Práctica 2 - Modelos de distribución de probabilidad

Emilio López Cano

22 de enero de 2017

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(PASWR2)
datos <- BATTERY
```

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

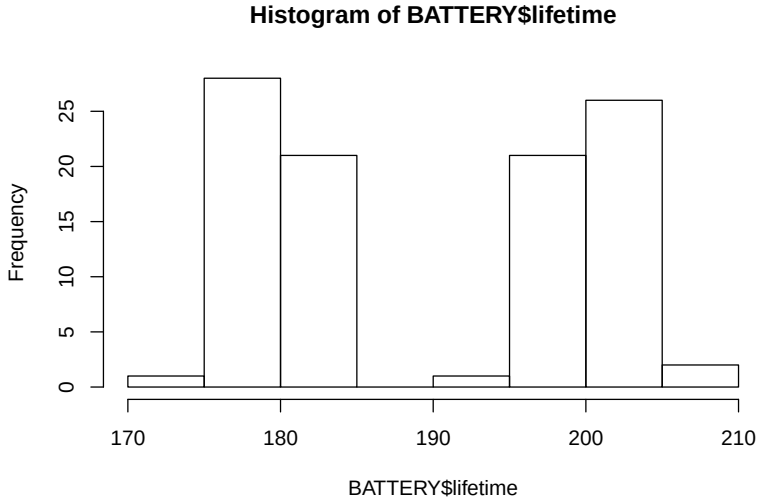
La variable de interés es `lifetime`, pero como sabemos que los datos se refieren a dos tipos distintos de baterías, posiblemente nos interese separarlos. En esta práctica vamos a realizar cálculo de probabilidades basados en este conjunto de datos para que se vea una aplicación, aunque tengamos que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

Actividad 1

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

Solución con R:

```
hist(BATTERY$lifetime)
```



Explicación:
La función hist realiza un histograma básico. Se puede elaborar más, por ejemplo coloreando las barras según el valor de 'facility'. En todo caso, se ve claramente que hay dos poblaciones mezcladas.

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

Solución con R:

```
tapply(datos$lifetime, datos$facility, summary)
```

```
## $A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   194.1   198.8   200.3   200.5   202.9   206.6
##
## $B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   174.2   178.5   179.6   179.7   181.1   183.6
```

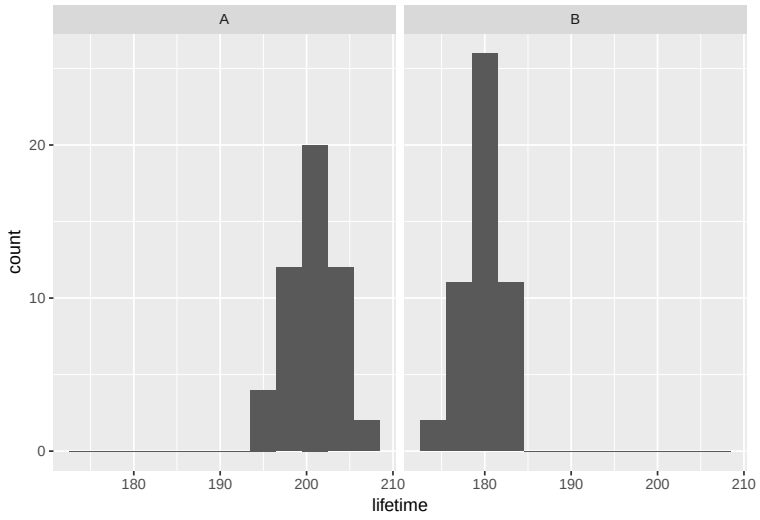
```
datosA <- subset(datos, facility == "A")
datosB <- subset(datos, facility == "B")
```

Explicación:
Está claro que los grupos que se veían en el histograma se corresponden con los dos valores de facility. Creamos los conjuntos filtrando con la función subset (hay otras formas, entre otras usar los corchetes para seleccionar dimensiones).

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

Solución con R:

```
library(ggplot2)
ggplot(datos, aes(lifetime)) + geom_histogram(binwidth=3) + facet_grid(.~facility)
```



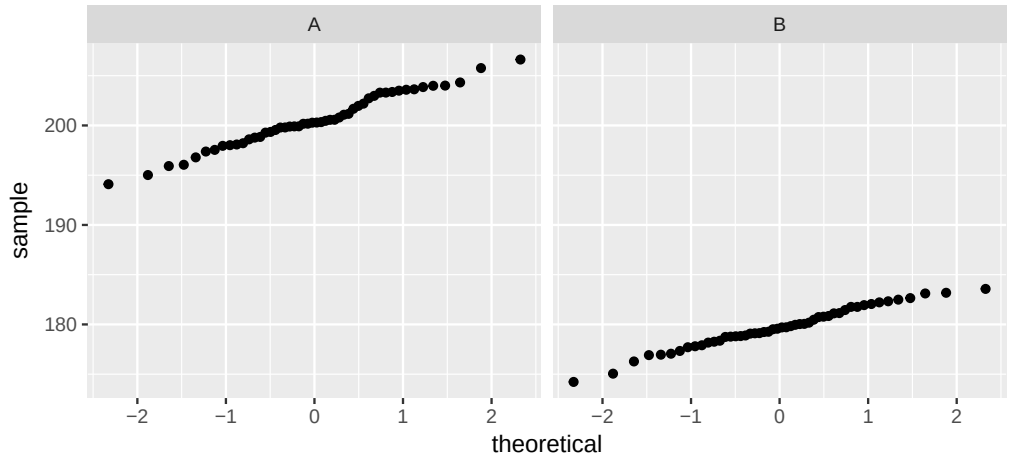

```
# hist(datosA$lifetime)
# hist(datosB$lifetime)
```

Explicación:
A simple vista no se puede confirmar con seguridad, pero sí parece que tienen forma aproximada de campana.

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

Solución con R:

```
ggplot(datos, aes(sample=lifetime)) + geom_qq()  + facet_grid(.~facility)
```




```
tapply(BATTERY$lifetime, BATTERY$facility, shapiro.test)
```

```
## $A
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.9848, p-value = 0.7632
##
##
## $B
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.98395, p-value = 0.7256
```

Explicación:
Los puntos de los gráficos cuantil-cuantil caen aproximadamente sobre una línea recta, pero lo más concluyente es contrastar la hipótesis de normalidad. El test de Shapiro-Wilk tiene un p-valor muy grande para ambos grupos, por lo que no podemos rechazar la normalidad.

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. **Solución con R**:

```
mA <- mean(datosA$lifetime); mA
```

```
## [1] 200.5087
```

```r
mB <- mean(datosB$lifetime); mB
```

```
## [1] 179.6805
```

```r
sA <- sd(datosA$lifetime); sA
```

```
## [1] 2.745777
```

```r
sB <- sd(datosB$lifetime); sB
```

```
## [1] 2.084977
```

Explicación:
Los estadísticos muestrales media y desviación típica son los mejores estimadores. Los guardamos para usarlos después.

- Calcula la probabilidad de que una batería tomada al azar del tipo A dure más de 210 horas **Solución con R**:

```r
pnorm(210, mA, sA, lower.tail = FALSE)
```

```
## [1] 0.0002734129
```

Explicación:
Usamos la función pnorm, con el valor del que queremos calcular la probabilidad, la media y la desviación típica calculada, y como queremos P[X>x], pedimos la cola superior.

- Calcula la probabilidad de que una batería tomada al azar del tipo B dure menos de 175 horas

Solución con R:

```r
(p <- pnorm(175, mB, sB))
```

```
## [1] 0.01238792
```

Explicación:
En este caso sí queremos la cola inferior, luego no decimos nada.

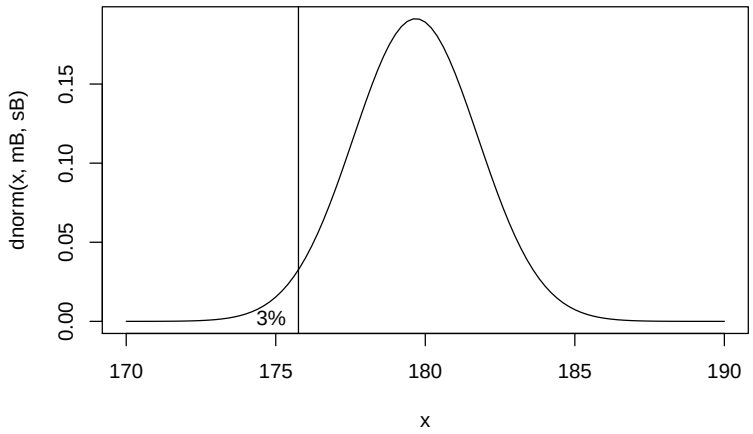
- Encuentra cuál es la duración máxima del 3% de las pilas del tipo B que duran menos (ayuda: esto es equivalente a encontrar el cuantil 0.03)

Solución con R:

```r
qqq <- qnorm(0.03, mB, sB); qqq
```

```
## [1] 175.7591
```

```r
curve(dnorm(x, mB, sB), 170, 190)
abline(v = qnorm(0.03, mB, sB))
text(qqq, 0.001, "3%", pos = 2)
```



Explicación:
Con la ayuda proporcionada la solución es trivial.

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).

Solución con R:

```
dbinom(0, 10, p)
```

```
## [1] 0.8828033
```

Explicación:
La función dbinom nos da justamente la probabilidad de una binomial para el primer argumento, el segundo es n y el tercero, p (lo habíamos guardado en el código anterior).

- 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. Ojo: en R, la variable X representa el número de fracasos hasta el primer éxito)

Solución con R:

```
dgeom(4, p)
```

```
## [1] 0.01178539
```

Explicación:
Como R nos devuelve la probabilidad del número de fracasos hasta el primer éxito, el argumento x de la función dgeom será 4, ya que si p es la probabilidad de éxito, y la primera defectuosa es la quinta, entonces antes ha habido 4 fracasos.

- 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)

Solución con R:

```
phyper(q = 0, m = 3, n = 17, k = 5, lower.tail = FALSE)
```

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

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

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

```
sum(dhyper(x = 1:5, m = 3, n = 17, k = 5))
```

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

Explicación:
Para el cálculo de la hipergeométrica, R espera los argumentos de la siguiente forma: primero el valor de la variable para el que queremos calcular la probabilidad, después el número de unidades que cumplen la característica de interés (en este caso, ser defectuosa), a continuación el número de elementos que no cumplen la característica, en nuestro caso 20-3, y por último el número de extracciones sin reposición que se hacen del conjunto. Como nos piden la probabilidad de que al menos una sea defectuosa, esto equivale a $P(X \geq 1)$, que podemos calcular de cualquiera de las formas indicadas.

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 $1000 \times p$ baterías, y el número de baterías defectuosas por día sigue una distribución de Poisson. Tomemos 12 como ese promedio de baterías defectuosas cada día.

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

Solución con R:

```
ppois(20, 12, lower.tail = FALSE)
```

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

```
1 - ppois(20, 12)
```

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

Explicación:
Siendo X una variable aleatoria que sigue una distribución de Poisson, lo que nos están pididendo es $P(X > 20)$. La función ppois nos devuelve la función de distribción, que es el suceso contrario del que buscamos. Por tanto la podemos calcular restándole ese valor de uno, o con el argumento lower.tail=FALSE.

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

Solución con R:

```r
dpois(0, 12)
```

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

Explicación:
En este caso usamos la función dpois, que nos da $P(X = x)$.

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

Respuesta:

Aplicando la propiedad aditiva de la distribución de Poisson, el número de baterías defectuosas en una semana es la suma de cinco distribuciones de Poisson, y por tanto sigue una distribución de Poisson cuyo parámetro es $\lambda = \sum_{1}^{5} \lambda_i = 5 \cdot 12 = 60$.

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 $b = 185$.

- 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.

Solución con R:

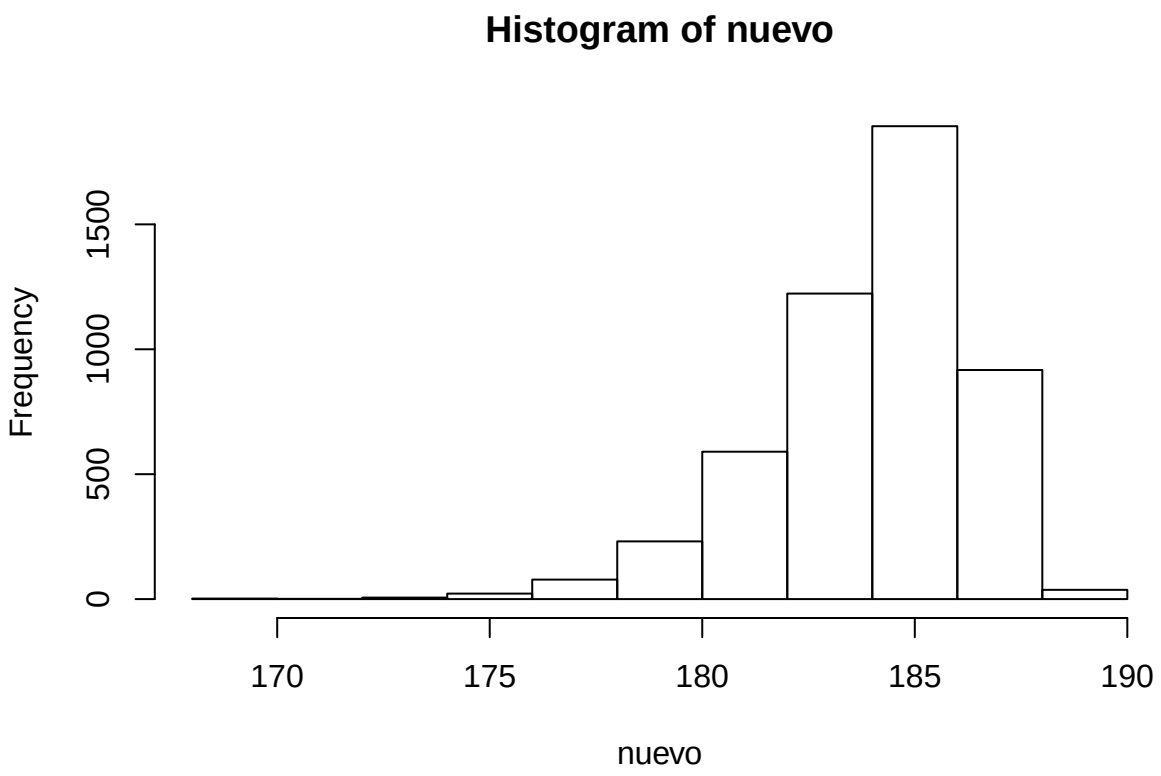
```r
a <- 100; b <- 185
set.seed(1)
nuevo <- rweibull(5*1000, a, b)
```

Explicación:
Guardamos los parámetros y se lo pasamos a la función que simula datos de una variable aleatoria (empiezan por r). el número de valores a simular es 5 días por 1000 baterías al día.

- 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)

Solución con R:

```r
hist(nuevo)
```




```r
mean(nuevo)
```

```
## [1] 183.9514
```

```r
sd(nuevo)
```

```
## [1] 2.394873
```

```r
b*gamma(1+2*a^(-1))
```

```
## [1] 182.9362
```

```r
mB
```

```
## [1] 179.6805
```

Explicación:
Podemos comprobarlo de dos formas: empíricamente con los datos de la simulación, y teóricamente con la expresión de la esperanza de una variable Weibul. Parece que sí se ha aumentado la media, en más de tres horas.

- 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 el nuevo proceso y compárala con el anterior (la p que calculamos en la actividad 2)

Solución con R:

```r
pweibull(175, a, b)
```

```
## [1] 0.003852956
```

Explicación:
Calculamos la probabilidad de ser defectuoso, es decir, que dure menos de 175 horas, y obtenemos un 0.39% de baterías defectuosas, que es menos de la tercera parte que antes (0.39% vs 1.24%). El histograma ya nos indicaba que la media estaba desplazada a la derecha de los valores centrales, lo que en este caso es beneficioso para el producto.

Entrega

Sube un fichero .Rmd con tu solución a cada actividad. Asegúrate de que el fichero compila en alguno de los formatos soportados. Opcionalmente, puedes subir el documento compilado en un formato que elijas (html, pdf, word)

Evaluación

El trabajo se evaluará de 0 a 10 puntos. Cada una de las 17 cuestiones dentro de cada actividad vale 0.5 puntos. En el punto y medio restante se evalúa las explicaciones, presentación, gráficos adicionales, etc.