Práctica 3

Alfredo Robledano Abasolo y Rubén Sierra Serrano

2022-11-24

Problema 1

El esquema de aceptación para comprar lotes que contienen un número grande de baterías consiste en probar no más de 75 baterías seleccionadas al azar y rechazar el lote completo si falla una sola batería. Se supone que la probabilidad de encontrar una que falle es de 0,001.

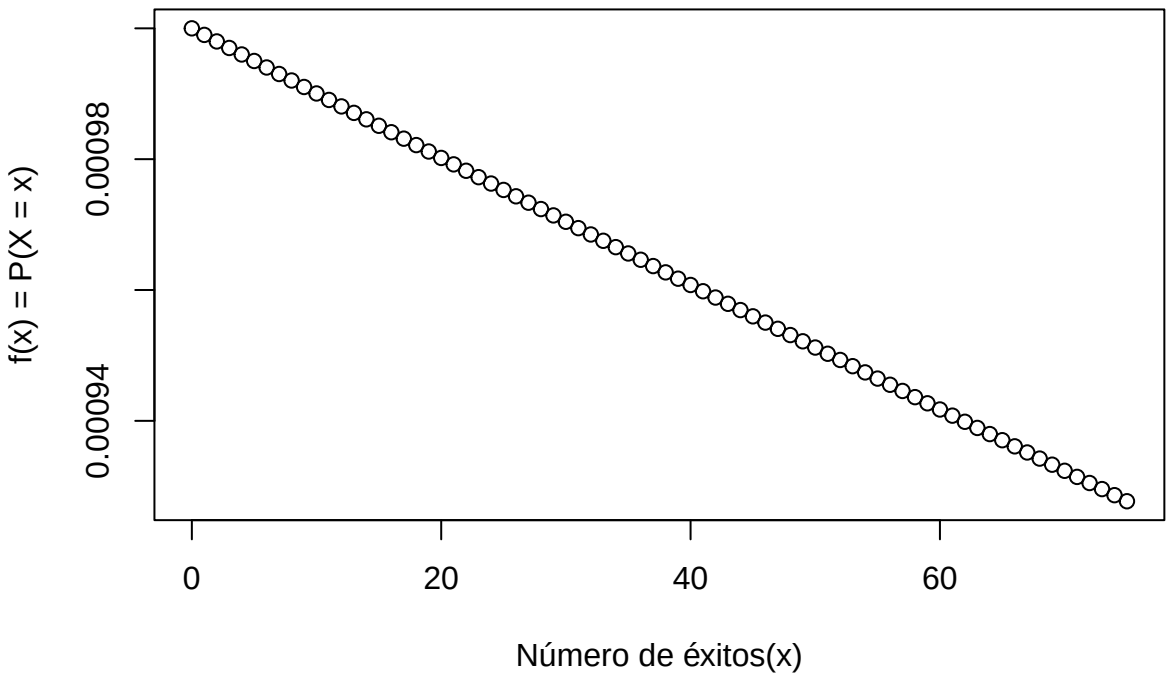
1. ¿Cuál es la probabilidad de que se acepte un lote?. Define la variable y representa la fucnión de masa / densidad correspondiente.
2. ¿Cuál es la probabilidad de que se rechace un lote en 10 pruebas o menos?. Define la variable y representa la fucnión de distribución correspondiente.

Es una variable discreta X = número de baterías defectuosas Se trata de una distribución geómetrica

```
var1 <- dgeom(x = 0:75,
              prob = 0.001,
              log = FALSE)
print(var1)
```

```
##  [1] 0.0010000000 0.0009990000 0.0009980010 0.0009970030 0.0009960060
##  [6] 0.0009950100 0.0009940150 0.0009930210 0.0009920279 0.0009910359
## [11] 0.0009900449 0.0009890548 0.0009880658 0.0009870777 0.0009860906
## [16] 0.0009851045 0.0009841194 0.0009831353 0.0009821522 0.0009811700
## [21] 0.0009801889 0.0009792087 0.0009782295 0.0009772512 0.0009762740
## [26] 0.0009752977 0.0009743224 0.0009733481 0.0009723747 0.0009714024
## [31] 0.0009704310 0.0009694605 0.0009684911 0.0009675226 0.0009665551
## [36] 0.0009655885 0.0009646229 0.0009636583 0.0009626946 0.0009617319
## [41] 0.0009607702 0.0009598094 0.0009588496 0.0009578908 0.0009569329
## [46] 0.0009559760 0.0009550200 0.0009540650 0.0009531109 0.0009521578
## [51] 0.0009512056 0.0009502544 0.0009493042 0.0009483549 0.0009474065
## [56] 0.0009464591 0.0009455126 0.0009445671 0.0009436226 0.0009426789
## [61] 0.0009417363 0.0009407945 0.0009398537 0.0009389139 0.0009379750
## [66] 0.0009370370 0.0009361000 0.0009351639 0.0009342287 0.0009332945
## [71] 0.0009323612 0.0009314288 0.0009304974 0.0009295669 0.0009286373
## [76] 0.0009277087
```

```
plot(0:75, var1,
  main = "Función de probabilidad geómetrica",
  ylab = "f(x) = P(X = x)", xlab = "Número de éxitos(x)")
```



La probabilidad de que el lote sea retirado es de: 0.0009277087 b) P(lote se rechace con 10 pruebas o menos):

```
var2 <- pgeom(10, 0.001)
print(var2)
```

```
## [1] 0.01094516
```

La probabilidad es de: 0.01094516

Problema 2

El esquema de aceptación para comprar lotes que contienen un número grande de baterías consiste en probar no más de 75 baterías seleccionadas al azar y rechazar el lote completo si falla una sola batería. Se supone que la probabilidad de encontrar una que falle es de 0,001.

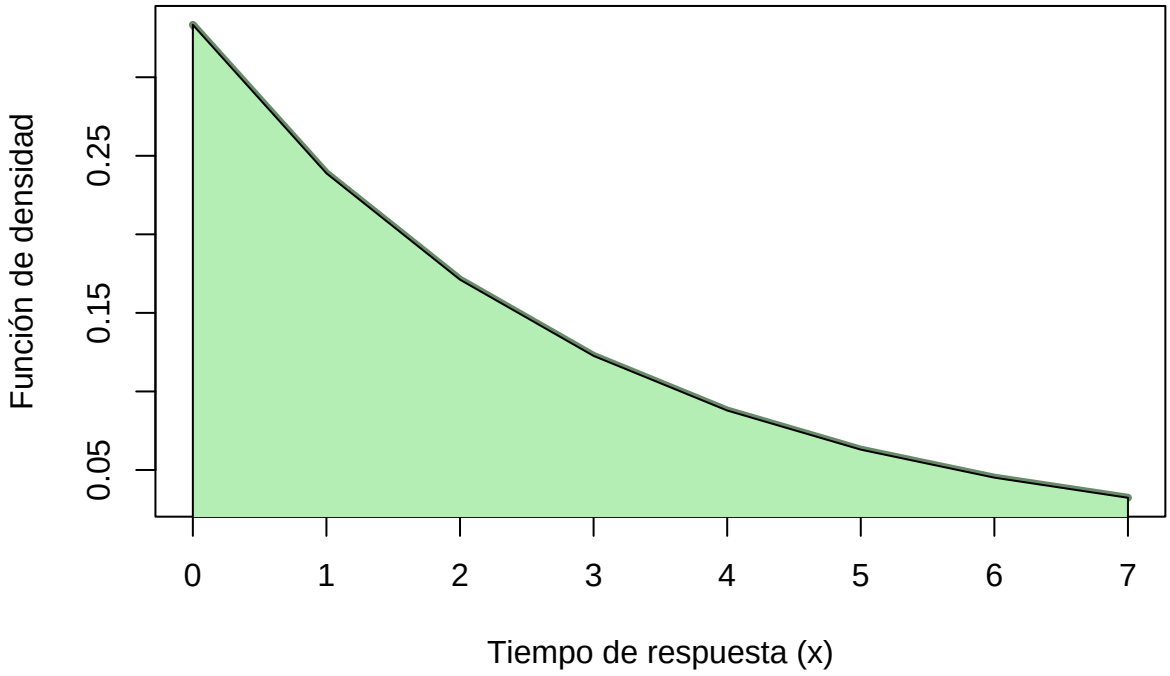
1. ¿Cuál es la probabilidad de que el tiempo de respuesta exceda los 5 segundos? ¿P(X > 5)?

```
print(pexp(5, 1/3, lower.tail = FALSE))
```

```
## [1] 0.1888756
```

2. Representa la función de masa / densidad correspondiente y explica su significado (¿Qué significa este gráfico?).

```r
x <- 0:7
y <- dexp(x, 1/3)
plot(x, y, type = "l",
ylab = "Función de densidad",
xlab = "Tiempo de respuesta (x)", lwd = 4, col = "darkseagreen4")

polygon(c(0, x, 7), c(0, y, 0), col = "darkseagreen2")
```



Significa que conforme aumenta el tiempo de respuesta, menor es la probabilidad de que esta siga aumentando.

Problema 3

Elegid una base de datos en alguno de los recursos web que estudiamos en el primer tema, os los recuerdo: https://archive.ics.uci.edu/ml/datasets.php y https://www.kaggle.com/datasets.Con la base de datos que hayáis elegido, debéis realizar los siguientes puntos o apartados:

1. Definir las variables (y sus tipos) y explicar en qué consiste la base de datos. Primero leemos la base de datos.

```r
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
```

```
datos <- read.csv('ds_salaries.csv')
```

Nos encontramos ante una base de datos que contiene los salarios de empleos agrupables en el campo de computer science.

A continuación, vamos a determinar el tipo de variables que tenemos en la base de datos, pueden ser o bien cuantitativas o bien categóricas, para ello, nos vamos a fijar en el tipo de datos de cada una de las variables, en caso de ser numérica serán cuantitativas y en caso de ser de texto será categórica.

```
summary(datos)
```

```
##        X           work_year    experience_level   employment_type
##  Min.   :  0.0   Min.   :2020   Length:607         Length:607
##  1st Qu.:151.5   1st Qu.:2021   Class :character   Class :character
##  Median :303.0   Median :2022   Mode  :character   Mode  :character
##  Mean   :303.0   Mean   :2021
##  3rd Qu.:454.5   3rd Qu.:2022
##  Max.   :606.0   Max.   :2022
##   job_title             salary         salary_currency    salary_in_usd
##  Length:607         Min.   :    4000   Length:607         Min.   :  2859
##  Class :character   1st Qu.:   70000   Class :character   1st Qu.: 62726
##  Mode  :character   Median :  115000   Mode  :character   Median :101570
##                     Mean   :  324000                      Mean   :112298
##                     3rd Qu.:  165000                      3rd Qu.:150000
##                     Max.   :30400000                      Max.   :600000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:607         Min.   :  0.00   Length:607         Length:607
##  Class :character   1st Qu.: 50.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 70.92
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

Por tanto, las variables enteras (int) son las cuantitativas que son: work-year, salary, salary_in_usd y remote_ratio. Y las variables character (chr) son las categóricas que son: experience_level, employment_type, job_title, salary_currency, employee_residence, company_location y company_size.

2. Realizar una limpieza básica (eliminar NA y comprobar si hay valores extraños). Empleamos la función apply() y aplicamos la función any() para determinar si hay algún valor Not Available (NA) o infinito (Inf) en los valores de las variables de nuestra base de datos.

```
apply(datos, 2, function(x) any(is.na(x)| is.infinite(x)))
```

```
##                   X           work_year   experience_level    employment_type
##               FALSE               FALSE              FALSE              FALSE
##           job_title              salary    salary_currency      salary_in_usd
##               FALSE               FALSE              FALSE              FALSE
## employee_residence       remote_ratio   company_location       company_size
##               FALSE               FALSE              FALSE              FALSE
```

Como devuelve FALSE en todas las variables significa que no hay valores Not Available (NA) o infinitos (Inf) en el data frame.

Utilizamos la función mean() para obtener la media de los salarios.

```
df <- as.data.frame(datos)

media_salario_usd <- mean(datos$salary_in_usd)
de_salario_usd <- sd(df$salary_in_usd)

print(media_salario_usd)
```

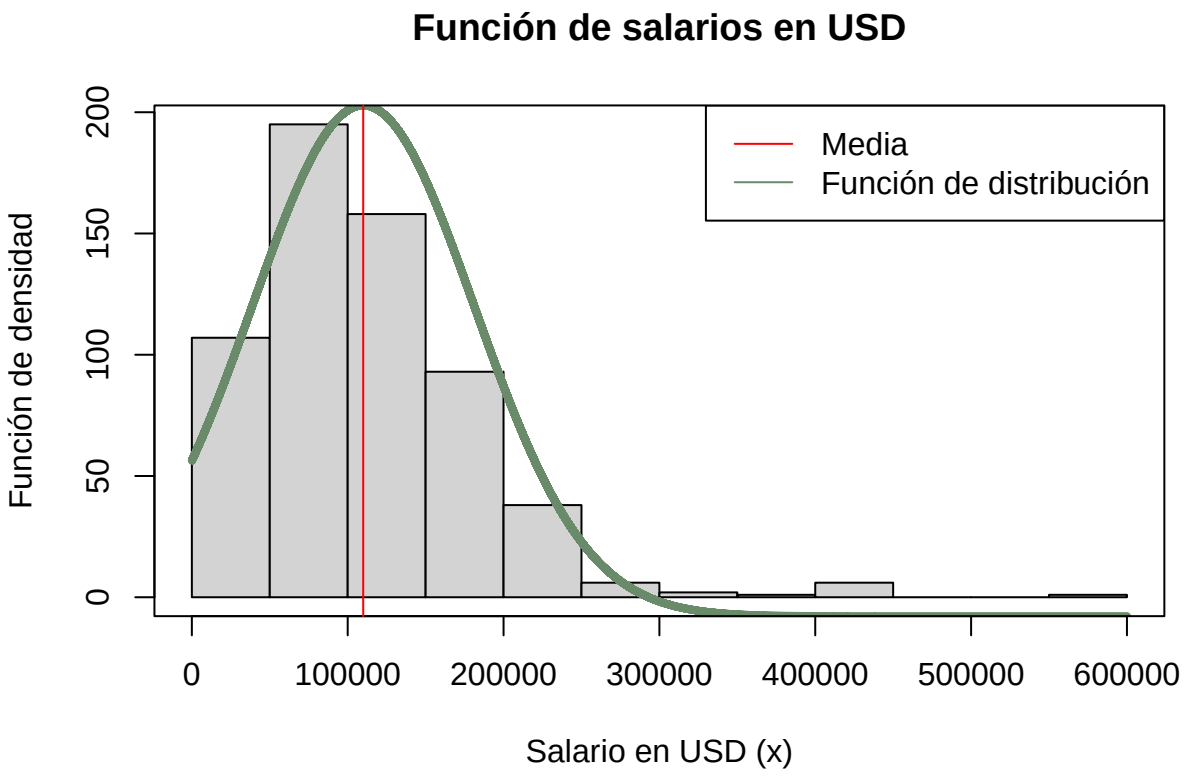
```
## [1] 112297.9
```

```
print(de_salario_usd)
```

```
## [1] 70957.26
```

Se trata de un valor razonable, teniendo en cuenta el valor que están tomando estos empleos con la digitalización de las empresas y la aparición de empresas que manejen grandes cantidades de datos.

```
options(scipen=999)

salario_usd <- dnorm(x = min(datos$salary_in_usd):max(datos$salary_in_usd),
                     mean = media_salario_usd,
                     sd = de_salario_usd,
                     log = FALSE)

hist(datos$salary_in_usd,
     ylab = "Función de densidad", xlab = "Salario en USD (x)",
     main = "Función de salarios en USD")

par(new=TRUE,yaxs="i", xaxt='n',yaxt='n',ann=FALSE )
plot(min(datos$salary_in_usd):max(datos$salary_in_usd), salario_usd,
     type = "l",  ylab = "Función de densidad", xlab = "Salario en USD (x)",
     lwd = 4, col = "darkseagreen4")

abline(v = media_salario_usd, lwd = 1,col = "red")

legend("topright",legend=c("Media", "Función de distribución"),
       col=c("red", "darkseagreen4"),lty=1)
```



Debido a la correlación entre el histograma y la función, podemos asegurar que se trata de una distribución normal.

```
media_distancia <- mean(datos$remote_ratio)
de_distancia <- sd(df$remote_ratio)

print(media_distancia)
```

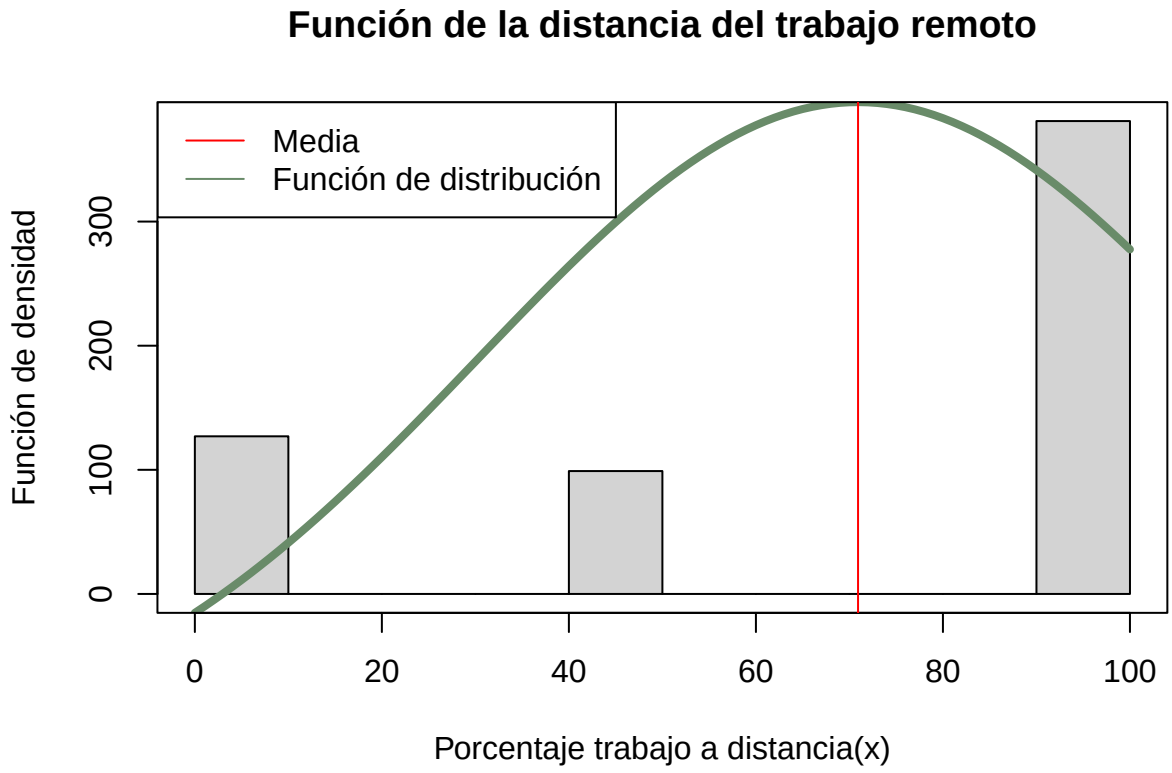
```
## [1] 70.92257
```

```
print(de_distancia)
```

```
## [1] 40.70913
```

```
options(scipen=999)

distancia <- dnorm(x = min(datos$remote_ratio):max(datos$remote_ratio),
                   mean = media_distancia,
                   sd = de_distancia,
                   log = FALSE)

hist(datos$remote_ratio,
     ylab = "Función de densidad", xlab = "Porcentaje trabajo a distancia(x)",
     main = "Función de la distancia del trabajo remoto")
```

```
par(new=TRUE,yaxs="i", xaxt='n',yaxt='n',ann=FALSE )

plot(min(datos$remote_ratio):max(datos$remote_ratio),distancia,
     type = "l", ylab = "Función de densidad", xlab = "Distancia(x)",
     lwd = 4, col = "darkseagreen4")

abline(v = media_distancia, lwd = 1,col = "red")

legend("topleft",legend=c("Media", "Función de distribución"),
       col=c("red", "darkseagreen4"),lty=1)
```



Observamos que la variable trabajo remoto no sigue ninguna distribucion que hayamos estudiado (hemos tratado de comparar el histograma con la distribución normal).