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

Propuesta análisis de datos. Deportistas australianos.

César González Fernández

2017-2018

Índice

1 Introducción

En el fichero Deportistas.csv aparecen datos sobre estatura, peso, índice de masa corporal, etc. en una muestra de deportistas profesionales australianos.

En concreto aparecen las siguientes variables:

- Sex: Sexo del deportista (0=hombre, 1=mujer);
- Ht: Altura (en cm.)
- Wt: Peso (en Kg.);
- BMI: Índice de masa corporal;
- Bfat: Porcentaje de grasa corporal;
- Sport: Deporte;

1.1 Análisis descriptivo de los datos.

Se desea realizar un análisis descriptivo de los datos.

En primer lugar deberemos cargar los datos en nuestro entorno. Al ser un fichero CSV, la carga es sencilla.

```
athletes.df <- read.csv("./deportistas.csv")
athletes.df$Sex <- c('Masculino', 'Femenino')[athletes.df$Sex + 1]
kable(head(athletes.df), align = 'c')
```

Sex	Ht	Wt	BMI	Bfat	Sport
Femenino	195.9	78.9	20.56	19.75	b_ball
Femenino	189.7	74.4	20.67	21.30	b_ball
Femenino	177.8	69.1	21.86	19.88	b_ball
Femenino	185.0	74.9	21.88	23.66	b_ball
Femenino	184.6	64.6	18.96	17.64	b_ball
Femenino	174.0	63.7	21.04	15.58	b_ball

En la tabla superior podemos ver una muestra de los datos cargados.

Haciendo uso de la función summary, podemos ver también algunos estadísticos de las diferentes columnas.

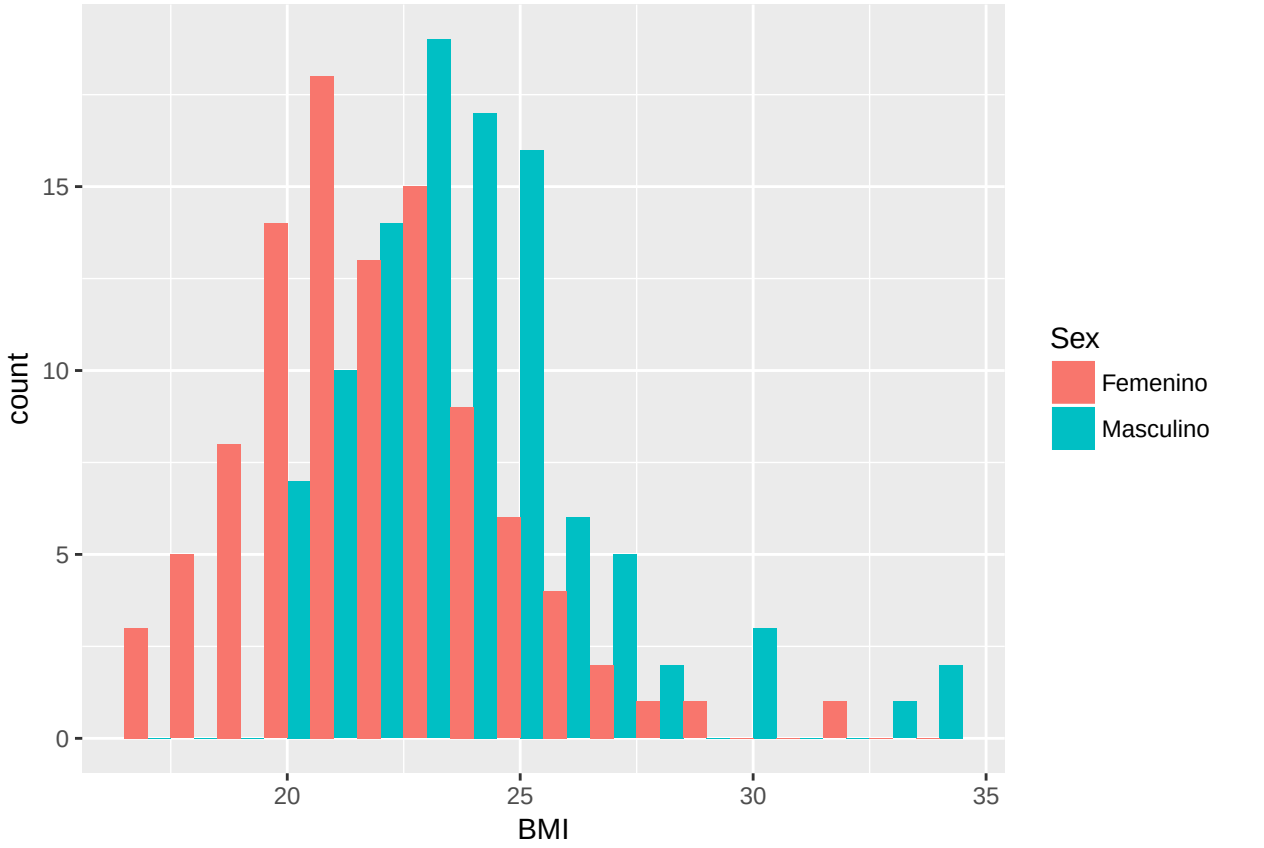
```
kable(summary(athletes.df))
```

Sex	Ht	Wt	BMI	Bfat	Sport
Length:202	Min. :148.9	Min. : 37.80	Min. :16.75	Min. : 5.630	row :37

Sex	Ht	Wt	BMI	Bfat	Sport
Class :character	1st Qu.:174.0	1st Qu.: 66.53	1st Qu.:21.08	1st Qu.: 8.545	t_400m :29
Mode :character	Median :179.7	Median : 74.40	Median :22.72	Median :11.650	b_ball :25
NA	Mean :180.1	Mean : 75.01	Mean :22.96	Mean :13.507	netball:23
NA	3rd Qu.:186.2	3rd Qu.: 84.12	3rd Qu.:24.46	3rd Qu.:18.080	swim :22
NA	Max. :209.4	Max. :123.20	Max. :34.42	Max. :35.520	field :19
NA	NA	NA	NA	NA	(Other):47

Como son los datos de las columnas BMI y Bfat en los que vamos a utilizar en los siguientes apartados, vamos a ver sus histogramas.

```
ggplot(athletes.df, aes(x=BMI, fill=Sex)) +
  geom_histogram(binwidth=1, position="dodge")
```

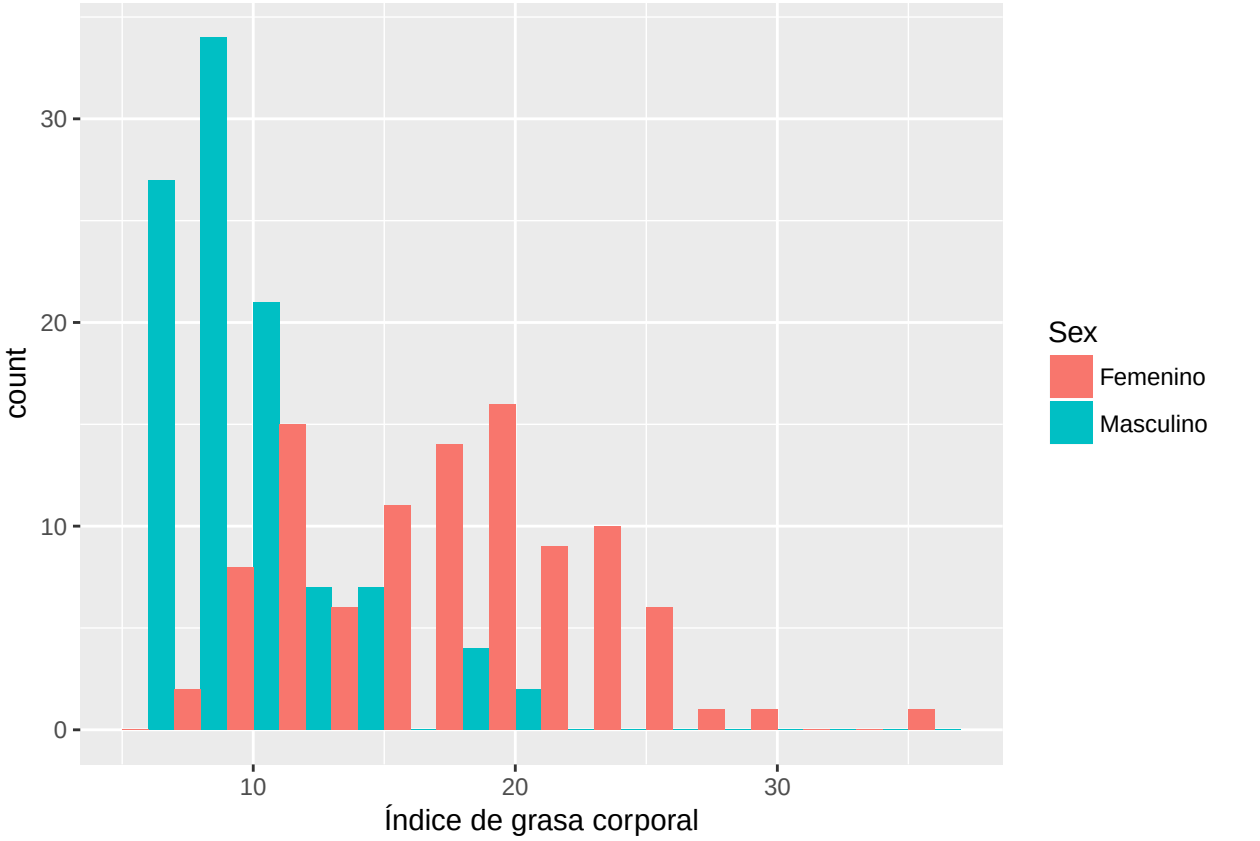



```
  xlab('Índice de masa corporal')
```

```
## $x
## [1] "Índice de masa corporal"
##
## attr(,"class")
```

```
## [1] "labels"
```

```
ggplot(athletes.df, aes(x=Bfat, fill=Sex)) +
  geom_histogram(binwidth=2, position="dodge") +
  xlab('Índice de grasa corporal')
```



1.2 Intervalo de confianza para BMI medio al 98% por sexo.

Vamos a calcular el interbalo de confianza para BMI medio al 98%. Vamos a hacer una diferenciación por sexos.

Empezaremos por los deportistas masculinos:

```
athletes.df.m = athletes.df %>% filter(Sex == 'Masculino')
kable(head(athletes.df.m), align = 'c')
```

Sex	Ht	Wt	BMI	Bfat	Sport
Masculino	172.7	67.0	22.46	8.47	swim
Masculino	176.5	74.4	23.88	7.68	swim
Masculino	183.0	79.3	23.68	6.16	swim
Masculino	194.4	87.5	23.15	8.56	swim

Sex	Ht	Wt	BMI	Bfat	Sport
Masculino	193.4	83.5	22.32	6.86	swim
Masculino	180.2	78.0	24.02	9.40	swim

```
kable(summary(athletes.df.m))
```

Sex	Ht	Wt	BMI	Bfat	Sport
Length:102	Min. :165.3	Min. : 53.80	Min. :19.63	Min. : 5.630	t_400m :18
Class :character	1st Qu.:179.8	1st Qu.: 73.95	1st Qu.:22.29	1st Qu.: 6.968	w_polo :17
Mode :character	Median :185.6	Median : 83.00	Median :23.56	Median : 8.625	row :15
NA	Mean :185.5	Mean : 82.52	Mean :23.90	Mean : 9.251	swim :13
NA	3rd Qu.:191.0	3rd Qu.: 90.30	3rd Qu.:25.16	3rd Qu.:10.010	b_ball :12
NA	Max. :209.4	Max. :123.20	Max. :34.42	Max. :19.940	field :12
NA	NA	NA	NA	NA	(Other):15

Calculamos el número de datos, la media muestral y la varianza:

```
n.m.bmi <- length(athletes.df.m$BMI)
mean.m.bmi <- mean(athletes.df.m$BMI)
var.m.bmi <- var(athletes.df.m$BMI)
```

Lo cuál nos da que n = 102, $\overline{X} = 23.9036275$ y $\sigma = 7.6589441$.

Dado que la muestra con la que trabajamos es grande (n > 30), podemos aplicar el Teorema de Límite Central y obtener:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \approx N(0,1)$$

Y por tanto calcular el intervalo de confianza como:

$$[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Siendo en nuestro caso $\alpha = 0.02$.

```
alpha <- 1 - 0.98
lim.inf.m.bmi <- mean.m.bmi - qnorm(1 - alpha / 2) *
  var.m.bmi / sqrt(n.m.bmi)
lim.sup.m.bmi <- mean.m.bmi + qnorm(1 - alpha / 2) *
  var.m.bmi / sqrt(n.m.bmi)
```

Asi tenemos que el intervalo de confianza al 98% se encuentra entre 22.1394451 y 25.6678098

Repetimos lo mismo, pero esta vez con los datos de las deportistas de genero femenino:

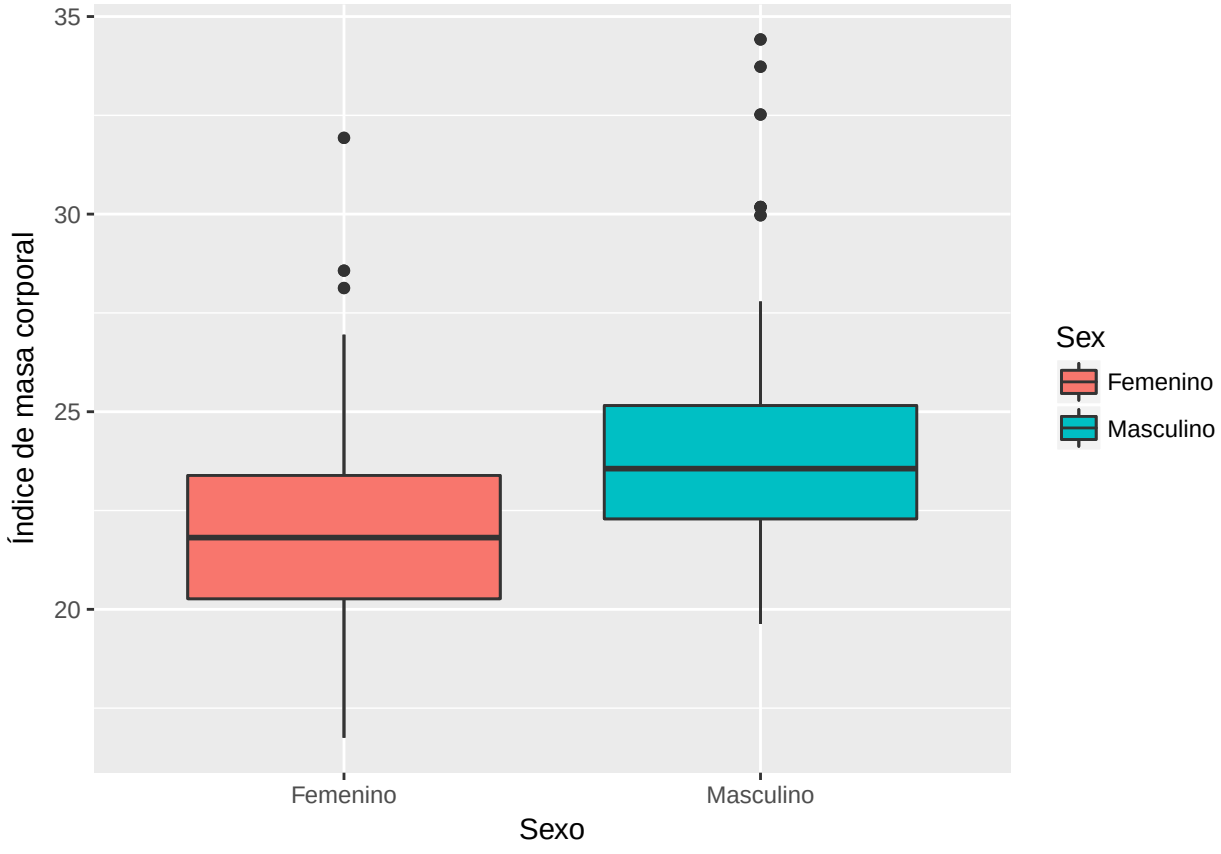
```
athletes.df.f = athletes.df %>% filter(Sex == 'Femenino')
n.f.bmi <- length(athletes.df.f$BMI)
mean.f.bmi <- mean(athletes.df.f$BMI)
var.f.bmi <- var(athletes.df.f$BMI)
lim.inf.f.bmi <- mean.f.bmi - qnorm(1 - alpha / 2) *
  var.f.bmi / sqrt(n.f.bmi)
lim.sup.f.bmi <- mean.f.bmi + qnorm(1 - alpha / 2) *
  var.f.bmi / sqrt(n.f.bmi)
```

Obtenemos que n = 100, $\overline{X} = 21.9892$ y $\sigma = 6.9697468$. Y que el intervalo de confianza al 98% se encuentra entre 20.3677944 y 23.6106056.

1.3 Diferencias entre el BMI de los deportistas masculinos y femeninos.

A la vista de los datos que tenemos, podrimos concluir lo siguiente:

La BMI de los deportistas masculinos es mayor que la de las deportistas femeninas.



Vamos a contrastar pues esta hipótesis. Diferenciaremos esntre:

- Hipótesis nula (H_0): La BMI de los deportistas masculinos es mayor que la de los deportistas femeninos. $BMI_m > BMI_f$
- Hipótesis alternativa (H_0): La BMI de los deportistas masculinos es menor o igual que la de los deportistas femeninos. $BMI_m <= BMI_f$

Nos encontramos en que las muestras con las que trabajamos se tratan de muestras independientes (deportistas masculinos vs deportistas femeninas).

En primer lugar, vamos a comprobar si podemos asumir que la varianza de ambas muestras son igules:

```
var.bmi.test <- var.test(athletes.df.f$BMI, athletes.df.m$BMI,
                         alternative = "two.sided", conf.level = 0.95)
var.bmi.test
```

```
##
##  F test to compare two variances
##
## data:  athletes.df.f$BMI and athletes.df.m$BMI
## F = 0.91001, num df = 99, denom df = 101, p-value = 0.6388
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6137557 1.3503374
## sample estimates:
## ratio of variances
##          0.9100141
```

El valor de p-value, 0.6387501, es muy alto, por lo que podemos asumir que las varianzas son iguales.

Asumiento lo anterior, podemos hacer una comparación de las medias:

```
bmi.test <- t.test(athletes.df.m$BMI, athletes.df.f$BMI,
                   alternative = "less", conf.level = 0.95,
                   var.equal = TRUE)
bmi.test
```

```
##
##  Two Sample t-test
##
## data:  athletes.df.m$BMI and athletes.df.f$BMI
## t = 5.0289, df = 200, p-value = 1
## alternative hypothesis: true difference in means is less than 0
```

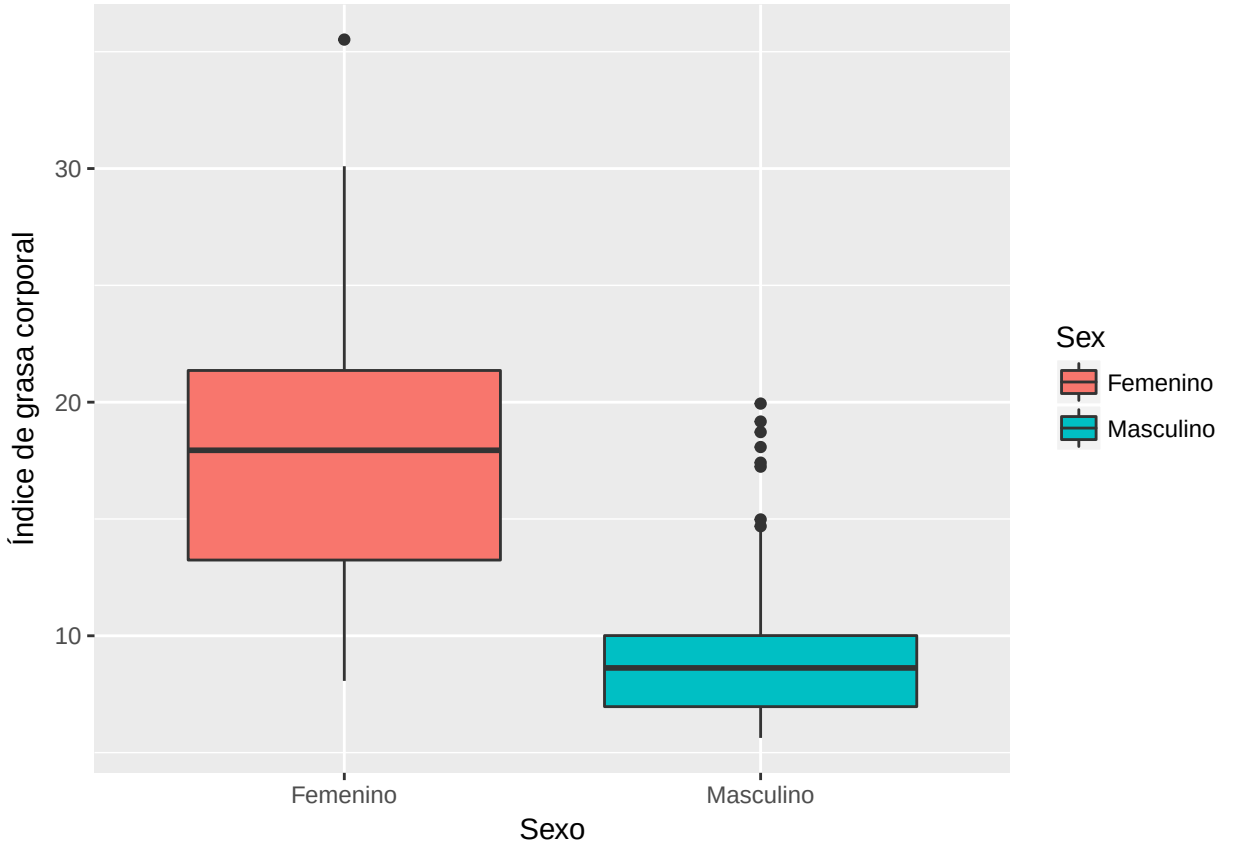
```
## 95 percent confidence interval:
##      -Inf 2.543512
## sample estimates:
## mean of x mean of y
##  23.90363  21.98920
```

Ahora tenemos que p-value es muy alta, es 0.9999995. Por tanto podemos decir que no hay evidencias de que la hipótesis nula, H_0, sea falsa.

1.4 Y con respecto a la grasa corporal, ¿qué podemos decir?

Por los datos de la muestra podemos decir que:

> El índice de grasa corporal de las deportistas femeninas es mayor que la de los deportistas masculinos.



Es este caso definiremos:

- H_0: El índice de grasa corporal de las deportistas femeninas es mayor que el de los deportistas masculinos. $Bfat_f > Bfat_m$
- H_1: El índice de grasa corporal de las deportistas femeninas es menor o igual que la de los deportistas masculinos. $Bfat_f <= Bfat_m$

Al igual que en el apartado anterior, empezamos estudiando si podemos asumir que la variana de ambas poblaciones es igual.

```
var.bfat.test <- var.test(athletes.df.f$Bfat, athletes.df.m$Bfat,
                          alternative = "two.sided", conf.level = 0.95)
var.bfat.test
```

```
##
##  F test to compare two variances
##
## data:  athletes.df.f$Bfat and athletes.df.m$Bfat
## F = 2.9317, num df = 99, denom df = 101, p-value = 1.5e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.977270 4.350234
## sample estimates:
## ratio of variances
##           2.931693
```

En este caso, p-value es muy pequeño, por tanto rechazamos la hipotesis de que ambas varianzas son iguales.

Sabiendo esto:

```
bfat.test <- t.test(athletes.df.f$Bfat, athletes.df.m$Bfat,
                    alternative = "less", conf.level = 0.95)
bfat.test
```

```
##
##  Welch Two Sample t-test
##
## data:  athletes.df.f$Bfat and athletes.df.m$Bfat
## t = 13.65, df = 158.87, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 9.640388
## sample estimates:
## mean of x mean of y
## 17.849100  9.250882
```

Siendo como es en este caso p-valor igual a , podemos rechazar la hipóstesis alternativa.