XS3310 Teoría Estadística

I Semestre 2021

Escuela de Estadística

2021-05-13

class: center, middle

¿Qué hemos visto hasta ahora?

Todo sobre estimadores puntuales + intro a los pivotes. Algunos intervalos de confianza clásicos.

¿Qué vamos a discutir hoy?

Otros intervalos de confianza clásicos. Intervalos de confianza para muestras grandes. Funciones estabilizadores de varianza.

Técnica del pivote

La semana pasada:

Esta técnica consiste en definir una variable aleatoria, U, que llamaremos **pivote**. El pivote debe cumplir con las siguientes condiciones:

- Para una muestra aleatoria $X_1, X_2, ..., X_n$ y un parámetro desconocido θ, U debe estar definido en términos de los elementos de la muestra aleatoria y del parámetro desconocido, donde el parámetro desconocido θ sea la única cantidad desconocida.
- La distribución de probabilidad de U debe ser conocido y no depender de θ.
- Se recomienda iniciar con un estadístico suficiente
- Se puede usar como guía las distribuciones con parámetros de posición - escala.

Ejemplo de un intervalo de confianza para la varianza desconocida.

Se toma el tiempo en milisegundos para que los nervios se recuperen después de administrar una droga a 4 ratones.

$$1.7, 1.6, 1.8, 1.9$$

Calcule un intervalo de confianza para el promedio este tiempo a un 90% de confianza.

Solución:

Para este caso tenemos que $\bar{X} = 1.75$ y $S^2 = 0.017$ y $S = \sqrt{0.017} = 0.13$.

En este caso vamos a tomar una $t_{3,0.05} = 2.353$ $(\alpha/2 = 0.1/2 = 0.05)$

$$\begin{aligned}
&P\left(\bar{X} - t_{3,0.05}\frac{s}{\sqrt{4}} \leq \mu \leq \bar{X} + t_{3,0.05}\frac{s}{\sqrt{n}}\right)\\
&= P\left(1.75 - 2.353\frac{0.13}{2} \leq \mu \leq 1.75 + 2.353\frac{0.13}{2}\right)\\
&= P\left(1.597 \leq \mu \leq 1.903\right) = 0.9.
\end{aligned}$$

Ejemplos de un intervalo de confianza para la diferencia de medias

En un hospital se realiza un estudio sobre la influencia del estrés en el peso de los bebés al nacer. Se consideran dos grupos de mujeres embarazadas: aquellas que viven en el campo y aquellas que viven en la ciudad, y se obtienen los siguientes datos sobre el peso de sus hijos.

	Muestra	Peso medio de los bebés	Desviación estándar
campo	$n_1 = 320$	$\overline{X}_1 = 3,8$	$\sigma_1 = 0,6$
ciudad	$n_2 = 240$	$\overline{X}_2 = 3,4$	$\sigma_2 = 0,8$

Asumiendo que los datos normales, determine cómo influye que la madre viva en el campo o en la ciudad en el peso de su hijo al nacer, utilizando un intervalo de confianza para la diferencia de medias con un nivel de confianza del 95%.

Solución: Intervalo de confianza del 95%:

$1 - \alpha = 0.95 \Rightarrow \alpha = 1 - 0,95 = 0,05$

Buscamos en el la tabla de normales aquel valor que cumpla $P(Z \leq z_{\alpha/2}) = 0.05/2 = 0.025$

Obtenemos que $\Rightarrow z_{\alpha/2} = z_{0,025} = 1,96$

$$
\begin{aligned}
&\left((\overline{\mathrm{X}}_1-\overline{\mathrm{X}}_2)-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{\mathrm{n}_1}+\frac{\sigma_2^2}{\mathrm{n}_2}},(\overline{\mathrm{X}}_1-\overline{\mathrm{X}}_2)+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{\mathrm{n}_1}+\frac{\sigma_2^2}{\mathrm{n}_2}}\right)\\
&=\left((3,8-3,4)-1,96\sqrt{\frac{0,6^2}{320}+\frac{0,8^2}{240}},(3,8-3,4)+1,96\sqrt{\frac{0,6^2}{320}+\frac{0,8^2}{240}}\right)\\
&=[0,279,0,520]
\end{aligned}
$$

Intervalos de confianza para poblaciones normales

Intervalo de confianza para σ^2

Nos interesa constuir un intervalo de confianza de probabilidad $1-\alpha$ para la variancia poblacional, σ^2. En este caso nuestra población sigue siendo Normal con parámetros μ y σ^2, donde ambos parámetros son desconocidos. Si queremos obtener una estimación por intervalo para σ^2 debemos primero obtener una variable aleatoria que funcione como pivote. En este caso esta variable aleatoria debe estar en términos de $X_1, X_2, ..., X_n$ y de σ^2, y su distribución no puede depender de σ^2. Con lo que conocemos de transformaciones a partir de muestras Normales podríamos utilizar el siguiente pivote:

$$
U=\frac{(n-1)S^2}{\sigma^2}
$$

Intervalo de confianza para σ^2

Sabemos que $U \sim \chi^2(n-1)$. Por lo tanto esta variable aleatoria cumple todas las condiciones necesarias para ser un pivote. Ahora debemos encontrar los valores de a y b tales que:

$$
P(U<a)=\frac{\alpha}{2} \qquad P(U>b)=\frac{\alpha}{2}
$$

El valor de a es aquel que acumula $1-\frac{\alpha}{2}$ a su derecha, mientras que el valor de b solo acumula $\frac{\alpha}{2}$ a la derecha. Por lo tanto tenemos que $a=\chi^2_{1-\frac{\alpha}{2},n-1}$ y $b=\chi^2_{\frac{\alpha}{2},n-1}$. En este caso no hay simetría que podamos utilizar para poner a en términos de b como hemos estado haciendo anteriormente. Ahora despejamos σ^2:

$$
\begin{aligned}
P(a\le U\le b)&=P\left(\chi^2_{1-\frac{\alpha}{2},n-1}\le\frac{(n-1)S^2}{\sigma^2}\le\chi^2_{\frac{\alpha}{2},n-1}\right)\\
&=P\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}\le\sigma^2\le\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right)=1-\alpha
\end{aligned}
$$

Intervalo de confianza para σ^2

Por lo tanto, con una confianza del $(1-\alpha)\%$ el intervalo $\left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right]$ incluye el valor de σ^2.

Ahora volvemos al caso donde tenemos dos poblaciones independientes, una de las cuales es $N(\mu_1, \sigma_1^2)$ y la otra es $N(\mu_2, \sigma_2^2)$. Supongase que se obtiene una muestra $X_1, X_2, ..., X_n$ de la primera población y otra $Y_1, Y_2, ..., Y_m$ de la segunda. Nuestro interés yace en encontrar una estimación por intervalo, con probabilidad $1-\alpha$, para el parámetro $\frac{\sigma_1^2}{\sigma_2^2}$.

Intervalo de confianza para $\frac{\sigma_1^2}{\sigma_2^2}$

Debemos construir un parámetro que incluya las dos muestras aleatorias y al parámetros desconocido. En el caso anterior habiamos usado una χ^2-cuadrado como pivote ya que esta incluía la muestra aleatoria y la varianza poblacional. Podriamos pensar en hacer lo mismo para cada muestra y de alguna forma juntar ambas χ^2-cuadrado. Para ello recordemos la forma de una distribución F:

$$F = \frac{\frac{V}{v_1}}{\frac{W}{v_2}}$$

donde $V \sim \chi^2(v_1)$ y $W \sim \chi^2(v_2)$. Sabemos de antemano que $\frac{(n-1)S_1^2}{\sigma_1^2}$ $\chi^2(n-1)$ y $\frac{(m-1)S_2^2}{\sigma_2^2}$ $\chi^2(m-1)$. Podemos usar estas variables aleatorias en la construcción de una F que incluya el parámetro de interés:

$$F = \frac{\frac{\frac{(n-1)S_1^2}{\sigma_1^2}}{n-1}}{\frac{\frac{(m-1)S_2^2}{\sigma_2^2}}{m-1}} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F(n-1, m-1)$$

Intervalo de confianza para $\frac{\sigma_1^2}{\sigma_2^2}$

Esta variable aleatoria cumple todas las condiciones para ser un pivote y podemos usarlo para encontrar los valores de a y b. Recordemos que se debe cumplir que $P(F < a) = P(F > b) = \frac{\alpha}{2}$. Podemos hacer uso de las tablas de la F, la cual acumula hacia la derecha. Por lo tanto tenemos que a es el valor F que acumula $1-\frac{\alpha}{2}$ a la derecha mientras que b acumula $\frac{\alpha}{2}$. Podemos notar estos valores como

$a = F_{1-\frac{\alpha}{2},n-1,m-1} \qquad b = F_{\frac{\alpha}{2},n-1,m-1}$

No obstante, vemos que hay un problema al usar las tablas para encontrar $F_{1-\frac{\alpha}{2},n-1,m-1}$ si la probabilidad que se busca es mayor a 0.20, el cual es el caso en la gran mayoría de los casos. Para ello podemos hacer uso de una propiedad de la distribución F, que dice que la inversa multiplicativa de una F también se distribuye F pero con los grados de libertad del numerador y denominador intercambiados. Por lo tanto podemos decir que: $a = F_{1-\frac{\alpha}{2},n-1,m-1} = F^{-1}_{\frac{\alpha}{2},m-1,n-1}$.

Intervalo de confianza para $\frac{\sigma_1^2}{\sigma_2^2}$

Finalmente podemos proceder a despejar el valor del parámetro de interés de la expresión $P(a \leq F \leq b)$ y obtenemos el siguiente intervalo:

$$\left[\frac{\frac{S_1^2}{S_2^2}}{F_{\frac{\alpha}{2},n-1,m-1}}, \frac{\frac{S_1^2}{S_2^2}}{F^{-1}_{\frac{\alpha}{2},m-1,n-1}}\right]$$

el cual incluye el valor de $\frac{\sigma_1^2}{\sigma_2^2}$ con una confianza del $(1-\alpha)\%$.

Ejercicio: Un instituto de investigaciones agronómicas siembra, en cinco parcelas diferentes, dos tipos de maíz híbrido. Las producciones en quintales métricos por hectárea son:

	1	2	3	4	5
Híbrido I	90	85	95	76	80
Híbrido II	84	87	90	92	90

Construir un intervalo de confianza para el cociente de varianzas con un error de significación de 0,10 .

Solución: Asumiendo que las poblaciones son normales defina $X_1 \equiv$ 'Producción de maíz del híbrido I' como $N(\mu_1, \sigma_1)$ y $X_2 \equiv$ ' Producción de maíz del híbrido II' sigue una distribución $N(\mu_2, \sigma_2)$

Entonces

$$\begin{array}{llll} n_1 = 5 & \bar{x}_1 = 85,20 & s_1^2 = 57,7 & \alpha/2 = 0,05 \\ n_2 = 5 & \bar{x}_2 = 88,6 & s_2^2 = 9,8 & s_1^2/s_2^2 = 57,7/9,8 = 5,89 \end{array}$$

$$F_{0,05;4,4} = 6,3883 \quad F_{0,95;4,4} = 1/F_{0,05;4,4} = 1/6,3883 = 0,1565$$

$$\begin{aligned} &\left[\frac{\frac{S_1^2}{S_2^2}}{F_{0.05,5-1,5-1}}, \frac{\frac{S_1^2}{S_2^2}}{F^{-1}_{0.05,5-1,5-1}}\right] \\ &= \left[\frac{5.89}{6.3883}, \frac{5.89}{0.1565}\right] \\ &= [0.92, 37.64] \end{aligned}$$

Intervalos de confianza para muestras grandes

Recordemos que cuando $n \to +\infty$ el Teorema del Límite Central nos dice que $\overline{X}$ converge en distribución a una $N(\mu, \frac{\sigma^2}{n})$. Por lo tanto, si queremos hacer una estimación por intervalo para μ, suponiendo que conocemos el valor de σ^2, podemos hacer uso del siguiente pivote:

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

De esta forma tenemos una cantidad pivote asintótica que se puede aplicar para este tipo de estimadores con muestras grandes. El desarrollo para obtener el intervalo es el mismo del caso de una estimación por intervalo para la media de una población normal. Por lo tanto tendremos un intervalo bilateral con probabilidad $1-\alpha$:

$$\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Intervalos de confianza para muestras grandes (Método de Wald)

Existen varios estimadores que se puede comprobar que tienen la forma de un $\overline{X}$ como por ejemplo $\hat{p}$, $\overline{X} - \overline{Y}$ y $\hat{p}_1 - \hat{p}_2$, para estimar p, $\mu_1 - \mu_2$ y $p_1 - p_2$ respectivamente. Cada uno de estos caso también tendrá una expresión específica para la variancia ($\frac{p(1-p)}{n}$, $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ y $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$, respectivamente).

Nótese que este intervalo supone que conocemos la varancia poblacional de nuestra población. Sin embargo, podemos hacer uso del Teorema de Slutsky para mantener el pivote que converge en distribución a una $N(0,1)$ si conocemos un estimador consistente para la variancia poblacional. Este reemplazaría la variancia poblacional por su estimador consistente en la fórmula del intervalo anterior.

Ejemplo: Se registraron los tiempos de compra de $n = 64$ clientes seleccionados al azar en un supermercado local. El promedio y varianza de los 64 tiempos de compra fueron 33 minutos y 256 minutos ^{2}, respectivamente. Estime μ, el verdadero promedio de tiempo de compra por cliente, con un coeficiente de confianza de $1 - \alpha = .90$ usando el método de Wald.

Solución: El intervalo tiene la forma

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \approx \bar{x} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right) \tag{1}$$

De la tabla de normales se deduce que $z_{\alpha/2} = z_{0.05} = 1.645$.

Por lo tanto el intervalo es

$$\left[33 - 1.645 \left(\frac{16}{8} \right), 33 + 1.645 \left(\frac{16}{8} \right) \right] \qquad = [29.71, 36.29]$$

Ejemplo Suponga que tienen datos te tiempos de espera en minutos en un banco con $n = 25$ clientes. La distribución del tiempo se observa que sigue una distribución $Poisson(\theta)$ con $\bar{X} = 5$. Cual es el intervalo de confianza para θ a un 95%.

Solución: Recuerde que si X es $Poisson(\theta)$, entonces $E[X] = Var(X) = \theta$.

Por el TLC se tiene que

$$\bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \approx \bar{X} \pm z_{\alpha/2}\left(\frac{\sqrt{\theta}}{\sqrt{n}}\right) \approx \bar{X} \pm z_{\alpha/2}\left(\frac{\sqrt{\bar{X}}}{\sqrt{n}}\right) \tag{2}$$

$$\begin{aligned}&\left[5 - 1.96\left(\frac{\sqrt{5}}{5}\right), 5 - 1.96\left(\frac{\sqrt{5}}{5}\right)\right]\\ &= [4.12, 5.88]\end{aligned}$$

Funciones estabilizadoras de varianza

El caso anterior requería hacer el supuesto que los datos eran normales para construir el intervalo.

Vamos a ver si podemos hacerlo mejor.

La pregunta que siempre debemos hacernos, es

¿Cómo transformar $\hat{\theta}$ para que tenga varianza constante?

Ejemplo: Retomemos el caso de los clientes en el banco.

En este caso hicimos el supuesto de normalidad y aproximamos la varianza con $\overline{X}$

Por el método Delta, se tiene que para $g\left(\bar{X}_n\right)$

$$\sqrt{n}(g(\overline{X} - g(\mu))) \xrightarrow{d} N(0, (g'(\mu)\sigma)^2)$$

$$(g'(\mu)\sigma)^2 = g'(\mu)^2\sigma^2(\mu)$$

Si se desea que la varianza sea constante con respecto a μ,

$$\begin{aligned}&g'(u)^2\sigma^2(\mu) = 1\\ &\implies g'(\mu) = \frac{1}{\sigma(\mu)} \quad (\sigma(\mu) > 0)\\ &\implies g(\mu) = \int_a^\mu \frac{dx}{\sigma(x)}dx\end{aligned}$$

donde a es una constante arbitraria que hace la integral finita (y fácil de calcular).

Del ejemplo anterior, recuerde que $\sigma^2 = \theta = \mu$, entonces se podría tomar que $\sigma(\mu) = \sqrt{\mu}$ por lo tanto definimos

$$g(\mu) = \int_0^{\mu} \frac{dx}{\sqrt{x}} = 2\sqrt{\mu}$$

Por el método Delta, sabemos que

$$\sqrt{n}(2\bar{X}_n^{\frac{1}{2}} - 2\theta^{\frac{1}{2}}) \xrightarrow{d} N(0,1)$$

O lo que es equivalente

$$2\bar{X}_n^{\frac{1}{2}} \xrightarrow{d} N\left(2\theta^{\frac{1}{2}}, \frac{1}{n}\right)$$

De esta manera un intervalo de confianza para la cantidad $2\sqrt{\theta}$ con nivel 95% sería

$$2\bar{X}_n^{\frac{1}{2}} \pm z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}}$$

Usando los valores que teníamos anteriormente tenemos que

$$\begin{aligned}
&\left[2\sqrt{5} - 1.96\frac{1}{\sqrt{25}}, 2\sqrt{5} - 1.96\frac{1}{\sqrt{25}}\right] \\
&= [4.27, 4.67]
\end{aligned}$$

Sin embargo esa cantidad no nos sirve para el ejercicio. Requerimos un intervalo para μ.

Para eso vamos tomar la función inversa de cada uno de los extremos $y = 2x^{\frac{1}{2}} \implies x = \frac{y^2}{4}$. Aplicando esta transformación al intervalo anterior, se obtiene

$$\begin{aligned}
&\left[\frac{1}{4}\left(2\bar{X}_n^{\frac{1}{2}} - \frac{1}{\sqrt{n}} z_{\frac{\alpha}{2}}\right)^2, \frac{1}{4}\left(2\bar{X}_n^{\frac{1}{2}} + \frac{1}{\sqrt{n}} z_{\frac{\alpha}{2}}\right)^2\right] \\
&= \left[\frac{1}{4}(4.27)^2, \frac{1}{4}(4.67)^2\right] \\
&= [4.56, 5.46]
\end{aligned}$$

Ejemplo: Suponga que $X_i \sim Binomial(n,p)$ encuentre un intervalo de confianza para p usando el método de Wald y el método delta al 95%, con $n = 25$ y $\bar{X} = 0.5$.

El estimador para p es $\bar{X}$.

Método Wald:

$$\bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \approx \bar{X} \pm z_{\alpha/2}\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \approx \bar{X} \pm z_{\alpha/2}\left(\frac{\sqrt{\bar{X}(1-\bar{X})}}{\sqrt{n}}\right)$$

Haciendo el cálculo correspondiente el intervalo queda en

$$[0.304, 0.696]$$

Método delta:

Hay que encontrar la $g(p)$ para estabilizar la varianza.

$$g(p) = \int_a^p \frac{1}{\sqrt{x(1-x)}}dx = 2\int_a^{\sqrt{p}} \frac{1}{\sqrt{1-u^2}}du = 2\arcsin(\sqrt{p})$$

$$\sqrt{n}(2\arcsin(\sqrt{\bar{X}_n}) - 2\arcsin(\sqrt{p})) \xrightarrow{d} N(0,1)$$

$$2\arcsin(\sqrt{\bar{X}_n}) \pm z_{\alpha/2}\frac{1}{\sqrt{n}}$$

usando los valores obtenemos

$$[1.179, 1.963]$$

$$y = 2\arcsin(\sqrt{x}) \Rightarrow x = \sin\left(\frac{y}{2}\right)^2$$

$$\left[\sin\left(\frac{1.179}{2}\right)^2, \sin\left(\frac{1.963}{2}\right)^2\right]$$
$$= [0.309, 0.691]$$

Ejercicio: Suponga que (X_i, Y_i) provienen de una distribución normal con correlación ρ.

Suponga que usted estima la correlación muestral con la siguiente fórmula:

$$r_n = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left\{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2\right\}^{1/2}}$$

Haciendo mucho cálculos, encontramos $Var(r_n) = 1 - \rho^2$. Por lo tanto,

$$\sqrt{n}\left(r_n - \rho\right) \rightsquigarrow N\left(0, \left(1-\rho^2\right)^2\right)$$

Si suponemos que $r_n = 0.3$ y $n = 25$, encuentre un intervalo de confianza para ρ al 95% usando un método de Wald y el método Delta.

Sugerencia:

Use el hecho que

$$\int \frac{1}{1-\rho^2}d\rho = \frac{1}{2}\log\frac{1+\rho}{1-\rho} = \operatorname{arctanh}\rho$$