TEMA 5

ESTIMACIÓN POR INTERVALO

5.1 INTRODUCCIÓN

Utilizando los procedimientos del tema anterior es posible construir un buen estimador del parámetro θ, pero al particularizarlo con la muestra extraída y obtener la estimación puntual no sabemos si ésta se aproxima o no al verdadero valor. Este desconocimiento se debe a la aleatoriedad del muestreo realizado. Por este motivo, sería conveniente acompañar a la estimación puntual de alguna medida que revele el posible error.

Esta medida es un intervalo $(\theta_{inf}, \theta_{sup})$ donde confiamos se encuentre incluido el verdadero valor θ del parámetro. Este intervalo, denominado **intervalo de confianza**, tiene dos límites que son funciones de los datos de la muestra y, por lo tanto, son aleatorios. Esto implica que podamos medir inicialmente la probabilidad de que el intervalo aleatorio cubra al parámetro y a dicha probabilidad la denominamos **nivel de confianza: 1-α**.

$$P\{\theta_{inf} < \theta < \theta_{sup}\} = 1 - \alpha$$

El intervalo $(\theta_{inf}, \theta_{sup})$ es aleatorio dependiendo de los posibles valores de la muestra aleatoria y, por ese motivo, la expresión anterior se debe interpretar como la probabilidad de que el intervalo aleatorio incluya una constante que es el valor real del parámetro θ.

Para construir el intervalo de confianza existen diferentes métodos pero el más sencillo y utilizado es el método pivotal que explicamos a continuación.

5.2 MÉTODO PIVOTAL

Se basa en encontrar un estadístico pivote del parámetro, es decir, una función T que dependa del estimador y del parámetro θ de interés que verifique estas tres condiciones:

1. T es una función monótona respecto del parámetro θ.

2. La distribución de T es perfectamente conocida e independiente del parámetro θ, por lo tanto, podemos cuantificar cualquier probabilidad sobre ella.
3. El estadístico T es calculable dada la muestra aleatoria.

El razonamiento para encontrar el intervalo de confianza para θ es el siguiente:

Fijado el nivel de confianza 1-α por el investigador, dado que la distribución de T es conocida, podremos encontrar dos valores $\varepsilon_1(\alpha)$ y $\varepsilon_2(\alpha)$ tales que la probabilidad de que la variable T esté entre ellos es exactamente el nivel de confianza fijado:

$$P\{\varepsilon_1(\alpha) < T(X_1,...,X_n,\theta) < \varepsilon_2(\alpha)\} = 1-\alpha$$

Como la función T es monótona respecto de θ, podemos despejar el parámetro desconocido dentro de la probabilidad anterior, mediante manipulaciones algebraicas:

$$P\{\hat{\theta}_{inf}(X_1,...,X_n,\varepsilon_1(\alpha)) < \theta < \hat{\theta}_{sup}(X_1,...,X_n,\varepsilon_2(\alpha))\} = 1-\alpha$$

Cuando obtengamos la muestra como T era calculable entonces los extremos inferior y superior también son calculables y obtendremos el intervalo de confianza a un nivel 1-α:

$$IC_{1-\alpha}(\theta) = (\theta_{inf}(X_1,...,X_n,\varepsilon_1(\alpha)), \theta_{sup}(X_1,...,X_n,\varepsilon_2(\alpha)))$$

Un problema añadido es la elección de ε_1 y ε_2 porque no es única. El argumento utilizado para su determinación está relacionado con la amplitud del intervalo, lo que produce mayor precisión. Denotamos por A la amplitud del intervalo de confianza:

$$A = \theta_{sup}(X_1,...,X_n,\varepsilon_2) - \theta_{inf}(X_1,...,X_n,\varepsilon_1)$$

La elección se realiza bajo la premisa de mínima amplitud, que equivale a máxima precisión del intervalo,

$$\min_{\varepsilon_1,\varepsilon_2} A \quad \text{sujeto a} \quad P\{\varepsilon_1 < T < \varepsilon_2\} = 1-\alpha$$

5.3 DISTRIBUCIONES RELACIONADAS CON LOS INTERVALOS DE CONFIANZA

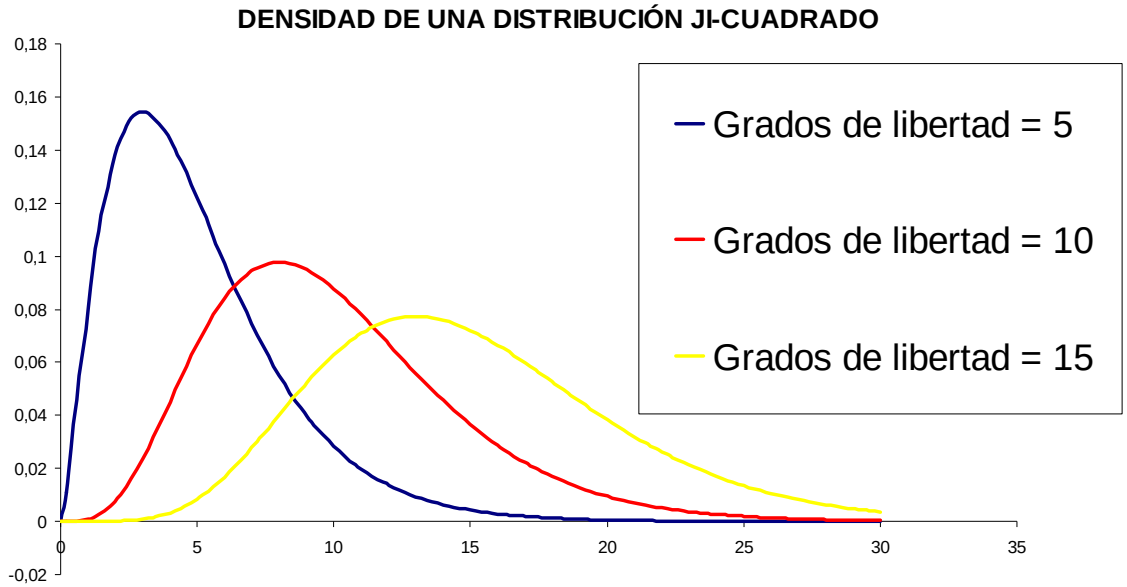
En este epígrafe presentamos dos distribuciones que vamos a utilizar para construir los intervalos de confianza notables.

5.3.1 Distribución Ji-cuadrado de Pearson

Una variable aleatoria X sigue una distribución Ji-cuadrado de Pearson con v grados de libertad si su función de densidad es:

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

Su representación gráfica viene dada por:



La media y la varianza de la distribución son:

$$E\left[X\right]=\nu$$
$$Var\left[X\right]=2\nu$$

El único parámetro de la distribución es v, que recibe el nombre de grados de libertad, que debe ser un número real positivo.

La distribución Ji-cuadrado se construye como una suma de cuadrados de variables aleatorias normales tipificadas e independientes y los grados de libertad son el número de sumandos, es decir, de normales independientes.

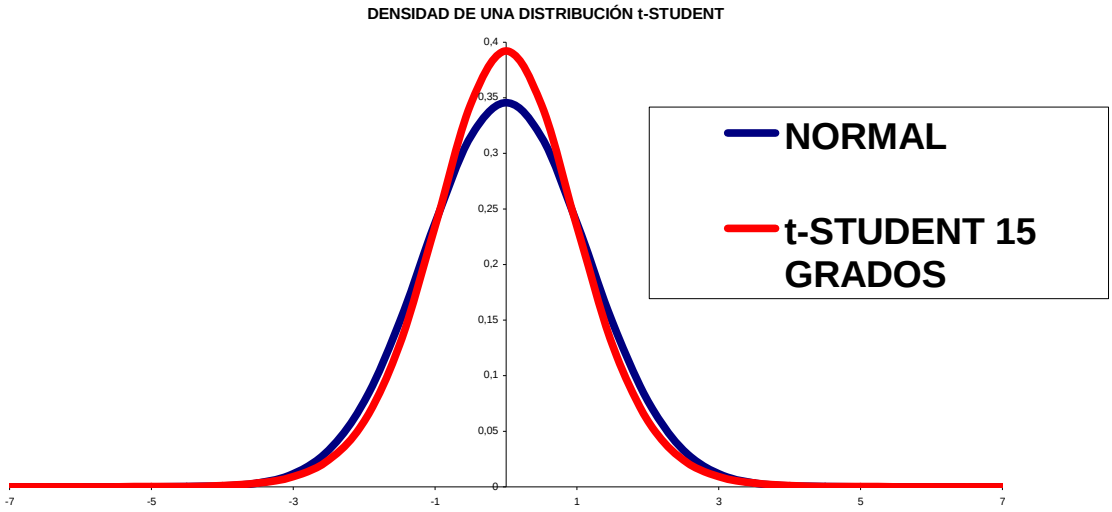
$$\text{Si}\quad X_1,X_2,...,X_\nu \sim i.i.d.\quad \mathrm{N}(0,1)\ \text{entonces}\quad X_1^2+X_2^2+\ldots+X_\nu^2\sim\chi_\nu^2$$

5.3.2 Distribución t-Student

Una variable aleatoria X sigue una distribución t de Student con v grados de libertad si su función de densidad es:

$$f_T(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{x^2}{\nu}\right]^{-\frac{\nu+1}{2}}$$

Su representación gráfica viene dada por:



La media y la varianza de la distribución son:

$$E[X] = 0$$

$$V[X] = \frac{\nu}{\nu - 2} \qquad \nu > 2$$

El único parámetro de la distribución es v, que recibe el nombre de grados de libertad, que debe ser un número real positivo.

La distribución t de Student se construye como el cociente entre una variable normal estándar y la raíz de una variable ji-cuadrado dividida por sus grados de libertad, cuando ambas variables aleatorias son independientes:

Sean X e Y variables aleatorias independientes

$$\begin{matrix} X \sim N(0,1) \\ Y \sim \chi^2_\nu \end{matrix} \Rightarrow T = \frac{X}{\sqrt{Y/\nu}} \sim t_\nu$$

Es una distribución simétrica respecto al cero, campaniforme pero con colas más pesadas.

5.3.3 Teorema de Fisher

En muchas ocasiones la variable aleatoria poblacional sigue una distribución normal y, por lo tanto, la muestra aleatoria simple seleccionada estará formada por variables aleatorias normales e independientes. Esto nos permitirá conocer la relación que existe entre los estimadores de la media y la varianza y sus distribuciones.

TEOREMA DE FISHER

Dada una muestra aleatoria simple extraída de una variable aleatoria normal de media μ y desviación típica σ, entonces se verifica que:

a) La media muestral $\overline{X}$ y la cuasivarianza muestral S_1^2 son dos variables aleatorias independientes.

b) La cuasivarianza muestral multiplicada por (n-1) y dividida por la varianza poblacional sigue una distribución Ji-cuadrado con n-1 grados de libertad:

$$\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{n-1}$$

Las consecuencias de este teorema son dos:

- El estadístico $\frac{(n-1)S_1^2}{\sigma^2}$ será utilizado como pivote para construir el intervalo de confianza para la varianza poblacional de una variable aleatoria normal.
- Al ser los dos estimadores (media y cuasivarianza muestral) independientes podemos usarlos para construir un nuevo pivote que tenga distribución t de Student:

$$\frac{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{1}{n-1}\frac{(n-1)S_1^2}{\sigma^2}}} = \frac{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}}{\frac{S_1}{\sigma}} = \frac{\overline{X}-\mu}{S_1}\sqrt{n} \sim t_{n-1}$$

Este pivote solo tiene un parámetro desconocido que es la media y la desviación típica ha sido estimada por la cuasidesviación típica, por lo tanto, podrá ser usado para construir el intervalo de confianza para la media de una variable aleatoria normal cuando su varianza poblacional es desconocida.

5.4 INTERVALOS DE CONFIANZA NOTABLES

A continuación presentaremos los casos más importantes y utilizados en el análisis de datos económicos. Nos centraremos básicamente en la media de una variable aleatoria y en la varianza sólo cuando la variable aleatoria poblacional siga una distribución normal.

A.1) Intervalo de confianza para la media de una normal con varianza conocida

Sea $(X_1,...,X_n)$ una m.a.s. extraída de una variable normal con media desconocida μ y desviación típica conocida σ. El estimador de la media poblacional es la media muestral $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Sabemos que la distribución de la media muestral, en este caso, es otra normal cuya media coincide con la poblacional y la desviación típica es la poblacional dividida por la raíz del tamaño muestral: $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Esto implica que si tipificamos podemos encontrar el pivote necesario para obtener el intervalo de confianza:

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

T es el pivote y cumple las tres condiciones necesarias. Así pues, fijado el nivel de confianza 1-α, podemos encontrar dos valores ε_1 y ε_2 en la tabla de la distribución normal tal que se cumpla:

$$P\left\{\varepsilon_1 < T = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} < \varepsilon_2\right\} = 1 - \alpha$$

A partir de esa expresión podemos despejar el parámetro de interés y obtenemos el intervalo de confianza:

$$P\left\{\varepsilon_1 \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < \varepsilon_2 \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

$$P\left\{-\overline{X} + \varepsilon_1 \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + \varepsilon_2 \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

$$P\left\{\overline{X} - \varepsilon_1 \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - \varepsilon_2 \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

$$IC_{1-\alpha}(\mu)=\left(\overline{X}-\varepsilon_2\frac{\sigma}{\sqrt{n}},\overline{X}-\varepsilon_1\frac{\sigma}{\sqrt{n}}\right)$$

En este caso concreto, la elección de los valores ε_1 y ε_2 para obtener mínima amplitud nos conduce a los valores simétricos de la distribución, es decir, $\varepsilon_1=-\varepsilon_2=-z_{\alpha/2}$. Esto implica que el intervalo de confianza para la media de una normal con varianza conocida a un nivel 1-α viene dado por:

$$IC_{1-\alpha}(\mu)=\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Este resultado es válido aunque la distribución de la variable aleatoria poblacional no tenga distribución normal siempre que el tamaño muestral n sea elevado porque el T.C.L nos permitiría aproximar la distribución del estimador media muestral a una normal. Por lo tanto, el I.C. calculado será válido para estimar la media de cualquier distribución de probabilidad siempre que el tamaño muestral sea elevado (n>30) y σ sea conocida.

A.2) I.C. para la media de una normal con varianza desconocida

Sea $(X_1,...,X_n)$ una m.a.s. extraída de una variable normal con media desconocida μ y desviación típica desconocida σ. El estimador de la media poblacional es la media muestral $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$. Sabemos que la distribución de la media muestral, en este caso, es otra normal cuya media coincide con la poblacional y la desviación típica es la poblacional dividida por la raíz del tamaño muestral: $\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\sim N(0,1)$. Sin embargo, esta función no puede ser pivote porque σ es desconocida, así pues tendremos que estimarla mediante la cuasidesviación típica S_1.

A.2.i) Si el tamaño muestral es elevado

La estimación de σ por la cuasidesviación típica S_1 no cambia la distribución del pivote porque es un estimador consistente y, por lo tanto, el error de muestreo será pequeño. El pivote que utilizaremos es el siguiente:

$$T=\frac{\overline{X}-\mu}{S_1}\sqrt{n}\approx N(0,1)$$

T es el pivote y cumple las tres condiciones necesarias. Así pues, fijado el nivel de confianza 1-α, podemos encontrar dos valores ε_1 y ε_2 en la tabla de la distribución normal tal que se cumpla:

$$P\left\{\varepsilon_1 < T = \frac{\overline{X}-\mu}{S_1}\sqrt{n} < \varepsilon_2\right\} = 1-\alpha$$

A partir de esa expresión podemos despejar el parámetro de interés y obtenemos el intervalo de confianza:

$$IC_{1-\alpha}(\mu) = \left(\overline{X} - \varepsilon_2 \frac{S_1}{\sqrt{n}}, \overline{X} - \varepsilon_1 \frac{S_1}{\sqrt{n}}\right)$$

En este caso concreto, la elección de los valores ε_1 y ε_2 para obtener mínima amplitud nos conduce a los valores simétricos de la distribución, es decir, $\varepsilon_1 = -\varepsilon_2 = -z_{\alpha/2}$. Esto implica que el intervalo de confianza para la media de una distribución cualquiera con varianza desconocida a un nivel 1-α viene dado por:

$$IC_{1-\alpha}(\mu) = \left(\overline{X} - z_{\alpha/2} \frac{S_1}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \frac{S_1}{\sqrt{n}}\right)$$

Este resultado es válido aunque la distribución de la variable aleatoria poblacional no tenga distribución normal siempre que el tamaño muestral n sea elevado porque el T.C.L nos permitiría aproximar la distribución del estimador media muestral a una normal. Por lo tanto, el I.C. calculado será válido para estimar la media de cualquier distribución de probabilidad siempre que el tamaño muestral sea elevado (n>30) y σ sea desconocida.

A.2.ii) Si el tamaño muestral es pequeño

El estimador cuasidesviación típica puede producir errores de estimación importantes y eso provoca que la distribución del pivote tenga que ser corregida. Utilizando el Teorema de Fisher y sus consecuencias, podemos concluir que:

$$T = \frac{\overline{X}-\mu}{S_1}\sqrt{n} \sim t_{n-1}$$

T es el pivote que cumple las tres condiciones necesarias y tiene una distribución t de Student con n-1 grados de libertad. Así pues, fijado el nivel de

confianza 1-α, podemos encontrar dos valores ε_1 y ε_2 en la tabla de la distribución t de Student tal que se cumpla:

$$P\left\{\varepsilon_1 < T = \frac{\overline{X}-\mu}{S_1}\sqrt{n} < \varepsilon_2\right\} = 1-\alpha$$

A partir de esa expresión podemos despejar el parámetro de interés y obtenemos el intervalo de confianza:

$$IC_{1-\alpha}(\mu) = \left(\overline{X} - \varepsilon_2 \frac{S_1}{\sqrt{n}}, \overline{X} - \varepsilon_1 \frac{S_1}{\sqrt{n}}\right)$$

En este caso concreto, la elección de los valores ε_1 y ε_2 para obtener mínima amplitud nos conduce a los valores simétricos de la distribución, es decir, $\varepsilon_1 = -\varepsilon_2 = -t_{n-1;\alpha/2}$. Esto implica que el intervalo de confianza para la media de una normal con varianza desconocida a un nivel 1-α viene dado por:

$$IC_{1-\alpha}(\mu) = \left(\overline{X} - t_{n-1;\alpha/2} \frac{S_1}{\sqrt{n}}, \overline{X} + t_{n-1;\alpha/2} \frac{S_1}{\sqrt{n}}\right)$$

B) Intervalo de confianza para la proporción p de una distribución Bernoulli con tamaño muestral elevado

Sea $(X_1,...,X_n)$ una m.a.s. extraída de una variable Bernoulli con proporción p desconocida. El estimador de la media poblacional es la media muestral, en este caso, proporción muestral: $\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Utilizando el teorema central del límite sabemos que la distribución aproximada de la media muestral, en este caso, es una normal cuya media coincide con la poblacional y la desviación típica es la poblacional dividida por la raíz del tamaño muestral: $\frac{\overline{X}-p}{\sqrt{p(1-p)}}\sqrt{n} \approx N(0,1)$. El pivote viene dado por:

$$T = \frac{\overline{X}-p}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n} \approx N(0,1)$$

T es el pivote y cumple las tres condiciones necesarias. Así pues, fijado el nivel de confianza 1-α, podemos encontrar dos valores ε_1 y ε_2 en la tabla de la distribución normal tal que se cumpla:

$$P\left\{\varepsilon_1 < T = \frac{\overline{X} - p}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n} < \varepsilon_2\right\} = 1-\alpha$$

A partir de esa expresión podemos despejar el parámetro de interés y obtenemos el intervalo de confianza:

$$IC_{1-\alpha}(p) = \left(\overline{X} - \varepsilon_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \overline{X} - \varepsilon_1\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

En este caso concreto, la elección de los valores ε_1 y ε_2 para obtener mínima amplitud nos conduce a los valores simétricos de la distribución, es decir, $\varepsilon_1 = -\varepsilon_2 = -z_{\alpha/2}$. Esto implica que el intervalo de confianza para la proporción de una distribución Bernoulli a un nivel 1-α viene dado por:

$$IC_{1-\alpha}(p) = \left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \overline{X} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

C) Intervalo de confianza para la varianza de una distribución normal

Sea $(X_1,...,X_n)$ una m.a.s. extraída de una variable normal con desviación típica desconocida σ. El estimador de la varianza poblacional es la cuasivarianza muestral $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Utilizando el teorema de Fisher sabemos que la distribución de la cuasivarianza muestral, en este caso, es una distribución ji-cuadrado con n-1 grados de libertad: $T = \frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{n-1}$.

T es el pivote y cumple las tres condiciones necesarias. Así pues, fijado el nivel de confianza 1-α, podemos encontrar dos valores ε_1 y ε_2 en la tabla de la distribución ji-cuadrado tal que se cumpla:

$$P\left\{\varepsilon_1 < T = \frac{(n-1)S_1^2}{\sigma^2} < \varepsilon_2\right\} = 1-\alpha$$

A partir de esa expresión podemos despejar el parámetro de interés y obtenemos el intervalo de confianza:

$$P\left\{\frac{\varepsilon_1}{(n-1)S_1^2} < \frac{1}{\sigma^2} < \frac{\varepsilon_2}{(n-1)S_1^2}\right\} = 1-\alpha$$

$$P\left\{\frac{(n-1)S_1^2}{\varepsilon_1} > \sigma^2 > \frac{(n-1)S_1^2\varepsilon_2}{\varepsilon_2}\right\} = 1-\alpha$$

$$\mathrm{IC}_{1-\alpha}\left(\sigma^2\right) = \left(\frac{(\mathrm{n}-1)\mathrm{S}_1^2}{\varepsilon_2}, \frac{(\mathrm{n}-1)\mathrm{S}_1^2}{\varepsilon_1}\right)$$

En este caso concreto, la elección de los valores ε_1 y ε_2 se realiza, por similitud a los casos anteriores, tomando los percentiles que dejan la misma probabilidad en la cola de la izquierda como en la cola de la derecha, es decir, $\varepsilon_1 = \chi^2_{\mathrm{n}-1;1-\alpha/2}$ y $\varepsilon_2 = \chi^2_{\mathrm{n}-1;\alpha/2}$. Esto implica que el intervalo de confianza para la varianza de una distribución normal a un nivel 1-α viene dado por:

$$\mathrm{IC}_{1-\alpha}\left(\sigma^2\right) = \left(\frac{(\mathrm{n}-1)\mathrm{S}_1^2}{\chi^2_{\mathrm{n}-1;\alpha/2}}, \frac{(\mathrm{n}-1)\mathrm{S}_1^2}{\chi^2_{\mathrm{n}-1;1-\alpha/2}}\right)$$