PEC 1

UOC

NOMBRE:

Borja Villena Pardo

Introducción

Supercebolla, el superhéroe con más capas, ha decidido hacer un estudio de todos de villanos del mundo Marvel y de DC, para saber bien a lo que se enfrentará en sus primeras aventuras.

Los datos para ese estudio proceden del siguiente enlace de Kaggle. El fichero para realizar la PEC 1 es *"data_pac1"* y lo encontraréis en formato csv. Esta base de datos contiene información relacionada con todos los villanos del mundo Marvel y de DC, con variables relacionadas con el género del villano, índice de audiencia, ranking de maldad, tipo de ser o nominaciones.

Las variables que se encuentran en el dataset son las siguientes:

- *Name* : Nombre del villano/a.
- *Gender* : Género del villano/a (Femenino/Masculino/Otro).
- *IGN_Rank* : Posición en el ranking de villano/as.
- *Main_Ft_Apperance* : Nombre de la película/serie con aparición
- *No_Feature_Film* : Número de películas/series protagonizadas
- *Rating* : Índice de audiencia.
- *Award_Wins* : Premios de villano/a ganados.
- *Nominations* : Nominaciones a villano/a, sin incluir las nominaciones ganadoras.
- *Human.Other* : Si el villano/a es ser humano o no.
- *Type* : Tipo de superpoder.

Os puede ser útil consultar el siguiente material:

- Manuales de R
- Actividades Resueltas del Reto 1 (Estadística Descriptiva y Muestreo)

Hay que entregar la práctica en fichero pdf o html (exportando el resultado final a pdf o html por ejemplo). Se recomienda generar el informe con Rmarkdown que genera automáticamente el pdf/html a entregar

Pregunta 1 (resolver con R). (3.5 puntos)

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).

```
#Espacio para solución
class(data_pac1$No_Feature_Films)
```

```
## [1] "integer"
```

```
cat("La variable 'No_Feature_Film' es de clase", class(data_pac1$No_Feature_Films),"\n",
    "y concretamente es una variable cuantitativa discreta.", "\n")
```

```
## La variable 'No_Feature_Film' es de clase integer
##  y concretamente es una variable cuantitativa discreta.
```

```
cat("Resumen:", "\n")
```

```
## Resumen:
```

```
print(summary(data_pac1$No_Feature_Films))
```

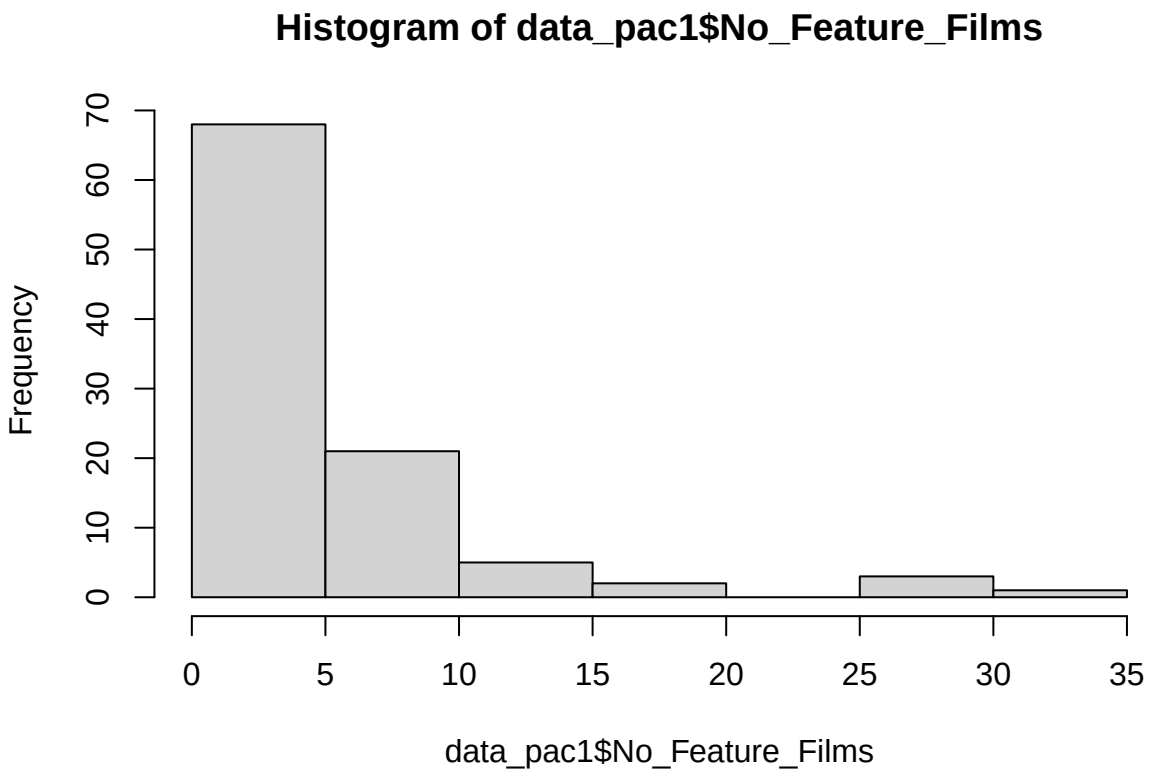
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    3.00    4.85    6.00   34.00
```

```
cat("Desviación típica =", sd(data_pac1$No_Feature_Films))
```

```
## Desviación típica = 6.418683
```

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

```
#Espacio para solución
hist(data_pac1$No_Feature_Films)
```




```
cat("Histograma asimétrico con cola hacia la derecha, se detecta un claro","\n",
    "ejemplo de caso unimodal en el intervalo de 0 a 5 películas/series","\n",
    "protagonizadas, y una clase vacía de 20 a 25 películas/series protagonizadas.")
```

```
## Histograma asimétrico con cola hacia la derecha, se detecta un claro
##  ejemplo de caso unimodal en el intervalo de 0 a 5 películas/series
##  protagonizadas, y una clase vacía de 20 a 25 películas/series protagonizadas.
```

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)

```
#Espacio para solución

data_pac1$No_Feature_Films <- as.numeric(data_pac1$No_Feature_Films)
data_pac1$Name <- as.character(data_pac1$Name)
data_pac1$Main_Ft_Apperance <- as.character(data_pac1$IGN_Rank)
nuevo_villano <- c("Nuevo villano",101,"Nuevo ", 50, 0.5, 0, 0, "male", "other", "normal")
data_pac1_new <- rbind(data_pac1, nuevo_villano)
data_pac1_new$No_Feature_Films <- as.numeric(data_pac1_new$No_Feature_Films)

cat("Mediana original =", median(data_pac1$No_Feature_Films), "\n")
```

```
## Mediana original = 3
```

```r
cat("Media original =", mean(data_pac1$No_Feature_Films), "\n")
```

```
## Media original = 4.85
```

```r
cat("Mediana después de agregar el nuevo villano =", median(data_pac1_new$No_Feature_Films), "\n")
```

```
## Mediana después de agregar el nuevo villano = 3
```

```r
cat("Media después de agregar el nuevo villano =", mean(data_pac1_new$No_Feature_Films), "\n")
```

```
## Media después de agregar el nuevo villano = 5.29703
```

```r
cat("Como hemos podido comprobar, concretamente en el caso planteado, la mediana", "\n",
    "se mantendría igual con un valor de 3, mientras que el valor de la medía si", "\n",
    "se vería incrementado en 44 décimas apróximadamente con respecto a la media","\n",
    "orginal. Por lo que la media cambia más que la mediana. Si observamos el ","\n",
    "histograma impreso en el apartado anterior de esta pregunta, podemos observar","\n",
    "que el intervalo de 0-5 presenta un caso unimodal, por lo que podemos intuir","\n",
    "que el hecho de que la moda no se vea afectada, tras añadir un nuevo villano,", "\n",
    "puede ser posible.")
```

```
## Como hemos podido comprobar, concretamente en el caso planteado, la mediana
##  se mantendría igual con un valor de 3, mientras que el valor de la medía si
##  se vería incrementado en 44 décimas apróximadamente con respecto a la media
##  orginal. Por lo que la media cambia más que la mediana. Si observamos el
##  histograma impreso en el apartado anterior de esta pregunta, podemos observar
##  que el intervalo de 0-5 presenta un caso unimodal, por lo que podemos intuir
##  que el hecho de que la moda no se vea afectada, tras añadir un nuevo villano,
##  puede ser posible.
```

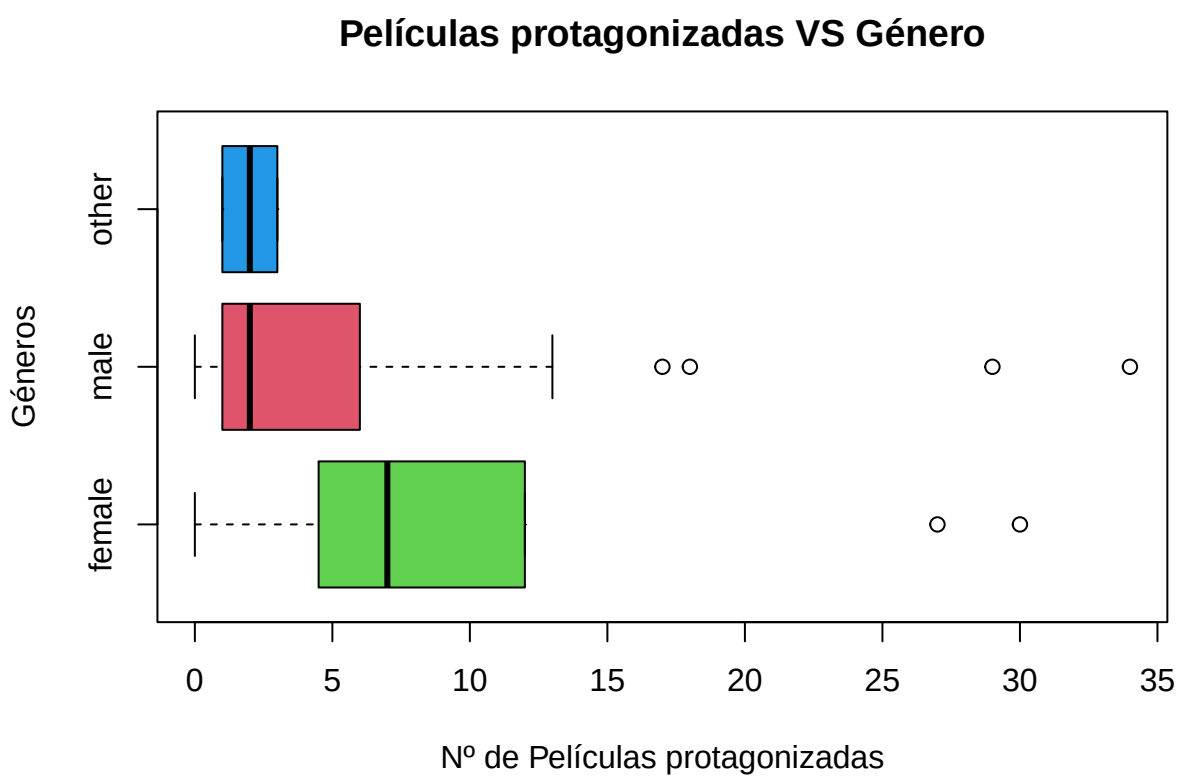
Pregunta 2 (resolver con R). (3.5 puntos)

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).

```r
#Espacio para solución

boxplot(data_pac1$No_Feature_Films ~ data_pac1$Gender, main="Películas protagonizadas VS Género",
        xlab= "Nº de Películas protagonizadas", ylab= "Géneros", col=c(123,234,180), horizontal=T)
```




```
cat("Analizando el boxplot, a primera vista nos damos cuenta que hay datos","\n",
    "atípicos tanto en el género 'male' como en 'female'. Por otro lado, observamos", "\n",
    "que sólo existe simetría con respecto a la mediana en el género 'other', a su vez","\n",
    "vemos una cola a la derecha en 'male' y una cola hacia la izquierda en","\n",
    "'female'. Comparando los diagramas de caja entre el género 'male' y género","\n",
    "'female', podemos intuir que los villanos masculinos protagonizan un porcentaje","\n",
    "mayor de películas que los villanos femeninos.")
```

```
## Analizando el boxplot, a primera vista nos damos cuenta que hay datos
##  atípicos tanto en el género 'male' como en 'female'. Por otro lado, observamos
##  que sólo existe simetría con respecto a la mediana en el género 'other', a su vez
##  vemos una cola a la derecha en 'male' y una cola hacia la izquierda en
##  'female'. Comparando los diagramas de caja entre el género 'male' y género
##  'female', podemos intuir que los villanos masculinos protagonizan un porcentaje
##  mayor de películas que los villanos femeninos.
```

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).

```
#Espacio para solución
cat("Las frecuencias absolutas de 'Género' son: ")
```

```
## Las frecuencias absolutas de 'Género' son:
```

```r
table(data_pac1$Gender)
```

```
##
## female   male  other
##     11     87      2
```

```r
cat("Las frecuencias relativas de 'Género' son: ")
```

```
## Las frecuencias relativas de 'Género' son:
```

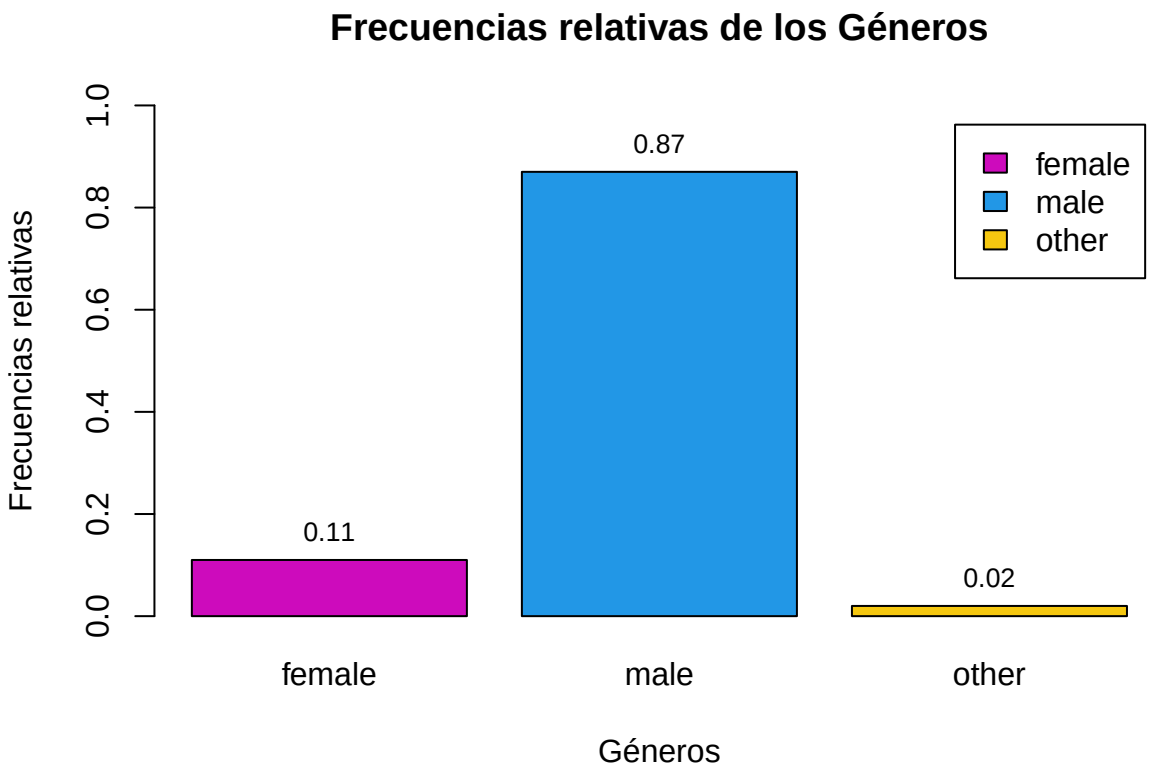
```r
prop.table(table(data_pac1$Gender))
```

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

```r
cat("Observamos que el número de variables del género masculino es mucho mayor","\n",
    "que del resto, siendo casi el 90% de la población, frente al 11 % de mujeres y","\n",
    "al 2% del género 'others'.")
```

```
## Observamos que el número de variables del género masculino es mucho mayor
##  que del resto, siendo casi el 90% de la población, frente al 11 % de mujeres y
##  al 2% del género 'others'.
```

```r
tabla <- prop.table(table(data_pac1$Gender))
barra_plot <- barplot(tabla, main = "Frecuencias relativas de los Géneros",
                      xlab = "Géneros", ylab= "Frecuencias relativas",
                      legend = rownames(tabla),ylim = c(0, 1),
                      col = c("30", "100", "175"))
text(x=barra_plot, y=tabla, label = tabla, pos=3, cex=0.8, col="black")
```



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)

```
#Espacio para solución
data_pac1_na <- na.omit(data_pac1)
data_pac1_na[data_pac1_na$Award_Wins > 80, c(1,6,8)]
```

```
##           Name Award_Wins Gender
## 15 The Govenor         84   male
## 89    Two-Face        162   male
## 91     Kingpin         81   male
## 99       Joker        121   male
```

Pregunta 3. (3 puntos)

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.

Los resultados de esta encuesta son datos de muestra ya que se ha encuestado sólo a un grupo del total de la población. En este caso la muestra consta de 30 variables (30 villanos encuestados con una sola respuesta

por cada uno de ellos). En caso contrario, si se hubiese encuestado al total de villanos, entonces si estaríamos hablando de datos de población.

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

La variable de estudio de la encuesta es el dato del año en el que decidió convertirse en villano cada uno de los encuestados. Por lo tanto, estamos hablando de una variable cuantitativa, y más concretamente en este caso se trataría de una variable cuantitativa discreta, puesto que el dato del año no se representa con decimales, sino es un número entero.

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?

Estaríamos realizando un tipo de muestreo por conglomerados, ya que tanto la selección de comarcas y de villanos se han realizado al azar. Hay que tener en cuenta, que si hubiesemos distribuido las comarcas y el número de villanos de manera que la proporción de los villanos de cada comarca fuese similar a la proporción dentro de la población, entonces estaríamos hablando de un muestreo por cuotas, pero no es el caso ya que la selección se ha hecho al azar y no respetando las proporciones.