# EAA1520 - Inferencia Estadística

M. Ignacia Vicuña - Felipe Ossa - Ricardo Olea

2do Semestre 2023

Capítulo 3

# Contenidos

# Intervalos de confianza

## Introducción

- Una estimación puntual, no proporciona por sí mismo información alguna sobre la precisión y confiabilidad de la estimación.
- Una medida de incerteza de $\hat{\theta}$ es $\text{Var}(\hat{\theta})$, sin embargo, depende de $\theta$ el cual es desconocido. En la práctica, uno reporta la estimación de $\text{Var}(\hat{\theta})$: $\widehat{\text{Var}(\hat{\theta})}$.
- Otra alternativa, es calcular un **intervalo de confianza**.

La construcción de Intervalos de Confianza (IC) se sustenta en la siguiente propiedad:

"Todo estimador del parámetro desconocido $\theta$ es una variable aleatoria. por lo cual debe tener asociada alguna distribución de probabilidad"

Las distribuciones muestrales más usuales son: Normal, Chi-cuadrado, T-Student, Fisher.

# Intervalos de confianza

## Definición e interpretación

Supongamos que $Y_1, Y_2, ..., Y_n$ es una m.a de tamaño $n$ de la población Normal$(\mu, \sigma^2)$. Sabemos que el EMV de $\mu$ es $\bar{Y}$ el cual tiene distribución Normal$(\mu, \frac{\sigma^2}{n})$, entonces podemos decir

$$P\left(\bar{Y} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$$

Dado que $\bar{Y}$ es una v.a entonces el intervalo resulta ser un **Intervalo Aleatorio.**

Además como la inferencia clásica asume que el parámetro $\mu$ es constante, entonces dicha probabilidad se interpreta como:

"La probabilidad que el intervalo aleatorio $\left(\bar{Y} - 2\frac{\sigma}{\sqrt{n}}, \bar{Y} + 2\frac{\sigma}{\sqrt{n}}\right)$ contenga al parámetro $\mu$ es de un 95,44%"

# Intervalos de confianza

## Definición e interpretación

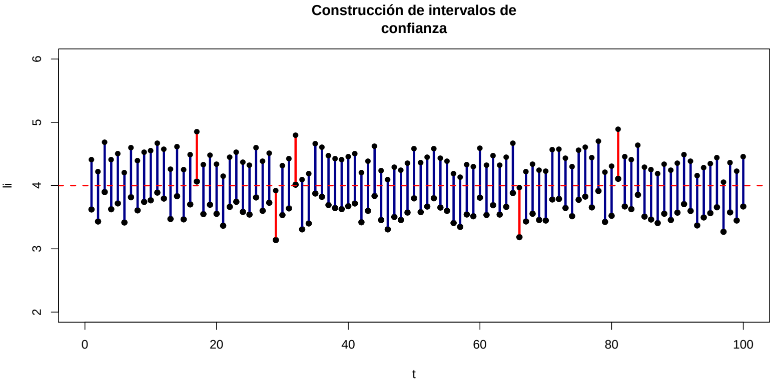
Bajo la interpretación frecuentista de probabilidad, afirmamos que:

"Cualesquiera sea el verdadero valor de $\mu$, de cada 100 intervalos aleatorios generados en forma equivalente al anterior, aproximadamente el 95% de ellos deberían contener la media $\mu$"

# Intervalos de confianza

## Definición e interpretación

Interpretación frecuentista,



# Intervalos de confianza

## Definición e interpretación

Supongamos que se extrae una muestra de tamaño $n = 36$, y se tiene que $\sigma = 3$ y $\bar{y} = 12.1$.

En tal caso, el intervalo resulta ser $(11.1, 13.1)$ el cual es una realización del intervalo aleatorio $(\bar{Y} - 1, \bar{Y} + 1)$ del ejemplo.

Note que es incorrecto afirmar que la probabilidad de que $\mu$ pertenezca al intervalo $(11.1, 13.1)$ es $95.44\%$

Sin embargo, la probabilidad del 95.44% del IC aleatorio induce una pseudo-probabilidad llamada Probabilidad Fiducial, o, en su expresión más común: **Confianza**

Sí entonces, es correcto afirmar que,

"Con una confianza del 95.44% se puede afirmar que el intervalo $(11.1, 13.1)$ contiene a la media $\mu$"

## Definición: Intervalo de Confianza

Sea $Y_1, Y_2, ..., Y_n$ una m.a de tamaño $n$ de la población $f(y, \theta)$. Se llama Intervalo de Confianza $(1-\alpha) \cdot 100\%$ para el parámetro desconocido $\theta$ a una realización de un intervalo aleatorio que con probabilidad de $(1-\alpha)$ contiene el parámetro $\theta$

Niveles de confianza usuales:

$$90\%, 95\%, 98\%$$

# Intervalos de confianza

## Definición e interpretación

Los IC pueden ser bilterales o unilaterales, por ejemplo, en el caso que el parámetro $\theta$ representa la media de una distribución entonces el formato de los respectivos IC será de la siguiente forma:

- $(\hat{\theta} - ee, \hat{\theta} + ee)$ Bilateral Simétrico
- $(\hat{\theta} - ee, \infty)$ Unilateral Inferior (izquierdo) o cota inferior
- $(-\infty, \hat{\theta} + ee)$ Unilateral Superior (derecho) o cota superior

$ee$ : error de estimación, error muestreo, precisión

Un método útil para construir los IC es el **Método Pivotal**

Los Pivotes son v.a construidas a partir de las distribuciones muestrales, los cuales tienen la siguiente propiedad:

- Es una función de los datos y del parámetro de interés
- La distribución de probabilidad dado el (los) parámetros no depende de dichos parámetros

Sea $Y_1, Y_2, ..., Y_n$ una m.a de tamaño $n$ de una población $f(y, \theta)$.

Sea la función $Q = q(Y_1, Y_2, ..., Y_n, \theta)$ la cual depende simultáneamente de la m.a y del parámetro $\theta$.

Si la distribución de $Q$ es independiente del parámetro $\theta$ entonces $Q$ es una **Cantidad Pivotal o Pivote**

## Ejemplos

Sea $Y_1, Y_2, ..., Y_n$ una m.a de tamaño $n$ de una población Normal$(\mu, \sigma^2)$. Interesa estimar el parámetro $\mu$, en este caso

- Un pivote para $\mu$ asumiendo $\sigma$ conocido es

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

- Un pivote para $\mu$ asumiendo $\sigma$ desconocido es

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \approx t(n-1)$$

Si se quiere estimar $\sigma^2$ y $\mu$ es desconocido, entonces un pivote para $\sigma^2$ es

$$W = \frac{(n-1)S^2}{\sigma^2} \approx \chi^2(n-1)$$

Sea $Y_1, Y_2, ..., Y_n$ una m.a de tamaño $n$ de la población $N(\mu, \sigma^2)$, entonces un IC$(1-\alpha)$% para $\sigma^2$ está dado por

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

donde $X^2_{1-\alpha/2}$ corresponde al percentil $1-\alpha/2$ de una distribución $\chi^2(n-1)$ y $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$

# Intervalos de confianza

## IC para $\mu$ con $\sigma$ conocido

Sea $Y_1, Y_2, ..., Y_n$ una m.a de tamaño $n$ de una población $\mathrm{N}(\mu, \sigma^2)$, entonces el IC de $(1-\alpha)\%$ para $\mu$ es

$$\left(\bar{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

donde $z_{1-\alpha/2}$ es el percentil $1-\alpha/2$ de una $N(0,1)$

# Intervalos de confianza

## IC para $\mu$ con $\sigma$ desconocido y $n$ pequeño

Sea $Y_1, Y_2, ..., Y_n$ una m.a de tamaño $n$ de una población $N(\mu, \sigma^2)$, entonces el IC de $(1-\alpha)\%$ para $\mu$ es

$$\left(\bar{y} - t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{y} + t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)$$

donde $t_{1-\alpha/2}(n-1)$ es el percentil $1-\alpha/2$ de una t-student$(n-1)$ y $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$

# Intervalos de confianza

## IC para $\mu$ con $\sigma$ desconocido y $n$ grande

Sea $Y_1, Y_2, ..., Y_n$ una m.a de tamaño $n$ de una población $\mathrm{N}(\mu, \sigma^2)$, para $n$ grande, la distribución t-student se aproxima por la $N(0,1)$, por lo tanto el IC de $(1-\alpha)\%$ para $\mu$ es

$$\left(\bar{y} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{y} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right)$$

## Nota importante:

De acuerdo al TLC para $n$ grande la Normalidad de la población no es necesaria, por lo tanto podemos levantar este supuesto, pudiendo así construir IC asintóticos para $\mu$ de otros modelos.

Sean $Y_1, ..., Y_n$ provenientes de una población con distribución $f(y, \theta)$. El EMV de $\theta$ distribuye asintóticamente $\text{N}(\theta, CCR(\theta))$. Un intervalo aproximado de nivel $(1-\alpha)\%$ para $\theta$ está dado por

$$\left(\hat{\theta} - z_{1-\alpha/2}\sqrt{CCR(\hat{\theta})}, \hat{\theta} + z_{1-\alpha/2}\sqrt{CCR(\hat{\theta})}\right)$$

## Ejemplos

| $f(y;\, \theta)$ | IC para $\mu$ |
|---|---|
| Ber($\pi$) | $\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}},\, p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$ |
| Poisson($\lambda$) | $\left(\bar{y} - z_{1-\alpha/2}\sqrt{\frac{\bar{y}}{n}},\, \bar{y} + z_{1-\alpha/2}\sqrt{\frac{\bar{y}}{n}}\right)$ |
| Exp($\lambda$) | $\left(\bar{y} - z_{1-\alpha/2}\frac{\bar{y}}{\sqrt{n}},\, \bar{y} + z_{1-\alpha/2}\frac{\bar{y}}{\sqrt{n}}\right)$ |

Sean $Y_1, ..., Y_n$ provenientes de una población con distribución $f(y, \theta)$. El EMV de $g(\theta)$ distribuye asintóticamente $\mathrm{N}\left(g(\theta), CCR(\theta)\cdot\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2\right)$. Un intervalo aproximado de nivel $(1-\alpha)\%$ para $g(\theta)$ está dado por

$$\left(g(\hat{\theta}) - z_{1-\alpha/2}\sqrt{CCR(\hat{\theta})\cdot\left(\frac{\partial g(\hat{\theta})}{\partial\theta}\right)^2}, g(\hat{\theta}) + z_{1-\alpha/2}\sqrt{CCR(\hat{\theta})\cdot\left(\frac{\partial g(\hat{\theta})}{\partial\theta}\right)^2}\right)$$

Por ejemplo, un IC Unilateral para $\mu$ con $\sigma$ desconocido y $n$ grande

| | |
|---|---|
| Unilateral Superior | $(-\infty, \bar{y} + z_{1-\alpha}\frac{S}{\sqrt{n}})$ |
| Unilateral Inferior | $(\bar{y} - z_{1-\alpha}\frac{S}{\sqrt{n}}, \infty)$ |

# Intervalos de confianza

## Toma de desiciones con IC

Sea $\theta$ el parámetro a estimar. A partir de IC se pueden evaluar supuestos acerca del parámetro $\theta$. Suponga que

- Interesa evaluar si $\theta = \theta_0$

  Un IC bilateral para $\theta$ de nivel de confianza $\alpha$ está dado por

  $$(\hat{\theta} - z_{1-\alpha/2}\sqrt{CCR(\hat{\theta})}, \hat{\theta} + z_{1-\alpha/2}\sqrt{CCR(\hat{\theta})})$$

  Si el valor de $\theta_0$ no está contenido en el IC bilateral, entonces con un nivel de significancia de $\alpha$ se puede concluir que $\theta \neq \theta_0$.

  De lo contrario, si el valor de $\theta_0$ está contenido en el IC bilateral, entonces con un nivel de significancia de $\alpha$ se concluye que $\theta = \theta_0$.

# Intervalos de confianza

## Toma de desiciones con IC

- Interesa evaluar si $\theta > \theta_0$

  Un IC unilateral inferior para $\theta$ de nivel de confianza $\alpha$ está dado por

  $$(\hat{\theta} - z_{1-\alpha}\sqrt{CCR(\hat{\theta})}, \infty)$$

  Si el valor de $\theta_0$ no está contenido en el IC unilateral, entonces con un nivel de significancia de $\alpha$ se puede concluir que $\theta > \theta_0$.

  De lo contrario, si el valor de $\theta_0$ está contenido en el IC unilateral, entonces con un nivel de significancia de $\alpha$ se conlcuye que $\theta \leq \theta_0$.

- Interesa evaluar si $\theta < \theta_0$

  Un IC unilateral superior para $\theta$ de nivel de confianza $\alpha$ está dado por

  $$(-\infty, \hat{\theta} + z_{1-\alpha}\sqrt{CCR(\hat{\theta})})$$

  Si el valor de $\theta_0$ no está contenido en el IC unilateral, entonces con un nivel de significancia de $\alpha$ se puede concluir que $\theta < \theta_0$.

  De lo contrario, si el valor de $\theta_0$ está contenido en el IC unilateral, entonces con un nivel de significancia de $\alpha$ se concluye que $\theta \geq \theta_0$.

¿Por qué establecer un nivel de confianza de 95% cuando es posible uno de 99%?

**Respuesta:** Por que el precio pagado por un nivel de confianza más alto es un intervalo más amplio.

De hecho, el único intervalo de confianza de 100% para $\mu$ es $(-\infty, \infty)$ el cual no es informativo, porque aún antes de un muestreo, sabemos que este intervalo cubre a $\mu$.

# Intervalos de confianza

## Precisión y selección del tamaño muestral

- Si consideramos que la longitud del intervalo especifica su precisión, entonces el nivel de confianza está inversamente relacionado con su precisión.
- Una estrategia consiste en especificar el nivel de confianza y la longitud del intervalo y luego determinar el tamaño muestral necesario.

## Determinación del tamaño muestral

Vimos que un intervalo de confianza para $\mu$ cuando $\sigma^2$ es conocido es

$$\left(\bar{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Denotemos por $A$ la amplitud del intervalo: $A = 2 \cdot z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 2 \cdot ee$ donde "*ee*" es el **Error de Estimación**.

Por lo tanto, para una precisión *ee* dada, es posible determinar el tamaño de muestra necesaria, con $\sigma$ y $\alpha$ fijos, dado por

$$n = \left( \frac{z_{1-\alpha/2} \sigma}{ee} \right)^2$$

En muchas oportunidades es interesante comparar las medias, proporciones o varianzas de dos poblaciones, por ejemplo:

- Las ventas medias mensuales de dos tiendas
- Las medias y varianzas de la producción de dos máquinas
- Efectividad de un nuevo medicamento, vía comparación antes y después del tratamiento

Sea $X_1, \ldots, X_n$ una muestra aleatoria con distribución Normal$(\mu_X, \sigma_x^2)$. Sea $Y_1, \ldots, Y_m$ una muestra aleatoria con distribución Normal$(\mu_Y, \sigma_y^2)$ independiente a la anterior.

## Varianzas cononocidas

Un estimador natural es $\bar{X} - \bar{Y}$ que el EMV de $\mu_X - \mu_Y$. Además,

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

Luego un IC para $\mu_X - \mu_Y$ es:

$$\left(\bar{X} - \bar{Y} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}, \bar{X} - \bar{Y} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}\right)$$

# Intervalos de confianza para dos muestras

IC para diferencia de medias con $\sigma_X^2 \neq \sigma_Y^2$

## Varianzas desconocidas

En el caso que las poblaciones tengan varianzas desconocidas, se tiene que

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim t_g$$

donde $g = \left[ \dfrac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}} \right]$

El intervalo de confianza para estimar $\mu_x - \mu_y$ está dado por

$$\left( \overline{X} - \overline{Y} - t_{g,1-\alpha/2}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}, \overline{X} - \overline{Y} - t_{g,1-\alpha/2}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}} \right)$$

$\sigma^2$conocido:

Ahora si las poblaciones tienen varianzas iguales tal que $\sigma_x^2 = \sigma_y^2 = \sigma^2$ entonces un IC para $\mu_X - \mu_Y$ es:

$$\left(\bar{X} - \bar{Y} - z_{1-\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X} - \bar{Y} + z_{1-\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}\right)$$

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

# Intervalos de confianza para dos muestras

## IC para diferencia de medias con $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

$\sigma^2$desconocido:

Se estima $\sigma^2$ a partir de los datos de las dos muestras,

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$$

donde,

$$\begin{aligned} S_X^2 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \\ S_Y^2 &= \frac{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{m-1} \end{aligned}$$

La función pivote está dada por:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{m+n-2}$$

Luego un IC para $\mu_X - \mu_Y$ es:

$$\left(\bar{X} - \bar{Y} - t_{n+m-2,1-\frac{\alpha}{2}} S_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X} - \bar{Y} + t_{n+m-2,1-\frac{\alpha}{2}} S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right)$$

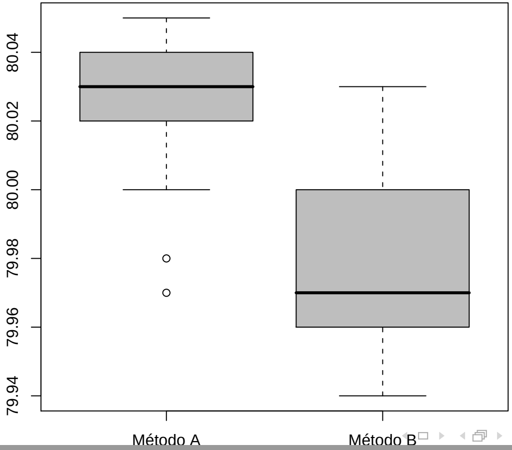
## Ejemplo

Dos métodos, $A$ y $B$, son usados para determinar el calor latente de fusión de hielo. Investigadores quieren saber si existe diferencia entre ambos métodos y ellos consideran que ambos métodos tienen la misma variabilidad.
La siguiente tabla contiene el cambio de calor total medido para hielo cuando su temperatura cambia de -0.72°C a 0°C.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Método A | 79.98 | 80.04 | 80.02 | 80.04 | 80.03 | 80.03 | 80.04 | 79.97 |
| | 80.05 | 80.03 | 80.02 | 80.00 | 80.02 | | | |
| Método B | 80.02 | 79.94 | 79.98 | 79.97 | 79.97 | 80.03 | 79.95 | 79.97 |

# Intervalos de confianza para dos muestras

## Ejemplo



## Intervalos de confianza para dos muestras

### Ejemplo

$\bar{X}_A = 80.02, \bar{X}_B = 79.98, S_A = 0.024, S_B = 0.031$

$S_p^2 = \frac{12S_A^2 + 7S_B^2}{19} = 0.0007178, S_p = 0.027$

Luego, $S_{\bar{X}_A - \bar{X}_B} = S_p\sqrt{\frac{1}{13} + \frac{1}{8}} = 0.012$

Un intervalo de 95% de confianza para $\mu_A - \mu_B$ está dado por:

$$(0.015, 0.065)$$

Poblaciones no necesariamente Normales, con varianzas desconocidas y distintas, muestras independientes y ambas de tamaño grande. Por TLC se tiene que un intervalo de confianza asintótico para $\mu_X - \mu_Y$ está dado por

$$\left(\bar{X} - \bar{Y} - z_{1-\alpha/2}\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}, \bar{X} - \bar{Y} + z_{1-\alpha/2}\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}\right)$$

Sean $X_1, ..., X_n$ provenientes de una población con distribución $f(x, \theta_1)$ y $Y_1, ..., Y_m$ provenientes de una población con distribución $f(y, \theta_2)$. El EMV de $\theta_1$ y $\theta_2$ distribuye asintóticamente $\mathrm{N}(\theta_1, CCR(\hat{\theta}_1))$ y $\mathrm{N}(\theta_2, CCR(\hat{\theta}_2))$ respectivamente. Un intervalo aproximado de nivel $(1-\alpha)\%$ para $\theta_1 - \theta_2$ está dado por

$$\left(\hat{\theta}_1 - \hat{\theta}_2 - z_{1-\alpha/2}\sqrt{CCR(\hat{\theta}_1) + CCR(\hat{\theta}_2)}, \hat{\theta}_1 - \hat{\theta}_2 + z_{1-\alpha/2}\sqrt{CCR(\hat{\theta}_1) + CCR(\hat{\theta}_2)}\right)$$

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria desde una población $N(\mu_X, \sigma_X^2)$ e $Y_1, Y_2, \ldots, Y_m$ una muestra aleatoria desde una población $N(\mu_Y, \sigma_Y^2)$ ambas muestras independientes.

Es posible construir un intervalo para comparar dos varianzas a través del cuociente de varianzas.

Luego la función pivote está dada por

$$\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n-1,m-1}$$

El intervalo de confianza para estimar $\frac{\sigma_X^2}{\sigma_Y^2}$ está dado por

$$\left(\frac{S_x^2}{S_y^2}\frac{1}{F_{n-1,m-1,1-\alpha/2}};\frac{S_x^2}{S_y^2}\frac{1}{F_{n-1,m-1,\alpha/2}}\right)$$