Máster en Data Science. URJC

Técnicas y Métodos de Ciencia de Datos

Practica 2 Modelos de distribución de probabilidad

Leda Duelo, Javier Llorente, Candela Vidal, Jaime Zamorano

26/11/2017

Índice

1 Introducción

El conjunto de datos `BATTERY` incluido en el paquete `PASWR2` contiene 100 observaciones de 2 variables correspondientes a la duración de dos tipos de baterías A y B (en horas). El conjunto de datos es un data.frame con las columnas `lifetime` y `facility`. Para realizar esta práctica, carga primero el conjunto de datos en tu espacio de trabajo, por ejemplo:

```
library(lattice)
library(ggplot2)
library(PASWR2)
datos <- BATTERY
```

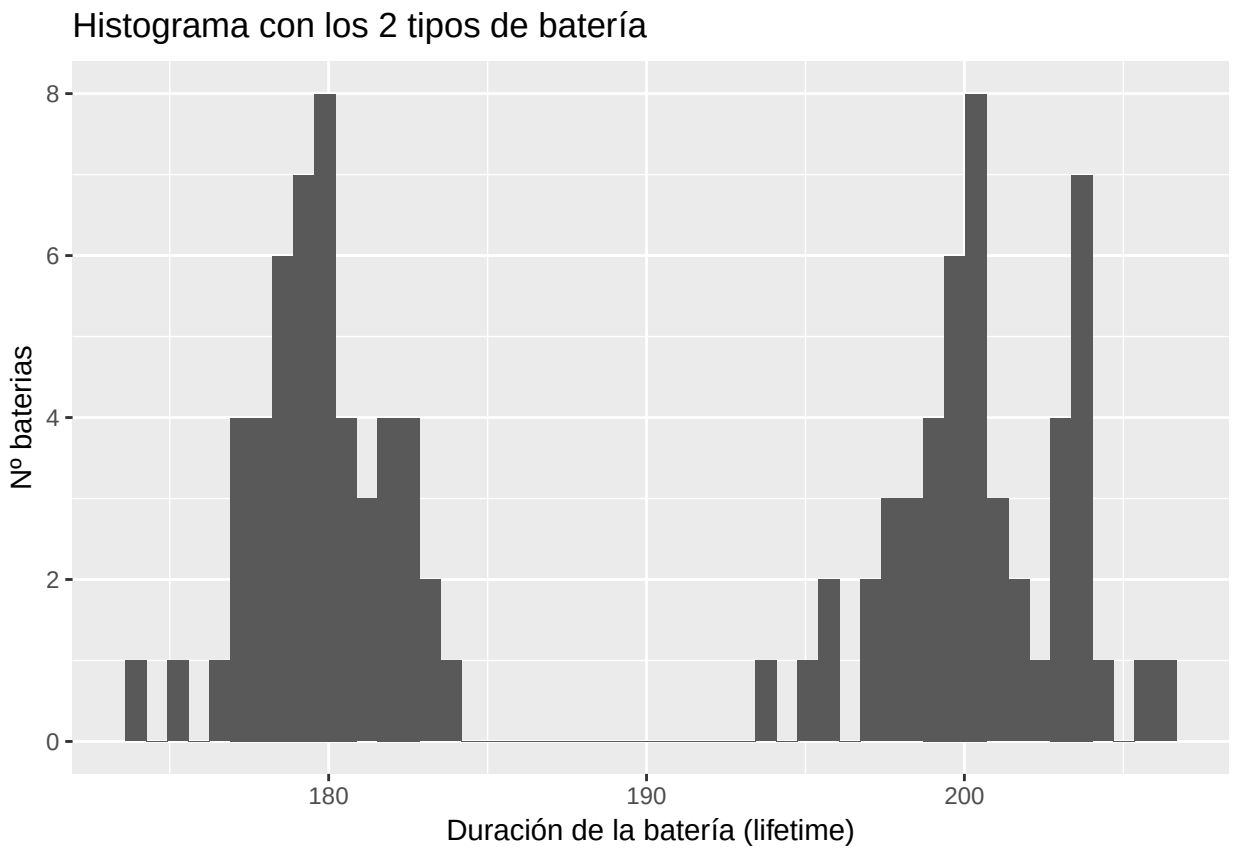
Fíjate que tienes que tener instalado el paquete `PASWR2` para poder acceder a este conjunto de datos.

La variable de interés es `lifetime`, pero como sabemos que los datos se refieren a dos tipos distintos de baterías, posiblemente nos interese separarlos. En esta práctica vamos a realizar cálculo de probabilidades basados en este conjunto de datos para que se vea una aplicación, aunque tengamos que hacer uso de algún concepto de inferencia.

2 Actividad 1

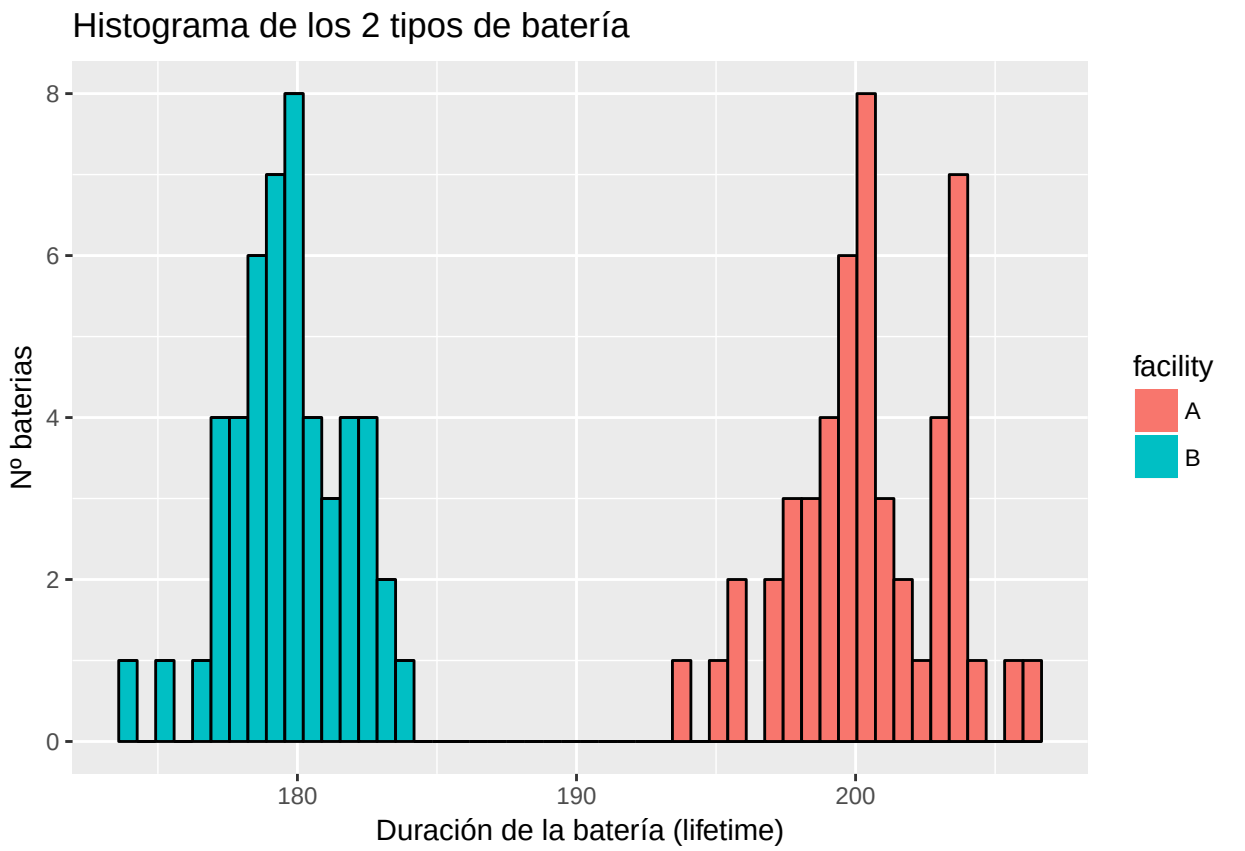
- Realiza un histograma de todas las filas de la variable `lifetime` y comprueba que efectivamente nos interesa separar los datos.

```
ggplot(datos, aes(datos$lifetime)) +
  geom_histogram(bins=50) +
  labs(x="Duración de la batería (lifetime)", y="Nº baterias",
       title="Histograma con los 2 tipos de batería")
```



El histograma muestra una distribución bimodal.Coloreamos los datos por tipo de batería para ver si es mejor crear un dataset por tipo de batería.

```
ggplot(datos,aes(datos$lifetime)) +
  geom_histogram(aes(fill = facility, col = I("black")),bins = 50) +
        labs(x="Duración de la batería (lifetime)", y="Nº baterias",
       title="Histograma de los 2 tipos de batería")
```



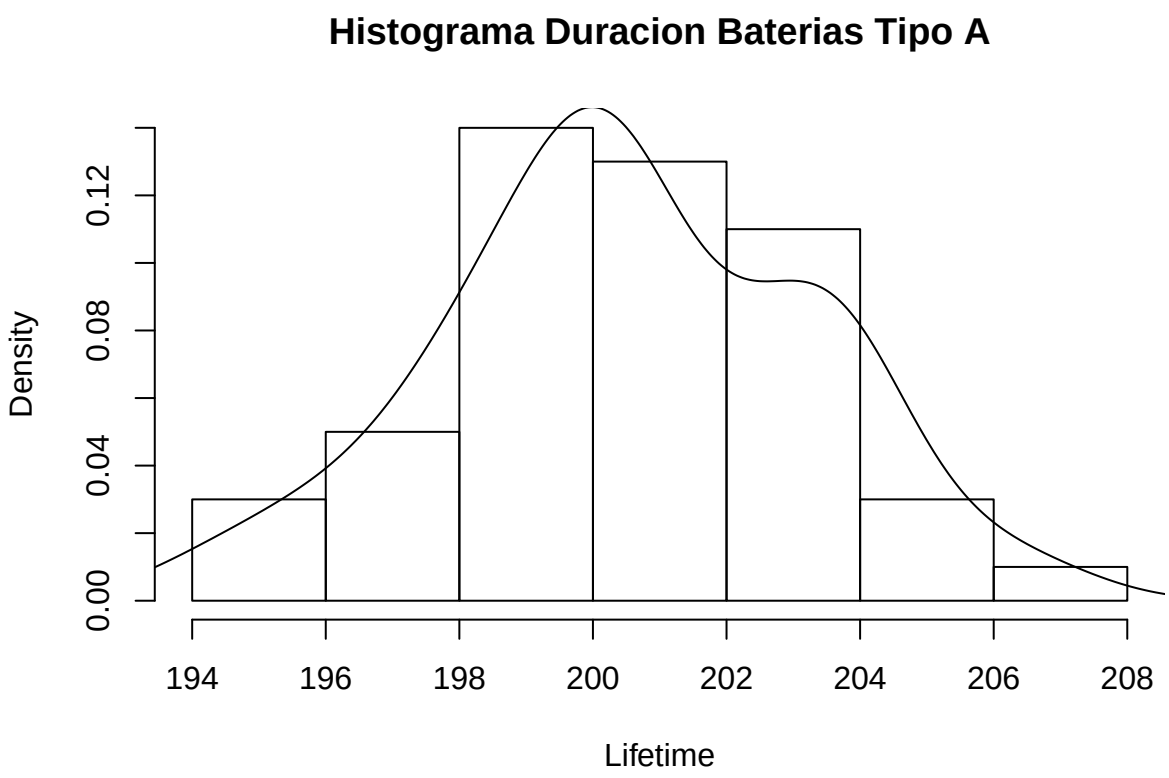
Es mejor hacer el análisis y crear un dataset por tipo de batería.

- Crea dos conjuntos de datos diferentes para los dos tipos de baterías, por ejemplo `datosA` y `datosB`.

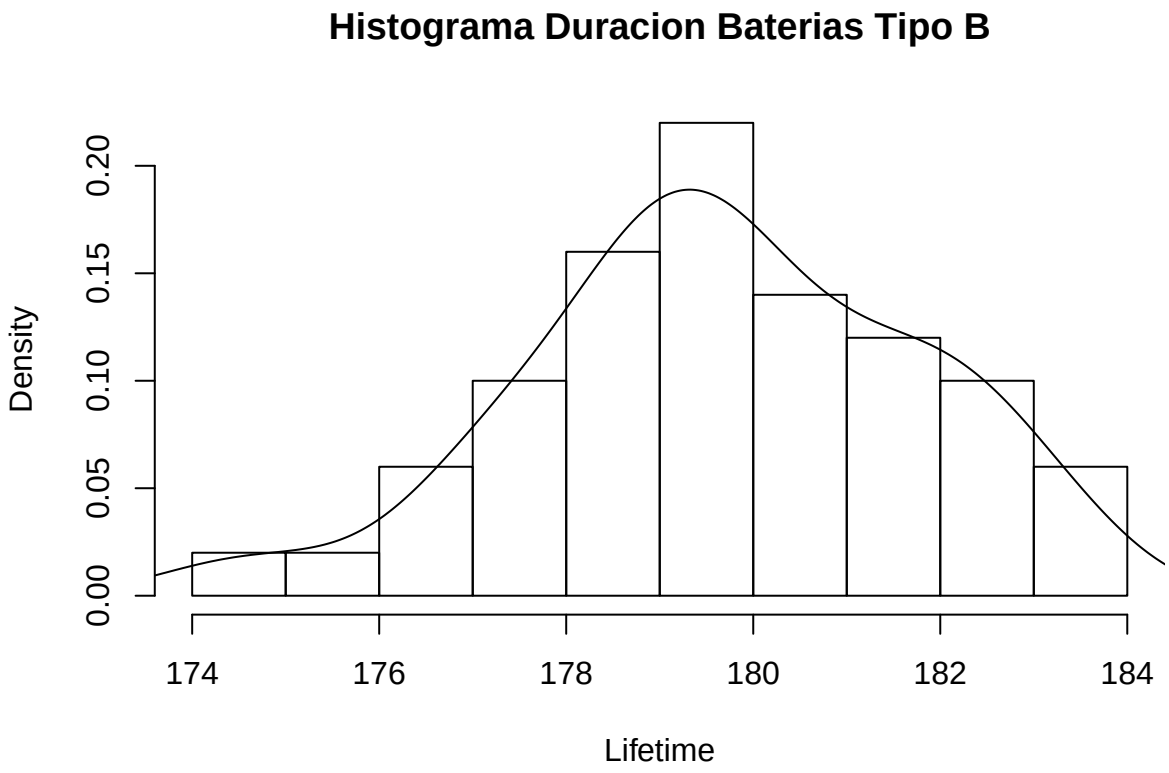
```
datosA = subset(datos, datos$facility == "A")
datosB = subset(datos, datos$facility == "B")
```

- Realiza ahora un histograma de cada uno de los tipos y comenta si te parece que los datos siguen una distribucion normal.

```
hist(datosA$lifetime,
     main="Histograma Duracion Baterias Tipo A",
     xlab="Lifetime",
     prob = TRUE)
lines(density(datosA$lifetime))
```



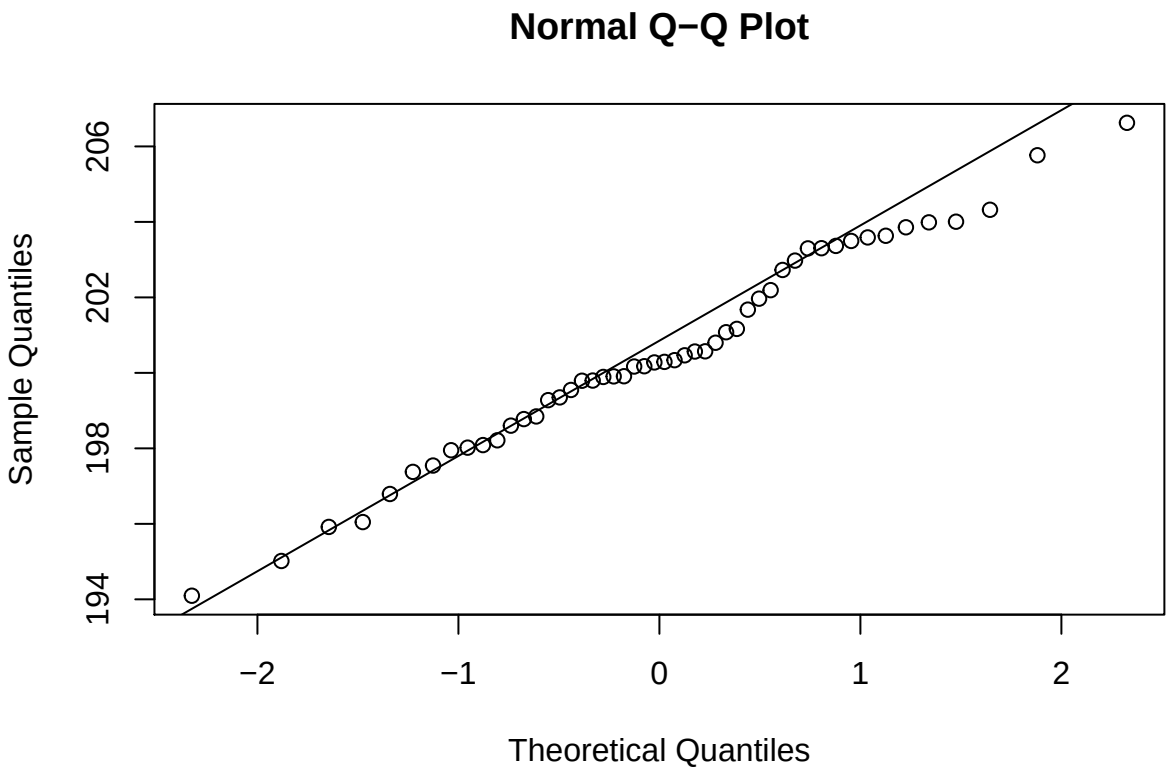

```
hist(datosB$lifetime,
     main="Histograma Duracion Baterias Tipo B",
     xlab="Lifetime",
     prob = TRUE)
lines(density(datosB$lifetime))
```



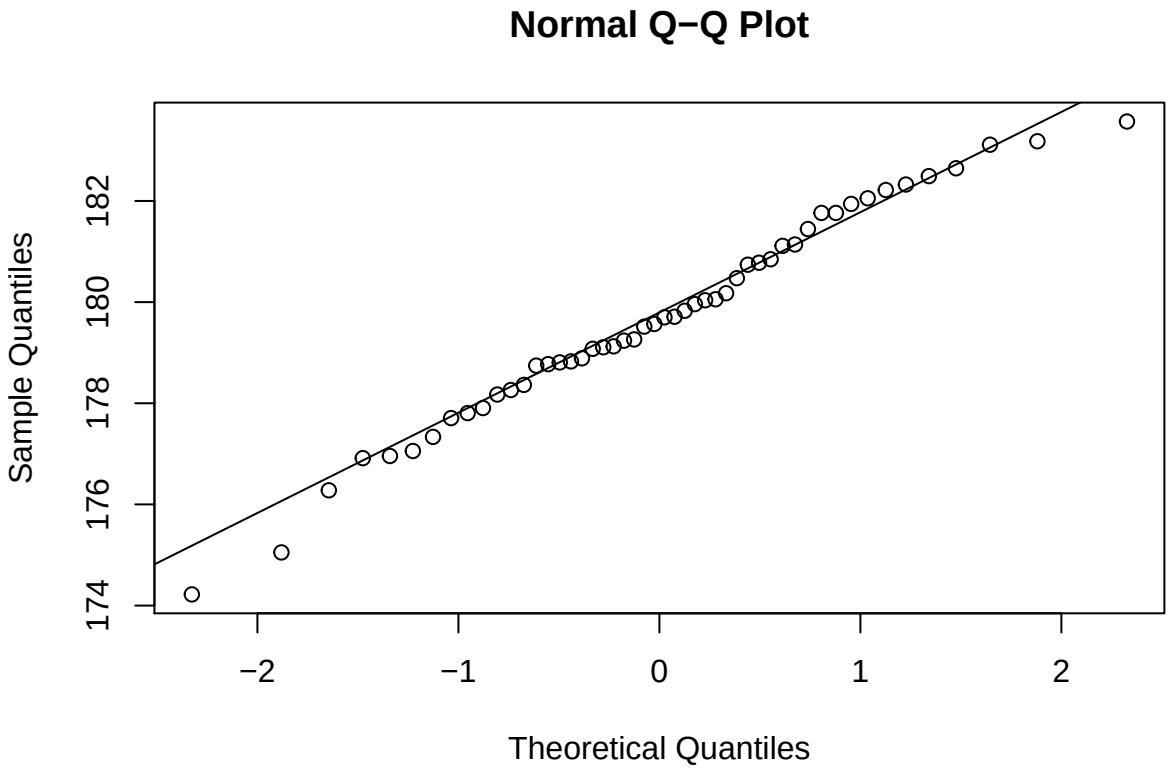
- Confirma tus conclusiones con alguna/s de las herramientas vistas en clase (test de normalidad, gráfico Quantil-Quantil, tests de normalidad,...)

Los histogramas muestrán que la duración de las baterias siguen una distribución normal. Lo comprobamos con los grafios quantil-quantil:

```
qqnorm(datosA$lifetime)
qqline(datosA$lifetime)
```




```
qqnorm(datosB$lifetime)
qqline(datosB$lifetime)
```



La escala probabilística es cercana a la recta, por lo que se puede afirmar que ambos tipos de batería siguen

una distribución normal.

3 Actividad 2

Ahora que sabemos que nuestros datos siguen aproximadamente una distribución normal, tendríamos que estimar sus parámetros μ y σ. A partir de ahí, podemos realizar cálculo de probabilidades de la normal.

- Realiza una estimación puntual de la media y la desviación típica de la población de cada tipo de baterías.

$$\mu_X = \mathbb{E}[X] = \sum x f(x) \tag{1}$$

$$\sigma = \sqrt{\sigma_X^2} = \sqrt{\text{Var}[X]} = \sqrt{\sum (x-\mu)^2 f(x)} \tag{2}$$

Con función de densidad:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

```
mediaA = round(mean(datosA$lifetime),2)
mediaB = round(mean(datosB$lifetime),2)

desvA = round(sd(datosA$lifetime),2)
desvB = round(sd(datosB$lifetime),2)
```

- Calcula la probabilidad de que una batería tomada al azar del tipo A dure más de 210 horas

$P[X > 210]$

```
prob_210 = pnorm(q = 210, mean = mediaA, sd = desvA, lower.tail = FALSE)
prob_210
```

```
## [1] 0.0002793509
```

- Calcula la probabilidad de que una batería tomada al azar del tipo B dure menos de 175 horas. Entendemos que menos de 175 horas es ≤ 174

$P[X < 175] = P[X \leq 174]$

```
prob_175 = pnorm(q = 174, mean = mediaB, sd = desvB)
prob_175
```

```
## [1] 0.003159335
```

- Encuentra cuál es la duración máxima del 3% de las pilas del tipo B que duran menos (ayuda: esto es equivalente a encontrar el cuantil 0.03 de la distribución)

```
durac_3 = qnorm(p = 0.03, mean = mediaB, sd = desvB)
durac_3
```

```
## [1] 175.7679
```

4 Actividad 3

Vamos a centrarnos ahora en las baterías de tipo B. Supongamos que una duración por debajo de 175 horas no es aceptable para el usuario de la batería. En la actividad anterior hemos calculado la probabilidad p de que esto suceda. Entonces, si tomamos una batería del tipo B al azar y comprobamos si dura menos de 175 horas, estamos realizando un experimento de Bernoulli con probabilidad p.

- Calcula la probabilidad de que en un lote de 10 baterías, no haya ninguna defectuosa (ayuda: distribución binomial).

$$P[X = k] = \left(\begin{array}{c} n \\ k \end{array} \right) .p^k .(1-p)^{(n-k)} \tag{4}$$

n numero de pruebas independientes de Bernuilli

$P[X < 1] = P[X = 0]$

```
pbinom(q = 0, size = 10,prob = prob_175)
```

```
## [1] 0.9688521
```

- Imagina que las baterías se fabrican en serie e independientemente. ¿Cuál es la probabilidad de que la batería producida en quinto lugar sea la primera defectuosa? (ayuda: distribución geométrica.)

$$P[X = x] = p(1-p)^x \tag{5}$$

x es numero de fracasos hasta el primer éxito

$P[X = 4]$

```
pgeom(4, prob = prob_175)
```

```
## [1] 0.01569718
```

- Supongamos que en una caja de 20 baterías van 3 defectuosas. ¿Cuál es la probabilidad de que al tomar una muestra sin reposición de 5 baterías al menos una sea defectuosa? (ayuda: distribución hipergeométrica)

$$P[X = x] = \frac{\left(\begin{array}{c} N - D \\ n - x \end{array} \right) \cdot \left(\begin{array}{c} D \\ x \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \tag{6}$$

muestreo sin remplazo de tamaño n de un conjunto con N elementos totales de los que D son de dicha categoría

$P[X \geq 1] = 1 - P[X < 1] = 1 - P[X = 0]$

```
1 - dhyper(x = 0, m = 3, n = 17, k = 5)
```

```
## [1] 0.6008772
```

5 Actividad 4

Seguimos con las baterías de tipo B, pero en vez de hacer experimentos de Bernoulli queremos estudiar el número de baterías defectuosas fabricadas cada día. Supongamos que se fabrican 1000 baterías cada día. Entonces, cada día en promedio se estarán produciendo aproximadamente $1000 » p$ baterías, y el número de baterías defectuosas por día sigue una distribución de Poisson. Tomemos 12 como ese promedio de baterías defectuosas cada día. (ayuda: repasa qué modelo de distribución modeliza estos recuentos de eventos raros con una tasa media por unidad de tiempo)

La distribución de Poisson describe muy bien los procesos donde se cuentan el número de ocurrencias de un evento por unidad (de tiempo, espacio,...)

$$P[X = k] = \frac{e^{-k} \lambda^k}{k!} \tag{7}$$

- ¿Cuál es la probabilidad de que un día se produzcan más de 20 baterías defectuosas?

$P[X > 20] = 1 - P[X \leq 20]$

```
1 - ppois(q = 20, lambda = 12)
```

```
## [1] 0.01159774
```

- ¿Cuál es la probabilidad de que un día no salga ninguna batería defectuosa de la fábrica?

$P[X < 1] = P[X = 0]$

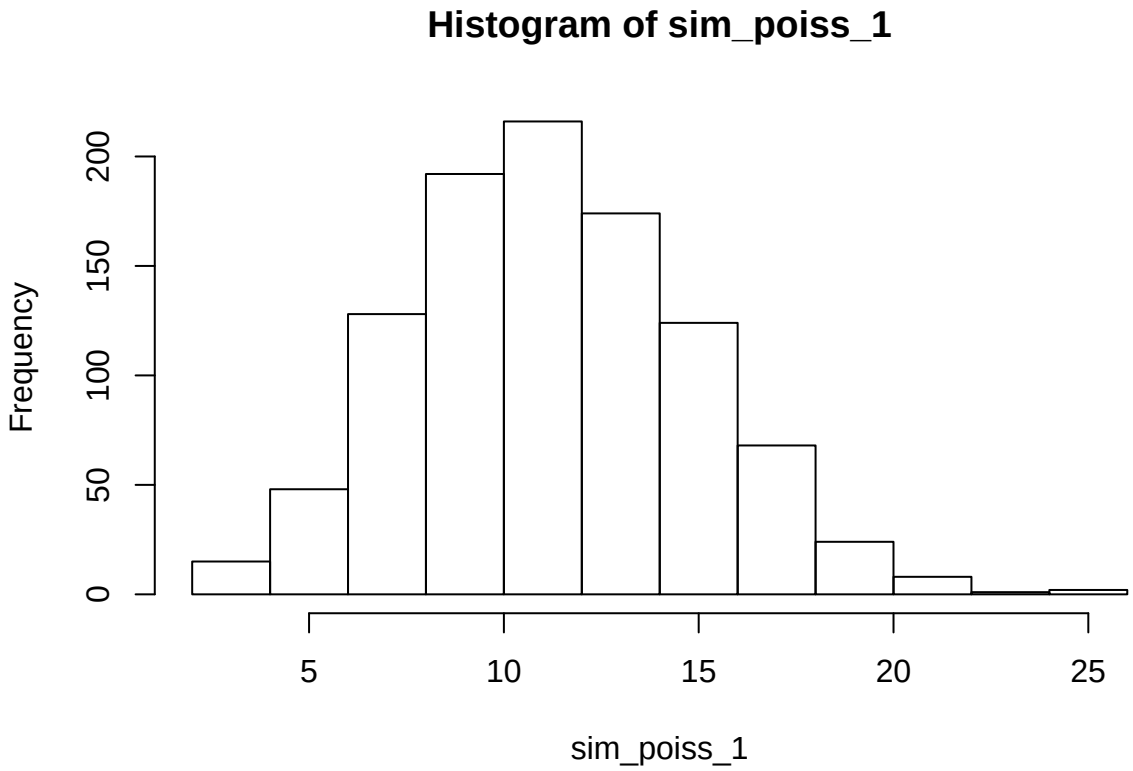
```
ppois(q = 0, lambda = 12)
```

```
## [1] 6.144212e-06
```

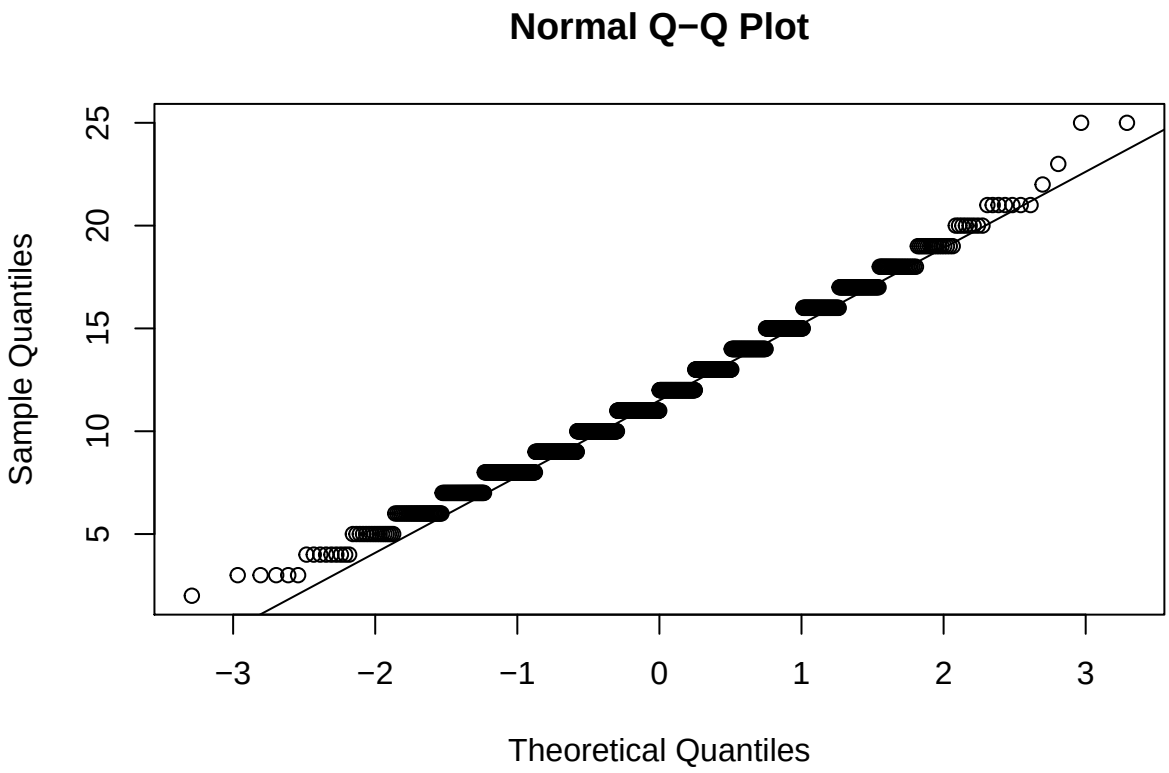
- La fábrica funciona de lunes a viernes. ¿Qué distribución sigue el número de baterías defectuosas por semana? Justifica qué propiedad se aplica.

Realizamos la simulacion de un día:

```
set.seed(1)
sim_poiss_1= rpois(1000,12)
hist(sim_poiss_1)
```

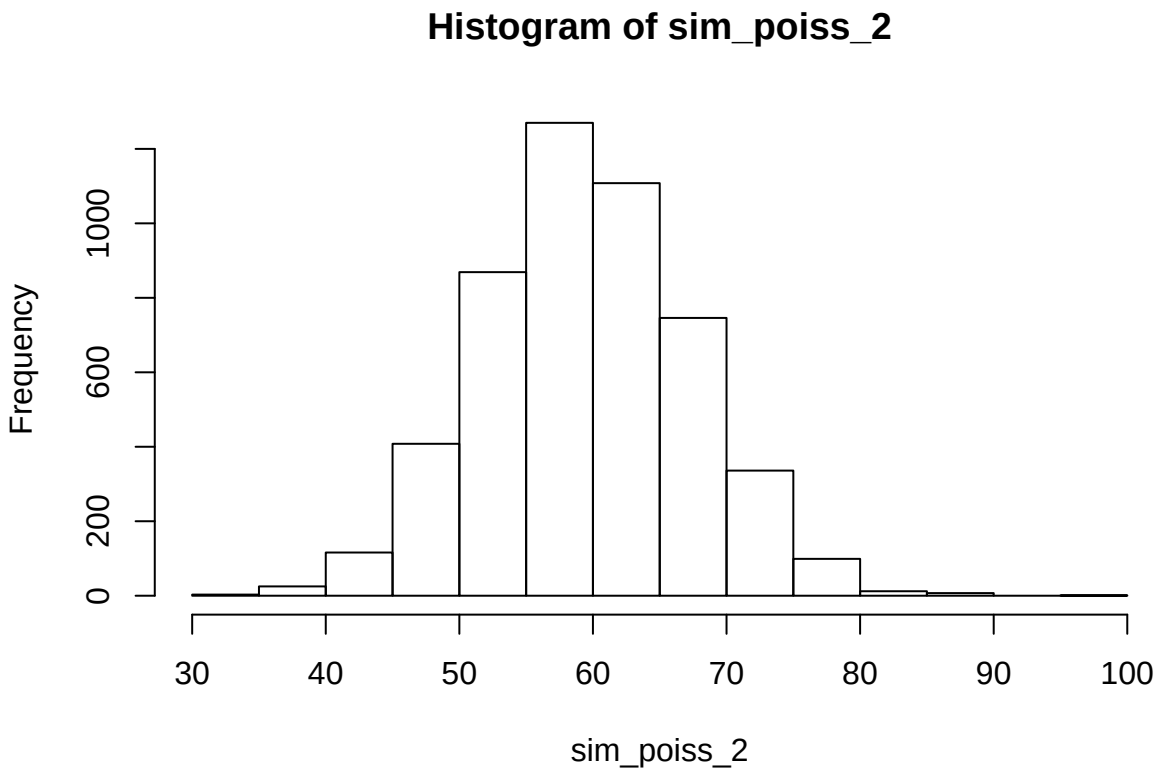



```
qqnorm(sim_poiss_1)
qqline(sim_poiss_1)
```

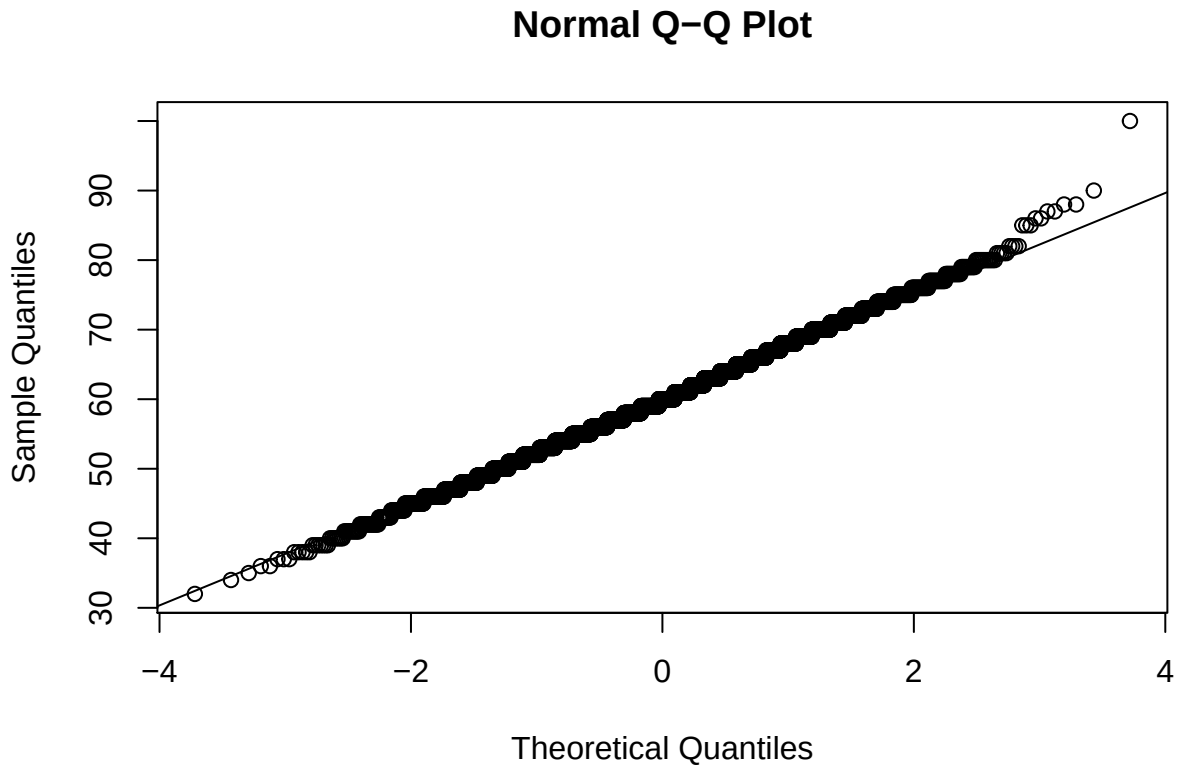


Realizamos ahora la simulación de una semana:

```
set.seed(1)
sim_poiss_2= rpois(5*1000,12*5)
hist(sim_poiss_2)
```




```
qqnorm(sim_poiss_2)
qqline(sim_poiss_2)
```



Al aumentar el número empieza a simular el comportamiento de una normal.

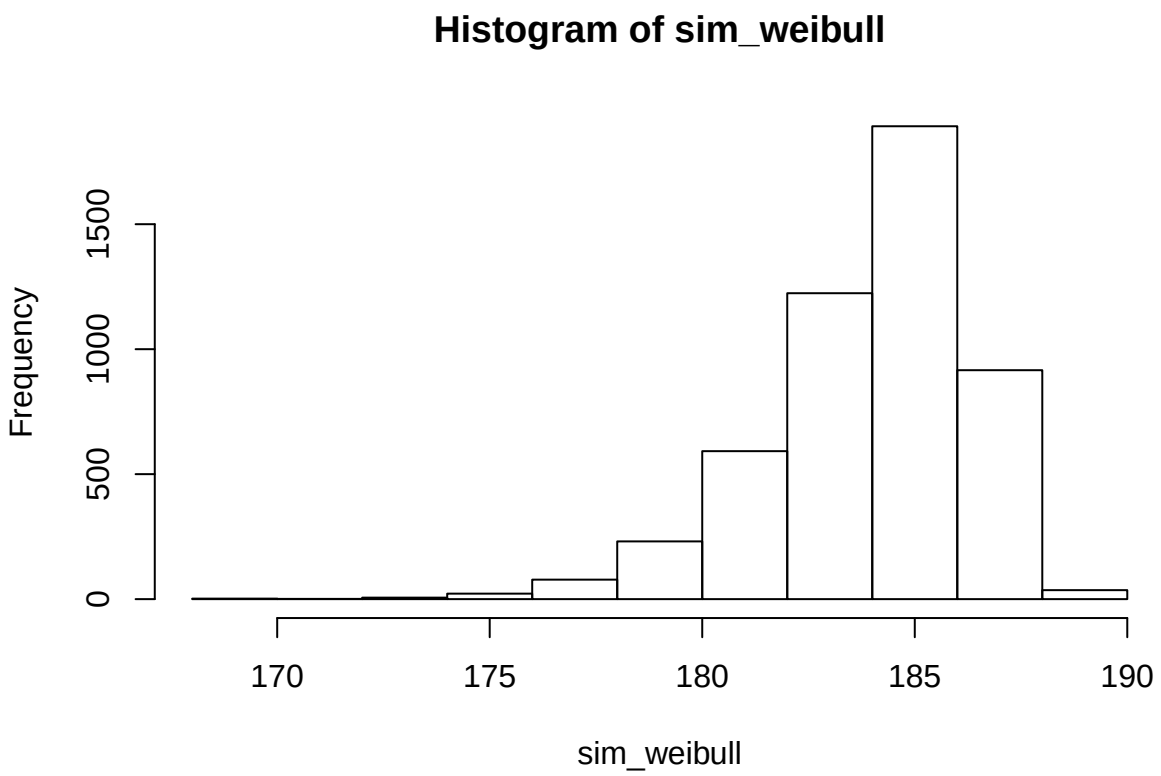
6 Actividad 5

El departamento de I+D de la empresa que fabrica las baterías tipo B está investigando nuevos materiales y métodos para mejorar la vida útil de las baterías. En particular, quieren llegar a diseñar una batería cuya duración siga una distribución de Weibull con parámetros $a = 100$ y $b = 185$.

```
a = 100
b = 185
```

- Realiza una simulación de la producción semanal de baterías (recuerda: 5 días de produccción, a 1000 baterías por día). Guarda los datos en un vector.

```
set.seed(1)
sim_weibull = round(rweibull(5000, a, b),2)
hist(sim_weibull)
```



- Con este nuevo proceso, ¿se mejora realmente la duración media de las baterías? (ayuda: puedes usar los datos simulados o la expresión de la esperanza de una Weibull)

```
media_weibull = round(mean(sim_weibull),2)
media_weibull
```

```
## [1] 183.95
```

- Los ingenieros no lo tienen muy claro (parece que la diferencia no es tanta en promedio y los nuevos

materiales son costosos). Para demostrarles que merece la pena, calcula la proporción de baterías defectuosas que producirá probablemente el nuevo proceso y compárala con el anterior (la p que calculamos en la actividad 2)

$$f(x) = \frac{a}{b}(\frac{x}{b})^{a-1}e^{-(\frac{x}{b})^a} \tag{8}$$

Para $x > 0$

```
func_weib = function(x) (a/b) * (x/b)^(a-1) * (exp(1)) ^ (-1*(x/b)^a)
prob_175_weib = integrate(func_weib,0,174)
prob_175_weib
```

```
## 0.0021741 with absolute error < 2.2e-09
```

```
prob_175_weib$value < prob_175
```

```
## [1] TRUE
```

Esta nueva distribución presenta una menor proporcion de baterías defectuosas que la primera.

7 Entrega del trabajo

Crea un informe R Markdown (.Rmd) y sube el fichero .pdf o .html al aula virtual con tu resolución, incluyendo fragmentos de código, así como el texto y gráficos que consideres conveniente. Opcionalmente, puedes subir el documento fuente .Rmd u otros ficheros.

Este es un trabajo a realizar en grupo. La entrega la debe realizar solo uno de los miembros del grupo de trabajo, y debe incluir los nombres y apellidos de todos los miembros del grupo.

8 Evaluación

El trabajo se valorará de 0 a 10 puntos, de modo que las cuestiones valen todas lo mismo (en total 7 puntos) y la presentación, explicaciones, y gráficos adicionales, 3 puntos.