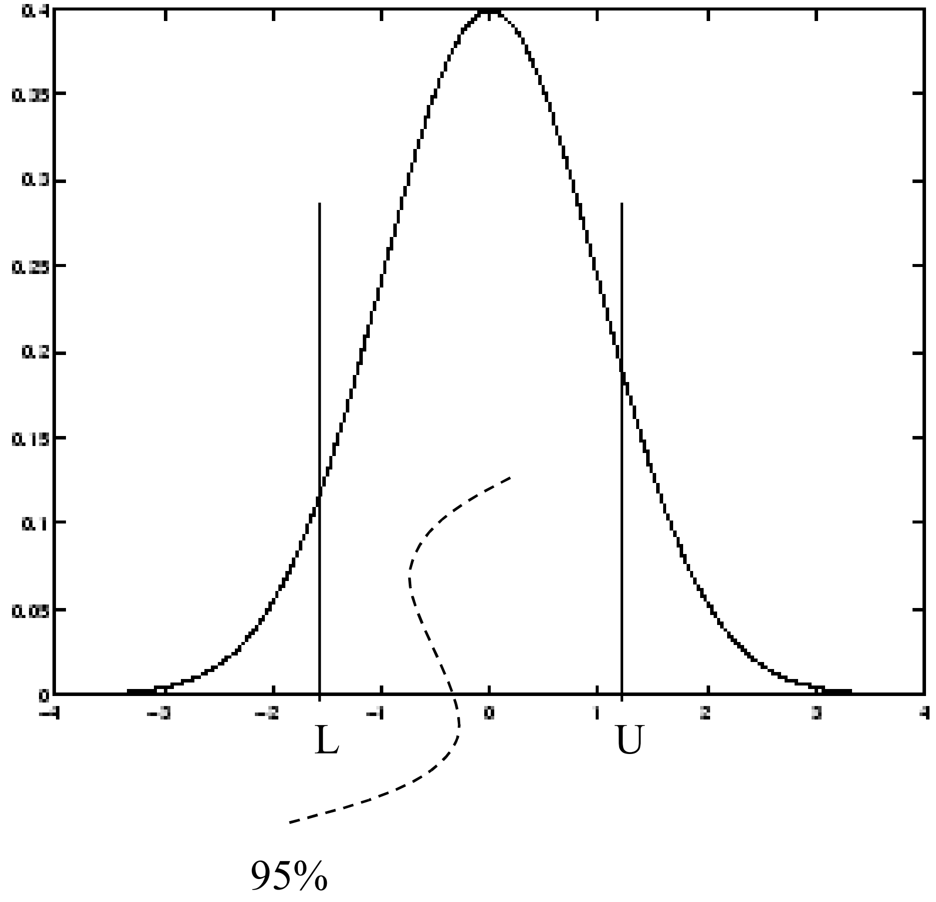
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

  - Si hacemos las transformaciones algebraicas correspondientes, tenemos:

$$P\left(\underbrace{\frac{\bar{X}-b}{\sigma/\sqrt{n}}}_{L} \leq \underbrace{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}_{Z} \leq \underbrace{\frac{\bar{X}-a}{\sigma/\sqrt{n}}}_{U}\right) = 1-\alpha$$

Análisis Gráfico

- En términos gráficos, si fijamos un nivel de confianza del 95% buscamos L y U tales que:



Condiciones de Borde (1)

- Si nos detenemos un poco en el análisis gráfico, veremos que existen infinitos valores de L y U que cumplen la condición. En efecto, fijando uno podemos obtener el otro.
- ¿Cómo elegimos entre uno u otro intervalo?
- Criterio: Largo del intervalo
- Claramente elegiremos aquellos intervalos que presenten un largo menor, ya que significa que son más precisos.

Condiciones de Borde (2)

- Idealmente preferiremos aquellos de largo mínimo.
- En el caso de las distribuciones simétricas, el problema está resuelto: Basta tomar como centro del intervalo el punto de simetría y tendremos un intervalo de largo mínimo.
- Luego, debemos encontrar L tal que $P(-U<Z<U) = 1-\alpha$, que es equivalente a buscar $P(Z>U) = 0{,}025 \Rightarrow U = 1{,}96$

Solución

- Luego, despejando tenemos que L = 1,96, pero:

$$L = \frac{\bar{X} - a}{\sigma/\sqrt{n}} \Rightarrow a = \bar{X} - 1,96\frac{\sigma}{\sqrt{n}}$$

- Y por otro lado:

$$b = \bar{X} + 1,96\frac{\sigma}{\sqrt{n}}$$

- Finalmente el intervalo buscado es:

$$[\bar{X} - 1,96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1,96\frac{\sigma}{\sqrt{n}}]$$

Observaciones del Resultado

- Notemos que el intervalo está centrado en la media muestral.
- Además crece con la varianza, y disminuye con el tamaño muestral, lo que es intuitivamente deseable.
- Si queremos aumentar la confianza, debemos estar dispuestos a que crezca el tamaño del intervalo.

Caso 2: Media μ para Normal con varianza desconocida

- Cuando la varianza es desconocida, se tiene que:

$$\frac{\bar{X} - \mu}{S_n/\sqrt{n-1}} \sim t_{n-1}$$

- De donde análogamente:

$$IC_\mu = \left[\bar{X} - t_{n-1}^{\alpha/2} \cdot \frac{S_n}{\sqrt{n-1}}, \bar{X} + t_{n-1}^{\alpha/2} \cdot \frac{S_n}{\sqrt{n-1}}\right]$$

Caso 3: Intervalo para una varianza

- Queremos resolver:

$$Pr(a \leq \sigma^2 \leq b) = 1 - \alpha$$

- Un estadístico que depende de σ^2 y cuya distribución es conocida es:

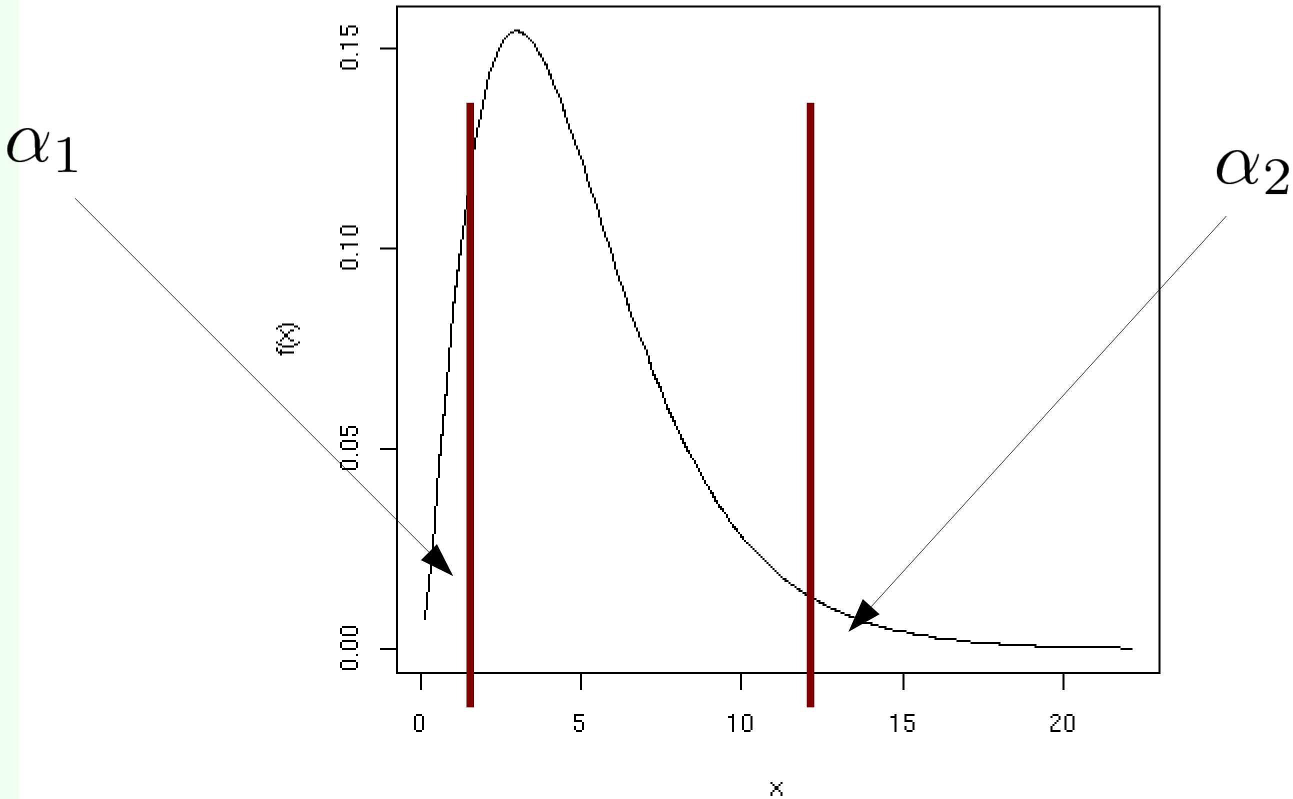
$$\frac{nS_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Por ende buscamos K_1 y K_2 tales que:

$$Pr(K_1 \leq \frac{nS_n^2}{\sigma^2} \leq K_2) = 1 - \alpha$$

Análisis Gráfico

- Si graficamos la distribución, nos encontramos que NO es simétrica:



Solución

- Debemos encontrar los valores de α_1 y α_2 tales que el largo del intervalo sea mínimo.
- Sin embargo, en la práctica se asigna arbitrariamente $\alpha/2$ a cada uno y se procede resolviendo cada cola de manera separada. De este modo se tiene que:

$$IC_{\sigma^2} = \left[\frac{nS_n^2}{K_{n-1}^{\alpha/2}}, \frac{nS_n^2}{K_{n-1}^{1-\alpha/2}} \right]$$

Caso 4: Intervalo para una proporción

- Queremos resolver:

$$Pr(a \leq p \leq b) = 1 - \alpha$$

- Del Teorema Central del Límite sabemos que:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

- Donde $\hat{p}$ es la proporción de éxitos en la muestra.
- El problema en este caso es que desconocemos “p”, de hecho, es precisamente lo que queremos estimar. Podemos desarrollar la expresión de la probabilidad, reemplazando Z.

Ecuación característica

- Se tiene que para tener un intervalo de largo mínimo se requiere:

$$Pr\left\{-u \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq u\right\} = 1-\alpha$$

$$Pr\left\{\left|\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq u\right\} = 1-\alpha$$

- Lo que da origen a una ecuación de 2º grado:

$$(n+u^2)p^2 - (u^2+2n\hat{p})p + n\hat{p} = 0$$

Solución

- Las soluciones de esta ecuación son:

$$p = \frac{u^2 + 2n\hat{p} \pm \sqrt{(u^2 + 2n\hat{p})^2 - 4(n + u^2)n\hat{p}}}{2(n + u^2)}$$

- Lo que se puede reescribir como:

$$p = \frac{n}{n + u^2}\left(\hat{p} + \frac{u^2}{2n}\right) \pm u\sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{u^2}{4n^2}}$$

- Cuando "n" es suficientemente grande:

$$IC_p = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right]$$

Funciones Pivotes

- Un estadístico T se dirá “pivote” para θ si $T=T(X_1,X_2,\ldots,X_n,\theta)$ y la f.d.p. de T NO depende de θ.
- Ejemplos de funciones pivote para diferentes parámetros y distribuciones:
  - Si $X \sim U[0,\theta]$, entonces $T = \mathrm{Min}\{X_i\}/\theta$
  - Si $X \sim \exp(\theta)$, entonces $T = 2\theta\Sigma X_i$
  - Si $f(x/\theta) = e^{-(x-\theta)}$ con $x \geq \theta$, entonces $T = \mathrm{Min}\{X_i\}-\theta$

Tamaños de Muestra

- Una de las aplicaciones más frecuentes para los intervalos de confianza es la determinación del tamaño de muestra para medias o proporciones.
- Por ejemplo, para el caso de las encuestas de opinión un número típico de tamaño muestral conservador inicial es 1100, con un nivel de confianza de 95% y un error del 3%.
- Este resultado se deriva del hecho de que la varianza máxima de una variable Bernoulli se alcanza cuando $p=0.5$

Notas

- Los intervalos de confianza para el caso Bayesiano se pueden derivar directamente de la distribución a posteriori.
- La interpretación correcta es que el intervalo de confianza "cubre" al parámetro el $(1-\alpha)100\%$ de las veces.
- Las distribuciones muestrales como la Normal, t-Student o Chi cuadrado se utilizan como distribuciones límite cuando el tamaño muestral es grande.
- Lo más complicado suele ser encontrar el estadístico que dependa del parámetro a estimar, pero cuya distribución NO dependa del mismo, o sea las funciones "pivotes".