EACS-Estadística descriptiva bivariante: Comparación de distribuciones

Índice

1. Introducción

Hay muchas situaciones en las que necesitamos comparar cómo se distribuye una variable en dos poblaciones distintas o en dos grupos de una misma población. Por ejemplo, ¿es el nivel de colesterol el mismo entre los que toman un tratamiento que entre los que no lo toman?, ¿es el efecto de ese tratamiento igual para hombres y mujeres? Estas preguntas son casos particulares de la pregunta general, ¿es la distribución de una variable numérica la misma en dos niveles de una variable categórica?

Podemos atacar este problema con las herramientas de la estadística descriptiva, comparando estadísticos y gráficos de la variable numérica entre los dos grupos que corresponden a dos niveles de una variable categórica. Si la muestra es representativa de un colectivo o población más grande, las discrepancias o similitudes de las distribuciones en la muestra nos darán indicaciones de lo que puede esperarse que ocurra en todo ese colectivo. Si además el muestreo ha sido aleatorio y se ha realizado de forma independiente en dos grupos, podemos utilizar las herramientas de la inferencia estadística para confirmar o rechazar la hipótesis de igualdad de medias (para la posición central) o la hipótesis de igualdad de las varianzas (para la dispersión) de esa variable en los dos grupos.

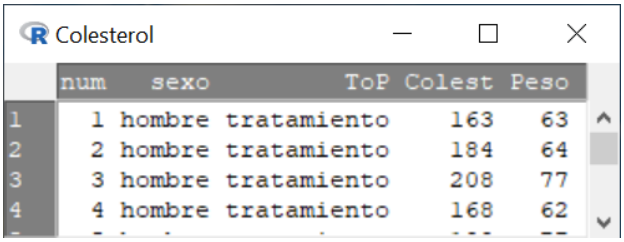
2. Comparación de la distribución de una variable numérica en dos grupos

Queremos comparar la distribución de una variable cuantitativa (variable *numérica* en R Commander) entre dos niveles de una variable cualitativa (variable *factor* en R Commander). Podemos recurrir a las herramientas de la estadística descriptiva (*Estadísticos* y *Gráficas*) que hemos utilizado en prácticas anteriores, pero utilizando la opción de realizar el cálculo o el gráfico *por grupos*.

Vamos a ilustrar el procedimiento con un ejemplo. Queremos comprobar si un nuevo tratamiento reduce el nivel de colesterol. Los datos están en el fichero *Colesterol.RData*. Empezaremos cargando este fichero en R Commander con la opción

Datos > Cargar conjunto de datos…

Comprobamos que el conjunto de datos activo es *Colesterol*. En la pantalla de *Mensajes* nos dice que el fichero tiene 200 filas y 5 columnas. Si lo visualizamos vemos los nombres de las 5 variables: *num*, que contiene el número de identificación del individuo, *sexo* (con niveles *hombre* y *mujer*), *ToP* (con niveles *tratamiento* y *placebo*), y las variables numéricas *Colest* (nivel de colesterol) y *Peso* (peso en kg).



	num	sexo	ToP	Colest	Peso
1	1	hombre	tratamiento	163	63
2	2	hombre	tratamiento	184	64
3	3	hombre	tratamiento	208	77
4	4	hombre	tratamiento	168	62

Los datos provienen de un ensayo clínico, en el que se han seleccionado 200 individuos, 100 hombres y 100 mujeres, siguiendo las especificaciones que garantizan que son representativos del colectivo para el que está diseñado el tratamiento. Para aplicar el tratamiento, se han seleccionado aleatoriamente a 50 mujeres y a 50 hombres. Los otros 50 han recibido el placebo. En ensayo es "doble ciego" y, por lo tanto, ni los pacientes (ni el equipo médico) saben si están recibiendo (o aplicando) tratamiento o placebo. Pasado el periodo de tiempo establecido para que el tratamiento pueda hacer efecto, se

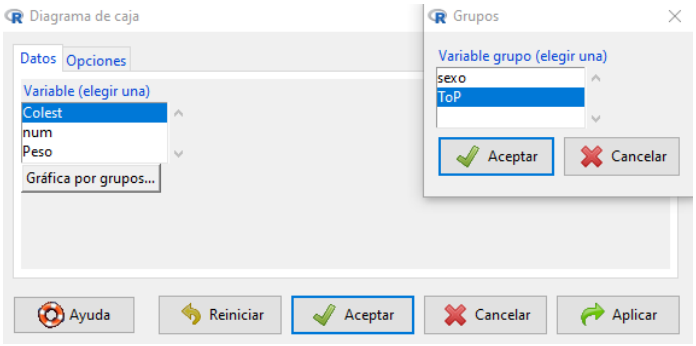
les ha medido a todos el nivel de colesterol en sangre y también se ha obtenido su peso. Por tanto, tenemos 100 individuos que han recibido tratamiento, 50 hombres y 50 mujeres, y otros 100 que han recibido placebo, también 50 hombres y 50 mujeres.

Nos interesa comparar el colesterol de los 100 individuos que han recibido tratamiento (lo llamaremos variable X) con el de los 100 individuos que han recibido el placebo (lo llamaremos variable Y). Notad que las variables X e Y **no** son dos variables del conjunto de datos en R Commander, sino que están en la misma variable *Colest*. X coincide con *Colest* en el nivel *ToP = "tratamiento"* e Y coincide con *Colest* en nivel *ToP = "placebo"*.

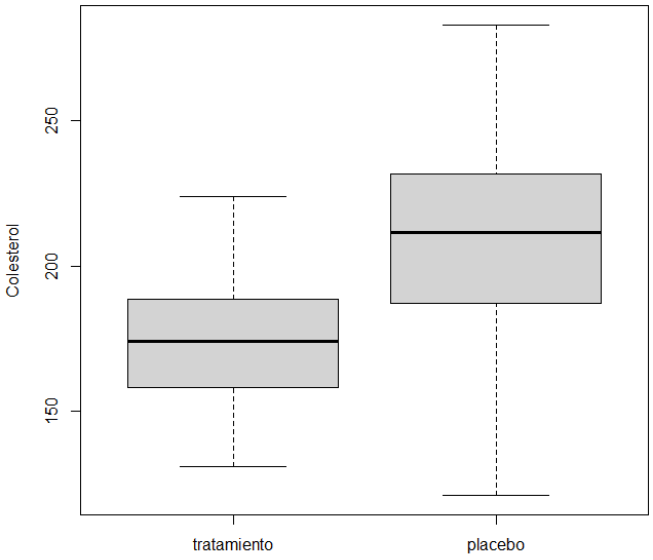
Para obtener los diagramas de caja de X e Y utilizaremos

Gráficas > Diagrama de caja…

En la pestaña *Datos* seleccionamos *Colest*, picamos en el botón *Gráfica por grupos…* y en la ventana que se abrirá, elegimos la variable *ToP*.



En la pestaña *Opciones*, seleccionamos *Automáticamente* en *Identificar atípicos*. Dejamos las opciones por defecto y pulsamos *Aceptar*. Obtendremos el siguiente gráfico

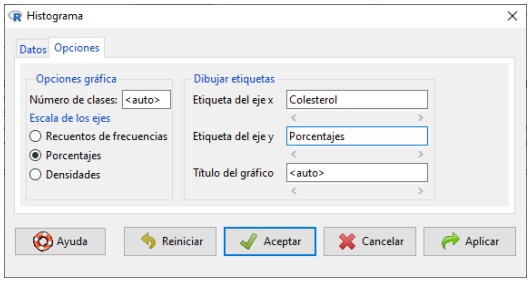


En este caso, no se han encontrado valores atípicos ni en los datos de tratamiento ni en los de placebo.

Para obtener los histogramas, seleccionamos

Gráficas > Histograma…

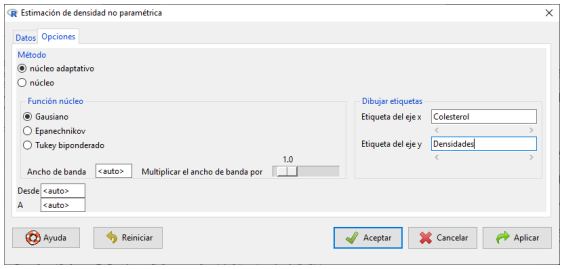
En la pestaña *Datos* seleccionamos *Colest*, picamos en el botón *Gráfica por grupos…* y en la ventana que se abrirá, elegimos la variable *ToP*. En la ventana *Opciones* pedimos que utilice *Porcentajes* en la *Escala de los ejes* y cambiamos las etiquetas de los ejes x e y (pondremos Colesterol y Porcentajes).



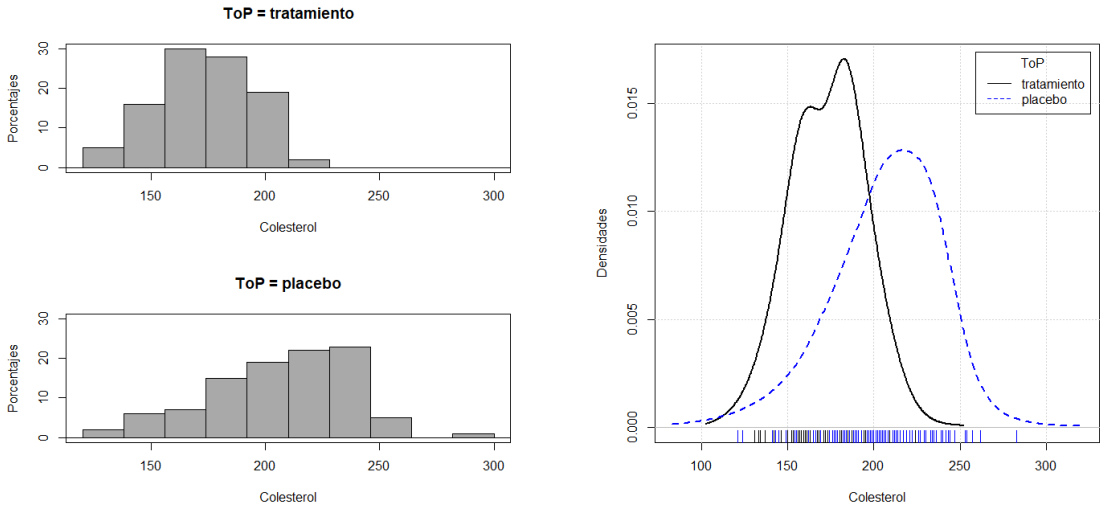
Para obtener las densidades estimadas, seleccionamos

Gráficas > Estimar densidad…

En la pestaña *Datos* seleccionamos *Colest*, picamos en el botón *Gráfica por grupos…* y en la ventana que se abrirá, elegimos la variable *ToP*. En la ventana *Opciones* cambiamos también las etiquetas de los ejes x e y.



A la izquierda podemos ver el histograma y a la derecha la estimación de la densidad que hemos obtenido.

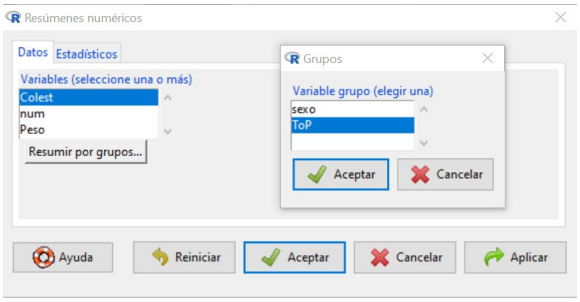


Tanto el histograma como la estimación de la densidad nos permiten comparar la forma de la distribución de la variable *Colest* en los dos grupos, *tratamiento* (variable X) y *placebo* (variable Y). En particular, permiten apreciar la simetría o asimetría de la distribución, y contrastar esta impresión con el valor del coeficiente de asimetría.

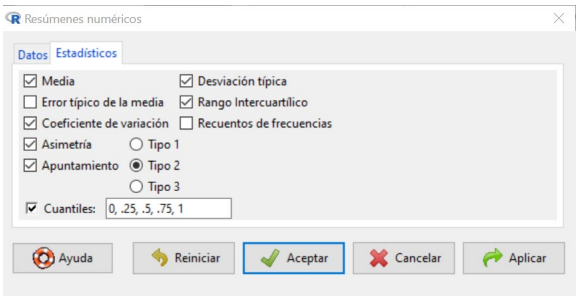
El botón de *Resumir por grupos* también permite comparar los estadísticos de posición, dispersión y forma. Para obtener los estadísticos usamos

Estadísticos > Resúmenes > Resúmenes numéricos

En la pestaña *Datos*, seleccionamos *Colest*. Con el botón *Resumir por grupos…* seleccionamos la variable que determina los dos grupos, en este caso *ToP*.



En la pestaña *Estadísticos* seleccionaremos los estadísticos que queremos comparar. Como estadísticos de posición, la media y los cuantiles 0, 0.25, 0.5, 0.75 y 1 (que son, respectivamente, mínimo, primer cuartil, mediana, tercer cuartil y máximo). Como estadísticos de dispersión, la desviación típica S, el rango intercuartílico y el coeficiente de variación. Como estadísticos de forma los coeficientes de Asimetría y de Apuntamiento.



Al pulsar *Aceptar* obtenemos la salida

```
              mean       sd   IQR        cv    skewness    kurtosis  0%   25%   50%    75% 100% Colest:n
tratamiento 173.61 21.01702 30.25 0.1210588 -0.03291617 -0.72719895 131 158.0 174.0 188.25  224      100
placebo     206.82 31.31733 43.25 0.1514231 -0.47987027  0.07846156 121 187.5 211.5 230.75  283      100
```

Los estadísticos de X están en la fila de tratamiento y los de Y en la de placebo.

Con toda la información proporcionada por estos estadísticos y gráficas, se puede hacer el siguiente comentario. **Vamos a comparar la distribución del nivel de colesterol entre 100 individuos que han recibido tratamiento y 100 individuos que han recibido placebo. Los diagramas de caja no han identificado datos atípicos en ninguno de los grupos. Los histogramas y las estimaciones de la densidad tampoco muestran datos muy extremos.**

La diferencia de medias entre los que han recibido placebo y los que han recibido tratamiento es 33.21 puntos (sus medias respectivas son 206.82 y 173.61). Además de tener una media más pequeña, el grupo de tratamiento también tiene menor dispersión absoluta (con una desviación típica de 21.02) y relativa (con un coeficiente de variación de 0.12) que el grupo de placebo (con una desviación típica de 31.32 y con un coeficiente de variación de 0.15). La mediana y el rango intercuartílico son menores en el grupo de tratamiento, así como los cuartiles 1 y 3, como puede verse en la tabla de estadísticos y en el diagrama de caja, lo que indica que la distribución de los valores del colesterol entre los que han tomado tratamiento es consistentemente menor que la de los que han tomado placebo. Esta impresión la confirman también los histogramas y las densidades estimadas.

Para los datos de tratamiento, la distribución parece bastante simétrica en los tres gráficos, y esta impresión la confirma un coeficiente de asimetría muy pequeño en valor absoluto (-0.033). Para los datos de placebo, el histograma muestra asimetría a la izquierda, y esta impresión la confirma un coeficiente de asimetría negativo y

relativamente grande (-0.48). Esta asimetría también puede apreciarse, aunque no tan claramente, en el diagrama de caja y en la estimación de la densidad.

NOTA 2.1: Para valorar si el valor absoluto de los coeficientes de asimetría o de curtosis es "suficientemente distinto" de 0, hemos indicado, en prácticas anteriores, unas fórmulas que proporcionan límites de referencia y que dependen sólo del tamaño muestral. Para n = 100 los valores de referencia son, respectivamente,

```
> 2*sqrt(6)/sqrt(100)
[1] 0.4898979

> 4*sqrt(6)/sqrt(100)
[1] 0.9797959
```

Ejercicio 2.1: Carga el conjunto de datos *PulsoT4* y analiza la diferencia en pulsaciones después del experimento (*Pulso2*) entre los individuos que han corrido y los que no. La variable categórica que determina esos dos grupos es *Correr*, con los niveles "sí" y "no".

Ejercicio 2.2: Utilizando el mismo conjunto de datos *PulsoT4,* compara las pulsaciones en reposo (*Pulso1*) entre hombres y mujeres. La variable categórica que determina esos dos grupos es *Sexo* con los niveles "hombre" y "mujer".

3. Comparación de medias en dos grupos

Llamamos *media muestral* de una variable X a la obtenida a partir del conjunto de datos (o muestra). La denotamos con $\bar{X}$. Llamamos *media poblacional* a la media de todo el colectivo o población de interés del que se ha extraído esa muestra, y la denotamos con μ_X. Suele ser difícil (a veces imposible) conocer la media poblacional, así que utilizamos $\bar{X}$ como un valor aproximado (esto es, como una *estimación*) de μ_X. Decimos que hacemos *inferencia estadística* cuando utilizamos la información obtenida de la muestra para estimar valores o para tomar decisiones sobre los valores de los parámetros de la población.

En la sección anterior hemos obtenido y comparado las medias muestrales de una variable en dos grupos distintos y hemos obtenido la diferencia entre sus medias muestrales, $\bar{X} - \bar{Y}$. Si las muestras son representativas de la población en los dos grupos esta diferencia, $\bar{X} - \bar{Y}$, se puede utilizar para hacer inferencia estadística sobre la diferencia entre las medias poblacionales $\mu_X - \mu_Y$. En particular, nos planteamos la siguiente pregunta ¿Es la diferencia observada entre medias muestrales suficientemente grande como para poder decir, con algún grado de confianza, que las medias poblacionales son distintas?

Formalmente, podemos plantear esta pregunta como una decisión entre dos hipótesis: la hipótesis (nula) de que estas dos medias son iguales

$$H_0: \mu_X = \mu_Y$$

frente a la hipótesis (alternativa) de que estas medias no son iguales

$$H_1: \mu_X \neq \mu_Y$$

La validez de la inferencia estadística depende de cómo se han obtenido los datos (esto es, del método de muestreo utilizado) y de la distribución de las variables muestreadas. Para poder aplicar este contraste asumimos que se ha realizado un muestreo aleatorio, que el muestreo en los dos grupos se ha realizado de forma independiente y que X e Y tienen una distribución Normal. Si no se puede asegurar la normalidad de X e Y, pero los tamaños muestrales son grandes, estos resultados son también válidos, aproximadamente.

En esta práctica vamos a realizar un contraste (o *test*) de la hipótesis de igualdad de medias. El test proporciona un *p-valor*, *p*, que podemos interpretar como una medida de concordancia de los datos con la hipótesis nula. Valores grandes de *p* indican que la diferencia entre las medias muestrales observada **no** es "suficientemente distinta" de 0. Valores pequeños de *p* indican que sí lo es, y en este caso diremos que hemos encontrado una diferencia *estadísticamente significativa*. Se suele usar la "regla del 5%" y se considera que la diferencia es *estadísticamente significativa* si $p < 0.05$, y diremos que *se rechaza* la hipótesis nula *con un nivel se significación del 5%*. En otro caso, *no rechazamos* la hipótesis nula, o, dicho de otra forma, *retenemos* la hipótesis nula.

NOTA 3.1: Una diferencia observada puede ser estadísticamente significativa y ser irrelevante (insignificante) a efectos prácticos. Por otra parte, podemos encontrar una diferencia muy relevante entre las medias muestrales pero que no sea estadísticamente significativa, bien porque los tamaños muestrales son pequeños o bien porque los procedimientos de muestreo no han sido los adecuados.

Vamos a describir cómo hacer el contraste de igualdad de medias con R Commander, usando el conjunto de datos *Colesterol* (fichero *Colesterol.RData*). Seguimos con la notación utilizada en la sección anterior. Los datos del colesterol están en la variable *Colest* para los dos grupos que se indican en la variable *ToP*. Los datos de los 100 individuos del grupo de *tratamiento* son la variable X, y son los valores de *Colest* en el grupo *ToP = "tratamiento"*. Los datos de los 100 individuos del grupo de *placebo* son la variable Y, y son los valores de *Colest* en el grupo *ToP = "placebo"*.

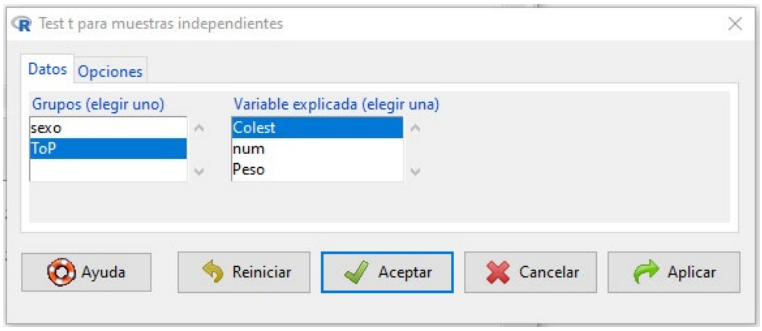
Denotamos la media poblacional de X con μ_X y la media poblacional de Y con μ_Y. En este ejemplo la hipótesis nula (que las medias son iguales) equivale a decir que el tratamiento, en media, no tiene ningún efecto.

El procedimiento de muestreo de este ensayo clínico garantiza que las dos muestras son independientes. Además, el tamaño de las muestras (100 datos de X y 100 datos de Y) permite aplicar el contraste, aunque no estemos seguros de que los datos en cada grupo sean normales. Fijamos el nivel de significación usual del 5% para rechazar la hipótesis nula. Esto es, rechazaremos la hipótesis nula si el p-valor $p < 0.05$.

El contraste de igualdad de medias para muestras independientes está en

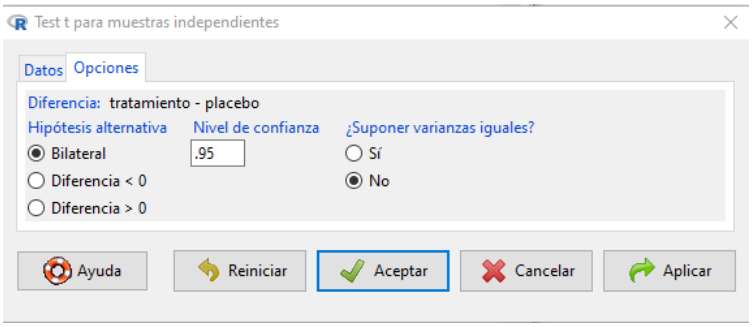
Estadísticos > Medias > Test t para muestras independientes…

En la pestaña *Datos* seleccionamos la variable que indica el grupo (*ToP*) y la que indica la variable (*Colest*).



En la pestaña *Opciones* dejamos marcada la opción *Bilateral.*

Si no podemos asegurar que la variable tiene la misma varianza en los dos grupos, que es la situación más frecuente, dejaremos marcada la opción *No* bajo la pregunta *¿Suponer varianzas iguales?* El contraste de igualdad de medias que no supone que las varianzas sean iguales se conoce como contraste de Welch.



NOTA 3.2: En la pestaña *Opciones* hay una casilla debajo del texto *Nivel de confianza.* Este nivel de confianza no tiene ningún efecto en la realización del contraste ni en el cálculo del p-valor. Solo se utiliza para obtener un intervalo de confianza, de ese nivel, para el valor medio de la diferencia: *Diferencia: tratamiento – placebo* (con nuestra notación corresponde a la diferencia $\mu_X - \mu_Y$).

Si pulsamos *Aceptar*, con las opciones que hemos indicado, se obtiene la salida

```
        Welch Two Sample t-test

data:  Colest by ToP
t = -8.8053, df = 173.14, p-value = 1.315e-15
alternative hypothesis: true difference in means between group tratamiento and group
placebo is not equal to 0
95 percent confidence interval:
 -40.65422 -25.76578
sample estimates:
mean in group tratamiento     mean in group placebo
                   173.61                    206.82
```

El p-valor *p* en este caso es prácticamente cero (p-value = 1.315e-15). Esto indica que hemos encontrado una diferencia estadísticamente significativa, para el nivel de significación prefijado del 5%, entre las estimaciones de las medias, que son 173.61 en el grupo de tratamiento y 206.82 en el grupo de placebo. Por lo tanto, rechazamos la hipótesis nula de que las medias son iguales en los dos colectivos en los que hemos muestreado.

No sólo rechazamos la hipótesis nula. El p-valor está dando información sobre la concordancia de los datos con la hipótesis nula. En este caso es prácticamente 0 ($p = 1.315 * 10^{-15}$) y por tanto indica mucha discrepancia. La hipótesis nula se hubiera rechazado para casi cualquier nivel de significación.

Ejercicio 3.1: Queremos saber si el peso medio de los individuos que han recibido tratamiento es el mismo que el de los que han recibido el placebo. A partir de las variables *Peso* y *ToP* en el conjunto de datos *Colesterol*, el contraste de igualdad de medias, sin asumir que las varianzas sean iguales, proporciona la siguiente salida:

```
        Welch Two Sample t-test

data:  Peso by ToP
t = -0.76217, df = 197.56, p-value = 0.4469
alternative hypothesis: true difference in means between group tratamiento and group
placebo is not equal to 0
95 percent confidence interval:
 -3.443898  1.523898
sample estimates:
mean in group tratamiento     mean in group placebo
                   64.69                     65.65
```

Aunque las estimaciones que se obtienen no son iguales, el p-valor obtenido (p = 0.4469) es muy grande. Tenemos que retener la hipótesis nula de que las medias poblacionales son iguales.

Se puede hacer notar que un p-valor tan alto está indicando un alto grado de concordancia de los datos con la hipótesis nula, y también que este resultado es el esperado, ya que la asignación de tratamiento o placebo se ha hecho de forma aleatoria.

Ejercicio 3.2: Queremos saber si el peso medio de los hombres y las mujeres en la población muestreada es el mismo. A partir de las variables *Peso* y *sexo* en el conjunto de datos *Colesterol*, el contraste de igualdad de medias, sin asumir que las varianzas sean iguales, proporciona la siguiente salida:

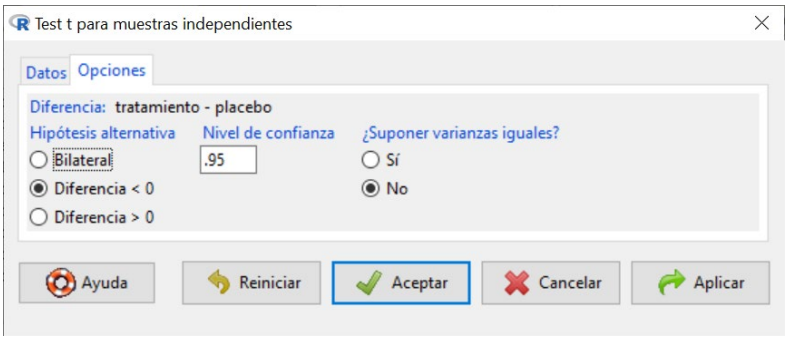
```
        Welch Two Sample t-test

data:  Peso by sexo
t = 10.294, df = 187.67, p-value < 2.2e-16
alternative hypothesis: true difference in means between group hombre and group mujer is
not equal to 0
95 percent confidence interval:
  8.471699 12.488301
sample estimates:
mean in group hombre  mean in group mujer
               70.41                59.93
```

El p-valor es prácticamente 0 y se tiene que rechazar que los hombres y las mujeres de la población muestreada tengan el mismo peso medio. La mejor estimación disponible de ese peso medio en el colectivo del que se ha obtenido la muestra es la que se ha calculado con los datos de la muestra: 70.41 para los hombres y 59.93 para las mujeres.

NOTA 3.3: Cuando se dispone de información adicional que permite asegurar que la media de X nunca puede ser mayor que la media de Y, la hipótesis alternativa puede ser unilateral, esto es

$$H_1: \mu_X < \mu_Y$$

En la pestaña *Opciones* se indica el orden de la diferencia, tal y como la va a considerar el procedimiento de R Commander. Si la variable grupo es *ToP*, utiliza sus etiquetas e indica: *Diferencia: tratamiento – placebo.* Con la notación que hemos usado en el ejemplo anterior, corresponde a la diferencia μ_X - μ_Y. Por tanto, tendríamos que seleccionar *Diferencia < 0*.



Si sabemos que la media de X nunca puede ser menor que la media de Y, la alternativa unilateral será

$$H_1: \mu_X > \mu_Y$$

y deberemos seleccionar *Diferencia > 0*.

Ejercicio 3.3: Vamos a utilizar los datos del conjunto *PulsoT4*. Queremos saber si la media de las alturas de los hombres y las mujeres en la población muestreada es la misma, frente a la alternativa de que la de los hombres es mayor (estamos asumiendo que nunca puede ser menor). A partir de las variables *Altura* y *Sexo* realiza el contraste de igualdad de medias e indica si se puede retener la hipótesis nula (que esas medias son iguales) o si se tiene que rechazar y aceptar la alternativa (que la altura media de los hombres es mayor) al nivel de significación habitual del 5%.

4. Comparación de varianzas en dos grupos

La comparación de las medias poblacionales de una variable en dos grupos nos ha permitido contrastar si esa variable tiene la misma tendencia central en los colectivos muestreados. La comparación de las varianzas poblacionales permitirá contrastar si la dispersión en torno a sus medias es la misma en los colectivos muestreados.

Llamando X a los valores de la variable en la primera población, e Y a los valores de esa misma variable en la segunda población, las varianzas de X e Y **en esas poblaciones** las denotamos σ_X^2 y σ_Y^2. Podemos contrastar la hipótesis (nula) de que las dos varianzas son iguales

$$H_0: \sigma_X^2 = \sigma_Y^2$$

frente a la hipótesis alternativa: que esas varianzas no son iguales

$$H_1: \sigma_X^2 \neq \sigma_Y^2$$

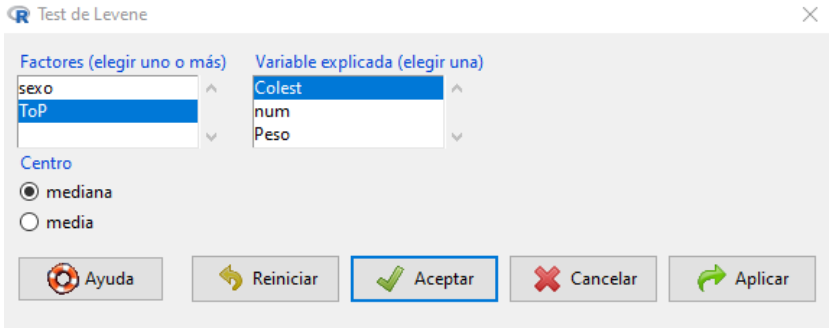
El contraste de igualdad de varianzas proporciona una medida p de concordancia de los datos de la muestra con la hipótesis nula, en este caso que las varianzas poblacionales son iguales. Valores muy pequeños (el límite usual es $p < 0.05$) nos llevarán a rechazar la hipótesis nula y por lo tanto a afirmar que las varianzas son distintas en esas poblaciones.

Vamos a describir cómo hacer el contraste de igualdad de varianzas usando el conjunto de datos *Colesterol*. Llamamos X al nivel de colesterol de los que toman el tratamiento en la población de referencia, y su varianza se denota σ_X^2. La variable Y es el colesterol de los que toman placebo en la población de referencia, y su varianza se denota σ_Y^2.

R Commander tiene implementados varios contrastes de igualdad de varianza. Usaremos el test de Levene, que es más robusto a fallos en la normalidad de las variables. Lo encontramos en

Estadísticos > *Varianzas* > *Test de Levene*

Seleccionamos la variable que indica el grupo (*ToP*) y la que indica la variable (*Colest*) y dejamos marcada la opción *mediana* bajo el texto *Centro*.



Pulsando *Aceptar*, proporciona dos salidas. En la primera proporciona las varianzas muestrales en cada grupo. Estas varianzas muestrales se pueden considerar como valores aproximados (estimaciones) de los valores de las varianzas poblacionales.

```
> Tapply(Colest ~ ToP, var, na.action=na.omit, data=Colesterol) # variances by group
tratamiento     placebo
   441.7151    980.7754
```

En la segunda, proporciona el p-valor del contraste de Levene

```
> leveneTest(Colest ~ ToP, data=Colesterol, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
       Df F value    Pr(>F)
group   1    10.2 0.001634 **
      198
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La varianza muestral en el grupo placebo (980.7754) es muy distinta, casi el doble, que la observada en el grupo tratamiento (441.7151). El p-valor del contraste de Levene es $p = 0.001634$, mucho menor que 0.05. Por lo tanto, con el nivel de significación usual del 5% se rechaza la hipótesis de que las varianzas sean iguales en los dos grupos.

La conclusión práctica es que el tratamiento también afecta a la varianza del nivel de colesterol en la población de la que hemos extraído la muestra. En la sección anterior hemos visto que el colesterol medio de los que toman tratamiento es menor que el de los que toman placebo. Con el contraste de Levene vemos que los niveles de colesterol tienen menor dispersión entre los que toman tratamiento que entre los que toman placebo.

Ejercicio 4.1: Vamos a utilizar los datos del conjunto *PulsoT4*. Asumimos que los datos del pulso en reposo (*Pulso1*) se han recogido de forma aleatoria y que los datos de chicos y chicas se han obtenido de forma independiente. Si consideramos las variables X "pulso en reposo de chicos" e Y "pulso en reposo de chicas" y fijamos un nivel de significación del 5%, ¿se puede rechazar que tengan la misma media? ¿Se puede rechazar que tengan la misma varianza?

5. Ejercicios

1. Se ha realizado un estudio para ayudar a comprender el efecto que tiene el hábito de fumar en los patrones de sueño. La variable considerada es el tiempo que tarda una persona en quedarse dormida, en minutos. El fichero *sueno.RData* contiene los datos de los tiempos para personas fumadoras y personas no fumadoras.
   1.1. Si asumimos que esta muestra se ha obtenido aleatoriamente de la población en estudio, ¿se puede mantener la hipótesis de que el hábito de fumar no afecta (en media) al tiempo que tardan en quedarse dormidos?
   1.2. Utilizando las medidas de posición, dispersión y forma por grupos, los histogramas por grupos, las estimaciones de densidad por grupos y los diagramas de caja por grupos, haz un estudio descriptivo de la variable tiempo en quedarse dormido, comparando su comportamiento entre los fumadores y los no fumadores. ¿Tienen esos tiempos la misma distribución entre fumadores y no fumadores?
   1.3. En particular, ¿tienen esos tiempos la misma dispersión entre fumadores que entre no fumadores? Haz un contraste de igualdad de varianzas y comenta el resultado que obtienes.
2. Con los datos del fichero *PulsoT4.RData* compara la variable *Peso* entre hombres y mujeres (variable *Sexo*).
   2.1. Haz un estudio descriptivo, utilizando las medidas de posición, dispersión y forma, los histogramas, las estimaciones de densidades y los diagramas de caja.
   2.2. Si asumimos que esta muestra se ha obtenido aleatoriamente de entre los alumnos de un centro de estudios y que los datos de chicos y chicas que se han obtenido son independientes, ¿se puede retener la hipótesis de que el peso medio de chicos y chicas es el mismo en este centro de estudios?
3. Con los datos del fichero *PulsoT4.Rdata*, compara la variable *Pulso1*, que contiene el pulso en reposo antes del experimento, diferenciando entre las personas que van a correr y las que no (variable *Correr*).
   3.1. Haz un estudio descriptivo, utilizando las medidas de posición, dispersión y forma, los histogramas, las estimaciones de densidades y los diagramas de caja.
   3.2. Si asumimos que esta muestra se ha obtenido aleatoriamente de entre los alumnos de un centro de estudios y que la decisión de correr o no se ha tomado con el resultado del lanzamiento de una moneda, ¿se puede rechazar la hipótesis de que la decisión de correr (o no) no afecta (en media) al pulso en reposo, antes del experimento, para los individuos de este centro de estudios?
4. Se están estudiando dos medicamentos, A y B, para combatir el virus de la gripe. Se han administrado por vía oral dosis únicas de 100 mg a adultos sanos. La variable estudiada es el tiempo requerido en minutos para alcanzar la concentración máxima en plasma. Los datos obtenidos están en el archivo *gripe.RData*.
   4.1. ¿Encuentras algún dato atípico en la variable tiempo requerido para alcanzar la concentración máxima para cada uno de los medicamentos? En caso afirmativo, justifica cuál o cuáles de ellos podrían eliminarse y/o corregirse. Utiliza el diagrama de caja para detectar atípicos y compara lo que obtienes cuando no distingues entre los dos medicamentos y cuando sí distingues entre ellos.
   4.2. En las indicaciones del procedimiento para medir la concentración máxima nos indican que valores inferiores a 50 o mayores de 350 son erróneos. Realiza el estudio descriptivo de la variable tiempo requerido para alcanzar la concentración máxima para cada uno de los medicamentos, una vez eliminados los datos erróneos.
   4.3. Si el experimento se ha realizado con una muestra aleatoria de la población estudiada y la asignación de medicamentos también ha sido aleatoria, con los datos obtenidos para esta muestra, ¿puede retenerse la hipótesis de que el tiempo medio para alcanzar la concentración máxima es el mismo para los dos tratamientos?
5. En el fichero *Colesterol.RData* se recogen los datos de un ensayo clínico para comprobar el efecto de un tratamiento para reducir el nivel de colesterol en sangre. Uno de los objetivos del ensayo clínico era determinar si el efecto del tratamiento es el mismo entre los hombres que entre las mujeres, y por eso se había aleatorizado teniendo en cuenta también el sexo de los participantes en el estudio.
   5.1. Analiza el efecto del tratamiento sobre el nivel de colesterol, pero sólo con las mujeres del estudio.
   5.2. ¿Se puede mantener o se debe rechazar la hipótesis de que el nivel de colesterol (variable *Colest*) en las mujeres que han recibido tratamiento es, en media, igual que las que no lo han recibido? ¿Se puede rechazar al 5% la hipótesis de que la varianza es la misma en los dos grupos?