Ampliación de Inferencia Estadística

Tercero Grado de Estadística

Universidad de Sevilla

Tema 1: Introducción a la Inferencia Bayesiana

Práctica-R-1

Problema 1.

Se sabe que el número de personas que entran en un bar sigue una distribución binomial de parámetros $n = 50$ y $p = 0.5$. Si i es el número de personas que entran se sabe que el número de mujeres sigue una distribución binomial de parámetros i y $p = 0.5$. Si un día hay 20 mujeres, se pide:

1. ¿cuál es la probabilidad de que en el bar haya 40 personas?
2. Calcular el número más probable de personas en el bar.

SOLUCIÓN.

Apartado 1.

Definamos las siguientes variables aleatorias discretas a partir del enunciado:

$$N = \text{ número de personas que entran en el bar} \sim Bi(50, 0.5)$$

$$N_{muj} = \text{Número de mujeres que entran en el bar}$$

Se sabe que $(N_{muj}|N = i) \sim Bi(i, 0.5)$. En concreto, en el apartado 1 nos piden calcular la siguiente probabilidad

$$P(N = 40|N_{muj} = 20).$$

Si un día hay 20 mujeres entonces al menos hay 20 personas como es obvio. Por tanto, aplicando el Teorema de Bayes para variables aleatorias discretas nos queda

$$\begin{aligned} P(N = 40|N_{muj} = 20) &= \frac{P(N = 40)P(N_{muj} = 20|N = 40)}{\sum_{k=20}^{50} P(N = k)P(N_{muj} = 20|N = k)} \\ &= \frac{\binom{50}{40}0.5^{50}\binom{40}{20}0.5^{40}}{\sum_{k=20}^{50}\binom{50}{k}0.5^{50}\binom{k}{20}0.5^{k}} \\ &= \frac{\binom{50}{40}\binom{40}{20}0.5^{40}}{\sum_{k=20}^{50}\binom{50}{k}\binom{k}{20}0.5^{k}} \end{aligned}$$

Para poder calcular esta probabilidad lo haremos en R.

```
s=is.vector(NA)
for (k in 20:50){
  s[k]=choose(50,k)*choose(k,20)*0.5^k}
round(print(s[40]/sum(s[20:50])),5)
```

```
## [1] 0.0001494289
```

```
## [1] 0.00015
```

Apartado 2.

Para ello usaremos el siguiente comando de R

```
ko=which.max(s)
sprintf("El número máximo de personas es %s",ko)
```

```
## [1] "El número máximo de personas es 30"
```

```
sprintf("con una probabilidad de %s",round(s[ko]/sum(s[20:50]),4))
```

```
## [1] "con una probabilidad de 0.153"
```

Problema 2.

Se sabe que los pacientes ingresados en UCI por la COVID-19 presentan una concentración de glucosa en sangre en grs/cc según una normal de media $\mu = 1$ y desviación típica $\sigma = 0.1$. En función de dicha concentración (un valor x) se sabe que la presión diastólica sigue una normal de media $\mu = 10x$ y desviación típica $\sigma = 1$. Si a un paciente se le observa una presión diastólica de 11.5, ¿cuál es la probabilidad de que tenga hipoglucemia?(Se entiende que hipoglucemia es una concentración de glucosa inferior a 0.8)

SOLUCIÓN.

Al igual que en el problema anterior se definen en primer lugar las variables aleatorias correspondientes. En este caso se tratan de v.a. absolutamente continuas de tipo Normal.

$$X = \text{ Concentración de glucosa en sangre en grs/cc} \sim N(\mu = 1, \sigma = 0.1);$$
$$Y = \text{Presión diastólica, donde se sabe que } Y|X = x \sim N(\mu = 10x, \sigma = 1).$$

Hay que calcular la siguiente probabilidad:

$$P(X \leq 0.8|Y = 11.5)$$

Dada la información que nos da el enunciado acerca de la densidad de Y condicionada a $X = x$ y como nos piden una probabilidad para X condicionada a un valor fijo de Y, esto nos obliga a calcular en primer lugar la función de densidad de X condicionada a $Y = 11.5$ mediante el teorema de Bayes para v.a. continuas y una vez que conozcamos esta densidad obtener el valor que nos piden valorando en su función de distribución el valor de 0.8 mediante el comando de R que nos lo proporciona.

Sabemos que el teorema de Bayes para v.a. continuas es:

$$f(x|y) = \frac{g(x)f(y|x)}{\int g(x)f(y|x)dx}.$$

Donde el valor de la integral del denominador juega el papel de constante normalizadora del numerador. Entonces, lo primero será ver si la expresión del numerador corresponde con la de alguna densidad conocida.

$$\begin{aligned} f(x|y = 11.5) &\propto \frac{1}{\sqrt{2\pi}\cdot 0.1}\exp\left\{-\frac{(x-1)^2}{2\cdot 0.1^2}\right\} \times \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(11.5-10x)^2}{2}\right\} \\ &\propto \exp\left\{-\frac{(x-1)^2}{2\cdot 0.1^2}\right\} \times \exp\left\{-\frac{(11.5-10x)^2}{2}\right\} \end{aligned}$$

El símbolo $\propto$ significa *proporcional a*. Esto nos permite ir eliminando todo lo que sea constante como ocurre en este caso con $1/(\sqrt{2\pi}\cdot 0.1)$ y $1/\sqrt{2\pi}$. Para ello iremos buscando el formato general de la densidad normal:

$$f(x) \propto \exp\left\{-\frac{1}{2}\frac{(x-\mu_0)^2}{\sigma_0^2}\right\}$$

Luego se tendría que desarrollando ahora el exponente, simplificando mediante un álgebra sencilla y quedándonos únicamente con lo que depende de x puesto que el resto sería constante y por paso a la proporcionalidad se pueden eliminar dichas constantes, nos quedaría:

$$\begin{aligned}
f(x|y=11.5) &\propto \exp\left\{-\frac{1}{2}\left[\frac{(x-1)^2}{0.1^2}+\frac{(10x-11.5)^2}{1}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{x^2+1-2x+0.1^2(100x^2+11.5^2-20\cdot 11.5x)}{0.1^2}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{2x^2-2x(1+0.1^2\cdot 10\cdot 11.5))}{0.1^2}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{x^2-2x(\frac{2.15}{2}))}{0.1^2/2}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{(x-1.075)^2}{0.1^2/2}\right]\right\}
\end{aligned}$$

Es decir, $X|Y=11.5 \sim N(\mu_0=1.075, \sigma_0^2=0.1^2/2)$. Por tanto, ahora con R ya podemos calcular la probabilidad que nos piden

```
round(pnorm(0.8, mean=1.075,sd=0.1/sqrt(2)),4)
```

```
## [1] 1e-04
```

Luego, una persona que tenga una presión diástolica de 11.5 es prácticamente imposible que tenga hipoglucemia. Hay que recordar que todos los valores dados en el enunciado son ficticios y puede que el resultado final no tenga nada que ver con la realidad.

Problema 3

Supongamos que las calificaciones en una evaluación son las siguientes:

$$Suspenso=1, Aprobado=2, Notable=3, Sobresaliente=4, MH=5$$

cuyas probabilidades a priori son 0.4,0.3,0.2,0.07,0.03 respectivamente. En caso la probabilidad de acertar una pregunta son respectivamente 0.15,0.5,0.65,0.9,0.999. Si corregimos un examen con tres respuestas acertadas, ¿qué calificación sería la más probable?

SOLUCIÓN.

Así creamos el vector de probabilidades a priori

```
p<-c(0.4,0.3,0.2,0.07,0.03)
```

Si suponemos que no tenemos información entonces `p<-c(0.2,0.2,0.2,0.2,0.2)`

A continuación creamos el vector de las respuestas condicionadas a la clasificación

```
pAc<-c(0.15,0.5,0.65,0.9,0.999)
```

Si corregimos un examen con tres respuestas acertadas, ¿qué calificación es la más probable?

Ahora el vector de tres respuestas correctas condicionadas a cada clasificación, es decir la verosimilitud es

```
pAAAc<-pAc^3
```

Ahora calculamos el numerador del teorema de Bayes

```
lcAAA<-p*pAAAc
```

La suma de todas nos daría la constante normalizadora

```
cte<-sum(lcAAA)
```

Por último, las probabilidades a posteriori serían el cociente:

```
aposterioric<-lcAAA/cte
print(round(aposterioric,2))
```

```
## [1] 0.01 0.21 0.31 0.29 0.17
```

Supongamos ahora que la entrevista consta de 10 preguntas y que la probabilidad de respuesta correcta en cada una de ellas no varía. Queremos construir una tabla donde por filas venga dada la calificación de 1 a 5 codificada como antes y por columnas el número de respuestas correctas, y cada entrada de la tabla se lea como sabiendo que ha respondido a j respuestas correctas nos da la probabilidad de clasificación. Ahora bien, debido a que puede ocurrir que no acierte ninguna pregunta y por tanto el número de acertadas sea 0 vamos a codificar las respuestas correctas entre 1 y 11.

Primero vamos a definir la tabla como una matrix de 5x10

```
prob<-matrix(NA,nrow=5,ncol=11)
```

Ahora bien, el número de respuestas correctas seguirá una binomial de parámetros 10 y la probabilidad correpondiente. Para facilitar el cálculo primero vamos a obtener el vector de constantes normalizadoras.

```
for (j in 1:11){
  cte[j]=0
for (i in 1:5){

  cte[j]=cte[j]+dbinom(j-1,10,pAc[i])*p[i]
  }}
```

A continuación se completa la tabla

```
for (j in 1:11){
  for (i in 1:5){
    prob[i,j]=p[i]*dbinom(j-1,10,pAc[i])/cte[j]
      } }
prob
```

```
##              [,1]         [,2]         [,3]         [,4]         [,5]
## [1,] 9.962240e-01 9.786472e-01 8.871378e-01 5.686297e-01 1.755632e-01
## [2,] 3.706202e-03 2.063127e-02 1.059787e-01 3.849325e-01 6.734669e-01
## [3,] 6.979402e-05 7.215399e-04 6.883324e-03 4.643117e-02 1.508643e-01
## [4,] 8.855351e-11 4.436547e-09 2.051069e-07 6.704846e-06 1.055755e-04
## [5,] 3.795151e-31 2.110529e-27 1.083052e-23 3.929895e-20 6.868757e-17
##              [,6]         [,7]         [,8]         [,9]        [,10]
## [1,] 3.143518e-02 4.526663e-03 5.616919e-04 5.386419e-05 2.914655e-06
## [2,] 6.833235e-01 5.575916e-01 3.920703e-01 2.130559e-01 6.532933e-02
## [3,] 2.842772e-01 4.308015e-01 5.625618e-01 5.677343e-01 3.232996e-01
## [4,] 9.640857e-04 7.080240e-03 4.480621e-02 2.191343e-01 6.047384e-01
## [5,] 6.962316e-14 5.675565e-11 3.986780e-08 2.164300e-05 6.629756e-03
##             [,11]
```

```
## [1,] 4.039982e-08
## [2,] 5.131308e-03
## [3,] 4.715963e-02
## [4,] 4.274939e-01
## [5,] 5.202151e-01
```