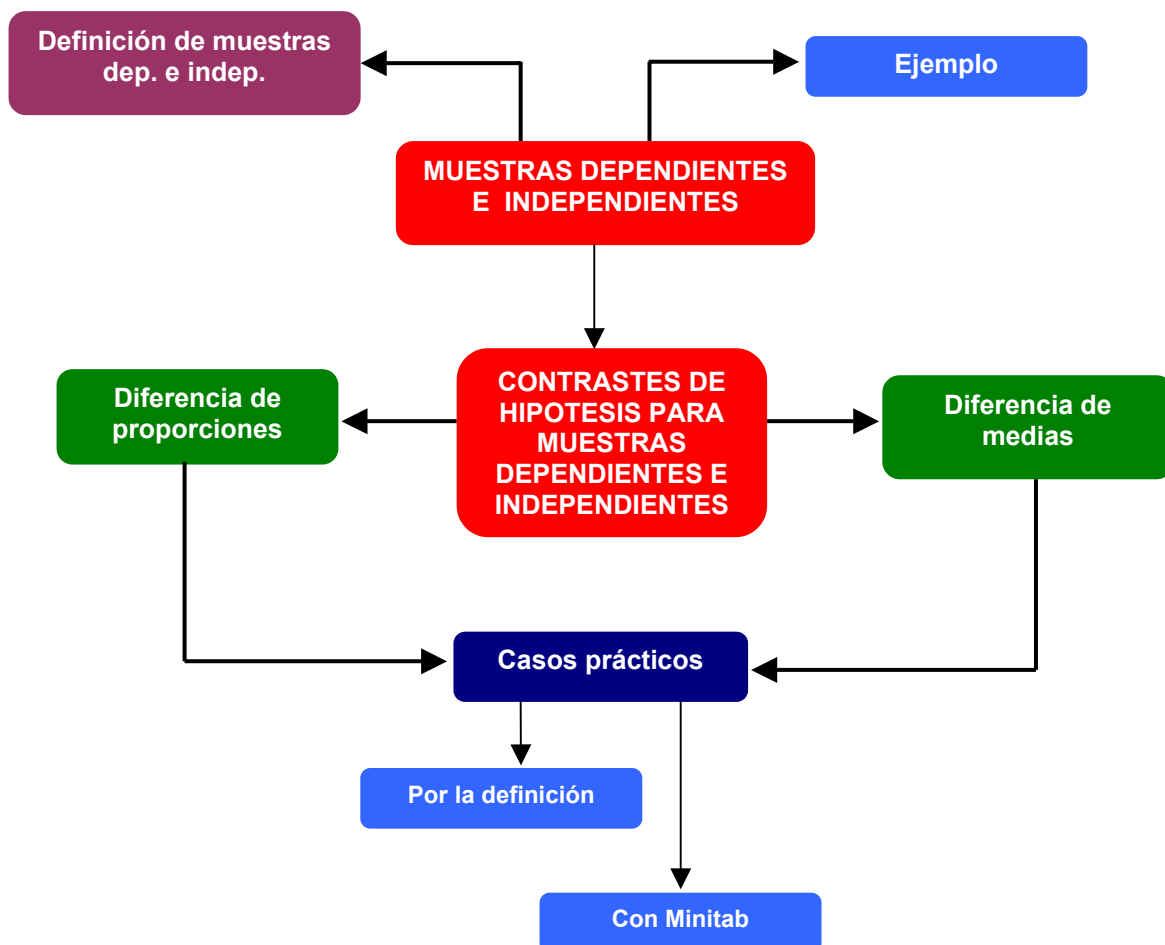


# CONTRASTE DE HIPÓTESIS DE DOS POBLACIONES

**Autores:** Ángel A. Juan ([ajuanp@uoc.edu](mailto:ajuanp@uoc.edu)), Máximo Sedano ([msedanoh@uoc.edu](mailto:msedanoh@uoc.edu)), Alicia Vila ([avilag@uoc.edu](mailto:avilag@uoc.edu)), Anna López ([alopezrat@uoc.edu](mailto:alopezrat@uoc.edu))

## MAPA CONCEPTUAL



## INTRODUCCIÓN

---

En este *math-block*, se pretende calcular e interpretar aquellos contrastes sobre la diferencia de medias y la diferencia de proporciones para dos poblaciones, que permita tomar decisiones acerca de qué población hay que tener en cuenta en comparación con la otra.

Además de calcular intervalos de confianza (rango de valores dentro del que se espera encontrar un determinado parámetro de la población), se realizará lo que llamaremos prueba de hipótesis acerca de una afirmación sobre un parámetro de la población. Para poner de manifiesto sus aplicaciones en la vida real, pondremos ejemplos de actividades en el ámbito económico-empresarial y en el informático. [2]

Hasta ahora, habíamos utilizado una sólo muestra aleatoria, comparando su media con un valor supuesto de la media poblacional, es decir, nos planteábamos si era posible que muestra con una media dada pudiera provenir de una población la media propuesta.

En este caso, extenderemos la idea anterior a dos muestras, preguntándonos si las medias de ambas son iguales o no, es decir, el planteamiento será razonar si es posible que las dos medias muestrales puedan provenir de dos poblaciones idénticas.

## OBJETIVOS

---

- Entender la diferencia entre muestras independientes y dependientes.
- Realizar los contrastes de diferencia de medias y de proporciones en dos muestras independientes.
- Saber interpretar los resultados estadísticos obtenidos.
- Tomar conclusiones de cualquier índole a través de los contrastes de hipótesis de dos poblaciones.

## CONOCIMIENTOS PREVIOS

---

Es recomendable haber leído, previamente, el *math-block* “Estimación puntual e intervalos de confianza” y “Contraste de hipótesis de una población”, así como el manual introductorio a Minitab y los ejercicios con Minitab asociados a los *math-blocks* anteriores.

## CONCEPTOS FUNDAMENTALES

---

### □ Diferencia entre muestras independientes y dependientes

Dos muestras son independientes o dependientes entre sí, en función de si las observaciones de las muestras se han obtenido de los mismos individuos u objetos o no.

Si ambas muestras se obtienen de distintos individuos, máquinas, empresas, objetos, etc...no hay nada en común en dichas muestras lo que hace que ambas sean "**independientes**".

Sin embargo, si las observaciones o valores de ambas muestras se obtienen de los mismos individuos, empresas, agentes, etc., diremos que hay algo en común en dichas muestras por lo que serán muestras "**dependientes**" o "**no independientes**".

### Ejemplo:

Supongamos que queremos comparar los beneficios empresariales del sector de la construcción entre el año 2001 y el año 2002. Para ello podemos tomar una muestra aleatoria formada por 50 empresas constructoras de todo el país y medimos sus beneficios en el año 2001.

A continuación, para poder comparar los beneficios del sector con el año 2002, se toma otra muestra aleatoria distinta con otras 30 empresas constructoras y analizamos sus beneficios en el año 2002.

En este caso se trata de muestras "**independientes**" puesto que las observaciones de ambas muestras se toman de distintos individuos, en este caso distintas empresas.

Sin embargo, si en el año 2002 observamos los beneficios de las mismas 50 empresas constructoras de la muestra del año 2001, estaríamos por tanto ante muestras "**dependientes**", o "**no independientes**".

Supongamos ahora que, al iniciar el semestre, seleccionamos al azar 30 alumnos matriculados en Estadística y les pasamos un test de conocimientos previos. Al final del semestre, seleccionamos otros 30 alumnos al azar y les pasamos un test de conocimientos adquiridos durante el curso. En tal caso, consideraríamos ambas muestras como independientes. Por el contrario, si el test de conocimientos adquiridos se realizase a los mismos 30 alumnos que hicieron el test inicial, entonces hablaríamos de muestras dependientes.

❑ **Contrastes de hipótesis en muestras dependientes**

**1. Contraste de diferencia de medias en dos muestras dependientes**

A las personas que sufren de tensión alta, se les recomienda seguir una dieta libre de sal. Queremos realizar un estudio para comprobar si esta dieta es efectivamente ventajosa. Para el estudio se estudió una muestra de 8 personas y se tomó la tensión antes de empezar la dieta y dos semanas después. Los resultados obtenidos fueron:

Antes	93	106	87	92	102	95	88	110
Después	92	102	89	92	101	96	88	105

Denotamos  $\mu_A$  y  $\mu_B$  a las medias poblacionales de tensión antes y después de empezar la dieta, respectivamente. De este modo, el contraste de hipótesis que debemos plantear es:

$$\begin{aligned} H_0 : \mu_A &= \mu_B \\ H_1 : \mu_A &< \mu_B \quad (\neq, >) \end{aligned} \quad (1)$$

**Observación:** En el caso que tuviéramos la creencia de que el hacer dieta supone una disminución de la presión de 2 puntos entonces el contraste deberíamos plantearlo como:

$$\begin{aligned} H_0 : \mu_A - \mu_B &= 2 \\ H_1 : \mu_A - \mu_B &< 2 \quad (\neq, >) \end{aligned}$$

Para realizar el contraste observamos en primer lugar que las muestras de antes,  $X_A$ , y después de la dieta,  $X_B$ , son dependientes, puesto que se han tomado del mismo individuo.

Para realizar este contraste consideramos la diferencia de ambas muestras:  $d = X_A - X_B$ . Denotaremos por  $\mu_d = \mu_A - \mu_B$  y  $\sigma_d$  a su media y desviación estándar respectivamente. Observamos pues que el contraste anterior es equivalente al contraste:

$$\begin{aligned} H_0 : \mu_d &= 0 \\ H_1 : \mu_d &< 0 \quad (\neq, >) \end{aligned}$$

- ❑ **Supuesto:**  $X_A$  y  $X_B$  siguen una distribución normal.
- ❑ **Observación:**  $d = X_A - X_B \sim N(\mu_A - \mu_B, \sigma_d)$ .

El **intervalo de confianza**, a nivel  $1-\alpha$ , para  $\mu_d = \mu_A - \mu_B$  viene dado por la expresión:

$$\bar{d} \pm t(n-1, \alpha/2) * S_d$$

donde  $t(n-1, \alpha/2)$  es el valor que, en una t-Student con  $n-1$  grados de libertad, deja a su derecha un área de  $\alpha/2$ , y  $S_d$  es la desviación estándar muestral de la v.a.  $d$ .

El **estadístico de contraste** para el test  $\begin{cases} H_0 : \mu_d = \mu_0 \\ H_1 : \mu_d \neq \mu_0 \end{cases}$  (o bien  $<$  ó  $>$ ) es:

$$t^* = \frac{\bar{d} - \mu_d}{S_d} \approx t - Student(n-1)$$

En nuestro ejemplo  $\mu_0 = 0$ .

En el caso de la observación donde sospechábamos que la tensión bajaba dos puntos,  $\mu_0 = 2$ .

Así siguiendo nuestro ejemplo:  $\bar{d} = -1$  y  $S_d = 2.390$ .

Entonces con un 95% de confianza  $\mu_d \in (-3,1)$ .

Y el estadístico de contraste es  $t^* = -1.18$ . Ahora bien, mirando la tabla de la  $t$  ( $7, 0.05$ )  $= 1.895$ . De este modo, como  $t^* < -1.18$  no tenemos evidencias significativas que realmente hacer dieta sea ventajoso.

## □ Contrastes de hipótesis en muestras independientes

### 1. Contraste de diferencia de medias en dos muestras independientes

Para realizar esta prueba, se requiere de tres suposiciones:

- Las poblaciones muestreadas tienen una distribución normal
- Las dos muestras son independientes
- Las desviaciones estándar de ambas poblaciones son iguales

Supongamos que un estadístico de recursos humanos desea analizar si los salarios por hora de los obreros semiespecializados son los mismos, mayores o menores en Madrid que en Barcelona. Los datos muestrales obtenidos son los siguientes:

Ciudad	Salarios medios por hora de la muestra	Desviación estándar de la muestra	Tamaño de la muestra
Madrid	8,95 euros	0,4 euros	200
Barcelona	9,1 euros	0,6 euros	175

Supongamos que la empresa desea probar la hipótesis en el nivel de significación del 5% de que (en promedio) no hay diferencia entre los salarios por hora de los trabajadores semiespecializados de las dos ciudades.

Llamamos  $\mu_M$  y  $\mu_B$  a las medias de salarios por hora de los trabajadores de Madrid y de Barcelona, respectivamente. Con esta notación el anterior contraste de hipótesis equivale a formular:

$$\begin{cases} H_0 : \mu_M = \mu_B \\ H_1 : \mu_M \neq \mu_B \text{ (o bien } < \text{ ó } > \text{)} \end{cases}$$

Notamos que en este ejemplo tomaremos el contraste bilateral, es decir, la hipótesis alternativa  $H_1$  es un "desigual" y no un "mayor que" o "menor que" puesto que no nos dan ninguna pista para saber en que lugar realmente creemos que en promedio el salario es mayor. Si en el enunciado se detallará que hay sospechas de que en Madrid se cobra un salario superior al de Barcelona entonces la hipótesis alternativa se traduciría por  $\mu_M > \mu_B$ , y a la inversa en caso contrario.

Observamos además que tal como hemos tomados las muestras éstas provienen de grupos independientes. Realizaremos pues un **contraste de hipótesis de muestras independientes**.

Denotamos:

$\bar{X}_M$ : Media de la muestra de los salarios de Madrid,

$S_M$ : Desviación estándar de la muestra de los salarios de Madrid,

$n_M$ : Número de individuos de la muestra de Madrid.

$\bar{X}_B$ : Media de la muestra de los salarios de Barcelona,

$S_B$ : Desviación estándar de la muestra de los salarios de Barcelona,

$n_B$ : Número de individuos de la muestra de Barcelona.

En nuestro ejemplo:  $\bar{X}_M = 8,95$ ,  $S_M = 0,4$ ,  $n_M = 200$  y  $\bar{X}_B = 9,1$ ,  $S_B = 0,6$ ,  $n_B = 175$ .

Bajo el supuesto que los salarios (por hora) se distribuyen mediante una distribución Normal tenemos:

$$\bar{X}_M - \bar{X}_B \approx N\left(\mu_M - \mu_B, \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_B^2}{n_B}}\right)$$

donde  $\sigma_M$  i  $\sigma_B$  son las desviaciones poblacionales de los salarios de Madrid y Barcelona, respectivamente.

El intervalo de confianza de nivel de confianza  $(1 - \alpha)$  para  $\mu_M - \mu_B$  viene dado por la expresión:

$$(\bar{X}_M - \bar{X}_B) \pm t(\min\{n_M - 1, n_B - 1\}, \alpha/2) \sqrt{\frac{S_M^2}{n_M} + \frac{S_B^2}{n_B}}$$

donde  $t(\min\{n_M - 1, n_B - 1\}, \alpha/2)$  es el valor que, en una t-Student con los grados de libertad indicados, deja a su derecha un área de  $\alpha/2$ , y  $S_M$ ,  $S_B$  son las desviaciones estándar de las muestras.

El **estadístico de contraste** para el test junto a su distribución es:

$$t^* = \frac{(\bar{X}_M - \bar{X}_B) - (\mu_M - \mu_B)_{H_0}}{\sqrt{\frac{S_M^2}{n_M} + \frac{S_B^2}{n_B}}} \approx t(\min\{n_M - 1, n_B - 1\}, \alpha/2)$$

La expresión  $(\mu_M - \mu_B)_{H_0}$  es el valor de la diferencia bajo la hipótesis nula. En nuestro ejemplo  $H_0: \mu_M - \mu_B = 0$  por lo tanto  $(\mu_M - \mu_B)_{H_0} = 0$ .

**Observación 1:**

En algunos casos lo que nos interesa es discutir si el promedio de las muestras difieren significativamente o no en un número  $k$ . Por ejemplo si en el enunciado del ejemplo anterior nos expusieran lo siguiente:

*“Por cuestiones de impuestos sabemos que en Madrid los salarios son 1Euro por hora más que en Madrid, pero sospechamos que son más de un euro”*

De este modo, el contraste de hipótesis se traduce formalmente como:

$$\begin{cases} H_0 : \mu_M - \mu_B = 1 \\ H_1 : \mu_M - \mu_B > 1 \end{cases}$$

Para contrastar esta hipótesis utilizamos el mismo estadístico  $t^*$  pero en este caso  $(\mu_M - \mu_B)_{H_0} = 1$ .

Sigamos con nuestro ejemplo . Si calculamos  $t^*$ :

$$t^* = \frac{(8,95 - 9,1) - 0}{\sqrt{\frac{0,4^2}{200} + \frac{0,6^2}{175}}} = -2,83$$

Entonces como  $\min(199,174) = 174$ , para 174 grados de libertad, si vamos a la tabla de la t-student a los grados de libertad más cercanos, 150, podemos ver que el área que hay por debajo de  $t^* = -2,83$ , será menor que 0,005 que es el área por debajo del valor  $t = -2,609$  por tanto el p-valor, si el contraste es unilateral, será menor que  $2 \cdot 0,005 = 0,01$ .

Como el p-valor es menor que el nivel de significación, si cogemos el 5%, por tanto rechazaremos la hipótesis nula y por tanto existe evidencia estadística de que sí existen diferencias significativas en los salarios de los trabajadores semiespecializados en las dos ciudades.

Si el contraste hubiera sido unilateral por la derecha o por la izquierda, es decir, en la hipótesis alternativa, hubiera aparecido  $>$  ó  $<$ , entonces el p-valor de  $t = -2,83$  sería menor que 0,005 y habría que compararlo con el nivel de significación para rechazar o no la hipótesis nula.

### Ejemplo:

En el campo de la informática, se hace un experimento en el que se miden las velocidades de los Pentium frente a los correspondientes AMD. Los resultados obtenidos son los siguientes:

$$\begin{aligned} \bar{X}_M &= 110 & \bar{X}_B &= 100 \\ S_M^2 &= 35 & S_B^2 &= 26 \\ n_M &= 61 & n_B &= 61 \end{aligned}$$

Contrastar la hipótesis de que la velocidad media es la misma para ambos procesadores. Nivel de significación del 1%.

Solución 1:

Estamos en el caso de dos muestras independientes de 50 elementos para cada una de ellas. El intervalo de confianza para la diferencia de medias viene dado por:

$$(\bar{X}_M - \bar{X}_B) \pm t(\min\{n_M - 1, n_B - 1\}, \alpha/2) \sqrt{S_M^2/n_M + S_B^2/n_B}$$

en nuestro caso, tenemos una t-student con 60 grados de libertad con  $\alpha/2 = 0.01/2 = 0.005$ , quedaría:

$$(110 - 100) \pm 2.6603 \sqrt{\frac{35}{61} + \frac{26}{61}}$$

$$10 \pm 2.66$$

El intervalo de confianza para la diferencia de medias al 99% es (7.34, 12.66).

Como el intervalo no contiene el valor 0, rechazamos que las medias de los Pentium y los AMD sean iguales.

### Solución 2:

Podemos realizar un contraste de hipótesis para contestar la cuestión de forma directa.

$$\begin{cases} H_0 : \mu_M = \mu_B \\ H_1 : \mu_M \neq \mu_B \end{cases}$$

El estadístico del contraste es: 
$$t^* = \frac{(\bar{X}_M - \bar{X}_B) - (\mu_{M_0} - \mu_{B_0})_{H_0}}{\sqrt{S_M^2/n_M + S_B^2/n_B}} = \frac{10}{1} = 10$$

El p-valor será la probabilidad de que en una distribución t-student con 60 grados de libertad obtengamos un valor superior a 10 o inferior a -10. El p-valor en este ejercicio es prácticamente 0. Podemos rechazar la hipótesis nula a cualquier nivel de significación ya que la probabilidad de equivocarnos al rechazar es prácticamente cero.

## 2. Contraste de diferencia de proporciones en dos muestras independientes.

Supongamos que con fines de la declaración del impuesto IRPF, el Ayuntamiento de una determinada ciudad ha estado utilizando dos métodos para listar propiedades. El primero requiere que el dueño de la propiedad aparezca en persona ante el recabador de la información; y el segundo método permite que el propietario envíe por correo una declaración fiscal con la información requerida. El Alcalde de la ciudad considera que el método en el cual se requiere la presencia de la persona produce menores errores que el otro. Autoriza la realización de un examen de 100 listas hechas con el primer método, donde el 71% no tiene errores y de 90 listas tomadas de los datos llegados por correo, donde el 64,4% no tiene errores.

El Ayuntamiento desea probar, al nivel de significación del 5%, si existe diferencia entre la información recogida entre los dos métodos.

En este caso queremos contrastar si hay diferencias o no entre las proporciones de errores en el método en el que se requiere presencia respecto a las que no se requiere presencia. Si llamamos  $P_A$  a la proporción de errores (poblacionales) cometidos con el método que se requiere presencia y  $P_B$  a la proporción de errores cometidos con el método sin presencia, el contraste anterior es equivalente a formular:

$$\begin{cases} H_0 : P_A = P_B \\ H_1 : P_A \neq P_B \quad (\text{o bien } < \text{ ó } >) \end{cases}$$

Las muestras en este caso son **independientes**. Este hecho es fundamental para que se cumplan los resultados que damos a continuación.

Denotamos:

$X_A$ : número de errores al realizar  $n_A$  pruebas en el método en el que se requiere presencia (poblacional).

$X_B$ : número de errores al realizar  $n_B$  pruebas en el método en el que NO se requiere presencia (poblacional).

Y definimos las proporciones de cada muestra como:  $p_A = X_A / n_A$ , y  $p_B = X_B / n_B$ .

En el ejemplo nos dan una realización de  $p_A$  y  $p_B$  al coger un par de muestras de la población. Estos valores son  $p'_A=0,71$  y  $p'_B=0,644$ .

Para muestras suficientemente grandes ( $n_A, n_B > 30$ ) se puede demostrar que:

$$(p_A - p_B) \approx N\left(P_A - P_B, \sqrt{\frac{P_A(1-P_A)}{n_A} + \frac{P_B(1-P_B)}{n_B}}\right)$$

Sabemos que:  $X_A \approx B(n_A, P_A)$  y  $X_B \approx B(n_B, P_B)$

Ahora bien, para muestras grandes (recordamos  $n \geq 20$ ,  $n^*p \geq 5$ , y  $n^*(1-p) \geq 5$ ) ambas se aproximan a una normal:

$$X_A \approx N(n_A P_A, \sqrt{n_A P_A (1-P_A)}) \quad \text{y} \quad X_B \approx N(n_B P_B, \sqrt{n_B P_B (1-P_B)})$$

Con lo cual este resultado junto a la definición de  $p_A$  y  $p_B$  obtenemos el resultado anterior.

El **intervalo de confianza**, a nivel  $1-\alpha$ , para  $p_A-p_B$  viene dado por la expresión:

$$(p'_A - p'_B) \pm z\left(\frac{\alpha}{2}\right) \sqrt{\frac{p'_A(1-p'_A)}{n_A} + \frac{p'_B(1-p'_B)}{n_B}}$$

donde  $z(\alpha/2)$  es el valor que, en una normal estándar, deja a su derecha un área de  $\alpha/2$ .

El **estadístico de contraste** para el test será:

$$Z^* = \frac{(p'_A - p'_B)}{\sqrt{p'_p (1-p'_p) \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Podemos escoger diferentes versiones del valor  $p'_p$ . (consultar literatura para ver opciones). Una posible buena aproximación que utilizamos en los ejemplos que siguen

es  $p'_p = \frac{n_A p'_A + n_B p'_B}{n_A + n_B}$  la cual es la estimación de la porción completa de éxitos de las

poblaciones combinadas.

De este modo para discutir el contraste en nuestro ejemplo calculamos:

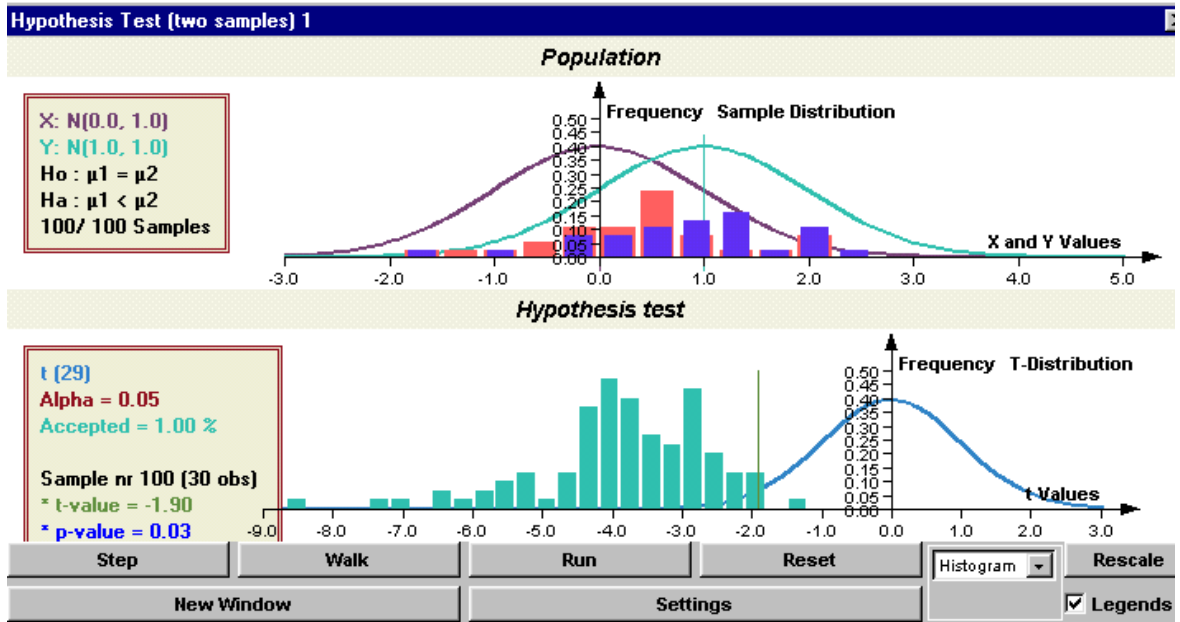
$$p'_p = \frac{n_A p'_A + n_B p'_B}{n_A + n_B} = \frac{100 * 0,71 + 90 * 0,644}{100 + 90} = \frac{71 + 58}{190} = 0,6789$$

$$Z^* = \frac{(p'_A - p'_B)}{\sqrt{p'_p (1 - p'_p) \left( \frac{1}{n_A} + \frac{1}{n_B} \right)}} = \frac{0,71 - 0,644}{\sqrt{0,6789 * (1 - 0,6789) * \left( \frac{1}{100} + \frac{1}{90} \right)}} = 0,9729$$

El último paso será calcular el p-valor de  $z = 0,9729$ . Como el contraste es bilateral por las dos colas, debemos buscar el área que hay por encima de  $z = 0,9729$  y el área que hay por debajo de  $z = -0,9729$  que será, p-valor =  $2 * 0,1660 = 0,332$ , porque el área por debajo de  $z = 0,9729$  es  $1 - 0,8340$ , mientras el área por debajo de  $z = -0,9729$  es  $0,1660$ .

Como el p-valor es  $0,332$  que es mayor que el nivel de significación del 5%, no rechazaremos la hipótesis nula, por lo tanto existe evidencia estadística de que los dos métodos de recogida de información sobre las propiedades de esta ciudad son igualmente fiables.

En el siguiente enlace: <http://fitbw2.rug.ac.be/iloapp/Applets/Ap6b.html>, podemos encontrar una representación gráfica de este concepto de Contraste de hipótesis para dos muestras. Obtendremos un gráfico similar al siguiente, donde podemos modificar los datos de entrada y observar las variaciones resultantes :



## CASOS PRÁCTICOS CON SOFTWARE

### 1. Contraste de diferencia de medias de dos muestras dependientes

Hemos pedido a 10 personas que evalúen, en base a unos criterios preestablecidos, la calidad y usabilidad de un determinado software informático. Las puntuaciones varían entre un mínimo de 0 y un máximo de 15. Pasados tres meses, las mismas 10 personas repiten el proceso de evaluación. Los resultados obtenidos, que introduciremos en las columnas C1 y C2, son los siguientes:

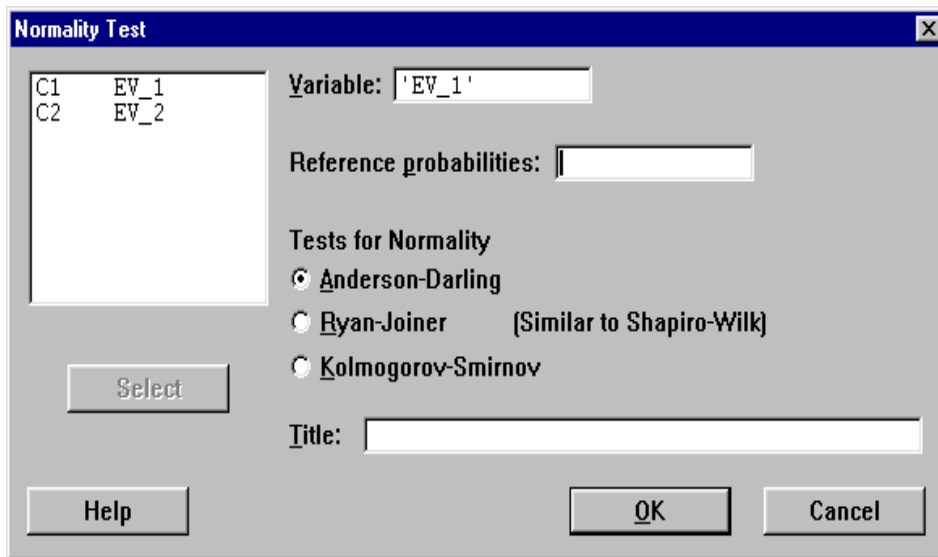
Persona	EV_1	EV_2
1	13,2	14,0
2	8,2	8,8
3	10,9	11,2
4	14,3	14,2
5	10,7	11,8
6	6,6	6,4
7	9,5	9,8
8	10,8	11,3
9	8,8	9,3
10	13,3	13,6

Nuestro objetivo es doble: por un lado, pretendemos calcular un intervalo de confianza, a nivel del 95%, para  $\mu_A - \mu_B$ ; por otro, contrastar las hipótesis:  $H_0: \mu_A - \mu_B = 0$  vs.  $\mu_A - \mu_B \neq 0$ .

En primer lugar, comprobaremos el supuesto de que las poblaciones siguen una distribución aproximadamente normal:

Seleccionamos: *Stat > Basic Statistics > Normality Test* :

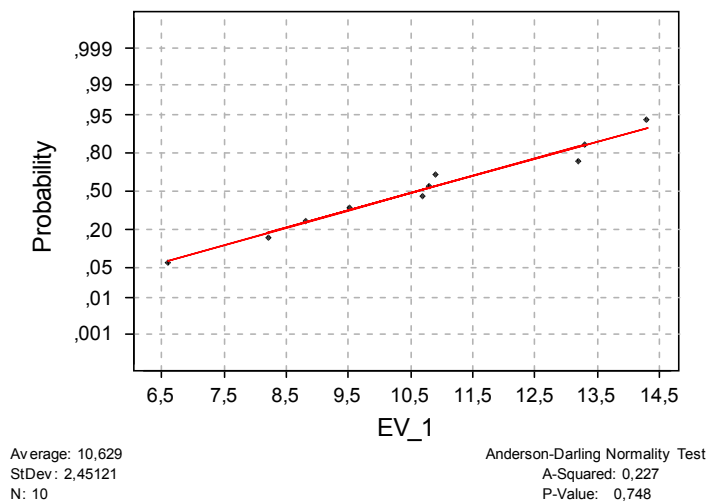
Completamos la ventana siguiente con cada una de las variables a estudiar:



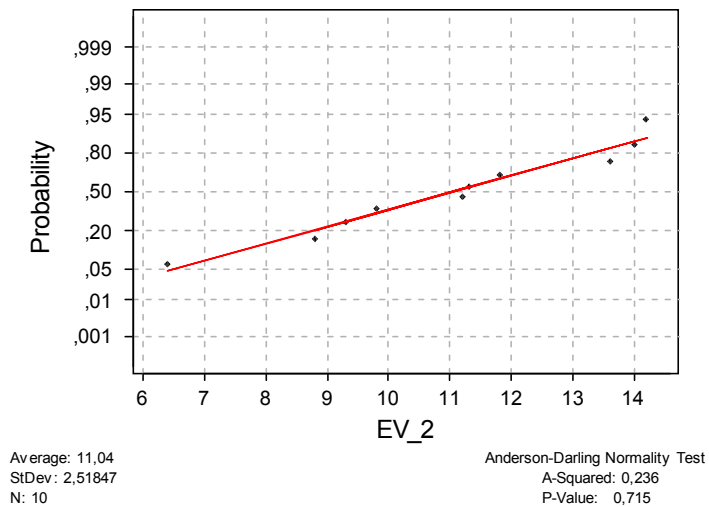
En los gráficos resultantes se observa que no hay indicios para dudar de que se cumple el supuesto de normalidad ya que los puntos se encuentran muy próximos a las respectivas rectas.

Además, los gráficos nos proporcionan también el p-valor asociado al **test de normalidad de Anderson-Darling**, siendo dicho p-valor suficientemente grande en ambos casos como para no descartar la hipótesis nula de este contraste: que los datos siguen una distribución normal.

Normal Probability Plot



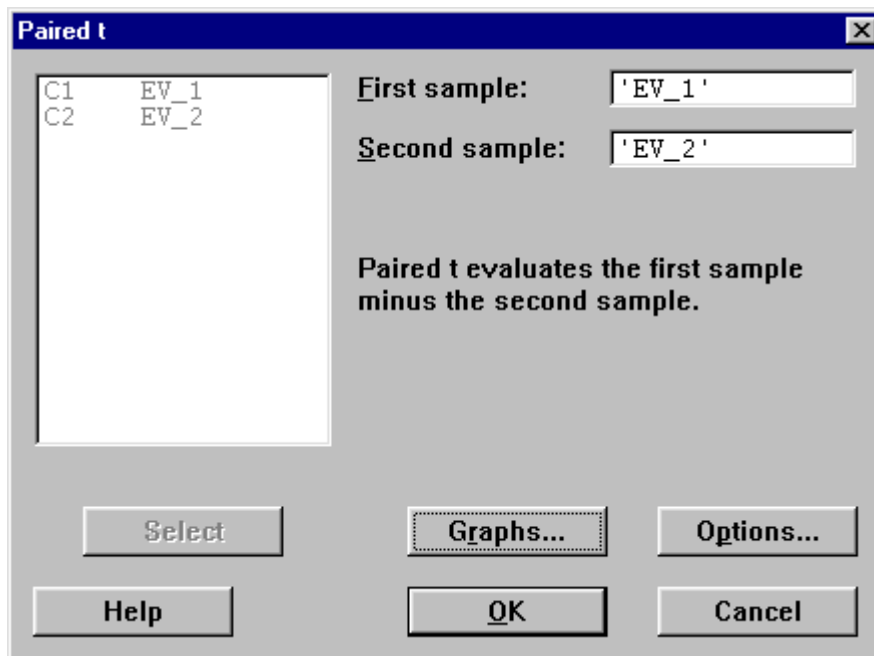
Normal Probability Plot

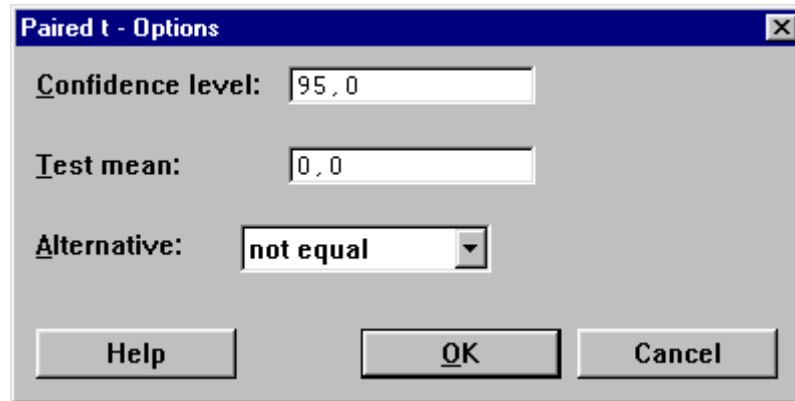


Pasamos pues a realizar las inferencias ya comentadas sobre  $\mu_A - \mu_B$  :

Seleccionamos: *Stat > Basic Statistics > Paired t* :

Completamos la ventana principal y la de opciones como se muestra en las imágenes:





### Paired T-Test and Confidence Interval

Paired T for EV\_1 - EV\_2

	N	Mean	StDev	SE Mean
EV_1	10	10,629	2,451	0,775
EV_2	10	11,040	2,518	0,796
Difference	10	-0,411	0,387	0,122

95% CI for mean difference: (-0,688; -0,134)

T-Test of mean difference = 0 (vs not = 0): T-Value = -3,36 P-Value = 0,008

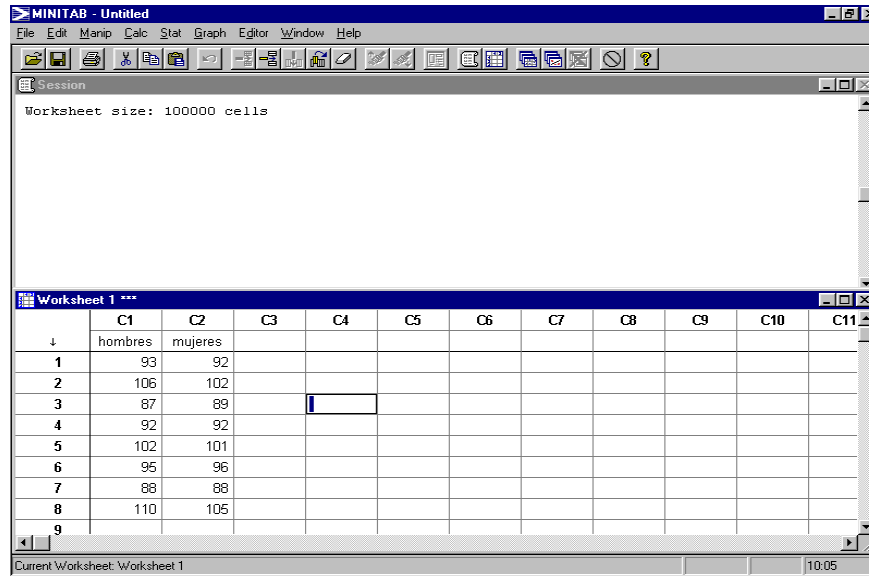
Los resultados obtenidos nos dicen que, en base a las observaciones registradas, hay una probabilidad de 0,95 de que  $\mu_A - \mu_B$  sea un valor del intervalo (-0,688 , -0,134). Además, con un p-valor de 0,008 también podemos afirmar que hay indicios suficientes como para descartar la hipótesis nula. Por tanto, parece sensato pensar que las dos medias poblacionales son distintas. Notar que esta conclusión es coherente con que el valor 0 no esté incluido en el intervalo de confianza hallado para la diferencia de ambas medias.

## 2. Contraste de diferencia de medias en dos muestras independientes

- Una agencia de valores desea analizar qué éxito han tenido sus nuevos comerciales en la obtención de nuevos clientes para la intermediación bursátil. Para ello, se tomaron dos muestras de 8 comerciales hombres y 8 comerciales mujeres donde se observó la cantidad de nuevas cuentas conseguidas por cada comercial (hombre o mujer) en el primer mes de trabajo.

Comerciales hombre	93	106	87	92	102	95	88	110
Comerciales mujer	92	102	89	92	101	96	88	105

Primero, insertamos los valores anteriores en el espacio de trabajo del Minitab:

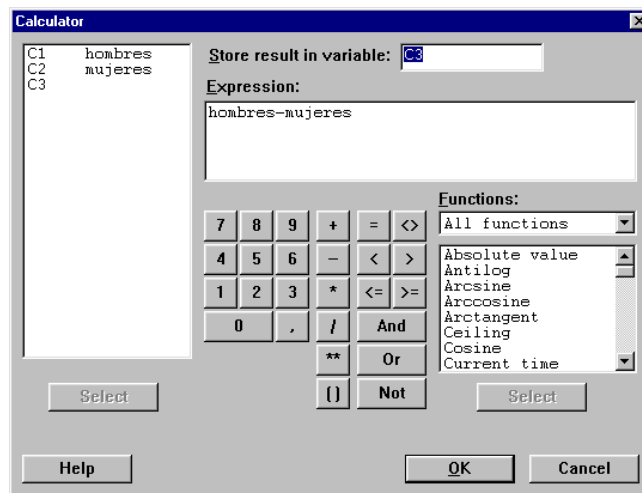


Worksheet size: 100000 cells

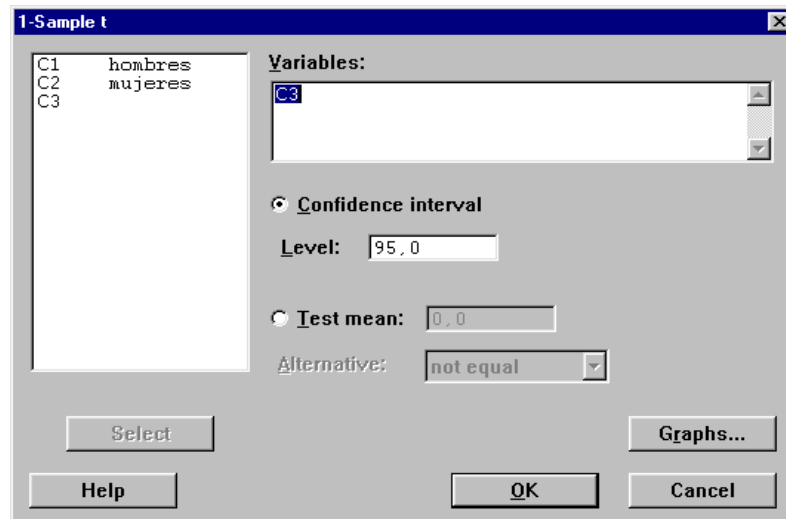
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
↓	hombres	mujeres									
1	93	92									
2	106	102									
3	87	89									
4	92	92									
5	102	101									
6	95	96									
7	88	88									
8	110	105									
9											

- a) Construir una nueva columna con las diferencias entre C1 y C2. Hallar el intervalo de confianza a nivel del 95% para la media de dichas diferencias.

Seleccionamos *Calc > Calculator* :



Así generamos una nueva columna formada por la diferencia entre los valores registrados. Seleccionamos ahora *Stat > Basic Statistics > 1-Sample t* :



T Confidence Intervals					
Variable	N	Mean	StDev	SE Mean	95.0 % CI
C4	8	1,000	2,390	0,845	( -1,000 ; 3,000)

De este resultado deducimos que en el 95% de los casos la diferencia de nuevos clientes conseguidos entre comerciales hombres y mujeres estará entre  $-1$  y  $3$ , es decir, un máximo de 3 nuevos clientes.

- b) Realizar un contraste de hipótesis, a un nivel de significación  $\alpha=0,05$ , para determinar si las dos medias muestrales son significativamente diferentes.

Planteamos el siguiente contraste de hipótesis bilateral aprovechando la columna de diferencias anterior:

$$H_0 : \mu_A = \mu_B;$$

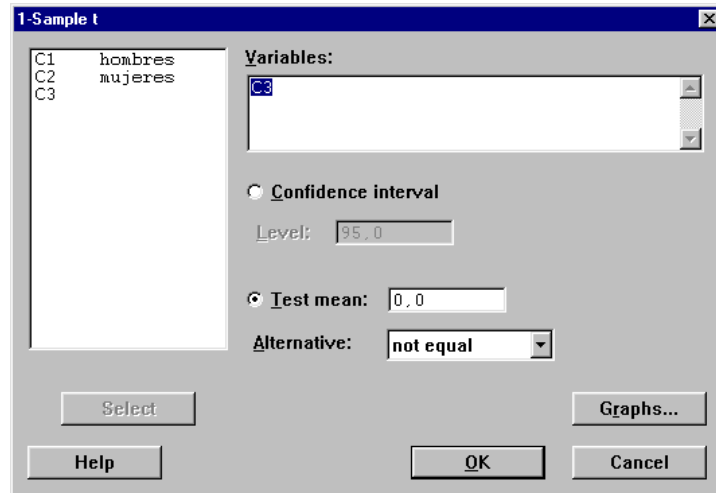
$$H_1 : \mu_A \neq \mu_B;$$

De donde,

$$H_0 : \mu_{B-A} = \mu_B - \mu_A = 0 ;$$

$$H_1 : \mu_{B-A} = \mu_B - \mu_A \neq 0;$$

Seleccionamos *Stat > Basic Statistics > 1-Sample t* :



Obteniendo el siguiente resultado:

T-Test of the Mean						
Test of $\mu = 0.000$ vs $\mu \text{ not } = 0.000$						
Variable	N	Mean	StDev	SE Mean	T	P
C4	8	1,000	2,390	0,845	1,18	0,28

Observar que el p-valor obtenido 0,28 es mucho mayor que 0,05 por lo cual no hay indicios suficientes para rechazar la hipótesis nula. Esto quiere decir que las dos medias no son significativamente diferentes.

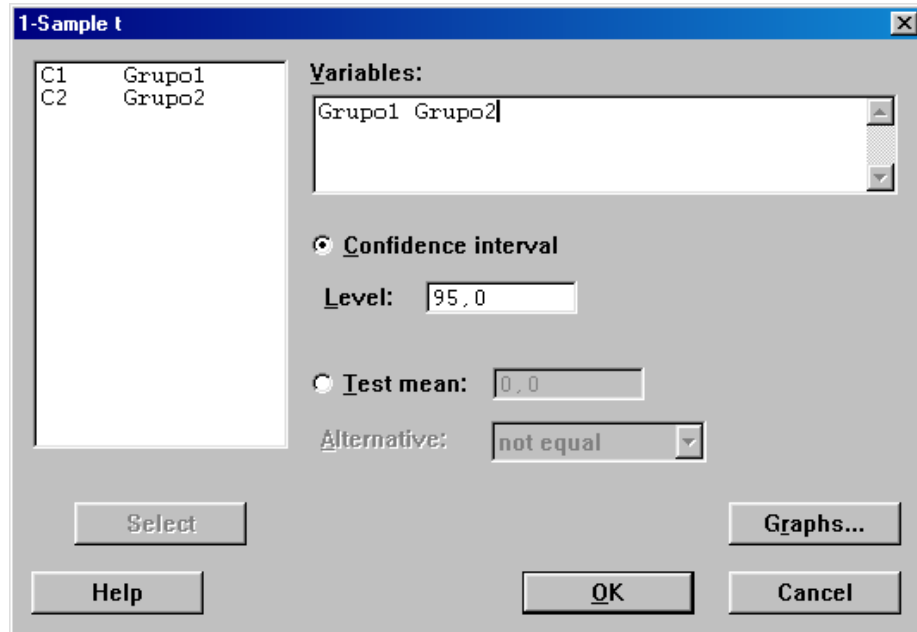
De ello se deduce que la productividad en la captación de nuevos clientes no depende de si el comercial es hombre o mujer en el primer mes de trabajo.

- Supongamos que disponemos los datos sobre las calificaciones obtenidas por dos grupos de estudiantes de Estadística de la UOC.

Grupo 1	Grupo 2
5	6.25
7.5	5.75
6	5
2.5	4.75
8	8
9	9
7	7.5
6	8
4	9
3.75	10
9	
10	
8.25	
9	
6	

- a) Calcular la un intervalo de confianza para cada una de las dos poblaciones al nivel de confianza del 95%. Comentar los resultados.

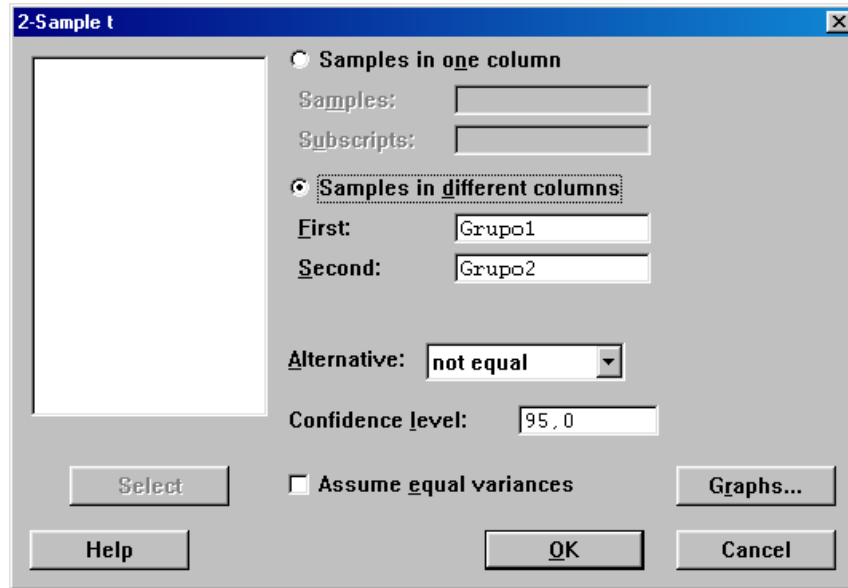
Para calcular un intervalo de confianza debemos usar las opciones *Stat > Basic Statistic > 1-Sample t*, pues no tenemos información acerca de la varianza de la población.



Variable	N	Mean	StDev	SE Mean	95,0 % CI
Grupo1	15	6,733	2,229	0,576	( 5,499; 7,968)
Grupo2	10	7,325	1,807	0,571	( 6,032; 8,618)

Si nos fijamos en los dos intervalos de confianza, estos se solapan. Esto implica que si estamos interesados en comparar las medias de ambas poblaciones, estas media pertenecen a intervalos con parte en comun, lo cual hace pensar que estas medias poblacionales, es decir, las medias del grupo1 y del grupo2 pueden ser iguales. En el siguiente apartado veremos si tras contrastar la hipótesis de igualdad de medias podemos concluir lo mismo.

- b) Calcular un intervalo de confianza para la diferencia de medias. Utilizando este intervalo contrastar la hipótesis de que la medias en los dos grupos no difieren.



**2-Sample t**

Samples in one column

Samples:

Subscripts:

Samples in different columns

First:

Second:

Alternative:

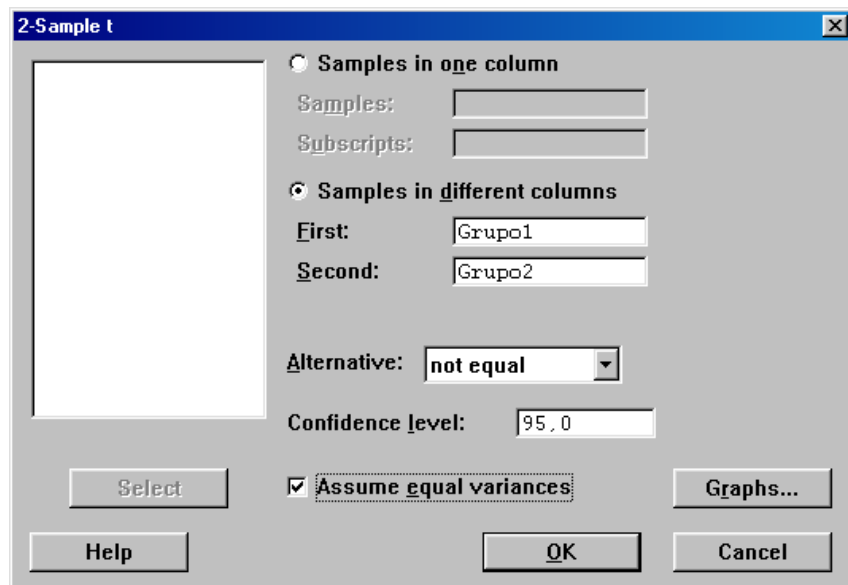
Confidence level:

Assume equal variances

```
Two sample T for Grupo1 vs Grupo2

N      Mean      StDev    SE Mean
Grupo1  15      6,73      2,23      0,58
Grupo2  10      7,33      1,81      0,57

95% CI for mu Grupo1 - mu Grupo2: ( -2,34;  1,16)
T-Test mu Grupo1 = mu Grupo2 (vs not =): T = -0,70  P = 0,49  DF = 23
Both use Pooled StDev = 2,07
```



**2-Sample t**

Samples in one column

Samples:

Subscripts:

Samples in different columns

First:

Second:

Alternative:

Confidence level:

Assume equal variances

```
Two sample T for Grupo1 vs Grupo2

N      Mean      StDev    SE Mean
Grupo1  15      6,73      2,23      0,58
Grupo2  10      7,33      1,81      0,57

95% CI for mu Grupo1 - mu Grupo2: ( -2,28;  1,09)
T-Test mu Grupo1 = mu Grupo2 (vs not =): T = -0,73  P = 0,47  DF = 21
```

- c) Que error de equivocarnos, si concluimos que hay diferencias entre las poblaciones, deberíamos estar dispuestos a asumir.

Si observamos por ejemplo el caso en el cual consideramos las varianzas iguales en las dos poblaciones, el error de equivocarnos al rechazar la hipótesis de igualdad de medias es de 0,47. Este error es muy alto, por lo que debemos concluir que no podemos rechazar la hipótesis nula de igualdad de medias.

- d) Comentar y contrastar las hipótesis que hemos asumido para poder realizar el experimento de comparar las dos muestras.

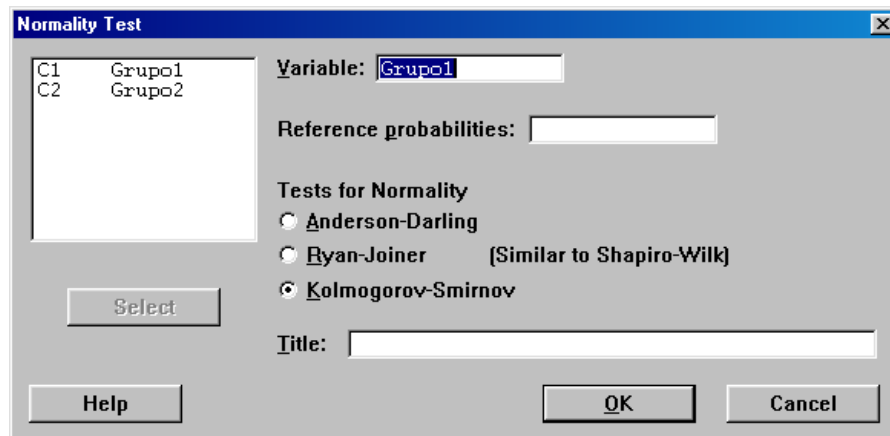
Las hipótesis que hemos utilizado para poder realizar el ejercicio son:

- Las dos muestras provienen de unas poblaciones normales.
- En el caso de suponer que las varianzas son iguales, estamos suponiendo que las dos distribuciones normales de las dos poblaciones tienen la misma varianza.

Para comprobar la primera hipótesis, la de la normalidad, podemos realizar un **test de Normalidad**, y ver si nuestros datos provienen de una distribución normal.

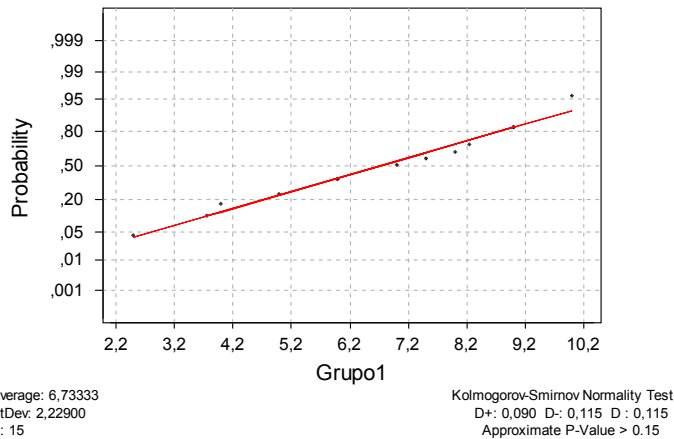
Para el caso de la primera muestra:

Seleccionar *Stat > Basic Statistic > Normality Test* :



obteniendo el siguiente contraste:

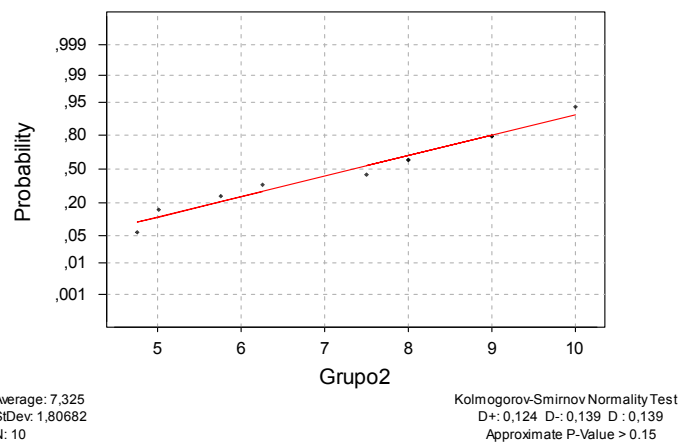
Normal Probability Plot



El p-valor del contraste es  $>0,15$ . Por lo tanto no podemos rechazar la hipótesis de que los datos provengan de una distribución normal.

Para la segunda muestra obtendríamos los siguientes resultados:

Normal Probability Plot



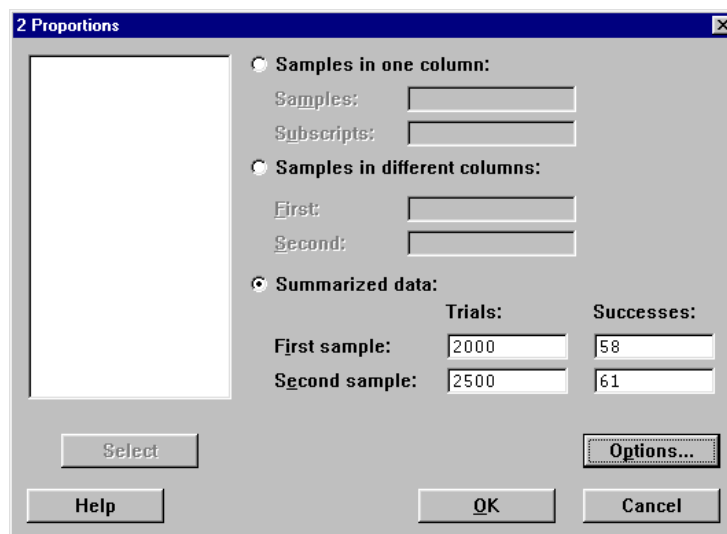
También obtenemos un valor superior a 0,15.

### 3. Contraste de diferencia de proporciones en dos muestras independientes

- De 2.000 empresas muestreadas aleatoriamente en el año 2002, 58 tenían alguna anomalía en sus cuentas auditadas en EE.UU. mientras que en 2000, de otra muestra de 2.500 empresas, 61 tenían algún error en la contabilización de sus cuentas. ¿la proporción de empresas con algún error en sus cuentas auditadas en 2002, fue significativamente distinta que la proporción de ellas en el año 2000?

Para realizar el contraste, vamos a calcular un intervalo de confianza para la diferencia de proporciones de empresas con algún error en sus cuentas de los dos años para poder comprobar si la diferencia entre los dos años es significativa o no.

Seleccionamos: *Stat > Basic Statistics > 2 Proportions* y completamos la ventana principal y la de opciones como sigue:



2 Proportions

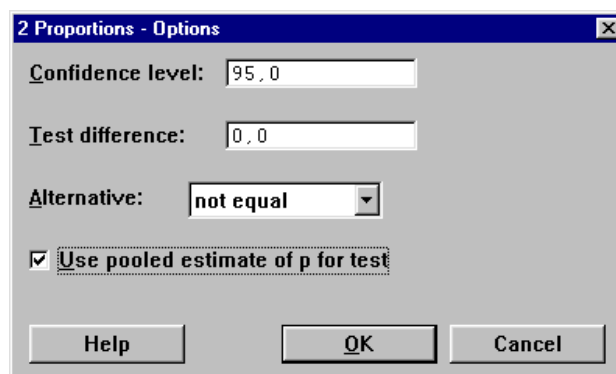
Samples in one column:  
Samples:   
Subscripts:

Samples in different columns:  
First:   
Second:

Summarized data:

	Trials:	Successes:
First sample:	<input type="text" value="2000"/>	<input type="text" value="58"/>
Second sample:	<input type="text" value="2500"/>	<input type="text" value="61"/>

Select Options... Help OK Cancel



2 Proportions - Options

Confidence level:

Test difference:

Alternative:

Use pooled estimate of p for test

Help OK Cancel

Sample	X	N	Sample p
1	58	2000	0,029000
2	61	2500	0,024400

Estimate for p(1) - p(2): 0,0046  
 95% CI for p(1) - p(2): (-0,00492175; 0,0141217)  
 Test for p(1) - p(2) = 0 (vs not = 0): Z = 0,96 P-Value = 0,339

El intervalo de confianza para la diferencia de proporciones, a nivel del 95%, está entre -0,0049 y 0,0141. Esto parece apuntar a que el porcentaje de empresas que tiene alguna anomalía en sus cuentas contables no es significativamente diferente en los dos años.

El estadístico de contraste es  $z = 0,96$  cuyo p-valor es 0,339 que al ser menor que el nivel de significación del 5%, el p-valor resulta coherente con la impresión anterior, por lo que no rechazaremos la hipótesis nula.

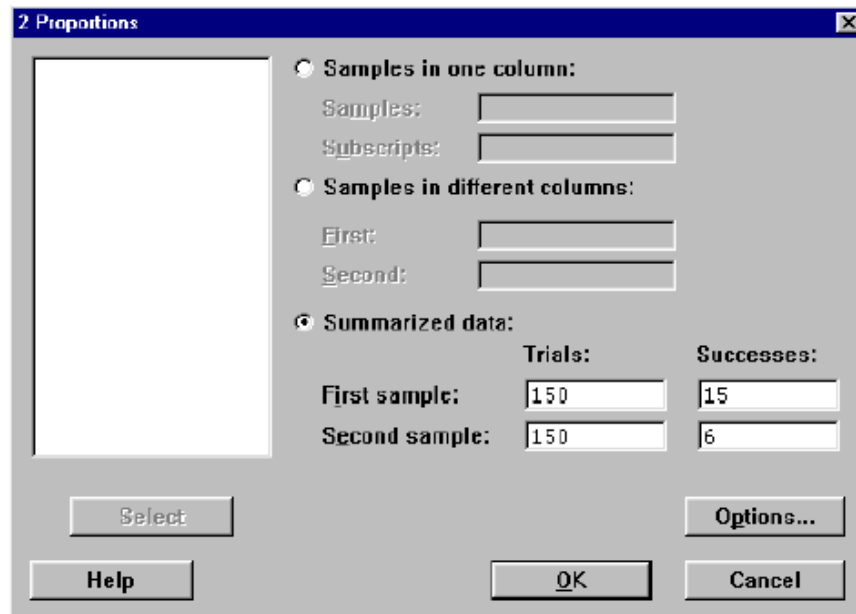
- En un anuncio publicitario de discos duros para ordenador, el fabricante asegura que sus precios son más económicos y que el porcentaje de sus discos defectuosos es igual al de la competencia. Para contrastar esta última afirmación hemos tomado dos muestras aleatorias, cada una de ellas compuesta por 150 unidades. Los resultados obtenidos se muestran en la tabla siguiente:

Fabricante	Nº de discos defectuosos	Nº de discos observados
Anunciante	15	150
Competencia	6	150

Es inmediato comprobar que se cumplen los supuestos para este caso, por lo que pasaremos a calcular un intervalo de confianza del 95% para la diferencia entre proporciones y a realizar el correspondiente test de hipótesis:

Seleccionamos: *Stat > Basic Statistics > 2 Proportions* :

Completamos la ventana principal y la de opciones como sigue:



**2 Proportions**

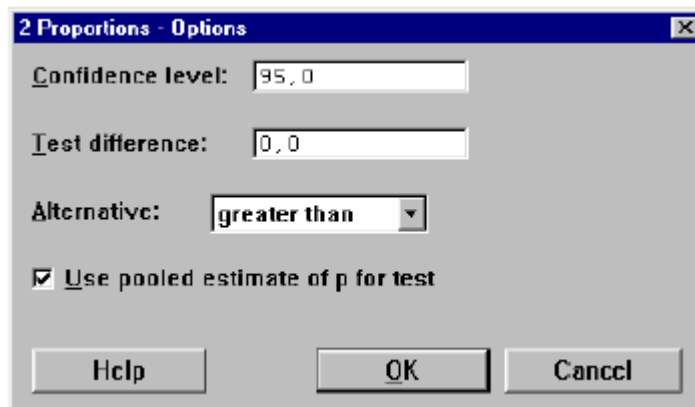
Samples in one column:  
 Samples:   
 Subscripts:

Samples in different columns:  
 First:   
 Second:

Summarized data:  

	Trials:	Successes:
First sample:	<input type="text" value="150"/>	<input type="text" value="15"/>
Second sample:	<input type="text" value="150"/>	<input type="text" value="6"/>

Select Options... Help OK Cancel



**2 Proportions - Options**

Confidence level:

Test difference:

Alternative:

Use pooled estimate of p for test

Help OK Cancel

Test and Confidence Interval for Two Proportions			
Sample	X	N	Sample p
1	15	150	0,100000
2	6	150	0,040000
Estimate for p(1) - p(2): 0,06			
95% CI for p(1) - p(2): (0,00265640; 0,117344)			
Test for p(1) - p(2) = 0 (vs > 0): Z = 2,04 P-Value = 0,021			

El intervalo de confianza para la diferencia de proporciones, a nivel del 95%, tiene por extremos los valores positivos 0,003 y 0,117 (observar que no contiene el valor 0, aunque por muy poco). Esto parece apuntar a que el porcentaje de defectos en los discos del anunciante es significativamente superior al porcentaje de la competencia. Para un nivel de significación del 0,05, el p-valor resulta coherente con la impresión anterior, por lo que resulta sensato rebatir la afirmación del anunciante (si bien las cosas cambiarían si tomásemos  $\alpha = 0,01$ ).

## **BIBLIOGRAFÍA**

---

- [1] D.A. Lind, R.D. Mason, W.G. Marchal (2001): "Estadística para Administración y Economía". Ed. Irwin McGraw-Hill.F.
- [2] Kvanli, A. "Introduction to Business Statistics" . South-Western
- [3] R. Johnson (1996): "Elementary Statistics". Ed. Duxbury
- [4] Richard I. Levin & David S. Rubin (1996): "Estadística para Administradores". Ed. Prentice Hall.
- [5] Cuadras, Carles M.: "Problemas de probabilidades y estadística Barcelona" : EUB, 1995.
- [6] Canavos, George C.: "Probabilidad y estadística : aplicaciones y métodos". Madrid: McGraw-Hill, DL 1992.

## **ENLACES**

---

- ❑ <http://www.unalmed.edu.co/~estadist/confinterval/intervalconf.htm> : Definición y applets que representan el concepto de Intervalo de confianza.
- ❑ <http://oak.cats.ohiou.edu/~wallacd1/sci.html> : Características y ejemplos de los intervalos de confianza para una única muestra
- ❑ <http://oak.cats.ohiou.edu/~wallacd1/shyp.html> : Características y ejemplos de contraste de hipótesis para una población
- ❑ [http://e-stadistica.bio.ucm.es/mod\\_intervalos/intervalos\\_applet\\_ghost.html](http://e-stadistica.bio.ucm.es/mod_intervalos/intervalos_applet_ghost.html) : Applets sobre intervalos de confianza
- ❑ [http://e-stadistica.bio.ucm.es/mod\\_contraste/contraste\\_applet.html](http://e-stadistica.bio.ucm.es/mod_contraste/contraste_applet.html) : Applet sobre contraste de hipótesis para muestras independientes.
- ❑ <http://halweb.uc3m.es/esp/Personal/personas/stefan/ESP/applet.htm> : Conjunto de applets interactivos de Estadística básica
- ❑ <http://ftbw2.rug.ac.be/iloapp/Applets/Ap6b.html> : Applet interactivo de contraste de hipótesis con dos muestras