ESTIMACIÓN

Se conoce como inferencia estadística al proceso de generalizar estos resultados muestrales a la población

Inferencia Estadística y Estimación

La estadística inferencial incluye aquellos métodos que permiten realizar inferencias o generalizaciones acerca de la población a partir de los resultados obtenidos para una muestra y empleando la teoría de probabilidades.

Podemos distinguir entre

- El Método Clásico para estimar el parámetro poblacional en el que las inferencias se basan en la información obtenida de una muestra aleatoria seleccionada de la población.

- El Método Bayesiano que utiliza el conocimiento subjetivo previo acerca de la distribución de probabilidad con los parámetros conocidos, junto con la información proporcionada por los datos muestrales

Utilizaremos los métodos clásicos, es decir que calcularemos los estadísticos a partir de muestras aleatorias y aplicando la teoría de las distribuciones muestrales estableceremos conclusiones respecto a parámetros poblacionales tales como media, proporción o varianza

Las áreas principales de la inferencia estadística son la estimación y las pruebas de hipótesis estadísticas. Para distinguir entre estas áreas consideremos los siguientes ejemplos:

- Un candidato para un puesto público desea estimar la proporción real de votantes que lo apoyan mediante la obtención de las opiniones sobre una muestra aleatoria de 100 votantes. La fracción de ellos que lo apoye puede utilizarse como una estimación de la proporción real de la población total de votantes. Un conocimiento de la distribución muestral de una proporción permite establecer el grado de precisión de la estimación. Este problema pertenece al área de la estimación.

- Un ama de casa se interesa en determinar si la cera para pisos marca A es más resistente que el de marca B. El ama de casa podría suponer que la marca A es mejor que la B y, después de realizar las pruebas apropiadas, aceptar o rechazar esta hipótesis. En este caso se intenta tomar una decisión correcta respecto a la hipótesis preestablecida. La teoría de muestreo permitirá obtener alguna medida de precisión para la decisión que se tome.

Métodos clásicos de estimación

Sea θ un parámetro poblacional y $\hat{\theta}$ el estadístico que se emplea para estimar el parámetro poblacional. Una estimación puntual de algún parámetro poblacional θ es un valor único $\hat{\theta}$ del estadístico $\hat{\theta}$.

Por ejemplo:

- el valor $\bar{x}$ del estadístico $\bar{X}$, calculado a partir de una muestra de tamaño n, es una estimación puntual de μ_X, la media poblacional de la variable aleatoria X.

- el valor $\hat{p} = \frac{x}{n}$ del estadístico $\hat{P}$, obtenido a partir de una muestra de tamaño n, es una estimación puntual de p, la proporción verdadera para un experimento binomial.

El estadístico que se usa para obtener una estimación puntual recibe el nombre de estimador o **función de decisión.**

De aquí que la función de decisión Varianza muestral (S^2) que depende de la muestra aleatoria, es un estimador de la varianza poblacional σ_X^2, y la estimación s^2 que se obtiene es la acción que se toma.
En general, muestras diferentes conducen a acciones o estimaciones (valor del estimador) diferentes.
No se espera que un estimador obtenga sin error un parámetro poblacional; no se espera que $\bar{X}$ estime con exactitud el valor de μ_X sino que no se aleje mucho del valor real.
Para una muestra en particular, es posible obtener una estimación más precisa de μ_X utilizando como estimador la mediana muestral ($\tilde{X}$)

¿ Cuáles son las propiedades de una buena función de decisión que influirán en la selección de un estimador sobre otro?

Sea $\hat{\Theta}$ un estimador cuyo valor $\hat{\theta}$ es una estimación puntual de algún parámetro poblacional desconocido θ ; sería deseable que la distribución muestral de $\hat{\Theta}$ tuviera una media igual al parámetro estimado. A un estimador que posee esta propiedad se le llama **estimador insesgado.**

Entonces ; se dice que un estadístico $\hat{\Theta}$ es un estimador insesgado del parámetro θ si

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta$$

La varianza muestral, S^2 es un estimador insesgado de la varianza σ_X^2

La desviación estándar muestral S es un estimador sesgado σ_X con la tendencia a hacer el sesgo insignificante en muestras grandes.

Si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores insesgados del mismo parámetro poblacional θ, se seleccionaría el estimador cuya distribución muestral tuviera la varianza más pequeña.

Si $\sigma^2_{\hat{\Theta}_1} < \sigma^2_{\hat{\Theta}_2}$ se dice que $\hat{\Theta}_1$ es un **estimador más eficiente** de θ que $\hat{\Theta}_2$

De modo que si se consideran todos los estimadores insesgados posibles para algún parámetro θ aquel con la varianza más pequeña recibe el nombre de estimador más eficiente de θ.

En la Figura 1 se presentan las distribuciones muestrales de cuatro estimadores $\hat{\Theta}_1$, $\hat{\Theta}_2$, $\hat{\Theta}_3$ y $\hat{\Theta}_4$ del parámetro θ. Se observa que $\hat{\Theta}_1$, $\hat{\Theta}_2$ y $\hat{\Theta}_3$ son insesgados dado que sus distribuciones se centran en θ. El estimador $\hat{\Theta}_1$ tiene una varianza más pequeña que $\hat{\Theta}_2$ y $\hat{\Theta}_3$ y por lo tanto es más eficiente. El estimador de θ que se seleccionaría en este caso es $\hat{\Theta}_1$ porque resulta el estimador insesgado más eficiente.

Para poblaciones normales se puede demostrar que tanto $\bar{X}$ y $\tilde{X}$ son estimadores insesgados de la media poblacional μ_X pero la varianza de $\bar{X}$ es menor que la de $\tilde{X}$. Entonces entre las estimaciones $\bar{x}$ y $\tilde{x}$ probablemente $\bar{x}$ esté más cerca del valor de μ_X para una muestra dada y, $\bar{X}$ resulta más eficiente que $\tilde{X}$.

Es probable que, incluso el estimador insesgado más eficiente no estime el parámetro poblacional con exactitud. Es cierto que la precisión se incrementa con muestras grandes, pero no hay razón para esperar que la estimación puntual de una muestra única debe ser exactamente igual que el parámetro poblacional que se supone que estima.

En la mayoría de las situaciones es preferible determinar un intervalo dentro del cual se esperaría encontrar el valor del parámetro. A eso se conoce con estimación por intervalo.

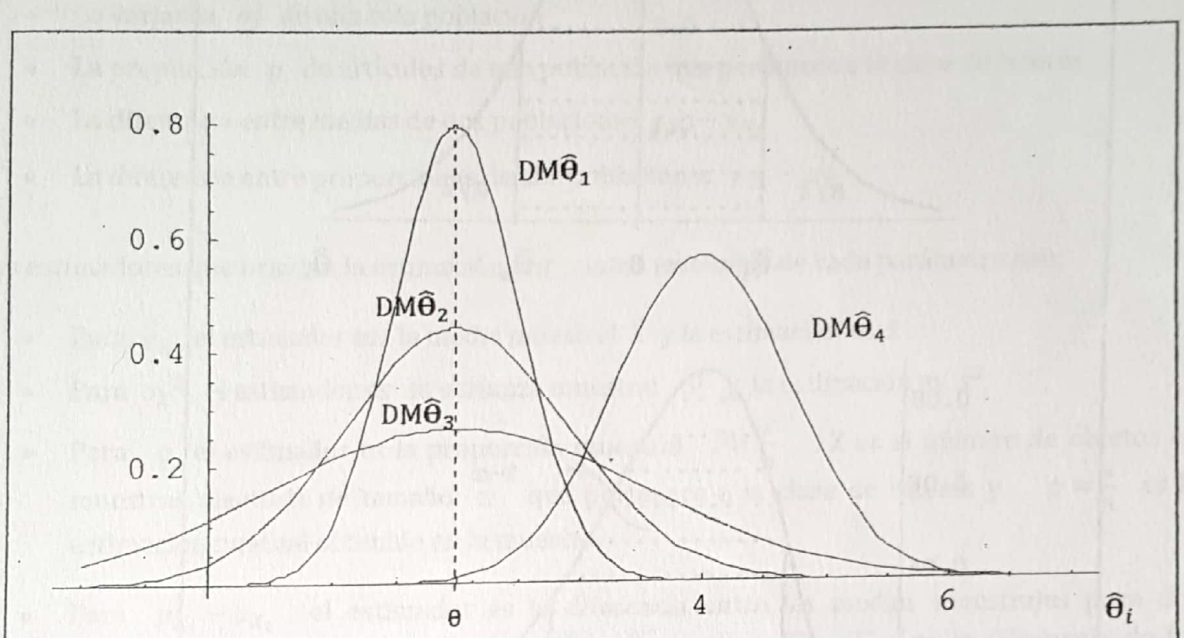

Figura 1: Distribuciones Muestrales de los estimadores $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ y $\hat{\theta}_4$

Estimación por Intervalo

Una estimación por intervalo de un parámetro poblacional θ es un intervalo de la forma $\hat{\theta}_i < \theta < \hat{\theta}_s$, donde los extremos inferior y superior $\hat{\theta}_i$ y $\hat{\theta}_s$ dependen del valor del estadístico $\hat{\theta}$ para una muestra particular y también de la distribución muestral de ese estadístico (DM $\hat{\theta}$). Cuanto más pequeño sea el intervalo, más se aproxima la estimación al parámetro θ.

Ya que muestras distintas generalmente dan valores distintos de $\hat{\theta}$ y por lo tanto de $\hat{\theta}_i$ y $\hat{\theta}_s$. Estos extremos del intervalo corresponden a algún valor de las variables aleatorias $\hat{\Theta}_i$ y $\hat{\Theta}_s$. Si θ es el parámetro que se estima con el estadístico $\hat{\Theta}$ con distribución muestral conocida, se pueden calcular $\hat{\theta}_i$ y $\hat{\theta}_s$ tal que $P(\hat{\Theta}_i < \theta < \hat{\Theta}_s)$ sea igual a cualquier valor que se desee especificar. Por ejemplo, si

$$P(\hat{\Theta}_i < \theta < \hat{\Theta}_s) = 1-\alpha \qquad (0<\alpha<1)$$

significa que si se tomaran todas las muestras posibles de tamaño n de una población la probabilidad que el parámetro θ esté en el intervalo $(\hat{\Theta}_i, \hat{\Theta}_s)$ es de $1-\alpha$. Esto es que la proporción de muestras (de entre todas las muestras posibles de tamaño determinado a extraer de la población) que incluirían al parámetro θ en el intervalo $(\hat{\Theta}_i, \hat{\Theta}_s)$ es de $1-\alpha$. Esta

probabilidad coincide con la probabilidad de seleccionar una muestra aleatoria que produzca un intervalo de confianza que incluya a θ. (Ver Figura 2).

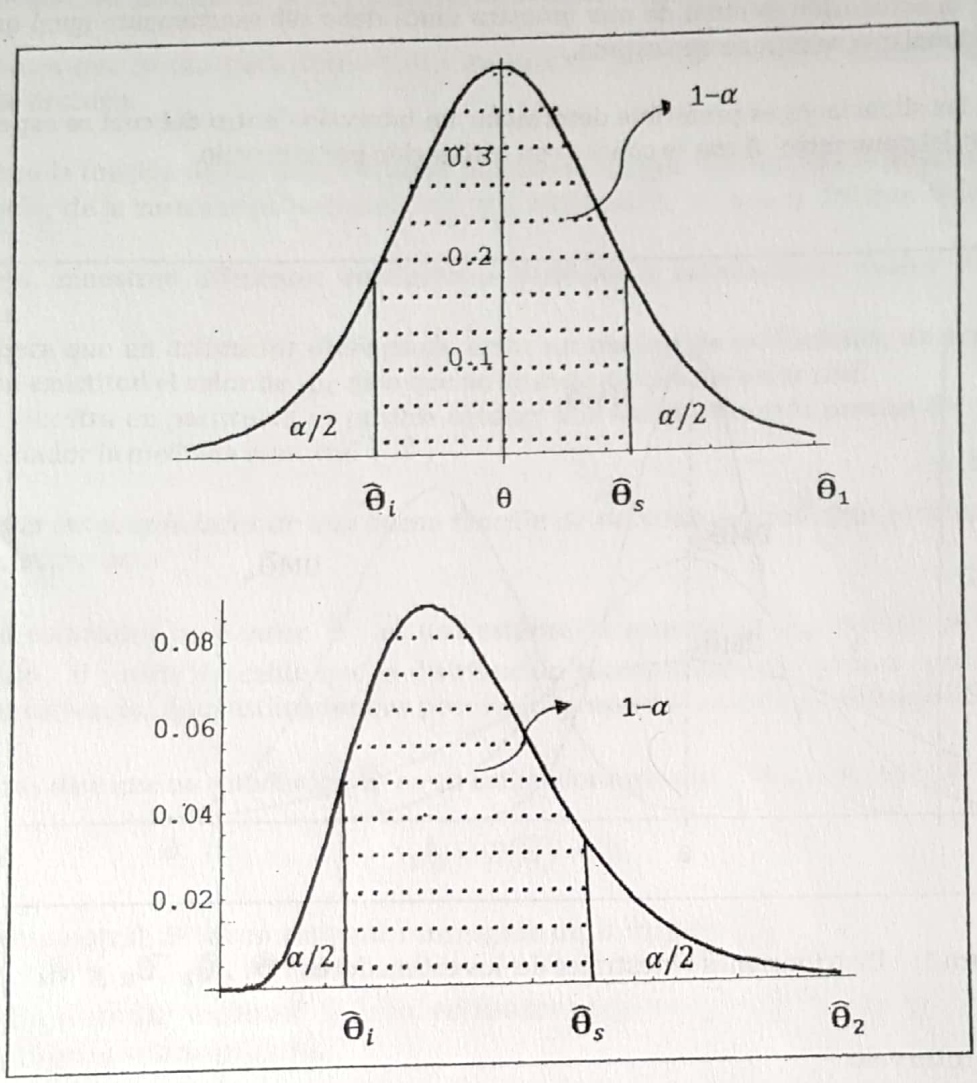

Figura 2: $P(\hat{\theta}_i < \theta < \hat{\theta}_s) = 1-\alpha \quad (0 < \alpha < 1)$

Si estimo θ, a partir de una muestra única, se obtiene una estimación puntual $\hat{\theta}$ del estadístico $\hat{\theta}$, la que permitirá calcular un intervalo $\hat{\theta_i} < \theta < \hat{\theta_s}$ para θ con $1-\alpha$ de probabilidades de que ese intervalo sea uno que contenga al parámetro poblacional θ.

El intervalo $\hat{\theta_i} < \theta < \hat{\theta_s}$ que se calcula a partir de la muestra seleccionada se denomina **intervalo de confianza del $(1-\alpha)$ 100 %**.

La fracción $1-\alpha$ recibe el nombre de **grado de confianza o coeficiente de confianza**

Los valores extremos $\hat{\theta_i}$ y $\hat{\theta_s}$ se llaman **límites de confianza inferior y superior**.

De modo que cuando $\alpha = 0{,}05$ se tiene un intervalo de confianza del 95%. Si $\alpha = 0{,}01$ se tiene un intervalo más amplio del 99%.

Cuanto mayor es el intervalo de confianza, se tiene más seguridad de que el intervalo dado contiene al parámetro desconocido. Por supuesto es mejor tener un 95% de confianza de que la vida promedio de determinado transistor de TV está entre 6 y 7 años que el 99% de confianza de

que está entre 3 y 10 años. Desde el punto de vista ideal, es preferible un intervalo de poca amplitud con un alto grado de confianza.

Las estimaciones por intervalo de confianza se basan en estimaciones puntuales.

A menudo necesitamos estimar los siguientes parámetros:

- La media μ_X de una sola población
- La varianza σ_X^2 de una sola población
- La proporción p de artículos de una población que pertenece a la clase de interés
- La diferencia entre medias de dos poblaciones $\mu_{X_1} - \mu_{X_2}$
- La diferencia entre proporciones de dos poblaciones $p_{X_1} - p_{X_2}$

Los estimadores que brindan la estimación por puntos razonable de cada parámetro son:

- Para μ_X, el estimador es la media muestral $\bar{X}$ y la estimación es $\bar{x}$
- Para $\sigma_X{}^2$, el estimador es la varianza muestral S^2 y la estimación es s^2
- Para p el estimador es la proporción muestral $\hat{P} = \frac{X}{n}$. X es el número de objetos en muestras aleatoria de tamaño n que pertenece a la clase de interés y $\hat{p} = \frac{x}{n}$ es la estimación puntual obtenida en la muestra.
- Para $\mu_{X_1} - \mu_{X_2}$, el estimador es la diferencia entre las medias muestrales para dos muestras aleatorias independientes $\overline{X_1} - \overline{X_2}$ y $\overline{x_1} - \overline{x_2}$ es la diferencia de las estimaciones puntuales obtenidas en una muestra única de cada población
- Para $p_{X_1} - p_{X_2}$, el estimador es la diferencia entre las proporciones para todas las muestras de tamaños n_1 y n_2 que se puedan tomar de las poblaciones $(\widehat{P_1} - \widehat{P_2})$, y la estimación es $\widehat{p_1} - \widehat{p_2}$ la diferencia de las estimaciones puntuales de las proporciones obtenidas a partir de dos únicas muestras aleatorias e independientes tomadas una de cada población.

Estimación de la media

El estadístico que se emplea para estimar la media poblacional μ_X, es la media muestral, $\bar{X}$. La DM $(\bar{X})$ se centra en $\mu_{\bar{X}} = \mu_X$ y la varianza de $\bar{X}$ es $\sigma_{\bar{X}}{}^2 = \frac{\sigma_X{}^2}{n}$, resulta menor que la varianza de los otros estimadores de μ_X. En este caso $\bar{x}$ la media de una muestra aleatoria se empleará como una estimación puntual de μ_X. Muestras grandes darán una distribución muestral de $\bar{X}$ con varianza pequeña, entonces la estimación puntual $\bar{x}$ es una estimación muy precisa de μ_X.

Si la muestra se selecciona de una población normal o, a falta de esto, si el tamaño de la muestra n es lo bastante grande ($n \geq 30$), se puede establecer un intervalo de confianza considerando que la DM $(\bar{X})$ es normal

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X{}^2}{n}\right)$$

Entonces la variable aleatoria $Z = \dfrac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$ tiene una distribución normal estándar: $Z \sim N(0,1)$

Escribimos $z_{\alpha/2}$ como el valor de z por encima del cual se encuentra un área de $\alpha/2$ y

$-z_{\alpha/2}$ es el valor de z debajo del cual se encuentra un área de $\alpha/2$.

Resulta $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$ según se observa en la Figura 3

Como $Z = \dfrac{\bar{X}-\mu_X}{\frac{\sigma_X}{\sqrt{n}}}$, entonces $P(-z_{\alpha/2} < \dfrac{\bar{X}-\mu_X}{\frac{\sigma_X}{\sqrt{n}}} < z_{\alpha/2}) = 1-\alpha$

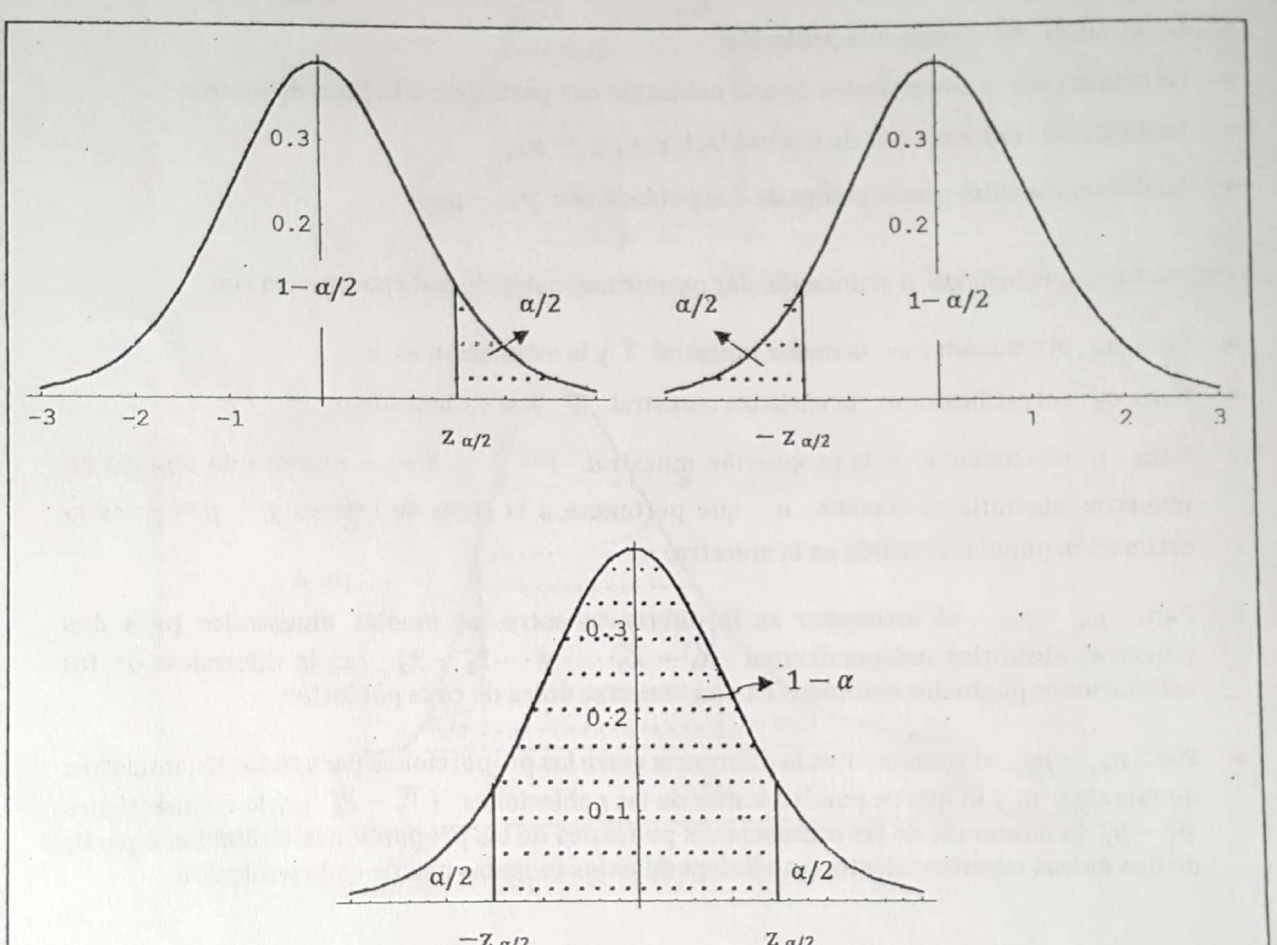

Figura 3: Interpretación de $z_{\alpha/2}$ y $-z_{\alpha/2}$ en la Distribución Normal Estándar. Representación del grado de confianza $1-\alpha$

Si trabajamos algebraicamente con la desigualdad, se obtiene:

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}) = 1-\alpha$$

Donde

$$\hat{\theta}_i = \bar{X} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \qquad \hat{\theta}_s = \bar{X} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}$$

Ahora, si se selecciona una muestra aleatoria de tamaño n de una población de varianza $\sigma_X{}^2$ conocida y se calcula la media $\bar{x}$ para la muestra, se obtiene el siguiente intervalo de confianza del $(1-\alpha)\,100\%$ para μ_X:

$$\bar{x} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{x} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \qquad \text{IC(1)}$$

Donde $z_{\alpha/2}$ es el valor de z de la distribución normal estándar a la derecha del cual se encuentra un área de $\alpha/2$.

Para muestras pequeñas que se seleccionan de poblaciones que no son normales, no se puede esperar un grado de confianza preciso. Para muestras de tamaño $n \geq 30$, sin importar la forma de la población, la teoría muestral garantiza buenos resultados.

Observemos que los límites inferior $(\widehat{\theta_i})$ y superior $(\widehat{\theta_s})$ del intervalo de confianza obtenido resultan:

$$\widehat{\theta_i} = \bar{x} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \qquad \widehat{\theta_s} = \bar{x} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}$$

Muestras diferentes darán valores diferentes de $\bar{x}$ y, por lo tanto producirán diferentes estimaciones por intervalos de confianza del parámetro μ_X. En la Figura 4, se observa el esquema de estos intervalos. Los círculos en el centro de cada intervalo indican la posición de la estimación puntual $\bar{x}$ para cada muestra aleatoria. Se observa que la mayor parte de los intervalos contienen a μ_X pero no todos. Todos los intervalos tienen la misma amplitud, ya que ésta depende, para cada determinación de $\bar{x}$, de la selección de $z_{\alpha/2}$ o del nivel de confianza.

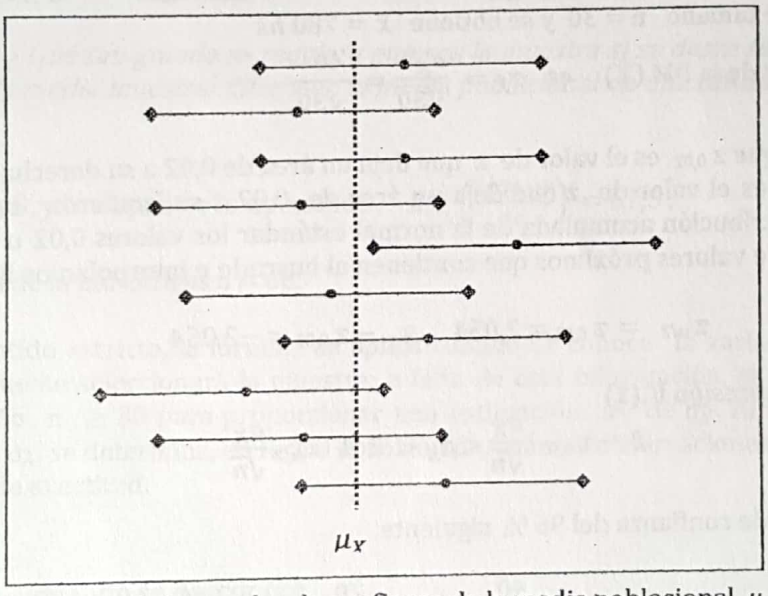

Figura 4: Intervalos de confianza de la media poblacional μ_X para diferentes muestras.

Cuanto más grande es el valor de $z_{\alpha/2}$ más amplios se hacen todos los intervalos y mayor es la confianza de que la muestra seleccionada producirá un intervalo que contenga el parámetro desconocido μ_X.

Para algunas muestras la estimación del intervalo de confianza será correcto y para otras no. En la práctica seleccionamos sólo una muestra y no conocemos la verdadera media de la población y tampoco podemos determinar si nuestra estimación es correcta. Podemos determinar la porción de muestras que producen resultados que nos llevan a construir intervalos de confianza que nos llevan a conclusiones correctas respecto a la media poblacional.

Una estimación del intervalo de confianza del 95%, puede interpretarse diciendo que, si se tomaran todas las muestras posibles del mismo tamaño n, el 95% de ellas proporcionarán un intervalo de confianza que incluye la media desconocida de la población en alguna parte del

intervalo construido alrededor de sus medias muestrales y sólo el 5% de ellas no incluirían la media verdadera en el intervalo obtenido. Entonces, tenemos una probabilidad del 95% de que el intervalo de confianza obtenido a partir de una muestra aleatoria sea uno de los que contenga el parámetro poblacional. Tenemos, en definitiva, el 95 % de confianza que hemos seleccionado una muestra cuyo intervalo contiene a la media.

Ejemplo 1

Un fabricante produce focos cuya vida útil en horas tiene una distribución aproximadamente normal y una desviación estándar de 40 horas. Si una muestra de 30 focos tiene una vida promedio de 780 horas, encuentra el intervalo de confianza del 96% para la duración media de los focos que produce esta empresa

La variable aleatoria es X : vida útil de los focos (en horas) con varianza $\sigma_X^2 = (40\ hs)^2$ y distribución aproximadamente normal

El intervalo de confianza para el parámetro μ_X debe determinarse con un nivel de confianza de $(1-\alpha)\ 100\% = 96\ \% \rightarrow grado\ de\ confianza \quad \alpha = 0{,}04 \rightarrow \frac{\alpha}{2} = 0{,}02$

El estadístico que se emplea para construir el intervalo de confianza es $\bar{X} \quad \bar{X} \sim N(\mu_X, \frac{\sigma_X{}^2}{n})$

Se toma un muestra de tamaño $n = 30$ y se obtiene $\bar{x} = 780\ hs$

La desviación estándar de la DM $(\bar{X})$ es $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{30}} = \frac{40}{\sqrt{30}}$

Si tenemos en cuenta que $z_{0,02}$ es el valor de z que deja un área de 0,02 a su derecha y de 0,98 a su izquierda y $-z_{0,02}$ es el valor de z que deja un área de 0,02 a su izquierda , buscamos en el cuerpo la tabla de distribución acumulada de la normal estándar los valores 0,02 o 0,98. En este caso tomamos el par de valores próximos que contienen al buscado e interpolamos linealmente. Se obtiene

$$z_{\alpha/2} = z_{0,02} = 2{,}054 \quad \text{y} \quad -z_{0,02} = -2{,}054$$

Reemplazando en la expresión **IC(1)**

$$\bar{x} - z_{\alpha/2}\ \frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{x} + z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}$$

Se obtiene el intervalo de confianza del 96 % siguiente

$$780 - 2{,}054\ \frac{40}{\sqrt{30}} < \mu_X < 780 + 2{,}054 \frac{40}{\sqrt{30}}$$

$$765 < \mu_X < 795$$

Error en la estimación de μ_X

El intervalo de confianza del $(1-\alpha)100\%$ proporciona una precisión de la exactitud de la estimación puntual. En general $\bar{x}$, la estimación puntual no será exactamente igual a μ_X y es errónea. El tamaño de ese error será el valor absoluto de la diferencia entre $\bar{x}$ y μ_X y se puede tener una confianza del $(1-\alpha)100\%$ de que esa diferencia no excederá $z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}$ (Ver Figura 5)

En el ***Ejemplo 1*** se tiene una confianza del 96 % de que la media muestral $\bar{x} = 780$ difiere de la media poblacional por una cantidad menor que 15.

Con frecuencia se desea determinar el tamaño de la muestra para asegurar que el error en la estimación de μ_X será menor que una cantidad especificada e. Esto significa que se debe seleccionar una muestra de tamaño n tal que $z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \leq e$. Así, resulta que si se utiliza $\bar{x}$ como una estimación de μ_X, se puede tener una confianza del $(1-\alpha)100\%$ de que el error no excederá una cantidad e cuando el tamaño de la muestra es

$$n \geq (z_{\alpha/2}\frac{\sigma_X}{e})^2$$

El resultado obtenido se redondea al número entero más cercano.

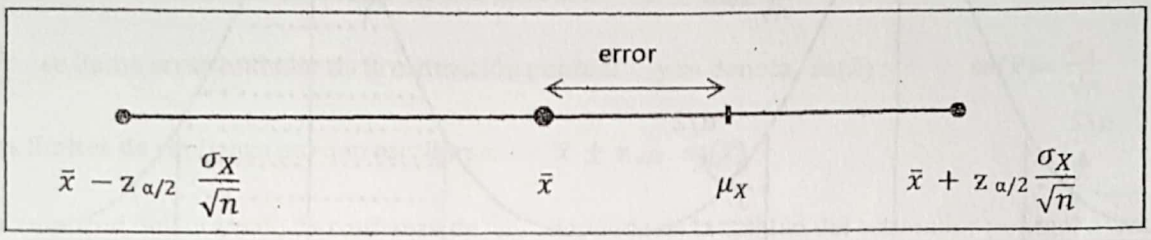

Figura 5: Error al estimar μ_X por $\bar{x}$

*Para el **Ejemplo 1** ¿ Qué tan grande se requiere que sea la muestra si se desea tener una confianza del 96% de que la media muestral difiera de la media poblacional en una cantidad menor que 10 horas ?*

En este caso $e = 10$ y se obtiene, reemplazando $n \geq \left(\frac{2{,}054 \cdot 40}{10}\right)^2 = 67{,}5$.

El tamaño mínimo de la muestra es n = 68.

Aunque, en el sentido estricto, la fórmula se aplica cuando se conoce la varianza σ_X^2 de la población de la que se seleccionará la muestra; a falta de esta información, se puede tomar una muestra de tamaño $n \geq 30$ para proporcionar una estimación s^2 de σ_X^2. Al emplear s como aproximación de σ_X se determina, en forma aproximada, cuántas observaciones se necesitan para el grado deseado de exactitud.

Estimación de μ_X cuando se desconoce σ_X^2

Para muestras aleatorias de una población normal, la variable aleatoria T / $T = \frac{\bar{X}-\mu_X}{\frac{S}{\sqrt{n}}}$

desviación estándar muestral

Tiene una distribución **t de Student** con **ν= n-1 grados de libertad**, donde S es el estadístico desviación estándar muestral.

Emplearemos las DM (T) para determinar el intervalo de confianza de μ_X. Entonces:

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1-\alpha$$

$t_{\alpha/2}$ es el valor de t con ν= n-1 grados de libertad que deja un área de $\alpha/2$ a su derecha y

$-t_{\alpha/2}$ es el valor de t con ν= n-1 grados de libertad que deja un área de $\alpha/2$ a su izquierda.

Como $T = \dfrac{\bar{X} - \mu_X}{\frac{S}{\sqrt{n}}}$, resulta:

$$P(-t_{\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{S}{\sqrt{n}}} < t_{\alpha/2}) = P(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu_X < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1 - \alpha$$

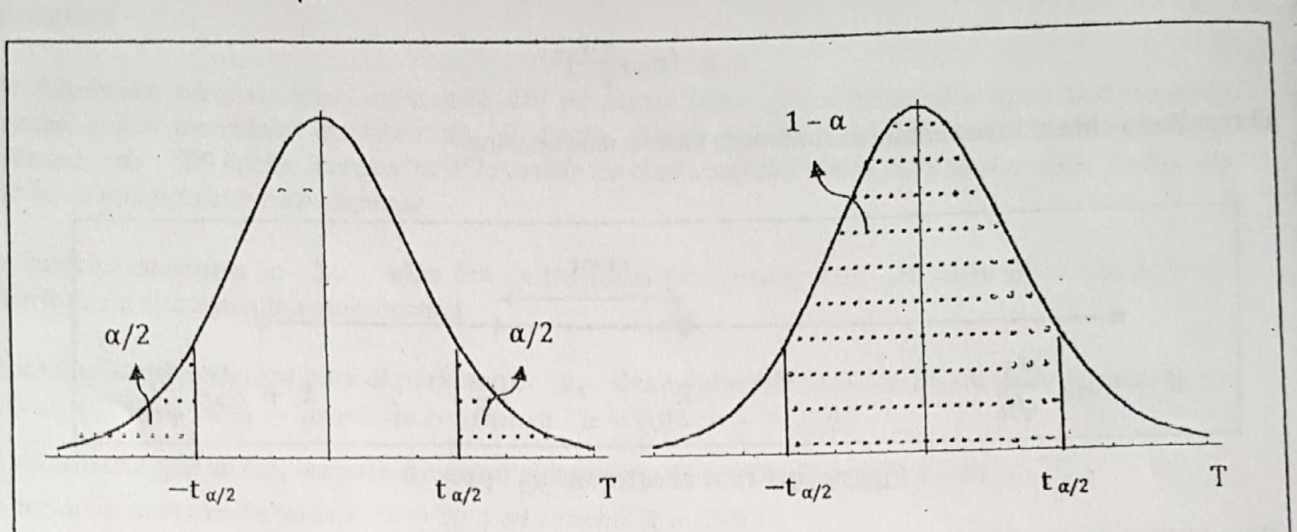

Figura 6: Ubicación de $-t_{\alpha/2}$ y $t_{\alpha/2}$ y $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$

Para el caso de una muestra aleatoria de tamaño n, de una población normal, se calculan $\bar{x}$ y s a partir de los valores de la variable en la muestra y se obtiene el siguiente intervalo de confianza del $(1 - \alpha)100\%$ para μ_X cuando se desconoce σ_X^2

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu_X < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} \quad \text{IC(2)}$$

Donde $t_{\alpha/2}$ como el valor de t a la derecha del cual se encuentra un área de $\alpha/2$.

Mientras la distribución de X se aproxime a la forma de campana, los intervalos de confianza pueden calcularse, cuando σ_X^2 se desconoce , utilizando la DM (T) y se pueden esperar muy buenos resultados.

Cuando no se pueda suponer normalidad y σ_X^2 es desconocida , si $n \geq 30$, el valor de s obtenido de la muestra puede reemplazar a σ_X y se recomienda emplear el siguiente intervalo de confianza

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu_X < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}} \quad \text{IC(3)}$$

En este caso se emplea la distribución de la variable aleatoria Z . A este intervalo se lo conoce como **intervalo de confianza para muestra grande.**

Error estándar de una estimación puntual

Las estimaciones puntuales proveen un número único que se obtiene a partir de un conjunto de datos experimentales y las estimaciones por intervalos de confianza proporcionan un intervalo a partir de los datos experimentales que contiene valores que se consideran razonables para el

parámetro. Esto es $(1-\alpha)100\%$ de tales intervalos calculados "cubren el parámetro" y se dice, por ejemplo, que se tiene una confianza del $(1-\alpha)100\%$ de que la media de la población se encuentre en el intervalo obtenido.

Estas dos aproximaciones al parámetro se relacionan una con otra.

Sea el estimador $\bar{X}$ de μ_X cuando se conoce σ_X^2. Una medida de la calidad de este estimador insesgado es su varianza $\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$ que determina una desviación estándar

Los límites de confianza obtenidos en este caso son: $\bar{x} \pm z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}$

$\frac{\sigma_X}{\sqrt{n}}$ se llama **error estándar de la estimación puntual** y se denota $se(\bar{x})$: $se(\bar{x}) = \frac{\sigma_X}{\sqrt{n}}$

Los límites de confianza pueden escribirse: $\bar{x} \pm z_{\alpha/2} . se(\bar{x})$

La amplitud del intervalo de confianza de μ_X depende de la calidad del estimador puntual a través de su error estándar.

En el caso donde no se conoce σ_X^2 y X es normal , s reemplaza a σ y surge un **error estándar estimado** de la estimación puntual $\bar{x}$ igual a $\frac{s}{\sqrt{n}}$ que se denota $\widehat{se}(\bar{x})$: $\widehat{se}(\bar{x}) = \frac{s}{\sqrt{n}}$

Los límites de confianza resultan : $\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm z_{\alpha/2} . \widehat{se}(\bar{x})$

En este caso también se observa que la amplitud del intervalo de confianza depende de la calidad de la estimación puntual y esto se evidencia a través de su error estándar estimado

Las amplitudes de los intervalos de confianza se hacen menores a medida que mejora la calidad de las correspondientes estimaciones puntuales. Un intervalo de confianza resulta una ampliación de la estimación puntual para considerar la precisión de la misma.

Estimación de la diferencia entre dos medias con varianzas conocidas

Se tienen dos poblaciones X_1 y X_2 Normales con medias μ_{X_1} y μ_{X_2} y varianzas conocidas $\sigma_{X_1}^2$ y $\sigma_{X_2}^2$ respectivamente. La estimación puntual del parámetro desconocido, la diferencia $\mu_{X_1} - \mu_{X_2}$ es un valor del estadístico $\bar{X}_1 - \bar{X}_2$. Para obtener ese valor, una estimación puntual de la diferencia de medias, se seleccionan dos muestras independientes, una de cada población de tamaños n_1 y n_2 respectivamente. Luego, con los datos obtenidos en cada muestra, se calculan $\bar{x}_1$ y $\bar{x}_2$ y la diferencia $\bar{x}_1 - \bar{x}_2$ de las estimaciones puntuales.

Sabemos que se puede esperar que la $DM(\bar{X}_1 - \bar{X}_2)$ sea aproximadamente en forma normal con

$$\text{Media} \quad \mu_{\bar{X}_1 - \bar{X}_2} = \mu_{X_1} - \mu_{X_2} \quad \text{y} \quad \text{Varianza} \quad \sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}$$

Por lo tanto, se puede afirmar con una probabilidad de $1-\alpha$ que la variable normal estándar Z/

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}}}$$

Caerá entre $-z_{\alpha/2}$ y $z_{\alpha/2}$, es decir $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$

Esto es :

$$P\left(-z_{\alpha/2} < \frac{(\overline{X_1} - \overline{X_2}) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha$$

Lo anterior conduce al siguiente intervalo de confianza del $(1-\alpha)100\%$ para la diferencia de medias $\mu_{X_1} - \mu_{X_2}$ en dos poblaciones con varianzas $\sigma_{X_1}^2$ y $\sigma_{X_2}^2$ conocidas

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}} < \mu_{X_1} - \mu_{X_2} < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}}$$

IC(4)

Con $\bar{x}_1$ y $\bar{x}_2$ las medias de muestras aleatorias independientes de tamaños n_1 y n_2 y $z_{\alpha/2}$ el valor de z que tiene un área igual a $\alpha/2$ a su derecha.

El grado de confianza es exacto cuando las muestras se seleccionan de poblaciones normales. Para poblaciones no normales, el teorema del límite central proporciona una buena aproximación para muestras de tamaño razonable.

El **error estándar** de la estimación puntual es:

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}}$$

Ejemplo 2

Se aplica una evaluación estandarizada de Química a un grupo de 50 niñas y 75 niños. Las niñas obtienen una calificación promedio de 76 puntos y los niños de 82. Encuentra un intervalo de confianza del 96 % para la diferencia $\mu_{X_1} - \mu_{X_2}$, donde μ_{X_1} es la calificación promedio de los niños y μ_{X_2} es la calificación promedio de las niñas. Supone que la deviación estándar de las poblaciones de niñas y niños son 6 y 8 respectivamente.

La estimación puntual de la diferencia de medias, obtenida a partir de muestras de tamaño 75 y 50 de las poblaciones de niños y niñas respectivamente es: $\bar{x}_1 - \bar{x}_2 = 82 - 76 = 6$

$(1-\alpha)\ 100\% = 96\ \% \rightarrow grado\ de\ confianza \quad \alpha = 0{,}04 \rightarrow \frac{\alpha}{2} = 0{,}02$

$z_{0,02} = 2{,}054$ (Se busca 0,98 en el cuerpo de la tabla de distribución acumulada de la normal)

Empleando la expresión **IC(4)**, el intervalo de confianza del 96 % es

$$6 - 2{,}054\sqrt{\frac{8^2}{75}+\frac{6^2}{50}} < \mu_{X_1} - \mu_{X_2} < 6 + 2{,}054\sqrt{\frac{8^2}{75}+\frac{6^2}{50}}$$

$$3{,}424 < \mu_{X_1} - \mu_{X_2} < 8{,}576$$

Estimación de la diferencia entre dos medias para varianzas desconocidas y muestra grande

Cuando $\sigma_{X_1}^2$ y $\sigma_{X_2}^2$ son desconocidas y se obtienen, a partir de las observaciones en las muestras s_1^2 y s_2^2, se obtiene un estadístico con una distribución normal estándar aproximada cuando las muestras son grandes. Resulta Z /

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

El intervalo de confianza es

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < \mu_{X_1} - \mu_{X_2} < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

IC(5)

Y el error estándar estimado

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Estimación de la diferencia entre dos medias para varianzas desconocidas e iguales

Dadas dos poblaciones normales con varianzas $\sigma_{X_1}^2$ y $\sigma_{X_2}^2$ desconocidas y tal que $\sigma_{X_1}^2 = \sigma_{X_2}^2 = \sigma^2$

Se puede definir una variable normal estándar

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Las variables aleatorias $\frac{(n_1-1){S_1}^2}{\sigma^2}$ y $\frac{(n_2-1){S_2}^2}{\sigma^2}$ tienen distribuciones ji cuadrada con $\nu_1 = n_1 - 1$ y $\nu_2 = n_2 - 1$ grados de libertad respectivamente. Estas variables aleatorias son independientes ya que las muestras se seleccionan independientemente de poblaciones distintas. La suma de estas variables define una variable aleatoria V que tiene una distribución ji cuadrado con $\nu = n_1 + n_2 - 2$ grados de libertad:

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

Las variables aleatorias Z y V son independientes y el cociente $\frac{Z}{\sqrt{\frac{V}{\nu}}}$ define una variable T con

distribución t de Student con $\nu = n_1 + n_2 - 2$ grados de libertad

$$T = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \Big/ \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}$$

Definimos $S_p^2 = \frac{(n_1-1){S_1}^2 + (n_2-1){S_2}^2}{n_1+n_2-2}$ y se obtiene el estadístico T

$$T = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sqrt{\frac{\sigma^2}{{S_p}^2}} \rightarrow T = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{{S_p}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Entonces:

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = P\left(-t_{\alpha/2} < \frac{(\overline{X_1} - \overline{X_2}) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{{S_p}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{\alpha/2}\right) = 1 - \alpha$$

$t_{\alpha/2}$ es el valor de t con $\nu = n_1 + n_2 - 2$ grados de libertad con un área igual a $\alpha/2$ a su derecha.

Esto conduce al siguiente intervalo de confianza del $(1-\alpha)100\%$ para la diferencia de medias $\mu_{X_1} - \mu_{X_2}$ en dos poblaciones normales con varianzas iguales pero de valor desconocido $\sigma_{X_1}^2 = \sigma_{X_2}^2 = \sigma^2$

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < \mu_{X_1} - \mu_{X_2} < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

IC(6)

Donde $\bar{x}_1$ y $\bar{x}_2$ son las medias obtenidas de muestras independientes de tamaños n_1 y n_2 tomadas de las poblaciones ($n_1 < 30$ y $n_2 < 30$).

Las desviaciones estándar s_1 y s_2 obtenidas de las muestras permiten obtener la estimación puntual del estimador común S_p^2: $s_p^2 = \frac{(n_1-1){s_1}^2 + (n_2-1)s^2}{n_1+n_2-2}$

El **error estándar estimado** de la estimación puntual es:

$$\widehat{se}(\bar{x}_1 - \bar{x}_2) = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Ejemplo 3

Se desea evaluar la eficacia de un índice numérico de diversidad de especies para indicar la degradación del agua debida al drenaje de ácido de una explotación minera en el cauce de un río. Desde el punto de vista conceptual; un alto índice de diversidad en las especies de macro invertebrados debe indicar un sistema de agua no contaminado, mientras que uno bajo, debe indicar un sistema de agua contaminada. Para este estudio, se seleccionan dos estaciones

independientes de muestreo, una localizada aguas arriba del punto de descarga y otra, aguas abajo. Para las 12 muestras recogidas aguas arriba el índice de diversidad de especies tuvo un valor promedio $\bar{x}_1 = 3{,}11$ *y una desviación estándar* $s_1 = 0{,}771$ *, mientras que en 10 muestras recogidas mensualmente aguas abajo, el valor del índice promedio fue* $\bar{x}_2 = 2{,}04$ *y la desviación estándar* $s_2 = 0{,}448$ *. Encuentra un intervalo de confianza para la diferencia entre las medias para los índices de diversidad aguas arriba y aguas debajo de la descarga de ácido. Asume que las poblaciones están distribuidas en forma aproximadamente normal y considera las varianzas iguales.*

La estimación puntual de la diferencia de medias, obtenida a partir de muestras de tamaño 12 y 10 , aguas arriba y aguas abajo, respectivamente es: $\bar{x}_1 - \bar{x}_2 = 3{,}11 - 2{,}14 = 1{,}07$

Las desviaciones estándar $s_1 = 0{,}771$ y $s_2 = 0{,}448$ obtenidas de las muestras permiten obtener la estimación puntual del estimador común S_p^2 :

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s^2}{n_1 + n_2 - 2} = \frac{11\ 0{,}711^2 + 9\ 0{,}448^2}{12+10-2} = 0{,}417$$

$(1 - \alpha)\ 100\% = 90\ \% \rightarrow grado\ de\ confianza \quad \alpha = 0{,}10 \rightarrow \frac{\alpha}{2} = 0{,}05$

$t_{0{,}05} = 1{,}725$ es el valor de t con $\nu = 12 + 10 - 2 = 20$ grados de libertad con un área igual a 0,05 a su derecha.

El intervalo de confianza del 90 % es

$$1{,}07 - 1{,}725\sqrt{0{,}417\left(\frac{1}{12} + \frac{1}{10}\right)} < \mu_{X_1} - \mu_{X_2} < 1{,}07 + 1{,}725\sqrt{0{,}417\left(\frac{1}{12} + \frac{1}{10}\right)}$$

$$0{,}593 < \mu_{X_1} - \mu_{X_2} < 1{,}547$$

Se tiene una confianza del 90 % que el intervalo (0,593 , 1,547) contiene la diferencia de las medias poblacionales para valores de los índices de diversidad de especies de las estaciones aguas arriba y aguas debajo del drenaje de ácido.

Como ambos límites de confianza son positivos podemos establecer, con el nivel de confianza del 90% , que en promedio , el índice para la estación ubicada aguas arriba del punto de descarga es mayor que el correspondiente a la localizada aguas abajo.

- Aún si las varianzas poblacionales son considerablemente diferentes, se obtienen resultados razonables cuando las poblaciones son normales y los muestras tienen igual tamaño.
- Desviaciones ligeras de la suposición de varianzas iguales y de normalidad no alteran el grado de confianza del intervalo.

Estimación de la diferencia entre dos medias para varianzas desconocidas y distintas

Dadas dos poblaciones aproximadamente normales con varianzas $\sigma_{X_1}^2$ y $\sigma_{X_2}^2$ desconocidas y tal que $\sigma_{X_1}^2 \neq \sigma_{X_2}^2$

Se puede definir el estadístico T' con una distribución aproximada t de Student con ν grados de libertad: