PEC1 Manuel Rojas García

2023-10-08

Con la libraria ya instalada, seleccionamos el dataset con el que vamos a trabajar y su ruta

```
library(readr)
file.choose()
```

```
## [1] "C:\\Users\\Manuel\\Desktop\\UOC\\SEMESTRE 3 (Sep 2023 - Feb 2024)\\Estadistica\\PEC 1\\archive\\
```

Creamos una varible donde almacenaremos la ruta donde se encuentra el csv

```
ruta_csv <- "C:\\Users\\Manuel\\Desktop\\UOC\\SEMESTRE 3 (Sep 2023 - Feb 2024)\\Estadistica\\PEC 1\\arc
```

Creamos una nueva variable donde almacenamos el dataset. Imprimimos por pantalla la variable con 100 valores.

```
villanos <- read_csv(ruta_csv)
```

```
## Rows: 100 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Name, Main_Ft_Apperance, Gender, Human/Other, Type
## dbl (5): IGN_Rank, No_Feature_Films, Rating, Award_Wins, Nominations
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(villanos, 100)
```

```
## # A tibble: 100 x 10
##    Name          IGN_Rank Main_Ft_Apperance   No_Feature_Films Rating Award_Wins
##    <chr>            <dbl> <chr>                          <dbl>  <dbl>      <dbl>
##  1 MADOK              100 Ant-Man and the Wa~                1   0.82          0
##  2 Fin Fang Foom       99 The invincible iro~                1   0.46          0
##  3 Mastermind          98 X-men:The animated~                0   0.94          1
##  4 Violator            97 Spawn                              1   0.36          2
##  5 Despero             96 Superman/Batman:Pu~                1   0.66          0
##  6 Omega Red           95 Hulk vs                            2   0.67          0
##  7 Annihilus           94 Hulk and the Agent~                2   0.61          0
##  8 Omni-Man            93 Invincible                         0   0.89          5
##  9 Parallax            92 Green Lantern                      2   0.45          3
## 10 The Adversary       91 Pinocchio                         12   0.73          7
## # i 90 more rows
## # i 4 more variables: Nominations <dbl>, Gender <chr>, `Human/Other` <chr>,
## #   Type <chr>
```

Pregunta 1:

Supercebolla sabe que hay que entrenar mucho y aparecer en muchas películas / series para ser un verdadero superhéroe. Se requiere analizar los siguientes puntos para ayudarle: a) ¿De qué tipo es la variable No_Feature_Film? Haga un resumen numérico (media, mediana, cuartiles, desviación típica, mínimo y máximo) de dicha variable. (1 punto).

La variable es una variable cuantitativa discreta, ya que solo contiene valores enteros.

Calculamos la media

```
media <- mean(villanos$No_Feature_Films)

media
```

```
## [1] 4.85
```

Calculamos la mediana

```
mediana <- median(villanos$No_Feature_Films)
mediana
```

```
## [1] 3
```

Calculamos los dos cuartines Q1, Q3 ya que el Q2 es la mediana.

```
Q1 <- quantile(villanos$No_Feature_Films,0.25)
Q1
```

```
## 25%
##   1
```

```
Q3 <- quantile(villanos$No_Feature_Films,0.75)
Q3
```

```
## 75%
##   6
```

Calculamos la desviación típica.

```
desviacion <- sd(villanos$No_Feature_Films)
desviacion
```

```
## [1] 6.418683
```

Calculamos el mínimo y máximo.

```
minimo <- min(villanos$No_Feature_Films)
minimo
```

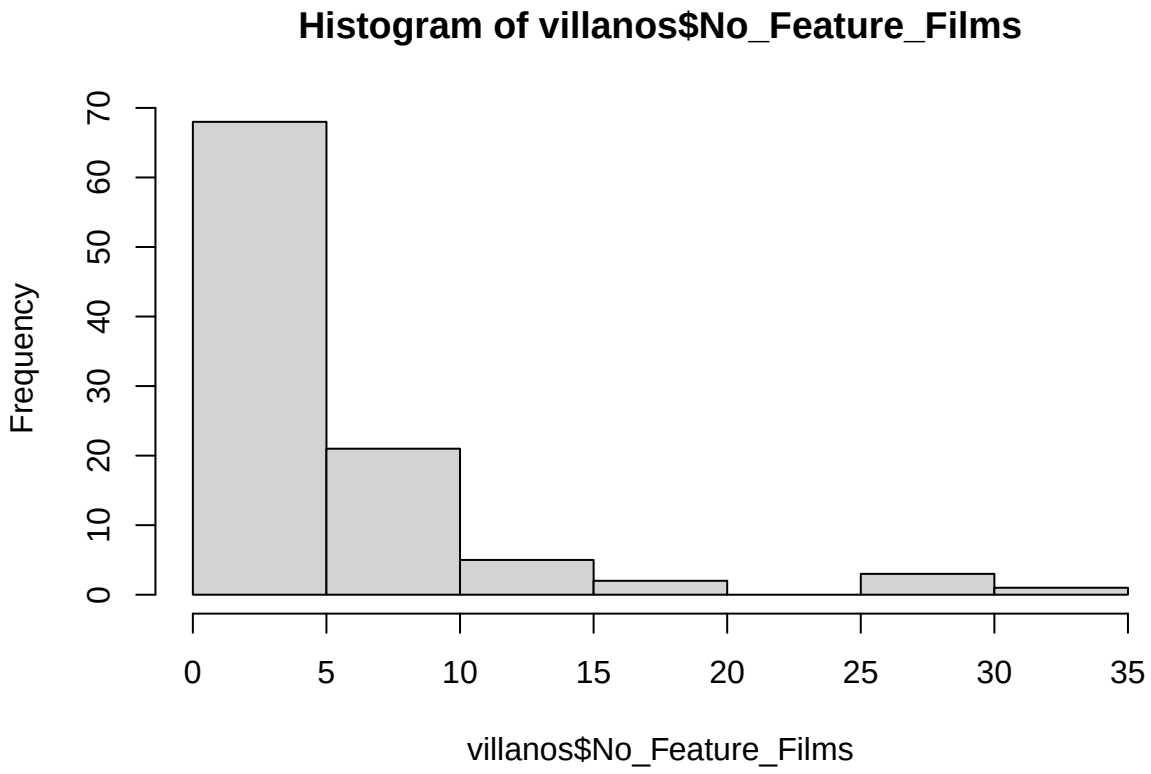
```
## [1] 0
```

```
maximo <- max(villanos$No_Feature_Films)
maximo
```

```
## [1] 34
```

b) Realice un histograma para representar los datos de la variable No_Feature_Films y comente el resultado. (1 punto).

```
hist(villanos$No_Feature_Films)
```



Podemos interpretar que el histograma tiene un pico en los primeros valores y ausencia en uno de los valores entre 20/25. Se podría determinar que tiene una asimetria a la derecha pero los valores de 25/30 son mayores que de 15/20. Habría que determinar si los valores ausentes son valores átipicos.

c) Si se añade un nuevo villano con No_Feature_Films de 50 a la lista original. ¿Qué cambiará más, la media o la mediana? Razona y desarrolla la respuesta. (1.5 puntos)

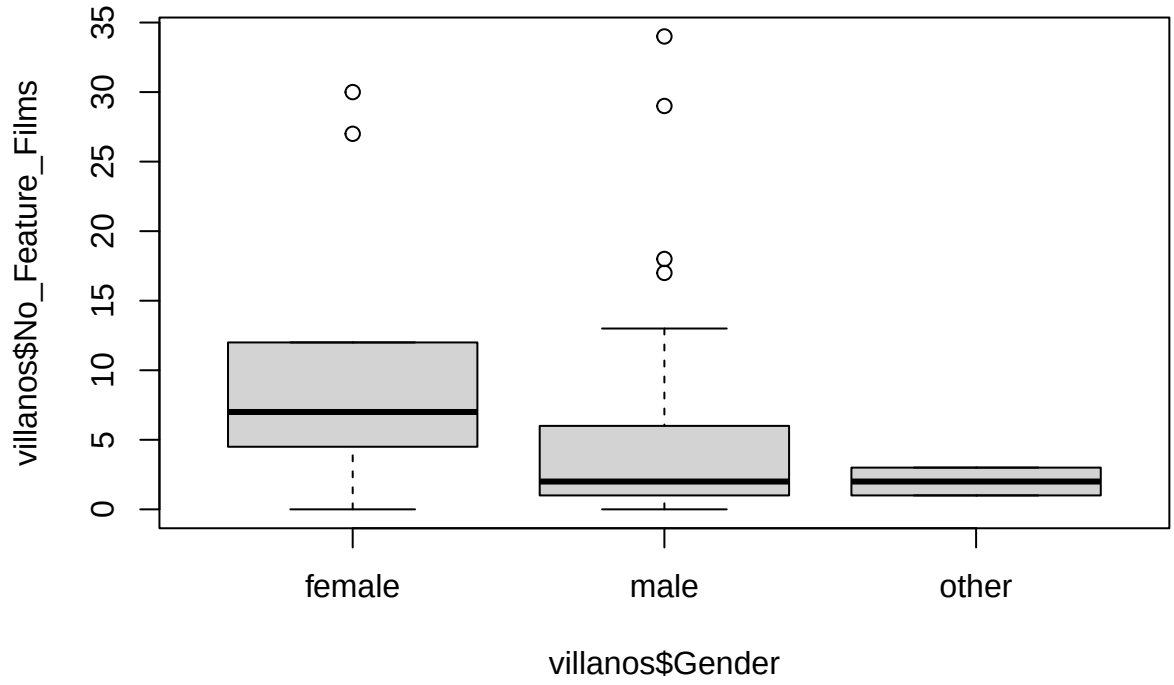
Afectaría sobre todo a la media, ya que se condideraría un valor atípico puesto que es el número máximo es 34. En resumen, afectaría más a la media que a la mediana ya que los valores alejados afectan más a la media y se dice que es poco resistente a los valores extremos.

Pregunta 2:

Supercebolla necesita más información sobre los perfiles de los villanos. Responde a las siguientes preguntas:

a) Realice un boxplot entre la variable "No_Feature_Films" y los distintos grupos de la variable "Gender". Comente el resultado. (1 punto).

```
boxplot(villanos$No_Feature_Films ~ villanos$Gender)
```



Podemos determinar que los valores femininos y masculinos comparten el mismo mínimo. La mediana está muy cerca de Q1 en ambos generos, pero sobre todo en el masculino. El máximo y los valores más atipicos pertenecen al genero másculino. En cuanto a los valores otros, podemos determinar que no tiene valores atipicos, que los valores máximos y mínimos coinciden con los cuartiles y que sus valores son asimétricos. En conceptos generales podemos determinar que el genero femenino tiene la mediana más alta ya que la caja está en una posición más elevada.

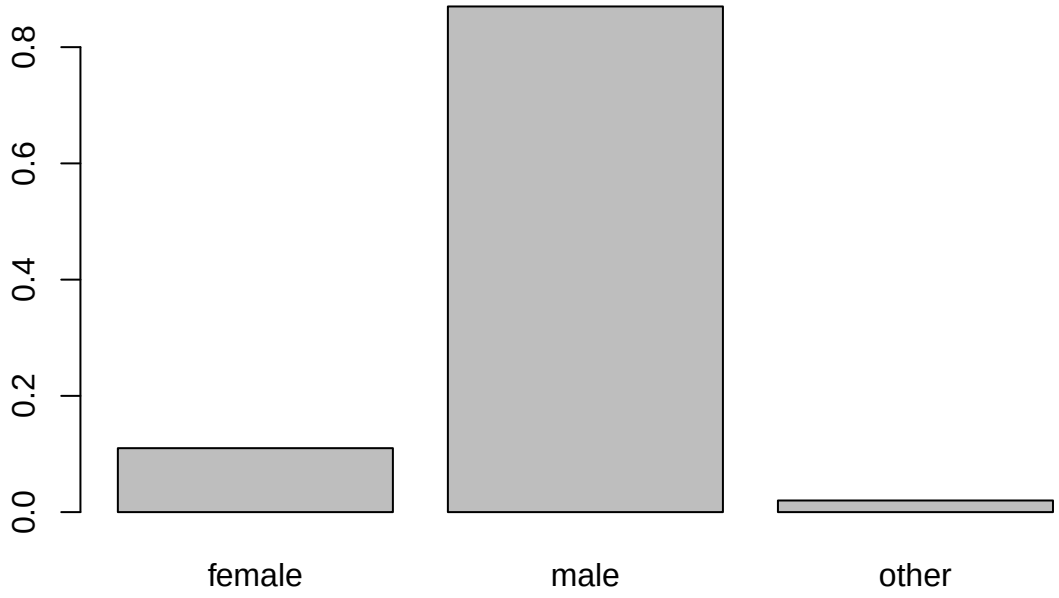
b) Haced una tabla de frecuencias relativas de la variable "Gender". Después usa la tabla para hacer un diagrama de barras. Comentad los resultados. (1 punto).

Calculamos la frecuencia relativa por genero y generamos la tabla.

```
frecuencia <- prop.table(table(villanos$Gender))
frecuencia
```

```
##
## female   male  other
##   0.11   0.87   0.02
```

```
barplot(frecuencia)
```



Podemos determinar que la frecuencia de que un villano sea hombre es muy elevada respecto a los demás valores.

c) ¿Cuáles son los villanos que han ganado más de 80 "Premios de villanos" (variable Award_Winds)? Mostrar en la salida únicamente la variable del nombre del villano (Name), Award_Wins y Gender. (1.5 punto) **fuente https://rpubs.com/hllinas/R_Filtrar_DataFrames**

```
masde80 <- subset(villanos, Award_Wins > 80, select = c("Name", "Award_Wins", "Gender"))
masde80
```

```
## # A tibble: 4 x 3
##   Name        Award_Wins Gender
##   <chr>            <dbl> <chr>
## 1 The Govenor         84 male
## 2 Two-Face           162 male
## 3 Kingpin             81 male
## 4 Joker              121 male
```

Pregunta 3:

Supercebolla ha realizado una encuesta, para entre otras cosas, que los villano/as de Marvel y DC indicaran el año en que decidieron convertirse en villanos por primera vez. Se encuestó en total a 30 villanos de Marvel y DC.

a) ¿Los resultados de esta encuesta son datos de población o datos de muestra? Razona la respuesta.

Teniendo en cuenta que podemos acceder al número total de villanos en el dataset y solo vamos a seleccionar 30, los resultados deben ser datos de muestra sobre un subconjunto de la población.

b) ¿Cuál es la variable de estudio de la encuesta? ¿Qué tipo de variable es, cuantitativa o cualitativa?

Será una variable cuantitativa, ya que se refiere a una cantidad numérica entera (el año) que puede medirse y cuantificarse.

c) Si se selecciona para otra encuesta 10 comarcas al azar y seleccionamos al azar 3 villanos de estas comarcas a los cuales llamamos por teléfono, ¿qué tipo de muestreo sería?

Al seleccionar 10 comarcas al azar podemos determinar que es un muestreo simple (aún sin saber su proceso de selección), pero al seleccionar 3 villanos de dichas comarcas estamos haciendo un estrato, por consiguiente, sería un muestreo de estratificación. Como indica los apuntes "La muestra se obtiene seleccionando una muestra aleatoria dentro de cada estrato". Pero después de como se ha resuelto un ejercicio del moodle debemos considerar que es un muestreo por conglomerado, aún sin conocer si los conglomerados son similares los unos a los otros.