2º Grado en Ingeniería Informática – Grupo A (Estándar)

Asignatura: Estadística

Profesor encargado: Prof. Dr. Ángel Serrano Sánchez de León

Examen liberatorio de materia
Fecha: 21/04/2015

Nombre y apellidos:

Indicar y explicar todos los pasos intermedios.

1. [1,5 puntos] Sea X una variable aleatoria discreta con la siguiente función de masa:

x	-3	6	9
f(x)	1/6	1/2	1/3

Se pide:

- **a)** Comprobar que f(x) es una función de masa de probabilidad.
- **b)** Calcular la función de distribución F(x).
- **c)** Calcular E(X).
- **d)** Calcular E(X^2).
- **e)** Calcular Var(X)
- **f)** Calcular de dos maneras diferentes E[$(2X + 1)^2$].

Solución:

a) f(x) será una función de masa de probabilidad correcta si la suma total es 1.

```
> x <- c(-3,6,9)
> f <- c(1/6,1/2,1/3)
> sum(f)
[1] 1
```

b) La función de distribución F(x) se calcula con la suma acumulada:

```
> F <- cumsum(f)
> F
[1] 0.1666667 0.6666667 1.0000000
```

x	-3	6	9
F(x)	1/6	2/3	1

Recordemos que R es sensible a mayúsculas y minúsculas, con lo que los vectores f y F son diferentes.

c) El valor esperado E(*X*) es:

$$\mathrm{E}(X) = \mu_X = \sum_{i=1}^{3} x_i f(x_i)$$

```
> e <- sum(x*f)
> e
[1] 5.5
```

d) El valor esperado E(X^2) es:

$$\mathrm{E}(X^2) = \mu_{X^2} = \sum_{i=1}^{3} x_i^2 f(x_i)$$

```
> e2 <- sum(x^2 * f)
> e2
[1] 46.5
```

e) La varianza Var(*X*) es:

$$\mathrm{Var}(X) = \sum_{i=1}^{3} (x_i - \mu_X)^2 f(x_i)$$

```
> v <- sum((x-e)^2 * f)
> v
[1] 16.25
```

Otra manera para calcular la varianza es:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \mu_{X^2} - \mu_X^2$$

```
> e2 - e^2
[1] 16.25
```

f) Para calcular este valor esperado, podemos operar un poco en la expresión y utilizar las propiedades de linealidad del valor esperado:

$$\mathrm{E}[(2X+1)^2] = \mathrm{E}(4X^2 + 4X + 1) = \mathrm{E}(4X^2) + \mathrm{E}(4X) + \mathrm{E}(1) = 4\mathrm{E}(X^2) + 4\mathrm{E}(X) + 1$$
$$= 4\mu_{X^2} + 4\mu_X + 1$$

```
> 4*e2 + 4*e + 1
[1] 209
```

Segunda manera: vamos a crear una nueva variable *Y* que valga precisamente $(2X + 1)^2$.

```
> y <- (2*x+1)^2
> y
[1]  25 169 361
```

Esta variable *Y* tiene la misma función de masa f(*x*) que *X*. Luego el valor esperado de *Y* será:

$$\mathrm{E}(Y) = \mu_Y = \sum_{i=1}^{3} y_i f(y_i)$$

```
> sum(y*f)
[1] 209
```

2. [1,5 puntos] Sean *X* e *Y* dos variables aleatorias con una densidad de probabilidad conjunta:

$$\mathrm{f}(x,y) = \begin{cases} 4xy & \text{si } 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{resto} \end{cases}$$

Se pide:

- **a)** Comprobar que f(*x*,*y*) es una función de densidad de probabilidad (paquete "pracma").
- **b)** Representar gráficamente f(*x*,*y*) entre $0 \le x \le 1$, $0 \le y \le 1$.
- **c)** Calcular la probabilidad de que *X* tome un valor en el rango entre 0 y 1/2, y que *Y* tome un valor en el rango entre 1/4 y 1/2.
- **d)** Calcular E(*X*) y E(*Y*).
- **e)** Calcular Var(*X*) y Var(*Y*).
- **f)** Calcular Cov(*X*,*Y*).

Solución:

a) Al ser una función de dos variables, f(*x*,*y*) representa una superficie en el espacio tridimensional. Para que f(*x*,*y*) sea una función de densidad de probabilidad correcta, el volumen debajo de dicha superficie debe ser igual a 1. Debemos comprobar que la siguiente integral doble vale 1:

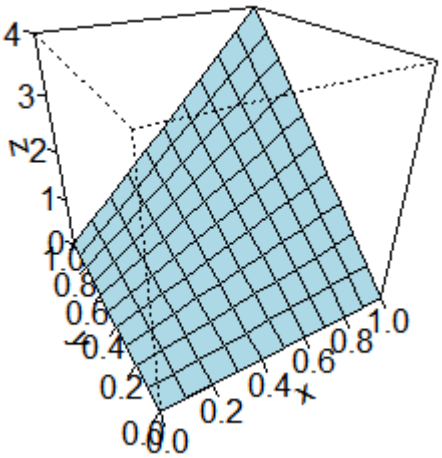
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{0}^{1}\int_{0}^{1} f(x,y)dxdy = 1$$

```
> require("pracma")
> f <- function(x,y) {4*x*y}
> integral2(f,0,1,0,1)
$Q
[1] 1

$error
[1] 5.551115e-17
```

b) Representemos gráficamente esta función:

```
> x <- seq(0,1,0.1)
> y <- x
> z <- outer(x,y,f)
> persp(x,y,z,col="lightblue",ticktype="detailed",theta=-30,phi=30)
```



c) Calcular esta probabilidad se realiza mediante la integral doble correspondiente:

$$P\left(0 \le X \le \frac{1}{2}, \frac{1}{4} \le Y \le \frac{1}{2}\right) = \int_{0}^{\frac{1}{2}} \int_{\frac{1}{4}}^{\frac{1}{2}} f(x,y)dxdy$$

```
> integral2(f,0,1/2,1/4,1/2)
$Q
[1] 0.046875

$error
[1] 1.734723e-18
```

Luego es una probabilidad del 4,6 %.

d) Los valores esperados son:

$$\mathrm{E}(X) = \mu_X = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x,y)dxdy = \int_{0}^{1} \int_{0}^{1} xf(x,y)dxdy$$

$$\mathrm{E}(Y) = \mu_Y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x,y)dxdy = \int_{0}^{1} \int_{0}^{1} yf(x,y)dxdy$$

```
> fx <- function(x,y) {f(x,y)*x}
> integral2(fx,0,1,0,1)
$Q
[1] 0.6666667

$error
[1] 1.873501e-16
```

```
> fy <- function(x,y) {f(x,y)*y}
> integral2(fy,0,1,0,1)
```

```
$Q
[1] 0.6666667

$error
[1] 9.020562e-17
```

Era previsible que los dos valores esperados fueran iguales ya que la función f(*x*,*y*) es invariable al cambio *x* ← → *y*.

e) Las varianzas son:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)^2 f(x,y)dxdy = \int_{0}^{1}\int_{0}^{1}(x-\mu_X)^2 f(x,y)dxdy$$

$$\mathrm{Var}(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-\mu_Y)^2 f(x,y)dxdy = \int_{0}^{1}\int_{0}^{1}(y-\mu_Y)^2 f(x,y)dxdy$$

```
> fvx <- function(x,y) {f(x,y)*(x-0.6666667)^2}
> integral2(fvx,0,1,0,1)
$Q
[1] 0.05555556

$error
[1] 1.734723e-17

> fvy <- function(x,y) {f(x,y)*(y-0.6666667)^2}
> integral2(fvy,0,1,0,1)
$Q
[1] 0.05555556

$error
[1] 1.12757e-17
```

Luego ambas varianzas valen 0,05. De nuevo los dos valores obtenidos son iguales por la simetría de la función.

f) La covarianza entre *X* e *Y* es:

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy = \int_{0}^{1}\int_{0}^{1}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy$$

```
> fcov <- function(x,y) {f(x,y)*(x-0.6666667)*(y-0.6666667)}
> integral2(fcov,0,1,0,1)
$Q
[1] 1.111958e-15

$error
[1] 0
```

Luego la covarianza sale 0.

3. [2 puntos] En la jornada 33 de la Liga de Fútbol Profesional española, que tendrá lugar el próximo 26 de abril de 2015, se enfrentan el Celta de Vigo y el Real Madrid, en el estadio del primero. La siguiente tabla resume los resultados de los últimos encuentros entre ambos equipos, tanto en la Liga como en la Copa del Rey. Con estos datos, se pide calcular:

a) El porcentaje de veces que el Celta de Vigo ha ganado contra el Real Madrid, así como el porcentaje de empates entre ambos equipos y el porcentaje de derrotas del Celta frente al Real Madrid.

b) De las ocasiones en las que el Celta ha ganado al Real Madrid, calcular el porcentaje de veces que el partido se celebró en el estadio del Celta. Repetir el cálculo para considerar los empates entre ambos equipos y las derrotas del Celta frente al Real Madrid, siempre en el caso de que el partido se celebrase en el estadio del Celta.

c) Dividiendo el espacio de probabilidad en el conjunto completo de tres sucesos "victoria del Celta", "empate" y "derrota del Celta", se pide calcular y dibujar el árbol de probabilidades.

d) Suponiendo que no hay otras variables que influyan en el resultado del encuentro, calcular la probabilidad de que el Celta gane en casa el próximo partido contra el Real Madrid, así como la probabilidad de que se produzca un empate. ¿Tienen sentido los resultados? **Será imprescindible comentarlos de manera crítica.**

NOTA: Para cargar los datos en RStudio, utilizar el siguiente comando (recordando que las barras de separación de directorios son las inclinadas hacia la derecha):

> futbol <- read.csv("ruta/dataset-celta-realmadrid.csv")

Liga	28/11/1999	**Celta de Vigo**	**1**	0	Real Madrid
Liga	09/04/2000	**Real Madrid**	**1**	0	Celta de Vigo
Liga	10/12/2000	**Real Madrid**	**3**	0	Celta de Vigo
Liga	05/05/2001	**Celta de Vigo**	**3**	0	Real Madrid
Liga	21/10/2001	Real Madrid	1	1	Celta de Vigo
Liga	02/03/2002	Celta de Vigo	0	**1**	**Real Madrid**
Liga	11/01/2003	Celta de Vigo	0	**1**	**Real Madrid**
Liga	31/05/2003	Real Madrid	1	1	Celta de Vigo
Liga	18/10/2003	Celta de Vigo	0	**2**	**Real Madrid**
Liga	29/02/2004	**Real Madrid**	**4**	2	Celta de Vigo
Liga	10/09/2005	Real Madrid	2	**3**	**Celta de Vigo**
Liga	29/01/2006	Celta de Vigo	1	**2**	**Real Madrid**
Liga	05/11/2006	Real Madrid	1	**2**	**Celta de Vigo**
Liga	01/04/2007	Celta de Vigo	1	**2**	**Real Madrid**
Liga	20/10/2012	**Real Madrid**	**2**	0	Celta de Vigo
Copa del Rey	12/12/2012	**Celta de Vigo**	**2**	1	Real Madrid
Copa del Rey	09/01/2013	**Real Madrid**	**4**	0	Celta de Vigo
Liga	10/03/2013	Celta de Vigo	1	**2**	**Real Madrid**
Liga	06/01/2014	**Real Madrid**	**3**	0	Celta de Vigo
Liga	11/05/2014	**Celta de Vigo**	**2**	0	Real Madrid
Liga	06/12/2014	**Real Madrid**	**3**	0	Celta de Vigo

Solución:

En este ejercicio los cálculos se basan en contar datos de la tabla. El conteo se puede hacer a mano, o bien mediante comandos de R.

a) Consideraremos las dos variables aleatorias siguientes:

- *E* = estadio en el que se juega el partido. Toma dos valores: Celta ("C") o Real Madrid ("RM").
- *G* = qué equipo gana el partido. Toma tres valores: Celta ("C"), Real Madrid ("RM") o empate ("0").

Nos preguntan cuántas victorias, empates y derrotas se han producido del Celta sobre el Real Madrid.

Calculemos cuántas veces ha jugado el Celta como equipo local:

```
> celta_local <- futbol[futbol$LOCAL=="Celta de Vigo",]
```

De estas, veamos cuántas veces ha ganado, empatado o perdido.

```
> victoria_local <-
sum(celta_local$RESULTADO_LOCAL>celta_local$RESULTADO_VISITANTE)
> empate_local <-
sum(celta_local$RESULTADO_LOCAL==celta_local$RESULTADO_VISITANTE)
> derrota_local <-
sum(celta_local$RESULTADO_LOCAL<celta_local$RESULTADO_VISITANTE)
```

Veamos ahora cuántas veces ha jugado el Celta como visitante en el estadio del Real Madrid.

```
> celta_visitante <- futbol[futbol$VISITANTE=="Celta de Vigo",]
```

De estas, veamos cuántas veces ha ganado, empatado o perdido.

```
> victoria_visitante <-
sum(celta_visitante$RESULTADO_LOCAL<celta_visitante$RESULTADO_VISITANT
E)
> empate_visitante <-
sum(celta_visitante$RESULTADO_LOCAL==celta_visitante$RESULTADO_VISITAN
TE)
> derrota_visitante <-
sum(celta_visitante$RESULTADO_LOCAL>celta_visitante$RESULTADO_VISITANT
E)
```

Por lo tanto, el número total de victorias, empates y derrotas del Celta es:

```
> totalVictorias <- victoria_local+victoria_visitante
> totalVictorias
[1] 6
> totalEmpates <- empate_local+empate_visitante
> totalEmpates
[1] 2
> totalDerrotas <- derrota_local+derrota_visitante
> totalDerrotas
[1] 13
```

El número de partidos considerados es:

```
> totalPartidos <- nrow(futbol)
> totalPartidos
```

```
[1] 21
```

En tanto por ciento:

```
> porcVictorias <- totalVictorias*100/totalPartidos
> porcVictorias
[1] 28.57143
> porcEmpates <- totalEmpates*100/totalPartidos
> porcEmpates
[1] 9.52381
> porcDerrotas <- totalDerrotas*100/totalPartidos
> porcDerrotas
[1] 61.90476
```

Luego de sus encuentros contra el Real Madrid en los últimos años, se deduce que:

- La probabilidad de que el Celta gane un partido contra el Real Madrid es P(*G*=C) = 28,6 %.
- La probabilidad de que el Celta empate un partido contra el Real Madrid es P(*G*=0)=9,5 %.
- La probabilidad de que el Celta pierda un partido contra el Real Madrid es P(*G*=RM) = 61,9 %.

Estas son las probabilidades a priori.

b) Lo que nos piden ahora son probabilidades condicionadas. En particular, la probabilidad de que el estadio sea el del Celta condicionado al suceso de victoria, empate o derrota del Celta.

```
> victoria_local*100/totalVictorias
[1] 66.66667
> empate_local*100/totalEmpates
[1] 0
> derrota_local*100/totalDerrotas
[1] 46.15385
```

Luego:

- El porcentaje de veces que una victoria del Celta frente al Real Madrid se produjo en el estadio del Celta es P(*E*=C| *G*=C)=66,7 %.
- El porcentaje de veces que un empate del Celta frente al Real Madrid se produjo en el estadio del Celta es P(*E*=C| *G*=0)=0 %.
- El porcentaje de veces que una derrota del Celta frente al Real Madrid se produjo en el estadio del Celta es P(*E*=C| *G*=RM)=46,2 %.

c) Para dibujar el árbol de probabilidades debemos considerar los dos tipos de sucesos: quién gana el partido (variable *G*) y en qué estadio tuvo lugar el partido (variable *E*). Nos falta por calcular las probabilidades siguientes:

- El porcentaje de veces de una victoria del Celta frente al Real Madrid en el estadio del Real Madrid es P(*E*=RM| *G*=C)=100 – 66,7 = 33,3 %.

- El porcentaje de veces de un empate del Celta frente al Real Madrid en el estadio del Real Madrid es P(*E*=C| *G*=0)=100 – 0 = 100 %.
- El porcentaje de veces que una derrota del Celta frente al Real Madrid se produjo en el estadio del Real Madrid es P(*E*=C| *G*=RM)=100 – 46,2 = 53,8 %.

Por tanto, el árbol de probabilidades es:

$$\text{Partidos} = \begin{cases} P(G=\text{C}) = 28{,}6\ \% & \begin{cases} P(E=\text{C} \mid G=\text{C}) = 66{,}7\ \% \\ P(E=\text{RM} \mid G=\text{C}) = 33{,}3\ \% \end{cases} \\ P(G=0) = 9{,}5\ \% & \begin{cases} P(E=\text{C} \mid G=0) = 0\ \% \\ P(E=\text{RM} \mid G=0) = 100\ \% \end{cases} \\ P(G=\text{RM}) = 61{,}9\ \% & \begin{cases} P(E=\text{C} \mid G=\text{RM}) = 46{,}2\ \% \\ P(E=\text{RM} \mid G=\text{RM}) = 52{,}8\ \% \end{cases} \end{cases}$$

d) Se trata de un caso típico del Teorema de Bayes:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

Lo que nos piden es calcular la probabilidad de que el Celta gane el próximo partido condicionado a que se juegue en su estadio, es decir:

$$P(G=\text{C} \mid E=\text{C}) = \frac{P(G=\text{C})\text{P}(E=\text{C} \mid G=\text{C})}{P(E=\text{C})}$$
$$= \frac{P(G=\text{C})\text{P}(E=\text{C} \mid G=\text{C})}{P(G=\text{C})\text{P}(E=\text{C} \mid G=\text{C}) + P(G=0)\text{P}(E=\text{C} \mid G=0) + P(G=\text{RM})\text{P}(E=\text{C} \mid G=\text{RM})}$$

```
> (0.286*0.667)/(0.286*0.667 + 0.095*0 + 0.619*0.462)
[1] 0.4001384
```

Luego la probabilidad de que el Celta gane el próximo partido en casa frente al Real Madrid es del 40 %.

Respecto del empate:

$$P(G=0 \mid E=\text{C}) = \frac{P(G=0)\text{P}(E=\text{C} \mid G=0)}{P(E=\text{C})}$$
$$= \frac{P(G=0)\text{P}(E=\text{C} \mid G=0)}{P(G=\text{C})\text{P}(E=\text{C} \mid G=\text{C}) + P(G=0)\text{P}(E=\text{C} \mid G=0) + P(G=\text{RM})\text{P}(E=\text{C} \mid G=\text{RM})}$$

```
> (0.095*0)/(0.286*0.667 + 0.095*0 + 0.619*0.462)
[1] 0
```

Luego la probabilidad de empate es del 0 %. Por lo tanto, se predice un 60 % de probabilidad de derrota del Celta frente al Real Madrid.

Análisis crítico del resultado obtenido:

- La probabilidad de ganar el Celta al Real Madrid la hemos estimado según la frecuencia de victorias a partir de los resultados de los últimos 21 partidos. La probabilidad es el límite de la frecuencia cuando el número de experimentos tiende a infinito, luego deberíamos incluir muchos más partidos en el cálculo. Normalmente ambos equipos se enfrentan dos veces al año, salvo la ocasión en la que jugaron además en la Copa del Rey. Por lo tanto, estos 21 partidos corresponden a un intervalo de tiempo que se extiende durante 15 años.

- Para poder realizar el cálculo hemos realizado numerosas suposiciones. Por ejemplo que esta probabilidad es constante en el tiempo. Sin embargo hay factores evidentes no tenidos en cuenta: las plantillas y los entrenadores de ambos equipos han variado completamente en estos 15 años. Pero también son importantes los factores ambientales, como datos climatológicos, la hora y fecha del encuentro, la afluencia de público, situación en la Liga, el tipo de competición (Liga/Copa), el árbitro, etc. Esto hace que las probabilidades a priori sean solo valores aproximados.

- Por otro lado, la probabilidad de jugar en el estadio del Celta condicionado al caso de empate ha sido estimada como del 0 %. Esto produce el resultado también del 0 % de probabilidad de empate condicionado al caso de jugar en el estadio del Celta. Una probabilidad de 0 % se produce para un suceso IMPOSIBLE, que no puede ocurrir nunca. Obviamente es muy aventurado pensar que es imposible que el Celta empate con el Real Madrid en el próximo partido. La solución a este dilema es una vez más incluir datos más partidos para que haya algún empate producido en el estadio del Celta.

- Como conclusión, la probabilidad de derrota del 40 % para el Celta debe considerarse un cálculo aproximado. Las casas de apuestas profesionales utilizan modelos estadísticos mucho más complejos para predecir los resultados de los encuentros y establecer así el valor de las apuestas.

- Como nota anecdótica, el resultado real del encuentro fue **Celta 2 – Real Madrid 4**, es decir, se cumplió el pronóstico de derrota del Celta frente al Real Madrid.

Para los ejercicios siguientes, habrá que cargar los datos en RStudio, mediante el siguiente comando:

> paises <- read.csv("ruta/dataset-paises.csv")

Se trata de un dataset con información sobre los 35 estados del continente americano. En concreto los 8 atributos del dataset son:

- ISO: Abreviatura de 3 letras que representa al país.
- PAIS: Nombre del país.
- REGION: Nombre del subcontinente o región al que pertenece el país.
- POBLACION: Expresada en miles de habitantes.
- PIB_PER_CAPITA: Producto interior bruto por habitante expresado en $.
- CON_SALIDA_AL_MAR: Vale 1 si el país tiene costa, 0 en caso contrario.
- SUPERFICIE: Expresada en km2.
- IDH: Índice de desarrollo humano, un indicador estadístico basado en la esperanza de vida, educación y nivel de vida digno.

4. [1,25 puntos] Respecto del atributo REGION, se pide:

a) Calcular la tabla de frecuencias relativas de los países según dicho atributo.
b) Dibujar un diagrama de barras con dicha tabla de frecuencias. Parámetros:
- Color: usar 4 colores en la escala rainbow.
- Escala del eje Y: debe estar expresada en %.
- Etiqueta del eje X: Región.
- Etiqueta del Eje Y: Número de países.

Solución:

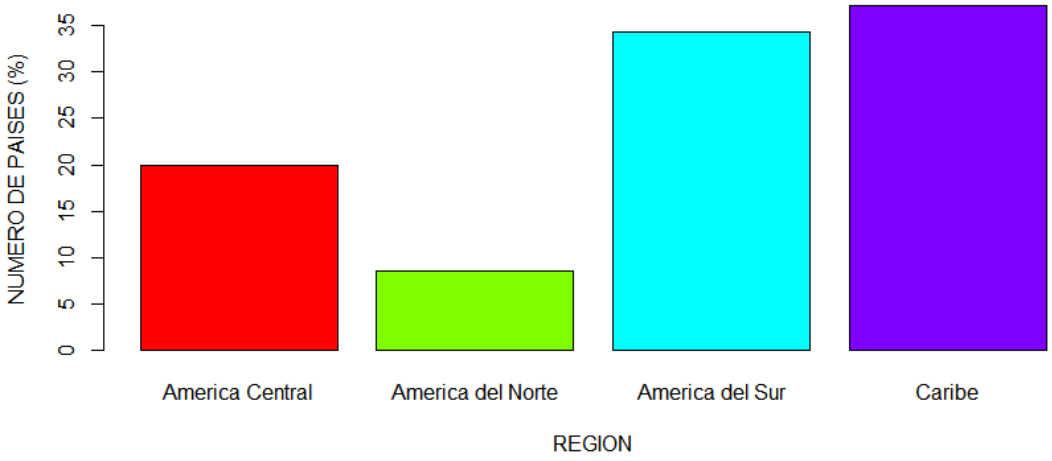
a) La tabla de frecuencias relativas expresadas en % sería la siguiente:

```
> frecRel <- table(paises$REGION)*100/sum(table(paises$REGION))

  America Central America del Norte   America del Sur           Caribe
         20.000000          8.571429         34.285714        37.142857
```

b) El diagrama de barras:

```
> barplot(frecRel,col=rainbow(4),xlab="REGION",ylab="NUMERO DE PAISES
(%)")
```



5. [1,25 puntos] Sobre el atributo IDH, se pide:

- **a)** Calcular su media y desviación típica (sin agrupar los valores).
- **b)** Calcular su nivel de asimetría y curtosis (paquete "moments").
- **c)** Dibujar su histograma, con los parámetros por defecto.
- **d)** Dibujar su diagrama de tallo y hojas. Explicar los parecidos y las diferencias con respecto al histograma del apartado anterior.

Solución:

a) El valor medio y la desviación típica de este parámetro son:

```
> m <- mean(paises$IDH)
> m
[1] 0.7293429
> s <- sd(paises$IDH)
> s
[1] 0.08534771
```

b) La asimetría y la curtosis son:

```
> require("moments")
> skewness(datos$IDH)
[1] -0.6725941
```

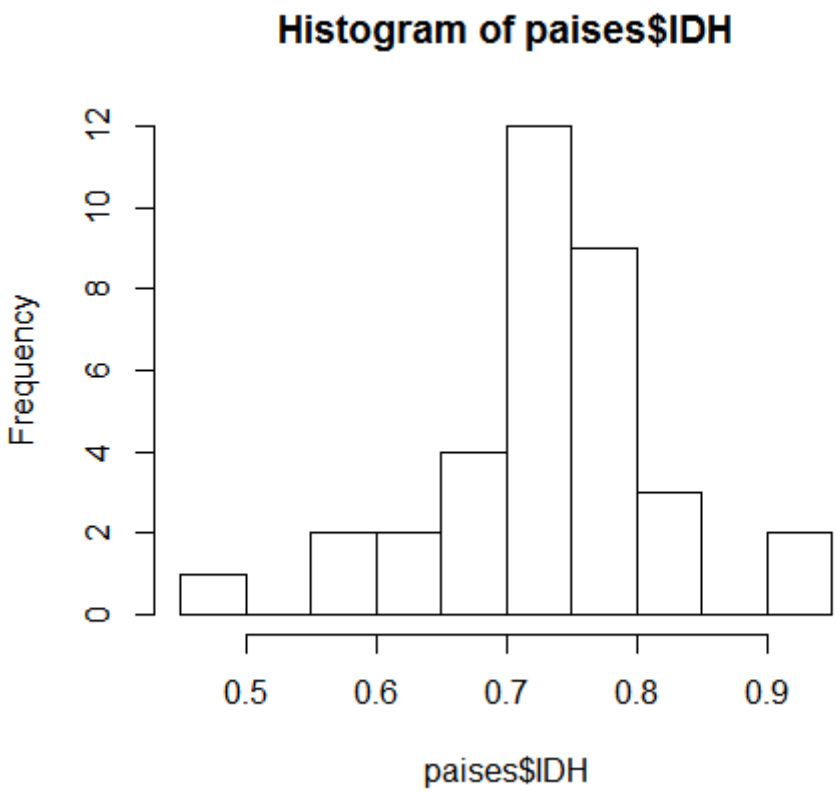
Tiene asimetría negativa, es decir, una cola más larga a la izquierda del máximo.

```
> kurtosis(datos$IDH)
[1] 5.026807
```

Es picuda, al ser la curtosis mayor que 4. Esto significa que los datos se concentran respecto del valor central y luego la distribución cae muy rápidamente.

c) El histograma es:

```
> hist(paises$IDH)
```



Efectivamente existe una gran concentración de valores en el intervalo (0.7, 0.8] y luego cae rápidamente a 0. Además la cola izquierda es mayor que la derecha, luego la asimetría es negativa.

d) El diagrama de tallo y hojas es:

```
> stem(paises$IDH)

  The decimal point is 1 digit(s) to the left of the |

  4 | 6
  5 |
  5 | 8
  6 | 034
  6 | 788
  7 | 001111223444
  7 | 5666777899
  8 | 112
  8 |
  9 | 01
```

El diagrama es similar al histograma, en tanto que los datos aparecen en forma de barras. Pero en esta ocasión podemos ver los valores numéricos que caen dentro de cara intervalo. Se nos

dice que la coma decimal se encuentra a una posición a la izquierda de la barra "|", que separa el tallo de las hojas. Por eso todos los valores son 0.4, 0.5, 0.6, etc.

Rápidamente vemos que el intervalo más frecuente es el de [0.70, 0.74] y el valor más repetido es 0.71 (frecuencia absoluta de 4). Vemos que el diagrama de tallo y hojas es más una representación textual de las frecuencias.

Sin embargo en el histograma representamos los intervalos mediante una barra cuya altura es proporcional a la frecuencia, pero no sabemos qué valores concretos han entrado dentro de cada intervalo. El histograma es por tanto una representación gráfica de las frecuencias.

6. [1,25 puntos] Parece existir una relación entre el IDH y el logaritmo de PIB_PER_CAPITA. Sabiendo que el logaritmo decimal se calcula en R con la función log10, se pide:

a) Representar gráficamente el IDH (eje Y) frente a log10(PIB_PER_CAPITA) (eje X) con los siguientes parámetros:
- Símbolo: abreviatura ISO del país. Para ello primero utilizar la función **plot** con el parámetro type="n", que solo dibuja los ejes con la escala adecuada pero no muestra ningún símbolo. Utilizar inmediatamente después la función **text** para dibujar sobre estos ejes los nombres de los países en sus coordenadas correspondientes. Los dos primeros parámetros de la función **text** son la posición X e Y en el diagrama, es decir, log10(PIB_PER_CAPITA) e IDH, y el tercero son las etiquetas (labels) que se escribirán en cada posición (en nuestro caso el código ISO).
- Etiqueta del eje X: log(PIB per cápita)
- Etiqueta del eje Y: IDH
- Título del gráfico: Relación entre IDH y PIB per cápita en estados de América.

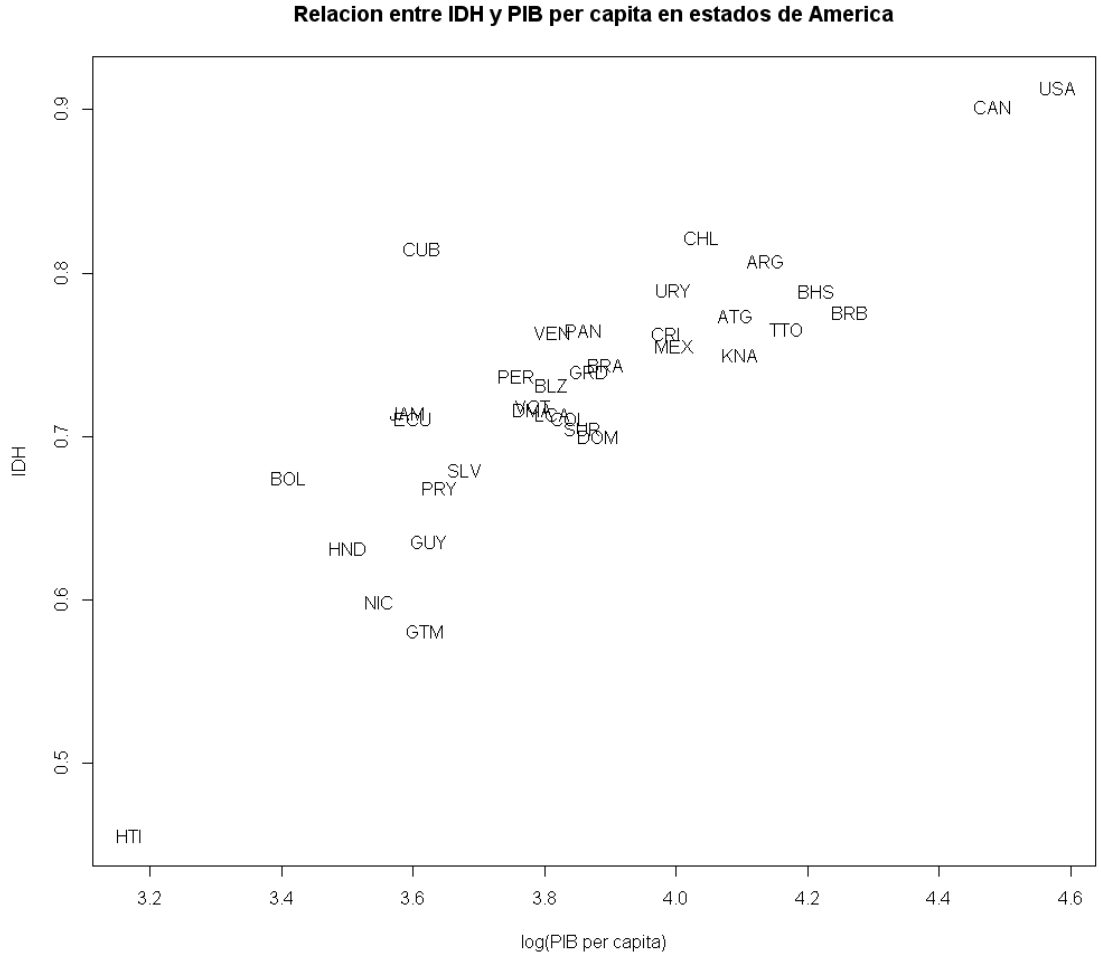
b) Calcula la correlación lineal entre IDH y log10(PIB_PER_CAPITA) y **comenta el resultado**.

c) Realiza un ajuste lineal entre IDH y log10(PIB_PER_CAPITA). En concreto: calcula los coeficientes de regresión y dibuja la recta de regresión sobre la gráfica del apartado anterior (color azul). **Será imprescindible comentar la figura.**

Solución:

a) La gráfica que nos piden es:

```
> plot(log10(paises$PIB_PER_CAPITA),paises$IDH,type="n",xlab="log(PIB
per capita)",ylab="IDH",main="Relacion entre IDH y PIB per capita en
estados de America")
> text(log10(paises$PIB_PER_CAPITA),paises$IDH,labels=paises$ISO)
```



b) La correlación lineal entre los dos parámetros que nos piden es:

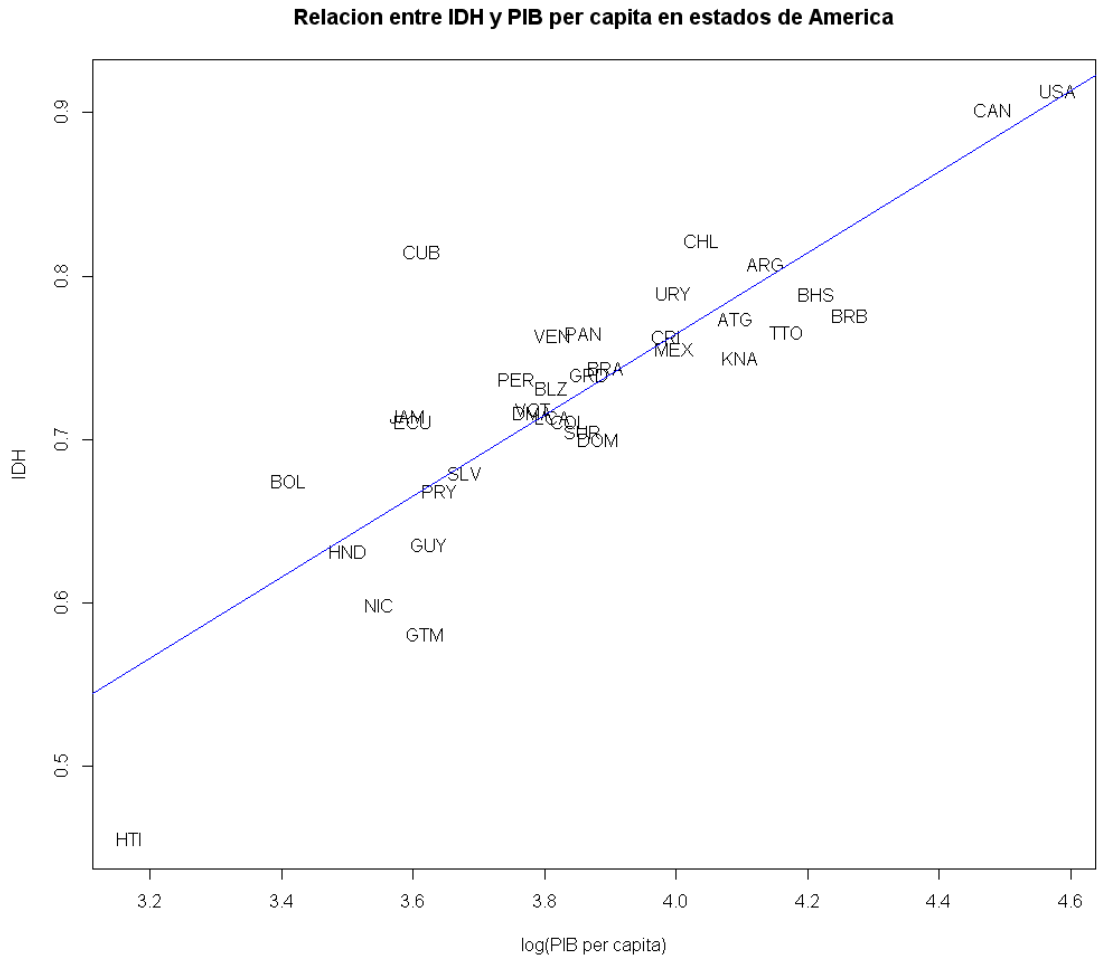
```
> cor(paises$IDH,log10(paises$PIB_PER_CAPITA))
[1] 0.8480559
```

Efectivamente se muestra una tendencia creciente entre el índice IDH y el logaritmo de la renta per cápita: cuando aumenta uno, aumenta el otro, y viceversa. El grado de correlación es bastante elevado, por ser próximo a 1. Por tanto el índice IDH es bueno para hablar del índice de desarrollo humano y de bienestar del país.

c) La recta de regresión sería:

```
> ajuste <- lm(paises$IDH ~ log10(paises$PIB_PER_CAPITA))
> abline(ajuste,col="blue")
> ajuste$coefficients
                (Intercept) log10(datos$PIB_PER_CAPITA)
                 -0.2280648                   0.2482127
```

Luego la recta de regresión es: IDH = -0,2290648 + 0,2482127 $\log_{10}$(PIB PER CAPITA)



Comentarios:

- Vemos cómo el ajuste lineal entre el parámetro IDH y el logaritmo de PIB per cápita es bastante bueno. Surgen grupos de países claramente diferenciados. Estados Unidos y Canadá lideran el ranking de riqueza, seguidos en la distancia de países como Barbados, Bahamas, Argentina y Chile. En el otro extremo, Haití es con diferencia el país más pobre, seguido también en la distancia por Bolivia, Honduras, Nicaragua y Guatemala.
- Los países que peor se ajustan a la recta son precisamente Haití, muy por debajo del valor predicho por el ajuste lineal, y Cuba, todo lo contrario.

7. [1,25 puntos] A partir de la variable cuantitativa IDH, se pide:

a) Crear una nueva variable IDH2 que sea categórica ordinal mediante la función **cut**. Habrá que utilizar los siguientes parámetros:

- Los intervalos en los que deben dividirse la variable son: [0, 0.535), [0.535, 0.7), [0.7, 0.8) y [0.8, 1).
- Las etiquetas (labels) de estos cuatro niveles deben ser: “Bajo”, “Medio”, “Alto” y “Muy alto”.

- Debe obligarse a que IDH2 sea ordinal mediante el valor correcto del parámetro ordered_result.

b) Calcular la tabla de frecuencias conjunta entre el atributo REGION y la variable recientemente creada IDH2

c) Representar en un diagrama de mosaico la tabla de frecuencias del apartado anterior usando los siguientes parámetros:

- En el eje X debe aparecer REGION, mientras que en el eje Y debe aparece IDH2.
- Pintar la gráfica con 4 colores de la escala topo.colors.
- Etiqueta del eje X: Región
- Etiqueta del eje Y: IDH
- Título del gráfico: IDH por regiones de América
- **Será imprescindible comentar el gráfico.**

Solución:

a) La nueva variable categoría ordinal se crea de la siguiente manera:

```
> IDH2 <- cut(paises$IDH,right=FALSE,breaks<-c(0,0.535,0.7,0.8,1),
labels<-c("Bajo","Medio","Alto","Muy alto"),ordered_result=TRUE)
> IDH2
 [1] Alto     Muy alto Alto     Alto     Alto     Medio    Alto
Muy alto
 [9] Muy alto Alto     Alto     Muy alto Alto     Alto     Alto
Medio
[17] Alto     Medio    Medio    Bajo     Medio    Alto     Alto
Medio
[25] Alto     Medio    Alto     Alto     Alto     Alto     Alto
Alto
[33] Muy alto Alto     Alto
Levels: Bajo < Medio < Alto < Muy alto
```

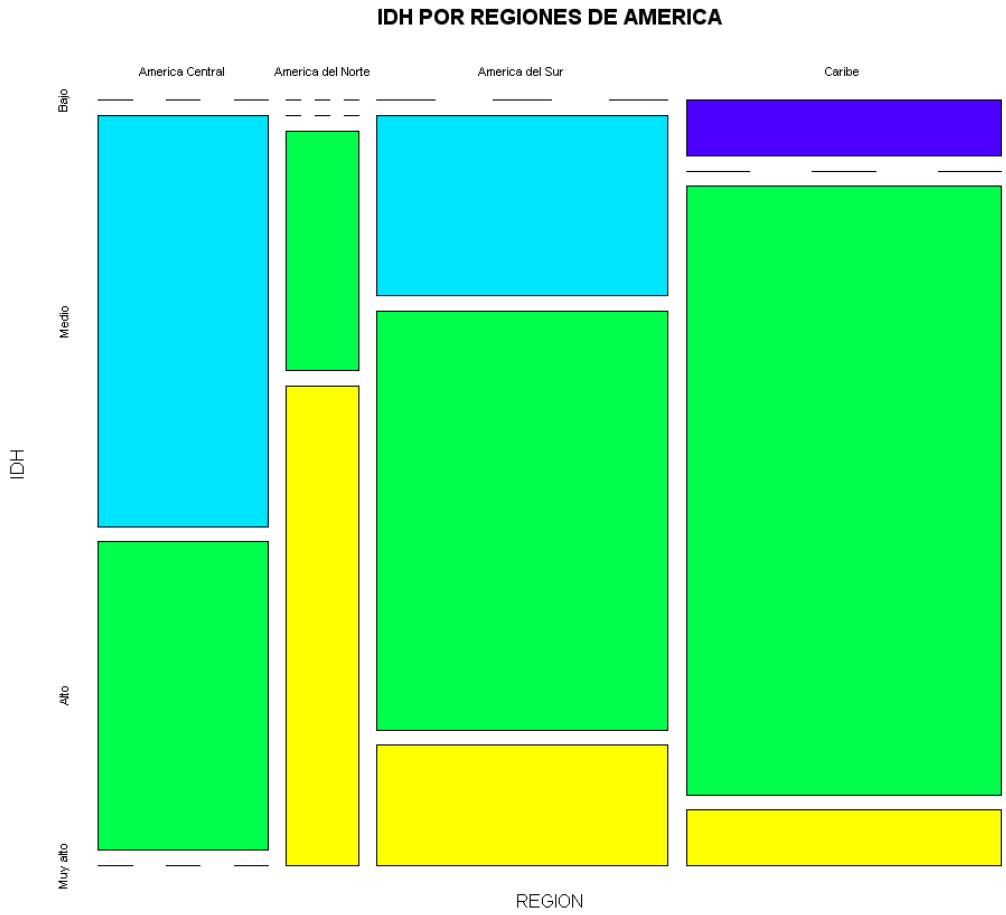
b) La tabla de frecuencias de doble entrada es:

```
> table(paises$REGION,IDH2)
                   IDH2
                    Bajo Medio Alto Muy alto
  America Central      0     4    3        0
  America del Norte    0     0    1        2
  America del Sur      0     3    7        2
  Caribe               1     0   11        1
```

c) El diagrama de mosaico es:

```
> mosaicplot(table(paises$REGION,IDH2),xlab="REGION",ylab="IDH",
col=topo.colors(4),main="IDH POR REGIONES DE AMERICA")
```

Según la escala de colores elegida, los colores son azul oscuro, azul claro, verde y amarillo, ordenados de manera ascendente.



Comentarios:

- América del Norte (EE.UU., Canadá y México) es la región con IDH más alto (color amarillo). Ningún país tiene IDH medio o bajo. Si bien, estos países son una minoría respecto del resto de países (son solo 3).
- América Central es central: los países se dividen entre IDH medio o alto y no hay ninguno con índice bajo o muy alto.
- En América del Sur hay mucha desigualdad, pues hay países con IDH muy alto, alto o medio. No hay ninguno con un nivel muy bajo.
- En la zona del Caribe una inmensa mayoría de países tienen un nivel alto, si bien un pequeño grupo de países tienen un nivel IDH muy alto y otro grupo tiene un nivel muy bajo. Es por tanto la zona con mayores desigualdades. Por otro lado, la proporción de países del Caribe es aproximadamente igual que de América del Sur (13 frente a 12).