PRÁCTICA CALIFICADA N°2

Curso: Estadística para Ingeniería (EST218)
Horario: 0508
Profesora: Osorio Martinez, Miluska Elena

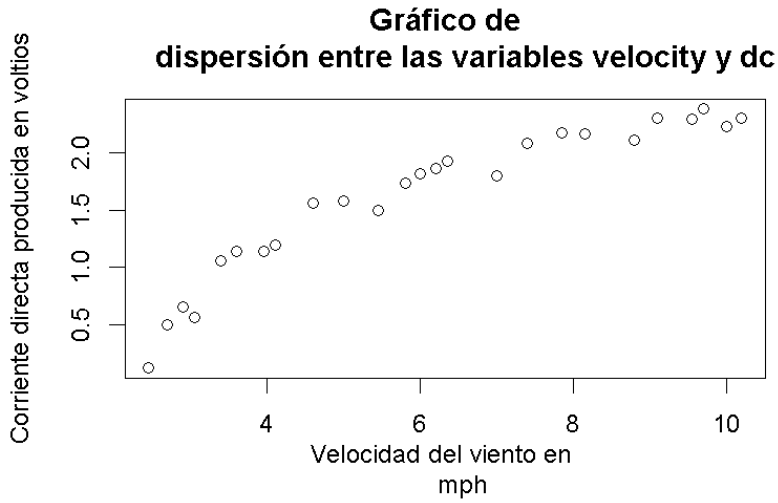
Integrantes:

Iván Alexander Aráoz Andrade
20201216
i.araoz@pucp.edu.pe

Pregunta 1:

Ítem a:

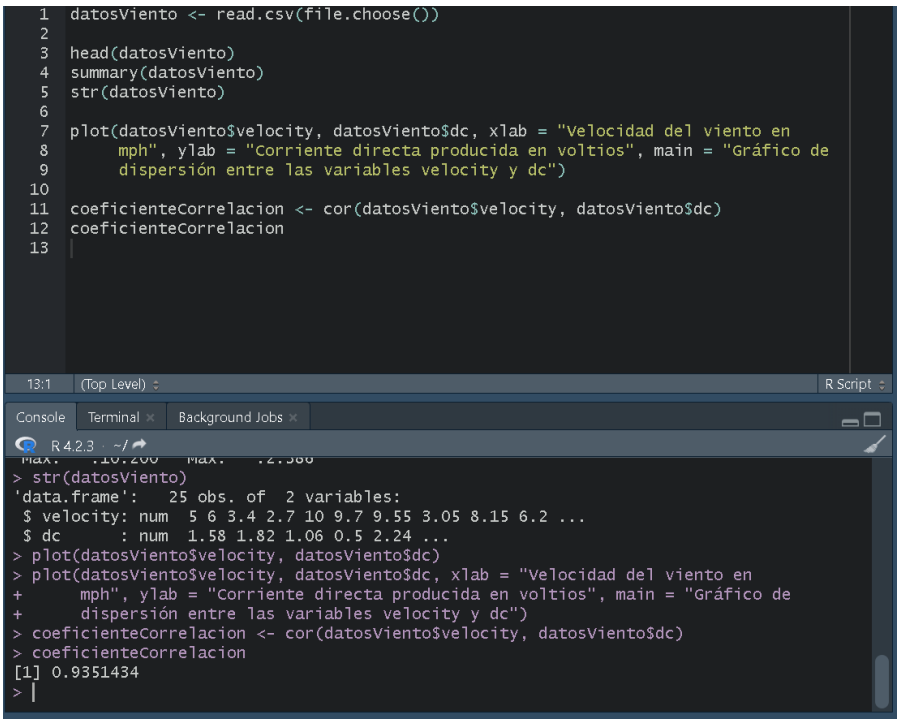
Después de importar los datos del archivo “vientos.csv”, creamos un gráfico de dispersión para visualizar gráficamente la relación entre las variables “velocity” y “dc”.



Del gráfico podemos estimar ciertas conclusiones, como una posible correlación positiva (puntos ascendiendo de izquierda a derecha). También se observa que los puntos están algo separados de la recta imaginaria de relación lineal, lo que indica que la relación no es perfecta.

Ahora procedemos a calcular el coeficiente de correlación de Pearson (nos indicará si los puntos tienen una tendencia a disponerse alineadamente, excluyendo rectas horizontales y verticales).
Después de calcularlo utilizando R-studio, obtenemos un r = 0.9351434. Como es un número cercano a 1, nos indica que existe una alta correlación (casi perfecta) entre las dos variables.

Analizando ambos enfoques, podemos concluir que ambas variables presentan una relación lineal alta y positiva (velocity vs dc). A continuación, se colocará el código utilizado.



```
datosViento <- read.csv(file.choose())

head(datosViento)
summary(datosViento)
str(datosViento)

plot(datosViento$velocity, datosViento$dc, xlab = "Velocidad del viento en
     mph", ylab = "Corriente directa producida en voltios", main = "Gráfico de
     dispersión entre las variables velocity y dc")

coeficienteCorrelacion <- cor(datosViento$velocity, datosViento$dc)
coeficienteCorrelacion
```

```
> str(datosViento)
'data.frame':   25 obs. of  2 variables:
 $ velocity: num  5 6 3.4 2.7 10 9.7 9.55 3.05 8.15 6.2 ...
 $ dc      : num  1.58 1.82 1.06 0.5 2.24 ...
> plot(datosViento$velocity, datosViento$dc)
> plot(datosViento$velocity, datosViento$dc, xlab = "Velocidad del viento en
+      mph", ylab = "Corriente directa producida en voltios", main = "Gráfico de
+      dispersión entre las variables velocity y dc")
> coeficienteCorrelacion <- cor(datosViento$velocity, datosViento$dc)
> coeficienteCorrelacion
[1] 0.9351434
>
```

Ítem b:

Estimaremos el modelo de regresión utilizando R-studio, obteniendo la siguiente información:

```
> modeloRegresion <- lm(dc ~ velocity, data = datosViento)
> summary(modeloRegresion)

Call:
lm(formula = dc ~ velocity, data = datosViento)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59869 -0.14099  0.06059  0.17262  0.32184

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.13088    0.12599   1.039     0.31
velocity     0.24115    0.01905  12.659 7.55e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2361 on 23 degrees of freedom
Multiple R-squared:  0.8745,    Adjusted R-squared:  0.869
F-statistic: 160.3 on 1 and 23 DF,  p-value: 7.546e-12

>
```

A continuación, interpretaremos los coeficientes de regresión estimados:

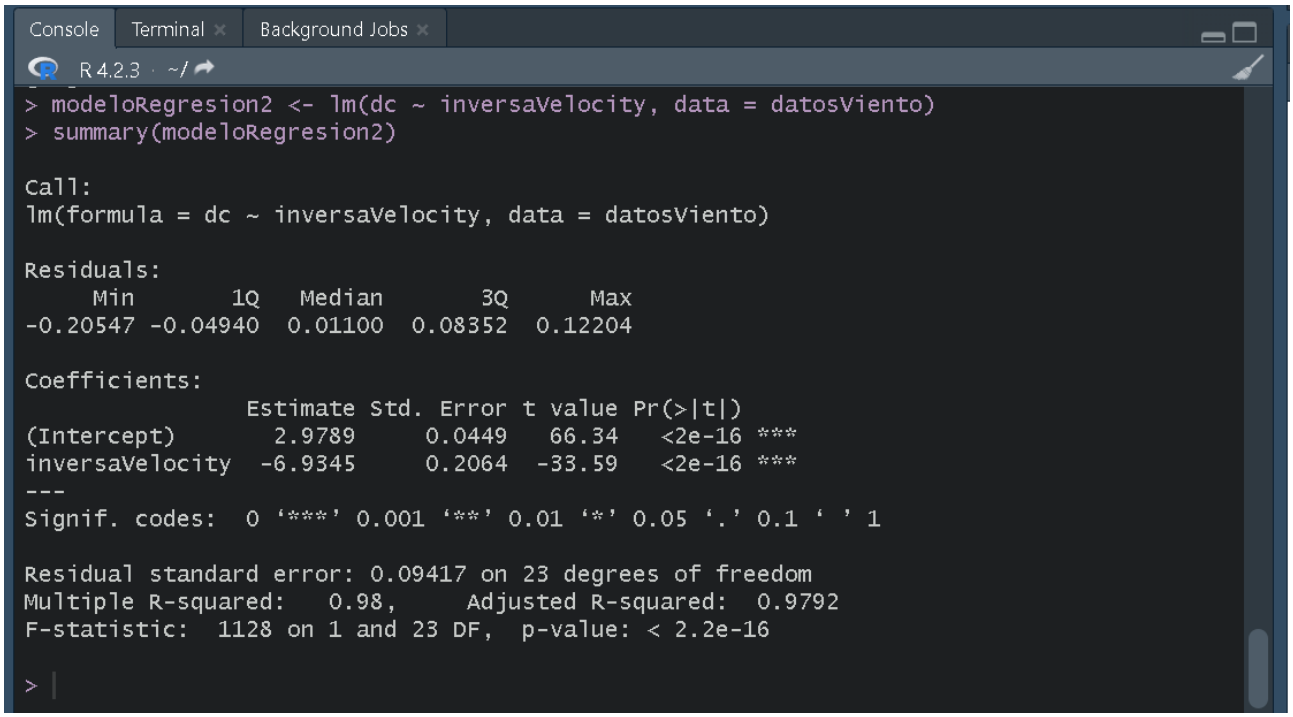
Intercepto: No es correcto interpretarlo.
Pendiente: Por cada 1 mph de velocidad que aumenta el viento, se estima que la corriente directa producida aumenta en 0.241115 voltios.

A continuación, se colocará el código utilizado.

```
14
15  modeloRegresion <- lm(dc ~ velocity, data = datosViento)
16  summary(modeloRegresion)
17
```

Ítem c:

Estimamos el modelo de regresión pedido utilizando otra variable, que contiene la información de velocity a la inversa (1/velocity). Obtenemos los siguientes resultados:



Podemos observar que el coeficiente de determinación R^2 en este modelo es 0.98 que es mayor al del modelo hallado anteriormente (0.8745).

Por tanto, podemos concluir que el modelo más adecuado para predecir la corriente producida por el molino es el que involucra a la inversa de la velocidad del viento como variable independiente.

A continuación, se colocará el código utilizado.

```
18
19  inversaVelocity <- 1/datosViento$velocity
20  datosViento$velocity
21  inversaVelocity
22
23  modeloRegresion2 <- lm(dc ~ inversaVelocity, data = datosViento)
24  summary(modeloRegresion2)
25
```

Ítem d:

El modelo elegido fue: yEstimado = 2.9789 – 6.9345(x)

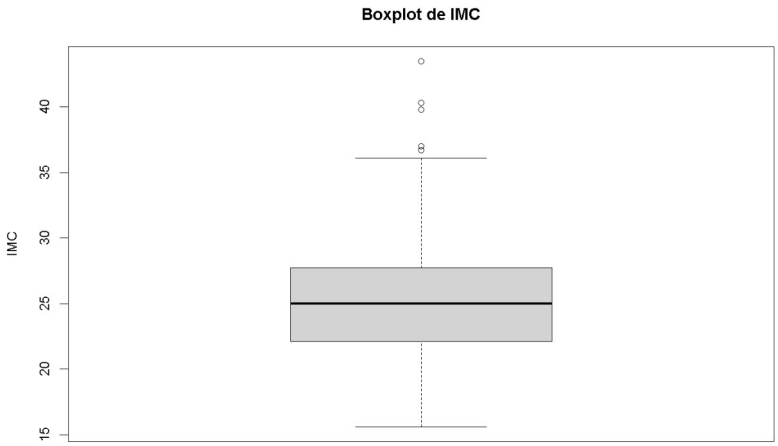
Reemplazando el dato en x tenemos: yEstimado = 2.9789 – 6.9345(1/6) = 1.8232

Por tanto, la corriente estimada producida por el molino con una velocidad del viento de 6 mph es 1.8232 voltios aproximadamente.

Pregunta 2:

Ítem a:

Primero importaremos los datos del archivo “datos_bmi.csv”. Luego, crearemos un diagrama de cajas (boxplot) para visualizar los datos del IMC.



Observamos que todos los datos atípicos se encuentran en la parte superior del boxplot, utilizando R-studio vemos que el límite superior es 36.1. Esto nos dice que todos los valores de IMC mayores a 36.1 serán considerados atípicos.

Luego, calculamos la cantidad de personas que tienen un IMC atípico en nuestra muestra (utilizando R), que resulta en 5 (podríamos verlo en el boxplot, pero de esta manera nos aseguramos).
Ya que hay 5 personas con IMC atípico hay 306 personas con un IMC normal. Nos piden la probabilidad de que exactamente 2 de las 12 personas elegidas al azar y sin reemplazo tengan un IMC atípico. Para responder esto podemos usar combinatorias (a mano o en R):

x: número de personas con IMC atípico seleccionadas
Piden: x = 2 → 2 atípicas y 10 normales (las combinatorias se observan en las imágenes)

Por tanto, la probabilidad de que exactamente dos de los 12 seleccionados al azar y sin reemplazo tengan un IMC que se considere atípico entre todos los individuos del estudio es 0.0124.

A continuación, se colocará el código utilizado.

```
datosObesidad <- read.csv(file.choose())

head(datosObesidad)
summary(datosObesidad)
str(datosObesidad)

boxplot(datosObesidad$bmi, main = "Boxplot de IMC", ylab = "IMC")
stats <- boxplot.stats(datosObesidad$bmi)
stats
limiteInferior <- stats$stats[1]
limiteInferior
limiteSuperior <- stats$stats[5]
limiteSuperior

individuosAtipicos <- sum(datosObesidad$bmi > limiteSuperior)
individuosAtipicos

probabilidad <- choose(5, 2)*choose(306, 10)/choose(311, 12)
probabilidad
```

```
> limiteSuperior <- stats$stats[5]
> limiteSuperior
[1] 36.1
> individuosAtipicos <- sum(datosObesidad$bmi > limiteSuperior)
> individuosAtipicos
[1] 5
> probabilidad <- choose(5, 2)*choose(306, 10)/choose(311, 12)
> probabilidad
[1] 0.01240076
```

Ítem b:

Primero, utilizamos R-studio para procesar los datos y determinar que cantidad de hombres y mujeres hay en la muestra. Obtenemos 185 hombres y 126 mujeres.

Ahora, definimos:

Evento: seleccionar al azar y sin reemplazo 8 individuos de la muestra
x: número de mujeres seleccionados en la muestra
$R_x = \{0; 1; 2; 3; 4; 5; 6; 7; 8\}$

Nos piden: $x > 4$ (que seleccionemos más mujeres que hombres)

Para calcularlo utilizaremos combinatorias: (5M y 3H) o (6M y 2H) o (7M y 1H) o (8M y 0H)

Haciendo los cálculos en R obtenemos 0.1787

Por tanto, concluimos que la probabilidad de que se hayan seleccionado más mujeres que hombres es de 0.1787 aproximadamente.

A continuación, se colocará el código utilizado.

```
44  probabilidad <- choose(5, 2)*choose(306, 10)/choose(311, 12)
45  probabilidad
46
47
48  frecuenciaSexo <- table(datosObesidad$sexo)
49  print(frecuenciaSexo)
50
51  probabilidadMujeres <- (choose(126, 5)*choose(185, 3) + choose(126, 6)*
52                            choose(185, 2) + choose(126, 7)*choose(185, 1) +
53                            choose(126, 8)*choose(185, 0))/choose(311, 8)
54  probabilidadMujeres
55
55:1  (Top Level)

Console  Terminal  Background Jobs
R 4.2.3 · ~/
> print(frecuenciaSexo)

Hombre  Mujer
   185    126
> probabilidadMujeres <- (choose(126, 5)*choose(185, 3) + choose(126, 6)*
+                           choose(185, 2) + choose(126, 7)*choose(185, 1) +
+                           choose(126, 8)*choose(185, 0))/choose(311, 8)
> probabilidadMujeres
[1] 0.1787214
>
```

Ítem c:

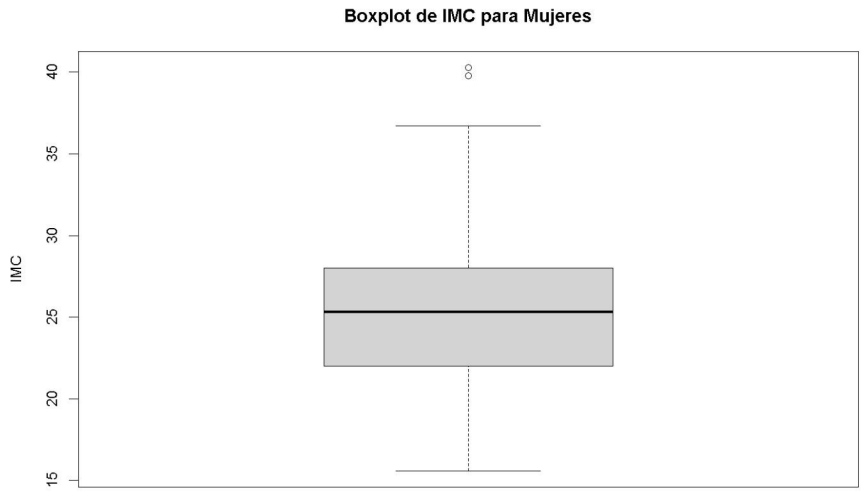
Primero, definimos:

A: se ha seleccionado una mujer con IMC que sea considerado atípico entre las mujeres del estudio
B: se ha seleccionado un hombre con IMC “ ” hombres del estudio

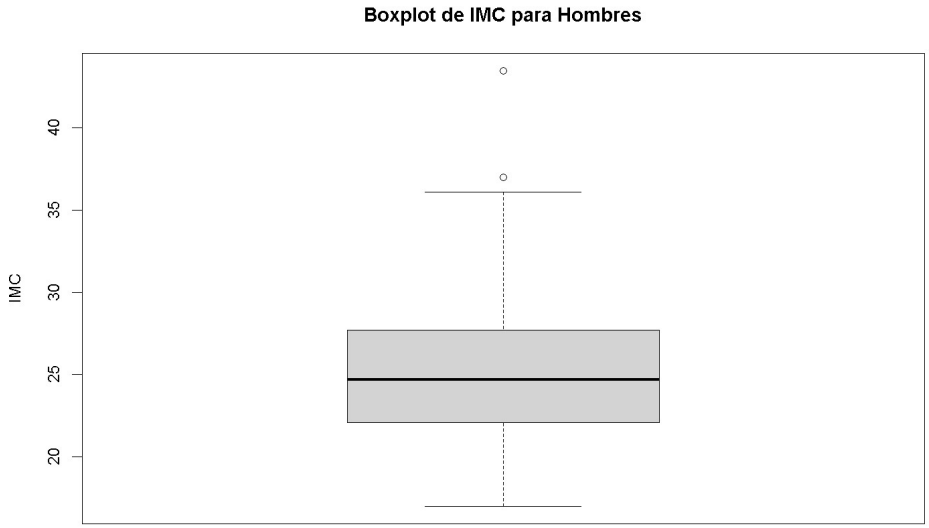
Nos piden la probabilidad de que suceda A o B, esto es:

$P(A) + P(B) – P(A \cap B)$

Ahora elaboramos un boxplot para mujeres y utilizando el mismo método que en el ítem a calculamos el número de atípicos entre las mujeres obteniendo 2.



Realizando un procedimiento similar con los hombres obtenemos 2 igualmente.



Ahora utilizaremos combinatorias para calcular todo lo que nos pidan:

Definimos:

Evento: se selecciona al azar y sin reemplazo 10 individuos de la muestra
x: número de mujeres seleccionadas con IMC atípico para las mujeres
y: número de hombres seleccionados con IMC atípico para las hombres

Calculamos (usando R) primero x=1:

1 atípico mujer y 9 restante
2C1*309C9/311C10 = 0.06244166

Calculamos y = 1:

1 atípico hombre y 9 restante
2C1 * 309C9/311C10 = 0.06244166

Calculamos la intersección:

1 atípico hombre y 1 atípico mujer y 8 restantes
2C1*2C1*307C8/311C10 = 0.003542901

Ahora solo debemos aplicar: $P(A) + P(B) - P(A \cap B)$

0.06244166 + 0.06244166 – 0.003542901 = 0.1213404

Por tanto, concluimos que la probabilidad de que se haya seleccionado una mujer con IMC que sea considerado atípico entre las mujeres del estudio o que se haya seleccionado un hombre con IMC considerado atípico entre loshombres del estudio es de 0.1213 aproximadamente.

A continuación, se colocará el código utilizado.

Boxplot Mujer

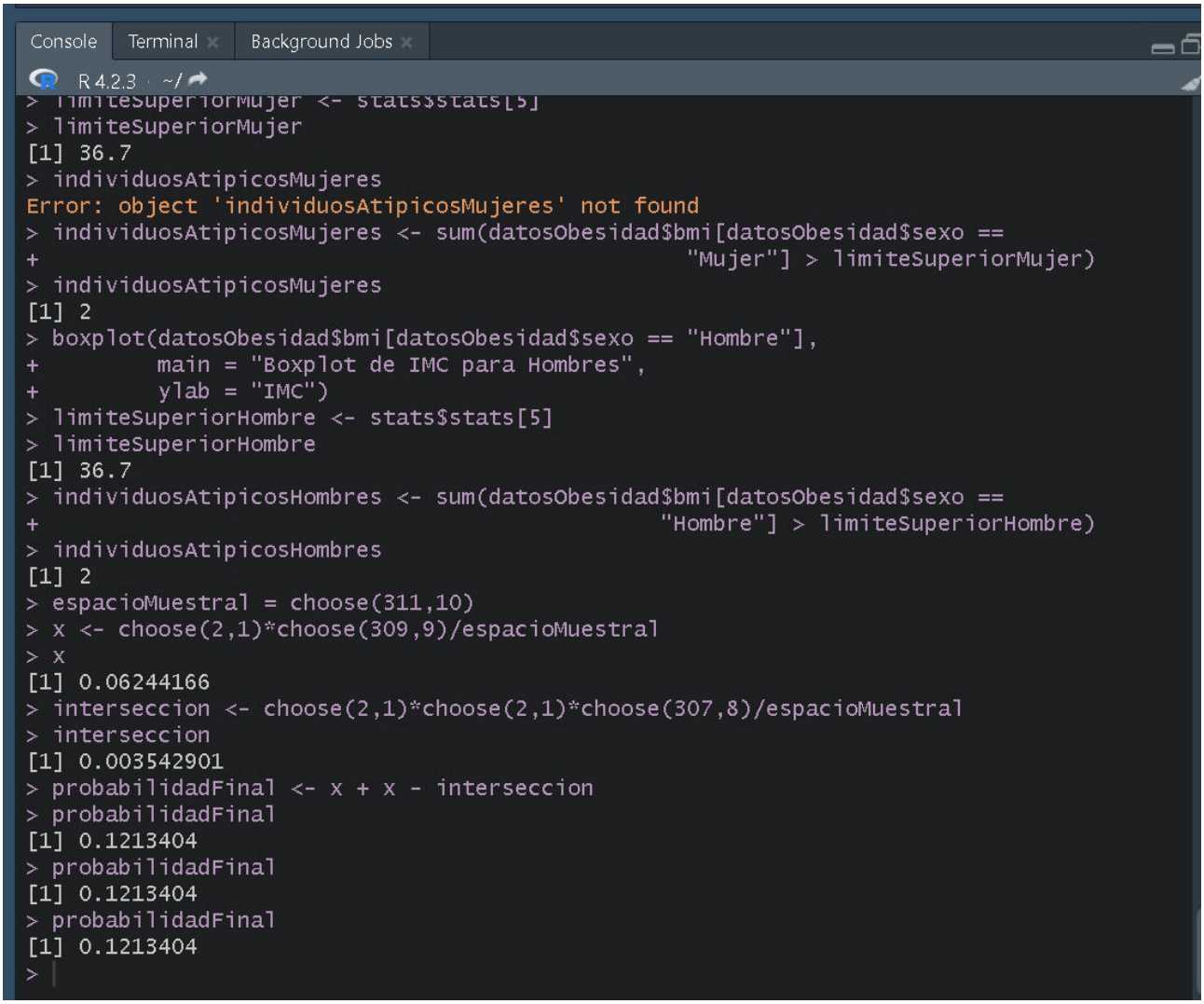
```
boxplot(datosObesidad$bmi[datosObesidad$sexo == "Mujer"],
        main = "Boxplot de IMC para Mujeres",
        ylab = "IMC")
stats <- boxplot.stats(datosObesidad$bmi[datosObesidad$sexo == "Mujer"])
limiteSuperiorMujer <- stats$stats[5]
limiteSuperiorMujer
individuosAtipicosMujeres <- sum(datosObesidad$bmi[datosObesidad$sexo ==
                                                   "Mujer"] > limiteSuperiorMujer)
individuosAtipicosMujeres
```

Boxplot Hombre

```
boxplot(datosObesidad$bmi[datosObesidad$sexo == "Hombre"],
        main = "Boxplot de IMC para Hombres",
        ylab = "IMC")
stats <- boxplot.stats(datosObesidad$bmi[datosObesidad$sexo == "Hombre"])
limiteSuperiorHombre <- stats$stats[5]
limiteSuperiorHombre
individuosAtipicosHombres <- sum(datosObesidad$bmi[datosObesidad$sexo ==
                                                   "Hombre"] > limiteSuperiorHombre)
individuosAtipicosHombres
```

Cálculos de combina-torias

```
espacioMuestral = choose(311,10)
x <- choose(2,1)*choose(309,9)/espacioMuestral
x
interseccion <- choose(2,1)*choose(2,1)*choose(307,8)/espacioMuestral
interseccion
probabilidadFinal <- x + x - interseccion
probabilidadFinal
```


Console
Terminal
Background Jobs
R 4.2.3 · ~/
> limiteSuperiorMujer <- stats$stats[5]
> limiteSuperiorMujer
[1] 36.7
> individuosAtipicosMujeres
Error: object 'individuosAtipicosMujeres' not found
> individuosAtipicosMujeres <- sum(datosObesidad$bmi[datosObesidad$sexo ==
+ "Mujer"] > limiteSuperiorMujer)
> individuosAtipicosMujeres
[1] 2
> boxplot(datosObesidad$bmi[datosObesidad$sexo == "Hombre"],
+ main = "Boxplot de IMC para Hombres",
+ ylab = "IMC")
> limiteSuperiorHombre <- stats$stats[5]
> limiteSuperiorHombre
[1] 36.7
> individuosAtipicosHombres <- sum(datosObesidad$bmi[datosObesidad$sexo ==
+ "Hombre"] > limiteSuperiorHombre)
> individuosAtipicosHombres
[1] 2
> espacioMuestral = choose(311,10)
> x <- choose(2,1)*choose(309,9)/espacioMuestral
> x
[1] 0.06244166
> interseccion <- choose(2,1)*choose(2,1)*choose(307,8)/espacioMuestral
> interseccion
[1] 0.003542901
> probabilidadFinal <- x + x - interseccion
> probabilidadFinal
[1] 0.1213404
> probabilidadFinal
[1] 0.1213404
> probabilidadFinal
[1] 0.1213404
>
Resultados ejecutados que se mencionan en la solución