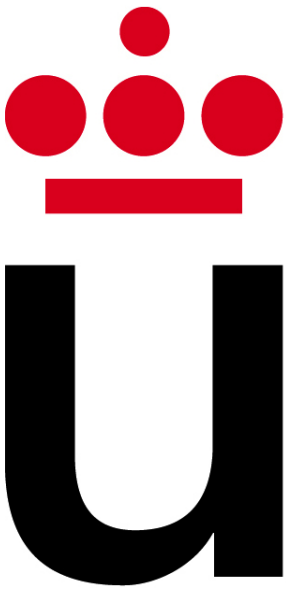
Caso Práctico II

Modelización y tratamiento de la incertidumbre

Probabilidades y variables aleatorias

José Ignacio Escribano



Móstoles, 10 de octubre de 2015

Índice de figuras

Índice

1. Introducción

La empresa WEBSYMASWEBS quiere conocer la calidad de las conexiones, midiendo el número de intentos hasta concertarse a una página web. Para ello contrata a una consultora para que hagan el informe. Además de la conexión, se mide el tráfico que se genera al realizar la conexión, y el grado de satisfacción de los clientes a través de una serie de cuestionarios.

2. Resolución del caso práctico

A continuación resolveremos las tres partes de las que consta el caso práctico: la primera es sobre el número de intentos al acceder a la página web; la segunda es de el tráfico que se registró al acceder a la página web; y la tercera, sobre una encuesta que se hizo para conocer el grado de satisfacción

2.1. Tarea I

La calidad de la conexión a una página web se mide mediante el número de intentos (incluido el exitoso) hasta que se consigue acceder a la página web. Existen tres niveles de calidad: conexión EXCELENTE, BUENA y MALA.
Se tiene una conexión EXCELENTE si se han necesitado 2 ó menos intentos, BUENA si se han necesitado 3 ó 4 intentos y, MALA si se han necesitado más de 5 intentos para conectarse a la página web.

De los 50 datos recogidos por la empresa, 35 corresponden a la conexión EXCELENTE, 13 a la conexión BUENA, y 2 corresponden a la conexión MALA. En porcentaje, el 70 % corresponden con la conexión EXCELENTE, el 26 % con la conexión BUENA y, el 4 % con la conexión MALA.

Para calcular la probabilidad de que un cliente necesite más de 7 intentos necesitamos la frecuencia absoluta de cada uno de los valores del soporte:

Número de intentos	Frecuencia absoluta
1	25
2	10
3	9
4	4
5	2

Así pues, la probabilidad viene dada por

$$P(X \geq 7) = 1 - P(X < 6) = 1 - \sum_{k=1}^{5} P(X = k) = 1 - \left(\frac{25}{50} + \frac{10}{50} + \frac{9}{50} + \frac{4}{50} + \frac{2}{50}\right) = 0$$

2.2. Tarea II

A parte de los datos del número de intentos para acceder a la página web, también se han recogido datos sobre el tráfico de datos que se generó durante la conexión, medidos en kilobytes. También se han creado tres categorías para clasificar a sus clientes: A = "Tráfico BAJO", B = "Tráfico MEDIO" y C = "Tráfico ALTO", pero estos datos se han perdido a un fallo al realizar la copia de seguridad, aunque se recuerdan que los datos seguía una distribución normal de media 250 y desviación típica 60.

La empresa desea que en cada categoría se halle exactamente un tercio de sus clientes. Para obtener los bordes de las distintas regiones, calculamos los percentiles que acumulan a su izquierda un $1/3$ y $2/3$ de probabilidad. Estos datos se corresponden con los valores 224.1564 y 275.8436, respectivamente. En la Figura 1 se muestra la distribución normal, junto con cada una de las regiones.

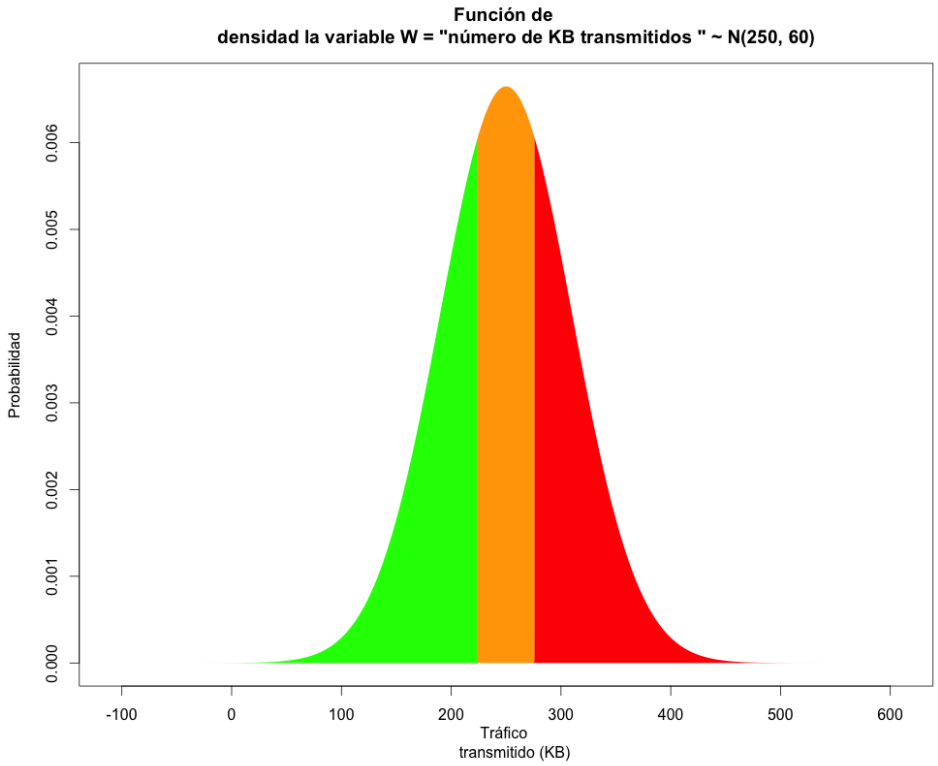


Figura 1: Distribución normal con media 250 y desviación típica 60. Se muestran tres zonas: la verde correspondiente con la categoría Tráfico BAJO; la naranja, con la categoría Tráfico MEDIO; y, la roja, con la categoría Tráfico ALTO. Cada una de las regiones tiene un tercio de probabilidad.

Por tanto, si denotamos con W el número de kilobytes trasmitidos, las categorías vienen dadas por el siguiente criterio:

$$\text{Tráfico} \begin{cases} \text{BAJO} & \text{si } W \leq 224.1564 \\ \text{MEDIO} & \text{si } 224.1564 < W \leq 275.8436 \\ \text{ALTO} & \text{si } W > 275.8436 \end{cases}$$

Además de las regiones, se necesitan calcular una serie de probabilidades:

- Probabilidad de que un cliente tenga una conexión MALA y haya generado un tráfico BAJO.

$$\begin{aligned} P(\text{MALA} \cap \text{BAJO}) &= P(\text{MALA}) \cdot P(\text{BAJO}) \\ &= P(\text{MALA}) \cdot P(W \leq 224.1564) \\ &= 0.04 \cdot 0.333 \\ &= 0.0133 \end{aligned}$$

  La probabilidad $P(\text{MALA})$ viene dada por el porcentaje calculado en la Tarea I, y la probabilidad $P(\text{BAJO})$ se ha calculado con R. Se ha supuesto independencia porque así lo supone el enunciado.

- Probabilidad de que, de 10 clientes seleccionados al azar, al menos 9 de ellos tengan conexión BUENA o EXCELENTE.

  Primero, calculamos la probabilidad de un cliente, elegido al azar, tenga conexión BUENA o EXCELENTE.

$$\begin{aligned} P(\text{BUENA} \cup \text{EXCELENTE}) = P(\text{BUENA}) + P(\text{EXCELENTE}) \\ - P(\text{BUENA} \cap \text{EXCELENTE}) \end{aligned}$$

  Pero, como tener una conexión BUENA y tener una conexión EXCELENTE son sucesos disjuntos tenemos que $P(\text{BUENA} \cap \text{EXCELENTE}) = 0$.

  Por tanto,

$$\begin{aligned} P(\text{BUENA} \cup \text{EXCELENTE}) &= P(\text{BUENA}) + P(\text{EXCELENTE}) \\ &= 0.70 + 0.26 \\ &= 0.96 \end{aligned}$$

  Ambas probabilidades vienen dadas por los porcentajes calculados en la Tarea I.

Nos queda calcular la probabilidad de que al menos 9 de 10 clientes, tengan una conexión BUENA o EXCELENTE.

En este caso, tenemos una distribución binomial con $n = 10$ y $p = 0.96$, que acabamos de calcular.

Por tanto, si llamamos H al número de clientes tenga una conexión BUENA o EXCELENTE,

$$\begin{aligned} P(H \geq 9) &= 1 - P(H < 9) \\ &= 1 - P(H \leq 8) \\ &= 1 - 0.058 \\ &= 0.941 \end{aligned}$$

- Probabilidad de que, de 10 clientes seleccionados al azar, no haya ninguno que tenga conexión EXCELENTE y haya generado un tráfico BAJO.

  Primero, calculamos la probabilidad de un cliente tenga conexión EXCELENTE y haya generado un tráfico BAJO.

$$\begin{aligned} P(\text{EXCELENTE} \cap \text{BAJO}) &= P(\text{EXCELENTE}) \cdot P(\text{BAJO}) \\ &= 0.7 \cdot P(W \leq 224.1564) \\ &= 0.7 \cdot 0.333 \\ &= 0.233 \end{aligned}$$

  De nuevo, volvemos a suponer independencia entre ambas variables.

  Nos queda calcular la probabilidad de que, entre 10 clientes elegidos al azar, no haya ninguno que tenga una conexión EXCELENTE y haya generado un tráfico BAJO.

  Como en el caso anterior tenemos una distribución binomial con $n = 10$ y $p = 0.233$.

  Sea H el número de clientes que tenga una conexión EXCELENTE y haya generado un tráfico BAJO. Entonces, la probabilidad pedida es

$$P(H = 0) = 0.070$$

2.3. Tarea III

Por último, se desea saber el grado de satisfacción de los clientes. Para ello, se distribuyen cuatro cuestionarios (C_1, C_2, C_3 y C_4). El cuestionario C_1 se ha distribuido al 40 % de los clientes, el C_2 al 30 %, el C_3 al 20 % y, el C_4 al 10 %. El porcentaje de clientes que no ha respondido al cuestionario es del 1 %, 2 %, 7 % y 4 %, respectivamente.

De nuevo, se piden una serie de probabilidades relativas a los cuestionarios.

Antes de calcular las probabilidades hacemos el modelo gráfico, que se muestra en la Figura 2.

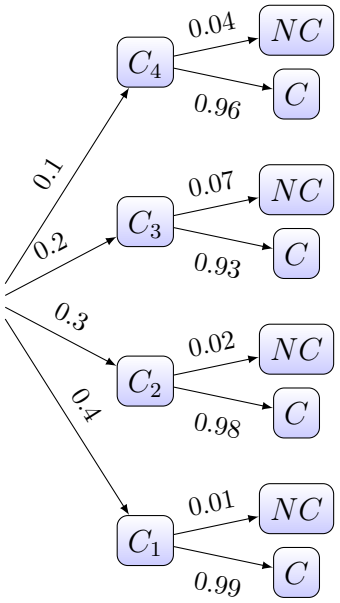


Figura 2: Modelo gráfico de la Tarea III. C_i denota el suceso de recibir el cuestionario C_i con $i = 1, 2, 3, 4$, NC es el suceso de no contestar al cuestionario, y C es el suceso de contestar al cuestionario

- Si elegimos un cliente al azar, ¿cuál es la probabilidad de que haya contestado al cuestionario?

  Si denotamos con C al suceso de contestar el cuestionario, la probabilidad la calculamos aplicando el Teorema de la Probabilidad Total. Así la probabilidad es

$$\begin{aligned} P(C) &= 0.4 \cdot 0.99 + 0.3 \cdot 0.98 + 0.2 \cdot 0.93 + 0.1 \cdot 0.1 \cdot 0.96 \\ &= 0.972 \end{aligned}$$

- Si elegimos un cliente al azar y resulta que no ha contestado a su cuestionario, ¿qué cuestionario es más probable que hubiese recibido, C_1, C_2, C_3 o C_4?

  Si denotamos con NC al hecho de no contestar el cuestionario y con C_i al suceso de recibir el cuestionario C_i. Entonces, aplicamos el teorema de Bayes para cada uno de los cuestionarios.

$$
\begin{aligned}
P(C_1|NC) &= \frac{P(NC|C_1)\cdot P(C_1)}{P(NC)} = \frac{0.01\cdot 0.4}{0.028} = 0.142\\
P(C_2|NC) &= \frac{P(NC|C_2)\cdot P(C_2)}{P(NC)} = \frac{0.02\cdot 0.3}{0.028} = 0.214\\
P(C_3|NC) &= \frac{P(NC|C_3)\cdot P(C_3)}{P(NC)} = \frac{0.07\cdot 0.2}{0.028} = 0.500\\
P(C_4|NC) &= \frac{P(NC|C_4)\cdot P(C_4)}{P(NC)} = \frac{0.04\cdot 0.1}{0.028} = 0.142
\end{aligned}
$$

  La probabilidad $P(NC)$ se puede calcular aplicando el Teorema de la Probabilidad Total, o bien, como suceso complementario al suceso de contestar el cuestionario.

  Por tanto, lo más probable es que contestara al cuestionario C_3.

- Si tomamos 4 clientes al azar, ¿cuál es la probabilidad de que al menos haya un cuestionario sin contestar?

  La probabilidad $P(NC) = 0.028$, ya está calculada anteriormente. Por lo que debemos considerar una distribución binomial con $n = 4$ y $p = 0.028$.

$$
\begin{aligned}
P(NC \geq 1) &= 1 - P(NC < 1)\\
&= 1 - P(NC = 0)\\
&= 1 - 0.892\\
&= 0.107
\end{aligned}
$$

- Misma pregunta que el caso anterior si la muestra es de 200 clientes.

  Repetimos lo mismo que en caso anterior, pero teniendo en cuenta que ahora nuestra distribución binomial tiene como parámetros $n = 200$ y $p = 0.028$.

$$
\begin{aligned}
P(NC \geq 1) &= 1 - P(NC < 1)\\
&= 1 - P(NC = 0)\\
&= 1 - 0.003\\
&= 0.997
\end{aligned}
$$

3. Conclusiones

En este caso práctico hemos puesto en práctica la teoría del cálculo de probabilidades así como los teoremas clásicos de ésta (Teorema de la Probabilidad Total y el Teorema de Bayes). También hemos utilizado dos de las distribuciones más usadas: la distribución normal y la distribución binomial. Para calcular las probabilidades pedidas, nos valimos de R para calcular éstas de forma fácil y eficiente.

4. Código R

```
# Seleccionamos el archivo
filename <- file.choose()

# Cargamos el archivo como CSV
data <- read.csv(filename,sep=" ", header=TRUE)

# Seleccionamos la variable x
x <- data$x

#-------------------------------------------------------------------------------
# Tarea I
#-------------------------------------------------------------------------------

# Definimos un array con las categorías
categories <- c("EXCELENTE", "BUENA", "MALA")

# Contamos el número de datos
n <- length(x)

# Definimos un vector para contar el número de datos de cada categoria
categories_count <- c(0,0,0)

# Recorremos el array
for(i in 1:n) {
  if(x[i] <= 2){ # Si la categia es EXCELENTE, aumentamos en 1 la cuenta
    categories_count[1] <- categories_count[1] + 1
  }else if(x[i] == 3 | x[i] == 4){ # Si la categia es BUENA, aumentamos en 1
                                   # la cuenta
    categories_count[2] <- categories_count[2] + 1
  }else if(x[i] >= 5){ # Si la categia es MALA, aumentamos en 1 la cuenta
    categories_count[3] <- categories_count[3] + 1
  }
}

# Calculamos los porcentajes de cada categoría
categories_percentage <- categories_count/n*100

# Contamos la frecuencia de cada valor
frequencies <- table(x)

# Calculamos la probabilidad de un cliente más de 7 intentos para conectarse
```

```
# a la web P(#intentos > 7) = 1 - P(#intentos <= 6) = 1 - [ P(#intentos = 1)
# + P(#intentos = 2) + P(#intentos = 3) + P(#intentos = 4) + P(#intentos = 5)
# + P(#intentos = 6) ]
1 - sum(table(x)/n)

#-------------------------------------------------------------------------
# Tarea II
#-------------------------------------------------------------------------

# Definimos la distribución normal con paramétros media = 250 y desviación
# típica 60

mean <- 250
sd <- 60

# Calculamos los los percentiles 100/3 y 200/3
p1 <- qnorm(1/3, 250, 60)
p2 <- qnorm(2/3, 250, 60)

# Dibujamos la distribución
curve(xlim=c(mean-6*sd, mean+6*sd), dnorm(x,mean,sd), lwd=0, main="Función de
    densidad la variable W = \"número de KB transmitidos \" ~ N(250, 60)",
    xlab="Tráfico transmitido (KB)", ylab="Probabilidad")

# Dibujamos las regiones "Tráfico bajo", "Tráfico medio" y "Tráfico alto"
cords1.x <- c(mean-6*sd, seq(mean-6*sd, p1, 0.01), p1)
cords1.y <- c(0, dnorm(seq(mean-6*sd, p1, 0.01), mean, sd), 0)
polygon(cords1.x,cords1.y,col='green', border=NA)

cords2.x <- c(p1, seq(p1, p2, 0.01), p2)
cords2.y <- c(0, dnorm(seq(p1, p2, 0.01), mean, sd), 0)
polygon(cords2.x,cords2.y,col='orange', border=NA)

cords3.x <- c(p2, seq(p2, mean+6*sd, 0.01), mean+6*sd)
cords3.y <- c(0, dnorm(seq(p2, mean+6*sd, 0.01), mean, sd), 0)
polygon(cords3.x,cords3.y,col='red', border=NA)

# Comprobamos si efectivamente las zonas tienen un tercio de probabilidad cada
# una. Para ello, simulamos 100000 muestras de nuestra distribución y contamos
# cuántas están cada de las zonas, y calculamos el porcentaje de cada ellas.

n_samples = 100000
samples <- rnorm(n_samples, mean, sd)
```

```
samples_BAJO = sum(samples <= p1)
samples_MEDIO = sum(samples > p1 & samples < p2)
samples_ALTO = sum(samples >= p2)

samples_percentages <- c(samples_BAJO, samples_MEDIO, samples_ALTO)/
  n_samples*100

# Probabilidad de un cliente tenga una conexión MALA y haya generado un
# tráfico BAJO

# P(MALA  BAJO) = P(MALA) * P(BAJO) puesto que dice el enunciado que son
# independientes

# P(MALA  BAJO) = P(MALA) * P(BAJO) = 0.04* P(W <= p1 = 224.15)

(categories_percentage[3]/100) * pnorm(p1, mean, sd)

# Probabilidad de que, de 10 clientes seleccionados al azar, al menos 9
# de ellos tengan conexión BUENA o EXCELENTE

# Primero calculamos la probabilidad de que se tenga una conexión BUENA o
# EXCELENTE.
# Es decir, P(BUENA  EXCELENTE) = P(BUENA) + P(EXCELENTE) -
# P(BUENA  EXCELENTE) = P(BUENA) + P(EXCELENTE),
# puesto que tener conexión BUENA y tener conexión EXCELENTE son
# sucesos disjuntos.

# P(BUENA  EXCELENTE) = P(BUENA) + P(EXCELENTE)
p <- (categories_percentage[2]/100) + (categories_percentage[1]/100)

# Una vez que hemos calculado esta probabilidad, calculamos la probabilidad
# de que al menos 9 de 10 clientes elegidos al azar tengan una conexión
# BUENA o EXCELENTE

# En este caso tenemos una distribución binomial con n = 10 y p = 0.96,
# calculada anteriormente

# P(número de clientes con conexión BUENA o EXCELENTE >= 9)
# = 1 - P(número de clientes con conexión BUENA o EXCELENTE < 9) =
# = 1 - P(número de clientes con conexión BUENA o EXCELENTE <= 8)

1 - pbinom(8, 10, p)
```

```
# Probabilidad de que, de 10 clientes seleccionados al azar, no haya ninguno
# que tenga conexión EXCELENTE y haya generado un tráfico BAJO.

# Calculamos la probabilidad de que haya un cliente que tenga conexión
# EXCELENTE y haya generado un tráfico BAJO.

# P(EXCELENTE  BAJO) = P(EXCELENTE) * P(BAJO), puesto que, según el
# enunciado, los sucesos son independientes.

# P(EXCELENTE  BAJO) = P(EXCELENTE) * P(BAJO) = 0.7 * P(W <= p1 = 224.15)

q <- (categories_percentage[1]/100) * pnorm(p1, mean, sd)

# Falta calcular que de 10 clientes elegidos al azar, no haya ninguno que
# tenga conexión EXCELENTE y haya generado un tráfico BAJO

# De nuevo, tenemos una distribución binomial con n = 10, p = 0.2333333,
# calculada anteriormente. Hay que calcular
# P(número de clientes tenga conexión EXCELENTE y haya generado tráfico
# BAJO = 0)

dbinom(0, 10, q)


#-------------------------------------------------------------------------
# Tarea III
#-------------------------------------------------------------------------

# Definimos las variables con las probabilidades de los cuestionario y los
# que no han sido contestados

C1 = 0.4
C2 = 0.3
C3 = 0.2
C4 = 0.1
C1_NC = 0.01
C2_NC = 0.02
C3_NC = 0.07
C4_NC = 0.04

# Probabilidad de que se haya contestado al cuestionario, si se elige un
# cliente al azar
```

```
# Aplicando el teorema de la probabilidad total, la probabilidad pedida es

C1 * (1 - C1_NC) + C2 * (1 - C2_NC) + C3 * (1 - C3_NC) + C4 * (1 - C4_NC)

# Si un cliente no ha contestado a su cuestionario, ¿qué cuestionario es más
# probable que hubiese recibido?

# Calculamos la probabilidad de que no contestar a un cuestionario

q2 <- C1 * C1_NC + C2 * C2_NC + C3 * C3_NC + C4 * C4_NC

# Aplicamos el teorema de Bayes para cada uno de los cuestionarios.
# P(C_i | NC) = P(NC | C_i) * P(C_i) / P(NC)

c <- c(C1, C2, C3, C4)
c_NC <- c(C1_NC, C2_NC, C3_NC, C4_NC)

bayes <- c*c_NC/q2

# Probabilidad de que al menos haya #un cuestionario sin contestar, si tomamos
# 4 clientes al azar

# La probabilidad de no contestar un cuestionario es:

q2 <- C1 * C1_NC + C2 * C2_NC + C3 * C3_NC + C4 * C4_NC

# Tenemos de nuevo una distribución binomial con n = 4, p = 0.028

# P(Cuestionarios sin contestar >= 1) = 1 - P(Cuestionarios sin contestar < 1) =
# P(Cuestionarios sin contestar = 0)

1 - dbinom(0, 4, q2)

# Si cambiamos 4 por 200, la probabilidad es

1 - dbinom(0, 200, q2)
```