VARIANZA Y PROPORRCIONES

UNIDAD N° 3

ESTIMACIÓN DE UNA PROPORCIÓN

Un estimador puntual de la proporción *p* en un experimento binomial está dado por la estadística $\hat{P} = X/n$, donde *X* representa el número de éxitos en *n* pruebas.

La proporción de la muestra $\hat{p} = x/n$ se utilizará como el estimador puntual del parámetro p

Si la proporción *p* desconocida no está demasiado cerca de 0 o de 1, se puede establecer un intervalo de confianza para *p* al considerar la distribución muestral de $\hat{P}$

Por el teorema del límite central, para *n* suficientemente grande, $\hat{P}$ está distribuida aproximadamente normal con media y varianza:

$$\mu_{\hat{P}} = p \qquad y \qquad {\sigma_{\hat{P}}}^2 = \frac{pq}{n}$$

Intervalo de confianza para p de una muestra grande

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

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

Donde z_α es el valor *z* que deja un área de $\alpha/2$ a la derecha

Cuando *n* es pequeña y la proporción desconocida *p* se considera cercana a 0 o a 1, el procedimiento del intervalo de confianza que se establece aquí no es confiable y no se debe utilizar.

EJEMPLO

- EN UNA MUESTRA DE UNA CIUDAD DE N=500 FAMILIAS, SE ENCUENTRA QUE X=340 POSEEN TELEVISIÓN SATELITAL.
- ENCUENTRE UN INTERVALO DE CONFIANZA DE 95% PARA LA PROPORCIÓN REAL DE FAMILIAS EN ESTA CIUDAD QUE POSEEN TELEVISIÓN SATELITAL.

<u>SOLUCIÓN</u>

- LA ESTIMACIÓN PUNTUAL DE *P* ES

$$\hat{p} = \frac{340}{500} = 0,68$$

- SABIENDO QUE $\alpha = 0,05$, $z_{\alpha/2} = 1,96$
- POR LO TANTO, EL INTERVALO DE CONFIANZA DE 95% PARA *P* ES

$$0,68 - 1,96\sqrt{\frac{0,68 \cdot 0,32}{500}} < p < 0,68 + 1,96\sqrt{\frac{0,68 \cdot 0,32}{500}}$$

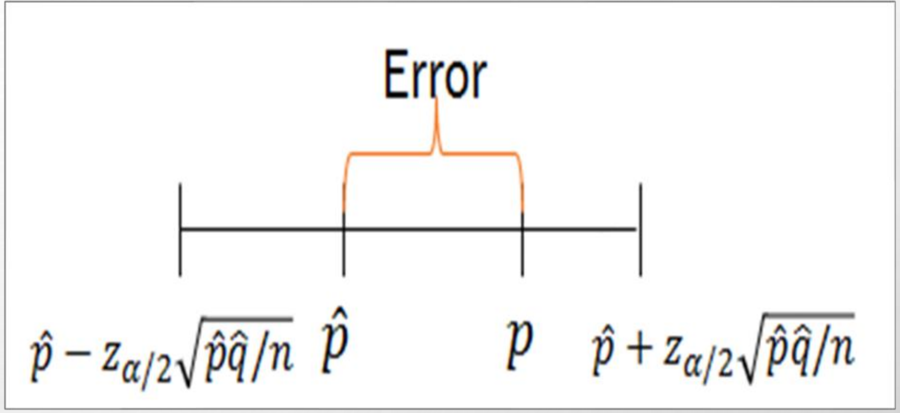
$$0,64 < p < 0,72$$

ESTIMACIÓN DE UNA PROPORCIÓN

Si p es el valor central de un intervalo de confianza de $(1-\alpha)100\%$ entonces $\hat{p}$ estima a p sin error

La mayor parte de las veces $\hat{p}$ no será exactamente igual a p y la estimación puntual es errónea

El tamaño de este error será la diferencia positiva que separa a $\hat{p}$ y p, y podemos tener una confianza del $(1-\alpha)100\%$ de que esta diferencia no excederá a $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$



Si $\hat{p}$ se utiliza como estimación de p, podemos tener una confianza del $(1-\alpha)100\%$ de que el error será menor que una cantidad específica e cuando el tamaño de la muestra es aproximadamente $n=\left(\frac{z_{\alpha/2}}{e}\right)^2\cdot\hat{p}\cdot\hat{q}$

Si se puede hacer una estimación cruda de p sin tomar una muestra, podemos usar este valor de $\hat{p}$ y después determinar n

A falta de tal estimación, podríamos tomar una muestra preliminar de tamaño $n\geq 30$ para proporcionar una estimación de p

EJEMPLO

- ¿QUÉ TAN GRANDE SE REQUIERE QUE SEA UNA MUESTRA (EN EL EJEMPLO ANTERIOR) SI QUEREMOS TENER UN 95% DE CONFIANZA DE QUE NUESTRA ESTIMACIÓN DE *P* ESTÁ DENTRO DE 0,02?

SOLUCIÓN

- CONSIDERAMOS QUE LAS 500 FAMILIAS COMPONEN UNA MUESTRA PRELIMINAR QUE PROPORCIONA UNA ESTIMACIÓN $\hat{p} = 0{,}68$, POR LO QUE TENEMOS:

$$n = \frac{{z_{\alpha/2}}^2 \hat{p}\hat{q}}{e^2} = \frac{1{,}96^2 \cdot 0{,}68 \cdot 0{,}32}{0{,}02^2} = 2089{,}83 \approx 2090$$

- POR LO TANTO, SI BASAMOS NUESTRA ESTIMACIÓN DE *P* SOBRE UNA MUESTRA DE 2090, PODEMOS TENER UNA CONFIANZA DE 95% DE QUE NUESTRA PROPORCIÓN MUESTRAL NO DIFERIRÁ DE LA PROPORCIÓN REAL POR MÁS DE 0,02

ESTIMACIÓN DE LA DIFERENCIA ENTRE DOS PROPORCIONES

Considere el problema donde deseamos estimar la diferencia entre dos parámetros binomiales p_1 y p_2
Nuestro problema es estimar la diferencia entre estas dos proporciones

Para ello seleccionamos muestras aleatorias independientes de tamaño n_1 y n_2 a partir de las poblaciones binomiales con medias $n_1 p_1$ y $n_2 p_2$ y varianzas $n_1 p_1 q_1$ y $n_2 p_2 q_2$ respectivamente, y formar las proporciones $\hat{p}_1 = x_1/n_1$ y $\hat{p}_2 = x_2/n_2$

Un estimador puntual de la diferencia entre las dos proporciones, $p_1 - p_2$, está dado por la estadística $\hat{P}_1 - \hat{P}_2$
Por tanto, la diferencia de las proporciones muestrales, $\hat{p}_1 - \hat{p}_2$ se utilizará como la estimación puntual de $p_1 - p_2$

Al elegir muestras independientes de las dos poblaciones, las variables $\hat{P}_1$ y $\hat{P}_2$ serán independientes y está distribuida de forma aproximadamente normal con media y varianza:

$$\mu_{\hat{P}_1-\hat{P}_2} = p_1 - p_2 \quad y \quad {\sigma^2}_{\hat{P}_1-\hat{P}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

INTERVALO DE CONFIANZA PARA P_1-P_2 DE UNA MUESTRA GRANDE

Si $\hat{p}_1$ y $\hat{p}_2$ son las proporciones de éxitos en muestras aleatorias de tamaño n_1 y n_2, respectivamente, $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$, un intervalo de confianza aproximado de $(1-\alpha)100\%$ para la diferencia de dos parámetros binomiales $p_1 - p_2$, está dado por:

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

Donde $z_{\frac{\alpha}{2}}$ es el valor de z que deja un área de $\frac{\alpha}{2}$ a la derecha

EJEMPLO

- SE CONSIDERA CIERTO CAMBIO EN UN PROCESO DE FABRICACIÓN DE PARTES COMPONENTES.
- SE TOMAN MUESTRAS DEL PROCEDIMIENTO EXISTENTE Y DEL NUEVO PARA DETERMINAR SI ÉSTE TIENE COMO RESULTADO UNA MEJORÍA.
- SI SE ENCUENTRA QUE 75 DE 1500 ARTÍCULOS DEL PROCEDIMIENTO ACTUAL SON DEFECTUOSOS Y 80 DE 2000 ARTÍCULOS DEL PROCEDIMIENTO NUEVO TAMBIÉN LO SON, ENCUENTRE UN INTERVALO DE CONFIANZA DE 90% PARA LA DIFERENCIA REAL EN LA FRACCIÓN DE DEFECTUOSOS ENTRE EL PROCESO ACTUAL Y EL NUEVO

EJEMPLO: SOLUCIÓN

- SEAN P_1 Y P_2 LAS PROPORCIONES REALES DE DEFECTUOSOS PARA LOS PROCESOS ACTUAL Y NUEVO, RESPECTIVAMENTE. DE AQUÍ, $\hat{p}_1 = \frac{75}{1500} = 0{,}05$ Y $\hat{p}_2 = \frac{80}{2000} = 0{,}04$ Y LA ESTIMACIÓN PUNTUAL DE $p_1 - p_2$ ES

$$\hat{p}_1 - \hat{p}_2 = 0{,}05 - 0{,}04 = 0{,}01$$

- PARA UNA REGIÓN DE $\frac{\alpha}{2} = 0{,}05$ TENEMOS $z_{0{,}05} = 1{,}64$. POR TANTO, AL SUSTITUIR EN ESTA FÓRMULA OBTENEMOS EL INTERVALO DE CONFIANZA DE 90%

$$0{,}01 - 1{,}64\sqrt{\frac{0{,}05 \cdot 0{,}95}{1500} + \frac{0{,}04 \cdot 0{,}96}{2000}} < p_1 - p_2 < 0{,}01 + 1{,}64\sqrt{\frac{0{,}05 \cdot 0{,}95}{1500} + \frac{0{,}04 \cdot 0{,}96}{2000}}$$

- QUE SE SIMPLIFICA A $-0{,}0017 < p_1 - p_2 < 0{,}0217$. COMO EL INTERVALO CONTIENE EL VALOR 0, NO HAY RAZÓN PARA CREER QUE EL NUEVO PROCEDIMIENTO PRODUCIRÁ UNA DISMINUCIÓN SIGNIFICATIVA EN LA PROPORCIÓN DE ARTÍCULOS DEFECTUOSOS COMPARADO CON EL MÉTODO EXISTENTE

ESTIMACIÓN DE LA VARIANZA

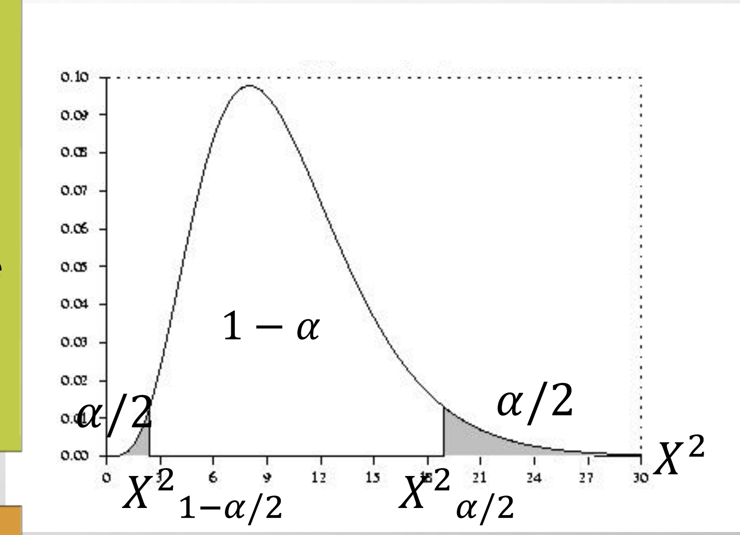
Si se extrae una muestra de tamaño *n* de una población normal con varianza σ^2, y se calcula la varianza muestral s^2 obtenemos un valor de estadística S^2
Esta varianza muestral calculada se usará como estimación puntual de σ^2
Por ello la estadística S^2 se llama estimador de σ^2

Se puede establecer una estimación por intervalos de σ^2 mediante el uso de la estadística

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

La estadística X^2 tiene una distribución ji cuadrada con $n-1$ grados de libertad cuando las muestras se eligen de una población normal

$$P\left({X^2}_{1-\alpha/2} < X^2 < {X^2}_{\alpha/2}\right) = 1-\alpha$$



${X^2}_{1-\alpha/2}$ y ${X^2}_{\alpha/2}$ son valores de la distribución ji cuadrada con $n-1$ grados de libertad, que dejan áreas de $1-\alpha/2$ y $\alpha/2$, respectivamente, a la derecha

$$P\left({X^2}_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < {X^2}_{\alpha/2}\right) = 1-\alpha$$

REVISIÓN: DISTRIBUCIÓN JI CUADRADA

Otro caso especial de la distribución gamma se obtiene al hacer $\alpha = \frac{\nu}{2}$ y $\beta = 2$, donde ν es un entero positivo.

Este resultado se llama ***distribución ji cuadrada***.

La distribución tiene un solo parámetro, ν, llamado ***grados de libertad***

La variable aleatoria continua X tiene una distribución ji cuadrada, con ν grados de libertad, si su función de densidad está dada por

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\,\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} & x > 0 \\ 0 & en\ cualquier\ otro\ caso \end{cases}$$

Donde ν es un entero positivo

La media y la varianza de la distribución ji cuadrada son

$$\mu = \nu \quad \text{y} \quad \sigma^2 = 2\nu$$

INTERVALO DE CONFIANZA PARA σ^2

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

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

Donde $X^2{}_{\alpha/2}$ y $X^2{}_{1-\alpha/2}$ son los valores X^2 con $v = n - 1$ grados de libertad, que dejan áreas de $\alpha/2$ y $1-\alpha/2$, respectivamente a la derecha

Un intervalo de confianza del $(1-\alpha)100\%$ para σ se obtiene al tomar la raíz cuadrada de cada extremo del intervalo para σ^2

EJEMPLO

- LOS SIGUIENTES SON LOS PESOS, EN DECAGRAMOS, DE 10 PAQUETES DE SEMILLAS DE PASTO DISTRIBUIDAS POR CIERTA COMPAÑÍA: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2 Y 46.0. ENCUENTRE UN INTERVALO DE CONFIANZA DE 95% PARA LA VARIANZA DE TODOS LOS PAQUETES DE SEMILLAS DE PASTO QUE DISTRIBUYE ESTA COMPAÑÍA, SUPONGA UNA POBLACIÓN NORMAL.

<u>SOLUCIÓN</u>

- PRIMERO ENCONTRAMOS LA VARIANZA: $S^2 = 0{,}286$
- CON EXCEL ES NECESARIO USAR LA DISTRIBUCIÓN GAMMA, CONSIDERANDO QUE $\alpha = \nu/2$ Y $\beta = 2$ (=DISTR.GAMMA.INV(PROBABILIDAD;α; β)

$$X^2{}_{0{,}025} = 19{,}023 \text{ Y } X^2{}_{0{,}975} = 2{,}700$$

- POR LO TANTO, EL INTERVALO DE CONFIANZA DE 95% PARA σ^2 ES:

$$\frac{9 \cdot 0{,}286}{19{,}023} < \sigma^2 < \frac{9 \cdot 0{,}286}{2{,}700}$$

- O SIMPLEMENTE $0{,}135 < \sigma^2 < 0{,}953$

DOS MUESTRAS: ESTIMACIÓN DE LA RAZÓN DE DOS VARIANZAS

Una estimación puntual de la razón de dos varianzas poblacionales $\frac{\sigma_1{}^2}{\sigma_2{}^2}$ está dada por la razón $\frac{s_1{}^2}{s_2{}^2}$ de las varianzas muestrales. De aquí, la estadística $\frac{S_1{}^2}{S_2{}^2}$ se denomina estimador de $\frac{\sigma_1{}^2}{\sigma_2{}^2}$

Si $\sigma_1{}^2$ y $\sigma_2{}^2$ son las varianzas de poblaciones normales, podemos establecer una estimación por intervalos $\frac{\sigma_1{}^2}{\sigma_2{}^2}$ mediante el uso de la estadística

$$F = \frac{\sigma_2{}^2 S_1{}^2}{\sigma_1{}^2 S_2{}^2}$$

La variable aleatoria *F* tiene una distribución *F* con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Por lo que podemos escribir:

$$P\left[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)\right] = 1 - \alpha$$

Donde $f_{1-\alpha/2}(v_1, v_2)$ y $f_{\alpha/2}(v_1, v_2)$ son los valores de la distribución *F* con v_1 y v_2 grados de libertad, que dejan áreas de $1 - \alpha/2$ y $\alpha/2$, respectivamente, a la derecha

DISTRIBUCIÓN F

La distribución *F* se usa en situaciones de dos muestras para extraer inferencias acerca de las varianzas de población.

Esta distribución también se aplica a muchos otros tipos de problemas en los que las varianzas muestrales están involucradas.

La estadística *F* se define como la razón entre dos variables aleatorias ji cuadrada independientes, dividida cada una entre su número de grados de libertad. De aquí podemos escribir

$$F = \frac{U/\upsilon_1}{V/\upsilon_2}$$

Donde *V* y *U* son variables aleatorias independientes que tienen distribuciones ji cuadrada con υ_1 y υ_2 grados de libertad, respectivamente.

A partir de esta definición se puede definir la distribución muestral *F* (no lo analizaremos en esta instancia)

Sea f_α por arriba del cual encontramos un área igual a α, considerando los diferentes grados de libertad υ_1 y υ_2

Al escribir $f_\alpha(\upsilon_1, \upsilon_2)$ para f_α con υ_1 y υ_2 grados de libertad, obtenemos

$$f_{1-\alpha}(\upsilon_1, \upsilon_2) = \frac{1}{f_\alpha(\upsilon_1, \upsilon_2)}$$

Si ${S_1}^2$ y ${S_2}^2$ son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones normales con varianzas ${\sigma_1}^2$ y ${\sigma_2}^2$, respectivamente, entonces

$$F = \frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_2}^2} = \frac{{\sigma_2}^2{S_1}^2}{{\sigma_1}^2{S_2}^2}$$

Tiene una distribución *F* con $\upsilon_1 = n_1 - 1$ y $\upsilon_2 = n_2 - 1$ grados de libertad

INTERVALO DE CONFIANZA PARA σ_1^2/σ_2^2

Si ${s_1}^2$ y ${s_2}^2$ son varianzas de muestras independientes de tamaño n_1 y n_2, respectivamente, de poblaciones normales, entonces un intervalo de confianza de $(1-\alpha)100\%$ para $\frac{{\sigma_1}^2}{{\sigma_2}^2}$ es:

$$\frac{{s_1}^2}{{s_2}^2}\frac{1}{f_{\alpha/2}(v_1,v_2)} < \frac{{\sigma_1}^2}{{\sigma_2}^2} < \frac{{s_1}^2}{{s_2}^2} f_{\alpha/2}(v_2,v_1)$$

Donde $f_{\alpha/2}(v_1,v_2)$ es un valor *f* con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad que deja un área $\alpha/2$ a la derecha $f_{\alpha/2}(v_2,v_1)$ es un valor *f* similar con $v_2 = n_2 - 1$ y $v_1 = n_1 - 1$ grados de libertad

Un intervalo de confianza del $(1-\alpha)100\%$ para σ_1/σ_2 se obtiene al tomar la raíz cuadrada de cada extremo del intervalo para σ_1^2/σ_2^2

EJEMPLO

- SE REALIZA UNA INVESTIGACIÓN EN DOS ESTACIONES DISTINTAS PARA DETERMINAR LA CANTIDAD DE UN QUÍMICO EN UN RÍO. SE REUNIERON 15 MUESTRAS DE LA ESTACIÓN 1 Y 12 MUESTRAS DE LA ESTACIÓN 2. 15 MUESTRAS DE LA ESTACIÓN 1 TUVIERON UN CONTENIDO PROMEDIO DEL QUÍMICO DE 3,84 MILIGRAMOS POR LITRO Y UNA DESVIACIÓN ESTÁNDAR DE 3,07 MILIGRAMOS POR LITRO, MIENTRAS QUE LAS 12 MUESTRAS DE LA ESTACIÓN 2 TUVIERON UN CONTENIDO PROMEDIO DE 1,49 MILIGRAMOS POR LITRO Y UNA DESVIACIÓN ESTÁNDAR DE 0,80 MILIGRAMOS POR LITRO.

- SUPONIENDO QUE LAS VARIANZAS DE LA POBLACIÓN SON DIFERENTES, JUSTIFIQUE ESTA SUPOSICIÓN MEDIANTE LA CONSTRUCCIÓN DE UN INTERVALO DE CONFIANZA DEL 98% PARA σ_1^2/σ_2^2 Y PARA σ_1/σ_2, DONDE σ_1^2 Y σ_2^2 SON LAS VARIANZAS POBLACIONALES DEL CONTENIDO DEL QUÍMICO EN LA ESTACIÓN 1 Y EN LA ESTACIÓN, RESPECTIVAMENTE.

EJEMPLO: SOLUCIÓN

- TENEMOS QUE $n_1 = 15$, $n_2 = 12$, $s_1 = 3{,}07$ Y $s_2 = 0{,}80$
- PARA UN INTERVALO DE CONFIANZA DEL 98%, $\alpha = 0{,}02$
- UTILIZANDO COMANDO DE EXCEL (=DISTR.F.INV(VALOR DE PROBABILIDAD; GRADOS LIBERTAD 1; GRADOS LIBERTAD 2))

$$f_{0,01}(14{,}11) \cong 4{,}30 \text{ Y } f_{0,01}(11{,}14) \cong 3{,}87$$

- POR LO TANTO, EL INTERVALO DE CONFIANZA DE 98% PARA σ_1^2/σ_2^2

$$\frac{3{,}07^2}{0{,}80^2}\frac{1}{4{,}30} < \frac{{\sigma_1}^2}{{\sigma_2}^2} < \frac{3{,}07^2}{0{,}80^2}\ 3{,}87$$

- QUE SE SIMPLIFICA A $3{,}425 < \frac{{\sigma_1}^2}{{\sigma_2}^2} < 56{,}991$
- AL CALCULAR LAS RAÍCES CUADRADAS DE LOS LÍMITES DE CONFIANZA, ENCONTRAMOS UN INTERVALO DE CONFIANZA DE 98% PARA σ_1/σ_2 ES

$$1{,}851 < \frac{\sigma_1}{\sigma_2} < 7{,}549$$

- COMO ESTE INTERVALO NO PERMITE LA POSIBILIDAD DE QUE $\frac{\sigma_1}{\sigma_2}$ SEA IGUAL A 1, ES CORRECTO SUPONER QUE $\sigma_1 \neq \sigma_2$ O ${\sigma_1}^2 \neq {\sigma_2}^2$