Test con nivel asintótico para hipótesis sobre la media de una población sin suponer normalidad.

El "test t para una muestra" que se ha deducido suponiendo distribución normal, puede ser usado para variables con cualquier distribución, gracias al TCL. En este caso el nivel del test ya no será el valor elegido α, sino que será aproximadamente α para muestras grandes. Para muestras grandes es indistinto usar la distribución N(0,1) o la distribución t de Student para calcular el valor P (o la región de rechazo), porque cualquiera de las dos distribuciones que se use, se consigue el nivel asintótico α. En la práctica generalmente se usa la distribución t (ya que el test "t" es el único programado en la mayoría de los software como en el Statistix): si la normalidad es cierta, el test tiene nivel exacto α, si los datos no son normales y n es grande, tiene nivel asintótico α. Si los datos no son normales y n es pequeño, es un error aplicar el test t.

Relación entre intervalos de confianza y test de hipótesis.

Si se desea estudiar las hipótesis

$$H_0 : \mu = \mu_o \qquad H_1 : \mu \neq \mu_o$$

se aplica el test "t" para una muestra que hemos visto con el nivel α que se prefije.
La siguiente regla de decisión:

- se calcula primero el IC para μ con nivel $1-\alpha$
- si $\mu_o \in$ IC se acepta H0, si $\mu_o \notin$ IC se rechaza H0

es equivalente (lleva a la misma decisión) que el test "t" (demostrarlo).

Este procedimiento para deducir un test a partir de un intervalo de confianza es muy general. Si conozco un método para encontrar un intervalo de confianza con nivel $1-\alpha$ para un parámetro cualquiera (que llamaremos θ), y si luego queremos estudiar hipótesis de la forma:

$$H_0 : \theta = \theta_o \qquad H_1 : \theta \neq \theta_o$$

la regla de decisión:

Si $\theta_o \in$ IC para θ se acepta H_0
si $\theta_o \notin$ IC para θ se rechaza H_0

es un test de hipótesis con nivel de sinificación (probabilidad de error tipo I) igual a α.

Esta relación entre intervalo y test es válida para hipótesis bilaterales. También se puede demostrar que recíprocamente si conocemos un método para encontrar tests para hipótesis bilaterales, se pueden deducir intervalos de confianza. La justificación de este hecho no es tan simple como la de su recíproco.

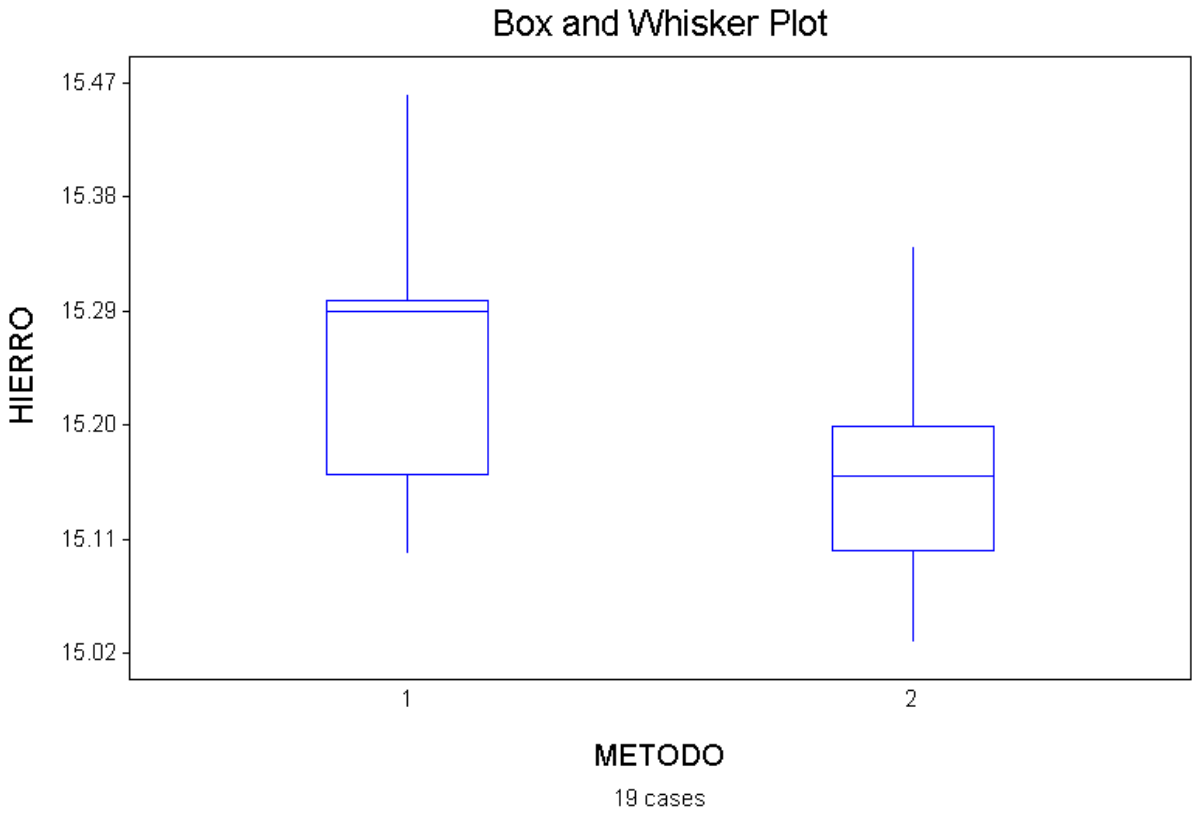
INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE DOS MEDIAS Y TEST PARA COMPARAR LAS MEDIAS DE DOS MUESTRAS INDEPENDIENTES

Ejemplo 3 Queremos comparar los resultados de dos métodos de medición del contenido de hierro de un mineral. Para ello hacemos 10 determinaciones por cada uno de los métodos del mismo trozo de mineral. Una de las mediciones la excluimos porque se detectó un problema en el proceso de medición, por lo que hay 9 determinaciones en uno de los métodos. Los resultados son:

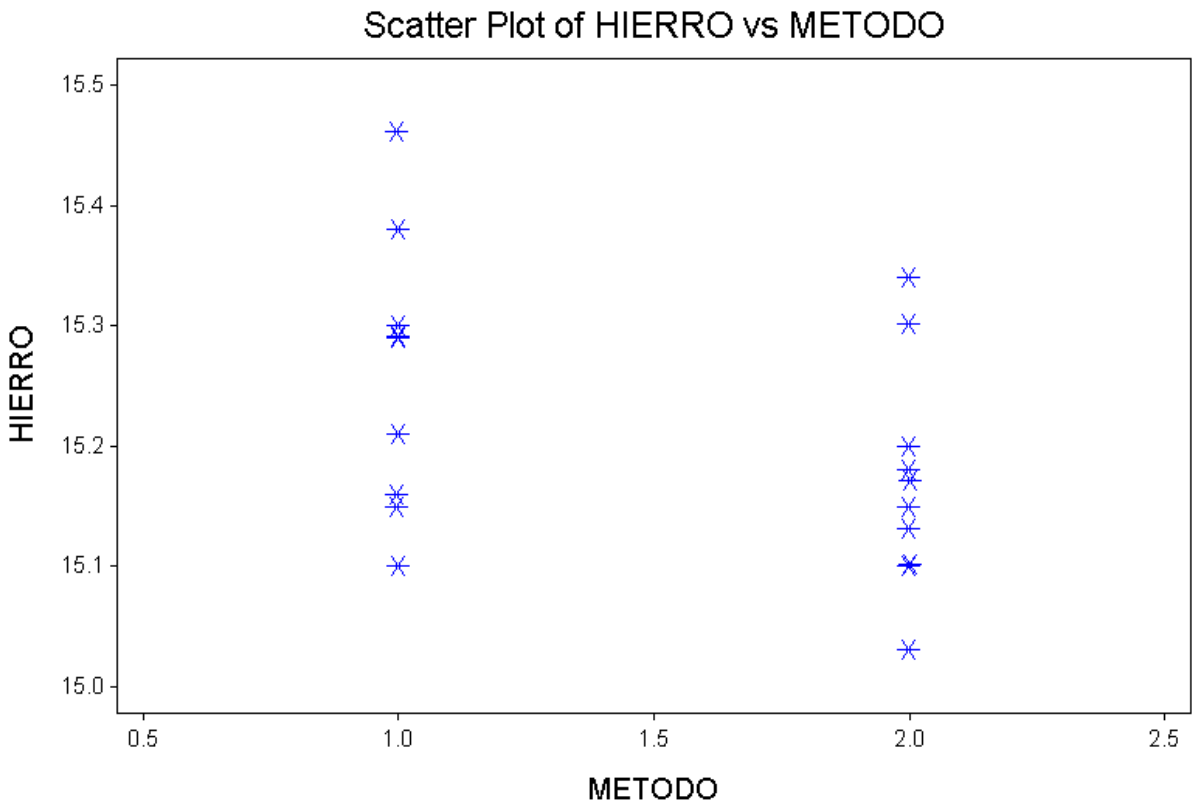
Determinaciones de hierro (en %) por dos métodos para un mismo trozo de un mineral.

	Método 1 (n_1=9)	Método 2 (n_2=10)
	15.30	15.10
	15.46	15.15
	15.38	15.34
	15.29	15.18
	15.16	15.13
	15.15	15.30
	15.10	15.20
	15.29	15.10
	15.21	15.03
		15.17
Promedio	**15.26**	**15.17**
DS	**0.116**	**0.093**

En la práctica lo primero que hacemos para comparar los valores observados con ambos métodos es aplicar algún método gráfico para ver los datos como puede ser un boxplot:



o simplemente representamos los puntos observados por ambos métodos:



También calculamos algunas medidas de resumen para ambos métodos como el promedio y la DS para cada uno (están presentadas en la tabla debajo de los datos).
Con el método 2 hemos obtenido un promedio menor (15.17) que con el método 1 (15.26). Esta diferencia puede deberse al azar o es "estadísticamente significativa"?
Para comparar las "verdaderas" medias de ambos métodos (no solamente las muestrales) podemos usar dos métodos de inferencia estadística:

a) Calcular un intervalo de confianza para la diferencia de las dos medias poblacionales
b) Aplicar un test de la hipótesis nula de que las dos medias son iguales.

Suponiendo que la experiencia indica que los errores de medición de ambos métodos tienen distribución aproximadamente gaussiana (de paso, los gráficos sirven para visualizar si hay algún outlier severo que contradiga esta suposición, lo que no ocurre en este ejemplo), podemos usar el siguiente modelo:

MODELO A: Modelo de dos muestras normales independientes:

$$X_1, X_2, ..., X_{n1} \text{ vs. as. i.i.d } N(\mu_1,\sigma_1^2)$$
$$Y_1, Y_2, ..., Y_{n2} \text{ vs. as. i.i.d } N(\mu_2,\sigma_2^2)$$

donde las vs. as. X_i's son independiente de las vs. as. Y_j's

Llamemos $\overline{X}$ y s_1 a la media y la DS de la muestra 1; $\overline{Y}$ y s_2 a la media y la DS de la muestra 2.

Sabemos que bajo el modelo de dos muestras normales independientes:

$$\overline{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$$

$$\overline{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

como las X's son independiente de las Y's, $\overline{X}$ es independiente de $\overline{Y}$ y por consiguiente:

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

Estandarizando:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Como en la practica pocas veces se conocen ,σ_1^2 y ,σ_2^2, parece natural reemplazar estas variancias "verdaderas" por sus estimadores, las varianzas muestrales s_1^2 y s_2^2 y considerar el cociente

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (15)$$

Pero, lamentablemente, aún suponiendo normalidad de las variables (modelo A) no se conoce exactamente la distribución de este cociente. No tiene distribución N(0,1) ni tampoco distribución t de Student (como quizás en un primer momento uno pueda creer, por analogía al caso de una muestra normal). El hecho de tener que estimar dos varianzas en lugar de una, hace que no valga un resultado similar al que usamos para una muestra (donde reemplazando σ^2 por s2 obteníamos una t_{n-1}, ver (8), pagina 67). El cociente (15) no sólo no es una t, su distribución no puede ser tabulada (o programada) porque depende de parámetros desconocidos del modelo (se puede demostrar que la distribución de (15) depende del cociente entre las dos varianzas σ_2^2/σ_1^2).
Si se hace una suposición más se puede encontrar una distribución t y por lo tanto un IC y un test con nivel exacto. La suposición que se agrega es que las varianzas (verdaderas) son iguales. Si las varianzas de las dos poblaciones son iguales es natural llamarlas σ^2 (en lugar de σ_1^2 y σ_2^2). Tenemos entonces este nuevo modelo:

MODELO B: Modelo de dos muestras normales independientes con varianzas iguales.

$X_1, X_2, ..., X_{n1}$ vs. as. i.i.d $N(\mu_1,\sigma^2)$
$Y_1, Y_2, ..., Y_{n2}$ vs. as. i.i.d $N(\mu_2,\sigma^2)$

donde las vs. as. X_i's son independiente de las vs. as. Y_j's

Llamemos, igual que para el modelo A $\overline{X}$ y s_1^2 a la media y la varianza de la muestra 1; $\overline{Y}$ y s_2^2 a la media y la varianza de la muestra 2.

Pero el modelo B tiene tres parámetros desconocidos (no cuatro como el modelo A): μ_1, μ_2 y σ^2
$\overline{X}$ es el estimador de μ_1, $\overline{Y}$ el estimador de μ_2.
Cuál es el estimador de σ^2?
Parece natural que el estimador de σ^2 se obtenga calculando un promedio de s_1^2 y de s_2^2 que le dé mayor peso a la varianza muestral de la muestra que tiene más observaciones. Se puede demostrar que el mejor estimador de σ^2 bajo el modelo B es:

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

y que bajo este modelo es cierto que

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n1-1)+(n2-1)} = t_{n1+n2-2} \qquad (16)$$

(demostrar (16)).

Intervalo de confianza para la diferencia de medias (basado en dos muestras independientes, normales con igual varianza).
A partir de (16) se puede deducir fácilmente (hacerlo) que

$$\left[\overline{X} - \overline{Y} - t_{n1+n2-2;\, \alpha/2} * \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad , \quad \overline{X} - \overline{Y} + t_{n1+n2-2;\, \alpha/2} * \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right]$$

es un IC para $\mu_1 - \mu_2$ con nivel de confianza $1-\alpha$.

También se puede deducir un test para comparar las dos medias:

Test t sobre la diferencia de medias, basado en dos muestras independientes (suponiendo normalidad y $\sigma_1 = \sigma_2$).
Suponemos el modelo B.
Vamos a considerar primero hipótesis bilaterales. Las hipótesis que queremos estudiar pueden ser:

$$H_0: \mu_1 = \mu_2 \qquad \text{vs} \qquad H_1: \mu_1 \neq \mu_2$$

o más general

$$H_0: \mu_1 - \mu_2 = \delta \qquad \text{vs.} \qquad H_1: \mu_1 - \mu_2 \neq \delta$$

donde δ es un valor propuesto, que generalmente vale cero.
Es intuitivamente razonable que si la diferencia $\overline{X} - \overline{Y}$ está cerca de δ (o de 0), aceptemos la hipótesis nula. ¿Pero que quiere decir cerca de cero?

Del resultado (16) deducimos el siguiente test:

Suponemos el modelo A y las hipótesis H_0: $\mu_1 - \mu_2 = \delta$ vs H_1: $\mu_1 - \mu_2 \neq \delta$

1er. paso: Calculo el estadístico T dado por

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{s_p * \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \qquad (17)$$

Nota: si H_0 es cierta, T tiene distribución de Student con n_1+n_2-2 grados de libertad.

2do. paso: Calculo el valor p usando la curva t de Student con n_1+n_2-2 grados de libertad El valor p es el área a dos colas (a partir del valor de t calculado en el paso anterior).

3er. paso: Si el valor $p < 0.05$ rechazo H_0 o equivalentemente afirmo que la diferencia entre las medias de las dos muestras es estadísticamente significativa.

Otra forma equivalente de hacer las cuentas del test "t" para dos muestras independientes es la siguiente.

1er. paso: Calculo el estadístico T dado por (17), igual que en la versión anterior.

2do. paso: Si $|T| > t_{n1+n2-2,\, \alpha/2}$ rechazo H_0
si $|T| \leq t_{n1+n2-2,\, \alpha/2}$ acepto H_0

Ejercicio: ver qué cambia en el test t anterior (cualquiera de las dos versiones) si la hipótesis alternativa es unilateral **(H_1: $\mu_1 - \mu_2 > \delta$)**

IC para diferencia de medias y test sobre la diferencia de medias, basado en dos muestras independientes, pero sin suponer $\sigma_1 = \sigma_2$ (método de Welch).

Cuando hay evidencia o se sospecha que las dispersiones de ambas muestras son diferentes se supone el modelo A y se usa un IC o un test aproximado. Ante la duda, es mejor usar este método que el anterior, ya que tiene una suposición menos.

Habíamos dicho que bajo el modelo A, la distribución del cociente (15) no es conocida. Sin embargo Welch (1947) verificó que la distribución puede aproximarse por una distribución t de Student con unos grados de libertad que dependen de las varianzas muestrales. La aproximación es la siguiente:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \tilde{a} \quad t_k \qquad (18)$$

donde los grados de libertad k se calculan con la siguiente expresión

$$k = \text{round}\left(\frac{\left(s_1^2 / n_1 + s_2^2 / n_2\right)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n2 - 1}} \right) \qquad (19)$$

Entonces tenemos el siguiente test:

Test de Welch,
Supongamos el modelo A y la hipótesis H_0: $\mu_1 - \mu_2 = \delta$

1er. paso: Se calcula el estadístico T'

$$T' = \frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

2do. paso: Calculo el valor P usando la curva t de Student con los grados de libertad calculado por (19), a dos colas o a una cola según sea bilateral o unilateral la hipótesis alternativa.

3er. Paso: Como en todo test, si $P < \alpha$ se rechaza H_0 , si $P \geq \alpha$ se acepta H_0

Intervalo de confianza para la diferencia de medias (basado en dos muestras independientes, normales, sin suponer igualdad de varianza).

A partir de (18) se puede deducir fácilmente que

$$\left[\overline{X} - \overline{Y} - t_{k\,;\,\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad , \quad \overline{X} - \overline{Y} + t_{k\,;\,\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

donde k está dado por (19) es un IC para $\mu_1 - \mu_2$ con nivel de confianza aproximado $1-\alpha$.

Test para comparar las varianzas de dos muestras normales independientes

Supongamos nuevamente el modelo A, pero ahora queremos comparar las varianzas de las dos poblaciones. Queremos estudiar las hipótesis:

$$H_0: \sigma_1^2 = \sigma_2^2 \qquad \text{vs} \qquad H_1: : \sigma_1^2 \neq \sigma_2^2$$

Es intuitivo que el test se va a basar en comparar las varianzas muestrales s_1^2 con s_2^2 El test no se basa en las diferencia entre estas varianzas sino en su cociente. El test es el siguiente:

1er. paso: Calculo el estadístico *F* dado por

$$F = \frac{s_1^2}{s_2^2} \qquad (20)$$

Nota: si H_0 es cierta, F tiene distribución F de Fisher (la tabuló Snedecor, pero la llamó F en homenaje a Fisher) con n_1-1grados de libertad en el numerador y n_2-1 grados de libertad en el denominador.

2do. paso:

Si $F > F_{n1-1, n2-1, \alpha/2}$ o $F < F_{n1-1, n2-1, 1-\alpha/2}$ rechazo H_0
en caso contrario acepto H_0

Nota: si la hipótesis alternativa fuese unilateral (H_1: $\sigma_1^2 > \sigma_2^2$) , ser rechazaría H_0 solamente en una cola: cuando $F > F_{n1-1, n2-1, \alpha}$

Ejercicio: Otra forma equivalente es, como siempre calcular el valor P y rechazar H_0 si $P<\alpha$, como se calcula P para el test a una cola? y para el test a dos colas?

Comentarios:

1) Tanto el test de t como el test de Welch y los IC para $\mu_1 - \mu_2$ son válidos, suponiendo normalidad, para cualquier tamaño de muestra.
Si la variable no tiene distribución normal y ambas muestras son grandes pueden usarse como aproximación, gracias al Teorema Central del Límite.

2) El test F para comparar dos varianzas es válido para cualquier tamaño de muestra suponiendo normalidad. Si la variable no tiene distribución normal, el test no es válido, aún para muestras grandes. No hay ningún teorema parecido al TCL que diga que este test tiene nivel asintótico correcto sin suponer normalidad. Esta es una limitación de este test.

3) Si se tienen todos los datos no se necesita aplicar las fórmulas dadas para aplicar los tests estudiados. El StatistiX (o cualquier programa que tenga procemientos estadísticos) hace todas las cuentas. Nuestra tarea es elegir cuál test se puede aplicar y saber cuál es incorrecto aplicar para cada problema.

CONTINUACIÓN DEL ANÁLISIS DE LOS DATOS DEL EJEMPLO 3.

Habíamos visto que :

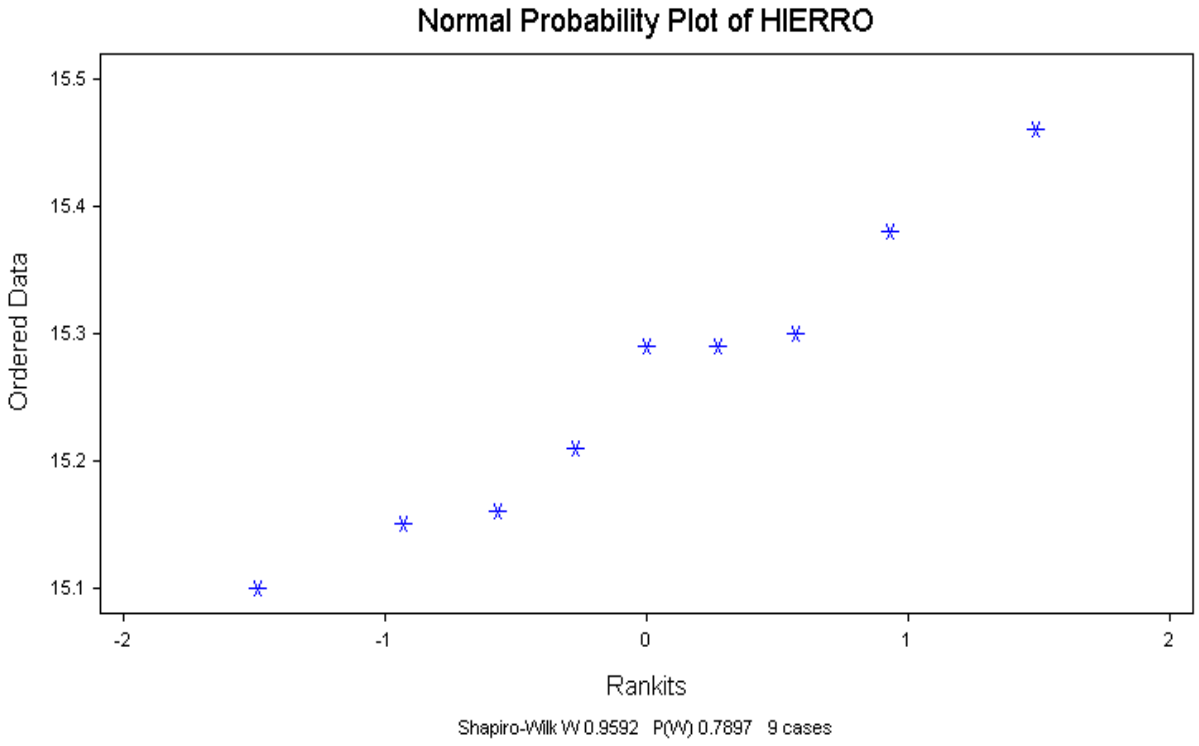
	Método 1	Método 2
Promedio	15.26	15.17
DS	0.116	0.093

El promedio es más alto con el método 1, podemos decir con estos datos que las medias de ambos métodos son diferentes?
Podemos aplicar el test t o el test de Welch?

Ambos tests suponen normalidad (las muestras son pequeñas). Supongamos que la experiencia previa indica que es razonable suponer errores normales. De cualquier modo, parar estar más seguros de que nada raro pasó en este experimento, conviene describir los datos con algún gráfico como los que hicimos (box plot o gráfico de puntos). Ya observamos que estos gráficos no muestran outliers ni nada que haga sospechar que la suposición de normalidad sea falsa.
Además de los gráficos hechos (box plot y gráfico de puntos), también se puede graficar un "normal probability plot", que debería parecer una recta si los datos son aproximadamente normales. También se puede estudiar formalmente la hipótesis de normalidad con un test. Un test propuesto para estudiar esta hipótesis, es el test de Wilk-Shapiro, que estudia la hipótesis:

H_0: $X_1, X_2, \ldots X_n$ son vs. as. i.i.d. con distribución Normal
H_1: tienen otra distribución.

Veremos como se grafica el "normal probability plot" y se aplica este test con el Statistix. Primero lo aplicaré a los datos del método 1. Para ello primero seleccionamos los datos del método 1: hacemos click en DATA, Omit/Select/Restore cases y ponemos Omit METODO=2. Así quedan seleccionados los nueve casos del método 1. Luego vamos a Statistics, Randomnes/Normality Tests, Normal Probability Plot, ponmose HIERRO en la ventana de las variables y se obtenemos el siguiente gráfico y en el renglón de abajo el valor P del test de Wilk-Shapiro:



Luego para seleccionar los datos del método 2, primero, para deshacer la selección anterior, vamos a Omit/Select/Restore cases y marcamos Restore; luego escribímos Omit METODO=1. Luego volvemos a hacer los pasos antes descriptos y obtenemos un gráfico similar, ahora con un valor P de 0.6243.

Se puede apreciar que en ambas salidas el valor P es alto (bastante >0.05), por lo que no hay ninguna evidencia para rechazar la hipótesis nula de normalidad.

Aceptemos normalidad. Nos queda aún la duda de si usar el Modelo A y aplicar el test t o el Modelo B y aplicar el test de Welch. Puede ser que también haya experiencia sobre este tema. Si hay experiencia, la usamos. En caso contrario, podemos observar las DS de ambas muestras. En este caso son 0.116 y 0.093. Con un poquito de experiencia, ya podemos pensar que estas DS son parecidas y podemos aplicar el test t. En caso de duda, podemos aplicar primero el test F para estudiar la hipótesis de igualdad de varianzas.

$$H_0: \sigma_1^2 = \sigma_2^2 \qquad \text{vs} \qquad H_1: : \sigma_1^2 \neq \sigma_2^2$$

Test:

1er. paso: Calculo el estadístico *F* dado por (comentario: por la forma en que están hechas las tablas conviene poner en el numerador el s^2 más grande):

$$F = \frac{s_1^2}{s_2^2} = \frac{0.116^2}{0.093^2} = 1.556$$

Si usamos el método "antiguo" de comparar con la tabla de F, buscamos en la tabla el valor que deja un área de 0.025 a derecha en la función de densidad F con 9-1=8 gl en el numerador y 10-1=9 en el denominador que es (ver tabla) 4.10 y como

$$1.556 < 4.10 \quad , \quad \text{no rechazamos } H_0$$

Si en cambio tenemos el StatistiX a mano, podemos calcular el valor P, en Probability Functions, F(x,dfnum,dfden), ponemos 1.556 en X, 8 en DFNUM y 9 en DFDEN y obtenemos:

F(1.556,8,9) = 0.26153.

Esta es el área de una cola bajo la densidad F, el valor P para el test bilateral es el doble:

P=2*0.262=0.52

y volvemos a llegar a la conclusión de que no hay evidencia en contra de la hipótesis de igualdad de varianzas.

Por la experiencia previa o por un análisis como el que hemos hecho, estamos en condiciones de suponer el modelo A y para contestar la pregunta que nos interesa (comparar las medias) estudiamos $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$. Para ello aplicamos el test t de Student para dos muestras normales independientes:

Calculamos primero el estimador de la varianza:

$$s_p^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2} = \frac{(9-1) * 0.1162^2 + (n_2 - 1) * 0.0932^2}{9 + 10 - 2} = 0.01095$$

o su raíz cuadrada:

$$s_p = \sqrt{0.01095} = 0.1046$$

y ahora aplicamos el test:

1er. paso: Calculamos el estadístico

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s_p * \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{15.260 - 15.170}{0.1046 * \sqrt{(\frac{1}{9} + \frac{1}{10})}} = \frac{0.09}{0.04806} = 1.87$$

2do. paso: Calculamos el valor p usando la curva t de Student con $n_1+n_2-2=17$ grados de libertad El valor p es el área a dos colas a partir del valor observado 1.87. Puedo calcular esta área con Statistix, Probability Function, T-2tail y obtengo

P=0.0788

3er. paso: Como P=0.0788 > 0.05 no puedo rechazar H_0 y digo que "a pesar de que el promedio de los valores observados con el método 2 es un poco más alto que con el método 1, la diferencia no es estadísticamente significativa (P=0.079)" o "a pesar de que el promedio de los valores observados con el método 2 es un poco más alto que con el método 1, esto no es evidencia suficiente para afirmar que los dos métodos tengan diferentes medias (P=0.079)" .

StatistiX puede hacer todas las cuentas. Marcando "Statistics", "One, Two and Multi-Sample tests", "Two Sample T test", se obtiene la siguiente salida:

```
TWO-SAMPLE T TESTS FOR HIERRO BY METODO

                          SAMPLE
  METODO         MEAN      SIZE       S.D.        S.E.
----------   ----------   ------   ---------   ---------
     1          15.260       9       0.1162      0.0387
     2          15.170      10       0.0932      0.0295
DIFFERENCE      0.0900

NULL HYPOTHESIS: DIFFERENCE =  0
ALTERNATIVE HYP: DIFFERENCE <> 0

ASSUMPTION             T       DF        P     95% CI FOR DIFFERENCE
-----------------   ------   ------   ------   ---------------------
EQUAL VARIANCES       1.87     17     0.0786     (-0.0115, 0.1915)
UNEQUAL VARIANCES     1.85     15.4   0.0838     (-0.0135, 0.1935)

                        F      NUM DF   DEN DF      P
TESTS FOR EQUALITY   -------   ------   ------   ------
      OF VARIANCES     1.55       8        9     0.2622

CASES INCLUDED 19    MISSING CASES 0
```

Ejercicio: buscar en esta salida del Statistix los tests que hemos aplicado.

Veamos ahora otro ejemplo en que tambien se desea comparar las medias de las determinaciones hechas con dos métodos.

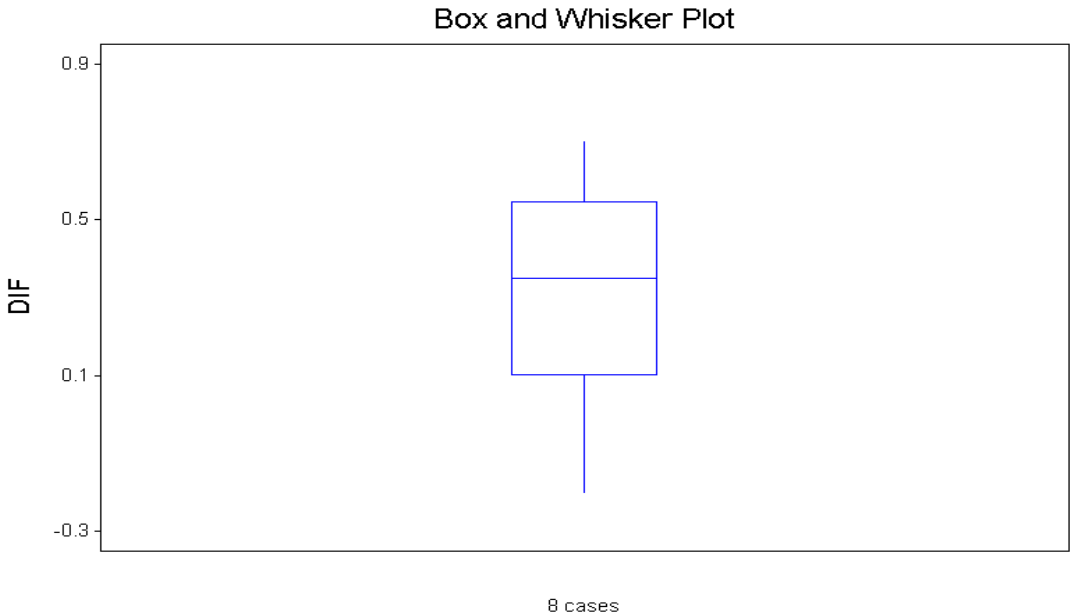
Ejemplo 4: Queremos comparar los resultados de dos métodos de medición de la concentración de zinc en alimentos. Para ello para cada una de 8 muestras de alimentos medimos el porcentaje de concentración de zinc por ambos métodos, obteniendo

Determinaciones de zinc (en %) por dos métodos para 8 muestras de alimentos.
Método 1: titulación con AEDT. Método 2: Espectrometría atómica.

Muestra	Método 1	Método 2
1	7.2	7.6
2	6.1	6.8
3	5.2	5.2
4	5.9	5.7
5	9.0	9.7
6	8.5	8.7
7	6.6	7.0
8	4.4	4.7
Promedio	**6.61**	**6.92**
DS	**1.57**	**1.72**

Queremos, igual que en el ejemplo 3 comparar las medias de los dos métodos. Podemos suponer el modelo A o el modelo B y aplicar el test t para dos muestras independientes o el test de Welch?

Muestra	Método 1	Método 2	Diferencia
1	7.2	7.6	0.4
2	6.1	6.8	0.7
3	5.2	5.2	0
4	5.9	5.7	-0.2
5	9.0	9.7	0.7
6	8.5	8.7	0.2
7	6.6	7.0	0.4
8	4.4	4.7	0.3
Promedio	**6.61**	**6.92**	**0.312**
DS	**1.57**	**1.72**	**0.314**



```
ONE-SAMPLE T TEST FOR DIF

NULL HYPOTHESIS: MU =  0
ALTERNATIVE HYP: MU <> 0

MEAN         0.3125
STD ERROR    0.1109
LO 95% CI    0.0503
UP 95% CI    0.5747
T              2.82
DF                7
P            0.0259

CASES INCLUDED 8    MISSING CASES 0
```

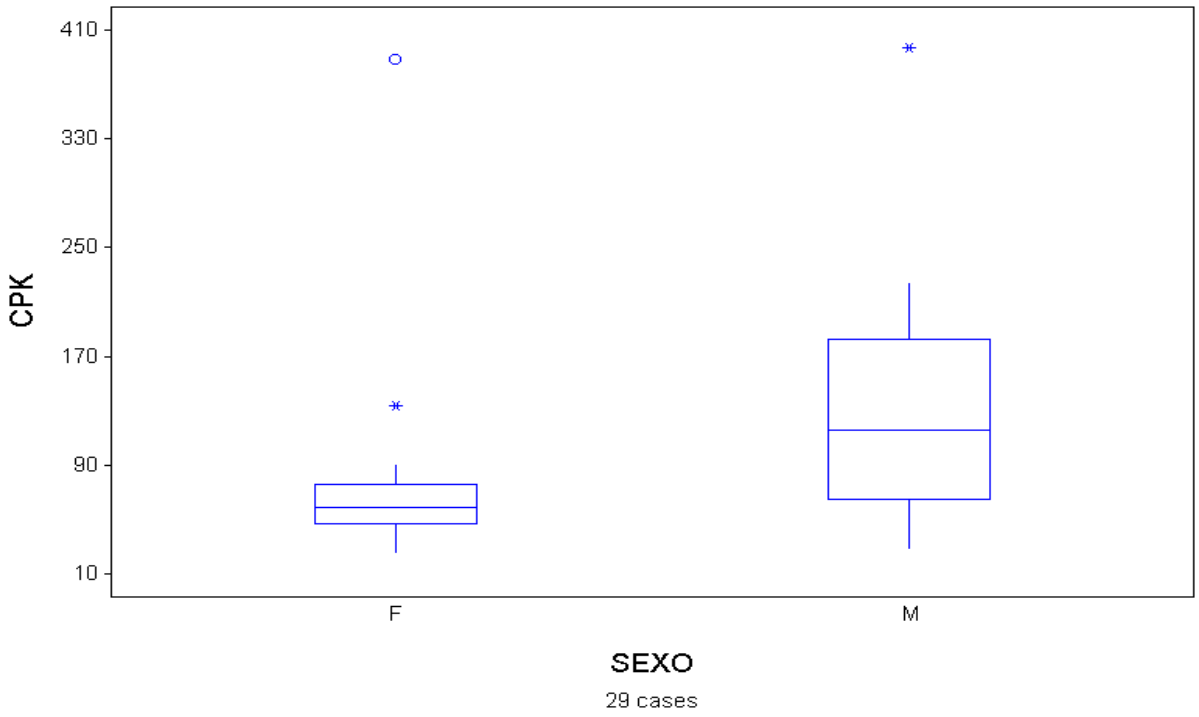
Otra forma de hacer el test directamente con Statistix sin calcular primero las diferencias es la siguiente: Vamos a Statistics, luego a One, Two, Multi- Sample Tests y luego a Paired T Test. Eligiendo las variables Metodo 1 y Metodo 2 obtenemos la misma salida:

```
PAIRED T TEST FOR METODO1 - METODO2

NULL HYPOTHESIS: DIFFERENCE =  0
ALTERNATIVE HYP: DIFFERENCE <> 0

MEAN        -0.3125
STD ERROR    0.1109
LO 95% CI   -0.5747
UP 95% CI   -0.0503
T             -2.82
DF                7
P            0.0259

CASES INCLUDED 8    MISSING CASES 0
```

TESTS NO PARAMÉTRICOS PARA DOS MUESTRAS INDEPENDIENTES, PARA UNA MUESTRA Y PARA MUESTRAS DE A PARES. (TEST DE MANN-WHITNEY, TEST DE SIGNOS, TEST DE WILCOXON).

Volvamos al problema de comparar las medias de dos muestras independientes.
Vimos el test t para comparar las media de dos muestras normales independientes, suponiendo $\sigma_1 = \sigma_2$ y el test de Welch que no hace esta última suposición.

Ejemplo 5: En la clase de estadística descriptiva uno de los ejemplos usados fueron los datos de análisis de sangre de 29 enfermos cardíacos (archivo PCRCPKTn). Supongamos que queremos comparar los valores de CPK para pacientes varones y mujeres. Comencemos por describir los datos. Para ello a continuación se muestra el box plot y medidas de resumen de CPK para cada sexo.



DESCRIPTIVE STATISTICS FOR SEXO = F

	CPK
N	11
MEAN	90.727
SD	102.91
1ST QUARTI	45.000
MEDIAN	59.000
3RD QUARTI	90.000

DESCRIPTIVE STATISTICS FOR SEXO = M

	CPK
N	18
MEAN	133.89
SD	88.912
1ST QUARTI	64.000
MEDIAN	116.00
3RD QUARTI	188.00

Dijimos en la clase de estadística descriptiva que, teniendo en cuenta que la variable tiene distribución asimétrica y con datos atípicos, es más informativo elegir como medidas de resumen la mediana y cuartiles (y no el promedio y la DS).
El box plot nos muestra que los valores de CPK en los hombres son mayores que para las mujeres. Las medianas son 116 y 59 respectivamente. ¿Pueden las diferencias que observamos ser atribuidas al azar o esto es poco probable? Queremos aplicar un test para comparar los valores de CPK en ambos sexos.
¿Se puede aplicar el test t para dos muestras independientes? Evidentemente no: el modelo de dos muestras normales no es una aproximación razonable parar estos datos. Para peores las muestras no son grandes (11 mujeres y 18 hombres) lo que hace que la aproximación que da el TCL no tiene por que ser buena.

Modelos no paramétricos

Se llaman modelos paramétricos a los que tienen un número finito de parámetros desconocidos. El modelo de una muestra normal por ejemplo tiene dos parámetros. Si conociésemos los valores de estos parámetros, quedaría totalmente especificada la distribución de las variables. Se llaman modelos no paramétricos a aquellos para los que no alcanza con dar los valores de un número finito de parámetros (digamos que tienen "infinitos parámetros desconocidos").
Un típico modelo no paramétricos para los datos de una muestra sería por ejemplo:

$$X_1, X_2, ..., X_n \text{ vs. as. i.i.d. con distribución continua}$$

En este modelo no alcanza con conocer un número finito de parámetros para que la distribución de las variables quede completamente especificada. Los modelos no paramétricos son más generales, con menos suposiciones que los modelos paramétricos. No suponen distribución normal, ni exponencial, ni gamma, ni ninguna otra familia particular de distribuciones. Pero alguna suposición siempre se hace. Otra denominación de los "modelos no parámetricos" es "modelos a distribución libre".

Un test no paramétrico para dos muestras independientes: test de Mann-Whitney.

Vamos a presentar un test no paramétrico para comparar dos muestras que se llama test de Mann-Whiney o test de la suma de rangos de Wilcoxon (Wilcoxon lo propuso primero para el caso $n_1=n_2$ y Mann y Whitney poco tiempo después para el caso general). Al igual que muchos otros tests no paramétricos, el test de Mann-Whitney se basan en los rangos (ranks) o número de orden de las observaciones.

El modelo que hay que suponer para poder aplicar el test de Mann-Whitney es el siguiente:

Modelo C:

$$X_1, X_2, ..., X_{n1} \text{ vs. as. i.i.d con distribución } F_1(x)$$
$$Y_1, Y_2, ..., Y_{n2} \text{ vs. as. i.i.d con distribución } F_2(x)$$

donde las vs. as. X_i's son independiente de las vs. as. Y_j's

Se puede apreciar que es mucho más general que el modelo A o el modelo B (no supone ninguna "forma" para la distribución).
La hipótesis nula que se estudia es:

$$H_0: \quad F_1 = F_2$$

Como dijimos, este test es un "test de rangos". Compara no las medias de los datos de las dos muestras, sino las medias de los rangos. Para aplicarlo primero se calculan los rangos (número de órden) juntando los datos de las dos muestras. Muestro a continuación los datos de CPK y sus rangos para las dos muestras (hombres y mujeres):

Sexo Femenino (n_1=11)		Sexo Masculino (n_2=18)	
Dato	Rango	Dato	Rango
26,00	1	29,00	2
33,00	3	47,00	5.5
45,00	4	51,00	7
47,00	5.5	64,00	12.5
53,00	8	64,00	12.5
59,00	9	69,00	14
60,00	10	87,00	15
63,00	11	91,00	17
90,00	16	112,00	18
134,00	20	120,00	19
388,00	28	152,00	21
		153,00	22
		166,00	23
		184,00	24
		200,00	25
		201,00	26
		223,00	27
		397,00	29
Suma rangos	**115.5**		**319.5**
Promedio	**10.5**		**17.75**

La idea del test es comparar no los promedios de los datos originales sino los promedios de los rangos. El estadístico en el que se basa el test es similar (no es exactamente igual pero toma valores muy próximos) al estadístico del test t pero aplicado a los rangos y no a los datos originales.
Para muestras grandes se demuestra que este estadístico tiene, bajo H_0 una distribución aproximadamente normal. Para muestras pequeñas se puede calcular el valor "P" exacto del test. El Statistix hace las cuentas. Para ello vamos a Statistics; One, Two and Multi-Sample Tests and Rank Sum Test. Obtenemos la siguiente salida:

```
WILCOXON RANK SUM TEST FOR CPK BY SEXO

                        SAMPLE
   SEXO       RANK SUM   SIZE     U STAT   MEAN RANK
----------  ---------  ------  ---------  ---------
F              115.50     11      49.500      10.5
M              319.50     18      148.50      17.8
TOTAL          435.00     29

NORMAL APPROXIMATION WITH CORRECTIONS FOR CONTINUITY AND TIES   2.203
TWO-TAILED P-VALUE FOR NORMAL APPROXIMATION                    0.0276

TOTAL NUMBER OF VALUES THAT WERE TIED          4
MAXIMUM DIFFERENCE ALLOWED BETWEEN TIES 0.00001

CASES INCLUDED 29    MISSING CASES 0
```

Podemos resumir los resultados que obtuvimos diciendo por ejemplo:

"La mediana de CPK para hombres es mayor que para mujeres 116 (P25,P75: 64,188) y 59 (45, 90) respectivamente. La diferencia entre los valores de CPK en ambos grupos es estadísticamente significativa (test de Mann-Whitney, p=0.028)."

La ventaja del test de Mann-Whiney es que el valor de p está bien calculado siempre que se cumpla el modelo C y la hipótesis H_0: $F_1=F_2$, no importa cuál sea la distribución de las variables. En este ejemplo el valor p del test de t no es correcto, así que si aplicásemos el test t no sabríamos cuál es la probabilidad de error tipo I.
Pero además de este problema, el test t puede tener alta probabilidad de error tipo II cuando hay outliers, y ser muy ineficiente para detectar diferencias. En este ejemplo, si **POR ERROR** aplicásemos el test t, no se detectaría diferencias entre hombres y mujeres. Muestro la salida del Statistix:

```
TWO-SAMPLE T TESTS FOR CPK BY SEXO

                          SAMPLE
   SEXO          MEAN      SIZE       S.D.        S.E.
----------   ----------   ------   ---------   ---------
F               90.727      11       102.91      31.030
M               133.89      18       88.912      20.957
DIFFERENCE     -43.162

NULL HYPOTHESIS: DIFFERENCE =  0
ALTERNATIVE HYP: DIFFERENCE <> 0

ASSUMPTION             T       DF        P     95% CI FOR DIFFERENCE
-----------------   ------   ------   ------   ---------------------
EQUAL VARIANCES      -1.20     27     0.2423     (-117.24, 30.919)
UNEQUAL VARIANCES    -1.15     18.9   0.2634     (-121.56, 35.239)

                        F      NUM DF   DEN DF      P
TESTS FOR EQUALITY   -------   ------   ------   ------
      OF VARIANCES     1.34      10       17     0.2864

CASES INCLUDED 29    MISSING CASES 0
```

El test t dá un valor P=0.24 y no podríamos rechazar H_0.

Pensar: ¿es el test de Mann-Whitney sensible a datos atípicos? ¿y el test "t"?

"Propaganda" del test de Mann-Whitney:

1. El valor de p es correcto siempre que se cumpla el modelo C y la hipótesis H_0: $F_1=F_2$, no importa cuál sea la distribución de las variables. Por lo tanto si rechazamos H_0 cuando $P<\alpha$, nos aseguramos que la probabilidad de error tipo I es α, cualquiera sea la distribución de la variable.
2. En el caso normal el test t es el test óptimo, en el sentido de que es el test más potente (de menor probabilidad de error tipo II). Pero el test de Mann- Whitney no es mucho menos potente (se puede demostrar que tiene una eficiencia relativa del 95.5% para muestras grandes). De cualquier modo, si uno sabe que la variable es normal, conviene aplicar el test t ("aprovecha mejor los datos" y por eso es un poco más potente).

3. Si la distribución no es normal, especialmente si hay datos atípicos el test de Mann Whitney es más potente que el test t y puede llegar a ser muchísimo más potente.
4. Cuando la muestra es grande y no es normal, podemos aplicar cualquiera de los dos tests. El valor p está bien calculado (para el caso de la t en forma aproximada) para ambos tests y por lo tanto también la probabilidad de error tipo I. Pero si la distribución es muy alejada de la normal, especialmente para datos con "outliers severos" conviene elegir el test de Mann-Whitney, porque tiene menor probabilidad de error tipo II.

Tests no paramétrico para hipótesis sobre la mediana de una población basados en una muestra sin suponer distribución normal: 1) test de signos, 2) test de rangos con signos de Wilcoxon ("Wilcoxon signed rank test").

Veremos dos tests no paramétricos para una muestra que luego serán usados para muestras de a pares.
Como ejemplo para presentar el test t para una muestra hemos considerado (ejemplo 1) los datos de 10 determinaciones de una aleación "standard" para la que se sabía que contenía 4.44% de níquel. Las 10 mediciones daban:

4.32 4.31 4.50 4.12 4.43 4.36 4.48 4.28 4.18 4.42

Supusimos un modelo normal y estudiamos la H_0 de que la media es 4.44 (que en este ejemplo significa que no hubo error sistemático en las mediciones).

Con estos datos ni un box-plot, ni un histograma, ni el test de normalidad de Wilk-Shapiro dan ninguna evidencia de que el modelo normal sea falso. Si por el contrario cambiamos el cuarto dato (que es el menor) por un valor bastante menor (digamos 3.90):

4.32 4.31 4.50 **3.90** 4.43 4.36 4.48 4.28 4.18 4.42

y con estos nuevos datos hacemos un box plot se detecta el valor 3.90 como outlier. Una posibilidad (intuitivamente razonable, en este ejemplo que hay un solo outlier) es excluirlo y aplicar el test t. Otra es aplicar un test no paramétrico que no suponga normalidad ni sea sensible a datos atípicos.
Los dos tests más usados para el caso de una muestra son el test del signo y el test de Wilcoxon. Empiezo por presentar el primero porque es una idea más simple (muy simple).

Test de signos: Para poder aplicar este test, no se necesita suponer casi nada. Podemos poner el siguiente modelo:

Modelo: $X_1, X_2, ..., X_n$ vs. as. i.i.d con una distribución F cualquiera

Llamemos $\tilde{\mu}$ a la mediana de la distribución F (o lo que es lo mismo la mediana de X_i).
Queremos estudiar las hipótesis:

$$H_0: \tilde{\mu} = 4.44 \quad ; \qquad H_1: \tilde{\mu} \neq 4.44$$

o en general

$$H_0: \tilde{\mu} = m_0$$

donde m_0 es un valor propuesto por el investigador.

La idea del test es: si la mediana fuese 4.44 se espera que haya tantos datos mayores que 4.44 como menores que 4.44. En estos 10 datos hay 2 que son >4.44 y 8 que son <4.44. Una forma equivalente es calcular las diferencias entre los datos observados y el valor propuesto en H0:

$$D_i=X_i-m_0$$

y contar cuantas de estas diferencias tienen signo positivo y cuantas signo negativo (por eso el nombre "test de signos"). En nuestro ejemplo hay 8 diferencias negativas y sólo 2 positivas.
Como dijimos, si H_0 fuese cierta habría igual probabilidad de que la diferencias sean >0 o <0. Excluyendo las diferencias nulas si las hubiere, la probabilidad de que cada diferencia sea <0 (o >0) sería 1/2. Si H_0 es cierta el número de diferencias negativas (o positivas) tiene una distribución binomial con p=1/2. Se calcula entonces, en el ejemplo, la probabilidad de que una variable Binomial (n=10,p=0.5) tome un valor 2 "o aún más extremo" $P(X\leq 2)$. Este es el valor p del test "a una cola". Pero como en nuestro ejemplo la H_1 es bilateral, tenemos que informar el valor P a dos colas, que es el anterior multiplicado por 2. Con el Statistix, en Probability Functions, obtenemos

$$\text{Binomial}(2,10,0.5)=0.05469$$

y el valor P del test es $P=2*0.05469\cong 0.11$ y no podemos rechazar H_0.

Distribuciones simétricas: Antes de presentar el test de Wilcoxon, veremos lo que quiere decir que una variable tenga distribución simétrica. Lo más facil de definir es que tenga distribución simétrica alrededor de cero. Si la variable es continua quiere decir que su función de densidad es una función par, o sea f(-x)=f(x). En el caso general X tiene distribución simétrica alrededor de cero si

$$P(X<-x) = P(X>x) \quad \text{para todo número real } x \qquad (21)$$

o sea la probabilidad de cada cola es la misma. Se puede demostrar (muy fácilmente) que esto es equivalente a pedir que

la distribución de la v.a. -X es la misma que la de la v.a. X (22)

Consideren como definición de v.a. simétrica alrededor de cero a (21) o (22) (la que les parezca más intuitiva). Y recuerden que en el caso continuo es equivalente a que la función de densidad cumple

$$f(-x) = f(x)$$

Se dice que una v.a. X tiene distribución simétrica alrededor de un valor m si la v.a. X-m tiene distribución simétrica alrededor de cero.

Ejercicio: (pensar intuitivamente, no les pido una demostración formal)

a) Si X es simétrica alrededor de m y es continua, como será la función de densidad?
b) Si X es simétrica alrededor de m, entonces mediana de X = m
c) Si existe E(X) entonces E(X)=m

Test de rangos con signos de Wilcoxon.

El test de signos sólo tiene en cuenta el signo de las diferencias entre los valores observados y el valor propuesto en H_0. El test de Wilcoxon tiene en cuenta también si la diferencia es "grande" o "pequeña". Pero requiere una suposición más: se necesita suponer simetría.

Modelo: $X_1, X_2, ..., X_n$ vs. as. i.i.d con distribución simétrica

Llamemos $\tilde{\mu}$ al centro de simetría de las vs.as. X_i, que también es la mediana de X_i.
Queremos estudiar la hipótesis nula:

$$H_0 : \tilde{\mu} = m_0$$

En el ejemplo, como antes las hipótesis son: H_0: $\tilde{\mu} = 4.44$; H_1: $\tilde{\mu} \neq 4.44$.
Calculamos primero las diferencias $D_i=X_i-m_0$, que en el ejemplo son:

-0.12 -0.13 0.06 -0.54 -0.01 -0.08 0.04 -0.16 -0.26 -0.02

Observemos que, como usamos en el test de signos, 8 de estas diferencias son negativas y sólo 2 son positivas. Ordenemos estas diferencias, pero teniendo en cuenta solamente su módulo:

-0.01 -0.02 0.04 0.06 -0.08 -0.12 -0.13 -0.16 -0.26 -0.54

Anotemos, al lado de cada diferencia, su rango:
-0.01(1) -0.02(2) 0.04(3) 0.06(4) -0.08(5) -0.12(6) -0.13(7) -0.16(8) -0.26(9) -0.54(10)

y ahora los separamos en dos grupos según el signo:

Rangos de las diferencias <0: 1 2 5 6 7 8 9 10
Rangos de las diferencias >0: 3 4

Si se cumple H_0 esperamos que los datos se distribuyan en forma simétrica a ambos lados del valor 4.44 , entonces se espera que haya aproximadamente el mismo número de diferencias positivas y negativas y que unas y otras tengan aproximadamente la misma media de rangos. Por lo tanto se espera que la suma de los rangos de las diferencias <0 sea similar al de las diferencias >0. En el ejemplo:
Suma de los rangos de las diferencias <0: 48
Suma de los rangos de las diferencias >0: 7

El estadístico del test de Wilcoxon es la diferencia entre estas dos sumas. Para ello se ha estudiando la distribución de estas diferencias bajo la hipótesis nula y se la tabuló. En vez de usar estas tablas, podemos usar el Statistix, que nos calcula directamente el valor p, obteniendo:

```
WILCOXON SIGNED RANK TEST FOR DIF - CERO

SUM OF NEGATIVE RANKS                              -48.000
SUM OF POSITIVE RANKS                               7.0000

EXACT PROBABILITY OF A RESULT AS OR MORE
EXTREME THAN THE OBSERVED RANKS (1 TAILED P-VALUE)  0.0186

NORMAL APPROXIMATION WITH CONTINUITY CORRECTION      2.039
TWO TAILED P-VALUE FOR NORMAL APPROXIMATION         0.0415
```

Para muestras grandes, Statistix sólo calcula el valor de P usando la distribución asintótica del estadístico del test. Para muestras pequeñas calcula también el valor exacto del test a una cola.

Mejor usar este valor exacto, pero como la H_1 que planteamos es bilateral, necesitamos el valor p a dos colas. Multiplicamos por 2 (2*0.0186)obteniendo: P = 0.037.
Conclusión en el ejemplo: Rechazamos H_0 y llegamos a la conclusión de que los datos no se distribuyen alrededor del verdadero valor de contenido de zinc. Pensamos que hay algún error sistemático en el método de medición.
Observar que la conclusión no coincide con la del test de signos.

Tests no paramétrico para muestras de a pares.

Ejemplo 6: En la clase pasada (ejemplo 4) consideramos mediciones de la concentración de zinc en 8 muestras por dos métodos diferentes. Lo usamos para ilustrar el test t para muestras de a pares. Consideremos ahora los mismos datos, pero donde un valor 6.8 fue cambiado por 8.5. Esto ha sido hecho artificialmente para ejemplificar, pero a veces estos errores groseros ocurren en los problemas reales.

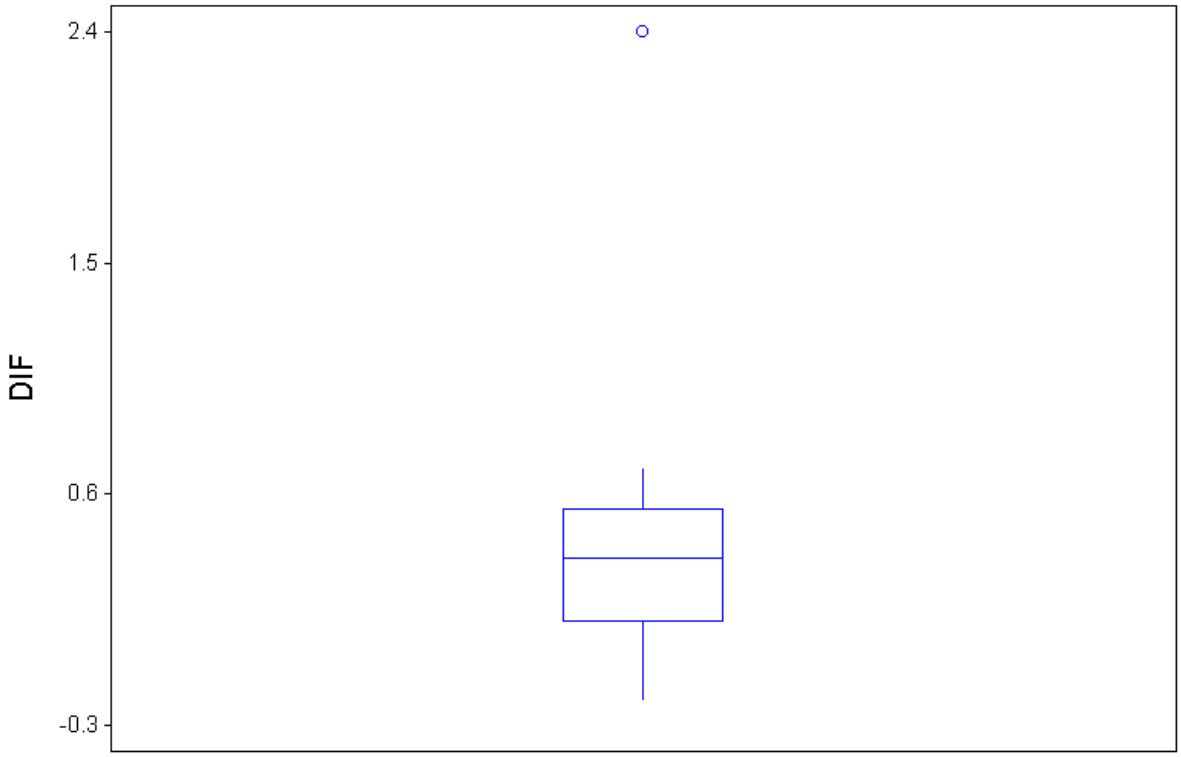
Muestra	Titulación con AEDT	Espectrometría atómica	Diferencia
1	7.2	7.6	0.4
2	6.1	**8.5** (antes 6.8)	2.4
3	5.2	5.2	0
4	5.9	5.7	-0.2
5	9.0	9.7	0.7
6	8.5	8.7	0.2
7	6.6	7.0	0.4
8	4.4	4.7	0.3
Promedio	**6.61**	**7.14**	**0.525**
DS	**1.57**	**1.81**	**0.805**

Si aplicásemos el test t para muestras de a pares a estos datos, obtendríamos:

```
 PAIRED T TEST FOR METODO1 - METODO2

NULL HYPOTHESIS: DIFFERENCE =  0
ALTERNATIVE HYP: DIFFERENCE <> 0

MEAN        -0.5250
STD ERROR    0.2846
LO 95% CI   -1.1979
UP 95% CI    0.1479
T             -1.84
DF                7
P            0.1076
```

y no detectaríamos diferencias entre ambos métodos.

Pero es un **grave error** aplicar el test t para muestras de a pares a estos datos, porque este test supone distribución normal de las diferencias, y acá hay un outlier severo. Si se grafica un box plot se lo observa claramente.



También el test de Wilk Shapiro detecta que los datos no son normales (ejercicio: hacerlo con el Statistix, dá P=0.0068).
En mi opinión, en este caso en que hay un único outlier muy claro, la mejor solución sería excluir el par de datos de la muestra nro. 2. Pero otra solución, especialmente útil cuando el o los valores atípicos no son tan evidentes y no es tan fácil decidir cuando ocurrió un error grosero y excluirlo, es utilizar algún test no paramétrico para muestras de a pares. Estos tests no son sensibles a outliers.
Sabemos que en el caso normal, hemos aplicado el test t de una muestra a las diferencias. Ahora procederemos en forma análoga: aplicaremos un test no parámétrico para una muestra a las diferencias y pondremos como H_0 que la mediana de las diferencias es cero.

H_0: mediana de la distribución de Di = 0
H_1: mediana de la distribución de Di ≠ 0

Vimos dos tests no paramétricos para una muestra. Apliquemos primero el test de signos. Para ello contamos cuántas diferencias tienen signo positivo y cuantas negativo. En nuestro ejemplo hay 6 diferencias positivas y sólo1 negativa (hay una más, que no se tiene cuenta porque vale cero). Se calcula entonces la probabilidad de que una variable Binomial(n=7,p=0.5) tome un valor 1 "o aún más extremo" P(X≤1). Este es el valor p del test "a una cola". Con el Statistix, en Probility Functions, obtenemos

Binomial(1,7,0.5)=0.0625

y el valor P del test a dos colas es P=2*0.0625=0.12

Esto lo hace automáticamente el Statistix, si vamos a Statistics, One, Two and Multi-Sample Tests, Sign test. Obtenemos

```
SIGN TEST FOR METODO1 - METODO2

NUMBER OF NEGATIVE DIFFERENCES             6
NUMBER OF POSITIVE DIFFERENCES             1
NUMBER OF ZERO DIFFERENCES (IGNORED)       1

PROBABILITY OF A RESULT AS
OR MORE EXTREME THAN OBSERVED         0.0625
```

Observemos que el valor que muestra el programa es para el test a una cola.

Otro test no paramétrico para muestras de a pares: test de rangos con signos de Wilcoxon.

Recordemos que las diferencias para el ejemplo anterior son:

0.4 2.4 0 -0.2 0.7 0.2 0.4 0.3

Igual que para el test del signo, no consideramos la diferencia=0 (no aporta nada a favor de que los valores del 1er método son más altos o más bajos). Luego ordenamos las diferencias, teniendo en cuenta su módulo:

-0.2 0.2 0.3 0.4 0.4 0.7 2.4

Calculamos los rangos de los módulos:

1.5 1.5 3 4.5 4.5 6 7

y ahora los separamos en dos grupos según el signo:

Rangos de las diferencias <0: 1.5
Rangos de las diferencias >0: 1.5 3 4.5 4.5 6 7

Si H_0 es cierta, se espera que la suma de los rangos de las diferencias <0 sea similar al de las diferencias >0
Suma de los rangos de las diferencias <0: 1.5
Suma de los rangos de las diferencias >0: 26.5

El estadístico del test de Wilcoxon es la diferencia entre estas dos sumas. Usamos tablas, o Statistix. Vamos a "Statistics", "One, Two and Multi Sample Tests", "Wilcoxon signed rank test":

```
WILCOXON SIGNED RANK TEST FOR METODO1 - METODO2

SUM OF NEGATIVE RANKS                          -26.500
SUM OF POSITIVE RANKS                           1.5000

EXACT PROBABILITY OF A RESULT AS OR MORE
EXTREME THAN THE OBSERVED RANKS (1 TAILED P-VALUE)  0.0156

NORMAL APPROXIMATION WITH CONTINUITY CORRECTION      2.028
TWO TAILED P-VALUE FOR NORMAL APPROXIMATION         0.0425
```

El valor exacto de P está calculado para el test a una cola, el del test a dos colas es el doble: P = 0.031.

Para este ejemplo de comparación de dos métodos de medición en distintas muestras (de a pares), lo que importa describir es las diferencias obtenidas con ambos métodos, así que una forma de redactar brevemente los resultados sería:

```
DESCRIPTIVE STATISTICS

                     DIF
N                      8
MEAN              0.5250
SD                0.8049
MINIMUM          -0.2000
1ST QUARTI        0.0500
MEDIAN            0.3500
3RD QUARTI        0.6250
MAXIMUM           2.4000
```

Las diferencias de las determinaciones obtenidas por ambos métodos (método2 menos método 1) tienen una mediana de 0.35 (P25,P75: 0.05, 0.62). Se rechaza la hipótesis de que estas diferencias se distribuyen alrededor del 0 (test de Wilcoxon, P=0.031).

Comparación del test del signo con el test de Wilcoxon: El test de signos y el test de Wilcoxon son "competidores": ambos pueden usarse para una muestra o para dos muestras de a pares cuando se sospecha que el modelo normal no se cumple y no se conoce la distribución de la variable estudiada.
¿Cuál conviene usar? ¿Cuál usa más información? Intuitivamente, Wilcoxon usa más información. ¿Cuál es más potente (tiene menor probabilidad de error II)? No se sabe…!
Si el modelo normal fuese cierto, el test t es el test óptimo (más potente). En este caso (normalidad) la eficiencia relativa (ER) del test de Wilcoxon con respecto al test t es $3/\pi \cong 0.955$ y la ER del test del signo con respecto al de t es $2/\pi \cong 0.637$ (son aproximaciones para muestras grandes). La siguiente tabla nos muestra la ER del test de Wilcoxon con respecto al test de signo para distintas distribuciones:

Distribución	ER del test de Wilcoxon con respecto al de signo
Normal	3/2 = 1.5
Uniforme	3
Doble exponencial	3/4=0.75

Se ve que, si los datos fuesen normales, se perdería poca eficiencia usando el test de Wilcoxon en lugar del test óptimo, pero se pierde bastante si se usa el test de signo. Sin embargo (esto va en contra de mi intuición), para algunas distribuciones, puede ser más potente el test del signo que el test de Wilcoxon, como es el caso de la distribución doble exponencial (una función de densidad similar a la exponencial, pero simétrica). Esta distribución tiene "colas muy pesadas" lo que quiere decir que pueden ocurrir valores muy alejados de la media (o mediana que coinciden para distribuciones simétricas). O sea es un modelo bajo el cuál se observan con alta probabilidad valores muy alejados del centro ("outliers severos" en un box plot).