INDESTAP: APRENDIENDO DE LOS DATOS. UN PROYECTO DE INNOVACIÓN DOCENTE EN ESTADÍSTICA APLICADA BASADO EN PROYECTOS DE INVESTIGACIÓN.

ÍNDICE:

1. Introducción al proyecto "Análisis del efecto de la dieta en el control de la diabetes mellitus tipo 2"

Se sabe que los carbohidratos son el mayor determinante de los niveles de glucosa postprandial, es decir, de los niveles de glucosa en sangre después de las comidas. Algunos estudios prueban que las dietas bajas en carbohidratos mejoran el control glicémico. De hecho, se sabe que las dietas cetogénicas (basadas en alimentos ricos en proteínas o grasas y, por tanto, bajas en carbohidratos) ayudan al control glicémico en pacientes con diabetes. Además, se intuye que pueden tener mejor efecto que las dietas con bajo índice glicémico, es decir, dietas basadas en alimentos con índice glucémico bajo que permiten mantener niveles de insulina bajos.

Recientemente, E.C. Westman y otros autores han estudiado el efecto de estos dos tipos de dietas: *low-carbohydrate, ketogenic diet* (*LCKD*, dieta cetogénica baja en hidratos de carbono) y *low-glycemic index diet* (*LGID*, dieta con bajo índice glicémico) en pacientes obesos con diabetes mellitus tipo 2. Los resultados de su investigación están publicados en:

> E.C. Westman, W.S. Yancy, J.C. Mavropoulos, M. Marquart and J.R. McDuffie. The effect of a low-carbohydrate, ketogenic diet versus a low-glycemic index diet on glycemic control in type 2 diabetes mellitus, *Nutrition & Metabolism*, 5:36, 2008.

Basándonos en este artículo, hemos simulado datos ficticios para 50 pacientes de forma que los estadísticos descriptivos de las características de interés coincidan con los observados en el estudio de E.C. Westman *et al* (2008). Los datos los podemos encontrar en el fichero *DietasDiabetes.xls.* En concreto, para cada uno de los pacientes se dispone de la siguiente información:

- *Program*. Dieta seguida {*LCKD*, *LGID*}
- *Age*. Edad
- *Gender*. Sexo
- *BMI*. Índice de masa corporal (Kg/m^2); dos medidas, al principio y al final del estudio.
- *Hemoglobin*. Nivel de hemoglobina A1c (%); dos medidas.
- *FastGluco*. Nivel de glucosa en sangre en ayunas (mg/dL); dos medidas.
- *FastInsulin*. Nivel de insulina en ayunas (μU/mL); dos medidas.
- *HDLCholesterol*. Nivel de colesterol HDL (mg/dL); dos medidas.
- *VLDLCholesterol*. Nivel de colesterol VLDL (mg/dL); dos medidas.
- *Triglycerides*. Nivel de triglicéridos (mg/dL); dos medidas.
- *Change_med*. Indica si ha habido una reducción o eliminación de la medicación tras el periodo de dieta
- *Adverse_effect*. Indica si el paciente ha experimentado algún efecto secundario durante el seguimiento de la dieta

En este proyecto, con el propósito de estudiar el efecto de los dos tipos de dietas descritas en pacientes obesos con diabetes mellitus tipo 2, analizaremos y compararemos las diferencias observadas en cada uno de los grupos entre las medidas basales (tomadas al inicio del estudio y representadas por *w0*) y las observadas después de seguir durante 24 semanas la dieta correspondiente (representadas por *w24*).

2. Análisis estadístico de algunas variables con R

En esta sección mostramos el análisis estadístico de algunas de las variables de interés mediante el software estadístico *R* y su interfaz gráfica *R-Commander* (ver Anexo).

2.1. Análisis exploratorio de datos

El primer paso de todo análisis estadístico es el análisis exploratorio de los datos que constituyen la muestra. Así pues, en esta primera sección veremos cómo calcular los principales estadísticos descriptivos y construir e interpretar distintos gráficos según el tipo de variable.

2.1.1. Descripción gráfica y numérica de una variable categórica

Las variables categóricas sólo pueden ser descritas numéricamente mediante las tablas de frecuencias, que indican el número (o porcentaje) de veces que se observa cada categoría en la muestra.

Por ejemplo, para construir la tabla de frecuencias correspondiente a la variable *Gender* seleccionamos el menú

Estadísticos / Resúmenes / Distribución de frecuencias

```
> .Table  # counts for Gender

Hombre  Mujer
    13     37

> round(100*.Table/sum(.Table), 2)  # percentages for Gender

Hombre  Mujer
    26     74
```

Tabla 1

La Tabla 1 nos indica que de los 50 pacientes estudiados, 13 son hombres (26%) y 37 son mujeres (74%).

Podemos representar gráficamente estos resultados mediante el diagrama de sectores (Figura 1) o el diagrama de barras (Figura 2), utilizando las opciones del menú *Gráficas*

Gráficas / Gráfica de sectores

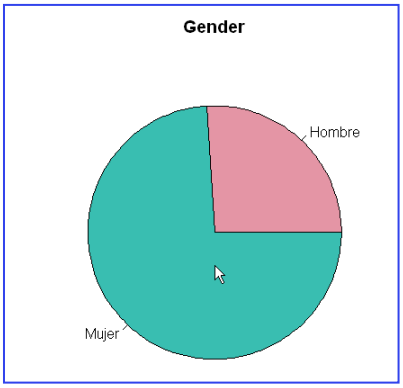


Figura 1

Gráficas / Gráfica de barras

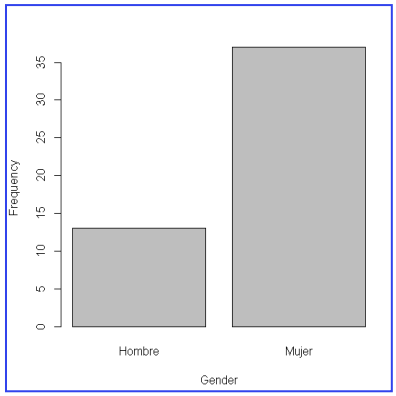


Figura 2

Si queremos cambiar el color o el título de los ejes ejecutaremos directamente la instrucción *pie* o *barplot*, respectivamente, cambiando las etiquetas de los ejes X e Y (xlab = "nombre_eje_X", ylab = "nombre_eje_Y") y añadiendo la opción col. Por ejemplo, podemos modificar la Figura 2 mediante la instrucción:

barplot(table(DietasDiabetes$Gender), xlab="Sexo", ylab="Frecuencia", col='BLUE')

2.1.2. Descripción gráfica y numérica de una variable cuantitativa

En el caso de variables cuantitativas, podemos completar la distribución de frecuencias con los estadísticos descriptivos. Los principales estadísticos descriptivos se encuentran en

Estadísticos / Resúmenes / Resúmenes numéricos

La salida proporcionada por *R-Commander* cuando pedimos los estadísticos descriptivos para la variable índice de masa corporal al comienzo del estudio *(BMI_w0)* es

```
> numSummary(DietasDiabetes[,"BMI_w0"], statistics=c("mean", "sd", "quantiles"),
quantiles=c(0,.25,.5,.75,1))
   mean       sd   0%   25%    50%    75% 100%  n
 37.604 4.451506 30.5 34.8 36.95 40.725 50.7 50
```

Tabla 2

En concreto, *R-Commander* nos proporciona los siguientes estadísticos descriptivos:

- Media (mean): media aritmética de los datos de la muestra.
- *Desviación típica (sd)*: medida de dispersión que nos informa cómo de alejados se encuentran los datos de la media muestral.
- *Mínimo y máximo* (*0% y 100%*): valores mínimo y máximo de la muestra.
- *Cuartiles Q1, Q2 y Q3* (*25%*, *50% y 75%*): valores que dividen la muestra ordenada en cuatro partes de igual tamaño en número de observaciones.

En el ejemplo de la variable *BMI_w0* podemos observar que:

- La media aritmética de los 50 índices de masa corporal al comienzo del estudio es 37.60 Kg/m^2 y
- su desviación típica es 4.45 Kg/m^2.
- El menor índice de masa corporal observado es igual a 30.50 Kg/m^2, mientras que el valor máximo observado es 50.70 Kg/m^2.
- El 50% de los datos (25) se encuentran por debajo del valor del segundo cuartil (mediana) 36.95 Kg/m^2, mientras que el 50% restante toman valores superiores a 36.95 Kg/m^2. Además, el 25% de los valores se encuentran por debajo de 34.80 Kg/m^2 y, de manera similar, un 25% de los datos toman valores superiores a 40.73 Kg/m^2.

Habitualmente es más útil obtener estos estadísticos separando los datos en grupos definidos por alguna variable categórica. Por ejemplo, si queremos obtener el resumen numérico de la variable *BMI_w0* en cada uno de los dos grupos que quedan definidos por el tipo de dieta seguida por los pacientes, basta con seleccionar la opción *Resumir por grupos* en la ventana de

Resúmenes numéricos y seleccionar como variable de agrupación la variable *Program,* con lo que se obtiene

```
> numSummary(DietasDiabetes[,"BMI_w0"], groups=DietasDiabetes$Program,
statistics=c("mean","sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
        mean       sd   0%  25%  50%  75% 100% data:n
LCKD 37.4619 3.581267 32.1 35.1 37.8 40.5 43.2     21
LGID 37.7069 5.048828 30.5 33.5 36.7 40.9 50.7     29
```

Tabla 3

Los gráficos más adecuados para representar una variable cuantitativa discreta son los diagramas de barras. En el caso de variables cuantitativas continuas utilizaremos el histograma (representación gráfica de la distribución de frecuencias agrupadas) o el diagrama de cajas (representación gráfica de la información obtenida en el resumen numérico: mínimo, máximo y cuartiles).

A continuación se muestra el histograma (Figura 3) y el diagrama de cajas (Figura 4) de la variable *BMI_w0.* Como podemos ver en el diagrama de cajas, el valor máximo observado del índice de masa corporal al inicio del estudio (50.7) es un valor extremo o outlier, que está caracterizado por el hecho de que dista del tercer cuartil más de 1.5 (Q3 - Q1) Kg/m^2; esto es, vez y media la longitud de la caja.

Gráficas / Histograma

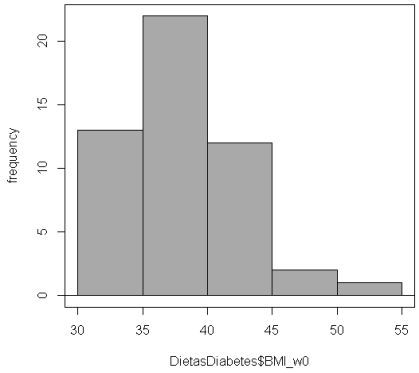


Figura 3

Gráficas / Diagrama de caja

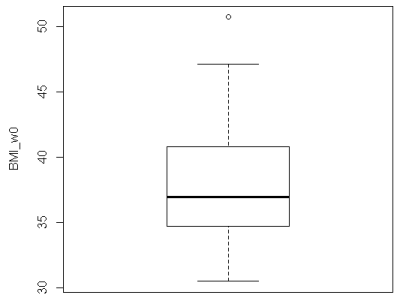


Figura 4

De nuevo podemos cambiar el aspecto de las gráficas añadiendo, en la instrucción generada por *R-Commander*, las opciones *col*, *xlab* e *ylab* (como hemos hecho con los diagramas de barras). En el histograma podemos, además, personalizar el número de intervalos cambiando la opción *breaks = "Sturges"* por *breaks = seq(extremo_inf, extremo_sup, amplitud)*.

Los diagramas de cajas nos permiten además hacer comparaciones entre grupos definidos por una variable categórica. Para ello debemos seleccionar la opción Gráfica por grupos e indicar el nombre de la variable de agrupación. Por ejemplo, para generar la Figura 5 hemos separado los valores observados de la variable *BMI_w0* según las distintas dietas seguidas por los pacientes.

Como hemos visto en la Tabla 3, no existen grandes diferencias entre los dos grupos respecto de la variable *BMI_w0* (índice de masa corporal al comienzo de la dieta). La diferencia más significativa se corresponde al valor máximo

observado en cada grupo. El valor máximo del grupo *LGID* (50.7) es considerablemente mayor que el del grupo *LCKD* (43.2), lo que implica una mayor dispersión de los datos de la muestra.

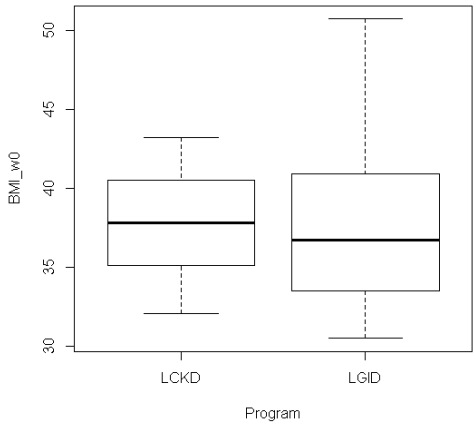


Figura 5

2.2. Descripción de la relación entre dos variables cuantitativas

En ocasiones es interesante analizar la relación entre dos variables cuantitativas. En esta sección veremos cómo

- determinar si dos variables cuantitativas están correlacionadas (valores mayores de una variable están asociados a valores mayores de la otra o al revés).
- medir la fuerza de la asociación lineal.
- predecir el valor de una variable a partir de un valor dado de la otra.

Como ejemplo ilustrativo analizaremos la relación entre el nivel de hemoglobina A1c antes y después de seguir durante 24 semanas alguna de las dos dietas propuestas (variables *Hemoglobin_w0* y *Hemoglobin_w24).* La Figura 6 muestra el diagrama de dispersión de los datos, que nos permite confirmar visualmente la existencia de una relación creciente entre las dos variables.

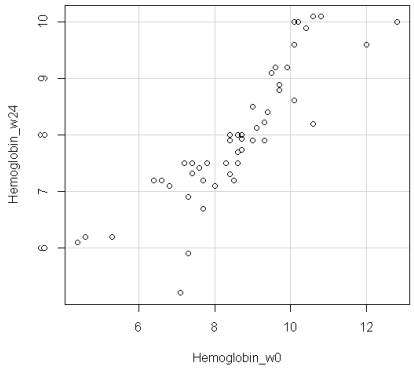
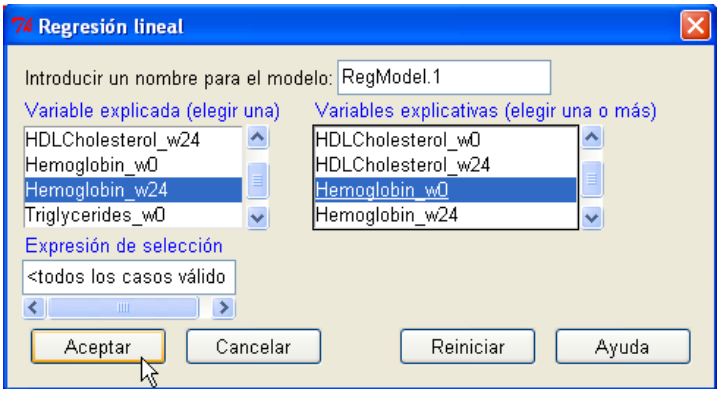


Figura 6

El menú de *R-Commander* que nos permite medir la fuerza de la asociación lineal y calcular la recta de regresión es

Estadísticos / Ajuste de modelos / Regresión lineal



Una vez seleccionadas la variable explicada (variable dependiente que irá en el eje vertical) y la variable explicativa (variable independiente, en el eje horizontal), *R-Commander* nos proporciona la siguiente información

```
> summary(RegModel.1)

Call:
lm(formula = Hemoglobin_w24 ~ Hemoglobin_w0, data = DietasDiabetes)

Residuals:
     Min       1Q   Median       3Q      Max
-1.85586 -0.29386  0.02207  0.41269  1.09974

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.69080    0.45091   5.968 2.81e-07 ***
Hemoglobin_w0 0.61480    0.05129  11.987 4.85e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5977 on 48 degrees of freedom
Multiple R-squared: 0.7496,  Adjusted R-squared: 0.7444
F-statistic: 143.7 on 1 and 48 DF,  p-value: 4.849e-16
```

Tabla 4

El coeficiente de correlación lineal de Pearson, r, lo podemos calcular a partir del coeficiente de determinación: $r = \sqrt{0.7496} = 0.87$. El signo de r es positivo si la recta es creciente y negativo en otro caso; en este caso es positivo. Como *r* toma un valor cercano a 1, podemos concluir que existe una relación lineal creciente fuerte entre las dos variables.

La recta de regresión o de mínimos cuadrados es $y = 2.69 + 0.61\,x$, donde y = *Hemoglobin_w24* y x = *Hemoglobin_w0.* Una vez calculada la recta de regresión podemos estimar el valor de y para un valor dado de x. Por ejemplo, el valor estimado del nivel de hemoglobina A1c después de seguir durante 24 semanas una de las dos dietas propuestas para un paciente con un nivel de hemoglobina inicial de 10 es: 2.69 + 0.61 * 10 = 8.79%.

2.3. Análisis estadístico de una variable continua: Inferencia sobre una media poblacional

Una vez realizado el análisis exploratorio de los datos, nos planteamos utilizar la información proporcionada por la muestra para extraer conclusiones que afectan a todos los individuos de la población. Es importante tener en cuenta

que no tenemos información completa (no hemos observado a toda la población) y, por tanto, existe incertidumbre. La estadística nos proporciona herramientas para trabajar en ambiente de incertidumbre. En concreto, nos permite estimar características de interés de la población y valorar el error que podemos cometer al extraer conclusiones.

En esta sección nos centraremos en la estimación de la media poblacional de una variable cuantitativa continua (estimación puntual e intervalos de confianza) y resolución de contrastes de hipótesis para la media.

A modo ilustrativo daremos respuesta a cuatro cuestiones que se plantean a continuación.

> ***2.3.a*** *Para cada uno de los grupos según la dieta seguida, calcula estimaciones puntuales para el nivel medio de hemoglobina A1c antes y después de seguir la dieta correspondiente durante 24 semanas. ¿Serían válidas las estimaciones por intervalos? Justifica tu respuesta.*

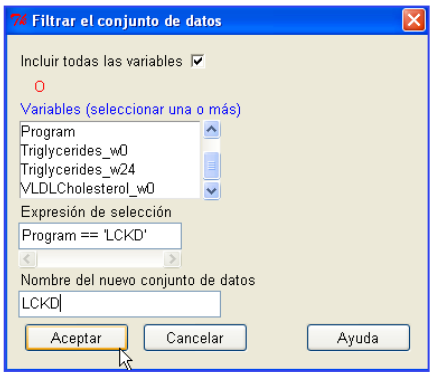
El resumen numérico por grupos de las variables *Hemoglobin_w0* y *Hemoglobin_w24* nos permite obtener estimaciones puntuales de los niveles medios de hemoglobina antes y después de seguir durante 24 semanas cada una de las dos dietas (ver Tabla 5).

```
> numSummary(DietasDiabetes[,c("Hemoglobin_w0", "Hemoglobin_w24")],
+   groups=DietasDiabetes$Program, statistics=c("mean", "sd", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))

Variable: Hemoglobin_w0
         mean       sd  0% 25% 50%  75% 100%  n
LCKD 9.214286 1.074842 7.1 8.6 9.5 10.1 10.8 21
LGID 8.217241 1.895685 4.4 7.3 8.4  9.1 12.8 29

Variable: Hemoglobin_w24
         mean       sd  0% 25%      50% 75% 100%  n
LCKD 8.440686 1.4213408 5.2 7.5 8.800000 9.6 10.1 21
LGID 7.681216 0.8673189 6.1 7.2 7.508421 8.0 10.0 29
```

Tabla 5

Para saber si es apropiado el uso de métodos paramétricos para el cálculo de intervalos de confianza, y dado que tenemos muestras de tamaño pequeño, debemos contrastar la Normalidad de las poblaciones de las cuales proceden las muestras.

Para ello crearemos, en primer lugar, un nuevo conjunto de datos (que llamaremos *LCKD*) que contenga únicamente la información correspondiente a los pacientes que siguieron la dieta *LCKD*. Para crear el nuevo conjunto de datos basta con *Filtrar el conjunto de datos* original eligiendo como expresión de selección *Program == 'LCKD'*.



A continuación contrastaremos la Normalidad de cada una de las variables mediante el test de Shapiro-Wilk (H_0: Normalidad)

Estadísticos / Resúmenes / Test de normalidad de Shapiro-Wilk

```
> shapiro.test(LCKD$Hemoglobin_w0)
        Shapiro-Wilk normality test

data:  LCKD$Hemoglobin_w0
W = 0.951, p-value = 0.3552

> shapiro.test(LCKD$Hemoglobin_w24)
        Shapiro-Wilk normality test

data:  LCKD$Hemoglobin_w24
W = 0.9274, p-value = 0.1222
```

Tabla 6

Para un nivel de significación $\alpha = 0.05$ no rechazamos la hipótesis nula, luego podemos actuar como si las dos variables se distribuyesen según una Normal: es apropiado el uso de métodos paramétricos. En concreto, es apropiado el cálculo de intervalos de confianza para el nivel medio de hemoglobina en los pacientes que han seguido la dieta *LCKD*.

El análisis de las variables *Hemoglobin_w0* y *Hemoglobin_w24* en los pacientes que siguen la dieta *LGID* se llevaría a cabo de la misma manera. Una vez creado el nuevo conjunto de datos conteniendo únicamente la información de los pacientes asociados a la dieta *LGID* (con nombre *LGID*), contrastamos la normalidad de los datos

```
> shapiro.test(LGID$Hemoglobin_w0)
        Shapiro-Wilk normality test

data:  LGID$Hemoglobin_w0
W = 0.9718, p-value = 0.6092

> shapiro.test(LGID$Hemoglobin_w24)
        Shapiro-Wilk normality test

data:  LGID$Hemoglobin_w24
W = 0.9396, p-value = 0.09764
```

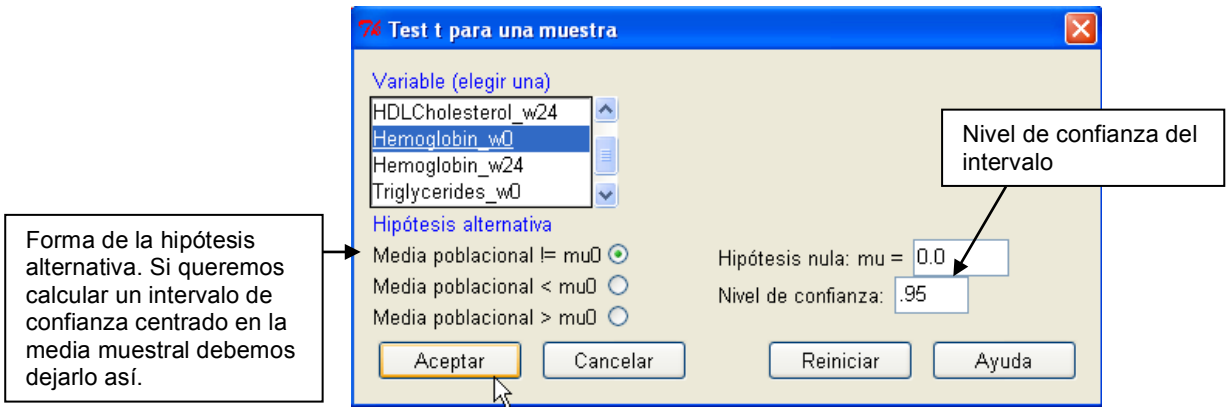
Tabla 7

Ambos resultados son no significativos. De nuevo podemos utilizar métodos paramétricos en el estudio de las dos variables *Hemoglobin_w0* y *Hemoglobin_w24*. Por tanto, las estimaciones por intervalos de los niveles medios de hemoglobina en el grupo *LGID* también serían válidas.

***2.3.b** Calcula e interpreta intervalos de confianza al 95% para el nivel medio de hemoglobina A1c en aquellas situaciones en las que sea apropiado según 2.3.a.*

Comenzaremos por la variable *Hemoglobin_w0* en el grupo *LCKD*. El menú de *R-Commander* que nos permite calcular intervalos de confianza es

Estadísticos / Medias / Test t para una muestra



La respuesta generada en la Ventana de resultados es:

```
> t.test(LCKD$Hemoglobin_w0, alternative='two.sided', mu=0.0, conf.level=.95)

        One Sample t-test

data:  LCKD$Hemoglobin_w0
t = 39.285, df = 20, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 8.725023 9.703548
sample estimates:
mean of x
 9.214286
```

Tabla 8

Así pues, podemos decir que la media poblacional del nivel de hemoglobina A1c al comienzo del estudio en el grupo *LCKD* está comprendida, con una confianza del 95%, entre el 8.73 y el 9.70%. Como podemos observar, el intervalo de confianza está centrado en el estimador puntual (media muestral): 9.21%.

Para la variable *Hemoglobin_w24* obtenemos el siguiente resultado

```
> t.test(LCKD$Hemoglobin_w24, alternative='two.sided', mu=0.0, conf.level=.95)

        One Sample t-test

data:  LCKD$Hemoglobin_w24
t = 27.2138, df = 20, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7.793699 9.087672
sample estimates:
mean of x
 8.440686
```

Tabla 9

El nivel medio de hemoglobina A1c después de 24 semanas de seguir la dieta *LCKD* está comprendido, con una confianza del 95%, entre el 7.79 y el 9.09%.

Análogamente, los resultados obtenidos para las variables *Hemoglobin_w0* y *Hemoglobin_w24* en los pacientes que siguen la dieta *LGID* son

```
> t.test(LGID$Hemoglobin_w0, alternative='two.sided', mu=0.0, conf.level=.95)

        One Sample t-test

data:  LGID$Hemoglobin_w0
t = 23.3431, df = 28, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7.496161 8.938321
sample estimates:
mean of x
 8.217241
```

Tabla 10

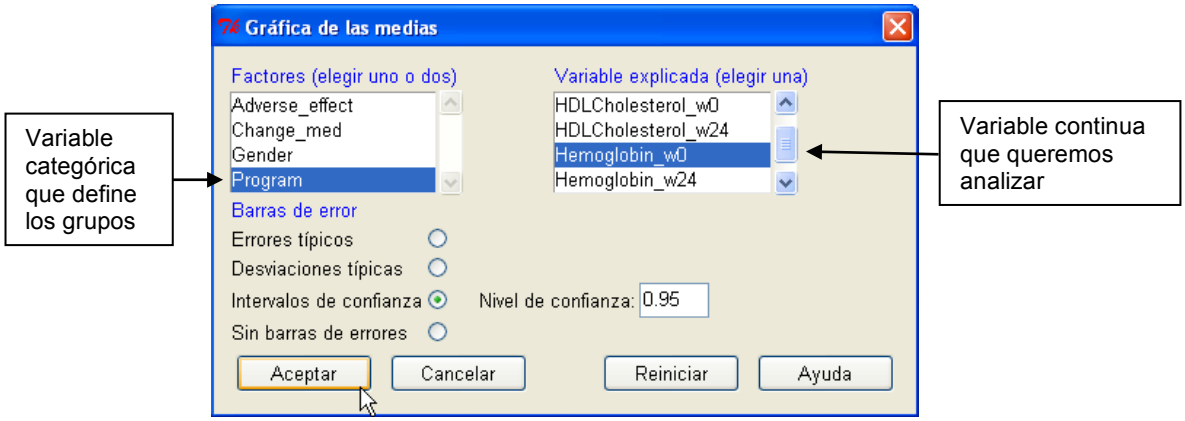
```
> t.test(LGID$Hemoglobin_w24, alternative='two.sided', mu=0.0, conf.level=.95)

        One Sample t-test

data:  LGID$Hemoglobin_w24
t = 47.6925, df = 28, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7.351306 8.011126
sample estimates:
mean of x
 7.681216
```

Tabla 11

A partir de estos resultados concluimos que el nivel medio de hemoglobina A1c al comienzo del estudio está comprendido, con una confianza del 95%, entre el 7.50 y el 8.94%. Después de seguir durante 24 semanas la dieta *LGID,* el nivel medio de hemoglobina A1c se encuentra entre el 7.35 y el 8.01%, con una confianza del 95%.

Una representación gráfica de los intervalos de confianza obtenidos en cada uno de los grupos nos puede ayudar a ver posibles diferencias entre las dietas. Para obtener dicha representación debemos volver al conjunto de datos original (*DietasDiabetes*) y seleccionar el menú

Gráficas / Gráfica de las medias



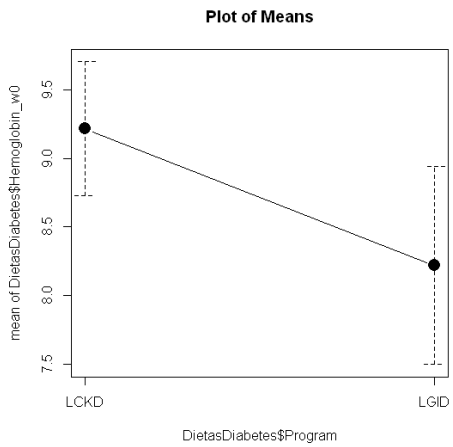


Figura 7

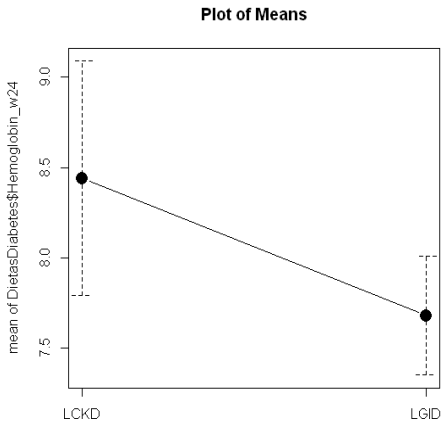


Figura 8

A partir de la Figura 7 podríamos pensar que el nivel medio de hemoglobina A1c al comienzo del estudio es mayor en el grupo *LCKD*. Lo mismo ocurre con los niveles medios de hemoglobina después de seguir durante 24 semanas la dieta *LCKD* o la dieta *LGID* (Figura 8). La pregunta en estos casos es: ¿la diferencia observada entre las muestras es debida al azar o a que vienen de poblaciones diferentes? Más adelante veremos técnicas estadísticas que nos permitirán contrastar si realmente existen diferencias significativas entre ambas dietas con respecto a los diferentes parámetros de interés.

> ***2.3.c*** *Se sabe que cuanto mayor es el nivel de hemoglobina A1c, mayor es el riesgo para el paciente de desarrollar complicaciones de la diabetes. De hecho, mantener un nivel de hemoglobina por debajo del 7% reduce significativamente la posibilidad de desarrollar complicaciones crónicas de la diabetes. ¿Podemos afirmar que el seguimiento de la dieta LCKD permite mantener un nivel de hemoglobina aceptable (es decir, por debajo del 7%)?*

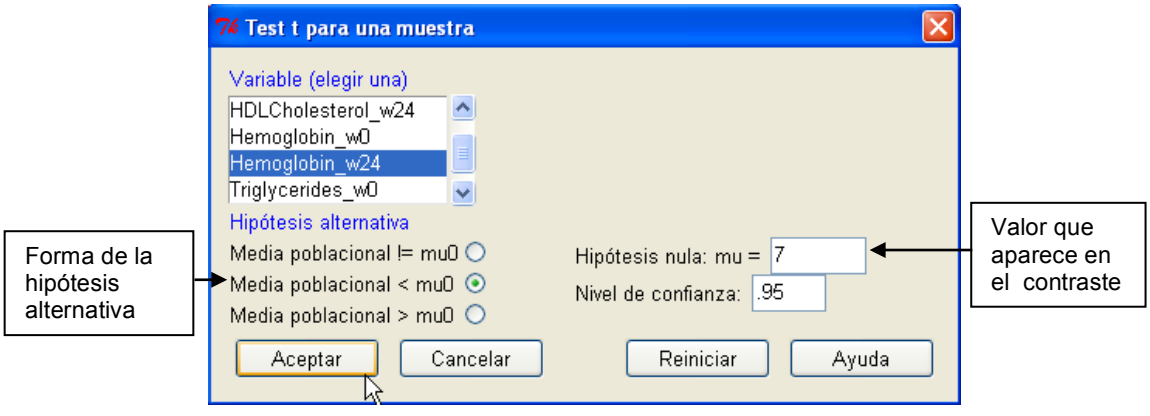
Supongamos que $X_{hemo\_w24\_LCKD}$ es la variable aleatoria que describe el nivel de hemoglobina A1c después de seguir durante 24 semanas la dieta *LCKD*. Como hemos visto anteriormente (Tabla 6), esta variable del conjunto de datos *LCKD* supera el test de normalidad. Sea $\mu_{hemo\_w24\_LCKD}$ su media poblacional.

Para dar respuesta a la pregunta anterior nos planteamos el siguiente contraste de hipótesis:

> H_0: La media poblacional de la variable $X_{hemo\_w24\_LCKD}$ es igual a 7 ($\mu_{hemo\_w24\_LCKD} = 7$)
>
> H_A: La media poblacional de la variable $X_{hemo\_w24\_LCKD}$ es inferior a 7 ($\mu_{hemo\_w24\_LCKD} < 7$)

El menú de *R-Commander* que nos permite resolver un contraste de hipótesis para la media poblacional de una variable aleatoria continua es, de nuevo,

Estadísticos / Medias / Test t para una muestra



```
> t.test(LCKD$Hemoglobin_w24, alternative='less', mu=7, conf.level=.95)

        One Sample t-test

data:  LCKD$Hemoglobin_w24
t = 4.6449, df = 20, p-value = 0.9999
alternative hypothesis: true mean is less than 7
95 percent confidence interval:
     -Inf 8.975628
sample estimates:
mean of x
 8.440686
```

Tabla 12

El p-valor del contraste es 0.9999, mayor que α = 0.05, por lo que no rechazamos H_0. El seguimiento de la dieta *LCKD* durante 24 semanas no garantiza un nivel de hemoglobina A1c por debajo del 7%.

2.3.d *La Organización Mundial de la Salud establece la obesidad como un índice de masa corporal superior a 30* Kg/m^2*. ¿Seguirían siendo obesos los pacientes que han seguido la dieta LCKD durante 24 semanas?*

En este caso debemos analizar la variable *BMI_w24* en el conjunto de datos *LCKD* (índice de masa corporal después de seguir durante 24 semanas la dieta *LCKD*), que representaremos por $X_{BMI\_w24\_LCKD}$.

Antes de llevar a cabo el análisis estadístico, debemos contrastar la normalidad de los datos.

```
> shapiro.test(LCKD$BMI_w24)

        Shapiro-Wilk normality test

data:  LCKD$BMI_w24
W = 0.9023, p-value = 0.03872
```

Tabla 13

Como el p-valor es menor que el nivel de significación α, rechazamos la hipótesis nula. Es decir, existe suficiente evidencia para afirmar que la variable $X_{BMI\_w24\_LCKD}$ no sigue una distribución Normal.

Los métodos inferenciales utilizados para el análisis de esta variable deben ser no paramétricos. En concreto, utilizaremos un test de Wilcoxon de rangos con signo para validar estadísticamente si la mediana poblacional queda o no por encima de 30 Kg/m^2

H_0: La mediana poblacional de $X_{BMI\_w24\_LCKD}$ es igual a 30 Kg/m^2
H_A: La mediana poblacional de $X_{BMI\_w24\_LCKD}$ es superior a 30 Kg/m^2

Para realizar este test debemos ejecutar directamente la instrucción

wilcox.test(LCKD$BMI_w24, alternative='greater', mu=30)

que nos proporciona la siguiente salida

```
> wilcox.test(LCKD$BMI_w24,alternative='greater',mu=30)

        Wilcoxon signed rank test with continuity correction

data:  LCKD$BMI_w24
V = 174, p-value = 0.02189
alternative hypothesis: true location is greater than 30
```

Tabla 14

El p-valor del test es menor que el nivel de significación α, por lo que rechazamos la hipótesis nula y concluimos que hay evidencia estadística de que la mediana poblacional del índice de masa corporal después de seguir durante 24 semanas la dieta *LCKD* es superior a 30 Kg/m^2 (obesidad).

2.4. Análisis estadístico de dos variables continuas: Comparación de dos medias poblacionales

En la sección anterior nos hemos centrado en el análisis de una única muestra de datos numéricos (es decir, en el análisis de una variable cuantitativa continua). Sin embargo, en la práctica, muchas investigaciones requieren comparar dos o más muestras. En esta sección nos centraremos en la comparación de dos medias poblacionales. En concreto, veremos cómo calcular intervalos de confianza para la diferencia de las medias poblacionales y resolver contrastes de hipótesis para la comparación de las medias. Como veremos a continuación, la elección del método de análisis en estas ocasiones dependerá no sólo de la normalidad de los datos (necesaria para aplicar métodos paramétricos) sino también de la forma en que los datos han sido obtenidos (muestras independientes vs muestras emparejadas).

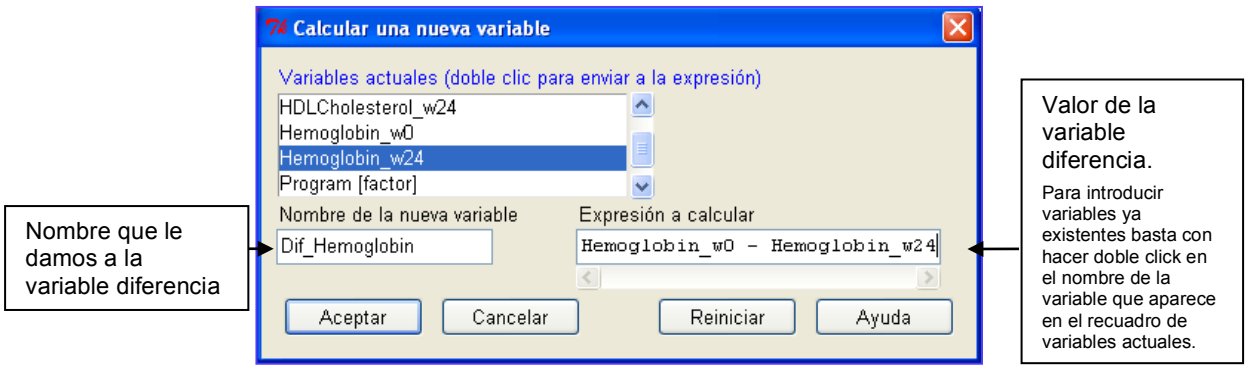
A modo ilustrativo daremos respuesta a las siguientes cuestiones:

2.4.a *¿Es efectiva la dieta LCKD para reducir el nivel de hemoglobina A1c? ¿Y la dieta LGID?*

Las dietas serán efectivas si el nivel medio de hemoglobina A1c al comienzo del estudio es mayor que el nivel medio de hemoglobina A1c después de seguir la dieta. Así pues, para contestar a la pregunta anterior debemos comparar, para cada una de las dietas, los niveles medios de hemoglobina antes y después de seguir la dieta.

Estamos ante un ejemplo de muestras emparejadas (se mide a la misma persona antes y después de seguir la dieta correspondiente) y, por tanto, trabajaremos con la variable diferencia, que mide el cambio ocurrido durante el periodo de dieta. Para calcular la variable diferencia seleccionamos el menú

Datos / Modificar variables del conjunto de datos activo / Calcular una nueva variable



Comencemos el análisis de la diferencia del nivel de hemoglobina (variable *Dif_Hemoglobin*) en el grupo *LCKD*. Para ello debemos filtrar el conjunto de datos original y crear un nuevo conjunto de datos (llamado *LCKD*) que contenga únicamente la información correspondiente a los pacientes que siguieron la dieta *LCKD*.

Aunque el conjunto de datos *LCKD* ya lo habíamos creado anteriormente, todavía no habíamos creado la variable diferencia. Si queremos trabajar con una variable nueva (en este caso *Dif_Hemoglobin*), debemos crearla primero en el conjunto de datos original (*DietasDiabetes*) y filtrar a continuación los datos, reemplazando el anterior conjunto de datos *LCKD* por el nuevo, que ya incluye a la nueva variable *Dif_Hemoglobin*.

Los estadísticos descriptivos para la variable *Dif_Hemoglobin* en el grupo *LCKD* son

```
> numSummary(LCKD[,"Dif_Hemoglobin"], statistics=c("mean", "sd", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
   mean        sd  0% 25% 50% 75% 100%  n
 0.7736 0.4131586 0.1 0.5 0.7   1  1.9 21
```

Tabla 15

Antes de realizar el análisis inferencial de la variable *Dif_Hemoglobin*, y dado que la muestra es pequeña (n = 21), debemos contrastar la Normalidad de los datos para saber si el uso de métodos paramétricos es adecuado o no

Estadísticos / Resúmenes / Test de normalidad de Shapiro-Wilk

```
> shapiro.test(LCKD$Dif_Hemoglobin)
        Shapiro-Wilk normality test

data:  LCKD$Dif_Hemoglobin
W = 0.9465, p-value = 0.2918
```

Tabla 16

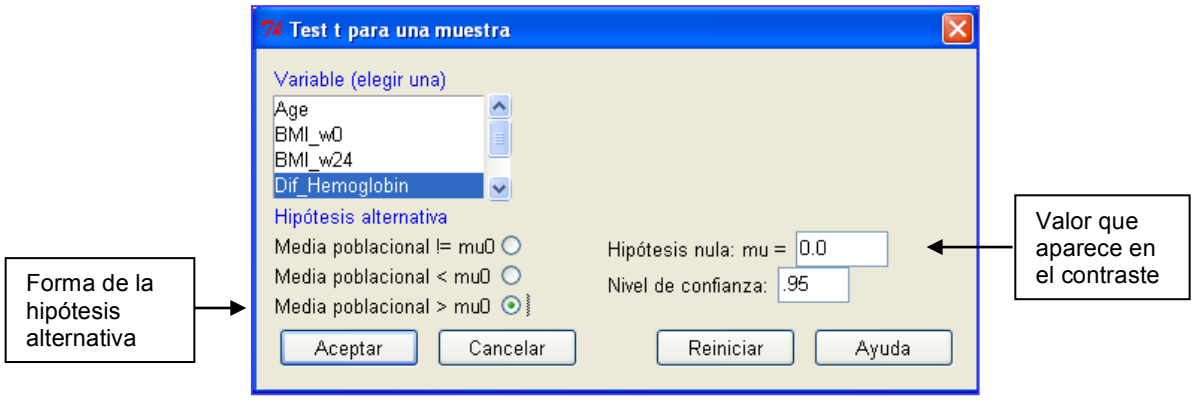
Para un nivel de significación $\alpha = 0.05$, no rechazamos la hipótesis nula (H_0: Normalidad); es decir, suponemos que la variable *Dif_Hemoglobin* en el grupo *LCKD* sigue un comportamiento Normal y, por tanto, el uso de métodos paramétricos es adecuado.

Sea $\mu_{Dif\_hemo\_LCKD} = \mu_{hemo\_w0\_LCKD} - \mu_{hemo\_w24\_LCKD}$ la media poblacional de la variable *Dif_Hemoglobin = Hemoglobin_w0 – Hemoglobin_w24* en el grupo *LCKD*. El contraste de hipótesis que nos permite conocer si el seguimiento de la dieta conlleva una reducción del nivel de hemoglobina es:

H_0: $\mu_{Dif\_hemo\_LCKD} = 0$
H_A: $\mu_{Dif\_hemo\_LCKD} > 0$

El menú de *R-Commander* que nos permite resolver este contraste es

Estadísticos / Medias / Test t para una muestra



La respuesta generada en la Ventana de resultados es:

```
> t.test(LCKD$Dif_Hemoglobin, alternative='greater', mu=0.0, conf.level=.95)

        One Sample t-test

data:  LCKD$Dif_Hemoglobin
t = 8.5804, df = 20, p-value = 1.943e-08
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.6181018       Inf
sample estimates:
mean of x
   0.7736
```

Tabla 17

El p-valor del test es menor que el nivel de significación α, por lo que rechazamos la hipótesis nula; es decir, existe suficiente evidencia para concluir que el nivel medio de hemoglobina A1c es mayor al comienzo del estudio que después de seguir durante 24 semanas la dieta *LCKD*.

El análisis de la variable *Dif_Hemoglobin* en el grupo *LGID* se lleva a cabo de la misma manera. Una vez creado el nuevo conjunto de datos conteniendo únicamente la información de los pacientes asociados a la dieta *LGID* (con nombre *LGID*), contrastamos la normalidad de la variable *Dif_Hemoglobin*

```
> shapiro.test(LGID$Dif_Hemoglobin)
        Shapiro-Wilk normality test

data:  LGID$Dif_Hemoglobin
W = 0.9807, p-value = 0.8559
```

Tabla 18

De nuevo, podemos suponer que la variable *Dif_Hemoglobin* sigue un comportamiento Normal en el grupo *LGID.*

El resumen numérico de la variable *Dif_hemoglobin* en el grupo *LGID* es

```
> numSummary(LGID[,"Dif_Hemoglobin"], statistics=c("mean", "sd", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
      mean       sd   0%  25% 50% 75% 100%  n
 0.5360254 1.088967 -1.7 -0.1 0.5 1.1  2.8 29
```

Tabla 19

Como podemos observar, la media muestral de la variable *Dif_Hemoglobin* es 0.54 > 0, lo que indica que los pacientes del estudio experimentaron una reducción en el nivel de hemoglobina. Veamos pues si este comportamiento observado en la muestra lo podemos generalizar a toda la población. El resultado obtenido para el contraste de hipótesis

H_0: $\mu_{Dif\_hemo\_LGID} = 0$
H_A: $\mu_{Dif\_hemo\_LGID} > 0$

es:

```
> t.test(LGID$Dif_Hemoglobin, alternative='greater', mu=0.0, conf.level=.95)

        One Sample t-test

data:  LGID$Dif_Hemoglobin
t = 2.6508, df = 28, p-value = 0.006532
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.1920292       Inf
sample estimates:
mean of x
0.5360254
```

Tabla 20

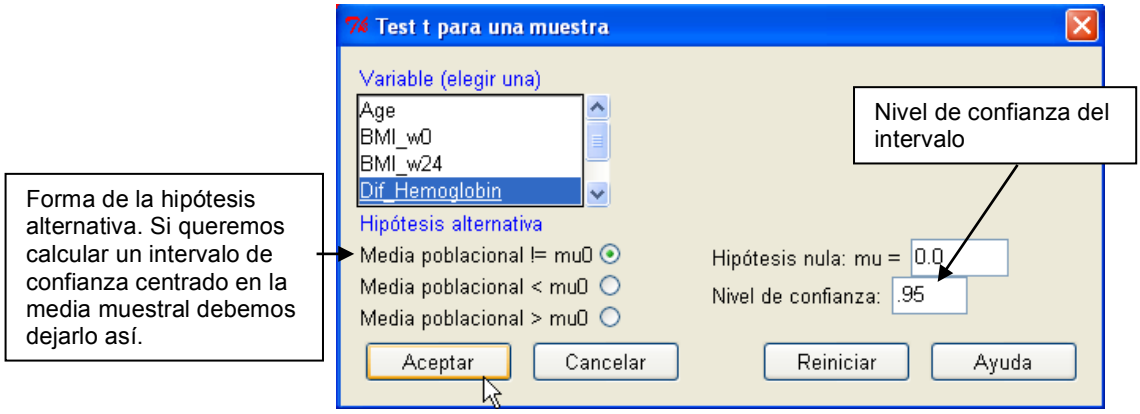
El p-valor del test es menor que el nivel de significación α, por lo que rechazamos la hipótesis nula: concluimos que la dieta *LGID* también es efectiva para reducir el nivel de hemoglobina A1c.

2.4.b *¿Sería correcto calcular un intervalo de confianza al 95% para la diferencia de los niveles medios de hemoglobina antes y después de seguir la dieta LCKD? En caso afirmativo, calcúlalo.*

En el apartado 2.4.a hemos visto que la variable *Dif_Hemoglobin* = *Hemoglobin_w0 – Hemoglobin_w24* sigue un comportamiento Normal en el grupo *LCKD* (Tabla 16), por lo que resulta apropiado el uso de métodos paramétricos para el cálculo de intervalos de confianza para su media poblacional $\mu_{Dif\_hemo\_LCKD} = \mu_{hemo\_w0\_LCKD} - \mu_{hemo\_w24\_LCKD}$.

El menú de *R-Commander* que nos permite calcular intervalos de confianza es, de nuevo,

Estadísticos / Medias / Test t para una muestra



```
> t.test(LCKD$Dif_Hemoglobin, alternative='two.sided', mu=0.0, conf.level=.95)

        One Sample t-test

data:  LCKD$Dif_Hemoglobin
t = 8.5804, df = 20, p-value = 3.885e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.5855325 0.9616675
sample estimates:
mean of x
   0.7736
```

Tabla 21

A partir de estos resultados concluimos que la diferencia entre el nivel medio de hemoglobina A1c al comienzo del estudio y el nivel medio de hemoglobina A1c después de seguir la dieta *LCKD* está comprendida, con una confianza del 95%, entre el 0.59 y el 0.96: $0.59 \leq \mu_{hemo\_w0\_LCKD} - \mu_{hemo\_w24\_LCKD} \leq 0.96$.

2.4.c *¿Podemos concluir que los pacientes que siguen la dieta LCKD experimentan un mayor descenso en el nivel de hemoglobina A1c que los que siguen la dieta LGID?*

En el apartado 2.4.a hemos comprobado que las dos dietas son efectivas para reducir el nivel de hemoglobina A1c ($\mu_{Dif\_hemo\_LCKD} > 0$ y $\mu_{Dif\_hemo\_LGID} > 0$), pero no hemos visto cuál de ellas es más efectiva.

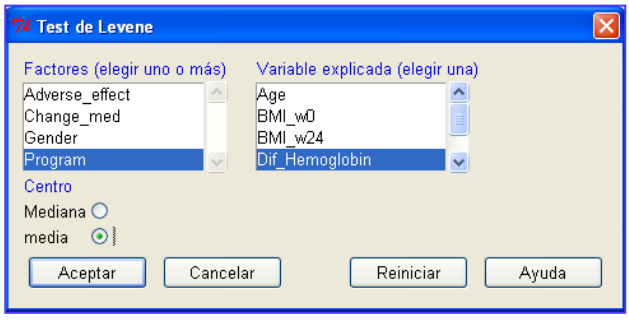
Para dar respuesta a la pregunta anterior nos planteamos el siguiente contraste de hipótesis:

H_0: $\mu_{Dif\_hemo\_LCKD} = \mu_{Dif\_hemo\_LGID}$
H_A: $\mu_{Dif\_hemo\_LCKD} > \mu_{Dif\_hemo\_LGID}$

Se trata pues de un contraste de hipótesis para la igualdad de medias de dos poblaciones independientes (los pacientes que siguen la dieta *LCKD* y los que siguen la dieta *LGID* no guardan ningún tipo de relación).

Como hemos visto anteriormente (Tablas 16 y 18), la variable *Dif_Hemoglobin* sigue un comportamiento Normal en los dos grupos (*LCKD* y *LGID*), por lo que el uso de métodos paramétricos es apropiado. A continuación debemos averiguar si las varianzas poblacionales son iguales o no, ya que el estadístico de contraste es diferente en cada caso. Para ello utilizamos la prueba de Levene (H_0: igualdad de varianzas)

Estadísticos / Varianzas / Test de Levene



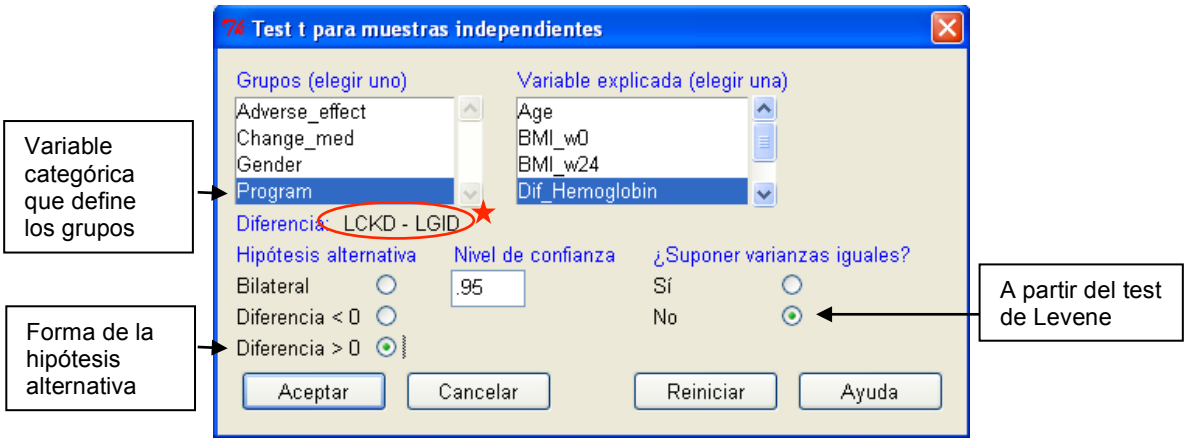
```
> leveneTest(DietasDiabetes$Dif_Hemoglobin, DietasDiabetes$Program, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value    Pr(>F)
group  1  10.352 0.002319 **
      48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabla 22

El p-valor del test es menor que $\alpha = 0.05$, por lo que rechazamos la hipótesis nula de igualdad de varianzas.

El menú de *R-Commander* que nos permite resolver un contraste de hipótesis para la igualdad de medias de dos poblaciones independientes es

Estadísticos / Medias / Test t para muestras independientes




```
> t.test(Dif_Hemoglobin~Program, alternative='greater', conf.level=.95,
+ var.equal=FALSE, data=DietasDiabetes)

        Welch Two Sample t-test

data:  Dif_Hemoglobin by Program
t = 1.073, df = 38.129, p-value = 0.145
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1356717        Inf
sample estimates:
mean in group LCKD mean in group LGID
         0.7736000          0.5360254
```

Tabla 23

El p-valor del contraste es 0.145 > $\alpha = 0.05$, por lo que no rechazamos la hipótesis nula; no existe suficiente evidencia para afirmar que la dieta *LCKD* conlleva un descenso mayor en el nivel de hemoglobina A1c.

★Es importante tener en cuenta que *R-Commander* ordena los grupos por orden alfabético, por lo que debemos comprobar qué grupo es el primero antes de elegir la forma de la hipótesis alternativa. En este caso el grupo *LCKD* es el primero y la hipótesis nula de mayor es correcta.

2.4.d *¿Existe una disminución significativa en los niveles de insulina en ayunas antes y después de seguir durante 24 semanas la dieta LCKD?*

Para contestar a esta pregunta debemos comparar, para los pacientes asociados a la dieta *LCKD*, el nivel de insulina en ayunas antes y después de seguir la dieta durante 24 semanas.

Antes de llevar a cabo el análisis estadístico, debemos averiguar si la variable diferencia (definida como *Dif_FastInsulin = FastInsulin_w0 - FastInsulin_w24*) sigue un comportamiento Normal en el grupo *LCKD*.

```
> shapiro.test(LCKD$Dif_FastInsulin)

        Shapiro-Wilk normality test

data:  LCKD$Dif_FastInsulin
W = 0.9066, p-value = 0.04701
```

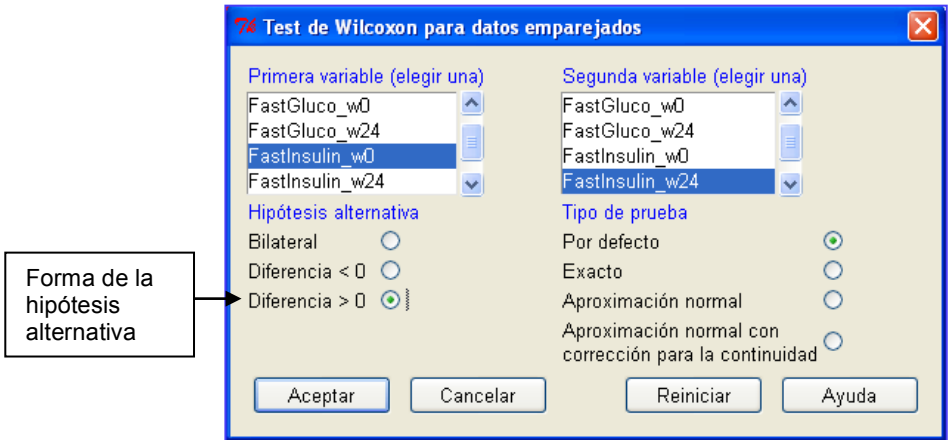
Tabla 24

Como el p-valor es menor que el nivel de significación $\alpha = 0.05$, rechazamos la hipótesis nula; es decir, concluimos que la variable *Dif_FastInsulin* no sigue un comportamiento Normal.

Así pues, utilizaremos un test no paramétrico para muestras emparejadas, el test de rangos con signo de Wilcoxon, que nos permite ver si existen diferencias significativas entre las dos muestras o no

H_0: El nivel de insulina en ayunas es el mismo en ambos casos
H_A: El nivel de insulina en ayunas al comienzo del estudio es mayor

Estadísticos / Test no paramétricos / Test de Wilcoxon para muestras pareadas



La salida proporcionada por *R-Commander* es:

```
> wilcox.test(LCKD$FastInsulin_w0, LCKD$FastInsulin_w24, alternative='greater',
+   paired=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  LCKD$FastInsulin_w0 and LCKD$FastInsulin_w24
V = 204, p-value = 0.001111
alternative hypothesis: true location shift is greater than 0
```

Tabla 25

Para un nivel de significación de 0.05, rechazamos H_0. Existe suficiente evidencia para afirmar que el nivel de insulina en ayunas antes del programa es mayor.

2.4.e *¿Podemos afirmar que los niveles de colesterol HDL después de seguir durante 24 semanas la dieta LCKD o la dieta LGID son iguales?*

Veamos, en primer lugar, si el uso de métodos paramétricos es apropiado. Como se trata de muestras independientes, debemos contrastar la Normalidad de la variable *HDLCholesterol_w24* en cada una de las poblaciones definidas por el tipo de dieta seguida mediante el test de Shapiro-Wilk

```
> shapiro.test(LCKD$HDLCholesterol_w24)

        Shapiro-Wilk normality test

data:  LCKD$HDLCholesterol_w24
W = 0.8909, p-value = 0.02336
```

Tabla 26

El p-valor es menor que $\alpha = 0.05$, rechazamos H_0 (Normalidad)

```
> shapiro.test(LGID$HDLCholesterol_w24)

        Shapiro-Wilk normality test

data:  LGID$HDLCholesterol_w24
W = 0.975, p-value = 0.7007
```

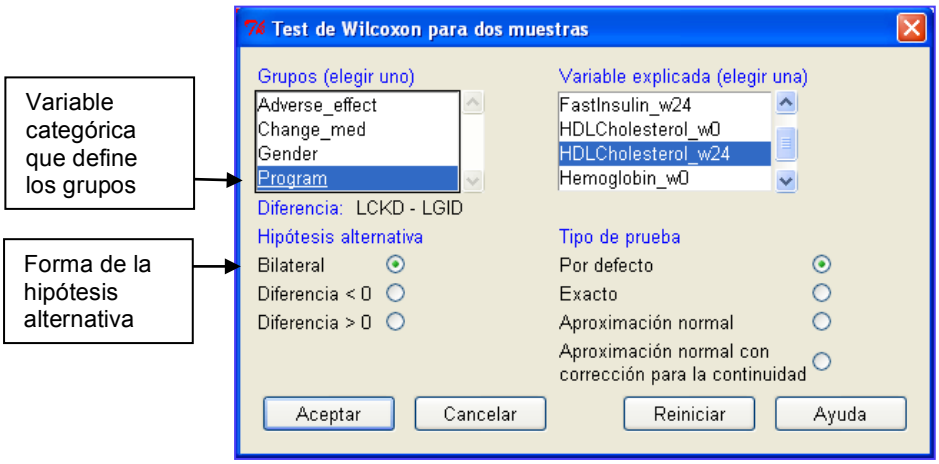
Tabla 27

El p-valor es mayor que $\alpha = 0.05$, no rechazamos H_0 (Normalidad)

Como no podemos asumir que las dos distribuciones son Normales, utilizaremos un test no paramétrico para su comparación, el test de Wilcoxon para muestras independientes

H_0: El nivel de colesterol HDL después de seguir la dieta *LCKD* o la dieta *LGID* durante 24 semanas es el mismo
H_A: El nivel de colesterol HDL después de seguir la dieta *LCKD* o la dieta *LGID* durante 24 semanas no es el mismo

Estadísticos / Test no paramétricos / Test de Wilcoxon para dos muestras



```
> wilcox.test(HDLCholesterol_w24 ~ Program, alternative="two.sided", data=DietasDiabetes)

        Wilcoxon rank sum test with continuity correction

data:  HDLCholesterol_w24 by Program
W = 431.5, p-value = 0.01289
alternative hypothesis: true location shift is not equal to 0
```

Tabla 28

Para un nivel de significación α = 0.05 rechazamos H_0. Hay evidencia de que los niveles de colesterol HDL después de seguir la dieta *LCKD* o *LGID* no son iguales.

2.5. Análisis estadístico de una variable continua en tres o más poblaciones

En esta sección extendemos el análisis desarrollado en las secciones anteriores para variables numéricas cuando se tienen tres o más muestras independientes. En concreto, veremos cómo resolver contrastes de hipótesis para la igualdad de las medias de una variable cuantitativa en diferentes poblaciones y cómo encontrar grupos homogéneos en aquellas situaciones en las que no podamos suponer que todas las poblaciones tienen la misma media.

2.5.a *En el apartado 2.4.c hemos visto que no existe suficiente evidencia para afirmar que la dieta LCKD conlleve un descenso mayor en el nivel de hemoglobina A1c. Es decir, el descenso medio en el nivel de hemoglobina A1c puede ser el mismo para las dos dietas ($\mu_{Dif\_hemo\_LCKD} = \mu_{Dif\_hemo\_LGID}$). Supongamos ahora que agrupamos a los pacientes en función de la dieta seguida y el sexo (factor que, en ocasiones, condiciona la respuesta a un fármaco, tratamiento, intervención dietética, etc.), formando así cuatro grupos (cuatro poblaciones): LCKD_Hombre, LCKD_Mujer, LGID_Hombre y LGID_Mujer. ¿Podemos concluir que el descenso medio en el nivel de hemoglobina A1c es el mismo en los cuatro grupos?*

Para dar respuesta a la pregunta anterior nos planteamos el siguiente contraste de hipótesis:

H_0: Las medias poblacionales son todas iguales ($\mu_{Dif\_hemo\_LCKD\_Hombre} = \mu_{Dif\_hemo\_LCKD\_Mujer} = \mu_{Dif\_hemo\_LGID\_Hombre} = \mu_{Dif\_hemo\_LGID\_Mujer}$)
H_A: Las medias poblacionales no son todas iguales

Como se trata de muestras independientes (los pacientes estudiados en cada grupo no guardan ningún tipo de relación entre sí), necesitamos una variable categórica que defina los cuatro grupos de interés. La variable *Program_Gender* indica para cada uno de los pacientes a que grupo pertenece en función de la dieta seguida *{LCKD, LGID}* y el sexo *{Hombre, Mujer}*.

Los estadísticos descriptivos para la variable *Dif_Hemoglobin* en los cuatro grupos definidos por la variable *Program_Gender* son

```
> numSummary(DietasDiabetes[,"Dif_Hemoglobin"], groups=DietasDiabetes$Program_Gender,
+   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean        sd   0%       25%      50%      75% 100% data:n
LCKD_Hombre 0.9000000 0.5830952  0.1  0.725000 0.850000 0.975000  1.9      6
LCKD_Mujer  0.7230400 0.3359203  0.2  0.500000 0.700000 0.977200  1.4     15
LGID_Hombre 0.9822556 1.0059953 -0.3  0.421842 0.968947 1.281579  2.8      7
LGID_Mujer  0.3940430 1.0974925 -1.7 -0.250000 0.450000 1.050000  2.4     22
```

Tabla 29

En la Tabla 29 podemos apreciar ciertas diferencias entre las medias muestrales. ¿Son esas diferencias suficientemente grandes como para inferir que las poblaciones de donde provienen las muestras son realmente poblaciones diferentes?

Antes de comenzar el análisis inferencial de la variable *Dif_Hemoglobin* y, dado que los tamaños muestrales son pequeños (n_1=6, n_2=15, n_3=7, n_4=22), debemos contrastar la Normalidad de la variable en las cuatro poblaciones. Para ello, podemos ejecutar directamente en *R-Commander* la instrucción

```
tapply(DietasDiabetes$Dif_Hemoglobin,DietasDiabetes$Program_Gender,
shapiro.test)
```

que aplica el test de Shapiro-Wilk a cada uno de los grupos definidos por la variable categórica *Program_Gender*. El resultado obtenido es

```
> tapply(DietasDiabetes$Dif_Hemoglobin,DietasDiabetes$Program_Gender,shapiro.test)
$LCKD_Hombre
        Shapiro-Wilk normality test

data:  X[[1L]]
W = 0.9197, p-value = 0.5031

$LCKD_Mujer
        Shapiro-Wilk normality test

data:  X[[2L]]
W = 0.9439, p-value = 0.4344

$LGID_Hombre
        Shapiro-Wilk normality test

data:  X[[3L]]
W = 0.9502, p-value = 0.7317

$LGID_Mujer
        Shapiro-Wilk normality test

data:  X[[4L]]
W = 0.9722, p-value = 0.7622
```

Tabla 30

Para un nivel de significación α = 0.05, no rechazamos la hipótesis de normalidad en ninguno de los cuatro grupos (p-valores asociados al test de Shapiro-Wilk mayores que α) y, por tanto, el uso de métodos paramétricos es adecuado.

A continuación debemos averiguar si las varianzas poblacionales son iguales o no, pues en el caso de varianzas iguales utilizaremos el test ANOVA para la comparación de las medias y, en caso contrario, el test de Welch. El resultado de la prueba de Levene (H_0: igualdad de varianzas) es

```
> leveneTest(DietasDiabetes$Dif_Hemoglobin, DietasDiabetes$Program_Gender, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  3  3.2116 0.03149 *
      46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabla 31

El p-valor del test es menor que $\alpha = 0.05$, por lo que rechazamos la hipótesis nula de igualdad de varianzas. Así pues, el método paramétrico utilizado para comparar las cuatro medias poblacionales es el test de Welch. Este test no aparece en el menú de *R-Commander*, por lo que ejecutaremos directamente la instrucción

```
oneway.test(Dif_Hemoglobin~Program_Gender, data=DietasDiabetes,
var.equal=FALSE)
```

La respuesta generada en la ventana de resultados es

```
> oneway.test(Dif_Hemoglobin~Program_Gender, data=DietasDiabetes, var.equal=FALSE)

        One-way analysis of means (not assuming equal variances)

data:  Dif_Hemoglobin and Program_Gender
F = 0.909, num df = 3.000, denom df = 13.961, p-value = 0.4616
```

Tabla 32

El p-valor del contraste es 0.4616 > 0.05, por lo que no rechazamos la hipótesis nula de igualdad de medias poblacionales; es decir, no encontramos diferencias significativa en los cuatro grupos en el descenso medio en el nivel de hemoglobina.

La instrucción

```
pairwise.t.test(DietasDiabetes$Dif_Hemoglobin,
DietasDiabetes$Program_Gender, pool.sd=FALSE)
```

nos permite detectar los grupos que pudieran ser distintos. En este caso, tal y como esperábamos, no encontramos ninguna diferencia: los p-valores asociados a cada una de las comparaciones dos a dos entre grupos son grandes.

```
> pairwise.t.test(DietasDiabetes$Dif_Hemoglobin,DietasDiabetes$Program_Gender, pool.sd=FALSE)

        Pairwise comparisons using t tests with non-pooled SD

data:  DietasDiabetes$Dif_Hemoglobin and DietasDiabetes$Program_Gender

            LCKD_Hombre LCKD_Mujer LGID_Hombre
LCKD_Mujer  1.00        -          -
LGID_Hombre 1.00        1.00       -
LGID_Mujer  0.90        0.99       0.99

P value adjustment method: holm
```

Tabla 33

2.5.b *¿Podemos afirmar que el nivel de insulina en ayunas al finalizar el estudio (semana 24) es el mismo en los cuatro grupos definidos por la variable Program_Gender? En caso contrario, define los grupos homogéneos.*

Los estadísticos descriptivos de la variable *FastInsulin_w24* en los cuatro grupos definidos por la variable *Program_Gender* son

```
> numSummary(DietasDiabetes[,"FastInsulin_w24"], groups=DietasDiabetes$Program_Gender,
+   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean       sd  0%   25%  50%    75% 100% data:n
LCKD_Hombre 10.883333 4.765886 5.2  7.50 10.3 14.675 16.8      6
LCKD_Mujer  14.940000 3.804096 7.6 13.65 15.4 18.550 18.8     15
LGID_Hombre  9.685714 2.336970 6.7  7.70 10.1 11.550 12.5      7
LGID_Mujer  12.359091 3.568067 6.3  9.65 13.4 14.900 17.5     22
```

Tabla 34

De nuevo podemos apreciar ciertas diferencias entre las medias muestrales de los cuatro grupos. Veamos a continuación si las diferencias observadas son significativas o no.

Mediante el test de Shapiro-Wilk (Tabla 35) podemos observar que la variable *FastInsulin_w24* no sigue un comportamiento Normal en las cuatro poblaciones (p-valor correspondiente al grupo *LCKD_Mujer* < $\alpha = 0.05$).

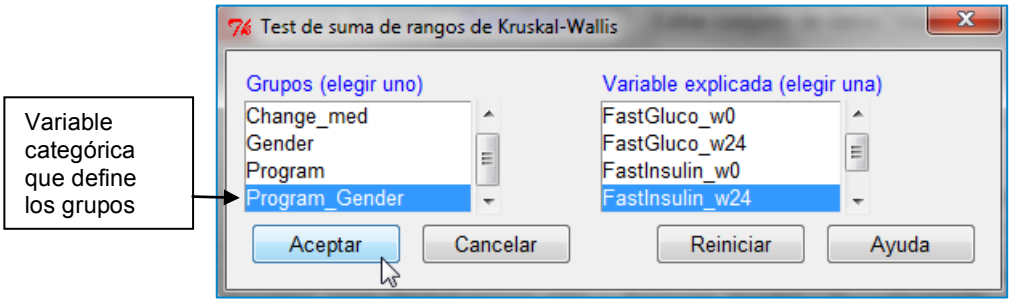
```
> tapply(DietasDiabetes$FastInsulin_w24,DietasDiabetes$Program_Gender,shapiro.test)
$LCKD_Hombre
        Shapiro-Wilk normality test

data:  X[[1L]]
W = 0.9059, p-value = 0.4098

$LCKD_Mujer
        Shapiro-Wilk normality test

data:  X[[2L]]
W = 0.8671, p-value = 0.03061

$LGID_Hombre
        Shapiro-Wilk normality test

data:  X[[3L]]
W = 0.9157, p-value = 0.4367

$LGID_Mujer
        Shapiro-Wilk normality test

data:  X[[4L]]
W = 0.9311, p-value = 0.1296
```

Tabla 35

Así pues, para dar respuesta a la pregunta anterior, utilizaremos un test no paramétrico, el test de Kruskal-Wallis, que nos permite resolver el siguiente contraste de hipótesis:

H_0: La mediana del nivel de insulina en ayunas al finalizar el estudio es la misma en los cuatro grupos.
H_A: La mediana del nivel de insulina en ayunas al finalizar el estudio no es la misma en los cuatro grupos.

Estadísticos / Test no paramétricos / Test de Kruskal-Wallis




```
> tapply(DietasDiabetes$FastInsulin_w24, DietasDiabetes$Program_Gender, median, na.rm=TRUE)
LCKD_Hombre  LCKD_Mujer LGID_Hombre  LGID_Mujer
       10.3        15.4        10.1        13.4

> kruskal.test(FastInsulin_w24 ~ Program_Gender, data=DietasDiabetes)

        Kruskal-Wallis rank sum test

data:  FastInsulin_w24 by Program_Gender
Kruskal-Wallis chi-squared = 9.9374, df = 3, p-value = 0.01911
```

Tabla 36

Para un nivel de significación $\alpha = 0.05$, rechazamos H_0. Existe suficiente evidencia para afirmar que el nivel de insulina en ayunas al finalizar el estudio (semana 24) no es el mismo en los cuatro grupos.

La instrucción

```
pairwise.wilcox.test(DietasDiabetes$FastInsulin_w24,
DietasDiabetes$Program_Gender, p.adjust= "bonf ")
```

nos permite encontrar grupos homogéneos utilizando una alternativa no paramétrica para la comparación múltiple de cada pareja de muestras. En este caso, el nivel de insulina en ayunas del grupo *LCKD_Mujer* tiene una distribución distinta.

```
> pairwise.wilcox.test(DietasDiabetes$FastInsulin_w24,DietasDiabetes$Program_Gender,
p.adjust="bonf")

        Pairwise comparisons using Wilcoxon rank sum test
data:  DietasDiabetes$FastInsulin_w24 and DietasDiabetes$Program_Gender

            LCKD_Hombre LCKD_Mujer LGID_Hombre
LCKD_Mujer  0.770       -          -
LGID_Hombre 1.000       0.035      -
LGID_Mujer  1.000       0.204      0.471

P value adjustment method: bonferroni
```

Tabla 37

2.6. Análisis estadístico de variables categóricas

En esta última sección nos centraremos en el análisis inferencial de variables categóricas. La elección del procedimiento estadístico que utilizaremos para analizar los datos dependerá de si estamos considerando una única variable categórica dicotómica (intervalos de confianza y contrastes de hipótesis para una proporción), una única variable categórica con tres o más categorías (bondad de ajuste) o si consideramos el comportamiento de una variable

categórica en varias poblaciones o la relación entre dos variables categóricas (tablas de contingencia).

A modo ilustrativo daremos respuesta a las tres siguientes cuestiones:

> ***2.6.a*** *Calcula un intervalo de confianza al 95% para la probabilidad de experimentar algún efecto secundario durante el seguimiento de la dieta LCKD.*

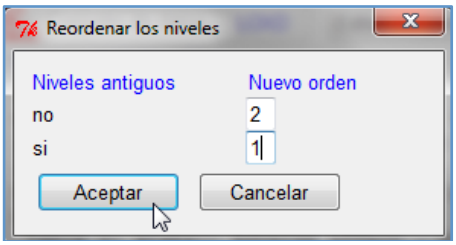
La tabla de frecuencias correspondiente a la variable *Adverse_effect* nos permite obtener una estimación puntual de dicha probabilidad. Como podemos observar en la Tabla 38, 11 pacientes de los 21 que siguieron la dieta *LCKD* (el 52.4%) experimentaron algún efecto secundario durante el seguimiento de la misma. Así pues, $\hat{\pi}$ = 0.52 es el estimador puntual de esa probabilidad.

```
> .Table <- table(LCKD$Adverse_effect)
> .Table  # counts for Adverse_effect

no si
10 11

> round(100*.Table/sum(.Table), 2)  # percentages for Adverse_effect

   no    si
47.62 52.38
```

Tabla 38

Cuando trabajamos con variables categóricas, es importante tener en cuenta que *R-Commander* ordena las diferentes categorías alfabéticamente. Además, si trabajamos con variables dicotómicas, *R-Commander* considera como éxito a la primera categoría y como fracaso a la segunda. En este caso, el éxito es "no (no efecto secundario)". Para cambiar el orden de las categorías podemos utilizar el menú

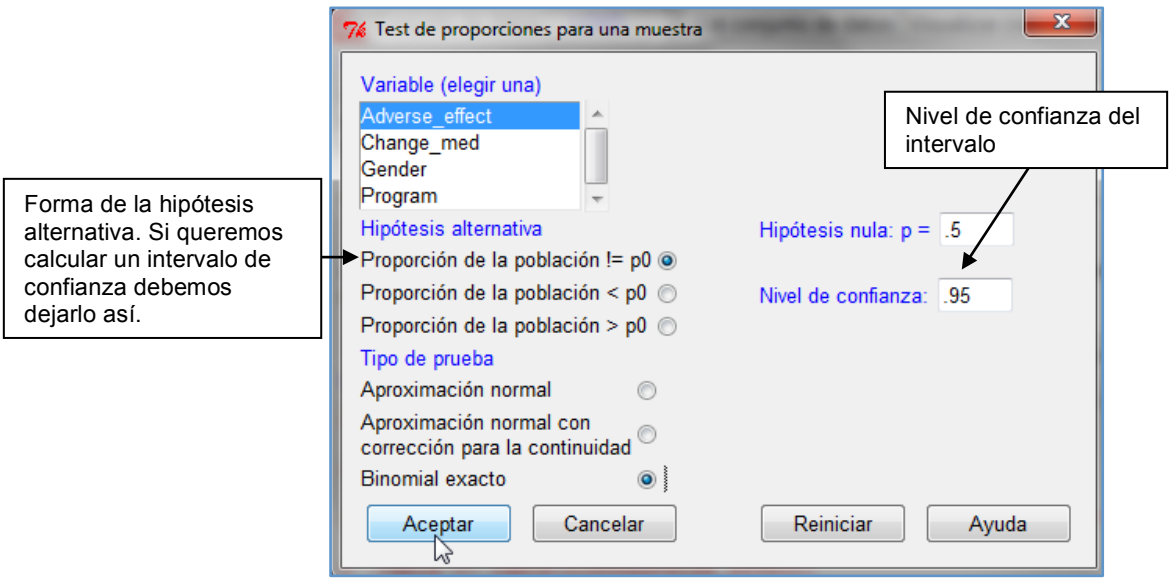
Datos -> Modificar variables del conjunto de datos activo ->
Reordenar niveles de factor



De esta forma, la categoría "si" es la primera categoría y representa éxito.

Una vez ordenadas las distintas categorías de la variable de interés, el menú de *R-Commander* que nos permite calcular intervalos de confianza y resolver contrastes de hipótesis para una proporción es:

Estadísticos / Proporciones / Test de proporciones para una muestra



La salida generada en la Ventana de resultados es

```
> binom.test(rbind(.Table), alternative='two.sided', p=.5, conf.level=.95)

        Exact binomial test

data:  rbind(.Table)
number of successes = 11, number of trials = 21, p-value = 1
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2978068 0.7428694
sample estimates:
probability of success
             0.5238095
```

Tabla 39

A partir de estos resultados concluimos que la probabilidad de experimentar algún efecto secundario durante el seguimiento de la dieta *LCKD* está comprendida entre 0.30 y 0.74 ($0.30 \leq \pi \leq 0.74$) con una confianza del 95%.

2.6.b *La Organización Mundial de la Salud establece tres tipos de obesidad en función del índice de masa corporal:*

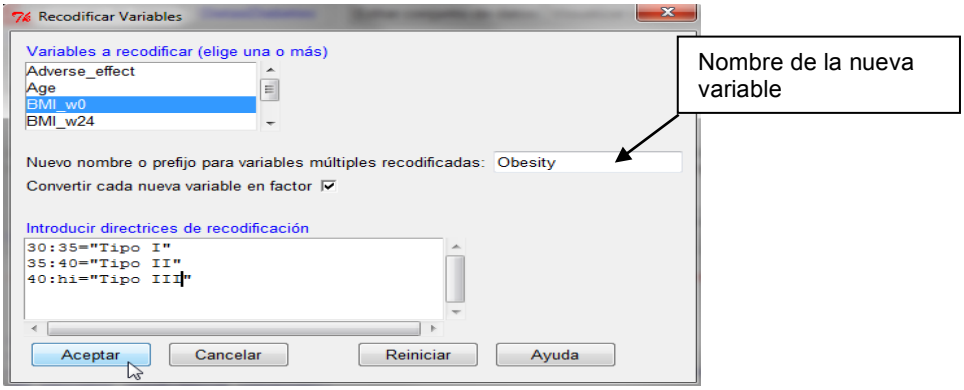
Obesidad	*BMI*
Tipo I	*[30,35)*
Tipo II	*[35,40)*
Tipo III	*≥ 40*

Según la Revista Española de Obesidad (Marzo, 2007, http://www.seedo.es/portals/seedo/consenso/Consenso_SEEDO_2007.pdf), las proporciones en las que se dan los diferentes grados de obesidad en España (población 25-60 años) son, aproximadamente, 83%, 13% y 4%.

¿Son compatibles los datos observados al comienzo del estudio con las proporciones establecidas por la Revista Española de Obesidad?

Para contestar a dicha pregunta debemos, en primer lugar, recodificar la variable *BMI_w0* de forma que cada valor sea asignado a una de las tres categorías {Tipo I, Tipo II, Tipo III}

Datos / Modificar variables del conjunto de datos activo / Recodificar variables

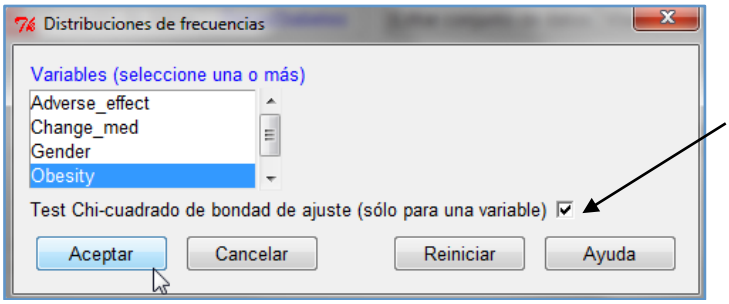


Una vez creada la variable categórica (con nombre *Obesity*) con las tres categorías de interés {C_1='Tipo I', C_2='Tipo II', C_3='Tipo III'}, nos planteamos resolver el siguiente contraste de bondad de ajuste:

H_0: $\pi_1 = 0.83$; $\pi_2 = 0.13$; $\pi_3 = 0.04$
H_A: No se cumple H_0

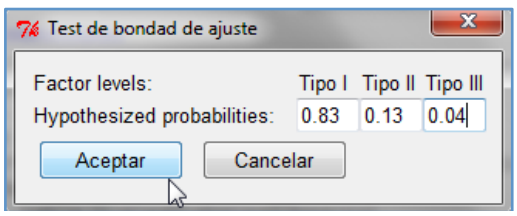
Para ello debemos seleccionar el menú

Estadísticos / Resúmenes / Distribución de frecuencias



y marcar la casilla Test Chi-cuadrado de bondad de ajuste

En la ventana que aparece a continuación debemos introducir las probabilidades de cada categoría que queremos contrastar



La salida obtenida es

```
> .Table <- table(DietasDiabetes$Obesity)
> .Table  # counts for Obesity

  Tipo I  Tipo II Tipo III
      13       22       15

> round(100*.Table/sum(.Table), 2)  # percentages for Obesity

  Tipo I  Tipo II Tipo III
      26       44       30

> .Probs <- c(0.83,0.13,0.04)

> chisq.test(.Table, p=.Probs)

        Chi-squared test for given probabilities

data:  .Table
X-squared = 141.0338, df = 2, p-value < 2.2e-16
```

Tabla 40

El p-valor del contraste es menor que el nivel de significación $\alpha = 0.05$, por lo que rechazamos la hipótesis nula; es decir, las proporciones poblacionales asociadas a los tres tipos de obesidad nos son compatibles con las proporciones establecidas. De hecho, si nos fijamos en la tabla de frecuencias, de los 50 pacientes que participaron en el estudio, 13 (26%) tenían obesidad de tipo I, 22 (44%) obesidad de tipo II y 15 (30%) obesidad de tipo III; las probabilidades de cada categoría observadas en la muestra se alejan mucho de las probabilidades a contrastar y, finalmente, concluimos que las proporciones poblacionales tampoco se ajustan.

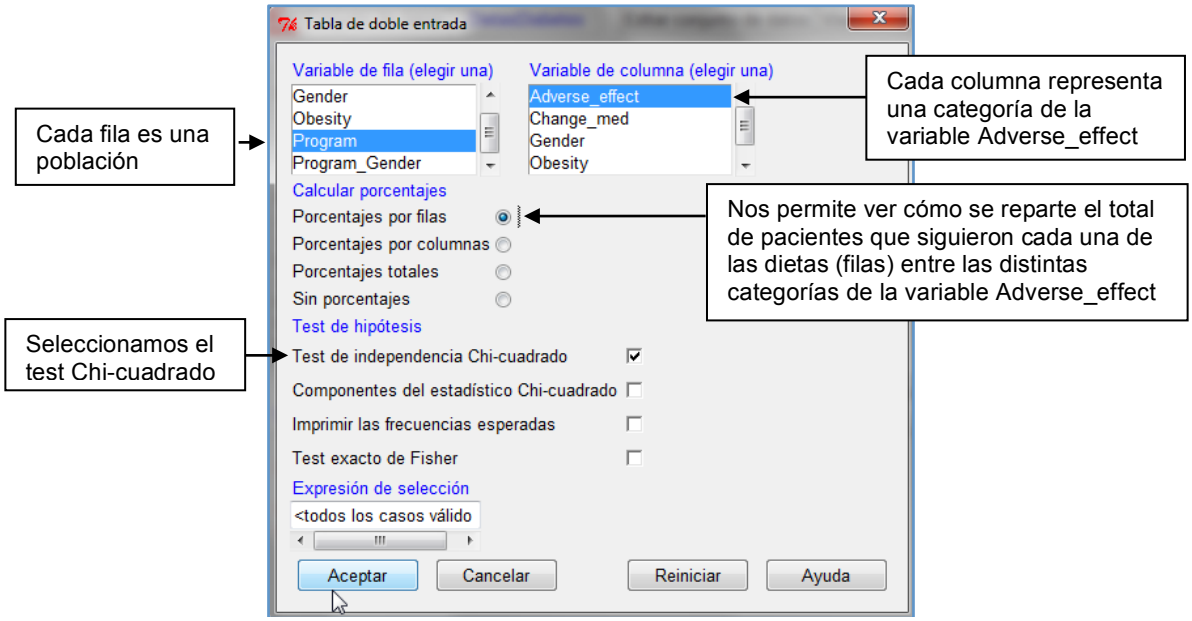
2.6.c *¿Podemos afirmar que la probabilidad de experimentar algún efecto secundario es la misma en ambas dietas?*

Para contestar a la pregunta anterior debemos resolver el siguiente contraste de hipótesis

H_0: Homogeneidad (la variable categórica *Adverse_effect* tiene la misma distribución en las dos poblaciones definidas por el tipo de dieta seguida)
H_A: No homogeneidad (la variable categórica *Adverse_effect* no tiene la misma distribución en las dos poblaciones)

y para ello utilizaremos una tabla de contingencia y el correspondiente test de la Ji-cuadrado desde el menú

Estadísticos / Tablas de contingencia / Tabla de doble entrada



La salida que proporciona ese cuadro de diálogo es

```
> .Table <- xtabs(~Program+Adverse_effect, data=DietasDiabetes)
> .Table
       Adverse_effect
Program no si
   LCKD 10 11
   LGID 15 14

> rowPercents(.Table) # Row Percentages
       Adverse_effect
Program   no   si Total Count
   LCKD 47.6 52.4   100    21
   LGID 51.7 48.3   100    29

> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
        Pearson's Chi-squared test

data:  .Table
X-squared = 0.0821, df = 1, p-value = 0.7745
```

Tabla 41

El p-valor del test es 0.7745 > 0.05, por lo que no rechazamos la hipótesis nula; es decir, no podemos descartar que la variable *Adverse_effect* se distribuya por igual en las dos poblaciones o, equivalentemente, que el riesgo de padecer efectos secundarios sea el mismo con las dos dietas.

3. Propuestas de trabajo para los estudiantes

Actividad 1: Describe gráfica y numéricamente la variable *Triglycerides* al comienzo del estudio (*Triglycerides_w0*) en cada uno de los grupos definidos por el tipo de dieta seguida.

Actividad 2: Repite el ejercicio anterior para la variable *Triglycerides_w24.* ¿Puede apreciarse alguna diferencia en los resultados obtenidos?

Actividad 3: ¿Existe relación lineal entre las variables *Triglycerides_w0 y Triglycerides_w24*? En caso afirmativo, calcula la recta de regresión que consideres adecuada.

Actividad 4: Representa gráfica y numéricamente la variable *Change_med*. A partir de los resultados obtenidos, ¿podrías pensar que el seguir una de las dos dietas propuestas permite eliminar o reducir la medicación?

Actividad 5: ¿Sería apropiado calcular intervalos de confianza para el nivel medio de triglicéridos al comienzo del estudio en el grupo que siguió la dieta *LCKD*? En caso afirmativo, calcula el intervalo de confianza al 90% y al 95%. ¿Existe alguna diferencia entre los dos intervalos calculados? ¿A qué se debe?

Actividad 6: Calcula un intervalo de confianza al 95% para la media poblacional de la variable *Triglycerides_w24* en el grupo *LCKD.* ¿Podemos pensar que el seguimiento de esta dieta conlleva una reducción del nivel de triglicéridos?

Actividad 7: Se considera que los niveles normales de triglicéridos son menores a 200 mg/dL. ¿Podemos concluir que el nivel de triglicéridos después de seguir de la dieta *LGID* cumple esa condición (es menor de 200 mg/dL)?

Actividad 8: Crea una nueva variable *Dif_BMI = BMI_w0 – BMI_w24* y calcula los estadísticos descriptivos en cada uno de los grupos definidos por el tipo de dieta seguida.

Actividad 9: ¿Podemos concluir que el seguimiento de la dieta *LCKD* conlleva una disminución del índice de masa corporal?

Actividad 10: ¿Podemos concluir que el seguimiento de la dieta *LGID* conlleva una disminución del índice de masa corporal?

Actividad 11: ¿Podemos concluir que la dieta *LCKD* es más efectiva que la dieta *LGID* para reducir el índice de masa corporal?

Actividad 12: ¿Podemos concluir que el índice de masa corporal al finalizar el estudio (semana 24) es el mismo en los cuatro grupos definidos por la variable *Program_Gender*? En caso contrario, define los grupos homogéneos.

Actividad 13: ¿Podemos concluir que la dieta *LCKD* es más efectiva que la dieta *LGID* para reducir/eliminar la medicación?

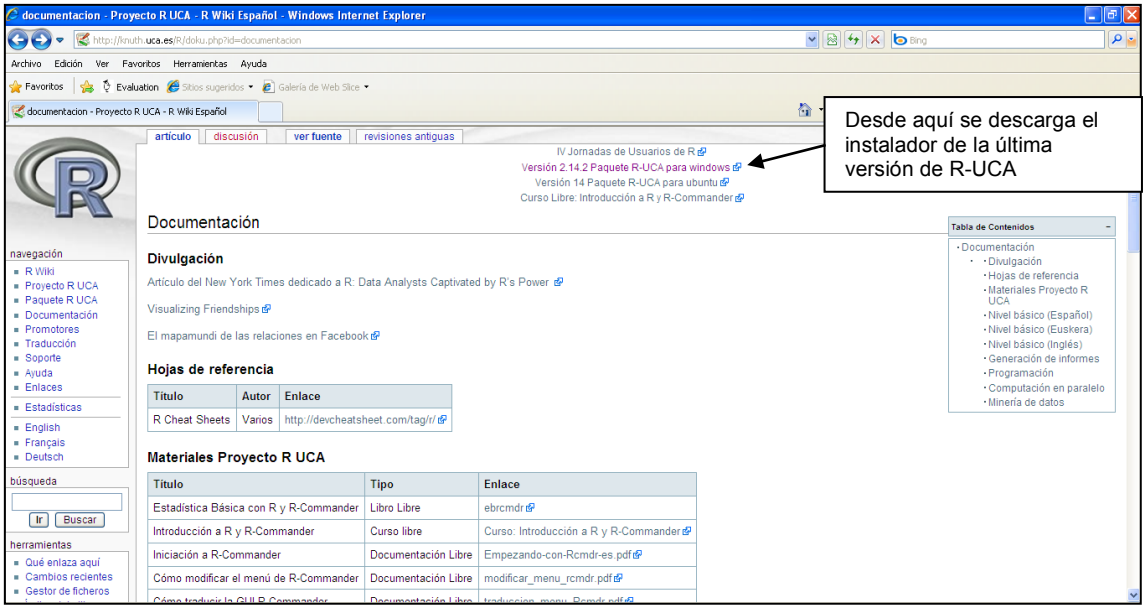
Actividad 14: Calcula un intervalo de confianza del 95% para la probabilidad de reducir/eliminar la medicación después de seguir durante 24 semanas la dieta *LCKD*.

ANEXO: INTRODUCCIÓN AL MANEJO DE DATOS CON R-COMMANDER

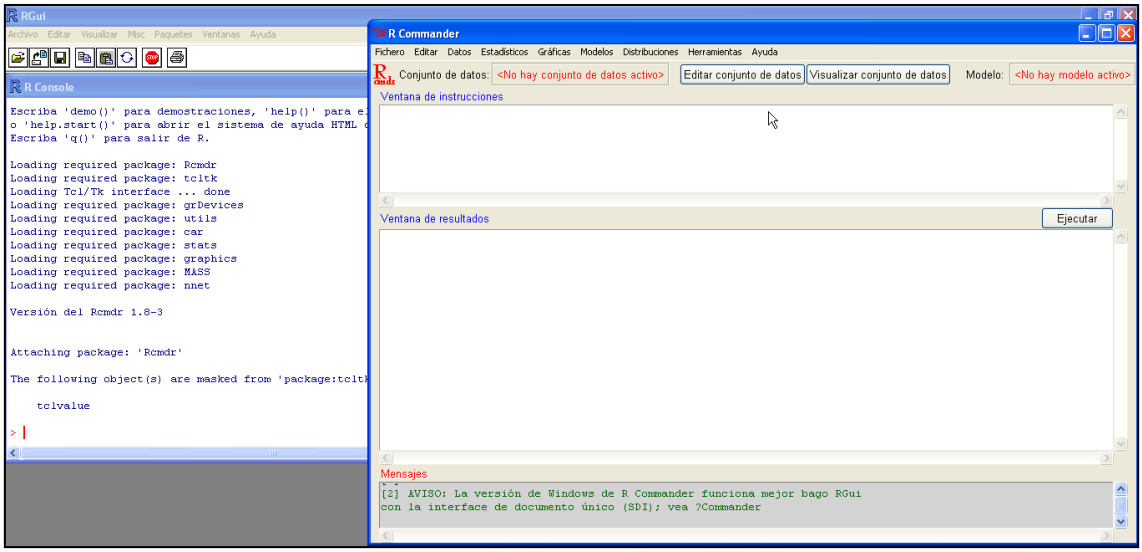
A.1. Introducción al programa

El programa *R-Commander* para *Windows* es una aplicación de libre acceso especializada en el tratamiento estadístico de datos. Para su instalación utilizaremos el paquete *R-UCA*, que instala en un único paso *R*, *R-Commander* y otros paquetes de uso frecuente. Podemos acceder a la última versión de *R-UCA* desde la página web

http://knuth.uca.es/R/doku.php?id=documentacion



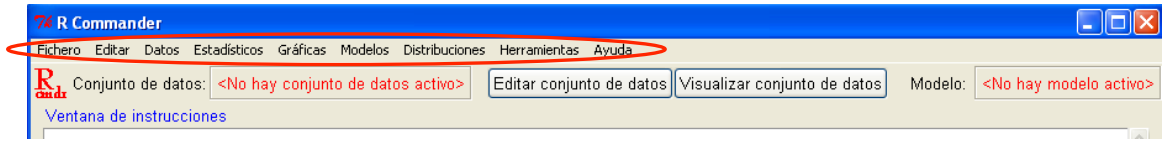
Cuando abrimos el *R* que tenemos instalado aparecen dos ventanas: la del interfaz *R* con su consola y la ventana del *R-Commander*, que es la que utilizaremos nosotros.



En la ventana del *R-Commander* aparecen los menús que nos permitirán introducir y manipular datos, realizar diferentes análisis estadísticos y crear gráficas. Esta ventana está dividida en tres subventanas:

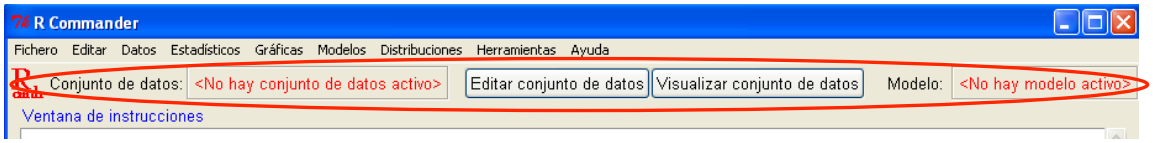
- Ventana de instrucciones: Aquí aparecen los comandos de *R* que se generan al realizar cualquier acción mediante un elemento de los menús. Es editable, así que podemos cambiar el comando y ejecutarlo de nuevo (seleccionando la línea o líneas del comando y presionando el botón Ejecutar). Además, podemos escribir comandos que no aparecen en los menús y ejecutarlos del mismo modo.
- Ventana de resultados: En esta ventana aparecen los resultados de los cálculos efectuados (excepto las gráficas que aparecen en una ventana gráfica).
- Mensajes: Aparecen mensajes de aviso y de errores en el comando.

Los menús incluidos en la ventana de R-Commander son:



Fichero	Abrir y guardar archivos de instrucciones, guardar resultados, guardar el entorno de trabajo de *R* y salir
Editar	Opciones de menú para editar (cortar, copiar, pegar,…) el contenido de las ventanas de instrucciones y resultados
Datos	Importar bases de datos, editar un nuevo conjunto de datos y manipular datos
Estadísticos	Opciones de menú para acceder a la mayoría de los procedimientos estadísticos básicos
Gráficas	Creación y edición de diversos tipos de gráficas
Modelos	Opciones de menú para obtener resúmenes, intervalos de confianza, test de hipótesis, etc. de modelos estadísticos
Distribuciones	Cálculo de probabilidades, cuantiles y gráficas de las distribuciones estadísticas habituales
Herramientas	Cargar paquetes de *R* no relacionados con el *R-Commander*
Ayuda	Manuales de introducción y ayuda de *R-Commander*

La interfaz de *R-Commander* incluye además una barra de herramientas debajo de los menús que nos permite:



- Ver el conjunto de datos activo (conjunto de datos con el que estamos trabajando) y cambiarlo por otro conjunto de datos guardado en la memoria.

- Abrir el editor de datos de *R* para visualizar o editar el conjunto de datos activo.
- Ver el nombre del modelo estadístico activo así como elegir otro modelo guardado en la memoria.

A.2. Manipulación de datos

- **Introducción de datos**

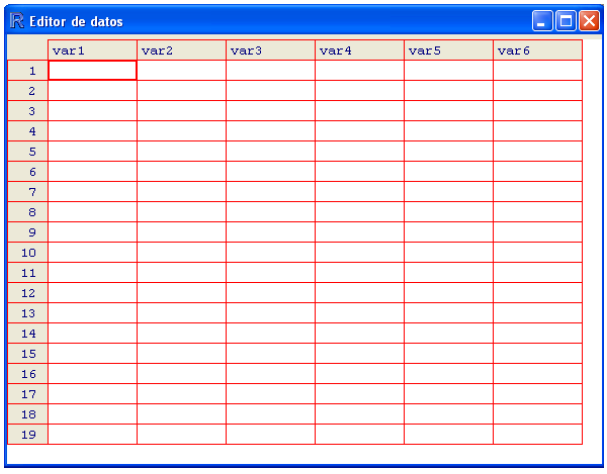
La mayoría de los programas estadísticos (incluido el *R-Commander*) necesitan los datos en formato tabla (o matriz). En las columnas de dicha tabla aparecen las variables que hemos observado en el experimento (y posiblemente otras que podemos calcular a partir de las observadas). Cada columna (o variable) tiene un nombre constituido por caracteres alfanuméricos sin espacios en blanco (si queremos alguna separación en el nombre se puede utilizar el punto y el guión bajo). Cada fila (que también llamaremos caso) está asociada a un individuo y contiene los valores de las variables observados para dicho individuo.

R Commander proporciona varias maneras de introducir datos en R:

- Podemos introducir los datos directamente mediante el menú

  Datos / Nuevo conjunto de datos

  el cual, tras indicar el nombre que queremos dar al conjunto de datos, abre el editor de datos de *R* donde pondremos el nombre de las variables y su tipo (variable numérica o categórica) e introduciremos los datos (los decimales se separan mediante el punto)



  Una vez introducidos los datos debemos cerrar la rejilla. Podemos guardar el conjunto de datos (en formato *R*) mediante el menú

  Datos / Conjunto de datos activo / Guardar el conjunto de datos activo

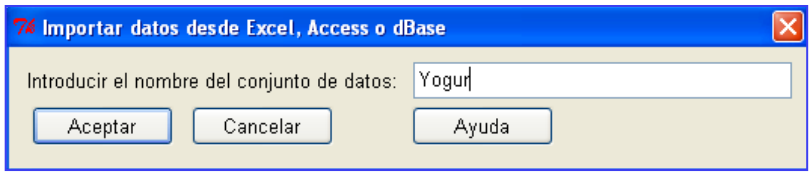
- Podemos importar datos desde ficheros de texto, desde algunos de los ficheros que crean otros paquetes estadísticos (SPSS, Minitab o STATA)

o desde hojas de cálculo o bases de datos (Excel, Access o dBase) mediante la opción

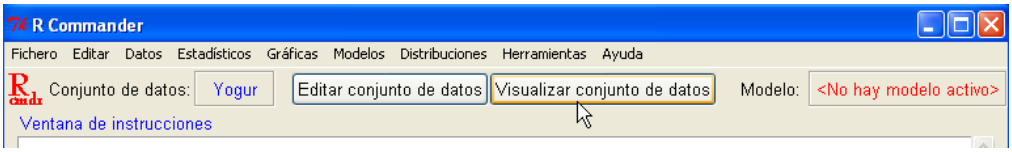
Datos / Importar datos / desde…

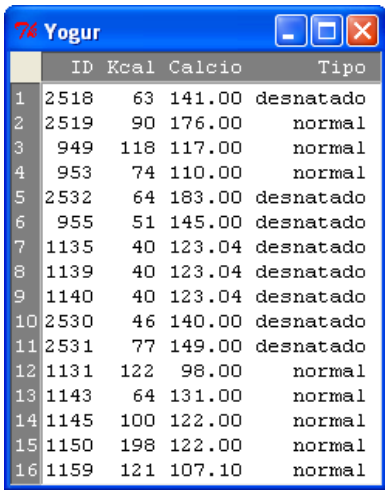
Como ejemplo ilustrativo importaremos el fichero de datos *yogur.xls*, donde se recoge el valor energético (Kcal. por cada 100 g. de porción comestible) y el contenido en calcio (mg. por cada 100 g. de porción comestible) de 16 tipos de yogures. Los datos han sido obtenidos de la Base de Datos Española de Composición de Alimentos (BEDCA, Agencia Española de Seguridad Alimentaria y Nutrición del Ministerio de Sanidad y Política Social y Ministerio de Ciencia e Innovación, http://www.bedca.net), en la que se describen las características nutritivas de los alimentos más frecuentes en la dieta mediterránea.

Para ello seleccionamos el menú Datos/Importar datos/desde conjunto de datos Excel, Access o dBase. A continuación se abre una ventana donde debemos indicar el nombre del conjunto de datos. En este caso decidimos llamarle *Yogur*.



Si pinchamos en el recuadro *Visualizar conjunto de datos* podemos visualizar los datos importados:





Yogur

	ID	Kcal	Calcio	Tipo
1	2518	63	141.00	desnatado
2	2519	90	176.00	normal
3	949	118	117.00	normal
4	953	74	110.00	normal
5	2532	64	183.00	desnatado
6	955	51	145.00	desnatado
7	1135	40	123.04	desnatado
8	1139	40	123.04	desnatado
9	1140	40	123.04	desnatado
10	2530	46	140.00	desnatado
11	2531	77	149.00	desnatado
12	1131	122	98.00	normal
13	1143	64	131.00	normal
14	1145	100	122.00	normal
15	1150	198	122.00	normal
16	1159	121	107.10	normal

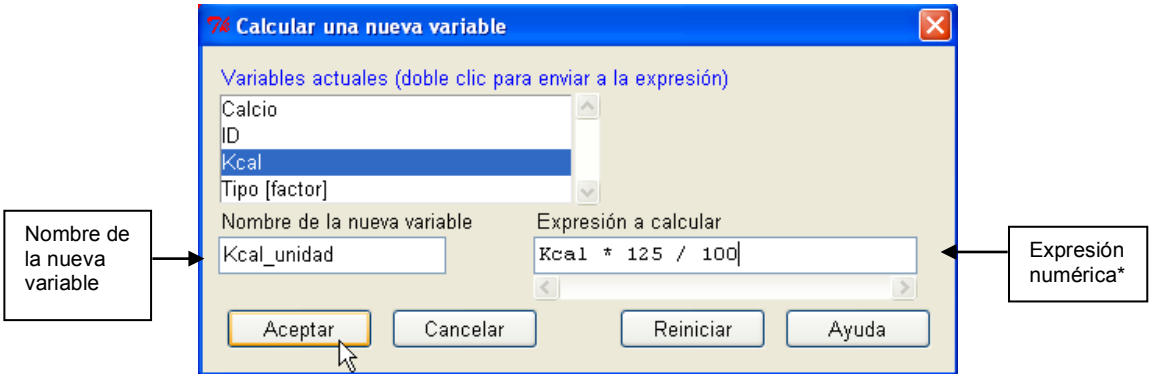
El conjunto de datos tiene cuatro variables: *ID* (código del producto en la base de datos), *Kcal*, *Calcio* y *Tipo* (desnatado o normal) y 16 filas, una por cada tipo de yogur.

Una vez obtenido el archivo de datos, y previamente al análisis estadístico, es posible modificar los datos originales presentes en el archivo. Dichos cambios pueden referirse a las variables o a los casos. Por ejemplo, puede interesarnos crear nuevas variables mediante transformaciones de las ya existentes, recodificar los valores de alguna variable o seleccionar aquellos casos que cumplen una determinada propiedad.

- **Transformación de variables**

R-Commander permite crear nuevas variables mediante transformaciones numéricas de las variables ya existentes. Por ejemplo, si deseamos conocer el valor energético y el contenido en calcio de los diferentes tipos de yogures por unidad de consumo (125 g.), podemos generar dos nuevas variables que contengan esta información.

Datos / Modificar variables del conjunto de datos activo / Calcular una nueva variable



Repitiendo el mismo proceso para la variable *Calcio* generaremos las dos nuevas variables de interés. Pinchando sobre el recuadro *Visualizar conjunto de datos* podemos ver los valores de las nuevas variables.

Yogur

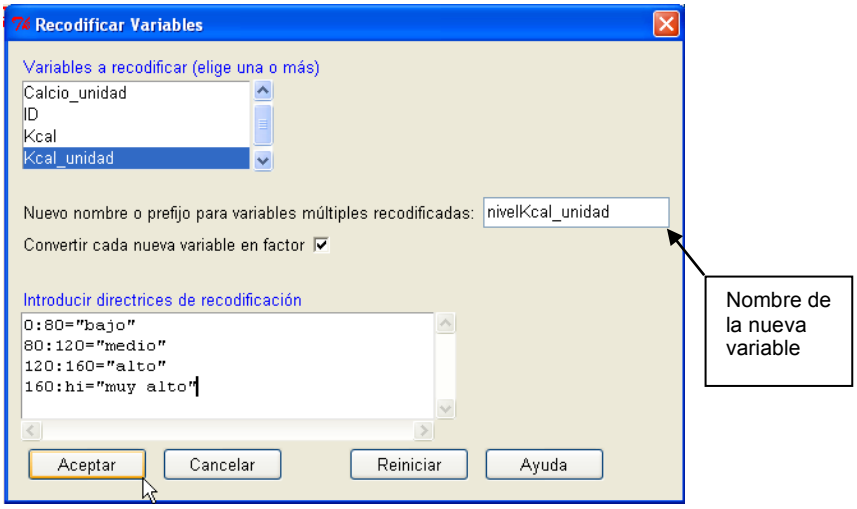
	ID	Kcal	Calcio	Tipo	Kcal_unidad	Calcio_unidad
1	2518	63	141.00	desnatado	78.75	176.250
2	2519	90	176.00	normal	112.50	220.000
3	949	118	117.00	normal	147.50	146.250
4	953	74	110.00	normal	92.50	137.500
5	2532	64	183.00	desnatado	80.00	228.750
6	955	51	145.00	desnatado	63.75	181.250
7	1135	40	123.04	desnatado	50.00	153.800
8	1139	40	123.04	desnatado	50.00	153.800
9	1140	40	123.04	desnatado	50.00	153.800
10	2530	46	140.00	desnatado	57.50	175.000
11	2531	77	149.00	desnatado	96.25	186.250
12	1131	122	98.00	normal	152.50	122.500
13	1143	64	131.00	normal	80.00	163.750
14	1145	100	122.00	normal	125.00	152.500
15	1150	198	122.00	normal	247.50	152.500
16	1159	121	107.10	normal	151.25	133.875

*Utilizaremos las variables ya existentes en el archivo, bien escribiendo su nombre en el recuadro *Expresión a calcular* o seleccionándola del listado de variables actuales que aparece en la parte superior del recuadro (con doble clic).
La expresión numérica, que puede involucrar a más de una variable, puede utilizar operadores aritméticos habituales (+, -, *, /, ^) y otras funciones matemáticas (*log*, *exp*, *sqrt*, *sin*, *cos*,…).

- **Recodificación de variables**

A partir de una variable podemos crear otra cuyos valores sean una recodificación de los de la primera.
Supongamos que queremos recodificar la variable *Kcal_unidad* en una nueva variable llamada *nivelKcal_unidad*, clasificándola en: bajo ≤ 80, medio (80,120], alto (120,160], muy alto > 160.

Datos / Modificar variables del conjunto de datos activo / Recodificar variables

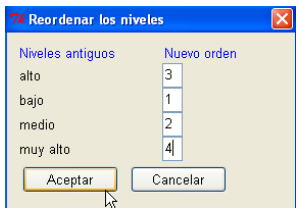


El menú anterior nos permite crear una nueva variable categórica (*nivelKcal_unidad*) que agrupa los valores de la variable *Kcal_unidad* en cuatro intervalos que dan lugar a las diferentes categorías {bajo, medio, alto, muy alto}*.

Yogur

	ID	Kcal	Calcio	Tipo	Kcal_unidad	Calcio_unidad	nivelKcal_unidad
1	2518	63	141.00	desnatado	78.75	176.250	bajo
2	2519	90	176.00	normal	112.50	220.000	medio
3	949	118	117.00	normal	147.50	146.250	alto
4	953	74	110.00	normal	92.50	137.500	medio
5	2532	64	183.00	desnatado	80.00	228.750	bajo
6	955	51	145.00	desnatado	63.75	181.250	bajo
7	1135	40	123.04	desnatado	50.00	153.800	bajo
8	1139	40	123.04	desnatado	50.00	153.800	bajo
9	1140	40	123.04	desnatado	50.00	153.800	bajo
10	2530	46	140.00	desnatado	57.50	175.000	bajo
11	2531	77	149.00	desnatado	96.25	186.250	medio
12	1131	122	98.00	normal	152.50	122.500	alto
13	1143	64	131.00	normal	80.00	163.750	bajo
14	1145	100	122.00	normal	125.00	152.500	alto
15	1150	198	122.00	normal	247.50	152.500	muy alto
16	1159	121	107.10	normal	151.25	133.875	alto

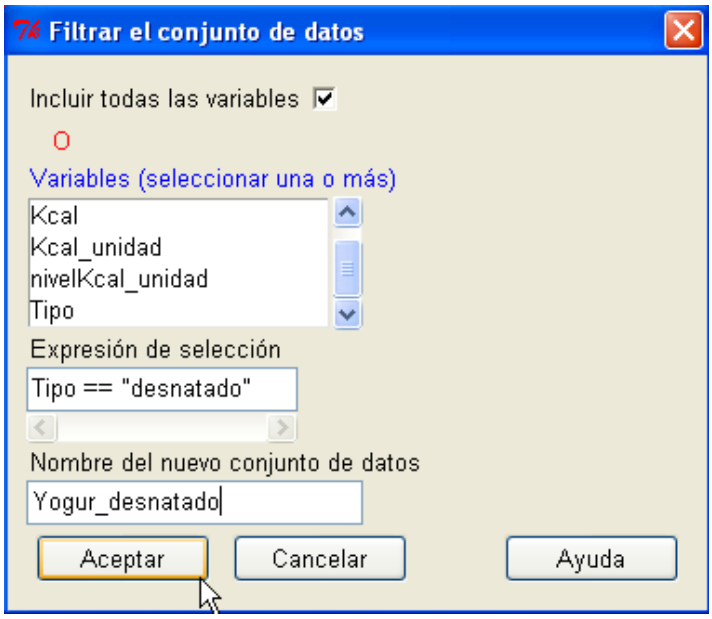
A la hora de trabajar con la variable categórica debemos tener en cuenta que R ordena las categorías por orden alfabético. Para cambiar el orden de las categorías debemos utilizar el menú *Datos / Modificar variables del conjunto de datos activo / Reordenar niveles de factor*



- **Filtrado de datos**

El programa *R-Commander* permite seleccionar determinados casos para un próximo proceso, de forma PERMANENTE, sobre la base de un criterio lógico. Para ello seleccionaremos el menú

Datos / Conjunto de datos activo / Filtrar el conjunto de datos activo



En este caso hemos creado un nuevo conjunto de datos, llamado *Yogur_desnatado*, que incluye únicamente los datos de los yogures desnatados.

Si hemos hecho bien la selección, en la ventana de mensajes aparecerá

NOTA: El conjunto de datos Yogur_desnatado tiene 8 filas y 7 columnas.

Si dice que tiene 0 filas es porque hemos puesto mal la condición de selección (falta un =, faltan las comillas, el nombre de la categoría que queremos seleccionar no está bien escrito,…). Para asegurarnos que hemos filtrado bien los datos es conveniente visualizar el conjunto de datos activo.

Yogur_desnatado

	ID	Kcal	Calcio	Tipo	Kcal_unidad	Calcio_unidad	nivelKcal_unidad
1	2518	63	141.00	desnatado	78.75	176.25	bajo
5	2532	64	183.00	desnatado	80.00	228.75	bajo
6	955	51	145.00	desnatado	63.75	181.25	bajo
7	1135	40	123.04	desnatado	50.00	153.80	bajo
8	1139	40	123.04	desnatado	50.00	153.80	bajo
9	1140	40	123.04	desnatado	50.00	153.80	bajo
10	2530	46	140.00	desnatado	57.50	175.00	bajo
11	2531	77	149.00	desnatado	96.25	186.25	medio