

Clase del lunes 10 de mayo de 2004
Favor de realizar las siguientes rutinas.
Los datos están en la página web se llaman collage.mtw

7.4 Inferencia acerca de la Varianza Poblacional.

Para hacer inferencia acerca de la varianza de una población Normal se requiere hacer uso de la distribución Ji-Cuadrado, la cuál será explicada brevemente antes de discutir la inferencia.

7.4.1. La Distribución Ji-Cuadrado

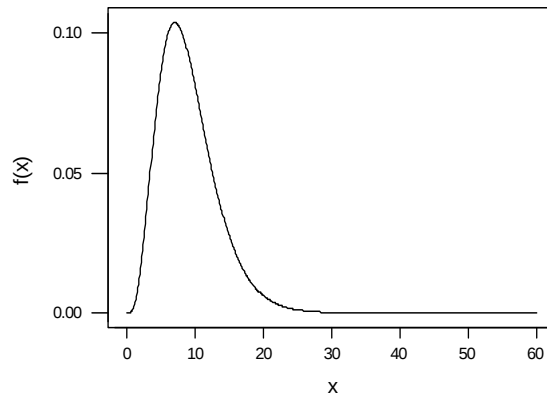
Sean $X_1, X_2, \dots, X_n$ observaciones de una muestra de tamaño n de una población normal $N(\mu, \sigma^2)$. Entonces:

$$\chi^2 = \frac{\sum_{i=1}^n (x - \bar{x})^2}{\sigma^2}$$

Se distribuye como una Ji-Cuadrado (χ^2) con **n-1 grados de libertad**. La distribución Ji-Cuadrado no es simétrica, pero a medida que los grados de libertad aumentan se va observando más simetría. En la figura 7.13 se muestra la gráfica de una χ^2 con 9 grados de libertad.

Se puede mostrar que el cuadrado de una normal estandarizada es una Ji-Cuadrado con un grado de libertad y que si se suma dos variables Ji-Cuadrado independientemente distribuidas, entonces se obtiene otra Ji-Cuadrado cuyos grados de libertad es igual a la suma de los grados de libertad de los otros dos.

Ji-Cuadrado con 9 grados de libertad



Hecho por Edgar Acuna

Figura 7.13. Gráfica de una Ji-Cuadrado con 9 grados de libertad

Recordando que la fórmula de la varianza muestral es $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$, se obtiene que

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Se acostumbra usar la notación $\chi^2_{(m)}$ para representar a una distribución Ji-Cuadrado con m grados de libertad.

Usos de la Ji-Cuadrado

- a) Para hacer inferencias acerca de la varianza poblacional. Es decir, para calcular Intervalos de Confianza y Prueba de hipótesis para la varianza poblacional.
- b) Para hacer pruebas de Bondad de Ajuste. O sea para probar si un conjunto de datos sigue una distribución pre-determinada.
- c) Para hacer análisis de tablas de contingencia

En este capítulo sólo se discutirá el primer uso, los otros dos se discutirán en el capítulo 8.

7.4.2 Intervalos de Confianza para la Varianza Poblacional

Partiendo de la siguiente fórmula, la cual puede justificarse con una gráfica:

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

Aquí $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ representan los valores de una Ji-Cuadrado con: n-1 grados de libertad tal que el área a la izquierda de dichos valores son $\alpha/2$ y $1 - \alpha/2$ respectivamente. Se puede llegar a establecer que un intervalo de confianza del $100(1-\alpha)\%$ para la varianza poblacional σ^2 de una población normal es de la forma:

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

Ejemplo 7.9 Los siguientes datos representan espesor de la membrana del plasma (medido en angstroms) de 20 especies de una planta:

80 90 85 82 75 58 70 84 87 81 87 61 73 84 85 70 78 95 77 52

Hallar un intervalo de confianza del 95 % para la varianza poblacional.

Solución:

MINITAB da este intervalo de confianza cuando se escoge **Graphical Summary** en la ventana de diálogo que aparece al oprimir **Graph**, después de haber elegido la secuencia **STAT** ▶ **Basic Statistics Display** ▶ **Descriptive Statistics**. Para los datos del ejemplo se obtienen los resultados que aparecen en la siguiente figura:

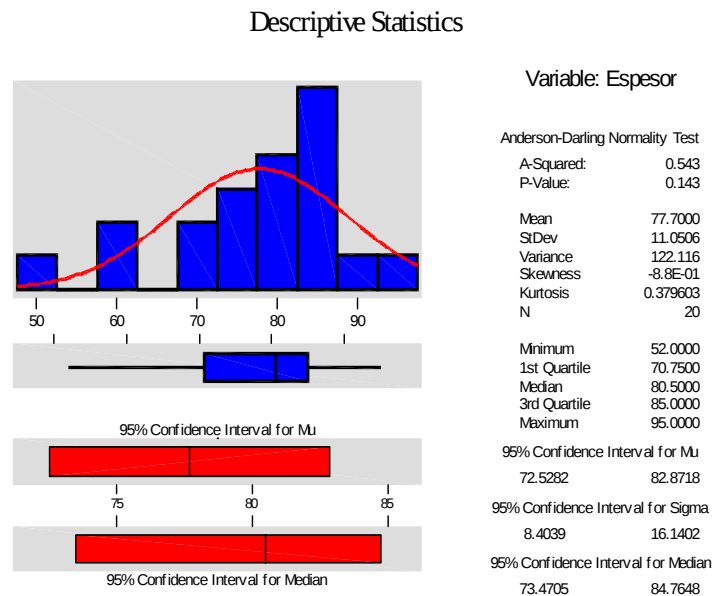


Figura 7.15. Resultados de **Graphical Summary** para el ejemplo 7.9

Interpretación:

Casi al final de la columna de cálculos se lee que un **intervalo de confianza del 95 % para sigma** es (8.4039, 16.1402). **Si se cuadra ambos valores se obtiene el intervalo de confianza para la varianza**, y se concluye de que hay un 95% de confianza de que la varianza del espesor de la membrana del plasma de todas las especies caen entre 70.6253 y 260.507.

7.5 Comparando la varianza de dos poblaciones

Supongamos que se tienen dos poblaciones normales con varianzas desconocidas σ_1^2 y σ_2^2 . Si de la primera población se toma una muestra de tamaño m que tiene una varianza muestral s_1^2 y de la segunda población se toma una muestra, independiente de la primera, de tamaño n que tiene una varianza muestral s_2^2 , se puede mostrar que la razón $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ se distribuye como una **F con m-1 grados de libertad en el numerador y n-1**

en el denominador. Esta es la base para la prueba de F de igualdad de varianza entre dos grupos. Las fórmulas para las pruebas de hipótesis son como sigue:

Caso I	Caso II	Caso III
$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$
$H_a: \sigma_1^2 < \sigma_2^2$	$H_a: \sigma_1^2 \neq \sigma_2^2$	$H_a: \sigma_1^2 > \sigma_2^2$
Prueba Estadística:		
$F = \frac{s_1^2}{s_2^2} \quad \text{con m-1 g.l. en el numerador y n-1 g.l en el denominador}$		
Decisión:		
Si $F_{cal} < F_{\alpha}$ entonces se rechaza H_0	Si $F_{cal} < F_{\alpha/2}$ o $F_{cal} > F_{1-\alpha/2}$ se rechaza H_0	Si $F_{cal} > F_{1-\alpha}$ se rechaza H_0

Nota: Se desea comprobar si dos grupos tienen varianzas iguales. Mas adelante se hará una prueba de 2 sample t para ver si el promedio o media son iguales. Préstale mucha atención.

MINITAB hace pruebas de igualdad de varianza de dos o más grupos. Para esto se selecciona la opción **2 Variances** del submenú **Basic Statistics** del menú **STAT**. Otra posibilidad es elegir **Homogeneity of Variances Test** del submenú **ANOVA** del menú **STAT**.

Ejemplo 7.11. En el siguiente ejemplo se trata de comparar las varianzas de los puntajes de aprovechamiento de los estudiantes de escuelas públicas y privadas. Los datos recolectados son:

Est	aprovech	escuela
1	580	pública
2	638	pública
3	642	privada
4	704	pública
5	767	privada

6	641	privada
7	721	privada
8	625	privada
9	694	pública
10	615	pública
11	617	pública
12	623	pública
13	689	privada
14	689	pública

Solución:

Las hipótesis son las siguientes:

H_0 : Varianza de los puntajes de estudiantes de escuela pública = varianza de puntajes de escuela privada.

H_a : Las varianzas no son iguales.

La ventana de diálogo de **2 Variances** se completará como lo muestra la figura 7.16. Oprimiendo el botón **Options** se puede elegir el nivel de confianza y poner un título a la gráfica que aparecerá:

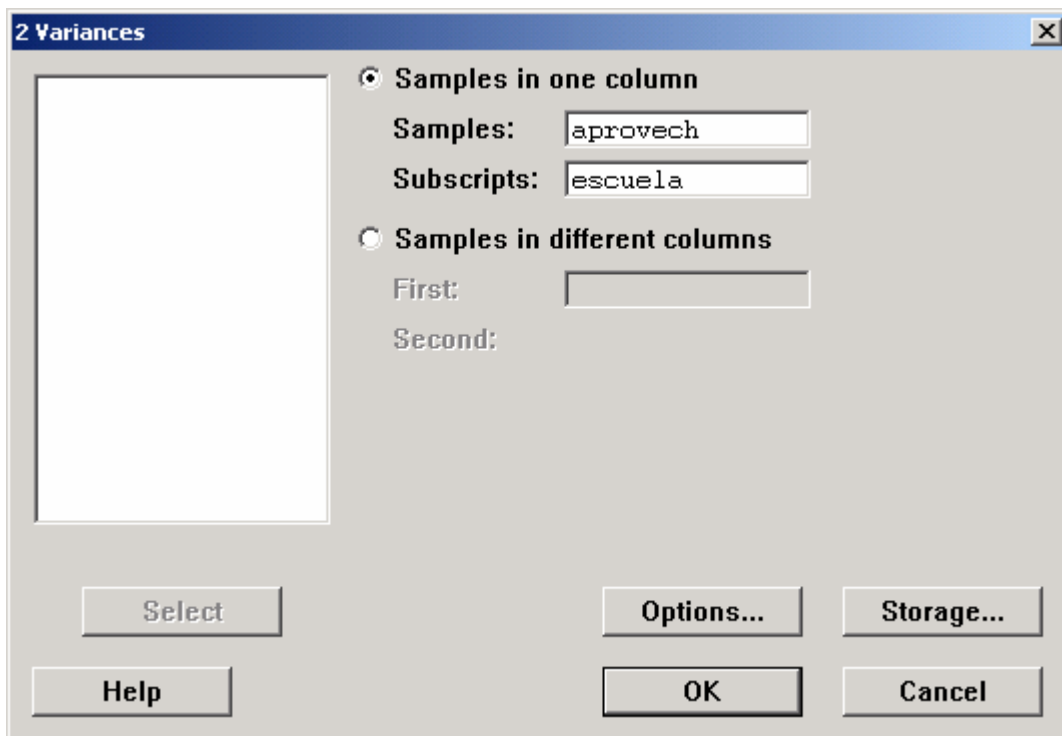


Figura 7.16. Ventana de diálogo de **2 variances** para el ejemplo 7.11

La ventana **session** mostrará los siguientes resultados:

```
MTB > %Vartest 'aprovech' 'escuela';
SUBC> Confidence 95.0;
```

```
SUBC> Title "Prueba de Igualdad de Varianza".
Executing from file: C:\PROGRAM FILES\MTBWIN\MACROS\Vartest.MAC
Macro is running ... please wait

Test for Equal Variances

Response      aprovech
Factors       escuela
ConfLvl       95.0000

Bonferroni confidence intervals for standard deviations

  Lower      Sigma      Upper    N  Factor Levels
32.4522    55.3477    158.347  6  privada
28.2368    45.1347    103.380  8  pública

F-Test (normal distribution)

Test Statistic: 1.504
P-Value       : 0.601

Levene's Test (any continuous distribution)

Test Statistic: 0.300
P-Value       : 0.594
```

Además aparecen una gráfica mostrando los intervalos de confianza para cada una de las desviaciones estándar y una comparación de la variabilidad de cada muestra, como aparece en la figura 7.17.

Interpretación: El “P-value” de la prueba de F es .601 mucho mayor que .05, luego se acepta la hipótesis nula y se concluye que los puntajes en la prueba de aprovechamiento en las escuelas pública y privada tienen igual varianza. De las gráficas se puede ver que los “boxplots” de ambos grupos tienen aproximadamente el mismo alargamiento.

Quiere decir que la variabilidad entre los grupos es la misma.

7.6 Comparación entre dos medias Poblacionales usando muestras independientes

Supongamos que se tiene dos poblaciones distribuidas normalmente con medias desconocidas μ_1 y μ_2 , respectivamente. Se puede aplicar una prueba t de Student para comparar las medias de dichas poblaciones basándonos en dos muestras independientes tomadas de ellas. La primera muestra es de tamaño m, con media $\bar{x}$ y varianza s_1^2 y la segunda muestra es de tamaño n, tiene media $\bar{y}$ y varianza s_2^2 .

Si las varianzas de las poblaciones son iguales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) entonces se puede mostrar que:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

se distribuye como una t con m+n-2 grados de libertad. En este caso la varianza poblacional σ^2 es estimada por una varianza combinada de las varianzas de las dos muestras tomadas, dada por la siguiente fórmula:

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

Un intervalo de confianza del 100(1- α) % para la diferencia $\mu_1 - \mu_2$ de las medias poblacionales será de la forma:

$$\bar{x} - \bar{y} \pm t_{(\alpha/2, n+m-2)} s_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

Las fórmulas para las pruebas de hipótesis son las siguientes:

Caso I	Caso II	Caso III
$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 = \mu_2$
$H_a: \mu_1 < \mu_2$	$H_a: \mu_1 \neq \mu_2$	$H_a: \mu_1 > \mu_2$
Prueba Estadística:		
$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{con m+n-2 grados de libertad}$		
Decisión:		
Si $t_{cal} < -t_\alpha$ entonces se rechaza H_0	Si $t_{cal} < t_{\alpha/2}$ o $t_{cal} > t_{1-\alpha/2}$ se rechaza H_0	Si $t_{cal} > t_{1-\alpha}$ se rechaza H_0

Las fórmulas se pueden generalizar para probar hipótesis de las diferencias de las dos medias es una cantidad especificada D_0 . En **MINITAB**, para hallar intervalos de confianza de diferencia de dos medias poblacionales y hacer prueba de hipótesis para comparar dos grupos se sigue la secuencia **STAT ► 2-sample t**.

Ejemplo 7.13. Se desea comparar si los estudiantes de escuelas privadas y públicas tiene igual rendimiento en la prueba de aprovechamiento matemático del College Board. Los datos aparecen en el ejemplo 7.11.

Solución:

En el ejemplo 7.11 se concluyó usando la prueba de F que que había igualdad de varianzas de las poblaciones de donde provenían las muestras. **(Nota: por tal motivo es que debe siempre verificar si las varianzas de los grupos son o no iguales)** Luego la ventana de diálogo **2 sample t** se completa como se muestra en la figura 7.20.

Notar que aparece seleccionada la opción **samples in one column** porque los datos de las dos muestras van en una misma columna (*aprovech*), y en otra columna (*escuela*) van los valores que permiten identificar a qué muestra pertenece el dato. La opción **Samples in different columns** se usa cuando las dos muestras están en columnas separadas. Notar además que la opción **Assume equal variances** aparece marcada.

Al oprimir el botón **Options** se puede elegir el nivel de confianza, el valor de la hipótesis que se quiere probar y la dirección de la hipótesis alterna tal como se muestra en la figura 7.21

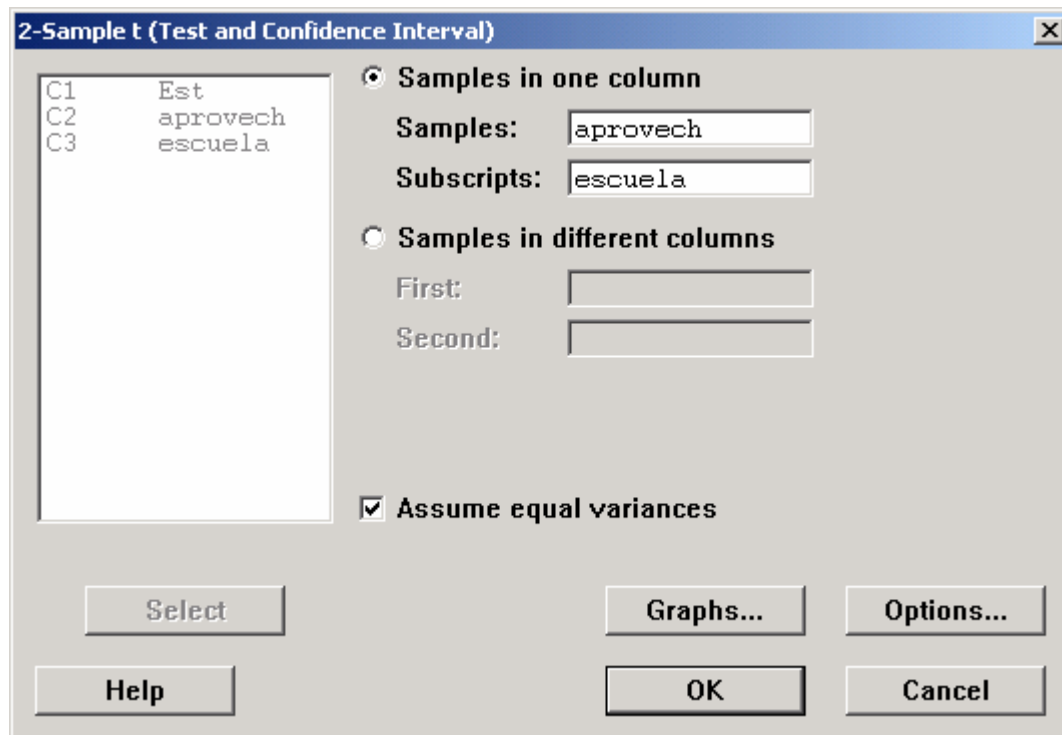


Figura 7.20. Ventana de diálogo de **2-sample t** para el ejemplo 7.13.

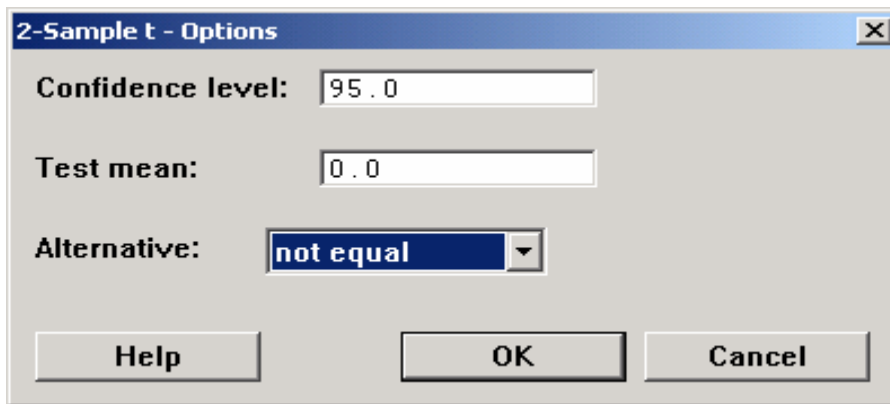


Figura 7.21. Ventana de diálogo de **Options** para 2-sample t

Nota: Se escoge not equal para verificar si el rendimiento de las escuelas publicas y privadas son diferentes.

Los siguientes resultados aparecerán en la ventana **session**:

Two-Sample T-Test and CI: aprovech, escuela

Two-sample T for aprovech

escuela	N	Mean	StDev	SE Mean
privada	6	680.8	55.3	23
pública	8	645.0	45.1	16

Difference = mu (privada) - mu (publica)

Estimate for difference: 35.8

95% CI for difference: (-22.6, 94.3)

T-Test of difference = 0 (vs not =): T-Value = 1.34 P-Value = 0.206 DF = 12

Both use Pooled StDev = 49.6

Interpretación:

El valor del “P-value” es .206 mayor que el nivel de significación $\alpha=.05$, por lo tanto NO se rechaza la hipótesis nula y se concluye de que **no hay evidencia de que los estudiantes de escuela pública tengan un rendimiento distinto que los de escuela privada** en las pruebas de aprovechamiento. El número de grados de libertad de la t es 12.

Notar que el intervalo de confianza del 95% para la diferencia es (-22.6,94.3) que contiene a cero. Este es otra manera de justificar que se acepta la hipótesis nula.

Eligiendo la opción **Graphs** de la ventana de diálogo se obtiene los boxplots de los dos grupos, como aparece en la siguiente figura:

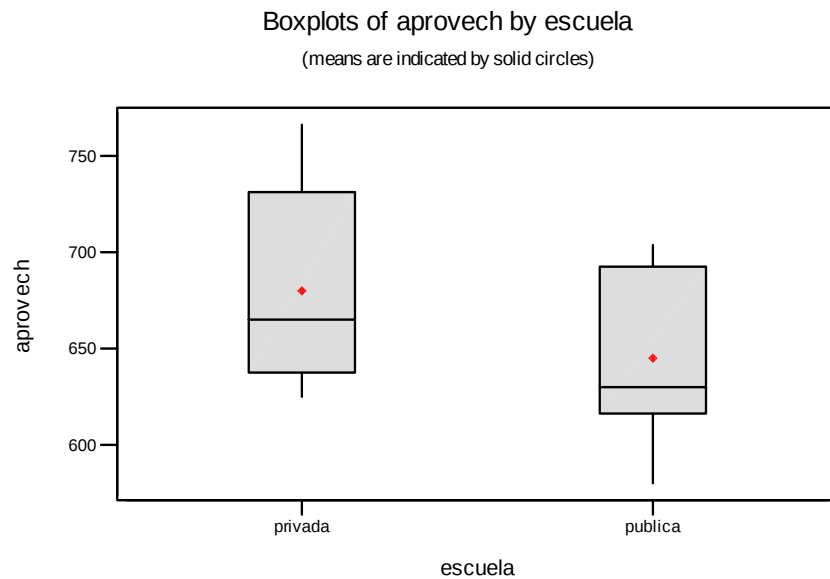


Figura 7.22. Comparación de dos grupos usando boxplots.

Interpretación: No se puede apreciar una marcada diferencia entre las medianas (representadas por las líneas dentro de las cajas), ni las medias (representadas por los puntos rojos) de los grupos. La variabilidad de los dos grupos también es bastante similar ya que los dos “boxplots” tienen alargamiento similar.

Si las varianzas de las poblaciones no son iguales, entonces se usa una prueba aproximada de t, donde el número de grados de libertad es calculado aproximadamente.

La prueba de t aproximada está dada por:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

donde los grados de libertad df son aproximados por la siguiente fórmula:

$$df = \frac{(c_1 + c_2)^2}{\frac{c_1^2}{m-1} + \frac{c_2^2}{n-1}}$$

con $c_1 = \frac{s_1^2}{m}$ y $c_2 = \frac{s_2^2}{n}$.