Práctica 6

Intervalos de confianza y contrastes de hipótesis en poblaciones normales

6.1. Introducción

El cálculo de intervalos de confianza y la realización de los contrastes de hipótesis que veremos en esta práctica están diseñados, y por lo tanto son válidos, cuando los datos sobre los que se realiza el estudio proceden de una distribución normal.

Por este motivo, antes de utilizar las técnicas correspondientes para el cálculo de intervalos y la realización de contrastes, es necesario comprobar la normalidad de los datos. Por este motivo, en la siguiente sección, veremos cómo realizar con R un gráfico y un test de hipótesis adecuados para realizar dicha comprobación.

6.2. Estudio de la normalidad de los datos muestrales

6.2.1. Gráfica de comparación de cuantiles

La gráfica de comparación de cuantiles compara los cuantiles de la muestra con los cuantiles correspondientes a la distribución poblacional teórica. La distribución de los datos muestrales se asemejará más a las distribución de contraste cuanto más parecida sea la recta que muestra el gráfico a la nube de puntos representada en el mismo.

En cualquier caso, las conclusiones sobre la normalidad de los datos no pueden obtenerse únicamente a partir de la representación gráfica, sino que es necesario realizar un contraste de hipótesis que las confirme. No obstante, la representación puede servir como apoyo visual a las conclusiones que se obtengan mediante los contrastes.

Además presenta una utilidad adicional y es que, en caso de que se rechace la normalidad de los datos, permite conocer qué datos se "apartan más de la normalidad" y son los que provocan el rechazo. De este modo podemos estudiar qué ocurre con ellos (si se han cometido errores al realizar las mediciones, si los casos que corresponden a esos

datos presentan un comportamiento distinto al resto, etc.) y actuar en consecuencia, por ejemplo, eliminándolos de la muestra si fuera necesario.

6.2.2. Test de Shapiro-Wilk

Para el estudio de la normalidad de los datos muestrales R-Commander dispone de distintos contrastes de hipótesis, entre los que se encuentra el test de Shapiro-Wilk. Este test es considerado uno de los más potentes (con menor probabilidad de cometer un error de tipo II) y puede aplicarse cuando el tamaño muestral es inferior a 50. Si el tamaño muestral es mayor o gual que 50, disponemos de otros contrastes aunque no los estudiaremos en esta práctica. De este modo, si tenemos una muestra aleatoria simple $X_1, \ldots, X_n$ procedente de una variable aleatoria X, el test de Shapiro-Wilk permite resolver el contraste de hipótesis

$$\begin{cases} H_0: & X \sim N(\mu, \sigma^2) \\ H_1: & X \nsim N(\mu, \sigma^2) \end{cases}$$

Esto es, el test contrasta la hipótesis nula de que los datos muestrales proceden de una población normal frente a la hipótesis alternativa de que los datos no proceden de una población normal.

El test de Shapiro-Wilk podemos encontrarlo en el menú de R-Commander: `Estadísticos`→`Resúmenes`→`Test de Normalidad`. En la ventana emergente elegimos la variable a contrastar y el test a utilizar que, como ya hemos dicho, será el test de Shapiro-Wilk. La salida del programa mostrará entonces el valor del estadístico W del test, así como el correspondiente *p-valor*. Como norma general, tanto en este contraste como en los que veremos a continuación, tendremos evidencia significativa para afirmar que la hipótesis alternativa es cierta, para un nivel de significación α, cuando el *p-valor* obtenido sea menor o igual que α.

¿Qué es realmente el *p-valor* y por qué debemos rechazar la hipótesis nula cuando el p-valor es inferior a la significación del contraste? Vamos a ver un ejemplo para entenderlo mejor.

Supongamos que un amigo nos dice que tiene poderes de adivinación y es capaz de acertar, en el 90 % de los casos, el resultado que se va a obtener al lanzar una moneda. Como es amigo nuestro, le concedemos el beneficio de la duda pero decidimos ponerle a prueba: vamos a lanzar una moneda 50 veces y le pedimos que, previamente a cada lanzamiento, haga una predicción sobre el resultado que se va a obtener. Tras hacer los 50 lanzamientos, encontramos que su predicción ha sido acertada únicamente en 30 de los 50 lanzamientos.

Si realmente tuviera fuera capaz de acertar el resultado del lanzamiento en el 90 % de las ocasiones, el número de aciertos en los 50 lanzamientos sería una variable aleatoria

$X \sim B(50, 0.9)$. ¿Cuál sería entonces la probabilidad de que, al hacer los 50 lanzamientos, acertara como mucho el resultado de 30 lanzamientos? Para responder a la pregunta tenemos que calcular $P(X \leq 30)$ que resulta ser igual a 0.00000002. Así pues, si realmente es cierto que puede adivinar el resultado en el 90 % de las ocasiones, la probabilidad de que al hacer nuestro experimento acierte como mucho en 30 acasiones sería extremadamente pequeña, tanto que debemos concluir que su afirmación no es cierta. Lo que hemos hecho, en realidad, es plantear el contraste de hipótesis:

$$\begin{cases} H_0: & p = 0.9 \\ H_1: & p < 0.9 \end{cases}$$

El p-valor correspondiente a este contraste para la muestra recogida, en la que se han producido 30 aciertos de 50 lanzamientos, es la probabilidad de obtener un resultado tan desfavorable a H_0, al menos, como el observado en la muestra. En nuestro caso, 0.00000002. Si la significación del contraste es $\alpha = 0.05$, al ser el p-valor obtenido inferior a la significación podemos decir que tenemos evidencia para rechazar la hipótesis nula, esto es, podemos afirmar que no es cierto que pueda predecir el resultado del lanzamiento con una eficacia del 90 %.

Ahora bien, ¿y si hubiera acertado el resultado de 44 de los 50 lanzamientos? En ese caso, si estuviera diciendo la verdad, que acierte a lo más 44 de los lanzamientos es algo bastante probable (nótese que $P(X \leq 44) = 0.383877$) por lo que no tendríamos evidencia para negar su afirmación.

Por tanto, como ya se ha dicho antes, de manera general al resolver un contraste de hipótesis tendremos evidencia para afirmar que la hipótesis alternativa es cierta, para un nivel de significación α, cuando el *p-valor* obtenido sea menor o igual que la significación del contraste. En caso contrario, no podremos afirmar que la hipótesis alternativa es correcta.

Ejemplo 1: Estudiar, para los datos del archivo *Mundo95*, la normalidad de los datos correspondiente a la tasa de nacimientos/defunciones en la región OCDE. Plantear y resolver un contraste de hipótesis adecuado con nivel de significación ($\alpha = 0.05$).

Si denotamos por X a la variable aleatoria que modela la tasa de nacimientos/defunciones en un país de la OCDE, el contraste a resolver es

$$\begin{cases} H_0: & X \sim N(\mu, \sigma^2) \\ H_1: & X \nsim N(\mu, \sigma^2) \end{cases}$$

Lo primero que debemos hacer es un filtrado del conjunto de datos para seleccionar los datos necesarios. Para ello seleccionamos el menú `Datos`→`Conjunto de datos activo`→`Filtrar el conjunto de datos activo`. En la ventana `Variables` seleccionamos la variable *nac_def* y en la expresión de selección escribimos `región=='OCDE'`. Como nombre del nuevo conjunto de datos ponemos, por ejemplo, *nac_def_OCDE*. Obtenemos entonces un conjunto de datos con los 21 casos que corresponden a la OCDE.

Aplicamos ahora el test de Shapiro-Wilk: para ello seleccionamos el menú `Estadísticos→Resúmenes→Test de normalidad`; en la ventana emergente, seleccionamos la variable *nac_def* y en el menú `Test de Normalidad` elegimos la opción `Shapiro-Wilk`.

Observamos en la salida que el p-valor obtenido es igual a 0.0248. Al ser inferior a la significación del contraste ($\alpha = 0.05$), tenemos evidencia para rechazar la hipótesis nula y afirmar que los datos no corresponden a una distribución normal.

No obstante, cabe la posibilidad de que el rechazo se deba a la existencia de algún valor anómalo en la muestra. Para comprobarlo vamos a construir un gráfico de comparación de cuantiles: seleccionamos el menú `Gráficas→Gráfica de comparación de cuantiles`. En la pestaña `Datos` seleccionamos la variable *nac_def* y en la pestaña `Opciones` seleccionamos la distribución `Normal` y la opción `Identificar Observaciones Automáticamente`, con `Número de puntos a identificar` igual a 1.

En la gráfica observamos que el programa identifica un punto correspondiente al caso 59, que es el que más *se aleja de la normalidad*. Vamos a eliminar ese caso del conjunto de datos. Para ello abrimos el editor de datos y ponemos el cursor en la fila 14 que es la que corresponde al caso 59. A continuación elegimos el menú `Editar→Borrar la fila actual` y pulsamos el botón `Aceptar`.

Volvemos a realizar el test de Shapiro-Wilk y obtenemos ahora un p-valor igual a 0.06335 por lo que no se rechaza, para un nivel de significación $\alpha = 0.05$, la normalidad de los datos una vez eliminado el caso 59.

6.3. Intervalos de confianza y contrastes de hipótesis sobre la media de una población normal

Ejemplo 2: : Un proceso químico debe producir cada día una cantidad media de 800 toneladas de un producto. El encargado de vigilar el buen funcionamiento del proceso sospecha que se está produciendo menos de esa cantidad y por ello ha observado las producciones diarias durante los últimos 5 días, que son: 802, 795, 752, 810, 783. ¿Hay evidencia para afirmar, al 5 % de significación, que la sospecha es correcta?

Consideramos la variable aleatoria X: producción diaria. Los datos corresponden a una única población por lo que, a la hora de introducirlos en el editor de dator de R-Commander, los pondremos en la primera columna del editor. A la columna le daremos el nombre *Producción*. La figura 6.1 muestra la forma en que deben quedar los datos una vez introducidos en el editor.

Aplicamos el test de Shapiro-Wilk a los datos de la variable *Producción* y obtenemos un p-valor igual a 0.4555, que es mayor que la significación del contraste $\alpha = 0.05$, por lo que no tenemos evidencia significativa para rechazar la hipótesis de que los datos siguen una distribución normal. Por tanto, asumimos que $X \sim N(\mu, \sigma^2)$ y pasamos a

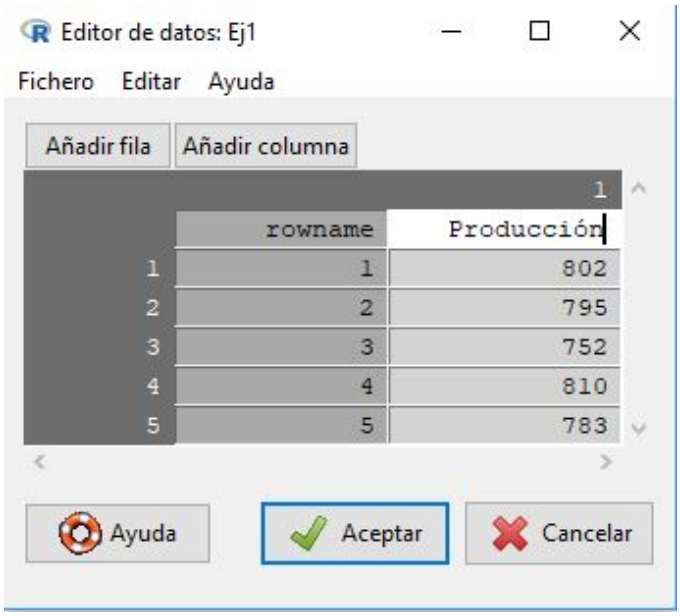


Figura 6.1: Introducción de los datos del Ejemplo 2.

resolver el contraste

$$\begin{cases} H_0: & \mu = 800 \\ H_1: & \mu < 800 \end{cases}$$

Seleccionamos el menú `Estadísticos`→`Medias`→`Test t para una muestra`. En la ventana emergente seleccionamos la variable que contiene los datos a los que aplicaremos el contraste (en este caso, la variable *Producción*); como `hipótesis alternativa` elegimos la opción `Media poblacional<mu0` y, finalmente, en la ventana `Hipótesis nula:mu=` indicamos el valor de prueba: 800.

Obtenemos entonces un p-valor igual a 0.1578. Puesto que $\alpha = 0.05$, el p-valor es mayor que la significación del contraste por lo que no tenemos evidencia significativa para rechazar la hipótesis nula y afirmar que la hipótesis alternativa es correcta. Esto es, no tenemos evidencia para afirmar que la producción media diaria es inferior a 800 toneladas.

Obsérvese que la salida nos proporciona también un intervalo de confianza al 95 % para la producción diaria media μ. Al haber realizado un contraste unilateral del tipo 'igual' frente a 'menor', la salida nos proporciona un intervalo de confianza para la producción media que está acotado únicamente en su extremo superior y que resulta ser $(-\infty, 809.9791)$.

Ejemplo 3: Los ingresos mensuales de un grupo de 9 personas seleccionadas al azar entre un gran número de individuos, han resultado ser: 1570, 1550, 1530, 1520, 1560, 1500, 1510, 1540, 1580. ¿Debe rechazarse la hipótesis de que la muestra procede de una población normal cuyos ingresos mensuales medios son de 1557 euros, para un

nivel de significación del 5 %? Calcular un intervalo de confianza al 95 % para la media poblacional.

En este caso la magnitud bajo estudio la modelamos mediante la variable aleatoria X : *ingresos mensuales de una persona elegida al azar*. El primer paso será crear un conjunto de datos con los datos del enunciado, de manera similar a cómo se hizo en el Ejemplo 2. Una vez hecho esto, aplicamos el test de Shapiro Wilk obteniendo un p-valor igual a 0.9136 por lo que no hay evidencia significativa para rechazar que los datos correspondan a una distribución normal. Por tanto, asumimos que $X \sim N(\mu, \sigma^2)$ y procedemos a realizar el contraste

$$\begin{cases} H_0: & \mu = 1557 \\ H_1: & \mu \neq 1557 \end{cases}$$

Para ello, seleccionamos nuevamente el menú `Estadísticos`→`Medias`→`Test T para una muestra`. En la ventana emergente seleccionamos la variable que contiene los datos. Como `hipótesis alternativa` elegimos la opción `Media poblacional!=mu0` y, en la ventana `Hipótesis nula:mu=`, indicamos el valor de prueba: 1557. Finalmente, en el cuadro `Nivel de confianza` indicamos el nivel de confianza del intervalo que deseamos calcular, dividido por 100, esto es, .95.

La salida del programa nos proporciona un p-valor igual a 0.09958 por lo que, al ser la significación del contraste $\alpha = 0.05$, deducimos que no hay evidencia para rechazar la hipótesis nula y afirmar que los ingresos mensuales medios son distintos de 1557 euros. Además, obtenemos que un intervalo de confianza, al 95 %, para la media poblacional viene dado por (1518.949, 1561.051).

Obsérvese que dicho intervalo contiene al valor 1557, cosa que era previsible pues no habíamos rechazado la hipótesis nula al hacer el contraste.

6.4. Intervalos de confianza y contrastes de hipótesis para muestras relacionadas o pareadas

Ejemplo 4: Se llevó a cabo un estudio para determinar el grado en el cual el alcohol entorpece la habilidad de pensamiento para llevar a cabo determinada tarea. Se seleccionaron al azar diez personas de distintas características y se les pidió que participaran en el experimento. Inicialmente, cada persona llevó a cabo la tarea sin nada de alcohol en su organismo y se midió el tiempo (en minutos) en realizarla. Posteriormente, la tarea volvió a llevarse a cabo después de que cada persona había consumido una cantidad suficiente de alcohol para tener un contenido en su organismo del 0.1 %. Los datos recogidos fueron los siguientes:

Sujeto	Tiempo antes	Tiempo después
1	28	29
2	22	35
3	55	57
4	45	51
5	32	36
6	35	58
7	40	51
8	25	34
9	37	48
10	20	30

¿Puede concluirse, al 5 %, que el consumo de alcohol incrementa el tiempo necesario para realizar la tarea?

En este caso no tenemos dos poblaciones independientes, sino que los datos del tiempo que se tarda en realizar la tarea, antes y después de la ingesta de alcohol, se miden sobre los mismos individuos y por lo tanto están relacionados. La variable bajo estudio es D = *tiempo necesario sin ingerir alcohol – tiempo necesario tras ingerir alcohol*. Supondremos que dicha variable sigue una distribución normal[1], esto es, que $D \sim N(\mu_D, \sigma_D^2)$. El contraste que debemos resolver es

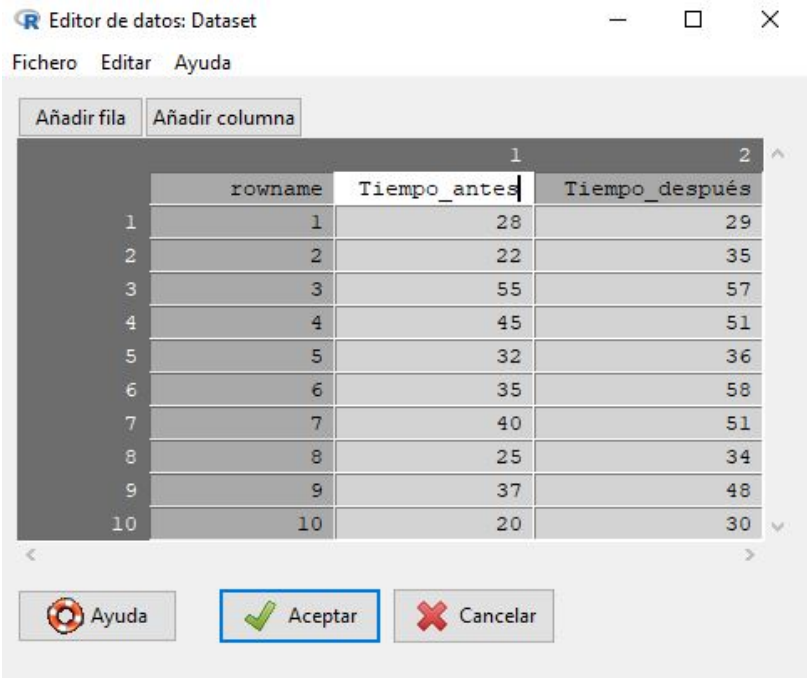
$$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D < 0 \end{cases}$$

En este caso debemos introducir los datos en dos columnas,a les que les daremos como nombre *Tiempo_antes* y *Tiempo_después*. Los datos correspondientes a cada sujeto deben introducirse en la misma fila puesto que corresponden a un mismo caso. La figura 6.2 muestra cómo deben introducirse los datos en el editor.

Una vez introducidos los datos, seleccionamos el menú `Estadísticos`→`Medias`→`Test t para datos relacionados`. En la ventana emergente seleccionaremos como primera variable *Tiempo_antes*, como segunda variable *Tiempo_después* y, en la pestaña `Opciones` indicamos como hipótesis alternativa `Diferencia <0`.

Obtenemos un p-valor igual a 0.0007992, que es menor que la significación del contraste ($\alpha = 0.05$) por lo que tenemos evidencia para rechazar la hipótesis nula y afirmar que la media de la diferencia de tiempos es negativa, esto es, que el tiempo medio que se tarda en realizar la tarea tras ingerir alcohol es superior al tiempo medio que se tarda en realizar la tarea antes de ingerir el alcohol.

[1]En este ejemplo y en los que siguen ignoraremos el paso previo consistente en contrastar la normalidad de los datos.



	rowname	Tiempo_antes	Tiempo_después
	1	1	2
1	1	28	29
2	2	22	35
3	3	55	57
4	4	45	51
5	5	32	36
6	6	35	58
7	7	40	51
8	8	25	34
9	9	37	48
10	10	20	30

Figura 6.2: Introducción de los datos del Ejemplo 4.

6.5. Intervalos de confianza y contraste de hipótesis para la diferencia de medias de dos poblaciones normales independientes

Ejemplo 5: Dividimos aleatoriamente un conjunto de pacientes afectados por cierto virus en dos grupos a los que se administra respectivamente un placebo y un tratamiento en estudio para combatir dicha afección vírica. Después de tratar a los pacientes durante 2 meses se mide la concentración de virus de cada uno de ellos. Los resultados se muestran en la siguiente tabla:

Sujeto	Administración	Nivel	Sujeto	Administración	Nivel
1	Placebo	23	2	Placebo	25
3	Placebo	26	4	Placebo	24
5	Placebo	26	6	Placebo	24
7	Placebo	22	8	Placebo	25
9	Placebo	27	10	Placebo	25
11	Tratamiento	22	12	Tratamiento	23
13	Tratamiento	22	14	Tratamiento	24
15	Tratamiento	19	16	Tratamiento	20
17	Tratamiento	21	28	Tratamiento	23
19	Tratamiento	22	20	Tratamiento	23

¿Puede decirse, al 5 % de significación, que el tratamiento hace el efecto deseado en

la infección, es decir, que disminuye el nivel de virus?

Consideramos ahora las variables X :*nivel de virus de los pacientes a los que se administró el placebo*, e Y : *nivel de virus de los pacientes a los que se administró el tratamiento*. Estas variables son independientes al haber sido distribuidos los pacientes aleatoriamente entre los dos grupos. Supondremos, además, que ambas variables se distribuyen normalmente, de tal modo que $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$. Para responder a la pregunta, tenemos que que contrastar la igualdad de medias de ambas poblaciones, esto es, tenemos que resolver, con $\alpha = 0.05$, el contraste

$$\begin{cases} H_0: & \mu_X = \mu_Y \\ H_1: & \mu_X > \mu_Y \end{cases}$$

o lo que es igual, el contraste:

$$\begin{cases} H_0: & \mu_X - \mu_Y = 0 \\ H_1: & \mu_X - \mu_Y > 0 \end{cases}$$

En este caso trabajamos con dos poblaciones independientes; para introducir los datos en el editor debemos poner los datos numéricos correspondientes a cada caso en una columna, a la que le daremos de nombre *Nivel*. En otra columna, a la que le pondremos como nombre *Admin*, escribiremos los datos correspondientes a la variable administración. Estos datos le servirán a `R` para saber a qué población corresponde cada caso. Para simplificar, en lugar de escribir *Placebo* y *Tratamiento*, escribiremos p y t. La figura 6.3 muestra cómo se deben escribir los datos en el editor.

Para resolver el contraste sobre la diferencia de medias debemos utilizar el menú `Estadísticos`→`Medias`→`Test t para muestras independientes`. Procedemos entonces a elegir, en el campo `Grupos` la variable de agrupación, que debe ser de tipo carácter y sirve para separar los datos en las dos poblaciones independientes. En nuestro caso es la variable *Admin*.

En el campo `variable explicada` elegiremos la variable *Nivel* que contiene los datos bajo estudio. Como hipótesis alternativa elegiremos `Diferencia >0` y, finalmente, tendríamos que indicar si las varianzas son iguales o no. Para determinar qué opción debemos elegir debemos resolver previamente el contraste

$$\begin{cases} H_0: & \sigma_X^2/\sigma_Y^2 = 1 \\ H_1: & \sigma_X^2/\sigma_Y^2 \neq 1 \end{cases}$$

Para ello selecconamos el menú `Estadísticos`→`Varianzas`→`Test F para dos varianzas`. En la ventana `Grupos` elegiremos la variable de agrupación *Admin* y en la ventana `Variable explicada` indicaremos la variable que contiene los datos de la variable bajo estudio, en nuestro caso, la variable *Nivel*. Elegiremos, además, la hipótesis alternativa `Bilateral` al tratarse de un contraste de la forma *igual* frente a *distinto*.

		1	2
	rowname	Nivel	Admin
1	1	23	p
2	2	26	p
3	3	26	p
4	4	22	p
5	5	27	p
6	6	25	p
7	7	24	p
8	8	24	p
9	9	25	p
10	10	25	p
11	11	22	t
12	12	22	t
13	13	19	t
14	14	21	t
15	15	22	t
16	16	23	t
17	17	24	t
18	18	20	t
19	19	23	t
20	20	23	t

Figura 6.3: Introducción de los datos del Ejemplo 5.

En el menú de R-Commander podemos elegir también el test de Barlett y el test de Levenne para contrastar la igualdad de varianzas. Ambos son test válidos para comparar las varianzas poblacionales de dos muestras independientes. No obstante, hemos elegido el test F porque corresponde con el test descrito en las clases teóricas.

Al realizar el contraste sobre las varianzas obtenemos un p-valor igual a 0.9546 por lo que obtenemos como conclusión que no hay evidencia para rechazar la igualdad de las varianzas de ambas poblaciones.

Ahora volvemos a repetir el procedimiento descrito anteriormente para realizar el contraste sobre la diferencia de medias y, teniendo en cuenta el resultado obtenido en el contraste anterior, indicaremos en la opción correspondiente que las varianzas son iguales. Es importante comprobar, en el campo `Diferencia` que realmente se va a contrastar la diferencia de medias deseada.

Obtenemos un p-valor igual a 0.0003018. Al ser menor que la significación del contraste ($\alpha = 0.05$) tenemos evidencia para rechazar la hipótesis nula, esto es, podemos afirmar que el nivel medio de virus de los pacientes que reciben el placebo es superior al de los pacientes que reciben el tratamiento y, por tanto, el tratamiento surte el efecto deseado.

Ejercicio 6: Cargar el fichero de datos Rcars. Plantear y resolver contrastes de hipótesis adecuados para estudiar si, para un nivel de significación del 10 %, existen evidencias significativas para afirmar que el consumo medio de los coches con 8 cilindros es superior al consumo medio de los coches con 4 cilindros.

Nota: Los menús que permiten realizar contrastes sobre las diferencias de medias y los cocientes de varianzas de dos poblaciones únicamente se activan cuando en el conjunto de datos existe una variable categórica con, exactamente, dos modalidades. Esta variable es la que permite definir qué casos corresponden a cada una de las poblaciones. Por ello, para realizar el ejercicio, será necesario realizar los siguientes pasos:

- Filtrar el conjunto de datos activo para construir un nuevo conjunto de datos que contenga, únicamente, los casos correspondientes a los coches que tienen 4 u 8 cilindros.
- Convertir la variable *cilindr* del nuevo conjunto de datos en una variable cualitativa mediante el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Convertir variable numérica en factor`.
- Realizar los contrastes de hipótesis necesarios para resolver el problema.

Ejercicio 7: Cargar el fichero de datos Rcars. Plantear y resolver contrastes de hipótesis adecuados para estudiar si, para un nivel de significación del 5 %, existen evidencias significativas para afirmar que el consumo medio de los coches con europeos es distinto del consumo medio de los coches japoneses.

Nota: En este caso, como paso previo a la resolución del problema, tenemos que construir un nuevo conjunto de datos que contenga únicamente los casos correspondientes a los coches europeos y japoneses. El primer paso consistiría pues en filtrar el conjunto de datos original para seleccionar estos casos. Si una vez hecho esto solicitamos un resumen del conjunto de datos activo observamos que, en el nuevo conjunto de datos, la variable origen sigue presentando tres modalidades: Europa, Japón y E.E.U.U., si bien la modalidad E.E.U.U. tiene frecuencia 0.

Para que en el nuevo conjunto de datos no aparezca la modalidad E.E.U.U., en la variable origen, procederemos como sigue: una vez obtenido este nuevo conjunto de datos, utilizamos el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Recodificar variables`. Elegimos la variable *origen* como variable a recodificar e introducimos las directrices de recodificación: *"Europa"* = *"Europa"* y *"Japón"* = *"Japón"*. El resultado de la recodificación podemos guardarlo en otra variable o bien en la misma variable *origen*. Una vez hecho esto podemos solicitar de nuevo un resumen del conjunto datos y observar que la variable origen creada sólo presenta las dos modalidades deseadas.