Máster en Data Science. URJC
Técnicas y Métodos de Ciencia de Datos

Examen Final

Carlos Sánchez Vega

2017

Índice

0.1 Análisis de los datos

El objetivo de esta práctica es es verificar si hay algunos tipos de deporte que estéen relacionados con menor indice de masa corporal para una base de datos de deportistas australianos.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
```

```
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(knitr)
library(ggplot2)
```

El primer paso será leer el fichero de datos

```
deportistas <- read.csv(file="/home/csanchez/Downloads/Deportistas.csv",
                        header=TRUE, sep=",")
```

A continuación, realizaremos un análisis de los datos: En primer lugar deberemos distinguir cuáles son las variables cualitativas y cuales son las cuantitativas en el caso de los datos de los deportistas:

- Deporte Practicado
- Sexo

En lo que respecta a las variables cuantitativas, se pueden destacar todas las demás columnas. Es decir:

- Ht (altura)
- Wt (peso)
- BMI (porcentaje de masa corporal)
- Bt (Porcentaje de grasa)

Por tanto, tendremos que hacer un análisis de las variables cuantitavivas en función de las cualitativas,ambas mencionadas anteriormente:

Para mejorar la legibilidad, voy a sustituir el valor de los valores “0” (por“Hombre”“) y”1” (por “Mujer”)

```
deportistas$Sex[deportistas$Sex=="0"]<-"Hombre"
deportistas$Sex[deportistas$Sex=="1"]<-"Mujer"
```

Para predecir las relaciones agruparemos los datos por las variables cualitativas, es decir, el deporte y el sexo. Empezaremos agrupando por deporte, calculando los deportes que tienen menor media de porcentaje de masa corporal:

```
deportistas %>%   group_by(Sport) %>%
  select(Sport, Ht, Wt, BMI, Bfat) %>%
  summarise_all(funs(mean, median, min, max))%>%arrange(BMI_mean)
```

```
## # A tibble: 10 x 17
##    Sport Ht_m~ Wt_m~ BMI_~ Bfat~ Ht_m~ Wt_m~ BMI_~ Bfat~ Ht_m~ Wt_m~ BMI_~
```

```
##    <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 gym     153  43.6  18.5 11.3    153  44.4  18.4 11.4    149  37.8  17.0
##  2 t_40~   175  64.0  20.7  8.63   176  64.7  20.9  6.99   162  49.2  16.8
##  3 tenn~   174  64.5  21.1 12.9    175  69.7  21.2 11.5    158  45.8  17.1
##  4 b_ba~   189  79.8  22.3 14.8    189  77.7  22.0 15.1    169  62.3  19.0
##  5 netb~   176  69.6  22.4 21.6    176  68.8  22.6 21.3    169  51.9  18.3
##  6 t_sp~   176  71.5  22.9  8.25   175  70.8  23.1  7.52   164  57.3  19.5
##  7 swim    181  75.1  22.9 10.6    181  75.0  22.7  9.30   165  55.1  19.0
##  8 row     182  78.5  23.5 15.6    182  78.7  23.7 16.6    156  49.8  19.7
##  9 w_po~   188  86.7  24.5 12.2    190  87.3  24.3 11.6    179  74.4  21.3
## 10 field   181  90.0  27.5 14.9    180  87.5  27.4 14.0    170  58.0  20.1
## # ... with 5 more variables: Bfat_min <dbl>, Ht_max <dbl>, Wt_max <dbl>,
## #   BMI_max <dbl>, Bfat_max <dbl>
```

Agrupando por deporte, se podría pensar que el deporte que conlleva menor porcentaje de masa corporal es "Gym"

```
deportistas %>%   group_by(Sex) %>%
  select(Sex, Ht, Wt, BMI, Bfat) %>%
  summarise_all(funs(mean, median, min, max))%>%arrange(BMI_mean)
```

```
## # A tibble: 2 x 17
##   Sex    Ht_m~ Wt_m~ BMI_~ Bfat~ Ht_m~ Wt_m~ BMI_~ Bfat~ Ht_m~ Wt_m~ BMI_~
##   <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mujer    175  67.3  22.0 17.8    175  68.1  21.8 17.9    149  37.8  16.8
## 2 Hombre   186  82.5  23.9  9.25   186  83.0  23.6  8.62   165  53.8  19.6
## # ... with 5 more variables: Bfat_min <dbl>, Ht_max <dbl>, Wt_max <dbl>,
## #   BMI_max <dbl>, Bfat_max <dbl>
```

Agrupando por sexo, se podría pensar que el sexo que posee menor porcentaje de masa corporal son las mujeres

```
deportistas %>%   group_by(Sex,Sport) %>%
  select(Sex,Sport, Ht, Wt, BMI, Bfat) %>%
  summarise_all(funs(mean, median, min, max))%>%arrange( Sex,BMI_mean)
```

```
## # A tibble: 17 x 18
## # Groups:   Sex [2]
##    Sex   Sport Ht_m~ Wt_m~ BMI_~ Bfat~ Ht_m~ Wt_m~ BMI_~ Bfat~ Ht_m~ Wt_m~
##    <chr> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 Homb~ t_40~   179  68.2  21.2  6.69   179  68.7  21.2  6.43   169  57.4
##  2 Homb~ tenn~   184  75.4  22.3  9.08   183  75.2  22.2  9.28   178  71.1
##  3 Homb~ b_ba~   196  88.9  23.2  8.89   195  88.6  23.2  8.64   186  75.5
```

```
##  4 Homb~ swim    186  81.7  23.7  8.30   184  83.0  23.3  8.47   173  67.0
##  5 Homb~ t_sp~   179  75.8  23.7  7.29   178  72.9  23.6  6.76   171  69.1
##  6 Homb~ w_po~   188  86.7  24.5 12.2    190  87.3  24.3 11.6    179  74.4
##  7 Homb~ row     188  86.8  24.6  9.41   188  88.2  25.1  9.36   165  53.8
##  8 Homb~ field   185  95.8  28.0 11.9    185  96.2  28.7 10.8    179  75.2
##  9 Mujer gym     153  43.6  18.5 11.3    153  44.4  18.4 11.4    149  37.8
## 10 Mujer t_40~   169  57.2  20.0 11.8    171  57.3  20.1 11.1    162  49.2
## 11 Mujer tenn~   169  58.2  20.4 15.0    168  56.1  20.5 15.3    158  45.8
## 12 Mujer t_sp~   170  59.7  20.6 10.9    170  59.8  20.2 10.9    164  57.3
## 13 Mujer b_ba~   182  71.3  21.4 20.3    185  69.1  21.2 19.9    169  62.3
## 14 Mujer swim    173  65.7  21.9 13.9    173  64.8  22.0 13.4    165  55.1
## 15 Mujer netb~   176  69.6  22.4 21.6    176  68.8  22.6 21.3    169  51.9
## 16 Mujer row     179  72.9  22.8 19.8    180  73.9  23.0 19.6    156  49.8
## 17 Mujer field   173  80.0  26.8 20.0    172  82.8  27.0 20.1    170  58.0
## # ... with 6 more variables: BMI_min <dbl>, Bfat_min <dbl>, Ht_max <dbl>,
## #   Wt_max <dbl>, BMI_max <dbl>, Bfat_max <dbl>
```

Por último, si agrupamos por sexo y deporte se puede determinar que el deporte que tiene asociado menor porcentaje de masa corporal para hombres son los "400 metros" de carrera y, en el caso de las mujeres, el deporte sería "gym".

No obstante, para poder determinar que exista alguna relación entre el deporte y el porcentaje de masa corportal, independientemente del sexo, seria necesario que existieran registros en todos los deportes para ambos sexos . Sin embargo, hay deportes como "gym" para hombres, de los que no se tienen datos.

0.2 Constraste de hipotesis

Queremos contrastar si los índices de masa corporal de cualquier deporte son iguales. Para hacer un contraste de hipótesis, se deben seguir los siguientes pasos:

- Hacer un contraste de medias
- Hacer un contraste de varianzas

La hipótesis nula consistiría en que la comparativa de las medias de dos deportes de la muestra, por ejemplo "b_ball" y "field", son iguales en los dos deportes contra la hipótesis alternativa,que reflejaría que las medias que son distintas. Formalmente el contraste que se pide es

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$$

Como las muestras de las que disponemos son grandes, podemos usar el t-test:

Seleccionamos las filas correspondientes a ambos deportes:

```
bball = deportistas%>% filter(Sport=="b_ball")
field = deportistas %>% filter(Sport=="field")
```

Comenzamos analizando el caso de la hipótesis nula: Calculamos su t-test:

```
sol.ttest.vareq=t.test(bball$BMI,field$BMI,alternative="two.side",
                       var.equal=TRUE,conf.level=0.95)
sol.ttest.vareq
```

```
##
##  Two Sample t-test
##
## data:  bball$BMI and field$BMI
## t = -5.7166, df = 42, p-value = 1.015e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.145401 -3.416747
## sample estimates:
## mean of x mean of y
##  22.25840  27.53947
```

Calculamos el p-valor

```
sol.ttest.vareq$p.value
```

```
## [1] 1.014546e-06
```

Calculamos su intervalo de confianza

```
sol.ttest.vareq$conf.int
```

```
## [1] -7.145401 -3.416747
## attr(,"conf.level")
## [1] 0.95
```

A continuación, nos centraremos en el caso de la hipótesis alternativa:

Calculamos su t-test:

```
sol.ttest.varneq=t.test(bball$BMI,field$BMI,alternative="two.side",
                        var.equal=FALSE,conf.level=0.95)
sol.ttest.varneq
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  bball$BMI and field$BMI
## t = -5.2234, df = 23.942, p-value = 2.38e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.368029 -3.194119
## sample estimates:
## mean of x mean of y
##  22.25840  27.53947
```

Calculamos su p-valor

```
sol.ttest.varneq$p.value
```

```
## [1] 2.379856e-05
```

Calculamos el intervalo de confianza

```
sol.ttest.varneq$conf.int
```

```
## [1] -7.368029 -3.194119
## attr(,"conf.level")
## [1] 0.95
```

En conclusión:

- Los p-valores de ambos contrastes son muy pequeños: 1.014546*10-06 (caso de que las medias sean iguales) y $2.379856 \times 10{-}5$ (caso en el que las medias sean diferentes).
- sus intervalos de confianza del 95[-7.145401 -3.416747] y [-7.368029 -3.194119] no contienen al cero

Por ello,tenemos evidencias para rechazar la hipótesis de que las medias de los dos deportes son iguales contra que son distintas.

Contraste para la igualdad de varianzas: Ahora haremos el contraste de las varianzas poblacionales de los dos deportes anteriormente mostrados. La hipótesis nula consistiría en decir que ambas tienen la misma varianza, contra la hipótesis alternativa, que consistiría en decir que las varianzas son distintas.

$$\begin{cases} H_0: & \sigma = \sigma_0 \\ H_1: & \sigma \neq \sigma_0 \end{cases}$$

Para realizar nuestros cálculos, usaremos el test de igualdad de varianzas de Fisher. Podríamos decir que en realidad lo que se contrasta es:

$$\begin{cases} H_0: & \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_0: & \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$$

Es decir. queremos analizar si los intervalos de confianza contienen el 1.

```
sol.var.test=var.test(bball$BMI,field$BMI,ratio=1,alternative="two.sided",
                      conf.level=0.95)
sol.var.test
```

```
##
##  F test to compare two variances
##
## data:  bball$BMI and field$BMI
## F = 0.21714, num df = 24, denom df = 18, p-value = 0.0006527
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.08676269 0.51349361
## sample estimates:
## ratio of variances
##          0.2171407
```

Hallamos su p-valor

```
sol.var.test$p.value
```

```
## [1] 0.0006526505
```

Hallamos su intervalo de confianza

```
sol.var.test$conf.int
```

```
## [1] 0.08676269 0.51349361
## attr(,"conf.level")
## [1] 0.95
```

Como el p-valor es 0.0006526505 y es menor que 0.05, podemos rechazar la hipótesis nula, que consiste en decir que las varianzas son iguales contra que son distintas.

En conclusión, podemos concluir que los dos deportes no tienen un porcentajes de masa corporal iguales y, por tanto,los porcentajes de masa corporal son dependientes del deporte.