Actividad 1: Ajuste de distribuciones

María del Carmen Rojas Moreno

02/04/2025

La práctica consiste en seleccionar un conjunto de datos con al menos 30 observaciones y ajustarlo a alguna de las distribuciones estudiadas.

1. Describir estadísticamente los datos (principales medidas de posición y dispersión e histograma).
2. Justificar la selección de distribuciones discretas o continuas.
3. Ajustar los datos a las distribuciones seleccionadas (al menos dos distribuciones).
4. Análisis del ajuste (contraste de medidas de bondad y criterios de informe).
5. Conclusión final (seleccionar de forma justificada una de las distribuciones si es posible y calculo de algún cuantil con su interpretación).

Previo: Paquetes necesarios

```
library(magrittr)
library(dplyr)
library(actuar)
library(fitdistrplus) #install.packages("fitdistrplus")
library(tidyverse)
library(stats)
library(univariateML)
library(readxl)
library(MASS)
library(survival)
library(e1071)
```

1. Describir estadísticamente los datos (principales medidas de posición y dispersión e histograma).

El conjunto de datos escogido recopila información de 2000 usuarios sobre el tráfico web de una página concreta. La variable que queremos ajustar es "tiempo" ("Time.on.Page" en el conjunto de datos), que nos indica el tiempo total que un usuario pasa en el sitio web medido en minutos.

```
data <- read.csv("website_wata.csv")
tiempo <-data$Time.on.Page
```

Podemos hacer un análisis descriptivo de las datos usando las siguientes funciones:

```r
summary(tiempo)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.06852  1.93504  3.31532  4.02744  5.41463 24.79618
```

```r
skewness(tiempo)
```

```
## [1] 1.485632
```

```r
kurtosis(tiempo)
```

```
## [1] 3.431247
```

```r
var(tiempo)
```

```
## [1] 8.337206
```

```r
sd(tiempo)
```

```
## [1] 2.887422
```

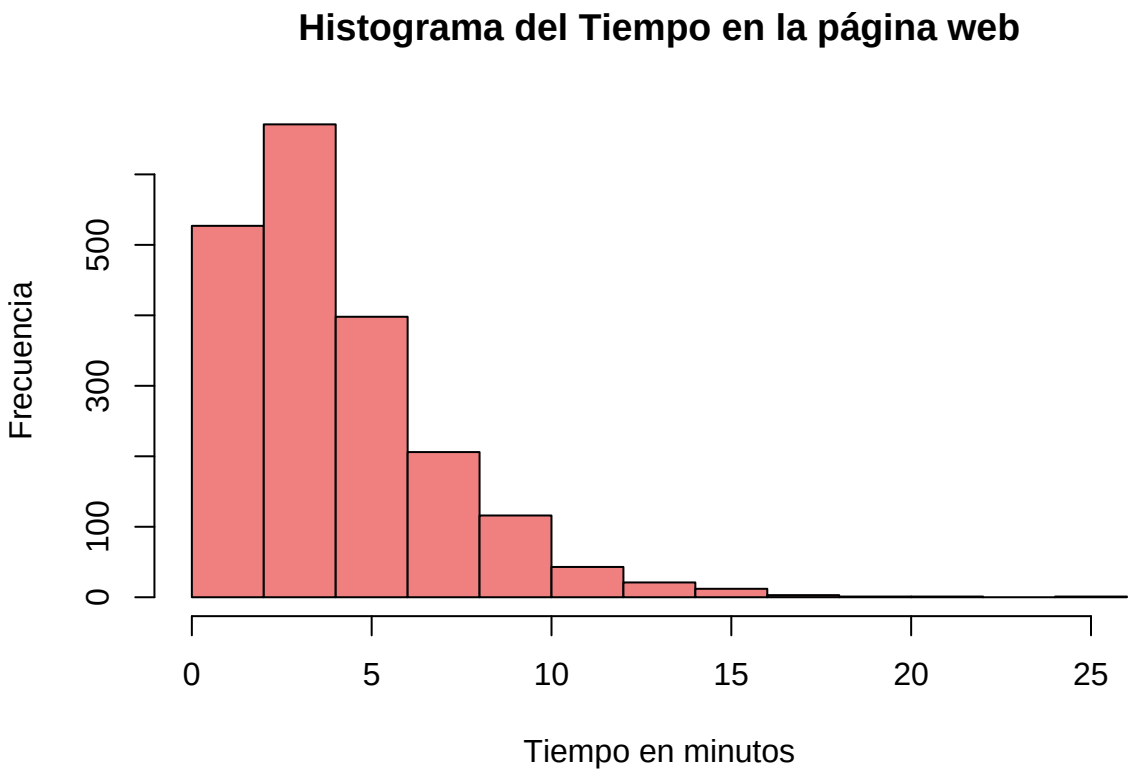
La función *summary* nos dice que la $media = 4.027439$ es mayor que la $mediana = 3.31532$, esto indica que hay asimetría positiva, es decir, la cola derecha es más larga que la izquierda. Además, como $skewness(tiempo) = 1.485632 > 0$, nos indica también que la distribución de los datos tiene una asimetría positiva.

Por otro lado, como $kurtosis(tiempo) = 3.431247 > 3$, los datos tienen una distribución con colas más pesadas y una mayor concentración de valores en la media si la comparamos con una distribución normal.

Asímismo, la varianza y la desviación típica indican que hay bastante variabilidad en los datos. Además, la función *summary* nos dice que los datos van de 0.06852 a 24.79618 (son mayores estrictos que cero).

A continuación representamos el histograma de los datos.

```r
hist(tiempo, main="Histograma del Tiempo en la página web",
     xlab= "Tiempo en minutos", ylab = "Frecuencia", col="lightcoral")
```



La gráfica que se muestra encaja con la descripción estadística que hemos realizado antes.

2. Justificar la selección de distribuciones discretas o continuas.

Nuestra variable es el tiempo total que un usuario pasa en el sitio web medido en minutos, seleccionaremos distribuciones continuas para el ajuste, pues el tiempo no toma valores discretos.

Para el ajuste probaremos distribuciones como la de Pareto, la log-normal y la de Weibull. La distribución exponencial la descartaremos ya que en una distribución exponencial, la varianza es igual al cuadrado de la media, y no se corresponde con lo que hemos obtenido en el análisis descriptivo de los datos realizado en el apartado anterior.

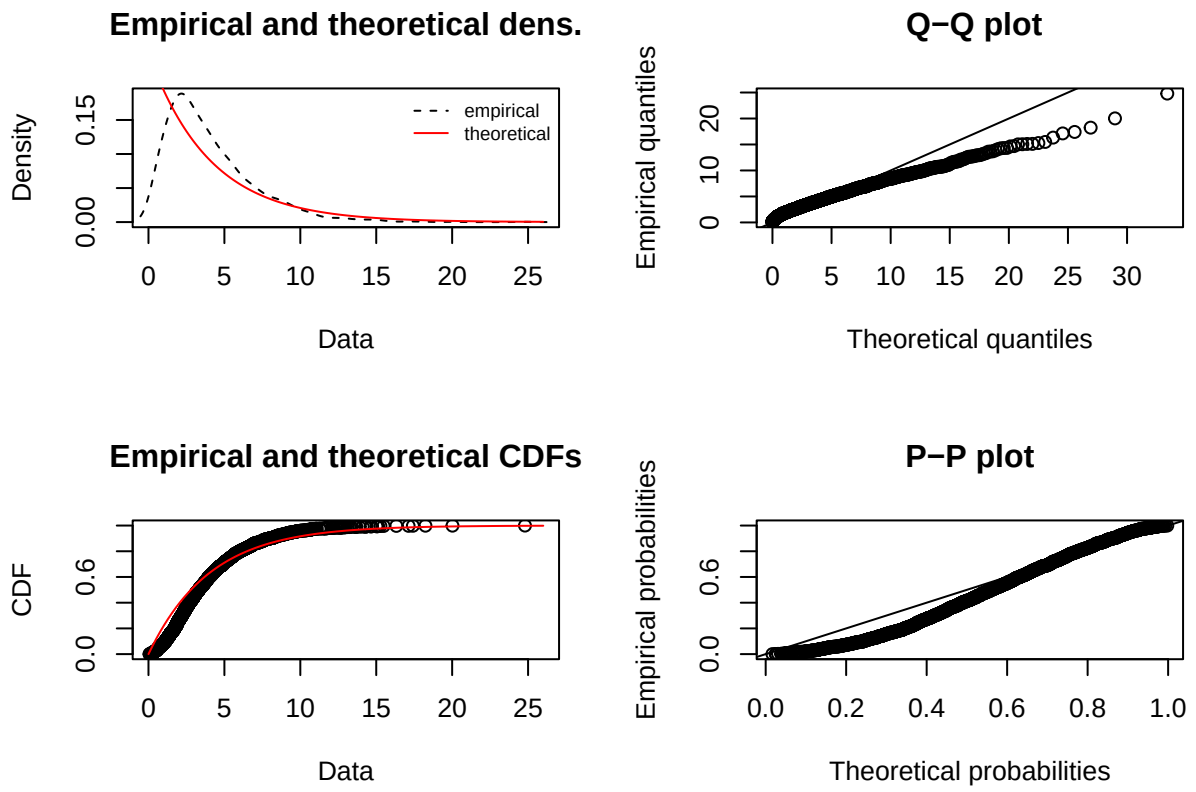
3. Ajustar los datos a las distribuciones seleccionadas (al menos dos distribuciones).

Ajuste Distribución de Pareto

```
fitp<- fitdist(tiempo, "pareto")
fitp
```

```
## Fitting of the distribution ' pareto ' by maximum likelihood
## Parameters:
##       estimate
## shape 15253207
## scale 61429191
```

```
plot(fitp,histo=FALSE,demp=TRUE)
```



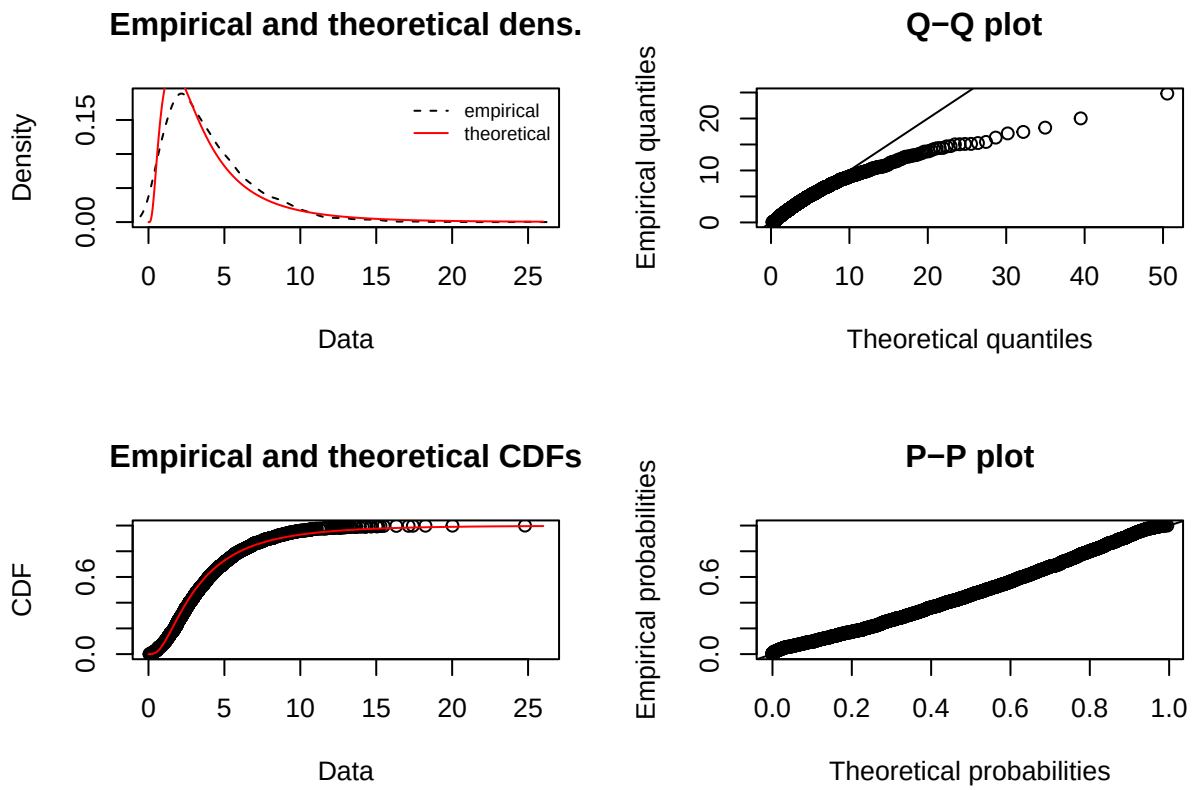
Obtenemos los parámetros estimados $\hat{\alpha} = 15253207$ y $\hat{\lambda} = 61429191$. En las gráficas se puede observar que la distribución de Pareto no se ajusta bien para la función de densidad. Además, podemos ver que este ajuste tampoco nos da buenos resultados en la gráfica Q-Q plot.

Ajuste Distribución log-normal

```
fitln <- fitdist(tiempo, "lnorm")
fitln
```

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters:
##         estimate Std. Error
## meanlog 1.119946 0.01800565
## sdlog   0.805237 0.01273183
```

```
plot(fitln,histo=FALSE,demp=TRUE)
```



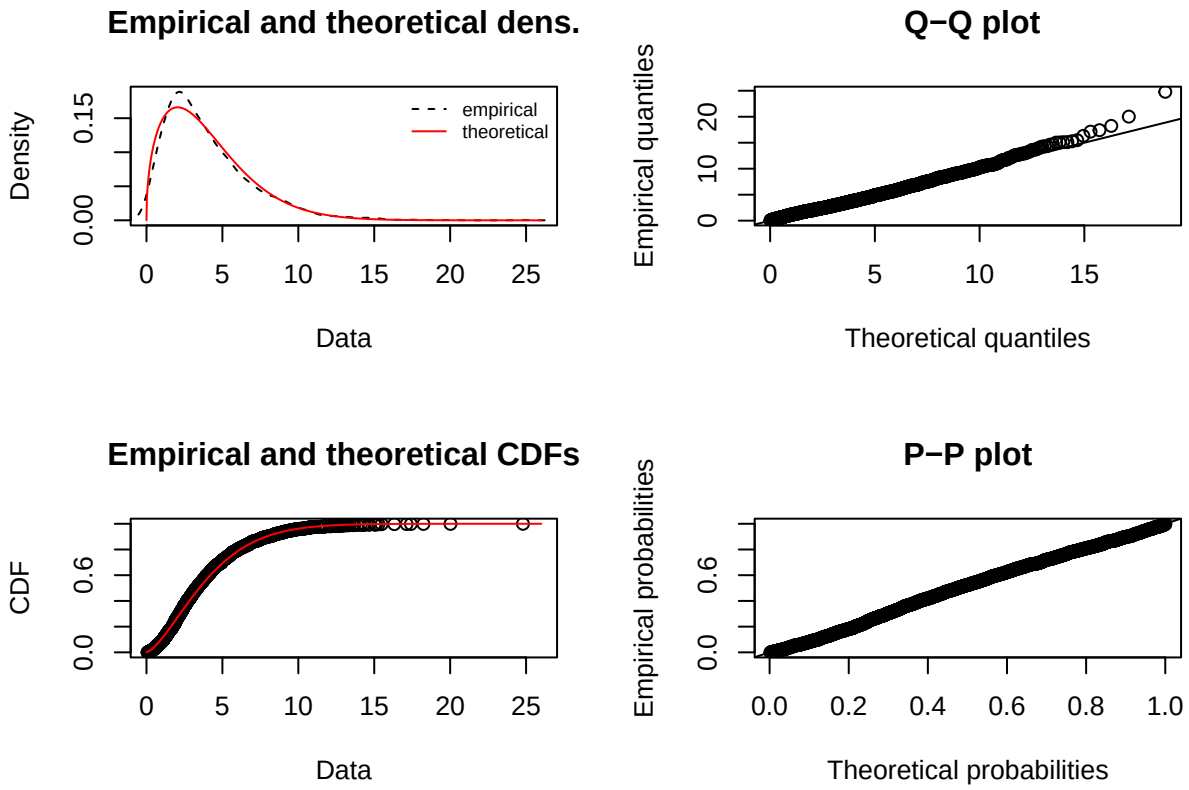
Obtenemos los parámetros estimados $\hat{\mu} = 1.119946$ y $\hat{\sigma} = 0.805237$. En las gráficas se puede observar que la distribución log-normal no se ajusta bien para la función de densidad, aunque parece que obtenemos mejores resultados que para la distribución de Pareto. Además, podemos ver que este ajuste tampoco nos da buenos resultados en la gráfica Q-Q plot.

Ajuste Distribución de Weibull

```
fitw <- fitdist(tiempo, "weibull")
fitw
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters:
##       estimate Std. Error
## shape 1.468129 0.02486420
## scale 4.464922 0.07181465
```

```
plot(fitw,histo=FALSE,demp=TRUE)
```



Obtenemos los parámetros estimados $\hat{\gamma} = 1.468129$ y $\hat{c} = 4.464922$. En las gráficas se puede observar que la distribución de Weibull se ajusta mejor para la función de densidad que las dos distribuciones que vimos anteriormente. Además, el gráfico Q-Q plot muestra que la distribución de Weibull proporciona un mejor ajuste que las distribuciones de Pareto y log-normal.

4. Análisis del ajuste (contraste de medidas de bondad y criterios de información).

Test de Kolmogorov Smirnov

```
parametrosp<-unname(fitp$estimate)
ks.test(tiempo, "ppareto", shape=parametrosp[1], scale=parametrosp[2])
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  tiempo
## D = 0.1507, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

En el caso de la distribución de Pareto obtenemos un p-valor menor que 0.05, luego rechazaríamos la hipótesis nula de que los datos siguieran una distribución de Pareto.

```
parametrosln<-unname(fitln$estimate)
ks.test(tiempo, "plnorm", meanlog=parametrosln[1], sdlog=parametrosln[2])
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  tiempo
## D = 0.043504, p-value = 0.001031
## alternative hypothesis: two-sided
```

En el caso de la distribución log-normal obtenemos también un p-valor menor que 0.05, luego rechazaríamos la hipótesis nula de que los datos siguieran una distribución log-normal.

```
parametrosw<-unname(fitw$estimate)
ks.test(tiempo, "pweibull", shape=parametrosw[1], scale=parametrosw[2])
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  tiempo
## D = 0.028715, p-value = 0.0739
## alternative hypothesis: two-sided
```

En el caso de la distribución de Weibull obtenemos un p-valor mayor que 0.05, luego no podríamos rechazar la hipótesis nula, por ello se podría decir que los datos siguen una distribución de Weibull.

Criterios de información

```
gofstat(list(fitp, fitw, fitln))
```

```
## Goodness-of-fit statistics
##                               1-mle-pareto 2-mle-weibull 3-mle-lnorm
## Kolmogorov-Smirnov statistic    0.1506977    0.02871498  0.04350372
## Cramer-von Mises statistic     14.4534021    0.54501967  1.57184927
## Anderson-Darling statistic     83.4476500    3.29462148 10.84321155
##
## Goodness-of-fit criteria
##                                 1-mle-pareto 2-mle-weibull 3-mle-lnorm
## Akaike's Information Criterion     9576.523      9157.193    9293.065
## Bayesian Information Criterion     9587.725      9168.395    9304.267
```

En la tabla se muestra que el menor AIC y BIC se obtiene para el ajuste de la distribución de Weibull.

5. Conclusión final (seleccionar de forma justificada una de las distribuciones si es posible y cálculo de algún cuantil con su interpretación).

Teniendo en cuenta las gráficas vistas en el apartado 3 y los resultados del apartado 4, donde obteníamos que no podíamos rechazar que los datos siguiesen una distribución de Weibull, y que el menor valor de los criterios de información correspondía tambien al ajuste de la distribución de Weibull, seleccionaremos esta distribución para llevar a cabo el ajuste.

A continuación, calculamos algunos cuantiles de la distribución usada para el ajuste:

```r
weibull<-qweibull(c(0.9,0.95,0.975,0.99), shape=parametrosw[1],scale = parametrosw[2])
weibull
```

```
## [1]  7.880124  9.427110 10.862948 12.635030
```

Podemos interpretar algunos cuantiles de los calculados. Por ejemplo, los cuantiles del 90% y 99% nos dicen que según la distribución ajustada:

- El 90% de los usuarios permanecen menos de 7.880124 minutos en la página web, mientras que el 10% restante pasa más tiempo.
- El 99% de los usuarios pasan menos de 12.635030 minutos en la página web, mientras que el 1% restante permanece más tiempo.

Por otro lado, si calculamos los cuantiles muestrales de los datos, tenemos los siguientes resultados:

```r
experimental<-quantile(tiempo, c(0.9,0.95,0.975,0.99))
error_weibull<-abs((experimental-weibull)/experimental)
datos<-data.frame(Experimental=experimental,Weibull=weibull,Error_Relativo=error_weibull)
print(datos)
```

```
##       Experimental   Weibull Error_Relativo
## 90%       7.975989  7.880124     0.01201926
## 95%       9.591644  9.427110     0.01715393
## 97.5%    11.058775 10.862948     0.01770781
## 99%      13.640119 12.635030     0.07368622
```

Los resultados que obtenemos para los cuantiles del 90%, 95% y 97.5% son buenos, pues el error relativo es bajo. Sin embargo, aunque la distribución Weibull nos aporta un buen ajuste general a los datos, para el cuantil 99% se tiene un error relativo mayor, indicando que la distrubución no se ajusta tan bien al final de la cola (como podíamos ver en el gráfico Q-Q plot).