Probabilidad y Estadística – INTERVALOS DE CONFIANZA – UTN

En la primera parte de la unidad trabajamos con estimadores puntuales para los parámetros, es decir un único valor para aproximar al parámetro desconocido.

En esta segunda parte lo que nos va a interesar es encontrar un intervalo de posibles valores para el parámetro de interés, de manera tal que el parámetro se encontrará dentro de él con cierto grado de "confianza". Que la manera de medir ese grado será por medio de probabilidades.

Solo buscaremos estos intervalos para los parámetros μ, σ^2, σ y p, es decir para la esperanza, varianza, desvío estándar y proporción.

• Comencemos buscando la estimación por intervalos de confianza para μ, la esperanza de una variable aleatoria.
Al igual que en la primera parte esta estimación la haremos en base a una muestra aleatoria $X_1, \ldots, X_n$:

1er Caso $X_1, \ldots, X_n$ v.a. i.i.d. con $X_i \sim N(\mu, \sigma^2)$, σ conocido!, para todo $i = 1, \ldots, n$:

Para construir el intervalo deseado, de manera tal que el parámetro se encuentre en él con cierta "probabilidad", tenemos que partir de un estimador del parámetro de interés cuya distribución las conozcamos para poder calcular probabilidad.

Recordemos que el estimador usual de la esperanza es $\overline{X}$ y como la m.a. es normal, tenemos que

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

pero para poder calcular probabilidades tiene que estar estandarizado:

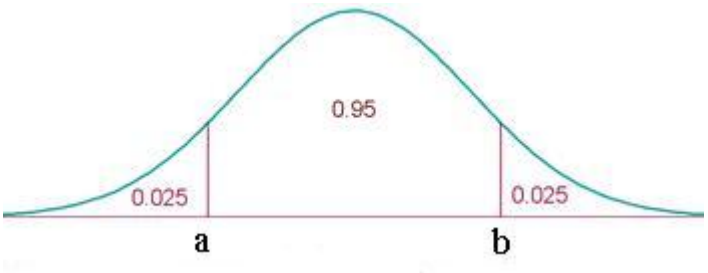
$$Z = \sqrt{n}\frac{\overline{X} - \mu}{\sigma} \sim N\left(0; 1\right),$$

Nuestro objetivo era encontrar un intervalo para μ, por lo tanto podemos acotar a Z por debajo y por arriba para luego despejar y encontrar las cotas que formen el intervalo para la esperanza.
A esta variable aleatoria Z cuya distribución es conocida y a partir de la cual despejaremos μ, se la denomina **pivote**.

Las cotas las tenemos que buscar de manera tal que la probabilidad sea alta, por ejemplo, quisiera encontrar a y b de modo que la probabilidad sea del 0,95, es decir:

$$P\left(a \leq Z \leq b\right) = 0{,}95$$

Entonces, quiénes deberían ser a y b para que el intervalo $[a; b]$ sea lo más chico posible? Viendo el gráfico de la función de densidad de una distribución normal estándar, como es simétrica, a tiene que ser el valor del recorrido que deja área a izquierda $0{,}05/2 = 0{,}025$ y b el valor que deja área a derecha 0,025, con lo cual a su izquierda deja área 0,975.



Si buscamos en la tabla de la $N(0;1)$, tenemos que $a = -1{,}96$ y $b = 1{,}96$. Para nombrar a estos valores a y b se usa la siguiente notación especial:

$$a = z_{0{,}025} \text{ y } b = z_{0{,}975},$$

es decir que son valores del recorrido de Z que acumulan probabilidad 0,025 y 0,0975.

Pero estos valores corresponden al caso puntual en que la probabilidad deseada sea 0,95, si lo queremos generalizar, a ese valor alto de probabilidad lo notaremos por

$$1 - \alpha$$

donde α es un número chico, en el caso particular en que $1 - \alpha = 0{,}95$, resulta que $\alpha = 0{,}05$. Por lo tanto

$$a = z_{\alpha/2} \text{ y } b = z_{1-\alpha/2}$$

o, como la normal es simétrica, podemos usar que

$$a = -z_{1-\alpha/2} \text{ y } b = z_{1-\alpha/2},$$

de esta manera la probabilidad deseada queda planteada como

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

Ahora escribamos quien es Z así podemos despejar μ

$$\begin{aligned}
1-\alpha &= P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) \\
&= P(-z_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X}-\mu}{\sigma} \leq z_{1-\alpha/2}) \\
&= P(-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \leq \overline{X}-\mu \leq \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}) \\
&= P(-\overline{X}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \leq -\mu \leq -\overline{X}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}) \text{ multiplico todo por menos:} \\
&= P(\overline{X}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \geq \mu \geq \overline{X}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}) \text{ lo escribo en orden:} \\
&= P(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \leq \mu \leq \overline{X}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2})
\end{aligned}$$

Por lo tanto decimos que encontramos un

Intervalo de Confianza de nivel $1-\alpha$ para μ: $\left[\overline{X}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}; \overline{X}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}\right]$

2do Caso $X_1, \ldots, X_n$ v.a. i.i.d. con $X_i \sim N(\mu, \sigma^2)$, σ desconocido!, para todo $i = 1, \ldots, n$:

Si bien este contexto es similar al anterior, si miramos el intervalo hallado, no lo podríamos calcular para este caso ya que no conocemos el valor de σ. Veamos qué se puede hacer en esta situación.

Recordemos que el despeje del intervalo surgió a partir de

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

pero como no conocemos el valor de σ debemos estimarlo, y eso lo haremos utilizando el estimador S. Resulta que al reemplazar σ por S, la variable aleatoria, NO tiene distribución $N(0;1)$:

$$T = \sqrt{n}\frac{\overline{X} - \mu}{S} \cancel{\sim N(0;1)},$$

Qué distribución tendrá esta variable pivote? Para poder responder a esta pregunta recordemos:

1. Distribución Ji-cuadrada o Chi-cuadrada con k grados de libertad - χ^2_k:

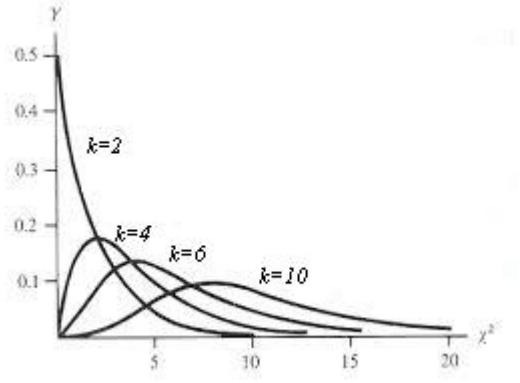
   Esta distribución surge de la siguiente manera, si uno tiene $Z_1, \ldots, Z_k$, v.a. i.i.d.$\sim N(0;1)$ y considera la suma de ellas elevadas al cuadrado, es decir

   $$X = Z_1^2 + Z_2^2 + \ldots + Z_k^2,$$

   entonces $\boxed{X \sim \chi^2_k}$.

   El nombre de la distribución incluye el cuadrado para enfatizar que necesariamente su recorrido son los números reales positivos, ya que se construye sumando los cuadrados de otras variables, y el parámetro, que es la cantidad de términos que se están sumando, es k.

   En el siguiente gráfico tenemos las funciones de densidad para algunos valores de k:

   y podemos ver que no es simétrica respecto al eje y.

   Por otra parte, la función de distribución acumulada se encuentra tabulada en la guía de tablas para diferentes valores de k en las páginas XX y XXI.

   Para qué sirve esta distribución? Recordemos que trabajamos con dos estimadores usuales, $\overline{X}$ y S^2. La clase pasada ya trabajamos con $\overline{X}$ pero todavía no mencionamos nada sobre la varianza muestral. Bueno, resulta

que así como al estandarizar la media muestral obtenemos distribución normal estándar, existe una manera de "estandarizar" S^2 para tener una distribución tabulada:

TEOREMA: Sean $X_1, \ldots, X_n$ v.a. i.i.d. $\sim N(\mu; \sigma)$, entonces si $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$,

$$\boxed{(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}},$$

porque al hacer el desarrollo de S^2 y algunas cuentas extras, queda que $(n-1)\frac{S^2}{\sigma^2}$ es la suma de $n-1$ variables aleatorias $N(0;1)$ elevadas al cuadrado.

2. Distribución t de Student con k grados de libertad - t_k:

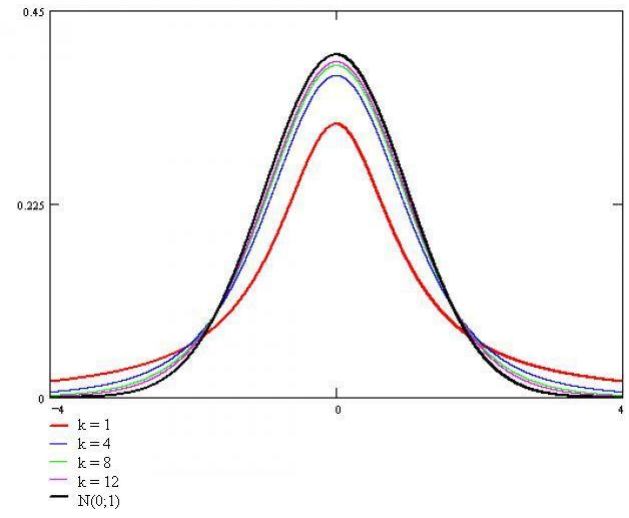
Finalmente esta variable aleatoria nos servirá para obtener la distribución del pivote de interés T.

La distribución t de Student surge a partir de dos variables aleatorias independientes, que son, $Z \sim N(0;1)$ y $V \sim \chi^2_k$ y que se construye de la siguiente manera:

$$X = \frac{Z}{\sqrt{V/k}},$$

entonces $\boxed{X \sim t_k}$.

Como está construida, esta distribución comparte la simetría con respecto al eje y de la normal estándar. En el siguiente gráfico vemos el comportamiento de la función de densidad para distintos valores de k:



a medida que aumenta k se parece a una $N(0;1)$.

Y los valores de las probabilidades acumuladas figuran también en la guía de fórmulas.

Tanto para la Ji-cuadrada como la t de Student la tabla funciona al revés que en las distribuciones ya vistas, de la siguiente manera: en la parte o fila superior se encuentra la probabilidad acumulada, en la columna de la izquierda el valor del parámetro y dentro de la tabla están los valores del recorrido.

Por último, a partir de la definición, se puede demostrar el siguiente teorema:

TEOREMA: Sean $X_1, \ldots, X_n$ v.a. i.i.d. $\sim N(\mu; \sigma)$, entonces, como $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$,

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

De esta manera tenemos la distribución del pivote a partir del cual vamos a poder despejar μ. Repitiendo el planteo del 1er caso:

$$\begin{aligned}
1-\alpha &= P(a \leq T \leq b) \\
&= P\left(t_{n-1;\alpha/2} \leq T \leq t_{n-1;1-\alpha/2}\right) \text{ donde } t_{n-1;\alpha/2} \text{ es el que acumula probabilidad } \alpha/2 \text{ y el otro } 1-\alpha/2. \\
&\quad \text{Como la distribución es simétrica:} \\
&= P\left(-t_{n-1;1-\alpha/2} \leq T \leq t_{n-1;1-\alpha/2}\right) \\
&= P\left(-t_{n-1;1-\alpha/2} \leq \sqrt{n}\frac{\overline{X} - \mu}{S} \leq t_{n-1;1-\alpha/2}\right) \\
&\quad \text{Despejando } \mu \text{ igual que en el 1er caso, nos queda:} \\
&= P\left(\overline{X} - \frac{S}{\sqrt{n}}t_{n-1;1-\alpha/2} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}}t_{n-1;1-\alpha/2}\right)
\end{aligned}$$

Por lo tanto, nos queda un

$$\boxed{\text{Intervalo de Confianza de nivel } 1-\alpha \text{ para } \mu\text{: } \left[\overline{X} - \frac{S}{\sqrt{n}}t_{n-1;1-\alpha/2}; \overline{X} + \frac{S}{\sqrt{n}}t_{n-1;1-\alpha/2}\right]}$$

3er Caso $X_1, \ldots, X_n$ v.a. i.i.d. con $E(X_i) = \mu$, $V(X_i) = \sigma^2$, σ conocido!, para todo $i = 1, \ldots, n$ y $n > 30$:

El pivote para esta situación será, usando el Teorema Central del Límite (TCL):

$$Z = \sqrt{n}\frac{\overline{X} - \mu}{\sigma} \overset{(a)}{\sim} N(0;1),$$

por lo tanto la probabilidad que tendremos será aproximada, pero podemos repetir los pasos hechos en el 1er Caso:

$$\begin{aligned}
1-\alpha = P\left(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) &\simeq P(-z_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X} - \mu}{\sigma} \leq z_{1-\alpha/2}) \\
&= P(-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \leq \overline{X} - \mu \leq \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}) \\
&= P(-\overline{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \leq -\mu \leq -\overline{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}) \text{ multiplico todo por menos:} \\
&= P(\overline{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \geq \mu \geq \overline{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}) \text{ lo escribo en orden:} \\
&= P(\overline{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2})
\end{aligned}$$

Por lo tanto decimos que encontramos un

Intervalo de Confianza de nivel aproximado $1-\alpha$ para μ: $\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}; \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]$

4to Caso $X_1, \ldots, X_n$ v.a. i.i.d. con $E(X_i) = \mu$, $V(X_i) = \sigma^2$, σ desconocido!, para todo $i = 1, \ldots, n$ y $n > 30$:

Para esta última situación necesitamos mencionar la siguiente propiedad, que se deduce del TCL:
Si $X_1, \ldots, X_n$ son v.a. i.i.d. con $E(X_i) = \mu$, $V(X_i) = \sigma^2$, entonces la distribución límite no se ve modificada al reemplazar σ por S, es decir que

$$\sqrt{n}\frac{\overline{X} - \mu}{S} \xrightarrow{\text{“}D\text{“}} N(0;1),$$

por lo tanto, si $n > 30$ $\sqrt{n}\frac{\overline{X} - \mu}{S} \overset{(a)}{\sim} N(0;1)$.

Así que estamos en condiciones de plantear el siguiente pivote para despejar μ:

$$Z = \sqrt{n}\frac{\overline{X} - \mu}{S} \overset{(a)}{\sim} N(0;1),$$

que es análogo al caso anterior con la única diferencia que figura S en lugar de σ, por lo tanto nos queda que

Intervalo de Confianza de nivel aproximado $1-\alpha$ para μ es: $\left[\overline{X} - \frac{S}{\sqrt{n}} z_{1-\alpha/2}; \overline{X} + \frac{S}{\sqrt{n}} z_{1-\alpha/2}\right]$

• Ahora buscaremos la estimación por intervalos de confianza para σ^2 y σ, la varianza y el desvío estándar de una variable aleatoria.

El único caso en el que la varianza muestral se puede “estandarizar” para que tenga una distribución conocida, es cuando se tiene una muestra aleatoria $X_1, \ldots, X_n$ de distribución $N(\mu, \sigma^2)$, sin importar si μ es conocido o no.

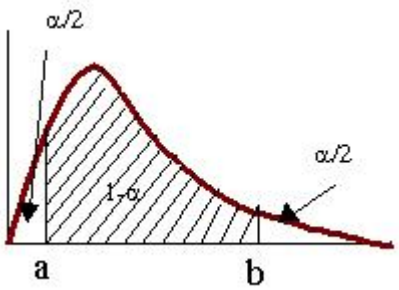
El pivote que nos servirá para el despeje de σ^2 es:

$$V = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Queremos encontrar a y b tales que

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

Recordemos que el gráfico de la función de densidad de una distribución Ji-cuadrada es asimétrico, como por ejemplo:



Por lo tanto las constantes serán: $a = \chi^2_{n-1;\alpha/2}$ y $b = \chi^2_{n-1;1-\alpha/2}$ (aquellos valores del recorrido que acumulen probabilidad $\alpha/2$ y $1-\alpha/2$ respectivamente).

Entonces, el planteo para despejar σ^2 será:

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{n-1;\alpha/2} \leq V \leq \chi^2_{n-1;1-\alpha/2}\right) \\ &= P\left(\chi^2_{n-1;\alpha/2} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{n-1;1-\alpha/2}\right) \\ &= P\left(\frac{\chi^2_{n-1;\alpha/2}}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{n-1;1-\alpha/2}}{(n-1)S^2}\right) \text{ invertimos las fracciones:} \\ &= P\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right) \text{ y si lo escribimos en orden, queda:} \\ &= P\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right). \end{aligned}$$

Por lo tanto para la varianza tenemos que el

Intervalo de Confianza de nivel $1-\alpha$ para σ^2 es: $\left[\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right]$

Por otra parte, de la última probabilidad podemos despejar σ, recordando que $\sigma > 0$:

$$\begin{aligned} 1-\alpha &= P\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right) \\ &= P\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}}\right) \end{aligned}$$

obteniendo el:

Intervalo de Confianza de nivel $1-\alpha$ para σ: $\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}}; \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}}\right]$

- Por último nos queda la estimación por intervalo de confianza para p, una proporción.

Para estudiar esta situación se necesitan tener $X_1, \ldots, X_n$ v.a. i.i.d., $X_i \sim Be(p)$ para todo $i = 1, \ldots, n$, con $n > 30$. Entonces como la $E(X_i) = p$ y la $V(X_i) = p(1-p)$, tenemos que por el TCL

$$Z_p = \sqrt{n}\frac{\overline{X} - p}{\sqrt{p(1-p)}} \overset{(a)}{\sim} N(0;1),$$

pero Z_p no sirve como pivote ya que queda una ecuación a partir de la cual no se puede despejar p, pero lo que sí podemos hacer es estimar p en el divisor, con $\hat{p} = \overline{X}$, lo cual no modificará la distribución aproximada, obteniendo el siguiente pivote:

$$Z = \sqrt{n}\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}} = \sqrt{n}\frac{\overline{X} - p}{\sqrt{\overline{X}(1-\overline{X})}} \overset{(a)}{\sim} N(0;1).$$

Ahora sí el planteo para despejar p:

$$\begin{aligned}
1-\alpha &= P\left(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) \\
&\simeq P\left(z_{1-\alpha/2} \leq \sqrt{n}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}} \leq z_{1-\alpha/2}\right) \\
&= P\left(-\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2} \leq \hat{p}-p \leq \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2}\right) \\
&= P\left(-\hat{p}-\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2} \leq -p \leq -\hat{p}+\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2}\right) \\
&= P\left(\hat{p}+\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2} \geq p \geq \hat{p}-\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2}\right) \\
&= P\left(\hat{p}-\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2} \leq p \leq \hat{p}+\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2}\right)
\end{aligned}$$

Por lo tanto nos queda el

Intervalo de confianza de nivel aproximado $1-\alpha$ para p: $\left[\hat{p}-\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2}; \hat{p}+\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}z_{1-\alpha/2}\right]$, donde $\hat{p}=\overline{X}$.

Definiciones: Dado un intervalo de confianza cualquiera $\left[L_i; L_s\right]$

1. Se define longitud del intervalo a $L = L_s - L_i$.
2. Se define error del intervalo a $Err = \frac{L}{2}$, es decir la longitud divido 2.

Observación: Si bien en esta materia usamos la notación $z_{1-\alpha}$ para indicar el valor del recorrido que deja probabilidad acumulada $1-\alpha$, también se suele notar a ese mismo valor como z_α, para abreviar escritura.