Ejemplo 10

En una planta de procesamiento de frutas, la máquina que se utiliza normalmente para envasar fresas congeladas produce paquetes con una media de 250 g/caja. Se está estudiando una nueva máquina que procesará mucho más rápido, pero que puede producir resultados más variables. Para investigar esta preocupación, el supervisor de control de calidad midió el contenido de 50 cajas producidas por cada máquina y encontró $s_O^2 = 25\ g^2$ y $s_N^2 = 64\ g^2$. A partir de sus resultados, ¿se confirman sus sospechas? [1]

Las hipótesis son:

$$H_0: \sigma_N^2 \leq \sigma_O^2$$

$$H_a: \sigma_N^2 > \sigma_O^2$$

Nota:

f.test2(sx, nx, sy, ny, alternative = "two.sided", conf.level = 0.95)

Detalles: sx y sy son las desviaciones estándar de las dos muestras y nx y ny son los tamaños de muestra correspondientes. El argumento opcional alternative da la hipótesis alternativa para la prueba. La alternativa por defecto es "two.sided" y las otras opciones posibles son "less" o "greater". El argumento opcional conf.level da el nivel de confianza que se utilizará en la prueba; el valor por defecto de 0,95 equivale a α = 0,05.

Descargue la función utilizando el comando source( ). La función sólo debe descargarse una vez por sesión.

> source (https://waveland.com/Glover-Mitchell/f.test2.txt)

```
> source("http://waveland.com/Glover-Mitchell/f.test2.txt")

## Downloaded: f.test2( ).

> f.test2(sx = sqrt(64), nx = 50, sy = sqrt(25), ny = 50, alternative = "greater")

##
##  F test to compare two variances
##
## data:  sx = 8, nx = 50; and sy = 5, ny = 50
## F = 2.56, num df = 49, denom df = 49, p-value = 0.0006496
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  1.59274     Inf
## sample estimates:
## ratio of variances
##               2.56
```

f.test2( ) proporciona las mismas dos formas que var.test( ) para responder si las varianzas son significativamente diferentes. El valor P es aproximadamente 0,00065, que es mucho más pequeño que $\alpha = 0,05$, por lo que hay una fuerte evidencia para rechazar la hipótesis nula de que la varianza de la nueva máquina es menor o igual que la varianza de la máquina antigua.

De forma equivalente, f.test2( ) proporciona el intervalo de confianza para el valor de F para la hipótesis alternativa especificada. En este ejemplo, el intervalo de confianza del 95% $[1,593,\infty)$ no contiene el valor F esperado de 1. Por tanto, se rechaza la hipótesis nula.

Ejemplo 11

Descargue el conjunto de datos de calcio de la Biblioteca de Datos e Historias: http://lib.stat.cmu.edu/DASL/Datafiles/Calcium.html. Los datos se recogieron para investigar si el aumento de la ingesta de calcio reduce la presión arterial. 21 personas participaron en este experimento, en el que diez de ellas tomaron un suplemento de calcio durante 12 semanas, mientras que los 11 restantes recibieron un placebo. Se midió la presión arterial de cada sujeto antes y después del periodo de 12 semanas. Trace el histograma de las variables Inicio y Fin. Compara los dos histogramas en cuanto a su tendencia central y la forma de su histograma [2].

Después de cargar "Calcium.txt" dar click → `Data → import data → from text file, clip board, or URL.`

Se crea los histogramas dando click → `Graphs → Histrogram` y se selecciona variable numérica

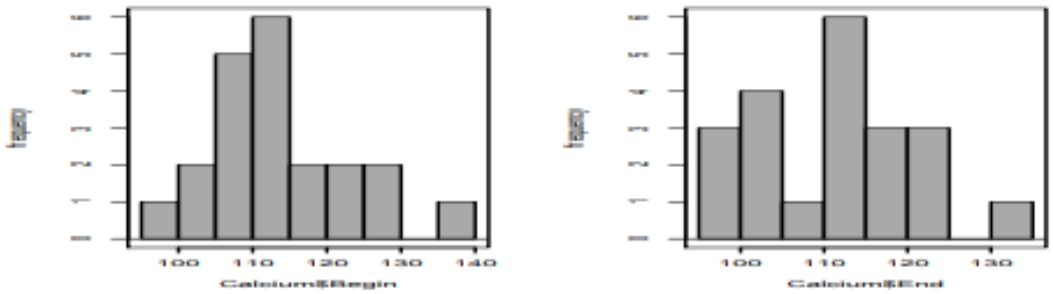


Figura 1. Histograma

El histograma de la presión arterial al inicio del estudio (antes del tratamiento) era unimodal y ligeramente sesgado hacia la derecha. La frecuencia de observaciones es alta, alrededor de 115. El histograma de presión arterial al final del experimento fue bimodal. La presencia de múltiples patrones en el histograma suele indicar que nuestra muestra no es homogénea y contiene subgrupos. En este caso, es trivial identificar los dos subgrupos: son el grupo de calcio y el grupo de placebo, por supuesto, así fue como se diseñó el experimento (es decir, los sujetos se dividieron en dos grupos de tratamiento). Si bien el experimento comenzó con un grupo homogéneo de sujetos (en un esfigmomanómetro), al final del experimento, los

sujetos asignados al grupo de calcio tenían una presión arterial media más baja, por lo que su distribución fue diferente a la del grupo placebo, por lo que hay dos modos.

Ejemplo 12

De la base de datos “BodyTemperature.txt” encuentre el resumen de datos de cinco números para todas las variables numéricas. Para las variables numéricas, proporcione los histogramas y boxplots [2].

Haciendo uso de la base de datos “BodyTemperature.txt” importarlo a R-Commander. Para encontrar cinco resúmenes numéricos para las variables numéricas ir a Statistics summaries Numerical Summeries y seleccionar todas las variables numéricas Edad, Ritmo cardíaco y temperatura y click en Ok para ver los resultados

	0%	25%	50%	75%	100%	n
Age	21.0	33.75	37.0	42.0	50.0	100
HeartRate	61.0	69.00	73.0	78.0	87.0	100
Temperature	96.2	97.70	98.3	98.9	101.3	100

La figura 2 muestra los histogramas y boxplots de "Edad", "Ritmo cardíaco" y "Temperatura". En el caso de la "Edad", el histograma está ligeramente inclinado hacia la izquierda; no hay ningún valor atípico; la tendencia central se sitúa en torno a los 35-40; podríamos utilizar la media de la muestra (37,62) o la mediana de la muestra (37,00) como medida de la tendencia central. Para la "Frecuencia cardíaca", el histograma es casi simétrico; de nuevo, no hay ningún valor atípico; la tendencia central se sitúa en torno a 70-75; como antes, podemos utilizar la media de la muestra (73,66) o la mediana de la muestra (73) como medida de la tendencia central. En el caso de la "Temperatura", parece haber una bi modalidad: hay una moda en torno a 98,5 y otra después de 100. Es posible que la muestra incluya un grupo de individuos con fiebre leve, aunque la población objetivo sean individuos sanos. Por otra parte, dado que sólo hay unos pocos (4) individuos con temperatura corporal superior a 100, podrían ser simplemente valores atípicos. El diagrama de caja muestra que dos de ellos pueden considerarse como valores atípicos (señalados con puntos). La tendencia central se sitúa en torno a 98-99.

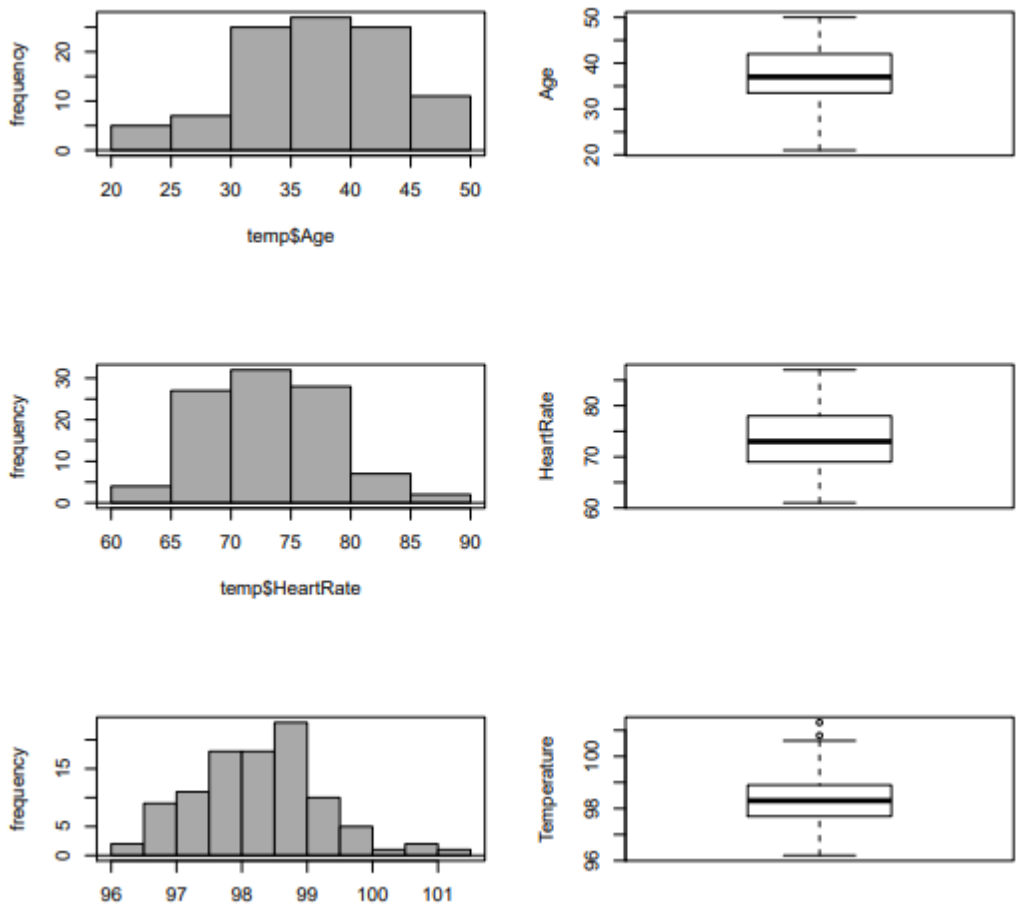


Figura 2. Histogramas y boxplots de "Edad", "Ritmo cardíaco" y "Temperatura".

Ejemplo 13

Utilizando el conjunto de datos "BodyTemperature.txt", crea el gráfico de dispersión de la temperatura corporal en función de la frecuencia cardíaca. Describe el patrón y comenta la posible relación entre las dos variables. Encuentra el coeficiente de correlación entre la temperatura corporal y la frecuencia cardíaca. Finalmente, crea boxplots de la temperatura corporal para hombres y mujeres por separado [2].

Después de cargar BodyTemperature en R-Commander, para crear el gráfico de dispersión, haga click en Graphs → scatterplot y seleccione “HeartRate" para la variable x y “Temperature" para la variable y. Para hacer un gráfico de dispersión con sólo la línea de mínimos cuadrados (es decir, la línea de tendencia) se debe desmarcar las opciones “Smooth line", “Show spread", and “Marginal boxplots", y después click OK. El gráfico de dispersión entre la temperatura corporal y la frecuencia cardíaca se muestra en el panel izquierdo de la figura 3. El gráfico sugiere que el aumento de la frecuencia cardíaca tiende a coincidir con el aumento de la temperatura corporal. Las dos variables parecen tener una relación lineal positiva. Para encontrar el coeficiente de correlación entre la temperatura corporal y la frecuencia cardíaca, ir a Statistics → Summaries → Correlation matrix., seleccionar

“Temperature" and “HeartRate", click OK. Debería obtener una correlación = 0,448. Este coeficiente de correlación está de acuerdo con lo que encontramos al examinar el gráfico de dispersión. De nuevo, para crear boxplots, ir a Graphs → boxplot, resalte “Temperature", click on “Plot by groups" to selecione Gender, click OK. Esto creará boxplots de temperatura por separado para hombres y mujeres. Los boxplots de temperatura por género se muestran en el panel derecho de la Figura 3. La temperatura corporal de los hombres tiende a ser ligeramente inferior. Además, la temperatura corporal de los hombres parece estar más dispersa en comparación con la de las mujeres.

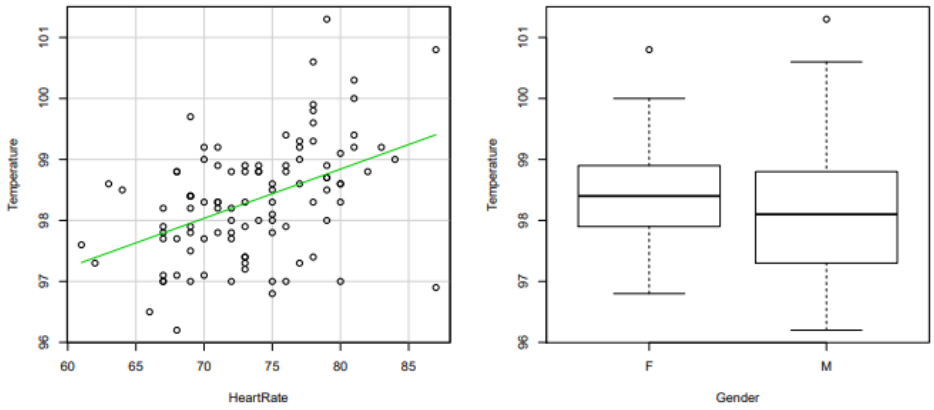


Figura 3. Gráfico de dispersión frecuencia cardiaca y temperatura (izquierdo) y Boxplots temperatura y genero (derecho)

Ejemplo 14

Haciendo uso del conjunto de datos "BodyTemperature.txt”. Para la variable de frecuencia cardíaca, queremos evaluar las siguientes hipótesis. Fijamos el nivel de significación (corte) en 0,01.

- Evalúe la hipótesis de que la media poblacional es inferior a 75. Escriba las hipótesis nula y alternativa
- Evalúe la hipótesis de que la media de la población es diferente de 75. Escriba las hipótesis nula y alternativa

Primero se cargar BodyTemperature en R-Commander, click Statistics → Means

→ Single-sample t-test y $\mu_0 = 75$ determinamos las hipotesis $H_A: \mu < 75$ vs $H_0: \mu = 75$ utilizamos una prueba t unilateral seleccionando la media de la población $\mu < \mu_0$ en este caso el valor p es 0.007, el cual es un nivel de corte menor de 0.01 con el que podemos rechazar la hipótesis nula con 0.01 nivel de significancia. Por lo tanto, la diferencia observada con respecto a 75 es estadísticamente significativa a este nivel.

```
        One Sample t-test

data:  NormTemp$HeartRate
t = -2.5222, df = 99, p-value = 0.006629
alternative hypothesis: true mean is less than 75
95 percent confidence interval:
     -Inf 74.54215
sample estimates:
mean of x
    73.66
```

Para la segunda evaluación donde $H_A: \mu \neq 75 \; vs \; H_0: \mu = 75$ utilizamos una prueba t unilateral seleccionando la media de la población diferente de 0, en este caso el valor de p = 0.013, cual es mas grande que corte de 0.01 entonces no podemos rechazar la hipótesis nula con 0.01 nivel de significancia. Por lo tanto, la diferencia observada con respecto a 75 no es estadísticamente significativa a este nivel.

```
        One Sample t-test

data:  NormTemp$HeartRate
t = -2.5222, df = 99, p-value = 0.01326
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 72.60581 74.71419
sample estimates:
```

Ejemplo 15

Una gerontóloga que investigaba varios aspectos del proceso de envejecimiento quería comprobar si mantenerse "delgado y mantenía", es decir, por debajo del peso corporal normal, alargaría la vida. Asignó aleatoriamente a ratas recién nacidas de una línea altamente endogámica una de las tres dietas: (1) acceso ilimitado a la comida, (2) el 90% de la cantidad de comida que una rata de ese tamaño comería normalmente, o (3) el 80% de la cantidad de comida que una rata de ese tamaño comería normalmente. Mantuvo a las ratas con las tres dietas durante toda su vida y registró su esperanza de vida (en años). ¿Hay pruebas de que la dieta afectó a la duración de la vida en este estudio?

Sin limites	**95% dieta**	**80 % dieta**
2.5	2.7	3.1
3.1	3.1	2.9
2.3	2.9	3.8
1.9	3.7	3.9
2.4	3.5	4.0

Datos en: http://waveland.com/Glover-Mitchell/Example08-1.txt.

```
> data.Ex08.1 <- read.table("http://waveland.com/Glover-Mitchell/Example08-1.txt",
+ header = TRUE)
> data.Ex08.1

##    Lifespan      Diet
## 1       2.5 Unlimited
## 2       3.1 Unlimited
## 3       2.3 Unlimited
## 4       1.9 Unlimited
## 5       2.4 Unlimited
## 6       2.7   90%Diet
## 7       3.1   90%Diet
## 8       2.9   90%Diet
## 9       3.7   90%Diet
## 10      3.5   90%Diet
## 11      3.1   80%Diet
## 12      2.9   80%Diet
## 13      3.8   80%Diet
## 14      3.9   80%Diet
## 15      4.0   80%Diet
```

Tenga en cuenta que el formato de los datos difiere del texto. Cada fila consiste en una respuesta Lifespan (Vida útil) a un nivel de tratamiento Diet (Dieta). Realizamos el ANOVA con el comando aov( ) y ponemos el resultado en la tabla aov.Ex08.1. Tenga en cuenta que estamos realizando un análisis de la varianza de la duración de la vida por los factores de la dieta. Los argumentos de aov( ) lo reflejan utilizando los nombres de las cabeceras de las columnas correspondientes en el marco de datos data.Ex08.1.

```
> aov.Ex08.1<- aov(Lifespan ~ Diet, data = data.Ex08.1)
```

Para ver realmente los resultados del ANOVA, utilice el comando summary( )

```
> summary(aov.Ex08.1)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Diet         2   3.15   1.573     7.7 0.0071 **
## Residuals   12   2.45   0.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observe los dos asteriscos al final de la fila de la Dieta en la tabla. Utilizando los códigos de significación en la salida, esto significa que la prueba F es significativa al nivel 0,01. De hecho, el valor P se indica en la tabla como Pr(>F) = 0,00707.

Para ver las medias por tratamiento, utilice la función tapply( ). Este comando se utiliza para aplicar una función (en este caso, la media) a grupos (en este caso, el tipo de dieta) dentro de una tabla o marco de datos. En otras palabras, tapply( ) puede considerarse como una abreviatura de "aplicar a la tabla".

Nota:

tapply(x, g, FUN = function)

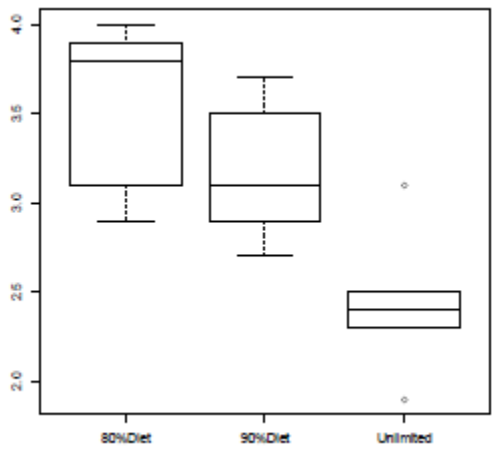
Detalles: x es típicamente un vector (columna) de datos y g es una lista de factores correspondientes de la misma longitud que x. FUN es la función que se aplicará a cada grupo de celdas definidas por las variables categóricas en g.

```
> tapply(data.Ex08.1$Lifespan, data.Ex08.1$Diet, mean)

##   80%Diet   90%Diet Unlimited
##      3.54      3.18      2.44
```

Boxplot (diagrama de caja) de los datos proporciona una indicación de si alguno de los tratamientos (dietas) es significativamente diferente. Puede que no nos tomemos el tiempo de hacer esto a mano, pero R lo hace fácil.

```
> boxplot(Lifespan ~ Diet, data = data.Ex08.1)
```



(El gráfico de cajas de la duración de la vida según la dieta indica que hay una diferencia entre la duración media de la vida de las dietas del 80% y de las ilimitadas. Otras comparaciones son menos claras.)

La prueba F fue significativa (véase la tabla ANOVA anterior) y el gráfico de caja indica que hay una diferencia entre algunas medias de vida. Para llevar a cabo las pruebas t por pares de Bonferroni-Holm para localizar cualquier diferencia en las medias de vida

Nota:
pairwise.t.test(x, g, p.adjust.method = "holm", conf.level = 0.95)

Detalles: x son los datos de medición o respuesta, g contiene los niveles de tratamiento correspondientes, y p.adjust.method = "holm" ajusta los valores P de la secuencia de pruebas según el método seleccionado. La abreviatura p.adj puede utilizarse en lugar de p.adjust.method. El argumento opcional conf.level da el nivel de confianza que se utilizará en la prueba; el valor por defecto de 0,95 equivale a α = 0,05. Hay argumentos adicionales argumentos adicionales para esta función que no son necesarios en este momento.

TukeyHSD(fit, ordered = TRUE, conf.level = 0.95)

Detalles: fit es la salida de la función aov( ) (ANOVA) realizada anteriormente en el marco de datos correspondiente. El argumento ordered especifica si se deben ordenar las medias en el análisis. El valor por defecto es FALSE, pero para ajustarse más al procedimiento del texto, utilice ordered = TRUE. El argumento opcional conf.level da el nivel de confianza que se utilizará en la prueba; el valor por defecto de 0,95 equivale a α = 0,05.

En este ejemplo, data.Ex08.1$Lifespan es el dato de respuesta y data.Ex08.1$Diet contiene el tratamiento correspondiente (grupo).

```
> pairwise.t.test(data.Ex08.1$Lifespan, data.Ex08.1$Diet, p.adj = "holm")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data.Ex08.1$Lifespan and data.Ex08.1$Diet
##
##           80%Diet 90%Diet
## 90%Diet   0.232   -
## Unlimited 0.007   0.047
##
## P value adjustment method: holm
```

El resultado de pairwise.t.test( ) se presenta en forma de tabla. La entrada en la columna 80%Dieta y la fila 90% Dieta es 0,232. Este es el valor P para la prueba de la diferencia de medias, por lo que no es significativo al nivel α = 0,05. Observaciones similares se aplican a las demás entradas. La tabla indica que la dieta ilimitada es diferente de las dietas del 80% y del 90%; los valores P correspondientes son 0,007 y 0,047, respectivamente. Este último es apenas significativo.

Otra forma de realizar comparaciones emparejadas puede ser la prueba de la diferencia significativa de Tukey. diferencia significativa de Tukey. La función R correspondiente es TukeyHSD( ).

En este ejemplo, la salida del ANOVA anterior reside en aov.Ex08.1. Las comparaciones por pares se llevan a cabo en el nivel a = 0,05 utilizando el nivel conf.por defecto = 0,95 como como sigue.

```
> TukeyHSD(aov.Ex08.1, ordered = TRUE)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = Lifespan ~ Diet, data = data.Ex08.1)
##
## $Diet
##                     diff        lwr     upr    p adj
## 90%Diet-Unlimited 0.74 -0.0227167 1.50272 0.057470
## 80%Diet-Unlimited 1.10  0.3372833 1.86272 0.006070
## 80%Diet-90%Diet   0.36 -0.4027167 1.12272 0.443277
```

La tabla de resultados proporciona un intervalo de confianza del 95% para la diferencia de cada par de medias y el valor P correspondiente. Las medias son significativamente diferentes al nivel $\alpha = 0{,}05$ sólo si 0 no está en el intervalo de confianza. La media de la dieta del 80% es diferente de la media de la dieta ilimitada porque el intervalo de confianza [0,337, 1,863] no contiene 0. El valor P correspondiente es 0,006. Utilizando la prueba de Tukey, las dietas del 90% y la ilimitada no logran ser significativamente diferentes en este pequeño estudio porque p = 0,057 y el intervalo de confianza es [-0,023, 1,502]. Este resultado es diferente al de la prueba de Bonferroni-Holm. La prueba de Tukey es más conservadora que la de Bonferroni-Holm. Hay una menor probabilidad de cometer un error de tipo I, pero una mayor probabilidad de cometer un error de tipo II (aceptar una falsa hipótesis nula). En este caso, un tamaño de muestra mayor podría haber dado un resultado más claro.

Ejemplo 16

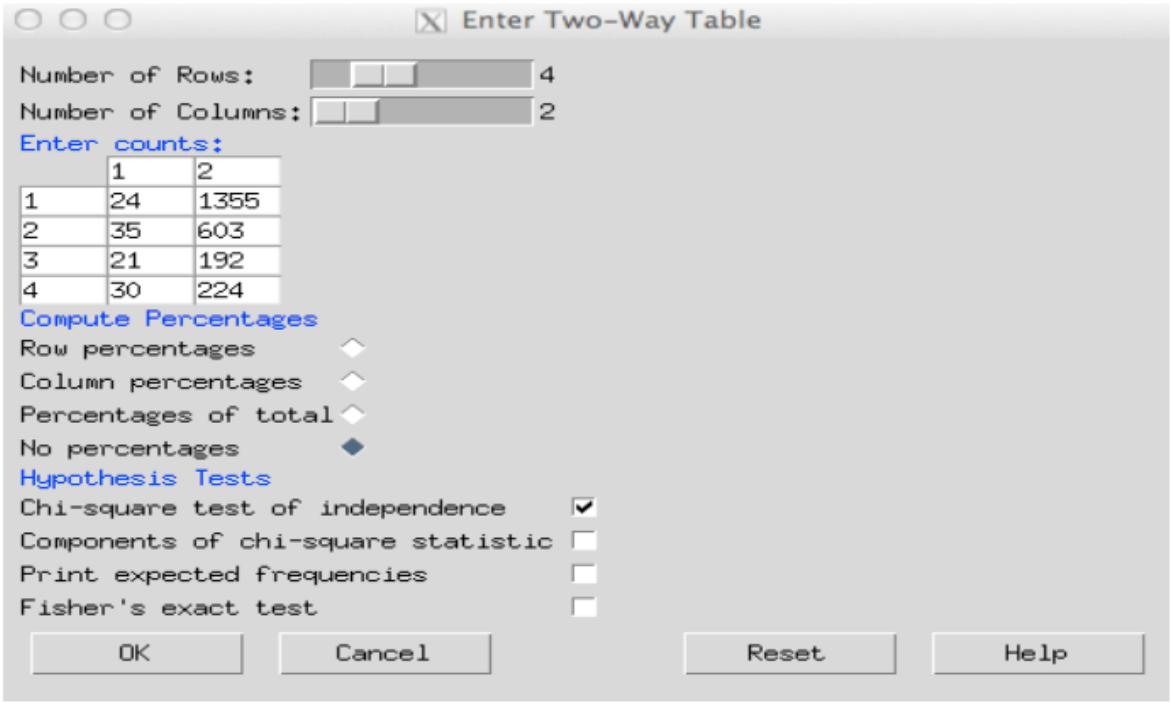
Considere los datos de la tabla X recogidos para investigar los ronquidos como factor de riesgo de enfermedades cardiacas. Utilice la prueba χ^2 de Pearson para examinar si la relación entre la gravedad de los ronquidos y el riesgo de enfermedad cardiaca es estadísticamente significativa.

Tabla 1. Frecuencias de las personas con enfermedad cardiaca para los diferentes niveles de ronquidos basado en una muestra de 2484 personas

Severidad del ronquido	Enfermedad Cardiaca	Total
Nunca	24	1379
Ocasionalmente	35	638
Casi cada noche	21	213
Cada noche	30	254

En R-Commander, damos click en Statistics → Contingency tables → Enter and analyze two-way tables, después se crea una tabla de 4x2, se ingresan las frecuencias de la tabla 1 y se presiona OK. Y el resultado de la prueba χ^2 de Pearson muestra independencia de la

relación entre ronquido y la enfermedad cardiaca es estadísticamente significativa (valor p = $1.08X10^{-15}$)




```
        Pearson's Chi-squared test

data:  .Table
X-squared = 72.7821, df = 3, p-value = 1.082e-15
```

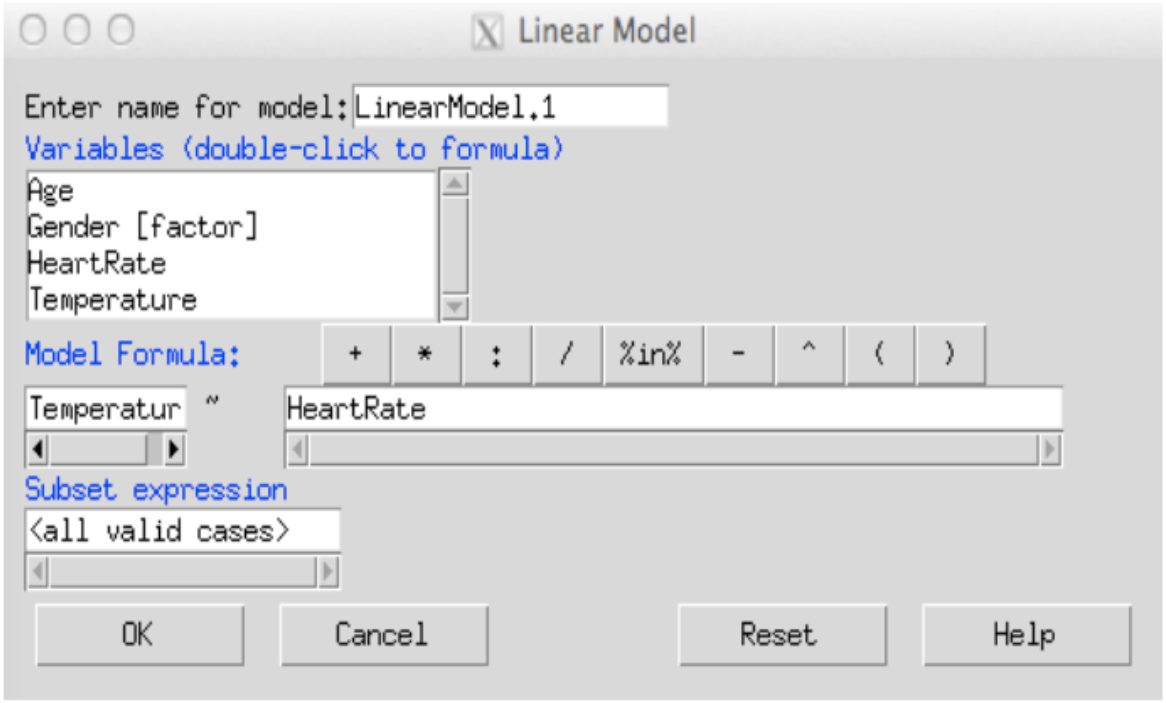
Ejemplo 17

Queremos examinar la relación entre la temperatura corporal Y y la frecuencia cardíaca X. Además, nos gustaría utilizar la frecuencia cardíaca para predecir la temperatura corporal.

a) Utilice el conjunto de datos "BodyTemperature.txt" para construir un modelo de regresión lineal simple para la temperatura corporal utilizando la frecuencia cardíaca como predictor.
b) Interprete la estimación del coeficiente de regresión y examine su significación estadística.
c) Encuentra el intervalo de confianza del 95% para el coeficiente de regresión.
d) Encuentre el valor de R2 y demuestre que es igual al coeficiente de correlación de la muestra.
e) Cree gráficos de diagnóstico sencillos para su modelo e identifique posibles valores atípicos.
f) Si la frecuencia cardíaca de una persona es de 75, ¿cuál sería su estimación de la temperatura corporal de esta persona?

En R-Commander cargamos los datos “BodyTemperature.txt”

a) Después de cargar los datos damos click Statistics → Fit models → Linear model y seleccionamos temperatura and frecuencia cardiaca como variable de respuesta y predictor respectivamente




```
Call:
lm(formula = Temperature ~ HeartRate, data = NormTemp)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50562 -0.46473  0.00543  0.48943  2.53943

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.39068    1.20144  76.900  < 2e-16 ***
HeartRate    0.08063    0.01627   4.956 3.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.86 on 98 degrees of freedom
Multiple R-squared: 0.2004,        Adjusted R-squared: 0.1923
F-statistic: 24.56 on 1 and 98 DF,  p-value: 3.011e-06
```

b) La estimación del coefiente de regresión de la frecuencia cardica es $\widehat{\beta_1} = 0.08$ esto que en promedio un amento unitario de la frecuencia cardiaca coincide con un

aumento de 0.08°F aumentado en la temperatura del cuerpo. Esta relación es estadísticamente significativa con p-valor = 3.01×10^{-6}.

c) Con un intervalo de confianza de 95% para β_1 es [0.08-2 x 0.016, 0.08 +2 x 0.016] = [0.048,0.12]. También se puede obtener el intervalo de confianza dando click Model → Confidence intervals.

d) R^2 = 0.2004 el cual es igual a cuadrado de coefieiente de correlación r=0.4477

e) Dando click Model → Graphs → Basic diagnostic plots para obtener gráficos de diagnóstico.

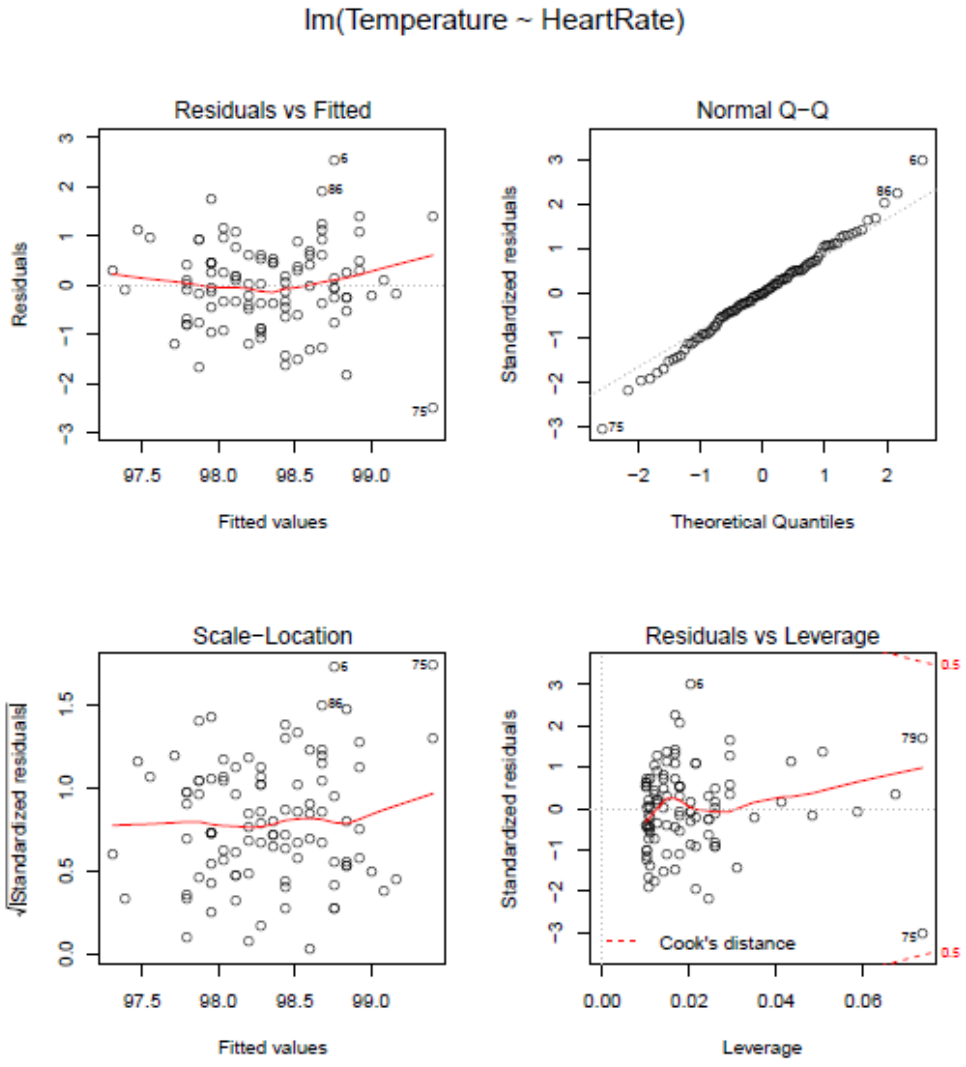


Figura 4. Gráficos de diagnóstico.

f) De acuerdo con el modelo

$$\hat{y} = 92.39 + 0.08x$$
$$= 92.39 + 0.08\ x\ 75$$
$$= 98.39$$

Referencias

[1] K. Mitchell y T. Glover, *An Introduction to biostatistics using R*.
[2] B. Shahbaba, *Biostatistics with R*. New York, NY: Springer New York, 2012. doi: 10.1007/978-1-4614-1302-8.