OPEN COURSEWARE
UNIVERSIDAD DE MURCIA

Práctica 6. Contrastes paramétricos en dos poblaciones

1. Comparación de dos varianzas con muestras independientes

En el apartado siguiente vamos a estudiar el problema de la comparación de dos medias poblacionales en el caso en que observemos dos variables aleatorias Normales (una en cada población), suponiendo que se han extraído dos muestras aleatorias (una de cada población) independientes. Veremos en dicho apartado que necesitamos saber si las varianzas poblacionales (que serán desconocidas) son iguales o distintas. Por este motivo estudiamos ahora el contraste de comparación de varianzas en el caso en que desconozcamos los valores de las medias poblacionales.

Este procedimiento estadístico solamente es válido cuando las dos muestras son aleatorias y las dos poblaciones son Normales.

Para realizar este test paramétrico hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**2 Variances**.

Ejemplo 1. Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del pulso de los hombres antes de correr es igual a la varianza poblacional del pulso de las mujeres antes de correr. Lo que se quiere es comparar la varianza poblacional de la variable **Pulse1** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste que tenemos que hacer es $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$, siendo X_1=*"Pulso de los hombres antes de correr"* y X_2=*"Pulso de las mujeres antes de correr"*. Como no hay relación alguna entre el grupo de hombres y el grupo de mujeres, podemos afirmar que las muestras son independientes. Por tanto, nos encontramos ante un contraste de comparación de dos varianzas poblacionales, con muestras independientes y medias poblacionales desconocidas. Ya hemos comprobado, en la Práctica 4, que las dos variables, X_1 y X_2, son Normales.

Para hacer este contraste seleccionamos **Stat** ⇒**Basic Statistics** ⇒**2 Variances**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse1**'; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna '**Sex**'. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la diferencia de desviaciones típicas poblacionales, $\sigma_1 - \sigma_2$. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro ejemplo, podemos dejar lo que aparece por defecto, es decir, 95.

Title: Aquí se puede escribir un título para el resultado del contraste. En nuestro ejemplo, podemos dejarlo en blanco.

Como resultado de este contraste obtenemos una nueva ventana que contiene dos gráficos y los resultados de dos tests de hipótesis sobre comparación de dos varianzas (el test F de Snedecor y el test de Levene). Podemos comprobar que el p-valor para el test F de Snedecor es $0'299$; claramente mayor que el nivel de significación, $\alpha = 0'05$, por lo que podemos aceptar la hipótesis nula; es decir, podemos aceptar que la varianza poblacional del pulso de los hombres antes de correr es igual a la varianza poblacional del pulso de las mujeres antes de correr. Con el test de Levene también aceptaríamos la hipótesis nula pues el p-valor es igual a $0'148$.

Ejemplo 2. Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del pulso de los hombres después de correr es igual a la varianza poblacional del pulso de las mujeres después de correr. Lo que se quiere es comparar la varianza poblacional de la variable **Pulse2** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste que tenemos que hacer es $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$, siendo X_1=*"Pulso de los hombres después de correr"* y X_2=*"Pulso de las mujeres después de correr"*.

Para hacer este contraste seleccionamos **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**2 Variances**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse2**'; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna '**Sex**'.

Se puede comprobar que el p-valor para el test F de Snedecor es $0'003$, claramente menor que el nivel de significación, $\alpha = 0'05$, por lo que tenemos que rechazar la hipótesis nula y, por tanto, aceptar que la varianza poblacional del pulso de los hombres después de correr es distinta de la varianza poblacional del pulso de las mujeres después de correr. Con el test de Levene llegamos a la misma conclusión pues el p-valor es igual a $0'011$.

También se puede realizar este contraste de hipótesis si sabemos los dos tamaños muestrales y los resultados de las dos cuasi-varianzas muestrales. Veámoslo con un nuevo ejemplo:

Ejemplo 3. Supongamos que, de una muestra aleatoria de 21 personas que son socias de una biblioteca, la media del número de horas por semana que pasan en la biblioteca es 10, con una cuasi-varianza de 9. Y para una muestra aleatoria independiente de la primera, de 16 personas que no son socias de la biblioteca, la media es 6, con una cuasi-varianza de 4. ¿Existe diferencia significativa entre las varianzas del número de horas semanales que pasan en la biblioteca los socios y los no socios?

Como la cuasi-varianza muestral en el grupo de los socios es mayor que en el grupo de los no socios, entonces S_1^2 será la cuasi-varianza en el grupo de los socios; es decir, X_1=*"Tiempo semanal que permanecen en la biblioteca los socios"* y X_2=*"Tiempo semanal que permanecen en la biblioteca los no socios"*. Hemos de suponer que las variables aleatorias X_1 y X_2 son Normales.

Así pues, se tienen los siguientes datos:

$$n_1 = 21\,, \quad S_1^2 = 9\,,$$
$$n_2 = 16\,, \quad S_2^2 = 4\,.$$

Vamos a decidir sobre el siguiente contraste de hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2\,,$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2\,.$$

Seleccionamos la opción **Stat** ⇒**Basic Statistics** ⇒**2 Variances**. Activamos la opción **Summarized data**, con lo cual se desactivan automáticamente las opciones **Samples in one column** y **Samples in different columns**. Dentro de **First**, en **Sample size** tenemos que teclear el tamaño muestral de la primera muestra, que es **21**, y en **Variance** tenemos que teclear el resultado de la cuasi-varianza de la primera muestra, que es **9**. Dentro de **Second**, en **Sample size** tenemos que teclear el tamaño muestral de la segunda muestra, que es **16**, y en **Variance** tenemos que teclear el resultado de la cuasi-varianza de la segunda muestra, que es **4**.

Tanto en la ventana de sesión como en el gráfico generado comprobamos que el p-valor para el test F de Snedecor es $0'114$, mayor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$) y, por tanto, aceptamos la hipótesis nula. En consecuencia, aceptamos que no existe diferencia significativa entre las varianzas del número de horas semanales que pasan en la biblioteca los socios y los no socios.

2. Comparación de dos medias con muestras independientes

En general, un contraste para decidir sobre la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_1 : \mu_1 \neq \mu_2$ es bastante frecuente y constituye uno de los primeros objetivos de cualquier investigador que se inicia en estadística. Los métodos de resolución del problema varían según las muestras sean independientes o apareadas, y según las varianzas poblacionales sean conocidas o desconocidas. Dentro del caso en que las varianzas poblacionales sean desconocidas, el método depende de si son iguales o distintas. El caso de muestras independientes y varianzas poblacionales conocidas no se puede hacer con ***Minitab***. Trataremos, a continuación, el resto de los casos.

2.1. Comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas pero iguales

Este procedimiento solamente es válido cuando las dos muestras son aleatorias y las dos poblaciones son Normales o los dos tamaños muestrales son grandes (en la práctica $n_1, n_2 \geq 30$).

Para realizar este test paramétrico hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional de los hombres antes de correr es igual al pulso medio poblacional de las mujeres antes de correr. Lo que se quiere es comparar la media poblacional de la variable **Pulse1** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste que tenemos que hacer es $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$, siendo X_1=*"Pulso de los hombres antes de correr"* y X_2=*"Pulso de las mujeres antes de correr"*.

En el **Ejemplo 1** de la sección 1 hemos comprobado que se puede aceptar que la varianza poblacional del pulso de los hombres antes de correr es igual a la varianza poblacional del pulso de las mujeres antes de correr. Por tanto, nos encontramos ante un contraste de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas pero iguales. Aunque las variables aleatorias X_1 y X_2 no fuesen Normales (que sí lo son, pues lo hemos comprobado en la Práctica 4), se puede aplicar este contraste debido a que los tamaños muestrales son suficientemente grandes: $n_1 = 57$ y $n_2 = 35$.

Para hacer este contraste seleccionamos **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different**

columns y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse1**'; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna '**Sex**'; y activamos **Assume equal variances** ya que hemos comprobado que las varianzas poblacionales son desconocidas pero iguales. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro ejemplo, podemos dejar lo que aparece por defecto, es decir, 95.

Test difference: Aquí se pone el valor con el que se compara la diferencia de medias poblacionales, μ_0. La hipótesis nula $H_0 : \mu_1 = \mu_2$ es equivalente a $H_0 : \mu_1 - \mu_2 = 0$, por lo que el valor con el que se compara la diferencia de medias poblacionales, en este ejemplo, es cero; es decir, $\mu_0 = 0$. En consecuencia, nosotros dejamos lo que aparece por defecto (cero).

Alternative: Aquí se especifica cuál es la hipótesis alternativa: **less than** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 < \mu_0$, **not equal** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 \neq \mu_0$ y **greater than** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 > \mu_0$. Tengamos en cuenta que con la opción **less than** el intervalo de confianza para $\mu_1 - \mu_2$ será del tipo $(-\infty, b)$, con la opción **not equal** el intervalo de confianza será del tipo (a, b) y con la opción **greater than** el intervalo de confianza será del tipo $(a, +\infty)$. En nuestro ejemplo, tenemos que dejar lo que aparece por defecto, que es **not equal**, ya que la hipótesis alternativa es $H_1 : \mu_1 \neq \mu_2$, que es equivalente a $H_1 : \mu_1 - \mu_2 \neq 0$.

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'006$, claramente menor que el nivel de significación, $\alpha = 0'05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos que el pulso medio poblacional de los hombres antes de correr es distinto del pulso medio poblacional de las mujeres antes de correr. Como la media muestral del pulso de las mujeres antes de correr ($76'9$) es mayor que la media muestral del pulso de los hombres antes de correr ($70'42$) podríamos, incluso, aceptar que la media poblacional del pulso de las mujeres antes de correr es mayor que la media poblacional del pulso de los hombres antes de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(-10'96, -1'91)$.

También se puede realizar este contraste de hipótesis si sabemos los dos tamaños muestrales, los resultados de las dos medias muestrales y los resultados de las dos cuasi-desviaciones típicas muestrales. Veámoslo con un nuevo ejemplo:

Con los datos del **Ejemplo 3** (de la sección 1) queremos decidir si existe diferencia significativa entre el número medio de horas semanales que permanecen en la biblioteca los socios y los no socios.

Como en dicho ejemplo hemos decidido aceptar que no existe diferencia significativa entre las varianzas poblacionales, entonces nos encontramos ante un contraste de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas pero iguales. Realizaremos el siguiente contraste de hipótesis:

$$H_0 : \mu_1 = \mu_2 \,,$$
$$H_1 : \mu_1 \neq \mu_2 \,.$$

Los datos son:

$$n_1 = 21\,, \quad \overline{X}_1 = 10\,, \quad S_1 = 3\,,$$
$$n_2 = 16\,, \quad \overline{X}_2 = 6\,, \quad S_2 = 2\,.$$

Seleccionamos la opción **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**. Activamos la opción **Summarized data**, con lo cual se desactivan automáticamente las opciones **Samples in one column** y **Samples in different columns**. Dentro de **First**, en **Sample size** tenemos que teclear el tamaño muestral de la primera muestra, que es **21**, en **Mean** tenemos que teclear el resultado de la media de la primera muestra, que es **10**, y en **Standard deviation** tenemos que teclear el resultado de la cuasi-desviación típica de la primera muestra, que es **3**. Dentro de **Second**, en **Sample size** tenemos que teclear el tamaño muestral de la segunda muestra, que es **16**, en **Mean** tenemos que teclear el resultado de la media de la segunda muestra, que es **6**, y en **Standard deviation** tenemos que teclear el resultado de la cuasi-desviación típica de la segunda muestra, que es **2**. Activamos **Assume equal variances** ya que hemos comprobado (en el **Ejemplo 3**, como ya hemos dicho) que las varianzas poblacionales son desconocidas pero iguales. Pulsamos en **Options** y en el cuadro de diálogo resultante dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es 0, el mínimo posible y, por supuesto, menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$), por lo que debemos rechazar la hipótesis nula. Aceptamos, en consecuencia, que existe diferencia significativa entre el número medio de horas semanales que permanecen en la biblioteca los socios y los no socios. Como la media muestral del número de horas semanales que permanecen en la biblioteca los socios (10) es mayor que la media muestral del número de horas semanales que permanecen en la biblioteca los no socios (6) podríamos, incluso, aceptar que la media poblacional del número de horas semanales que permanecen en la biblioteca los socios es mayor que la media poblacional del número de horas semanales que permanecen en la biblioteca los no socios. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(2'326, 5'674)$.

2.2. Comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas y distintas

Igual que en el apartado anterior, este procedimiento solamente es válido cuando las dos muestras son aleatorias y las dos poblaciones son Normales o los dos tamaños muestrales son grandes (en la práctica $n_1, n_2 \geq 30$).

Para realizar este test paramétrico hay que seleccionar, igual que antes, **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**. Hay que rellenar el cuadro de diálogo de manera similar al apartado anterior, con la salvedad de que, en este caso, hay que desactivar la opción **Assume equal variances**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional de los hombres después de correr es igual al pulso medio poblacional de las mujeres después de correr. Queremos comparar la media poblacional de la variable **Pulse2** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste que tenemos que hacer es $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$, siendo X_1=*"Pulso de los hombres después de correr"* y X_2=*"Pulso de las mujeres después de correr"*.

En el **Ejemplo 2** de la sección 1 hemos comprobado que se puede aceptar que la varianza poblacional del pulso de los hombres después de correr es distinta de la varianza poblacional del pulso de las mujeres después de correr. Por tanto, nos encontramos ante un contraste de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas

y distintas. Aunque las variables aleatorias X_1 y X_2 no fuesen Normales, se puede aplicar este contraste debido a que los tamaños muestrales son suficientemente grandes: $n_1 = 57$ y $n_2 = 35$.

Para hacer el contraste seleccionamos **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna **'Pulse2'**; y en **Subscripts** seleccionamos, de la lista de la izquierda, la columna **'Sex'**. Si se pulsa el botón **Options** aparece un cuadro de diálogo similar al ejemplo anterior. En este cuadro de diálogo dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'007$, claramente menor que el nivel de significación, $\alpha = 0'05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos que el pulso medio poblacional de los hombres después de correr es distinto del pulso medio poblacional de las mujeres después de correr. Como la media muestral del pulso de las mujeres después de correr $(86'7)$ es mayor que la media muestral del pulso de los hombres después de correr $(75'9)$ podríamos, incluso, aceptar que la media poblacional del pulso de las mujeres después de correr es mayor que la media poblacional del pulso de los hombres después de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(-18'65, -3'02)$.

3. Comparación de dos medias con muestras apareadas

Este procedimiento solamente es válido cuando las dos muestras son aleatorias y la variable aleatoria diferencia, $D = X_1 - X_2$, es Normal o el tamaño muestral común, n, es grande (en la práctica, $n \geq 30$).

Para realizar este test paramétrico hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**Paired t**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional antes de correr es igual al pulso medio poblacional después de correr. Lo que se quiere es comparar la media poblacional de la variable **Pulse1** con la media poblacional de la variable **Pulse2**. El contraste que tenemos que hacer es $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$, siendo X_1=*"Pulso antes de correr"* y X_2=*"Pulso después de correr"*. Como las dos variables están observadas en los mismos individuos, podemos afirmar que las muestras están relacionadas; es decir, son apareadas o asociadas. Por tanto, nos encontramos ante un contraste de comparación de dos medias poblacionales con muestras apareadas. Aunque la variable aleatoria diferencia, $D = X_1 - X_2$, no fuese Normal, se puede aplicar este contraste debido a que los tamaños muestrales son suficientemente grandes: $n_1 = n_2 = n = 92$.

Para hacer este contraste seleccionamos **Stat** ⇒**Basic Statistics** ⇒**Paired t**. Activamos la opción **Samples in columns**; en **First sample** seleccionamos, de la lista de variables de la izquierda, la columna **'Pulse1'**; en **Second sample** seleccionamos, de la lista de variables de la izquierda, la columna **'Pulse2'**. Si pulsamos el botón **Options** nos aparece un cuadro de diálogo similar al de la opción anterior (**2-Sample t**⇒**Options**). En este cuadro de diálogo dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es igual a 0, el mínimo posible y, por supuesto, menor que el nivel de significación, $\alpha = 0'05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos, por tanto, que el pulso medio poblacional antes de correr es distinto del pulso medio poblacional después de correr. Como la

media muestral del pulso después de correr ($80'00$) es mayor que la media muestral del pulso antes de correr ($72'87$) podríamos, incluso, aceptar que la media poblacional del pulso después de correr es mayor que la media poblacional del pulso antes de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, en este caso, es $(-9'92, -4'34)$.

4. Comparación de dos proporciones

Consideramos una variable aleatoria dicotómica o dicotomizada (con resultados denominados *éxito* y *fracaso*) evaluada en dos poblaciones distintas. Extraemos sendas muestras aleatorias independientes de tamaños n_1 y n_2. Queremos realizar contraste $H_0 : p_1 = p_2$ frente a $H_1 : p_1 \neq p_2$, donde p_i es la proporción de *éxitos* en la población i, para $i = 1, 2$.

Si los resultados de la variable aleatoria dicotómica o dicotomizada son numéricos, ***Minitab*** toma como suceso *éxito* al número más alto; y si los resultados son de tipo texto, ***Minitab*** toma como suceso *éxito* a la cadena de texto que esté más cerca del final del alfabeto. Por ejemplo, si los resultados son *SI* y *NO*, entonces el resultado *SI* sería el suceso *éxito*. Si los resultados son *1* y *2*, entonces el resultado *2* sería el suceso *éxito*.

Para realizar la comparación de dos proporciones poblacionales hay que seleccionar **Stat ⇒Basic Statistics ⇒2 Proportions**.

Recordemos que en la hoja de datos **Pulse.mtw** la variable **Smokes** tiene solamente dos resultados: 1=Fumador, 2=No Fumador. Por otra parte, la variable **Sex** también tiene solamente dos resultados: 1=Hombre, 2=Mujer. Comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la proporción poblacional de hombres no fumadores es igual a la proporción poblacional de mujeres no fumadoras. ***Minitab*** toma como suceso *éxito* de la variable **Smokes** el resultado **2** (es decir, No Fumador) pues es el resultado más alto de los dos. Lo que se quiere es comparar la proporción poblacional de éxitos de la variable **Smokes** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). El contraste que tenemos que hacer es $H_0 : p_1 = p_2$ frente a $H_1 : p_1 \neq p_2$.

Para hacer este contraste seleccionamos **Stat ⇒Basic Statistics ⇒2 Proportions**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna **'Smokes'**; y en **Subscripts** seleccionamos, de la lista de la izquierda, la columna **'Sex'**. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la diferencia de proporciones poblacionales, $p_1 - p_2$. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro ejemplo, podemos dejar lo que aparece por defecto, es decir, 95.

Test difference: Aquí se pone el valor con el que se compara la diferencia de proporciones poblacionales, p_0. La hipótesis nula $H_0 : p_1 = p_2$ es equivalente a $H_0 : p_1 - p_2 = 0$, por lo que el valor con el que se compara la diferencia de proporciones poblacionales, en este ejemplo, es cero; es decir, $p_0 = 0$. En consecuencia, nosotros dejamos lo que aparece por defecto (cero).

Alternative: Aquí se especifica cuál es la hipótesis alternativa: **less than** significa que la hipótesis alternativa es $H_1 : p_1 - p_2 < p_0$, **not equal** significa que la hipótesis alternativa es $H_1 : p_1 - p_2 \neq p_0$ y **greater than** significa que la hipótesis alternativa es $H_1 : p_1 - p_2 > p_0$. Tengamos en cuenta que con la opción **less than** el intervalo de confianza para $p_1 - p_2$ será del tipo $(-\infty, b)$, con

la opción **not equal** el intervalo de confianza será del tipo (a, b) y con la opción **greater than** el intervalo de confianza será del tipo $(a, +\infty)$. En nuestro ejemplo, tenemos que dejar lo que aparece por defecto, que es **not equal**, ya que la hipótesis alternativa es $H_1 : p_1 \neq p_2$, que es equivalente a $H_1 : p_1 - p_2 \neq 0$.

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'198$, mayor que el nivel de significación, $\alpha = 0'05$, por lo que debemos aceptar la hipótesis nula. Aceptamos, en consecuencia, que la proporción poblacional de hombres no fumadores es igual a la proporción poblacional de mujeres no fumadoras. El intervalo de confianza al 95 % para la diferencia de proporciones poblacionales, $p_1 - p_2$, es $(-0'308592, 0'0639809)$.

También se puede realizar este contraste de hipótesis si sabemos los dos tamaños muestrales y el número de éxitos en cada una de las dos muestras. Veámoslo con un ejemplo:

Con objeto de comparar dos pequeñas empresas A y B de encuadernación de libros, se extrajo una muestra aleatoria de 250 libros encuadernados en A y otra muestra aleatoria de 200 libros encuadernados en B, y se encontró que 50 de los libros encuadernados en A, y 32 de los encuadernados en B tenían algún defecto en su encuadernación. ¿Son igualmente buenas las dos empresas de encuadernación?

Lo que queremos comprobar es si la proporción poblacional de libros defectuosos encuadernados en la empresa A es igual a la proporción poblacional de libros defectuosos encuadernados en la empresa B.

Para hacer este contraste seleccionamos **Stat** ⇒**Basic Statistics** ⇒**2 Proportions**. Activamos la opción **Summarized data**. Dentro de **First**, en **Events** tenemos que teclear el número de éxitos en la primera muestra, que es **50**, y en **Trials** tenemos que teclear el tamaño de la primera muestra, que es **250**. Dentro de **Second**, en **Events** tenemos que teclear el número de éxitos en la segunda muestra, que es **32**, y en **Trials** tenemos que teclear el tamaño de la segunda muestra, que es **200**. En el cuadro de diálogo de **Options** dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'269$, mayor que el nivel de significación usual ($\alpha = 0'05$) por lo que debemos aceptar la hipótesis nula. Por tanto, aceptamos que la proporción poblacional de libros defectuosos encuadernados en la empresa A es igual a la proporción poblacional de libros defectuosos encuadernados en la empresa B; es decir, las dos empresas de encuadernación son igualmente buenas. El intervalo de confianza al 95 % para la diferencia de proporciones poblacionales, $p_1 - p_2$, es $(-0'0309929, 0'110993)$.

5. Ejercicios propuestos

6.1.

a) Crea un nuevo proyecto de ***Minitab***.

b) Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2).

c) Utilizando el test de Levene, ¿se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional del número anual de transacciones de referencia de las bibliotecas públicas es igual a la varianza poblacional del número anual de transacciones de referencia de las bibliotecas universitarias? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del número anual de transacciones de referencia de las bibliotecas públicas es igual a la media poblacional del número anual de transacciones de referencia de las bibliotecas universitarias? ¿Por qué?

e) Utilizando el test F de Snedecor, ¿se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas públicas es igual a la varianza poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas universitarias? ¿Por qué?

f) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas públicas es igual a la media poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas universitarias? ¿Por qué?

g) Graba el proyecto con el siguiente nombre: **Ejercicio6-1.mpj**

6.2. En la tabla siguiente aparece el precio, en euros, de una muestra aleatoria de 15 libros que se prestan pocas veces (X_1) y el precio, en euros, de una muestra aleatoria de 15 libros que se prestan muchas veces (X_2).

x_{1i}	x_{2i}
75	110
32	30
30	45
34	69
42	46
57	53
51	97
36	43
82	42
45	37
58	48
66	45
40	105
35	61
51	57

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **PrecioLibros.mtw**

c) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional del precio de los libros que se prestan poco es igual a la varianza poblacional del precio de los libros que se prestan mucho? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del precio de los libros que se prestan poco es igual a la media poblacional del precio de los libros que se prestan mucho? ¿Por qué?

e) Graba el proyecto con el siguiente nombre: **Ejercicio6-2.mpj**

6.3. En la siguiente tabla aparece el número de palabras por resumen de una muestra aleatoria de 30 artículos científicos escritos en francés (X_1) y el número de palabras por resumen de una muestra aleatoria de 30 artículos científicos escritos en inglés (X_2).

x_{1i}	70	65	68	74	79	67	75	80	62	69
	61	57	71	74	82	91	70	64	72	67
	74	70	81	85	70	74	75	71	69	54
x_{2i}	80	47	59	67	89	57	72	78	74	72
	104	118	89	87	79	78	101	120	107	95
	85	87	90	98	89	75	90	101	85	94

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **LongitudResumenes.mtw**

c) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional de la longitud de los resúmenes de artículos escritos en francés es igual a la varianza poblacional de la longitud de los resúmenes de artículos escritos en inglés? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional de la longitud de los resúmenes de artículos escritos en francés es igual a la media poblacional de la longitud de los resúmenes de artículos escritos en inglés? ¿Por qué?

e) Graba el proyecto con el siguiente nombre: **Ejercicio6-3.mpj**

6.4. Dos expertos califican una muestra aleatoria de 30 libros según su calidad (1=muy mala, 2=mala, 3=regular, 4=buena, 5=muy buena). En la tabla siguiente aparece la opinión del primer experto (X_1) y la opinión del segundo experto (X_2).

x_{1i}	x_{2i}	x_{1i}	x_{2i}
2	1	4	4
5	4	4	3
4	5	5	4
2	3	5	3
3	3	1	2
1	5	2	5
3	3	2	3
1	3	3	2
4	2	4	1
2	5	4	2
3	2	1	3
4	3	2	4
3	3	1	2
1	3	5	5
2	5	5	2

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **Opinion.mtw**

c) Calcula, en una nueva columna, los resultados de la variable diferencia $D = X_1 - X_2$.

d) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la muestra de las diferencias, $d_i = x_{1i} - x_{2i}$, es aleatoria? ¿Por qué?

e) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la variable diferencia, $D = X_1 - X_2$, es Normal? ¿Por qué?

f) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional de los resultados de la opinión del primer experto es igual a la media poblacional de los resultados de la opinión del segundo experto? ¿Por qué?

g) Graba el proyecto con el siguiente nombre: **Ejercicio6-4.mpj**

6.5. Elegimos al azar 30 matrimonios y observamos el número de veces que los hombres han visitado alguna biblioteca en los tres últimos meses (X_1) y el número de veces que las mujeres han visitado alguna biblioteca en los tres últimos meses (X_2). Los resultados se muestran en la siguiente tabla.

x_{1i}	x_{2i}	x_{1i}	x_{2i}	x_{1i}	x_{2i}
12	8	8	10	25	14
30	11	14	15	12	16
10	12	20	12	8	10
20	16	13	19	23	20
15	10	11	6	14	17
14	9	7	7	8	10
11	12	6	7	12	23
9	10	8	6	27	10
7	7	15	20	32	27
5	4	42	35	14	18

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **VisitasBiblioteca.mtw**

c) Calcula, en una nueva columna, los resultados de la variable diferencia $D = X_1 - X_2$.

d) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la muestra de las diferencias, $d_i = x_{1i} - x_{2i}$, es aleatoria? ¿Por qué?

e) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la variable diferencia, $D = X_1 - X_2$, es Normal? ¿Por qué?

f) ¿Podemos afirmar que hay diferencia significativa entre los hombres y las mujeres de los matrimonios en cuanto al número de veces que van a la biblioteca? ¿Por qué?

g) Graba el proyecto con el siguiente nombre: **Ejercicio6-5.mpj**

6.6. En la siguiente tabla aparece el número de usuarios diarios de la biblioteca A (variable X_1) y el número de usuarios diarios de la biblioteca B (variable X_2) en 10 días elegidos al azar.

x_{1i}	x_{2i}
51	45
72	58
35	32
70	56
75	68
98	76
100	88
80	69
72	57
90	75

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **UsuariosDiarios.mtw**

c) Calcula, en una nueva columna, los resultados de la variable diferencia $D = X_1 - X_2$.

d) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la muestra de las diferencias, $d_i = x_{1i} - x_{2i}$, es aleatoria? ¿Por qué?

e) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la variable diferencia, $D = X_1 - X_2$, es Normal? ¿Por qué?

f) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del número de usuarios diarios de la biblioteca A es igual a la media poblacional del número de usuarios diarios de la biblioteca B? ¿Por qué?

g) Graba el proyecto con el siguiente nombre: **Ejercicio6-6.mpj**

6.7. Se quiere saber si la proporción de libros escritos en español es la misma en dos bibliotecas universitarias (la de la facultad de matemáticas y la de la facultad de filosofía). Se toma una muestra aleatoria simple de 100 libros de la biblioteca de la facultad de matemáticas y se encuentra que 35 de ellos están escritos en español y el resto en otros idiomas. Se extrae otra muestra aleatoria simple de 150 libros de la biblioteca de la facultad de filosofía y se observa que 60 están escritos en español. ¿Qué conclusión se puede extraer?