Técnicas y métodos de la ciencia de datos. Master Data Science

Curso 2016/2017

Practica de análisis de datos

Alejandro Pérez Barreiro

Una aseguradora de coches está considerando la idea de establecerse en Móstoles, más concretamente en la zona del Campus de la Universidad Rey Juan Carlos, con la esperanza de conseguir clientes entre los conductores que viven, trabajan o en general que acceden a esta zona. La compañía tiene la teoría de que la póliza de los coches de color rojo debería ser más alta debido a la reciente publicación de un artículo en la revista Noticias del Mundo, donde se argumenta que aquellos conductores "más agresivos" y por lo tanto más propensos a tener accidentes prefieren el rojo como color de coche. Debido a este hecho la compañía pretende fijar en 900 euros la póliza para coches rojos y en 600 euros la póliza para coches de otro color. Dado que la aseguradora desconoce las probabilidades de tener un accidente con un coche rojo o de otro color, ha decidido llevar a cabo un muestreo examinando el número de coches accidentados de uno y otro color. Una vez llevado a cabo el estudio se han obtenido los siguientes datos:

	Accidentados	No accidentados
Rojos	16	30
Otros colores	40	110

- Calcula un intervalo de confianza al 98% para la probabilidad de coches rojos accidentados. Obtén el intervalo análogo para el caso de la probabilidad de accidentado entre los coches de otro color.

Solución: Tenemos dos muestras aleatorias simples de una v.a con distribución *Ber(p)*: $X_1, \dots, X_n$. En este caso el tamaño de muestra es suficientemente grande tanto para coches rojos, *nr = 16 + 30 = 46*, como para coches de otro color, *no = 40 + 110 = 150*. Por tanto aplicando el Teorema Central del Límite se tiene que:

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0,1) \;\; con\, \hat{p} = \; \bar{X}$$

Y su respectivo intervalo de confianza será:

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Calculemos estos intervalos con R:

```
nr = 46
no = 150
pr = 16/nr # Probabilidad coches rojos accidentados
po = 40/no # Probabilidad coches otro color accidentados
intr = c(pr - qnorm(0.99,0,1)*sqrt((pr*(1-pr))/nr),pr + qnorm(0.99,0,1)*sqrt((pr*(1-pr))/nr)); intr
into = c(po - qnorm(0.99,0,1)*sqrt((po*(1-po))/no),po + qnorm(0.99,0,1)*sqrt((po*(1-po))/no)); into
```

```
[1] 0.1844613 0.5111909
[1] 0.1826696 0.3506637
```

Por tanto el intervalo de confianza será *[0.1844613, 0.5111909]* para los coches rojos y *[0.1826696, 0.3506637]* para los de otro color.

- Plantea y realiza el contraste de hipótesis que permita determinar si hay evidencias para pensar que la probabilidad de tener accidentes entre los coches rojos es mayor que entre los coches de otro color.

Solución: Al igual que en el ejercicio anterior tenemos dos m.a.s Bernouilli, la de coches rojos y los de otro color, con probabilidad de éxito p_r y p_o respectivamente. El objetivo será comparar ambas proporciones poblacionales. Puesto que cada proporción muestral sigue aproximadamente una normal:

$$\frac{\hat{p}_r - p_r}{\sqrt{\hat{p}_r(1-\hat{p}_r)/nr}} \approx N(0,1), \qquad \frac{\hat{p}_o - p_o}{\sqrt{\hat{p}_o(1-\hat{p}_o)/no}} \approx N(0,1),$$

con $\hat{p}_r = \frac{\sum X_i}{nr}$ y $\hat{p}_r = \frac{\sum Y_i}{no}$ se puede construir un estadístico de contraste y resolver un contraste que compare ambas proporciones teóricas a partir de las dos muestras.

El contraste a realizar en este caso es bilateral, H_0: $p_r = p_o$ frente a H_1: $p_r \neq p_o$

Este contraste se hace en R con la función *prop.test*:

```
```{r}
prop.test(x=c(16,40), n=c(46,150), alternative="two.sided")
```
```

```
        2-sample test for equality of proportions with continuity correction

data:  c(16, 40) out of c(46, 150)
X-squared = 0.77335, df = 1, p-value = 0.3792
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.08780712  0.25012596
sample estimates:
   prop 1    prop 2
0.3478261 0.2666667
```

Por tanto como el p-valor es 0.3792 que es mayor que el error de tipo I establecido (5%), se tiene que no se rechaza la hipótesis nula, es decir, no hay evidencias para asegurar que la proporción de accidentados de coches rojos sea distintas que la de los coches de color

- Supón ahora que se desea realizar un análisis Bayesiano. A priori, dado que se desconoce la probabilidad de sufrir accidente con coches de color rojo o de otro color, se asume una distribución inicial Uniforme (0,1) (es decir, una Be(1, 1)). Obtén la distribución a posteriori para la proporción de coches rojos que tienen accidentes, y también la distribución a posteriori para la proporción de accidentes entre coches de otro color.

Solución: Como sabemos por la teoría:

Posteriori = priori x verosimilitud

En nuestro caso la distribución a priori es una *Beta(1,1), calculemos la de verosimilitud:*

$$P(X = x|p_r) = \binom{n_r}{x} {p_r}^x (1 - p_r)^{n_r - x} \propto {p_r}^x (1 - p_r)^{n_r - x}$$

Por tanto como si p_r sigue una distribución *Beta(1,1)* entonces $f(p_r) = 1$, se tiene que:

$$posteriori = verosimilitud = {p_r}^x (1 - p_r)^{n_r - x} = Be(x + 1, n_r - x + 1)$$

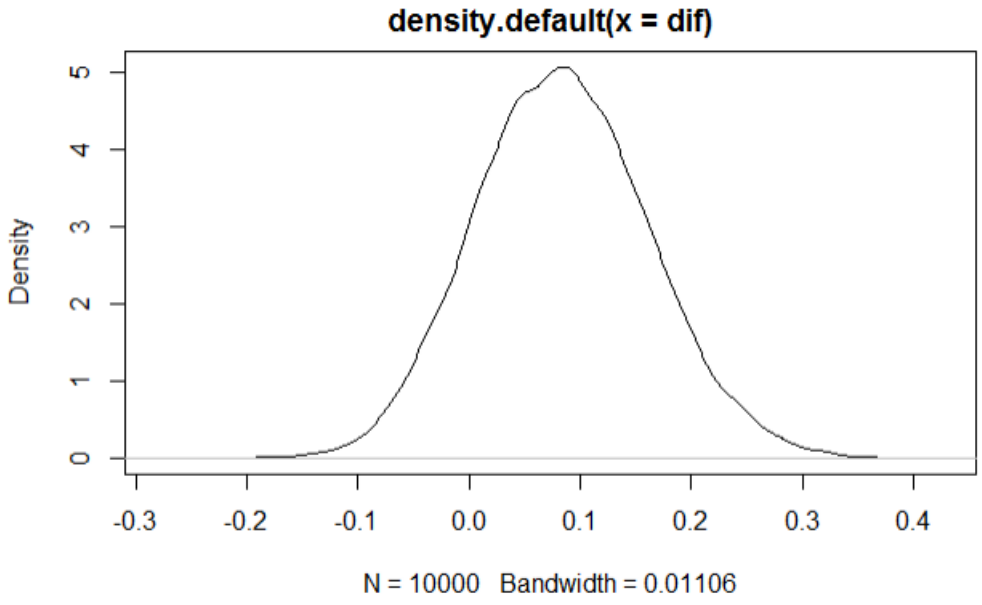
Por la tanto la distribución a posteriori para la proporción de coches rojos accidentados será una $Be(16+1,46-16+1)=Be(17,31)$.

La distribución a posteriori para los coches de otro color se haría de forma análoga obteniéndose una $Be(40+1,150-40+1)=Be(41,111)$

- Usa simulación para aproximar la distribución a posteriori de la diferencia entre la proporción de accidentes entre coches rojos y de otro color, obtén un intervalo de credibilidad al 98% para esta diferencia. ¿Qué se puede decir a la luz de los resultados obtenidos?

Solución: Para este ejercicio primero voy a simular 10000 valores de una Be(17,31) y de una Be(41,111) para después calcular su diferencia. A partir de esa diferencia calculare la media y la varianza de esos datos y los representare gráficamente para ver a que distribución se aproxima. Por ultimo calculare el cuantil 0.025 y 0.975 para el intervalo.

```{r}
p1 <- rbeta(10000,17,31)
p2 <- rbeta(10000,41,111)
dif <- p1 - p2
mean(dif)
var(dif)
plot(density(dif))
quantile(dif,c(0.025,0.975))
```



```
[1] 0.08390404
[1] 0.0060117
      2.5%      97.5%
-0.0622778  0.2421375
```

En función a los resultados se puede decir que la distribución a posteriori se aproxima a un normal de media 0.084 y varianza 0.006.

En cuanto al intervalo de confianza del 95% será [-0.062, 0.242]. Puesto que este intervalo contiene al 0 no se puede decir que haya diferencia probabilística entre la proporción de coches rojos accidentados y coches de otro color.