Intervalos de Confianza para la diferencia de medias

INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS

Sean x_{11}, x_{12}, ... x_{1n1}, una muestra aleatoria de n_1 observaciones tomadas de una primera población con valor esperado μ_1, y varianza σ^2_1 ; y x_{21}, x_{22}, ... x_{2n2}, una muestra aleatoria de n_2 observaciones tomada de la segunda población con valor esperado μ_2 y varianza σ^2_2. Si $\overline{x}_1$ y $\overline{x}_2$ son las medias muestrales, la estadística $\overline{x}_1 - \overline{x}_2$ es un estimador puntual de $\mu_1 - \mu_2$, y tiene una distribución normal si las dos poblaciones son normales, o aproximadamente normal si cumple con las condiciones del teorema del limite central (tamaños de muestras relativamente grandes). Por lo tanto,

$$z = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Para calcular el intervalo de confianza para la diferencia de dos medias se debe saber si las varianzas poblacionales son conocidas o desconocidas, y en caso de que sean desconocidas, se debe probar si son iguales o diferentes. Cada uno de estos tres casos se analizarán por separado

Varianzas conocidas pero diferentes, $\sigma_1 \neq \sigma_2$

Si las varianzas poblacionales son conocidas y diferentes, los pasos a seguir para encontrar el intervalo de confianza son los siguientes:

a) El estadístico usado como estimador puntual de la diferencia de medias $\mu_1 - \mu_2$, será T = $\overline{x}_1 - \overline{x}_2$, que es un estimador suficiente

b) La variable aleatoria asociada con el estimador será la variable normal estándar dada por:

$$z = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

c) Para calcular el intervalo de confianza se debe tener en cuenta el nivel de confianza que se quiere considerar.

Teorema. Si $\bar{x}_1 - \bar{x}_2$ son las medias de dos muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones que tienen varianzas conocidas σ_1^2 y σ_2^2, respectivamente, entonces el intervalo de confianza para $\mu_1 - \mu_2$ es:

$$\bar{x}_1 - \bar{x}_2 - Z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + Z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ejemplo. Construya un intervalo de confianza del 94% para la diferencia real entre las duraciones de dos marcas de focos, si una muestra de 40 focos tomada al azar de la primera marca dio una duración media de 418 horas, y una muestra de 50 focos de otra marca dieron una duración media de 402 horas. Las desviaciones estándares de las dos poblaciones son 26 horas y 22 horas, respectivamente.

Solución. Tenemos que: $\bar{x}_1 = 418$, $\bar{x}_1 = 402$,

$\sigma_1 = 26, \sigma_2 = 22, n_1 = 40, n_2 = 50, Z = 1.88$

El intervalo de confianza es, entonces:

$$\bar{x}_1 - \bar{x}_2 - Z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + Z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(418 - 402) - 1.88\sqrt{\frac{26^2}{40} + \frac{22^2}{50}} \leq \mu_1 - \mu_2 \leq (418 - 402) + 1.88\sqrt{\frac{26^2}{40} + \frac{22^2}{50}}$$

$$6.3 \leq \mu_1 - \mu_2 \leq 25.7$$

Varianzas desconocidas e iguales $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$

Cuando las varianzas son desconocidas, se debe realizar previamente una prueba estadística para verificar si éstas son iguales o diferentes. Para hacerlo debemos hacer uso de la distribución F, bien sea mediante el cálculo de la probabilidad de que la muestra tomada provenga de dos poblaciones con varianzas iguales, o mediante el uso de un intervalo de confianza para la relación de dos varianzas, . Como se desconocen las varianzas de la población, se usan las varianzas de las muestras como estimadores.

El procedimiento a seguir para el cálculo del intervalo de confianza para la diferencia de dos medias será el siguiente:

a) El estadístico usado como estimador puntual de la diferencia de medias $\mu_1 - \mu_2$ será $\overline{x}_1 - \overline{x}_2$, que es un estimador suficiente.

b) La variable aleatoria asociada con el estimador será la variable definida como (se usa t en caso de muestras pequeñas):

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde s_p es un estimador combinado de las s^2, "mejor" que s_1^2, s_2^2 por separado, donde

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

c) Para calcular el intervalo de confianza se debe tener en cuenta el nivel de confianza que se quiere considerar y los grados de libertad que se calculan

$$g.l. = n_1 + n_2 - 2$$

De nuevo, manipulando la expresión anterior en forma similar al caso previo se llega al siguiente teorema que nos define el intervalo de confianza para la diferencia entre dos medias $\mu_1 - \mu_2$ con varianzas desconocidas pero iguales:

Teorema. Si x_1, x_2, s_1^2, s_2^2 son las medias y las varianzas de dos muestras aleatorias de tamaños n_1, n_2, respectivamente, tomadas de dos poblaciones normales e independientes con varianzas desconocidas pero iguales, entonces un intervalo de confianza para la diferencia entre medias $\mu_1 - \mu_2$ es:

$$\bar{x}_1 - \bar{x}_2 - t\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Ejemplo. La siguiente tabla presenta los resultados de dos muestras aleatorias para comparar el contenido de nicotina de dos marcas de cigarrillos.

	Marca A	Marca B
n_i	10	8
$\bar{X}_i$	3.1	2.7
S_i	0.5	0.7

Suponiendo que los conjuntos de datos provienen de muestras tomadas al azar de poblaciones normales con varianzas desconocidas e iguales, construya un intervalo de confianza del 95% para la diferencia real de nicotina de las dos marcas.

Solución. Como las varianzas son iguales, calculamos s_p^2 que está dado por:

$$s_p^2 = \frac{(9)0.5^2 + (7)0.7^2}{16} = 0.355 \quad \Rightarrow \quad s_p = 0.596$$

El intervalo de confianza del 95% está dado por (t(0.025,g.l.16) = 2.21):

$$\bar{x}_1 - \bar{x}_2 - t\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$3.1 - 2.7 - 2.21\,(0.596)\sqrt{\frac{1}{10} + \frac{1}{8}} \leq \mu_1 - \mu_2 \leq 3.1 - 2.7 + 2.21\,(0.596)\sqrt{\frac{1}{10} + \frac{1}{8}}$$

$$-0.2 \leq \mu_1 - \mu_2 \leq 1.0$$

Varianzas desconocidas y diferentes $\sigma_1^2 \neq \sigma_2^2$

a) El estadístico usado como estimador puntual de la diferencia de medias $\mu_1 - \mu_2$, será $\bar{x}_1 - \bar{x}_2$, que es un estimador suficiente

b) La variable aleatoria asociada con el estimador será la variable t definida como:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

c) El intervalo de confianza esta dado por el siguiente teorema, basado en la distribución t con n grados de libertad.

Teorema. . Si x_1, x_2, s_1^2, s_2^2 son las medias y las varianzas de dos muestras aleatorias de tamaños n_1, n_2, respectivamente, tomadas de dos poblaciones normales e independientes con varianzas desconocidas y diferentes, entonces un intervalo de confianza para la diferencia entre medias $\mu_1 - \mu_2$ es (nuevamente para el caso de muestras pequeñas):

$$\bar{x}_1 - \bar{x}_2 - t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Los grados de libertad están dados por:

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

Nota: el valor obtenido se redondea al entero más próximo.

Nota.

Si llevamos a cabo un cálculo de intervalo de confianza para diferencia de medias, suponiendo que las varianzas no son iguales, en el dado caso que sí lo fueran, perderíamos muy poco, y el intervalo obtenido sería un poco conservador.

El caso de que supongamos que las varianzas son iguales, siendo que no lo son, nos produce un error mayor que puede ser considerable por lo que una sugerencia es usar varianzas diferentes como regla general.

Problema. Cierto metal se produce, por lo común, mediante un proceso estándar. Se desarrolla un nuevo proceso en el que se añade una aleación a la producción del metal. Los fabricantes se encuentran interesados en estimar la verdadera diferencia entre las tensiones de ruptura de los metales producidos por los dos procesos. Para cada metal se seleccionan 12 ejemplares y cada uno de éstos se somete a una tensión hasta que se rompe.

La siguiente tabla muestra las tensiones de ruptura de los ejemplares, en kilogramos por centímetro cuadrado:

Proceso estándar	446	401	476	421	459	438	481	411	456	427	459	445
Proceso nuevo	462	448	435	465	429	472	453	459	427	468	452	447

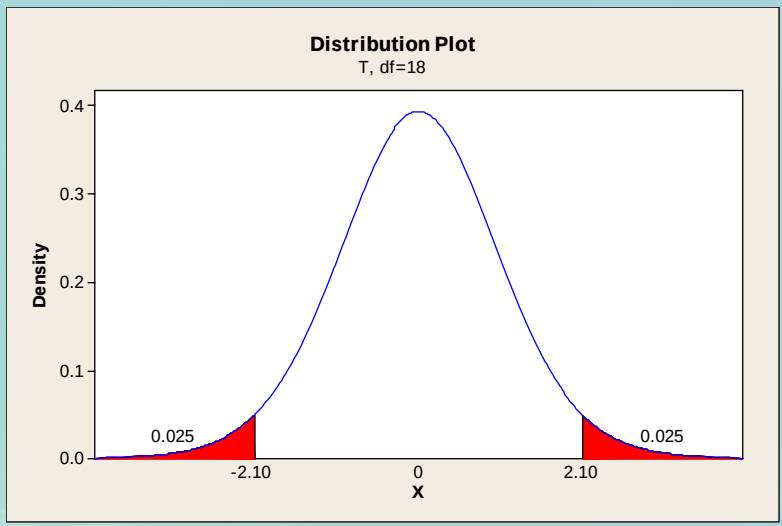
Si se supone que el muestreo se llevó a cabo sobre dos distribuciones normales e independientes, obtener los intervalos de confianza estimados del 95 y 99% para la diferencia entre los dos procesos. Interprete los resultados.

Solución:

Calculamos los valores que necesitamos.

n	Media	s
12	443.3	24.8
12	451.4	14.9

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left[\left(s_1^2/n_1\right)^2/(n_1-1)\right] + \left[\left(s_2^2/n_2\right)^2/(n_2-1)\right]} = 18$$



95% de confianza

$$t_1 = 2.10,\ t_2 = -2.10$$

$$\bar{x}_1 - \bar{x}_2 - t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

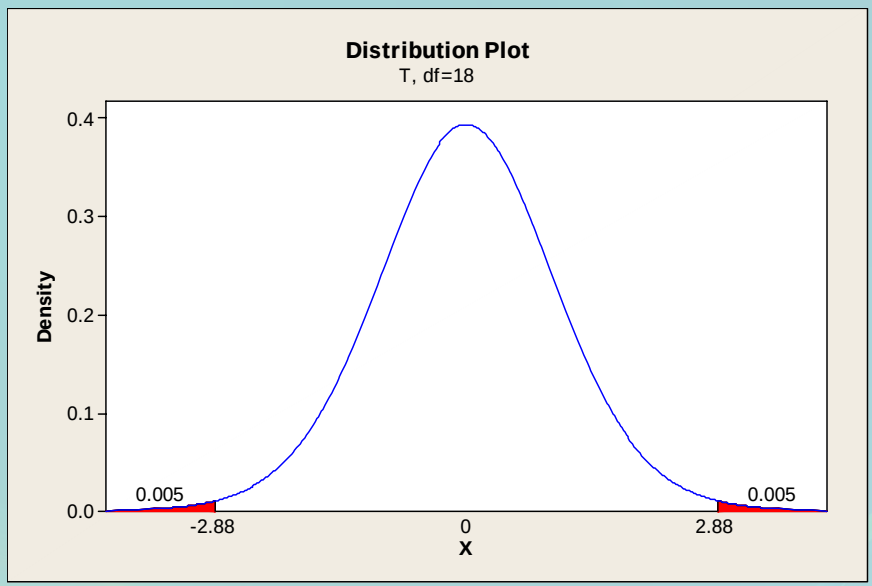
Por lo tanto:

$$(451.4-443.3)-2.10\sqrt{\frac{14.9^2}{12}+\frac{24.8^2}{12}} \le \mu_1-\mu_2 \le (451.4-443.3)+2.10\sqrt{\frac{14.9^2}{12}+\frac{24.8^2}{12}}$$

$$-25.65 \le \mu_1-\mu_2 \le 9.49$$

Y para 99% de confianza

$t_1 = 2.88,\ t_2 = -2.88$



$$(451.4-443.3)-2.88\sqrt{\frac{14.9^2}{12}+\frac{24.8^2}{12}} \le \mu_1-\mu_2 \le (451.4-443.3)+2.88\sqrt{\frac{14.9^2}{12}+\frac{24.8^2}{12}}$$

$$-32.16 \le \mu_1-\mu_2 \le 15.99$$