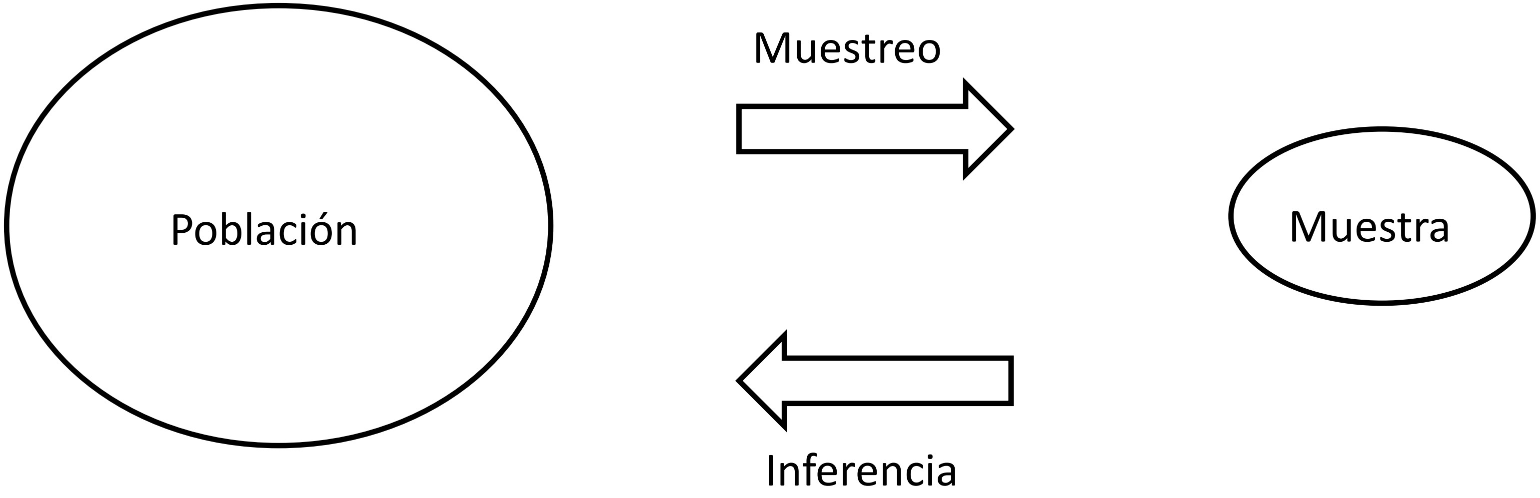
Inferencia estadística

La inferencia estadística consiste en aquellos métodos por medio de los cuales se realizan inferencias o generalizaciones acerca de una población.



Métodos de inferencia

Métodos

Clásico

Bayesiano

Problemas en inferencia

Problemas

- Estimación
  - Puntual
  - Por intervalo
- Prueba de hipótesis

Estimación puntual

Una estimación puntual de algún parámetro θ de la población es un valor $\hat{\theta}$ de una estadística $\widehat{\Theta}$.

Ejemplo:

Población

μ : media

p : proporción

.

.

.

Muestra

$\hat{x}$: media

$\hat{p}$: proporción

.

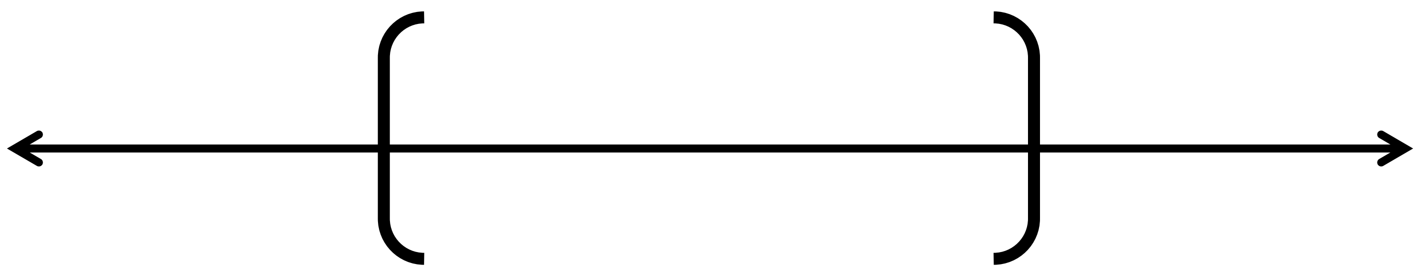
.

.

Estimación por intervalo

Una estimación por intervalo de un parámetro θ de la población es un intervalo de la forma $\hat{\theta}_{inf} < \theta < \hat{\theta}_{sup}$ donde $\hat{\theta}_{inf}$ y $\hat{\theta}_{sup}$ dependen de la estadística $\widehat{\Theta}$ y de la distribución de muestreo de $\widehat{\Theta}$.

θ



$\hat{\theta}_{inf}$ $\hat{\theta}_{sup}$

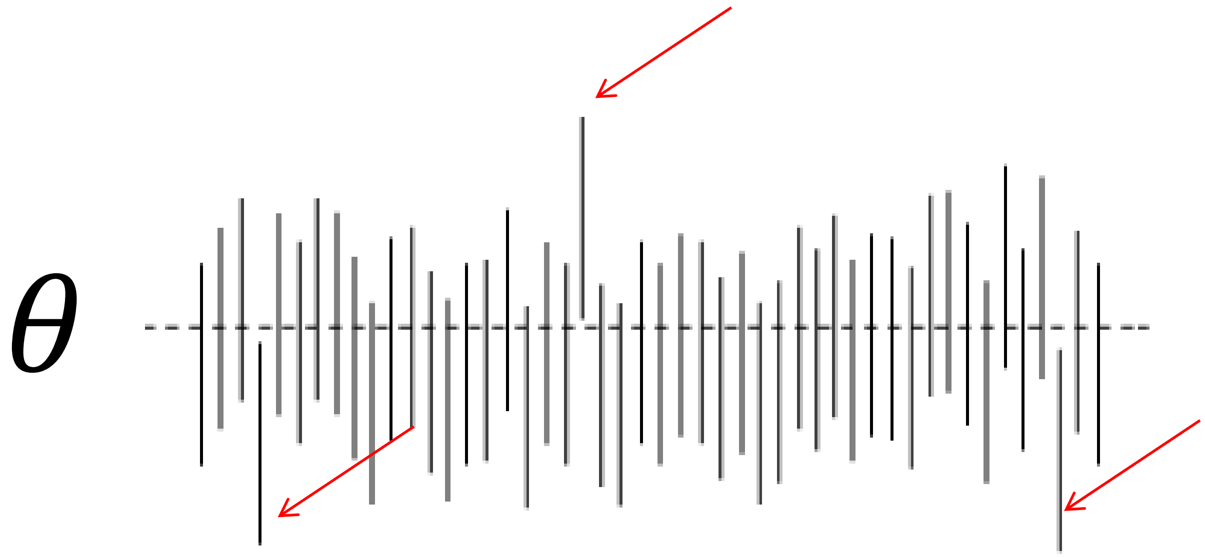
Nivel de confianza y significancia de una estimación por intervalo

$$\mathrm{NC} + \alpha = 1$$

$$\Downarrow \qquad\qquad \Downarrow$$

0.90	0.10
0.95	0.05
0.99	0.01

Representación del nivel de confianza



Ejemplo: si se construyen 100 intervalos de la forma $\hat{\theta}_{inf} < \theta < \hat{\theta}_{sup}$ con un nivel de confianza $NC = 0.97$ entonces se espera que 97 de los 100 intervalos contengan al parámetro θ.

Entendiendo lo que es un intervalo de confianza

http://www.youtube.com/watch?v=tFWsuO9f74o

Intervalos de confianza más utilizados

- Intervalos de confianza
  - Para una muestra
    - μ
    - p
    - σ^2
  - Para dos muestras
    - $\mu_1 - \mu_2$
    - $\mu_1 - \mu_2$ pareadas
    - $p_1 - p_2$
    - σ_1^2 / σ_2^2

Intervalos de confianza para una muestra

1. Intervalo de confianza para μ con σ conocida

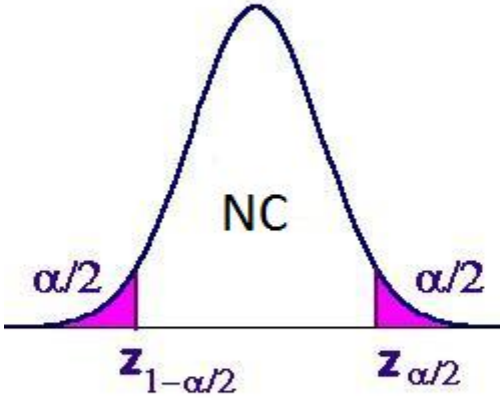
Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una población normal con varianza σ^2 conocida, un intervalo de confianza del $(1-\alpha)100\%$ para μ está dado por

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

donde $z_{\alpha/2}$ es el valor z que deja un área de $\alpha/2$ a la derecha.

Valores usuales de $z_{\alpha/2}$ para intervalos de confianza

Nivel de confianza	α	$z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576



Ejemplo 1

Las siguiente mediciones se registraron para el tiempo de secado, en horas, de cierta marca de pintura de látex:

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponga que las mediciones representan una muestra aleatoria de una población normal con una desviación de 1.2 minutos. Encuentre un intervalo de confianza para el tiempo medio de secado con una confianza del 95%.

Solución:

La muestra proviene de una distribución normal.

$\alpha = 0.05$

$n = 15$

$\bar{x} = 3.79$ minutos

$\sigma = 1.2$ minutos

Continuación ejemplo 1

Usando la fórmula

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

Reemplazando

$$3.79 - 1.96\frac{1.2}{\sqrt{15}} < \mu < 3.79 + 1.96\frac{1.2}{\sqrt{15}}$$

$$3.18 < \mu < 4.40$$

Con un nivel de confianza del 95% se espera que el tiempo de secado medio (μ) esté entre 3.18 horas y 4.40 horas.

De 100 intervalos construidos a partir de muestras de tamaño 15 se espera que 95 de ellos contengan a μ.

¿Cómo interpretar un intervalo de confianza?

http://www.youtube.com/watch?v=hP6flJdoIxc

2. Intervalo de confianza para μ con σ desconocida

Si $\bar{x}$ y s son la media y la desviación de una muestra aleatoria de tamaño n de una población normal con varianza desconocida, un intervalo de confianza del $(1-\alpha)100\%$ para μ está dado por

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

donde $t_{\alpha/2,n-1}$ es el valor t que deja un área de $\alpha/2$ a la derecha con $n-1$ grados de libertad.

Ejemplo 2

Una máquina produce piezas metálicas de forma cilíndrica. Se toma una muestra de las piezas y los diámetros son 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, y 1.03 centímetros. Encuentre un intervalo de confianza del 99% para el diámetro medio de las piezas de esta máquina, suponga una distribución aproximadamente normal

Solución:

La muestra proviene de una distribución normal.

$\alpha = 0.01$

$n = 9$

$\bar{x} = 1.0056$ cm

$s = 0.0246$ cm

Continuación ejemplo 2

Usando la fórmula

$$\bar{x} - t_{\alpha/_{2},n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/_{2},n-1}\frac{s}{\sqrt{n}}$$

Reemplazando

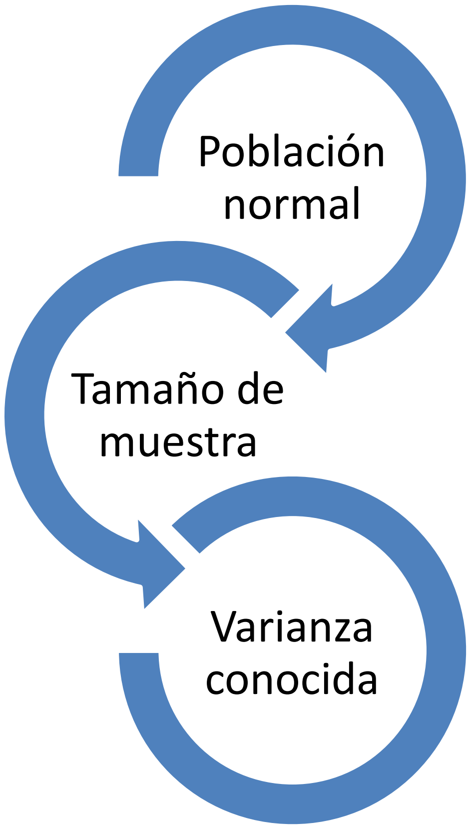
$$1.0056 - 3.355\frac{0.0246}{\sqrt{9}} < \mu < 1.0056 + 3.355\frac{0.0246}{\sqrt{9}}$$

$$0.9781 < \mu < 1.0331$$

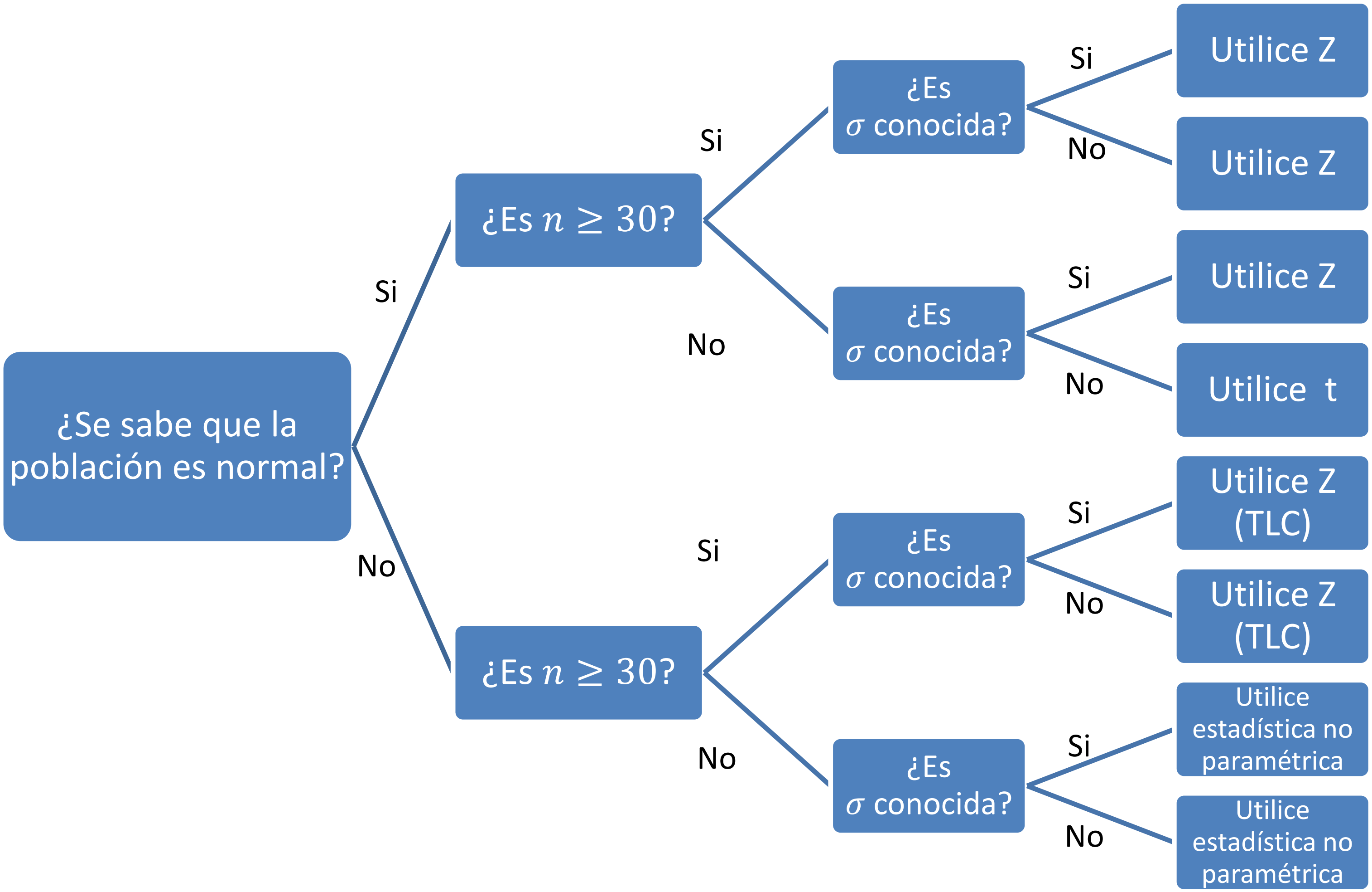
Con un nivel de confianza del 99% se espera que el diámetro medio (μ) esté entre 0.9781 cm y 1.0331 cm.

De 100 intervalos construidos a partir de muestras de tamaño 9 se espera que 99 de ellos contengan a μ.

Aspectos relevantes para construir intervalos de confianza para μ



Aspectos a tener en cuenta para construir IC para μ



¿Cómo determinar el tamaño de muestra para estimar μ?

http://www.youtube.com/watch?v=xSVS-0YR1JY

3. Intervalo de confianza para la proporción p

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, un intervalo de confianza aproximado de $(1-\alpha)100\%$ para p está dado por

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

donde $z_{\alpha/2}$ es el valor z que deja un área de $\alpha/2$ a la derecha.

Nota: usar solo cuando $n\hat{p} \geq 50$ y $n(1-\hat{p}) \geq 50$

Ejemplo 3

El gerente de una estación de televisión debe determinar en la ciudad qué porcentaje de casas tienen más de un televisor. Una muestra aleatoria de 500 casas revela que 275 tienen dos o más televisores. ¿Cuál es el intervalo de confianza del 90% para estimar la proporción de todas las casas que tienen dos o más televisores?

Solución:

$$\alpha = 0.10$$
$$n = 500$$
$$\hat{p} = \frac{275}{500} = 0.55$$

Continuación ejemplo 3

Usando la fórmula

$$\hat{p} - z\alpha_{/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z\alpha_{/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Reemplazando

$$0.55 - 1.645\sqrt{\frac{0.55(1-0.55)}{500}} < p < 0.55 + 1.645\sqrt{\frac{0.55(1-0.55)}{500}}$$

$$0.5134 < p < 0.5866$$

Con un nivel de confianza del 90% se espera que la proporción (p) de casas con dos o más televisores esté entre 0.5134 y 0.5866.

¿Cómo determinar el tamaño de muestra para estimar p?

http://www.youtube.com/watch?v=4G2kKHX5O8U

4. Intervalo de confianza para la varianza σ^2

Si s^2 es la varianza de una muestra aleatoria de tamaño n de una población normal, un intervalo de confianza aproximado de $(1-\alpha)100\%$ para σ^2 está dado por

$$\frac{(n-1)s^2}{x^2_{\alpha/2,\ n-1}} < \sigma^2 < \frac{(n-1)s^2}{x^2_{1-\alpha/2,\ n-1}}$$

donde los denominadores son obtenidos de una chi-cuadrada.

Nota: un intervalo de confianza para σ se puede obtener tomando $\sqrt{}$ a los límites del intervalo anterior.

Ejemplo 4

La duración en horas de un conjunto de bombillas se muestra a continuación:

1470 1510   1690 1740 1900 2000 2030 2010 2190
2200 2290 2380 2390 2480 2500 2580 2700

Construir un intervalo de confianza del 95% para la varianza poblacional σ^2.

Solución:

$\alpha = 0.05$
$n = 17$
$s^2 = 138098.5$

Continuación ejemplo 4

Usando la fórmula

$$\frac{(n-1)s^2}{x^2_{\alpha/2,\ n-1}} < \sigma^2 < \frac{(n-1)s^2}{x^2_{1-\alpha/2,\ n-1}}$$

Reemplazando

$$\frac{(17-1)138098.5}{28.845} < \sigma^2 < \frac{(17-1)138098.5}{6.908}$$

$$76601.7 < \sigma^2 < 319857.6$$

Con un nivel de confianza del 95% se espera que la varianza poblacional esté entre 76601.7 y 316558.2

Intervalos de confianza para dos muestras

5. Intervalo de confianza para la diferencia entre dos medias

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias muestrales de muestras aleatorias independientes de tamaño n_1 y n_2 de poblaciones normales con varianzas conocidas σ_1^2 y σ_2^2 , respectivamente, un intervalo de confianza de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ está dado por

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ejemplo 5

Se lleva a cabo un experimento en que se comparan dos tipos de motores, A y B. Se mide el rendimiento en kilómetros por galón de gasolina. Se realizan 50 experimentos con el motor A y 75 con el motor B. La gasolina que se utiliza y demás condiciones se mantienen constantes. El rendimiento promedio de gasolina para el motor A es de 36 Km/galón y el promedio para el motor B es de 42 Km/galón. Encuentre un intervalo de confianza de 96% sobre $\mu_B - \mu_A$, donde μ_B y μ_A son los rendimientos de gasolina poblacional para los motores B y A respectivamente. Suponga desviaciones estándar de seis y ocho para los motores A y B respectivamente.

Solución:

$$\alpha = 0.04$$

$$n_1 = 50 \qquad \bar{x}_1 = 36 \qquad \sigma_1 = 6$$

$$n_2 = 75 \qquad \bar{x}_2 = 42 \qquad \sigma_2 = 8$$

Continuación ejemplo 5

Usando la fórmula

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Reemplazando

$$(42 - 36) - 2.05\sqrt{\frac{8^2}{75} + \frac{6^2}{50}} < \mu_B - \mu_A < (42 - 36) + 2.05\sqrt{\frac{8^2}{75} + \frac{6^2}{50}}$$

$$3.43 < \mu_B - \mu_A < 8.57$$

Con un nivel de confianza del 96% se espera que $\mu_B - \mu_A$ esté entre 3.43 Km/galón y 8.57 Km/galón. **Como el 0 NO está en el intervalo hay evidencias de que μ_B es mayor que μ_A.**

6. Intervalo de confianza para la diferencia entre dos medias con varianzas poblacionales desconocidas

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias muestrales de muestras aleatorias independientes de tamaño n_1 y n_2 de poblaciones aproximadamente normal con varianzas iguales pero desconocidas, entonces un intervalo de confianza de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ está dado por

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2,v}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2,v}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

donde $v = n_1 + n_2 - 2$

$$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

7. Intervalo de confianza para la diferencia entre dos medias con varianzas poblacionales desconocidas

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias muestrales de muestras independientes de tamaño n_1 y n_2 de poblaciones aproximadamente normal con varianzas desconocidas y diferentes, entonces un intervalo de confianza de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ está dado por

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

donde

$$v = \frac{\left({s_1^2}/{n_1} + {s_2^2}/{n_2}\right)^2}{\dfrac{\left({s_1^2}/{n_1}\right)^2}{n_1 - 1} + \dfrac{\left({s_2^2}/{n_2}\right)^2}{n_2 - 1}}$$

Casos para intervalos de confianza de $\mu_1 - \mu_2$

- Casos
  - Varianzas conocidas
  - Varianzas desconocidas pero iguales
  - Muestras pequeñas con varianzas desconocidas y diferentes.

8. Intervalo de confianza para la diferencia de medias pareadas

Si $\bar{d}$ y s_d son la media y la desviación estándar de las diferencias distribuidas normalmente de n pares aleatorios de mediciones, entonces un intervalo de confianza de $(1-\alpha)100\%$ para $\mu_d = \mu_1 - \mu_2$ está dado por

$$\bar{d} - t_{\alpha/2,v}\frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{\alpha/2,v}\frac{s_d}{\sqrt{n}}$$

donde $v = n - 1$.

Ejemplo 8

Un estudio publicado en Chemoshere reporta los niveles de la toxina TCDD de 20 veteranos de Vietnam residentes en Massachusetts que se expusieron a un agente tóxico. Los niveles de la toxina en el plasma y en el tejido adiposo son dados en tabla siguiente. Construir un intervalo de confianza del 95%.

Table 9.1: Data for Example 9.13

Veteran	TCDD Levels in Plasma	TCDD Levels in Fat Tissue	d_i	Veteran	TCDD Levels in Plasma	TCDD Levels in Fat Tissue	d_i
1	2.5	4.9	−2.4	11	6.9	7.0	−0.1
2	3.1	5.9	−2.8	12	3.3	2.9	0.4
3	2.1	4.4	−2.3	13	4.6	4.6	0.0
4	3.5	6.9	−3.4	14	1.6	1.4	0.2
5	3.1	7.0	−3.9	15	7.2	7.7	−0.5
6	1.8	4.2	−2.4	16	1.8	1.1	0.7
7	6.0	10.0	−4.0	17	20.0	11.0	9.0
8	3.0	5.5	−2.5	18	2.0	2.5	−0.5
9	36.0	41.0	−5.0	19	2.5	2.3	0.2
10	4.7	4.4	0.3	20	4.1	2.5	1.6

Source: Schecter, A. et al. "Partitioning of 2,3,7,8-chlorinated dibenzo-*p*-dioxins and dibenzofurans between adipose tissue and plasma lipid of 20 Massachusetts Vietnam veterans," *Chemosphere*, Vol. 20, Nos. 7–9, 1990, pp. 954–955 (Tables I and II).

Continuación ejemplo 8

Calculando $\bar{d}$ y s_d

$$\bar{d} = -0.87. \qquad s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} = \sqrt{\frac{168.4220}{19}} = 2.9773.$$

Reemplazando

$$\bar{d} - t_{\alpha/2,v}\frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{\alpha/2,v}\frac{s_d}{\sqrt{n}}$$

$$-0.8700 - (2.093)\left(\frac{2.9773}{\sqrt{20}}\right) < \mu_D < -0.8700 + (2.093)\left(\frac{2.9773}{\sqrt{20}}\right)$$

$$-2.2634 < \mu_D < 0.5234$$

Como el intervalo de confianza del 95% incluye el cero se concluye que no existe diferencia entre la cantidad de toxina presente en el plasma y en el tejido adiposo.

9. Intervalo de confianza para la diferencia de proporciones

Si $\hat{p}_1$ y $\hat{p}_2$ son las proporciones de éxito de dos muestras aleatorias independientes de tamaño n_1 y n_2, entonces un intervalo de confianza de $(1-\alpha)100\%$ para $p_1 - p_2$ está dado por

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

en que $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$

Ejemplo 9

Se quiere determinar si un cambio en el método de fabricación de una piezas ha sido efectivo o no. Para esta comparación se tomaron 2 muestras, una antes y otra después del cambio en el proceso y los resultados obtenidos son los siguientes.

	Antes del cambio	Después del cambio
N° de piezas defectuosas	75	80
N° de piezas analizadas	1500	2000

Construir un intervalo de confianza del 90% para decidir si el cambio tuvo efecto positivo o no.

Solución:

$$\hat{p}_1 = 75/1500 = 0.05 \qquad \hat{p}_2 = 80/2000 = 0.04$$

Continuación ejemplo 9

Usando la fórmula

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Reemplazando $\hat{p}_1 = 75/1500 = 0.05$ $\hat{p}_2 = 80/2000 = 0.04$

se tiene que $$-0.0017 < p_1 - p_2 < 0.0217.$$

Con un nivel de confianza del 90% se espera que $p_1 - p_2$ esté entre -0.0017 y 0.0217.