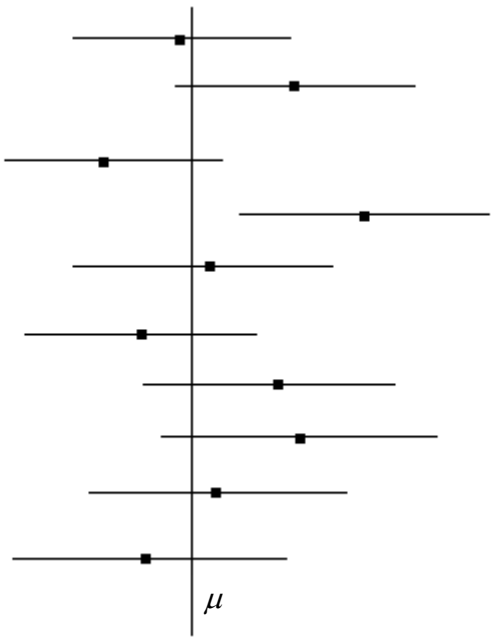
# 9- Intervalos de confianza

## 9.1 – Introducción

Se ha visto como construir a partir de una muestra aleatoria un estimador puntual de un parámetro desconocido. En esos casos necesitábamos dar algunas características del estimador, como por ejemplo si era insesgado o su varianza.
A veces resulta más conveniente dar un ***intervalo de valores posibles*** del parámetro desconocido, de manera tal que dicho intervalo contenga al verdadero parámetro con determinada probabilidad.
Específicamente, a partir de una muestra aleatoria se construye un intervalo $(\hat{\Theta}_1, \hat{\Theta}_2)$ donde los extremos $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estadísticos, tal que $P(\theta \in (\hat{\Theta}_1, \hat{\Theta}_2)) = 1-\alpha$ donde $\theta$ es el parámetro desconocido a estimar y $\alpha$ es un valor real entre cero y uno dado de antemano. Por ejemplo si $\alpha = 0.05$, se quiere construir un intervalo $(\hat{\Theta}_1, \hat{\Theta}_2)$ tal que $P(\theta \in (\hat{\Theta}_1, \hat{\Theta}_2)) = 0.95$, o escrito de otra forma $P(\hat{\Theta}_1 \leq \theta \leq \hat{\Theta}_2) = 0.95$
Esta probabilidad tiene el siguiente significado: como $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son estadísticos, los valores que ellos toman varían con los valores de la muestra, es decir si $x_1, x_2, ..., x_n$ son los valores medidos de la muestra entonces el estadístico $\hat{\Theta}_1$ tomará el valor $\theta_1$ y el estadístico $\hat{\Theta}_2$ tomará el valor $\theta_2$. Si medimos nuevamente la muestra obtendremos ahora valores $x_1', x_2', ..., x_n'$ y por lo tanto $\hat{\Theta}_1$ tomará el valor $\theta_1'$ y el estadístico $\hat{\Theta}_2$ tomará el valor $\theta_2'$, diferentes en general de los anteriores. Esto significa que si medimos la muestra 100 veces obtendremos 100 valores diferentes para $\hat{\Theta}_1$ y $\hat{\Theta}_2$ y por lo tanto obtendremos 100 intervalos distintos, de los cuales aproximadamente 5 de ellos no contendrán al verdadero parámetro.
Al valor $1-\alpha$ se lo llama ***nivel de confianza*** del intervalo. También se suele definir como nivel de confianza al $(1-\alpha)100\%$
La construcción repetida de un intervalo de confianza para $\mu$ se ilustra en la siguiente figura



## 9.2 – Intervalo de confianza para la media de una distribución normal, varianza conocida.

El método general para construir intervalos de confianza es el siguiente llamado ***método del pivote:***

Supongamos el siguiente caso particular, sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de tamaño *n* de una v.a. *X* donde $X \sim N(\mu, \sigma^2)$, $\sigma^2$ conocido, se quiere construir un intervalo de confianza para $\mu$ de nivel $1-\alpha$. Supongamos $\alpha = 0.05$.

1- tomamos un estimador puntual de $\mu$, sabemos que $\hat{\mu} = \overline{X}$ es un estimador con buenas propiedades.

2- a partir de $\hat{\mu} = \overline{X}$ construimos el estadístico $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$. Notar que ***Z (pivote)*** contiene al verdadero parámetro $\mu$ y que bajo las condiciones dadas $Z \sim N(0,1)$

3- como conocemos la distribución de *Z*, podemos plantear: hallar un número $z$ tal que $P(-z \le Z \le z) = 0.95$

Por la simetría de la distribución normal estándar podemos escribir

$$P(-z \le Z \le z) = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1 = 0.95 \quad \Rightarrow \quad \Phi(z) = 0.975 \quad \Rightarrow \quad z = 1.96$$

Por lo tanto $P(-1.96 \le Z \le 1.96) = P\left(-1.96 \le \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right) = 0.95$

Despejamos $\mu$:

$$P\left(-1.96 \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right) = P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) =$$

$$= P\left(-1.96\frac{\sigma}{\sqrt{n}} - \overline{X} \le -\mu \le 1.96\frac{\sigma}{\sqrt{n}} - \overline{X}\right) = P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Entonces

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = P\left(\mu \in \left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)\right) = 0.95$$

Es decir el intervalo de confianza para $\mu$ es $\left(\overline{X} - 1.96\dfrac{\sigma}{\sqrt{n}}; \overline{X} + 1.96\dfrac{\sigma}{\sqrt{n}}\right)$ y tiene nivel de confianza 0.95 o 95%.

Aquí $\hat{\Theta}_1 = \overline{X} - 1.96\dfrac{\sigma}{\sqrt{n}}$ y $\hat{\Theta}_2 = \overline{X} + 1.96\dfrac{\sigma}{\sqrt{n}}$

Repetimos el procedimiento anterior y construimos un intervalo de confianza para $\mu$ con nivel de confianza $1-\alpha$

1-Partimos de la esperanza muestral $\overline{X} = \frac{1}{n}\sum_{1=1}^{n} X_i$ para una muestra aleatoria $(X_1, X_2, ..., X_n)$ de tamaño *n*. Sabemos que es un estimador insesgado y consistente de $\mu$.
2-Construimos el estadístico

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

La variable aleatoria *Z* cumple las condiciones necesarias de un pivote
Para construir un intervalo de confianza al nivel de confianza 1-$\alpha$ partiendo del pivote *Z*, comenzamos por plantear la ecuación

$$P(-z \le Z \le z) = 1-\alpha,$$

donde la incógnita es el número real *z*.

Si reemplazamos la v.a. *Z* por su expresión tenemos:

$$P\left(-z \le \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \le z\right) = P\left(-z\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le z\frac{\sigma}{\sqrt{n}}\right) = P\left(-\overline{X} - z\frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X} + z\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

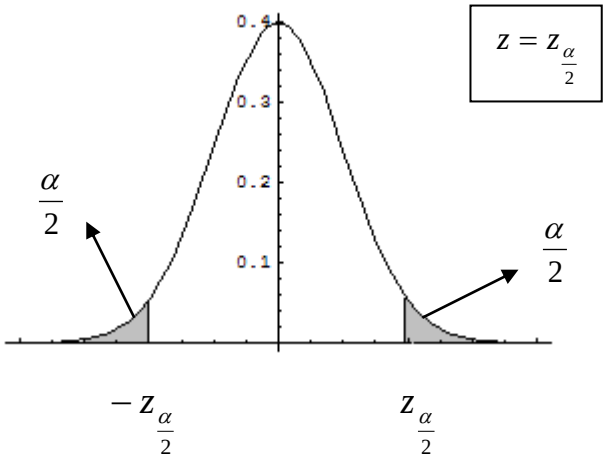
Multiplicando todos los miembros de la desigualdad por -1 (el orden de los miembros se invierte) llegamos a:

$$P\left(\overline{X} - z\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Evidentemente, si definimos

$$\begin{cases} \hat{\Theta}_1 = \overline{X} - z\frac{\sigma}{\sqrt{n}} \\ \hat{\Theta}_2 = \overline{X} + z\frac{\sigma}{\sqrt{n}} \end{cases},$$ hemos construido dos estadísticos $\hat{\Theta}_1$ y $\hat{\Theta}_2$ tales que $P(\hat{\Theta}_1 \le \mu \le \hat{\Theta}_2) = 1-\alpha$,

es decir hemos construido el intervalo de confianza bilateral deseado $[\hat{\Theta}_1, \hat{\Theta}_2]$. Todos los elementos que forman los estadísticos $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son conocidos ya que el número *z* verifica la ecuación anterior, es decir (ver figura):



$P(-z \le Z \le z) = \Phi(z) - \Phi(-z) = 1-\alpha$ donde $\Phi(z)$ es la *Fda* para la v.a. $Z \sim N(0,1)$

Recordando que $\Phi(-z) = 1 - \Phi(z)$, esta ecuación queda:
$\Phi(z) - \Phi(-z) = 2\Phi(z) - 1 = 1-\alpha$, o bien (ver figura anterior),

$\Phi(z) = 1 - \frac{\alpha}{2}$ o de otra forma $P(Z > z) = \frac{\alpha}{2}$.

Al valor de *z* que verifica esta ecuación se lo suele indicar $z_{\frac{\alpha}{2}}$. En consecuencia, el intervalo de confianza bilateral al nivel de significación $1-\alpha$ queda:

$$\left[\hat{\Theta}_1, \hat{\Theta}_2\right] = \left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

En consecuencia:

Si $(X_1, X_2, \ldots, X_n)$ una muestra aleatoria de tamaño *n* de una v.a. *X* donde $X \sim N(\mu, \sigma^2)$, $\sigma^2$ conocido, un intervalo de confianza para $\mu$ de nivel $1-\alpha$ es

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] \qquad (8.1)$$

Ejemplo:
Un ingeniero civil analiza la resistencia a la compresión del concreto. La resistencia está distribuida aproximadamente de manera normal, con varianza 1000 $(psi)^2$. Al tomar una muestra aleatoria de 12 especímenes, se tiene que $\bar{x} = 3250$ psi.

a) Construya un intervalo de confianza del 95% para la resistencia a la compresión promedio.
b) Construya un intervalo de confianza del 99% para la resistencia a la compresión promedio. Compare el ancho de este intervalo de confianza con el ancho encontrado en el inciso a).

Solución:
La v. a. de interés es $X_i$: "resistencia a la compresión del concreto en un espécimen *i*"
Tenemos una muestra de $n = 12$ especímenes.
Asumimos que $X_i \sim N(\mu, \sigma^2)$ para $i = 1,2,3,\ldots,12$ con $\sigma^2 = 1000$
a) Queremos un intervalo de confianza para $\mu$ de nivel 95%. Por lo tanto $\alpha = 0.05$
El intervalo a utilizar es $\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$.
Buscamos en la tabla de la normal estándar el valor de $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Reemplazando:

$$\left[3250 - 1.96 \times \frac{\sqrt{1000}}{\sqrt{12}}, 3250 + 1.96 \times \frac{\sqrt{1000}}{\sqrt{12}}\right] = \left[3232.10773,\ \ 3267.89227\right]$$

b) repetimos lo anterior pero ahora $\alpha = 0.01$

El intervalo a utilizar es $\left[ \overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$.

Buscamos en la tabla de la normal estándar el valor de $z_{\frac{\alpha}{2}} = z_{0.005} = 2.58$

Reemplazando:

$$\left[ 3250 - 2.58 \times \frac{\sqrt{1000}}{\sqrt{12}}, 3250 + 2.58 \times \frac{\sqrt{1000}}{\sqrt{12}} \right] = \left[ 3226.44793, \ \ 3273.55207 \right]$$

La longitud del intervalo encontrado en a) es: 35.78454
La longitud del intervalo encontrado en b) es: 47.10414
Notar que la seguridad de que el verdadero parámetro se encuentre en el intervalo hallado es mayor en el intervalo b) que en el a), pero la longitud del intervalo b) es mayor que la del intervalo a). Al aumentar el nivel de confianza se perdió ***precisión en la estimación***, ya que a menor longitud hay mayor precisión en la estimación.

En general la longitud del intervalo es $L = 2 z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Notar que:

a) si $n$ y $\sigma$ están fijos, a medida que $\alpha$ disminuye tenemos que $z_{\frac{\alpha}{2}}$ aumenta, por lo tanto $L$ aumenta.
b) si $\alpha$ y $\sigma$ están fijos, entonces a medida que $n$ aumenta tenemos que $L$ disminuye.

Podemos plantearnos la siguiente pregunta relacionada con el ejemplo anterior: ¿qué tamaño $n$ de muestra se necesita para que el intervalo tenga nivel de confianza 95% y longitud la mitad de la longitud del intervalo hallado en a)?
Solución: el intervalo hallado en a) tiene longitud 35.78454, y queremos que el nuevo intervalo tenga longitud 17.89227 aproximadamente. Planteamos:

$$L = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq 17.89227 \qquad \Rightarrow \qquad 2 \times 1.96 \times \frac{\sqrt{1000}}{\sqrt{n}} \leq 17.89227$$

Despejando $n$ :

$$\left( 2 \times 1.96 \times \frac{\sqrt{1000}}{17.89227} \right)^2 \leq n \qquad \Rightarrow \qquad n \geq 48$$

O sea, hay que tomar por lo menos 84 especímenes para que el intervalo tenga la longitud pedida.

En general, si queremos hallar $n$ tal que $L = 2 z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq l$ , donde $l$ es un valor dado, entonces despejando $n$

$$n \geq \left( \frac{2 z_{\frac{\alpha}{2}} \sigma}{l} \right)^2$$

Si estimamos puntualmente al parámetro $\mu$ con $\overline{X}$ estamos cometiendo un error en la estimación menor o igual a $\frac{L}{2} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, que se conoce como ***precisión del estimador***

Ejemplo: Se estima que el tiempo de reacción a un estímulo de cierto dispositivo electrónico está distribuido normalmente con desviación estándar de 0.05 segundos. ¿Cuál es el número de mediciones temporales que deberá hacerse para que la confianza de que el error de la estimación de la esperanza no exceda de 0.01 sea del 95%?

Nos piden calcular $n$ tal que $\frac{L}{2} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < 0.01$ con $\alpha = 0.05$.

Por lo tanto $n \geq \left( z_{0.025} \frac{0.05}{0.01} \right)^2$.

Además $z_{0,025}$=1.96. Entonces $n \geq \left( z_{0.975} \frac{0.05}{0.01} \right)^2 = (1.96 \times 5)^2 = 96.04$.

O sea hay que tomar por lo menos 97 mediciones temporales.

Para muestras tomadas de una población normal, o para muestras de tamaño $n \geq 30$, de una población cualquiera, el intervalo de confianza dado anteriormente en (8.1), proporciona buenos resultados.
En el caso de que la población de la que se extrae la muestra no sea normal pero $n \geq 30$, el nivel de confianza del intervalo (8.1) es ***aproximadamente*** $1-\alpha$.
Pero para muestras pequeñas tomadas de poblaciones que no son normales no se puede garantizar que el nivel de confianza sea $1-\alpha$ si se utiliza (8.1).

Ejemplo:
Supongamos que *X* representa la duración de una pieza de equipo y que se probaron 100 de esas piezas dando una duración promedio de 501.2 horas. Se sabe que la desviación estándar poblacional es $\sigma$=4 horas. Se desea tener un intervalo del 95% de confianza para la esperanza poblacional $E(X) = \mu$.
Solución:
En este caso, si bien no conocemos cuál es la distribución de *X* tenemos que el tamaño de la muestra es $n = 100 > 30$ (muestra grande) por lo tanto el intervalo buscado es

$$\left[ \overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

Puesto que 1-$\alpha$=0.95 $\rightarrow \alpha = 1 - 0.95 = 0.05 \rightarrow \frac{\alpha}{2} = 0.025$

De la tabla de la normal estandarizada obtenemos $z_{0,025}$=1.96. Entonces reemplazando:

$$\left[ \overline{X} - 1.96 \frac{4}{\sqrt{100}}, \overline{X} + 1.96 \frac{4}{\sqrt{100}} \right]$$

Para el valor particular $\overline{x}$ =501.2 tenemos el intervalo

$$\left[\bar{x}-1.96\frac{4}{\sqrt{100}},\bar{x}+1.96\frac{4}{\sqrt{n}}\right]=\left[501.2-1.96\frac{4}{10},501.2+1.96\frac{4}{10}\right]=\left[500.4,\ \ 502.0\right].$$

Al establecer que $\left[500.4,\ \ 502.0\right]$ es un intervalo al 95% de confianza de $\mu$ estamos diciendo que la probabilidad de que el intervalo $\left[500.4,\ \ 502.0\right]$ contenga a $\mu$ es 0.95. O, en otras palabras, la probabilidad de que la muestra aleatoria $(X_1,X_2,...,X_n)$ tome valores tales que el intervalo aleatorio $\left[\overline{X}-1.96\frac{4}{\sqrt{100}},\overline{X}+1.96\frac{4}{\sqrt{100}}\right]$ defina un intervalo numérico que contenga al parámetro fijo desconocido $\mu$ es 0.95.

### 9.3 - <u>Intervalo de confianza para la media de una distribución normal, varianza desconocida</u>

Nuevamente como se trata de encontrar un intervalo de confianza para $\mu$ nos basamos en la esperanza muestral $\overline{X}=\frac{1}{n}\sum_{1=1}^{n}X_i$ que sabemos es un buen estimador de $\mu$. Pero ahora no podemos usar como pivote a

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

porque desconocemos $\sigma$ y una condición para ser pivote es que, excepto por el parámetro a estimar ( en este caso $\mu$ ), todos los parámetros que aparecen en él deben ser conocidos. Entonces proponemos como pivote una variable aleatoria definida en forma parecida a *Z* pero reemplazando $\sigma$ por un estimador adecuado.
Ya vimos que la varianza muestral definida

$$S^2=\frac{1}{n-1}\sum_{1=1}^{n}\left(X_i-\overline{X}\right)^2,$$

donde $\overline{X}$ es la esperanza muestral, es un estimador insesgado de la varianza poblacional $V(X)$, es decir, $E\left(S^2\right)=V(X)=\sigma^2\ \ \forall n$. Entonces estimamos $\sigma$ con $S$ y proponemos como pivote a la variable aleatoria

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}.$$

Pero para poder usar a *T* como pivote debemos conocer su distribución.
Se puede probar que la distribución de *T* es una distribución llamada ***Student con parámetro n-1***.

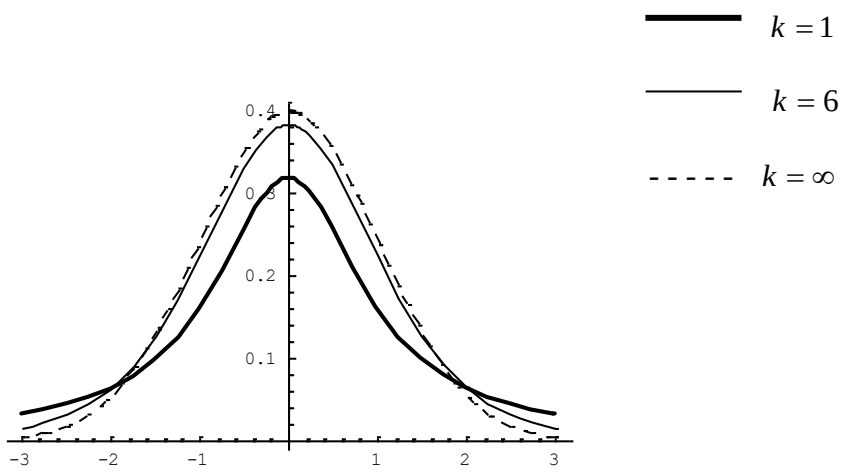
<u>Nota</u>: Una v.a. continua tiene distribución ***Student con k grados de libertad***, si su *f.d.p.* es de la forma

$$f(x)=\frac{\Gamma\left[\frac{(k+1)}{2}\right]}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\frac{1}{\left[\left(\frac{x^2}{k}\right)+1\right]^{\frac{k+1}{2}}} \qquad -\infty < x < \infty$$

Notación: $T \sim t_k$

La gráfica de la *f.d.p.* de la distribución Student tiene forma de campana como la normal, pero tiende a cero más lentamente. Se puede probar que cuando $k \to \infty$ la *fdp* de la Student tiende a la *fdp* de la $N(0,\ 1)$.

En la figura siguiente se grafica *f(x)* para diferentes valores de *k*




Anotaremos $t_{\alpha,k}$ al cuantil de la Student con *k* grados de libertad que deja bajo la *fdp* a derecha un área de $\alpha$, y a su izquierda un área de $1-\alpha$.

Luego, para construir el intervalo de confianza buscado a partir del pivote *T* procedemos como en los casos anteriores:

Comenzamos por plantear la ecuación

$$P(-t \le T \le t) = 1\text{-}\alpha,$$

donde la incógnita es el número real *t*.

Si reemplazamos la v.a. *T* por su expresión, tenemos sucesivamente (multiplicando por $S/\sqrt{n}$ y restando $\overline{X}$):

$$P\left(-t \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t\right) = P\left(-t\frac{S}{\sqrt{n}} \le \overline{X}-\mu \le t\frac{S}{\sqrt{n}}\right) = P\left(-\overline{X}-t\frac{S}{\sqrt{n}} \le -\mu \le -\overline{X}+t\frac{S}{\sqrt{n}}\right) = 1\text{-}\alpha$$

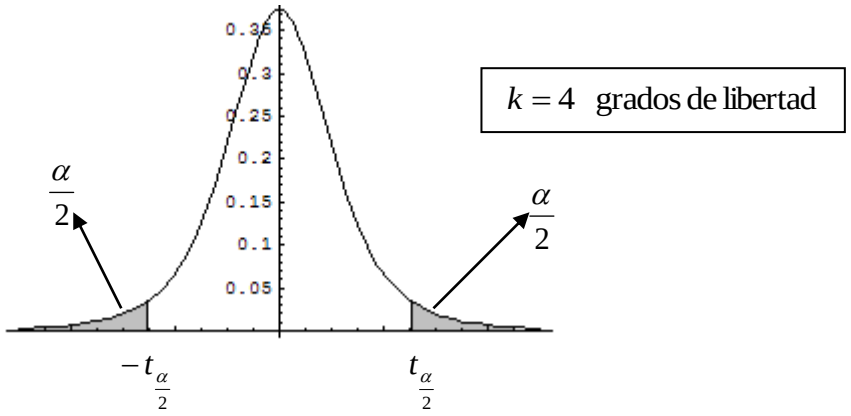
Multiplicando todos los miembros de la desigualdad por -1 (el orden de los miembros se invierte) llegamos a:

$$P\left(\overline{X}-t\frac{S}{\sqrt{n}}\le\mu\le\overline{X}+t\frac{S}{\sqrt{n}}\right)=1\text{-}\alpha$$

Evidentemente, si definimos

$\begin{cases}\hat{\Theta}_1=\overline{X}-t\dfrac{S}{\sqrt{n}}\\ \hat{\Theta}_2=\overline{X}+t\dfrac{S}{\sqrt{n}}\end{cases}$, hemos construido dos estadísticos $\hat{\Theta}_1$ y $\hat{\Theta}_2$ tales que $P\left(\hat{\Theta}_1\le\mu\le\hat{\Theta}_2\right)=1\text{-}\alpha$,

veamos quien es el número *t* que verifica la ecuación, es decir (ver figura):




$P(-t\le T\le t)=F(t)-F(-t)=1\text{-}\alpha$ donde $F(t)$ es la *Fda* para la v.a. $T\sim t_{n-1}$.

Por la simetría de la distribución t de Student se deduce fácilmente de la figura anterior que $F(-t)=1-F(t)$, entonces:

$F(t)-F(-t)=2F(t)-1=1\text{-}\alpha$, o bien (ver figura anterior),

$F(t)=1-\dfrac{\alpha}{2}$.

Al valor de *t* que verifica esta ecuación se lo suele indicar $t_{\frac{\alpha}{2},n-1}$. En consecuencia, el intervalo de confianza bilateral al nivel de significación $1\text{-}\alpha$ queda:

$\left[\overline{X}-t_{\frac{\alpha}{2},n-1}\dfrac{S}{\sqrt{n}},\overline{X}+t_{\frac{\alpha}{2},n-1}\dfrac{S}{\sqrt{n}}\right]$ con $F\left(t_{\frac{\alpha}{2},n-1}\right)=1-\dfrac{\alpha}{2}$.

En consecuencia:

Si $(X_1,X_2,...,X_n)$ una muestra aleatoria de tamaño *n* de una v.a. *X* donde $X\sim N(\mu,\sigma^2)$, $\sigma^2$ desconocido, un intervalo de confianza para $\mu$ de nivel $1-\alpha$ es

$$\left[\overline{X}-t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}},\overline{X}+t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]\qquad(8.2)$$

Ejemplo:
Se hicieron 10 mediciones sobre la resistencia de cierto tipo de alambre que dieron valores $x_1, x_2, \ldots, x_{10}$ tales que $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 10.48$ ohms y $S = \sqrt{\frac{1}{9}\sum_{i=!}^{10}(x_i - \bar{x})^2}$ = 1.36 ohms. Supóngase que $X \sim N(\mu,\sigma^2)$.
Se desea obtener un intervalo de confianza para la esperanza poblacional $\mu$ al 90 %.

Tenemos que $1-\alpha = 0.90 \rightarrow \alpha = 0.1 \rightarrow \alpha/2 = 0.05$
De la Tabla de la *t* de Student tenemos que $t_{0.05,9} = 1.8331$. Entonces el intervalo de confianza buscado es:

$$\left[\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right] = \left[10.48 - 1.8331\frac{1.36}{\sqrt{10}}, 10.48 + 1.8331\frac{1.36}{\sqrt{10}}\right]$$

Esto es: $[9.69,\ 11.27]$.

Si $\sigma^2$ es desconocido y el tamaño de la muestra grande, entonces se puede probar que al reemplazar $\sigma$ por *S*, el estadístico

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0,1) \quad \textbf{\textit{aproximadamente}}$$

y puedo construir el intervalo para $\mu$ como antes:

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$ , pero su nivel es aproximadamente $1-\alpha$

## 9.4 – Intervalo de confianza para la diferencia de dos medias, varianzas conocidas

Supongamos que tenemos dos variables aleatorias ***independientes*** normalmente distribuidas:

$$\begin{cases} X_1 \sim N(\mu_1, \sigma_1^2) \\ X_2 \sim N(\mu_2, \sigma_2^2) \end{cases}$$ y suponemos que las varianzas $\sigma_1^2$ y $\sigma_2^2$ son conocidas.

Sean además
$(X_{11}, X_{12}, \ldots, X_{1n_1})$ una muestra aleatoria de tamaño $n_1$ de $X_1$
$(X_{21}, X_{22}, \ldots, X_{2n_2})$ una muestra aleatoria de tamaño $n_2$ de $X_2$.

Deseamos construir un intervalo al nivel de confianza $1-\alpha$ para la diferencia de esperanzas $\mu_1 - \mu_2$.
Ya sabemos cuál es la distribución del promedio de variables aleatorias normales independientes:

$$\begin{cases} \overline{X}_1 = \dfrac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} \sim N\left(\mu_1, \dfrac{\sigma_1^2}{n_1}\right) \\ \overline{X}_2 = \dfrac{1}{n_2}\sum_{i=1}^{n_2} X_{2i} \sim N\left(\mu_2, \dfrac{\sigma_2^2}{n_2}\right) \end{cases}$$

Consideremos ahora la diferencia $\overline{Y} = \overline{X}_1 - \overline{X}_2$. Si $\overline{X}_1$ y $\overline{X}_2$ tienen distribución normal y son independientes, su diferencia también es normal, con esperanza igual a la diferencia de las esperanzas y la varianza es la suma de las varianzas:

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Por lo tanto

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1),$$ es decir, tiene distribución normal estandarizada.

La v.a. *Z* cumple con toda las condiciones para servir de pivote y construiremos nuestro intervalo en forma análoga a cómo hicimos en los casos anteriores:
Comenzamos por plantear la ecuación

$$P(-z \le Z \le z) = 1-\alpha,$$

donde la incógnita es el número real *z*.

Reemplazamos la v.a. *Z* por su expresión y tenemos sucesivamente (multiplicando por $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ y restando $\overline{X}_1 - \overline{X}_2$):

$$P\left(-z \le \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \le z\right) = P\left(-z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2) \le z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) =$$

$$= P\left(-(\overline{X}_1 - \overline{X}_2) - z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le -(\mu_1 - \mu_2) \le -(\overline{X}_1 - \overline{X}_2) + z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha$$

Multiplicando todos los miembros de la desigualdad por -1 (el orden de los miembros se invierte) llegamos a:

$$P\left(\overline{X}_1 - \overline{X}_2 - z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le (\mu_1 - \mu_2) \le \overline{X}_1 - \overline{X}_2 + z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha$$

Evidentemente, si definimos

$$\begin{cases} \hat{\Theta}_1 = \overline{X}_1 - \overline{X}_2 - z\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} \\ \hat{\Theta}_2 = \overline{X}_1 - \overline{X}_2 + z\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} \end{cases},$$

habremos construido dos estadísticos $\hat{\Theta}_1$ y $\hat{\Theta}_2$ tales que $P\left(\hat{\Theta}_1 \le (\mu_1 - \mu_2) \le \hat{\Theta}_2\right) = 1\text{-}\alpha$, es decir habremos construido el intervalo de confianza bilateral deseado $\left[\hat{A}_1, \hat{A}_2\right]$. Todos los elementos que forman los estadísticos $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son conocidos ya que el número *z* verifica la ecuación anterior, es decir:

$P(-z \le Z \le z) = \Phi(z) - \Phi(-z) = 1\text{-}\alpha$ donde $\Phi(z)$ es la *Fda* para la v.a. $Z \sim N(0,1)$

o bien, según vimos:

$\Phi(z) = 1 - \dfrac{\alpha}{2}$ que anotamos $z_{\frac{\alpha}{2}}$

En consecuencia, el intervalo de confianza bilateral al nivel de significación $1\text{-}\alpha$ queda:

$$\left[\overline{X}_1 - \overline{X}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\quad \overline{X}_1 - \overline{X}_2 + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Por lo tanto

Si $X_1$ y $X_2$ son dos variables aleatorias ***independientes*** normalmente distribuidas: $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ y suponemos que las varianzas $\sigma_1^2$ y $\sigma_2^2$ son conocidas. Un intervalo de confianza para la diferencia $\mu_1 - \mu_2$ de nivel $1 - \alpha$ es

$$\left[\overline{X}_1 - \overline{X}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\quad \overline{X}_1 - \overline{X}_2 + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] \qquad (8.3)$$

Ejemplo:

Se utilizan dos máquinas para llenar botellas de plástico con detergente para máquinas lavaplatos. Se sabe que las desviaciones estándar de volumen de llenado son $\sigma_1 = 0.10$ onzas de líquido y $\sigma_2 = 0.15$ onzas de líquido para las dos máquinas respectivamente. Se toman dos muestras aleatorias, $n_1 = 12$ botellas de la máquina 1 y $n_2 = 10$ botellas de la máquina 2. Los volúmenes promedio de llenado son $\bar{x}_1 = 30.87$ onzas de líquido y $\bar{x}_2 = 30.68$ onzas de líquido.

Asumiendo que ambas muestras provienen de distribuciones normales

Construya un intervalo de confianza de nivel 90% para la diferencia entre las medias del volumen de llenado.

Solución:

Como $1 - \alpha = 0.90$ entonces $\alpha = 0.10$

Por lo tanto $z_{\frac{\alpha}{2}} = z_{0.05} = 1.65$

El intervalo será $\left[(30.87-30.68)-1.65\sqrt{\frac{0.10^2}{12}+\frac{0.15^2}{10}};(30.87-30.68)+1.65\sqrt{\frac{0.10^2}{12}+\frac{0.15^2}{10}}\right]$

O sea $\left[0.09837;\ \ 0.281620\right]$

Si se conocen las desviaciones estándar y los tamaños de las muestras son iguales (es decir $n_1 = n_2 = n$), entonces puede determinarse el tamaño requerido de la muestra de manera tal que la longitud del intervalo sea menor que *l*

$$L = 2z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^{\ 2}}{n}+\frac{\sigma_2^{\ 2}}{n}} \le l \qquad \Rightarrow \qquad n \ge \left(\frac{2z_{\frac{\alpha}{2}}}{l}\right)^2\left(\sigma_1^{\ 2}+\sigma_2^{\ 2}\right)$$

Si $\sigma_1$ y $\sigma_2$ son ***desconocidos***, $n_1 \ge 30$ y $n_2 \ge 30$, entonces se puede probar que al reemplazar $\sigma_1$ por $S_1$ y $\sigma_2$ por $S_2$, el estadístico

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^{\ 2}}{n_1}+\frac{S_2^{\ 2}}{n_2}}} \approx N(0,1) \quad \textit{aproximadamente}$$

y puedo construir el intervalo para $\mu_1 - \mu_2$ como antes:

$$\left[\overline{X}_1 - \overline{X}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\ ,\ \ \overline{X}_1 - \overline{X}_2 + z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right], \qquad (8.4)$$

pero su nivel es aproximadamente $1-\alpha$

Para muestras tomadas de dos poblaciones normales, o para muestras de tamaño $n_1 \ge 30$ y $n_2 \ge 30$, de dos poblaciones cualesquiera, el intervalo de confianza dado anteriormente en (8.3), proporciona buenos resultados.
En el caso de que la población de la que se extrae la muestra no sea normal pero $n_1 \ge 30$ y $n_2 \ge 30$, el nivel de confianza del intervalo (8.3) es ***aproximadamente*** $1-\alpha$ .

Ejemplo:
De una muestra de 150 lámparas del fabricante A se obtuvo una vida media de 1400 hs y una desviación típica de 120 hs. Mientras que de una muestra de 100 lámparas del fabricante B se obtuvo una vida media de 1200 hs. y una desviación típica de 80 hs.
Halla los límites de confianza del 95% para la diferencia las vidas medias de las poblaciones A y B.

Solución:

Sean las variables aleatorias:
$X_1$ : "duración en horas de una lámpara del fabricante A"
$X_2$ : "duración en horas de una lámpara del fabricante B"
No se dice cuál es la distribución de estas variables, pero como $n_1 = 150$ y $n_2 = 100$ podemos usar el intervalo dado en (8.4)

Tenemos que $\bar{x}_1 = 1400$, $\bar{x}_2 = 1200$, $s_1 = 120$ y $s_2 = 80$.
Además $1-\alpha = 0.95 \rightarrow z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Entonces el intervalo es

$$\left[1400-1200-1.96\sqrt{\frac{120^2}{150}+\frac{80^2}{100}}\ ;\ 1400-1200+1.96\sqrt{\frac{120^2}{150}+\frac{80^2}{100}}\right] = \left[175.2077;\ \ 224.7922\right]$$

Observación: como este intervalo no contiene al cero, podemos inferir que hay diferencia entre las medias con probabilidad 0.95, es más, podemos inferir que la media del tiempo de duración de las lámparas del fabricante A es mayor que la media del tiempo de duración de las lámparas del fabricante B con probabilidad 0.95 .

## 9.5 – Intervalo de confianza para la diferencia de dos medias, varianzas desconocidas

Nuevamente supongamos que tenemos dos variables aleatorias ***independientes*** normalmente distribuidas:

$\begin{cases} X_1 \sim N(\mu_1, \sigma_1^2) \\ X_2 \sim N(\mu_2, \sigma_2^2) \end{cases}$ y suponemos que las varianzas $\sigma_1^2$ y $\sigma_2^2$ son ***desconocidas*** .

Sean además
$(X_{11}, X_{12}, ..., X_{1n_1})$ una muestra aleatoria de tamaño $n_1$ de $X_1$
$(X_{21}, X_{22}, ..., X_{2n_2})$ una muestra aleatoria de tamaño $n_2$ de $X_2$ .
Pero ahora $n_1$ ***o*** $n_2$ ***no son mayores que 30***
Supongamos que es razonable suponer que las varianzas desconocidas son iguales, es decir
$\sigma_1 = \sigma_2 = \sigma$
Deseamos construir un intervalo al nivel de confianza $1-\alpha$ para la diferencia de esperanzas $\mu_1 - \mu_2$

Sean $\bar{X}_1$ y $\bar{X}_2$ las medias muestrales y $S_1^2$ y $S_2^2$ las varianzas muestrales. Como $S_1^2$ y $S_2^2$ son los estimadores de la varianza común $\sigma^2$, entonces construimos un ***estimador combinado*** de $\sigma^2$. Este estimador es

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

Se puede comprobar que es un estimador insesgado de $\sigma^2$.
Se puede probar que el estadístico

$T = \dfrac{\vec{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ tiene distribución Student con $n_1 + n_2 - 2$ grados de libertad

Por lo tanto se plantea la ecuación

$$P\left(-t_{\frac{\alpha}{2},n_1+n_2-2} \le T \le t_{\frac{\alpha}{2},n_1+n_2-2}\right) = 1-\alpha$$

o

$$P\left(-t_{\frac{\alpha}{2},n_1+n_2-2} \le \frac{\vec{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \le t_{\frac{\alpha}{2},n_1+n_2-2}\right) = 1-\alpha$$

Despejamos $\mu_1 - \mu_2$ y queda la expresión

$$P\left(\overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2},\,n_1+n_2-2}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + t_{\frac{\alpha}{2},\,n_1+n_2-2}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) = 1-\alpha$$

Entonces

Si $X_1$ y $X_2$ son dos variables aleatorias ***independientes*** normalmente distribuidas:
$X_1 \sim \mathrm{N}(\mu_1, {\sigma_1}^2)$, $X_2 \sim \mathrm{N}(\mu_2, {\sigma_2}^2)$ y suponemos que las varianzas $\sigma_1^2$ y $\sigma_2^2$ son ***desconocidas e iguales***, es decir $\sigma_1 = \sigma_2 = \sigma$
Un intervalo de confianza para la diferencia $\mu_1 - \mu_2$ de nivel $1-\alpha$ es

$$\overline{X}_1 - \overline{X}_2 - t_{\alpha/2,n_1+n_2-2}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \;;\; \overline{X}_1 - \overline{X}_2 + t_{\alpha/2,n_1+n_2-2}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (8.5)$$

Ejemplo:
Se piensa que la concentración del ingrediente activo de un detergente líquido para ropa, es afectada por el tipo de catalizador utilizado en el proceso de fabricación. Se realizan 10 observaciones con cada catalizador, y se obtienen los datos siguientes:
Catalizador 1: 57.9, 66.2, 65.4, 65.4, 65.2, 62.6, 67.6, 63.7, 67.2, 71.0
Catalizador 2: 66.4, 71.7, 70.3, 69.3, 64.8, 69.6, 68.6, 69.4, 65.3, 68.8
a) Encuentre un intervalo de confianza del 95% para la diferencia entre las medias de las concentraciones activas para los dos catalizadores. Asumir que ambas muestras fueron extraídas de poblaciones normales con varianzas iguales.
b) ¿Existe alguna evidencia que indique que las concentraciones activas medias dependen del catalizador utilizado?
Solución:
Sean las variables aleatorias

$X_1$ : " concentración del ingrediente activo con catalizador 1"

$X_2$ : " concentración del ingrediente activo con catalizador 2"

Asumimos que ambas variables tienen distribución normal con varianzas iguales

Estamos e3n las condiciones para usar (8.5)

Tenemos que $\bar{x}_1 = 65.22$ , $\bar{x}_2 = 68.42$ , $s_1 = 3.444$, $s_2 = 2.224$ , $n_1 = n_2 = 10$

Calculamos $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \dfrac{9\times 3.444^2 + 9\times 2.224^2}{10+10-2} = 8.4036$

Por lo tanto $S_p = \sqrt{8.4036} = 2.89890$

Buscamos en la tabla de la Student $t_{\frac{\alpha}{2},n_1+n_2-2} = t_{0.025,18} = 2.101$

Entonces el intervalo es

$$\left[65.22 - 68.42 - 2.101\times 2.89890\sqrt{\frac{1}{10}+\frac{1}{10}};65.22 - 68.42 + 2.101\times 2.89890\sqrt{\frac{1}{10}+\frac{1}{10}}\right] =$$

$$= \left[-5.9237;-0.476301\right]$$

b) Existe alguna evidencia que indique que las concentraciones activas medias dependen del catalizador utilizado, pues el 0 no pertenece al intervalo.

En muchas ocasiones ***no es razonable suponer que las varianzas son iguales***. Si no podemos garantizar que las varianzas son iguales, para construir un intervalo de confianza de nivel $1-\alpha$ para $\mu_1 - \mu_2$ utilizamos es estadístico

$$T^* = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

Se puede probar que $T^*$ tiene ***aproximadamente*** una distribución Student con $\nu$ grados de libertad donde

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1-1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$ si $\nu$ no es entero, ***se toma el entero más próximo a*** $\nu$

Por lo tanto planteamos la ecuación

$$P\left(-t_{\frac{\alpha}{2},\nu} \le T^* \le t_{\frac{\alpha}{2},\nu}\right) = 1-\alpha$$

Y despejando $\mu_1 - \mu_2$ el intervalo es

$$\left[\bar{X}_1 - \bar{X}_2 - t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\quad \bar{X}_1 - \bar{X}_2 + t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right]$$

Entonces

Si $X_1$ y $X_2$ son dos variables aleatorias ***independientes*** normalmente distribuidas: $X_1 \sim N(\mu_1, {\sigma_1}^2)$, $X_2 \sim N(\mu_2, {\sigma_2}^2)$ y suponemos que las varianzas $\sigma_1^2$ y $\sigma_2^2$ son ***desconocidas*** y ***distintas***

Un intervalo de confianza para la diferencia $\mu_1 - \mu_2$ de nivel ***aproximadamente*** $1-\alpha$ es

$$\left[\bar{X}_1 - \bar{X}_2 - t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\quad \bar{X}_1 - \bar{X}_2 + t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right] \qquad (8.6)$$

Donde

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

Ejemplo:

Una muestra de 6 soldaduras de un tipo tenía promedio de prueba final de resistencia de 83.2 ksi y desviación estándar de 5.2. Y una muestra de 10 soldaduras de otro tipo tenía resistencia promedio de 71.3 ksi y desviación estándar de 3.1. supongamos que ambos conjuntos de soldaduras son muestras aleatorias de poblaciones normales. Se desea encontrar un intervalo de confianza de 95% para la diferencia entre las medias de las resistencias de los dos tipos de soldaduras.

Solución:

Ambos tamaños muestrales son pequeños y las muestras provienen de poblaciones normales. No podemos asumir igualdad de varianzas. Entonces aplicamos (8.6)

Tenemos que $\bar{x}_1 = 83.2$, $\bar{x}_2 = 71.3$, $s_1 = 5.2$, $s_2 = 3.1$, $n_1 = 6$; $n_2 = 10$

Como $1-\alpha = 0.95$ entonces $\frac{\alpha}{2} = 0.025$

Además $\nu = \dfrac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^1/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}} = \dfrac{\left(\frac{5.2^2}{6} + \frac{3.1^2}{10}\right)^2}{\frac{\left(5.2/6\right)^2}{5} + \frac{\left(3.1/10\right)^2}{9}} = 7.18 \approx 7$

Entonces buscamos en la tabla de la Student $t_{0.025,7} = 2.365$

Por lo tanto el intervalo es

$$\left[ \bar{X}_1 - \bar{X}_2 - t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\quad \bar{X}_1 - \bar{X}_2 + t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \right] =$$

$$= \left[ 83.2 - 71.3 - 2.365\sqrt{\frac{5.2^2}{6}+\frac{3.1^2}{10}};\ 83.2 - 71.3 + 2.365\sqrt{\frac{5.2^2}{6}+\frac{3.1^2}{10}} \right] = \left[ 6.37,\ 17.43 \right]$$

## 9.6 – Intervalo de confianza para $\mu_1 - \mu_2$ para datos pareados

Hasta ahora se obtuvieron intervalos de confianza para la diferencia de medias donde se tomaban dos muestras aleatorias independientes de dos poblaciones de interés. En ese caso se tomaban $n_1$ observaciones de una población y $n_2$ observaciones de la otra población.

En muchas situaciones experimentales, existen solo *n* unidades experimentales diferentes y los datos están ***recopilados por pares***, esto es cada unidad experimental está formada por ***dos observaciones***.

Por ejemplo, supongamos que se mide el tiempo en segundos que un individuo tarda en hacer una maniobra de estacionamiento con dos automóviles diferentes en cuanto al tamaño de la llanta y la relación de vueltas del volante. Notar que cada individuo es la unidad experimental y de esa unidad experimental se toman dos observaciones que ***no serán independientes***. Se desea obtener un intervalo de confianza para la diferencia entre el tiempo medio para estacionar los dos automóviles.

En general, supongamos que tenemos los siguientes datos $(X_{11}, X_{21});(X_{12}, X_{22});\ldots;(X_{1n_1}, X_{2n})$.

Las variables aleatorias $X_1$ y $X_2$ tienen medias $\mu_1$ y $\mu_2$ respectivamente.

Sea $D_j = X_{1j} - X_{2j}$ con $j = 1,2,\ldots,n$.

Entonces

$$E(D_j) = E(X_{1j} - X_{2j}) = E(X_{1j}) - E(X_{2j}) = \mu_1 - \mu_2$$

y

$$V(D_j) = V(X_{1j} - X_{2j}) = V(X_{1j}) + V(X_{2j}) - 2Cov(X_{1j}, X_{2j}) = \sigma_1^2 + \sigma_2^2 - 2Cov(X_1, X_2)$$

Estimamos $E(D_j) = \mu_1 - \mu_2$ con $\bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j = \frac{1}{n}\sum_{j=1}^{n}(X_{1j} - X_{2j}) = \bar{X}_1 - \bar{X}_2$

En lugar de tratar de estimar la covarianza, estimamos la $V(D_j)$ con $S_D^2 = \frac{1}{n-1}\sum_{j=1}^{n}(D_j - \bar{D})^2$

Anotamos $\mu_D = \mu_1 - \mu_2$ y $\sigma^2{}_D = V(D_j)$

Asumimos que $D_j \sim N(\mu_D, \sigma_D{}^2)$ con $j = 1,2,\ldots,n$

Las variables aleatorias en pares diferentes son independientes, no lo son dentro de un mismo par.

Para construir el intervalo de confianza notar que

$$T = \frac{\overline{D} - \mu_D}{S_D / \sqrt{n}} \sim t_{n-1}$$

entonces al plantear la ecuación $P(-t \le T \le t) = 1\text{-}\alpha$, deducimos que $t = t_{\frac{\alpha}{2},n-1}$

Por lo tanto el intervalo de confianza para $\mu_D = \mu_1 - \mu_2$ de nivel $1-\alpha$ se obtendrá al sustituir $T$ en la ecuación anterior y despejar $\mu_D = \mu_1 - \mu_2$

El intervalo resultante es

$$\left[ \overline{D} - t_{\frac{\alpha}{2},n-1} \frac{S_D}{\sqrt{n}};\ \overline{D} + t_{\frac{\alpha}{2},n-1} \frac{S_D}{\sqrt{n}} \right]$$

Entonces

Cuando las observaciones se dan de a pares $(X_{11}, X_{21});(X_{12}, X_{22});...;(X_{1n_1}, X_{2n})$, y las diferencias $D_j = X_{1j} - X_{2j}$ son tales que $D_j \sim \mathrm{N}(\mu_D, {\sigma_D}^2)$ para $j = 1,2,...,n$, un intervalo de confianza de nivel $1-\alpha$ para $\mu_D = \mu_1 - \mu_2$ es

$$\left[ \overline{D} - t_{\frac{\alpha}{2},n-1} \frac{S_D}{\sqrt{n}};\ \overline{D} + t_{\frac{\alpha}{2},n-1} \frac{S_D}{\sqrt{n}} \right] \qquad (8.7)$$

Ejemplo:

Consideramos el ejemplo planteado al comienzo. Deseamos un intervalo de nivel 0.90

Sean las variables aleatorias

$X_{1j}$: "tiempo en segundos que tarda el individuo *j* en estacionar automóvil 1" con $j = 1,2,...,n$

$X_{2j}$: "tiempo en segundos que tarda el individuo *j* en estacionar automóvil 2" con $j = 1,2,...,n$

Medimos estas variables de manera que tenemos las siguientes observaciones

| | ***Automóvil 1*** | ***Automóvil 2*** | ***diferencia*** |
|---|---|---|---|
| ***sujeto*** | ***(observación*** $x_{1j}$***)*** | ***(observación*** $x_{2j}$***)*** | $D_j$ |
| 1 | 37.0 | 17.8 | 19.2 |
| 2 | 25.8 | 20.2 | 5.6 |
| 3 | 16.2 | 16.8 | -0.6 |
| 4 | 24.2 | 41.4 | -17.2 |
| 5 | 22.0 | 21.4 | 0.6 |
| 6 | 33.4 | 38.4 | -5.0 |
| 7 | 23.8 | 16.8 | 7.0 |
| 8 | 58.2 | 32.2 | 26.0 |
| 9 | 33.6 | 27.8 | 5.8 |
| 10 | 24.4 | 23.2 | 1.2 |
| 11 | 23.4 | 29.6 | -6.2 |
| 12 | 21.2 | 20.6 | 0.6 |
| 13 | 36.2 | 32.2 | 4.0 |
| 14 | 29.8 | 53.8 | -24.0 |

A partir de la columna de diferencias observadas se calcula $\overline{D} = 1.21$ y $S_D = 12.68$

Además $t_{\frac{\alpha}{2},n-1} = t_{0.05,13} = 1.771$, entonces el intervalo para la diferencia $\mu_D = \mu_1 - \mu_2$ de nivel 0.90 es

$$\left[1.21 - 1.771 \times \frac{12.68}{\sqrt{14}};\ 1.21 + 1.771 \times \frac{12.68}{\sqrt{14}}\right] = \left[-4.79;\ 7.21\right]$$

## 9.7 – Intervalo de confianza para la varianza de una distribución normal

Supongamos que se quiere hallar un intervalo de confianza para la varianza $\sigma^2$ de una distribución normal.
Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de una v.a. *X*, donde $X \sim N(\mu, \sigma^2)$.

Tomamos como estimador puntual de $\sigma^2$ a $S^2 = \frac{1}{n-1}\sum_{1=1}^{n}\left(X_i - \overline{X}\right)^2$

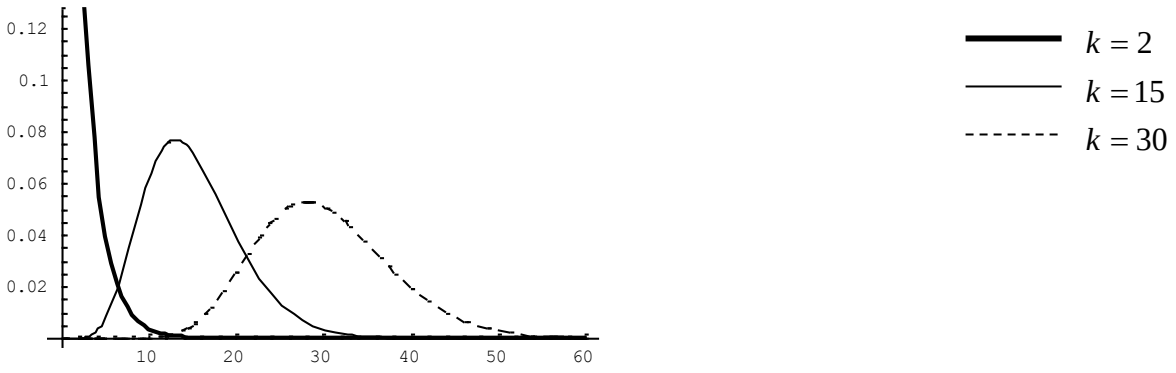
Luego a partir de este estimador puntual construimos el estadístico $X = \frac{(n-1)S^2}{\sigma^2}$

Este estadístico contiene al parámetro desconocido a estimar $\sigma^2$ y tiene una distribución conocida, se puede probar que *X* tiene una distribución llamada ***ji-cuadrado con n-1 grados de libertad***

Observación: Si X es una v.a. continua se dice que tiene distribución ***ji-cuadrado con k grados de libertad*** si su f.d.p. es

$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{(k/2)-1} e^{-x/2} \qquad x > 0$$

Notación: X~ $\chi_k^2$
La distribución ji-cuadrdo es asimétrica. En la figura siguiente se grafica la densidad para diferentes valores de *k*



Anotaremos $\chi^2_{\alpha,k}$ al cuantil de la ji-cuadrado con *k* grados de libertad que deja bajo la *fdp* a derecha un área de $\alpha$, y a su izquierda un área de $1-\alpha$.

Propiedades:

1- Se puede probar que si $X_1, X_2, ..., X_n$ son variables aleatorias independientes con distribución $N(0,1)$ entonces $Z = X_1^2 + X_2^2 + ... + X_n^2$ tiene distribución ji-cuadrado con *n* grados de libertad.

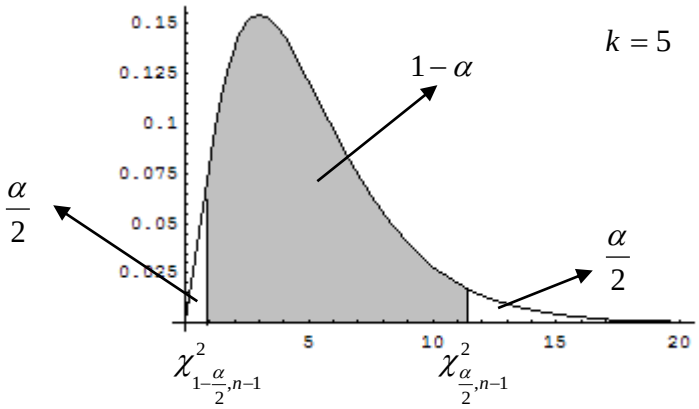
2- Si $X_1, X_2, ..., X_n$ son variables aleatorias independientes tal que $X_i$ tiene distribución ji-cuadrado con $k_i$ grados de libertad, entonces $Z = X_1 + X_2 + ... + X_n$ tiene distribución ji-cuadrado con *k* grados de libertad donde $k = k_1 + k_2 + ... + k_n$

3- Si $X \sim \chi^2_k$ entonces ***para k grande*** $\sqrt{2X} \sim N\left(\sqrt{2k-1},\ 1\right)$ aproximadamente.

Para desarrollar el intervalo de confianza planteamos hallar dos números *a* y *b* tales que

$$P(a \le X \le b) = 1-\alpha \qquad \text{es decir} \qquad P\left(a \le \frac{(n-1)S^2}{\sigma^2} \le b\right) = 1-\alpha$$

Se puede probar que la mejor elección de *a* y *b* es: $a = \chi^2_{1-\frac{\alpha}{2},n-1}$ y $b = \chi^2_{\frac{\alpha}{2},n-1}$




Por lo tanto

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

y despejando $\sigma^2$ se llega a

$$P\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) = 1-\alpha$$

Entonces

Si $(X_1, X_2, ..., X_n)$ es una muestra aleatoria de una v.a. *X*, donde $X \sim N(\mu, \sigma^2)$, un intervalo de confianza para $\sigma^2$ de nivel $1-\alpha$ es

$$\left( \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}};\ \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right) \qquad (8.8)$$

Observación: un intervalo de confianza para $\sigma$ de nivel $1-\alpha$, es $\left( \sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}};\ \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}} \right)$

Ejemplo:
Un fabricante de detergente líquido está interesado en la uniformidad de la máquina utilizada para llenar las botellas. De manera específica, es deseable que la desviación estándar $\sigma$ del proceso de llenado sea menor que 0.15 onzas de líquido; de otro modo, existe un porcentaje mayor del deseable de botellas con un contenido menor de detergente. Supongamos que la distribución del volumen de llenado es aproximadamente normal. Al tomar una muestra aleatoria de 20 botellas, se obtiene una varianza muestral $S^2 = 0.0153$. Hallar un intervalo de confianza de nivel 0.95 para la verdadera varianza del volumen de llenado.

Solución:
La v.a. de interés es *X*: " volumen de llenado de una botella"
Se asume que $X \sim N(\mu, \sigma^2)$ con $\sigma$ desconocido.
Estamos en las condiciones para aplicar (8.8)

Tenemos que $1-\alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \chi^2_{1-\frac{\alpha}{2},n-1} = \chi^2_{0.975,19} = 8.91$ y $\chi^2_{\frac{\alpha}{2},n-1} = \chi^2_{0.025,19} = 32.85$

Además $S^2 = 0.0153$

Por lo tanto el intervalo es

$$\left( \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}};\ \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right) = \left( \frac{(20-1)\times 0.0153}{32.85};\ \frac{(20-1)\times 0.0153}{8.91} \right) = (0.00884;\ 0.0326)$$

Y un intervalo para $\sigma$ es $\left(\sqrt{0.00884};\ \sqrt{0.0326}\right) = (0.09;\ 0.1805)$

Por lo tanto con un nivel de 0.95 los datos ***no apoyan la afirmación que*** $\sigma < 0.15$

## 9.8 – Intervalo de confianza para el cociente de varianzas de dos distribuciones normales

Supongamos que se tienen dos poblaciones normales e independientes con varianzas desconocidas $\sigma_1^2$ y $\sigma_2^2$ respectivamente. Se desea encontrar un intervalo de nivel $1-\alpha$ para el cociente de las dos varianzas $\frac{\sigma_1^2}{\sigma_2^2}$.

Se toma una muestra aleatoria de tamaño $n_1$ de una de las poblaciones y una muestra de tamaño $n_2$ de la otra población. Sean $S_1^2$ y $S_2^2$ las dos varianzas muestrales.
Consideramos el estadístico

$$F = \frac{S_2^2 / \sigma_2^{\,2}}{S_1^2 / \sigma_1^{\,2}}$$

Notar que *F* contiene al parámetro de interés $\frac{\sigma_1^2}{\sigma_2^2}$, pues $F = \frac{S_2^2 \times \sigma_1^2}{S_1^2 \times \sigma_2^2}$

Se puede probar que *F* tiene una distribución llamada Fisher con $n_2 - 1$ y $n_1 - 1$ grados de libertad.

Observación:
Sea *X* una variable aleatoria continua, se dice que tiene distribución Fisher con *u* grados de libertad en el numerador y *v* grados de libertad en el denominador si su *fdp* es de la forma

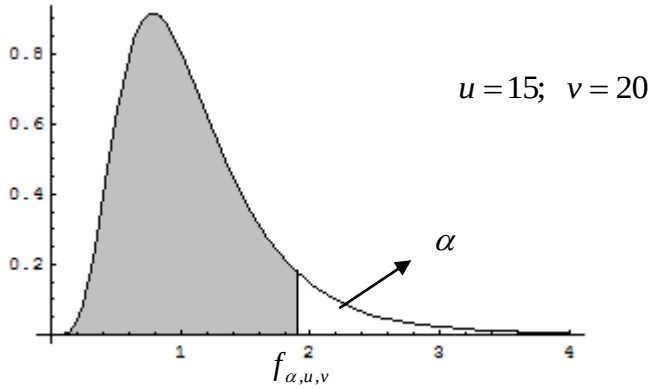
$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{\frac{u}{2}} x^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{\frac{u+v}{2}}} \qquad 0 < x < \infty$$

En particular si *W* e *Y* son variables aleatorias independientes ji-cuadrado con *u* y *v* grados de libertad respectivamente, entonces el cociente

$$F = \frac{W/u}{Y/v}$$

Tiene una distribución Fisher con *u* grados de libertad en el numerador y *v* grados de libertad en el denominador.
Notación: $F \sim F_{u,v}$

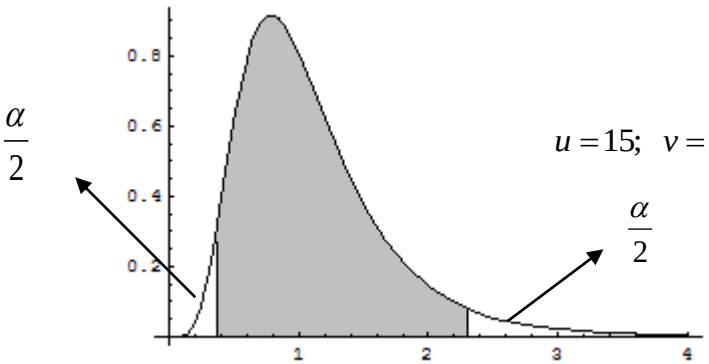
La gráfica de una distribución Fisher es similar a la de una ji-cuadrado, es asimétrica. Anotamos $f_{\alpha,u,v}$ al cuantil que deja a su derecha un área de $\alpha$ bajo la curva de densidad.



Existe la siguiente relación entre los cuantiles de una $F_{u,v}$ y de una $F_{v,u}$

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,v,u}}$$

Planteamos la siguiente ecuación $P(a \le F \le b) = 1-\alpha$ y se pede probar que la mejor elección de $a$ y $b$ es : $a = f_{1-\frac{\alpha}{2},n_2-1,n_1-1}$ y $b = f_{\frac{\alpha}{2},n_2-1,n_1-1}$




$\frac{\alpha}{2}$

$u = 15;\ \ v = 20$

$\frac{\alpha}{2}$

$f_{1-\frac{\alpha}{2},n_2-1,n_1-1}$ $f_{\frac{\alpha}{2},\ n_2-1,n_1-1}$

Entonces $P\left( f_{1-\frac{\alpha}{2},n_2-1,n_1-1} \le \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \le f_{\frac{\alpha}{2},n_2-1,n_1-1} \right) = 1-\alpha$

Despejando el cociente $\frac{\sigma_1^2}{\sigma_2^2}$ queda :

$$P\left( \frac{S_1^2}{S_2^2} f_{1-\frac{\alpha}{2},n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} f_{\frac{\alpha}{2},n_2-1,n_1-1} \right) = 1-\alpha$$

Por lo tanto

Si se tienen dos poblaciones normales e independientes con varianzas desconocidas $\sigma_1^2$ y $\sigma_2^2$ respectivamente, entonces un intervalo de nivel $1-\alpha$ para el cociente de las dos varianzas $\frac{\sigma_1^2}{\sigma_2^2}$ es

$$\left[ \frac{S_1^2}{S_2^2} f_{1-\frac{\alpha}{2},n_2-1,n_1-1};\ \frac{S_1^2}{S_2^2} f_{\frac{\alpha}{2},n_2-1,n_1-1} \right] \quad (8.9)$$

Ejemplo:
Una compañía fabrica propulsores para uso en motores de turbina. Una de las operaciones consiste en esmerilar el terminado de una superficie particular con una aleación de titanio. Pueden emplear-

se dos procesos de esmerilado, y ambos pueden producir partes que tienen la misma rugosidad superficial promedio. Interesaría seleccionar el proceso que tenga la menor variabilidad en la rugosidad de la superficie. Para esto se toma una muestra de 12 partes del primer proceso, la cual tiene una desviación estándar muestral $S_1 = 5.1$ micropulgadas, y una muestra aleatoria de 15 partes del segundo proceso, la cual tiene una desviación estándar muestral $S_2 = 4.7$ micropulgadas. Se desea encontrar un intervalo de confianza de nivel 90% para el cociente de las dos varianzas.
Suponer que los dos procesos son independientes y que la rugosidad de la superficie está distribuida de manera normal.

Solución:
Estamos en las condiciones para aplicar (8.9)
Buscamos en la tabla de la Fisher $f_{1-\frac{\alpha}{2},n_2-1,n_1-1} = f_{0.95,14,11} = \dfrac{1}{f_{0.05,11,14}} = \dfrac{1}{2.58} = 0.39$

y $f_{\frac{\alpha}{2},n_2-1,n_1-1} = f_{0.05,14,11} = 2.74$

Entonces el intervalo es

$$\left[\frac{5.1^2}{4.7^2}0.39;\ \frac{5.1^2}{4.7^2}2.74\right] = [0.46;\ 3.23]$$

Como este intervalo incluye al 1, no podemos afirmar que las desviaciones estándar de los dos procesos sean diferentes con una confianza de 90%.

## 9.9 – Intervalo de confianza para una proporción

Sea una población de tamaño *N* (eventualmente puede ser infinito) de cuyos individuos nos interesa cierta propiedad *A*. Supongamos que la probabilidad de que un individuo de la población verifique *A* es $p = P(A)$ .El significado del parámetro *p* es, en consecuencia, el de proporción de individuos de la población que verifican la propiedad *A*. Podemos definir una variable aleatoria $X_i$ que mide a los individuos de la población la ocurrencia o no de la propiedad $A$.
La variable aleatoria tendrá la distribución:

$$p(x) = \begin{cases} p(1) = P(X_i = 1) = p \\ \\ p(0) = P(X_i = 0) = 1 - p, \end{cases}$$

es decir, $X_i$ es una v.a. que toma sólo dos valores: 1 (si el individuo verifica *A*) con probabilidad *p* y 0 (cuando no verifica *A*) con probabilidad 1-*p*. Esto es equivalente a decir que $X_i$ tiene una distribución binomial con parámetros 1 y *p*: $X_i \sim B(1,p)$.

Supongamos que consideramos una muestra aleatoria $(X_1, X_2 ..., X_n)$ de tamaño $n$. Si formamos el estadístico $X = X_1 + X_2 + ... + X_n$ , es evidente que esta v.a. mide el número de individuos de la muestra de tamaño *n* que verifican la propiedad *A*. Por lo tanto por su significado *X* es una v.a. cuya distribución es binomial con parámetros *n* y *p*: $X \sim B(n,p)$. De acuerdo con esto, la variable

aleatoria $\hat{P}$ definida: $\hat{P} = \frac{X}{n}$ representa la proporción de individuos de la muestra que verifican la propiedad *A*.

Observemos que siendo $X_i \sim B(1,p)$ es $E(X_i) = p$. Y, dado que $X \sim B(n,p)$, también es $E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$, es decir $\hat{P}$ es un estimador insesgado de $p$. Esto es de esperar pues $\hat{P} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Pero además, es fácil ver que $\hat{P}$ es estimador consistente de $p$. En efecto, tenemos que $E(\hat{P}) = p$, pero también es

$$V(\hat{P}) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}.$$

Deseamos construir un intervalo de confianza de *p*. Es razonable basarnos en el estimador insesgado $\hat{P}$. Consideramos como pivote a la variable aleatoria

$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$ cuya distribución es, para *n* suficientemente grande, aproximadamente $N(0,1)$. En efecto:

Siendo $\hat{P} = \frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_n}{n}$, es $E(\hat{P}) = \sum_{i=1}^{n} E\left(\frac{X_i}{n}\right) = p$ y $V(\hat{P}) = \sum_{i=1}^{n} V\left(\frac{X_i}{n}\right) = \frac{p(1-p)}{n}$

Por lo tanto:

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

El pivote puede ponerse en una forma más conveniente si tenemos en cuenta que, según vimos recién, $\hat{P}$ es estimador consistente de *p* y en consecuencia, en el denominador reemplazamos el parámetro desconocido *p* por su estimador $\hat{P}$, y se puede probar que :

$Z = \frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \approx N(0,1)$. aproximadamente si *n* es grande

Partiendo de este pivote podemos seguir los mismos pasos de los casos anteriores para llegar al siguiente intervalo de confianza al nivel $1-\alpha$ de *p*:

$$\left[\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right] \quad \text{con} \quad \Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

Entonces

Si $\hat{P}$ es la proporción de observaciones de una muestra aleatoria de tamaño *n* que verifican una propiedad de interés, entonces un intervalo de confianza para la proporción p de la población que cumple dicha propiedad de nivel aproximadamente $1-\alpha$ es

$$\left[\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}},\ \ \hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right] \qquad (8.10)$$

Observaciones:

1- Este procedimiento depende de la aproximación normal a la distribución binomial. Por lo tanto el intervalo (8.10) se puede utilizar si $n\hat{P} > 10$ y $n(1-\hat{P}) > 10$, es decir, ***la muestra debe contener un mínimo de diez éxitos y diez fracasos.***

2- La longitud del intervalo es $L = 2z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$, pero esta expresión está en función de $\hat{P}$

Si nos interesa hallar un valor de *n* de manera tal que la longitud *L* sea menor que un valor determinado, podemos hacer dos cosas:

a) tomar una muestra preliminar, con ella estimar *p* con $\hat{P}$ y de la expresión anterior despejar *n*, lo que lleva a

$$L = 2z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le l \quad \Rightarrow \quad n \ge \left(\frac{2z_{\frac{\alpha}{2}}}{l}\right)^2 \hat{P}(1-\hat{P})$$

b) si no tomamos una muestra preliminar, entonces acotamos $\hat{P}(1-\hat{P}) \le 0.5\times(1-0.5)$, entonces

$$L = 2z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le 2z_{\frac{\alpha}{2}}\sqrt{\frac{0.5(1-0.5)}{n}} \le l \quad \Rightarrow \quad n \ge \left(\frac{z_{\frac{\alpha}{2}}}{l}\right)^2$$

Ejemplo:

Un fabricante de componentes compra un lote de dispositivos de segunda mano y desea saber la proporción de la población que están fallados. Con ese fin experimenta con 140 dispositivos elegidos al azar y encuentra que 35 de ellos están fallados.

a) Calcular un intervalo de confianza del 99% para la proporción poblacional *p*.

b) ¿De qué tamaño deberá extraerse la muestra a fin de que la proporción muestral no difiera de la proporción poblacional en más de 0.03 con un 95% de confianza?

Solución:

a) El tamaño de la muestra es $n = 140$ (muestra grande)

La proporción muestral es $\hat{P} = \frac{35}{140} = 0.25$

El nivel de confianza es $1-\alpha = 0.99 \rightarrow \alpha = 0.01 \rightarrow \frac{\alpha}{2} = 0.005$.

De la tabla de la normal estandarizada vemos que $z_{0.005} = 2.58$. Entonces el intervalo buscado es:

$$\left[0.25 - 2.58\sqrt{\frac{0.25(1-0.25)}{140}},\ \ 0.25 + 2.58\sqrt{\frac{0.25(1-0.25)}{140}}\right] = [0.15558,\ 0.34441]$$

b) Buscamos el tamaño *n* de la muestra tal que con un 95% de confianza la proporción muestral $\hat{P}$ esté a una distancia 0.03 de la proporción poblacional *p*, es decir buscamos *n* tal que

$\frac{L}{2} \le 0.03$ , por lo tanto como $\alpha = 0.05 \ \rightarrow\ \frac{\alpha}{2} = 0.025$ ***si tomamos la muestra anterior como preliminar*** :

$$n \ge \left(\frac{2z_{\frac{\alpha}{2}}}{l}\right)^2 \hat{P}(1-\hat{P}) = \left(\frac{2\times 1.96}{2\times 0.03}\right)^2 0.25(1-0.25) = 800.3333$$

Por lo tanto hay que tomar una muestra de tamaño por lo menos 801. Como ya se tomó una muestra de tamaño 140, hay que tomar otra adicional de tamaño $801 - 140 = 661$
Supongamos que no tomamos una muestra inicial, entonces directamente planteamos

$$n \ge \left(\frac{z_{\frac{\alpha}{2}}}{l}\right)^2 = \left(\frac{1.96}{2\times 0.03}\right)^2 = 1067.1111$$

Entonces hay que tomar una muestra de tamaño 1068 por lo menos.

## 9.10 – <u>Intervalo de confianza para la diferencia entre dos proporciones</u>

Supongamos que existen dos proporciones de interés $p_1$ y $p_2$ y es necesario obtener un intervalo de confianza de nivel $1-\alpha$ para la diferencia $p_1 - p_2$.
Supongamos que se toman dos muestras independientes de tamaños $n_1$ y $n_2$ respectivamente de dos poblaciones.
Sean las variables aleatorias
$X_1$ : "número de observaciones en la primera muestra que tienen la propiedad de interés"
$X_2$ : "número de observaciones en la segunda muestra que tienen la propiedad de interés"
Entonces $X_1$ y $X_2$ son variables aleatorias independientes y $X_1 \sim B(n_1,p_1)$ ; $X_2 \sim B(n_2,p_2)$
Además $\hat{P}_1 = \frac{X_1}{n_1}$ y $\hat{P}_2 = \frac{X_2}{n_2}$ son estimadores puntuales de $p_1$ y $p_2$ respectivamente.
Vemos que $E(\hat{P}_1 - \hat{P}_2) = p_1 - p_2$ y $V(\hat{P}_1 - \hat{P}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
Aplicando la aproximación normal a la binomial podemos decir que

$Z = \dfrac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$ , y como en el caso de intervalo para una proporción estima-

mos $\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$ con $\sqrt{\dfrac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \dfrac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$ y entonces

$Z = \dfrac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \dfrac{\hat{P}_2(1-\hat{P}_2)}{n_2}}} \approx N(0,1)$ aproximadamente.

Planteamos la ecuación $P(-z \le Z \le z) \approx \Phi(z) - \Phi(-z) = 1-\alpha$, lo que lleva a $z = z_{\frac{\alpha}{2}}$, y con una deducción análoga a las anteriores se llega al intervalo

$$\left[ \hat{P}_1 - \hat{P}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}};\ \hat{P}_1 - \hat{P}_2 + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \right]$$

Entonces

Si $\hat{P}_1$ y $\hat{P}_2$ son las proporciones muestrales de una observación de dos muestras aleatorias independientes de tamaños $n_1$ y $n_2$ respectivamente que verifican la propiedad de interés, entonces un intervalo de confianza de nivel $1-\alpha$ aproximadamente es

$$\left[ \hat{P}_1 - \hat{P}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}};\ \hat{P}_1 - \hat{P}_2 + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \right] \quad (8.11)$$

Ejemplo:
Se lleva a cabo un estudio para determinar la efectividad de una nueva vacuna contra la gripe. Se administra la vacuna a una muestra aleatoria de 3000 sujetos, y de ese grupo 13 contraen gripe. Como grupo de control se seleccionan al azar 2500 sujetos, a los cuales no se les administra la vacuna, y de ese grupo 170 contraen gripe. Construya un intervalo de confianza de nivel 0.95 para la diferencia entre las verdaderas proporciones de individuos que contraen gripe.

Solución:
Sean las variables aleatorias
$X_1$ : "número de personas que contraen gripe del grupo que recibió la vacuna"
$X_2$ : "número de personas que contraen gripe del grupo que no recibió la vacuna"
Entonces $X_1 \sim B(n_1,p_1)$ ; $X_2 \sim B(n_2,p_2)$ donde $n_1 = 3000$ ; $n_2 = 2500$
Además $\hat{P}_1 = \dfrac{13}{3000}$ ; $\hat{P}_2 = \dfrac{170}{2500}$
Y $1-\alpha = 0.95 \quad \rightarrow \quad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Entonces

$$\left[\hat{P}_1-\hat{P}_2-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}_1\left(1-\hat{P}_1\right)}{n_1}+\frac{\hat{P}_2\left(1-\hat{P}_2\right)}{n_2}};\ \ \hat{P}_1-\hat{P}_2+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}_1\left(1-\hat{P}_1\right)}{n_1}+\frac{\hat{P}_2\left(1-\hat{P}_2\right)}{n_2}}\right]=$$

$$=\left[\frac{13}{3000}-\frac{170}{2500}-1.96\sqrt{\frac{\frac{13}{3000}\left(1-\frac{13}{3000}\right)}{3000}+\frac{\frac{170}{2500}\left(1-\frac{170}{2500}\right)}{2500}};\right.$$

$$\left.\frac{13}{3000}-\frac{170}{2500}+1.96\sqrt{\frac{\frac{13}{3000}\left(1-\frac{13}{3000}\right)}{3000}+\frac{\frac{170}{2500}\left(1-\frac{170}{2500}\right)}{2500}}\right]=\left[-0.0738112;\ \ -0.0535222\right]$$