EJERCICIOS

EJERCICIOS RESUELTOS DE ESTADÍSTICA I

Unión de Estudiantes de Ciencias Económicas | AECUC3M

Estadística I
Ejercicios Temas 1 y 2
Curso 2009/10 – Soluciones

1. La tabla siguiente muestra el número de mujeres (en miles) que se encontraban activas en el año 1986 en EEUU por sectores profesionales:

Profesión	# de mujeres
Arte/Entretenimiento	901
Derecho	698
Educación	2833
Ingeniería	347
Salud	1937
Otros	355

   a) Obtén la distribución de frecuencias relativas para este conjunto de datos. ¿Qué porcentaje de mujeres trabajaron en el área de Derecho?

   b) Construye un diagrama de barras para estos datos usando las frecuencias relativas obtenidas en (a).

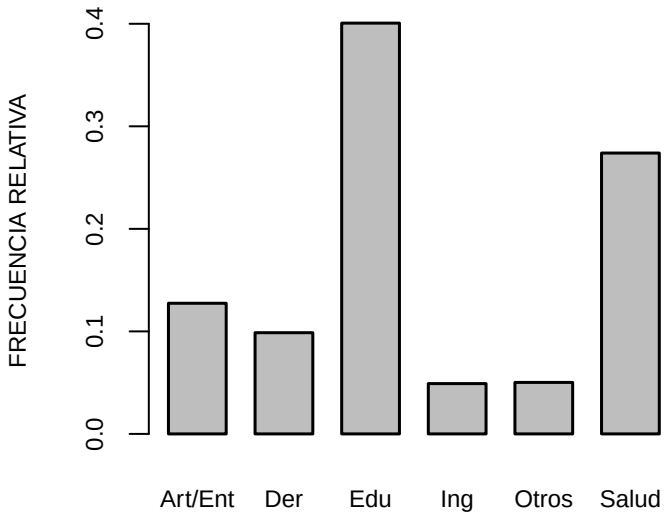
Solución:

a)

Clase	Freq. abs., n_i	Freq. rel., f_i
Art/Ent	901	0.13
Der	698	0.10
Edu	2833	0.40
Ing	347	0.05
Salud	1937	0.27
Otros	355	0.05

El 10 % de las mujeres trabajaron en el área de Derecho.

b) Diagrama de barras



2. La tabla inferior muestra las ganancias semanales de una compañía de marketing de hamburguesas (en miles de euros):

Ganancias				
3145	15879	6914	4572	11374
12764	9061	8245	10563	8164
6395	8758	17270	10755	10465
7415	9637	9361	11606	7836
13517	7645	9757	9537	23957
8020	8346	12848	8438	6347
21333	9280	7538	7414	11707
9144	7424	25639	10274	4683
5089	6904	9182	12193	12472
8494	6032	16012	9282	3331

a) Construye 6 intervalos de igual longitud que comprendan el rango de Ganancias 0-30000, especifica la marca de clase de cada intervalo y obtén la distribución de frecuencias absolutas para estos datos. Identifica el intervalo modal (el de mayor cantidad de observaciones).

b) Partiendo de la tabla del apartado (a), representa gráficamente los datos dibujando un histograma. Describe la simetría de la distribución.

c) Partiendo de lo que conoces del apartado (b), ¿qué tipo de medidas numéricas, de entre las estudiadas, serían las más adecuadas para describir el centro y la dispersión de los datos? Justifica tu respuesta.

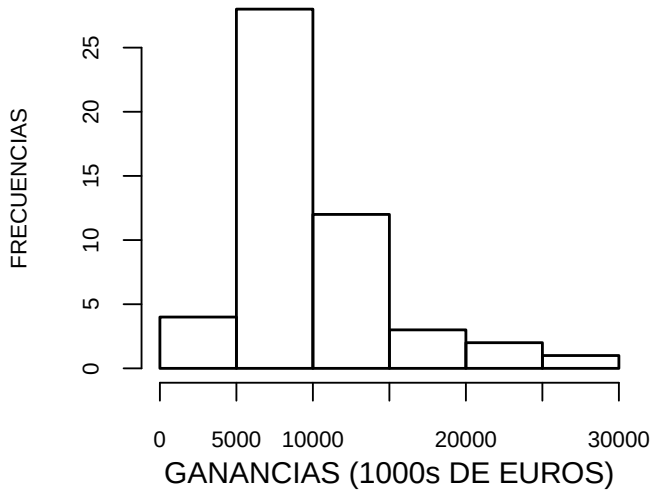
Solución:

a)

Clase $[l_{i-1}, l_i)$	Marca de clase x_i	Frec. absoluta, n_i
$[0, 5000)$	2500	4
$[5000, 10000)$	7500	28
$[10000, 15000)$	12500	12
$[15000, 20000)$	17500	3
$[20000, 25000)$	22500	2
$[25000, 30000)$	27500	1

El segundo intervalo, $[5000, 10000)$, es la clase modal.

b) La distribución es asimétrica a la derecha.



c) La mediana (centro) y el RIC (dispersión) son más apropiados que la media y la desviación típica para describir el centro y la variación, respectivamente, en distribuciones asimétricas.

3. Los siguientes datos muestran las temperaturas tomadas en cierta ciudad durante el mes de Abril:

Temperatura (°F)								
47	49	51	49	60	46	50	58	46
55	45	47	42	42	68	53	56	56
35	43	54	76	55	50	68	49	46
56	37	38	69	62	60	50	70	72
62	66	49	46	62	52	43	61	53
51	49	30	52	57	69	50	55	52
54	48	60	65	37	53	48	80	
63	51	69	68	63	18	59	38	
43	66	52	39	75	58	45	66	
49	47	46	55	45	60	46	49	

a) Construye la tabla de la distribución de frecuencias (absolutas) haciendo intervalos de amplitud igual a 10 y comenzando por el valor 10. ¿Cuántos registros de temperatura fueron al menos de 60°F?

b) Partiendo del apartado (a), construye un histograma de area igual a 1 para este conjunto de datos. ¿Existen observaciones que podrían ser atípicas? Describe la forma de la distribución omitiendo las observaciones que son potencialmente atípicas.

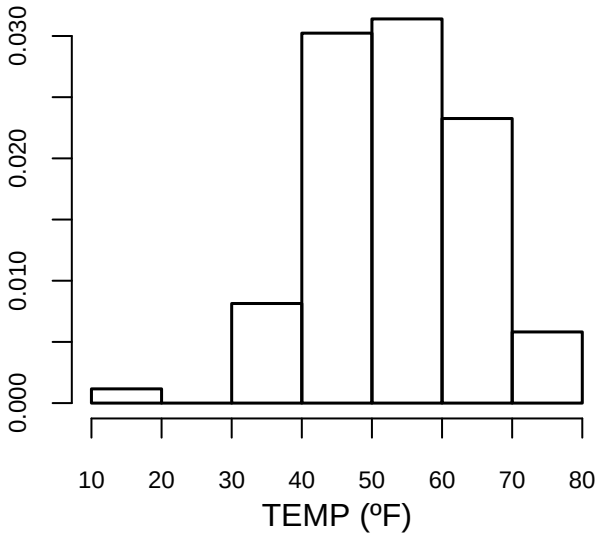
Solución:

a)

Clase $[l_{i-1}, l_i)$	Marca de clase x_i	Frec. absoluta, n_i
$[10, 20)$	15	1
$[20, 30)$	25	0
$[30, 40)$	35	7
$[40, 50)$	45	26
$[50, 60)$	55	27
$[60, 70)$	65	20
$[70, 80)$	75	5

$20 + 5 = 25$ registros fueros iguales o superiores a 60°F.

b) Mirando el histograma podríamos decir que la observación más pequeña, con valor 18, es potencialmente atípica. Ignorando tal observación, la distribución tiene una forma simétrica ($\bar{x} \cong M$).



4. La tabla siguiente muestra la Estatura (en metros) de 50 mujeres españolas:

Estatura (en metros)				
1.56	1.59	1.63	1.62	1.65
1.61	1.59	1.51	1.62	1.62
1.53	1.49	1.57	1.54	1.53
1.59	1.58	1.57	1.47	1.64
1.55	1.59	1.53	1.56	1.53
1.47	1.57	1.60	1.54	1.56
1.50	1.62	1.59	1.62	1.54
1.68	1.52	1.62	1.62	1.49
1.65	1.53	1.59	1.56	1.54
1.58	1.52	1.63	1.56	1.62

a) Obtén la distribución de frecuencias (absolutas) de los datos haciendo 10 intervalos que comprendan al rango 1.45-1.70. ¿Cuántas mujeres tienen estatura inferior a 1.55m? ¿Qué porcentaje de mujeres tiene una estatura de al menos 1.65m?

b) Realiza el histograma y describe la simetría de éste.

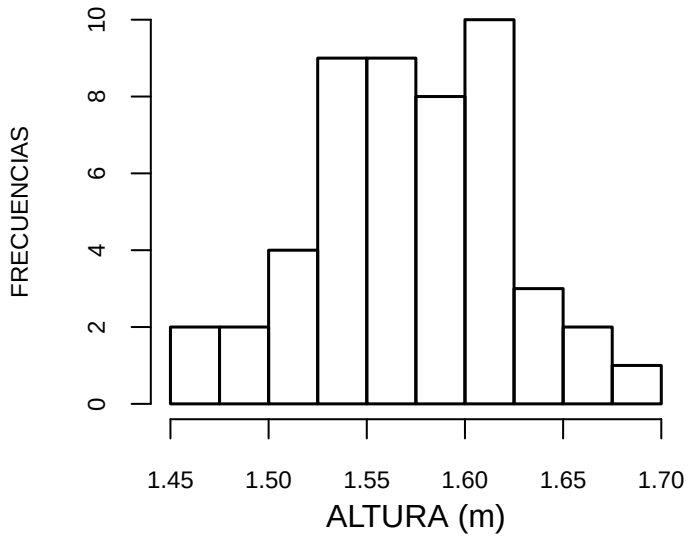
Solución:

a)

Clase $[l_{i-1}, l_i)$	Marca de clase x_i	Frec. absoluta n_i
$[1.450, 1.475)$	1.4625	2
$[1.475, 1.500)$	1.4875	2
$[1.500, 1.525)$	1.5125	4
$[1.525, 1.550)$	1.5375	9
$[1.550, 1.575)$	1.5625	9
$[1.575, 1.600)$	1.5875	8
$[1.600, 1.625)$	1.6125	10
$[1.625, 1.650)$	1.6375	3
$[1.650, 1.675)$	1.6625	2
$[1.675, 1.700)$	1.6875	1

$2 + 2 + 4 + 4 = 17$ mujeres tienen estatura inferior a 1.55m. $\frac{2+1}{50} = 6\,\%$ de las mujeres tiene al menos una estatura de 1.65m.

b) La distribución es aproximadamente simétrica ($\bar{x} \cong M$).



5. Estamos interesados en el número de transacciones mensuales realizadas por una cooperativa de crédito. Se han recogido los siguientes datos:

# de transacciones				
17	25	32	41	43
31	28	27	39	36
25	19	21	28	26
30	32	26	27	34
21	24	20	25	31

   a) Obtén una tabla de distribución de frecuencias observadas, realizando seis intervalos iguales de amplitud 5 y comenzando desde el valor 15.

   b) Determina sus correspondientes frecuencias relativas.

   c) A partir de los apartados (a) y (b) obtén las correspondientes frecuencias acumuladas. Identifica las clases modales.

   Solución:
   (a), (b), (c)

$[l_{i-1}, l_i)$	Marca de clase x_i	n_i	f_i	N_i	F_i
$[15, 20)$	17.5	2	0.08	2	0.08
$[20, 25)$	22.5	4	0.16	6	0.24
$[25, 30)$	27.5	9	0.36	15	0.60
$[30, 35)$	32.5	6	0.24	21	0.84
$[35, 40)$	37.5	2	0.08	23	0.92
$[40, 45)$	42.5	2	0.08	25	1.00

   El tercer intervalo, $[25, 30)$, es la clase modal.

6. El director de una compañía desea estudiar si la experiencia se traduce en una mayor rapidez al hacer una tarea. Para ello, lleva a cabo un experimento con 25 empleados, a los cuales les solicita que realicen una tarea 10 veces. A los mismos 25 empleados les hace la misma solicitud, pero con 20 repeticiones. Y de nuevo, con 50 repeticiones. La tabla siguiente muestra el tiempo medio (en minutos) necesitado por los empleados para realizar cada repetición en cada caso.

Tiempo (10 repeticiones)		Tiempo (20 repeticiones)		Tiempo (50 repeticiones)	
15	19	16	11	10	8
21	20	10	10	5	10
30	22	12	13	7	8
17	20	9	12	9	7
18	19	7	8	8	8
22	18	11	20	11	6
33	17	8	7	12	8
41	16	9	6	9	6
10	20	5	9	7	4
14	22	15	10	6	15
18	19	10	10	8	7
25	24	11	11	14	20
23		9		9	

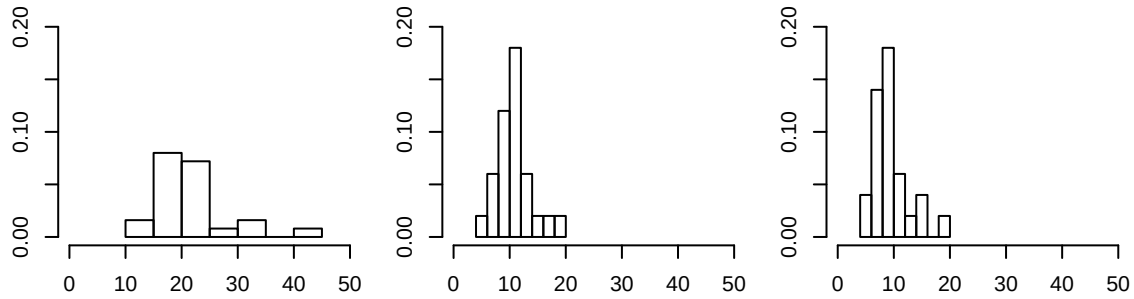
a) Representa gráficamente los tres histogramas, uno para cada subconjunto de datos.

b) Compara los histogramas del apartado (a). ¿Opinas que existe una relación entre el número de veces que se repite la tarea y el tiempo medio necesario para realizarla? Justifica tu respuesta.

c) Calcula y compara los coeficientes de variación, CV, para los tres conjuntos. ¿Cuáles son las unidades de los CV?

Solución:

a) Los histogramas son



b) Sí, parece que al incrementarse el número de repeticiones, se necesita un menor tiempo medio para llevar a cabo las tareas.

c) El CV no tiene unidad (o se expresa en %'s)

$CV_x = 6.3306/20.3306 = 30.26\,\%$ $CV_y = 3.2259/10.36 = 31.14\,\%$ $CV_z = 3.4559/8.88 = 38.92\,\%$

7. Los siguientes datos corresponden al número de accidentes de trabajo por mes:

$$1 \quad 3 \quad 4 \quad 5 \quad 2 \quad 2 \quad 6 \quad 7 \quad 2 \quad 0 \quad 1$$

a) ¿Son estos datos cualitativos o cuantitativos? En el primer caso, ¿son cualitativos ordinales o nominales? En el segundo, son cuantitativos discretos o continuos?

b) Calcula la media, la mediana y la moda para este conjunto de datos. ¿Qué unidades tienen estas medidas?

c) Calcula la (cuasi) varianza, la (cuasi) desviación típica, el rango, el rango inter-cuartílico y el coeficiente de variación. ¿Cuáles son sus unidades?

Solución:

a) Cuantitativos discretos

b)

$$\begin{aligned} \bar{x} &= 3 \\ M &= 2 \quad (0,1,1,2,2,\boxed{2},3,4,5,6,7) \\ \text{moda} &= 2 \end{aligned}$$

Las unidades de todas estas medidas son las unidades de los datos: accidentes por mes.

c)

$$\begin{aligned} s_x^2 &= \frac{\sum_{i=1}^n x_i^2 - n(\bar{x})^2}{n-1} = \frac{149 - 11(3)^2}{11-1} = 5 \\ s_x &= 2.24 \\ R &= 7 - 0 = 7 \\ RIC &= 5 - 1 = 4 \\ CV &= \frac{2.24}{3} = 0.75 \end{aligned}$$

Las unidades son: s_x^2 (unidades^2=(accidentes por mes)^2) and CV (ninguno o en %).

8. Los siguientes datos muestran el número de helados vendidos por hora en una heladería durante diferentes horas de apertura:

35	47	22	15	13	28	39	41	43	36	24	23
17	19	21	31	35	37	41	43	47	5	12	19

a) Obtén la media, la mediana y la moda de estas observaciones.

b) Obtén la (cuasi) varianza, la (cuasi) desviación típica, el rango, el rango inter-cuartílico y el coeficiente de variación.

c) Representa gráficamente el diagrama de cajas para este conjunto de datos.

d) ¿Existe alguna observación atípica? ¿Es simétrica la distribución? (Pista: compara la media y la mediana)

Solución:

a)

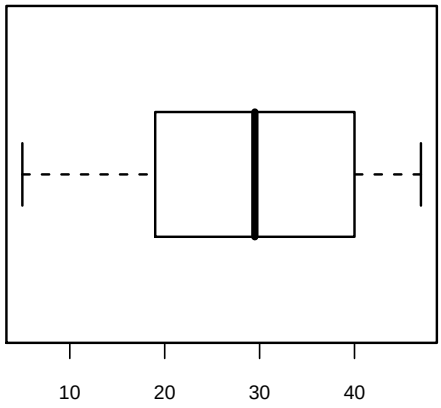
$$\begin{aligned} \bar{x} &= 28.875 \\ M &= 29.5 \quad (5,12,13,15,17,\boxed{19,\ 19},21,22,23,24,\boxed{28,\ 31},35,35,36,37,\boxed{39,\ 41},41,43,43,47,47) \end{aligned}$$

La Moda no es única

b)

$$\begin{aligned} s_x^2 &= \frac{\sum_{i=1}^n x_i^2 - n(\bar{x})^2}{n-1} = \frac{23463 - 24(28.875)^2}{24-1} = 150.1141 \\ s_x &= 12.2521 \\ R &= 47 - 5 = 42 \\ RIC &= x_{(18.75)} - x_{(6.25)} = (39 + 0.75(41-39)) - 19 = 40.5 - 19 = 21.5 \\ CV &= \frac{12.2521}{28.875} = 0.4243 \end{aligned}$$

c) El diagrama de cajas es



d) No. No hay observaciones que sean:

- Mayores que $Q_3 + 1.5RIC = 40 + 31.5 = 71.5$
- Menores que $Q_1 - 1.5RIC = 19 - 31.5 = -12.5$

La distribución es aproximadamente simétrica ($\bar{x} \cong M$).

9. La tabla siguiente muestra las calificaciones de un grupo de alumnos en el examen de una asignatura:

Calificaciones				
8.4	7.7	6.7	9.4	9.0
8.1	5.6	8.9	7.7	8.8
7.4	7.6	2.8	8.0	5.8
6.6	7.7	8.9	8.1	7.8
7.7	7.2	9.4	9.3	7.9
9.3	2.1			

a) Obtén la media, la mediana y la moda.

b) Obtén la (cuasi) varianza, la (cuasi) desviación típica, el rango, el rango inter-cuartílico y el coeficiente de variación.

c) Representa gráficamente el diagrama de cajas de estas calificaciones.

d) ¿Existe alguna observación atípica? ¿Es simétrica la distribución? (Pista: compara la media y la mediana)

Solución:

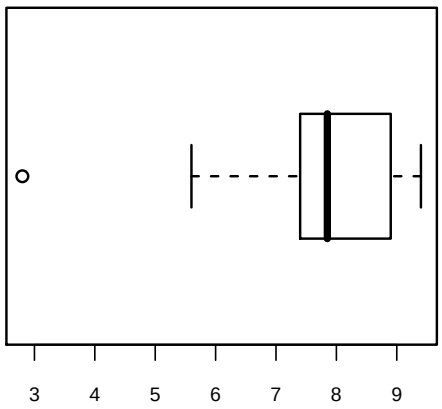
a) Datos ordenados: $2.8, 5.6, 5.8, 6.6, 6.7, \boxed{7.2,\ 7.4}, 7.6, 7.7, 7.7, 7.7, 7.7, \boxed{7.8,\ 7.9}, 8.0, 8.1, 8.1, 8.4, 8.8, \boxed{8.9,\ 8.9}, 9.0, 9.3, 9.3, 9.4, 9.4$

$$\begin{aligned} \bar{x} &= 7.761538 \\ M &= 7.85 \\ \text{mode} &= 7.7 \end{aligned}$$

b)

$$\begin{aligned} s_x^2 &= \frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n-1} = \frac{1618.4 - 23(7.761538)^2}{23-1} = 2.0849 \\ s_x &= 1.4439 \\ R &= 9.4 - 2.8 = 6.6 \\ RIC &= x_{(20.25)} - x_{(6.75)} = 8.9 - (7.2 + 0.75(7.4 - 7.2)) = 8.9 - 7.35 = 1.55 \\ CV &= \frac{7.761538}{1.4439} = 5.3754 \end{aligned}$$

c) El diagrama de cajas es



d) Sí. La observación 2.8 es atípica porque es:

- Mayor que $Q_3 + 1.5RIC = 8.9 + 1.5RIC = 11.225$
- Menor que $Q_1 - 1.5RIC = 7.35 - 1.5RIC = 5.025$

Excluyendo la observación atípica, la distribución es ligeramente asimétrica a la izquierda ($\bar{x} < M$).

10. Un agente de control de calidad de una compañía de neumáticos de coches estima que el peso medio de los neumáticos es de 20 kg, con una desviación típica de 1 kg. Además, sabemos que el 68 % de los neumáticos pesan entre 19 y 21 kg, y casi todos entre 17 y 23 kg.

a) ¿Qué puedes decir acerca de la forma de la distribución de los pesos a partir de la información que se aporta? *Pista: utiliza la regla empírica.*

Solución:

a) Tenemos que

$$\begin{aligned} (19, 21) &= (20 \pm 1 \cdot 1) = (\bar{x} \pm 1s) \quad 68\,\% \\ (17, 23) &= (20 \pm 3 \cdot 1) = (\bar{x} \pm 3s) \quad 99.7\,\% \end{aligned}$$

Por tanto, la regla empírica se verifica y podemos concluir que la distribución es acampanada.

Estadística I

Ejercicios Tema 3

Curso 2009/10

---

1. a) Distribuciones de frecuencias marginales relativas:

# de h: \ nota:	Suspenso	Aprobado	Notable	Sobresaliente	D. marg. de # de h
2	0.20	0.15	0.08	0.03	0.46
3	0.12	0.07	0.02	0.02	0.23
4	0.04	0.10	0.02	0.00	0.16
5	0.00	0.05	0.05	0.05	0.15
D. marg. de nota	0.36	0.37	0.17	0.10	1

b) Distribuciones de "nota" condicionadas a los distintos valores de "número de horas de estudio":

Nota | # horas= 2:	Suspenso	Aprobado	Notable	Sobresaliente	Total
$fr_{x_i\|y=2}$	0.435	0.326	0.174	0.065	1

Nota | # horas= 3:	Suspenso	Aprobado	Notable	Sobresaliente	Total
$fr_{x_i\|y=3}$	0.522	0.304	0.87	0.87	1

Nota | # horas= 4:	Suspenso	Aprobado	Notable	Sobresaliente	Total
$fr_{x_i\|y=4}$	0.250	0.625	0.125	0.000	1

Nota | # horas= 5	Suspenso	Aprobado	Notable	Sobresaliente	Total
$fr_{x_i\|y=5}$	0.000	0.333	0.333	0.333	1

Distribuciones de "número de horas de estudio" condicionadas a los distintos valores de "nota":

# horas| nota = Suspenso	$fr_{y_j\|x=Sus.}$	# horas| nota = Aprobado	$fr_{y_j\|x=Apr.}$
2	0.556	2	0.405
3	0.333	3	0.189
4	0.111	4	0.270
5	0.000	5	0.135
Total	1	Total	1

# horas| nota = Notable	$fr_{y_j\|x=Not.}$	# horas| nota = Sobresaliente	$fr_{y_j\|x=Sob.}$
2	0.471	2	0.3
3	0.118	3	0.2
4	0.118	4	0.0
5	0.294	5	0.5
Total	1	Total	1

2. a) Distribuciones de frecuencias marginales relativas:

# de hijos \ renta:	0-1000	1000-2000	2000-3000	> 3000	D. marg. de # de hijos
0	0.15	0.05	0.03	0.02	0.25
1	0.10	0.20	0.10	0.05	0.45
2	0.05	0.10	0.05	0.03	0.23
≥ 3	0.02	0.03	0.02	0.00	0.07
D. marg. de renta	0.32	0.37	0.18	0.12	1

b) Distribución condicionada de $Y|X = 2$:

renta | # hijos= 2:	0-1000	1000-2000	2000-3000	> 3000	Total
$fr_{y_i\|x=2}$	0.218	0.435	0.217	0.130	1

c) Distribución condicionada de $X|1000 < Y < 2000$:

# hijos| renta= $1000 < Y < 2000$	$fr_{x_i\|1000<y<2000}$
0	0.135
1	0.541
2	0.270
≥ 3	0.054
Total	1

3. a) Distribución conjunta de frecuencias absolutas:

		Núm. compras por semana				
		0	1	2	3	4
Núm. tarjetas	1	24	39	27	18	9
	2	9	24	24	27	21
	3	3	9	18	24	24

b) Distribución marginal de Y:

Núm. compras por semana	0	1	2	3	4	Total
n_j	36	72	69	69	54	300

Media del número de compras por semana:

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{5} y_j \cdot n_j = (0 \cdot 36 + 1 \cdot 72 + 2 \cdot 69 + 3 \cdot 69 + 4 \cdot 54)/300 = 2.11.$$

Varianza del número de compras por semana:

$$s_y^2 = \frac{1}{300-1}\left(\sum_{j=1}^{5} y_j^2 \cdot n_j - 300\overline{y}^2\right) = \frac{1}{300-1}\left((0^2 \cdot 36 + 1^2 \cdot 72 + 2^2 \cdot 69 + 3^2 \cdot 69 + 4^2 \cdot 54) - 300 \cdot 2.11^2\right) = 1.6634$$

Desviación típica del número de compras por semana: $s_y = \sqrt{s_y^2} = \sqrt{1.6634} = 1.29$

c) Distribución del número de tarjetas de crédito:

# tarjetas de crédito	n_i
1	122
2	107
3	81
Total	300

Número más frecuente de tarjetas de crédito (moda): 1.

d) Distribución del número de compras semanales pagadas con tarjetas de crédito que realizan las personas que poseen tres tarjetas:

Núm. compras por semana | num. tarjetas=3	0	1	2	3	4	Total
$fr_{y_j\|x=3}$	0.037	0.111	0.222	0.296	0.296	1

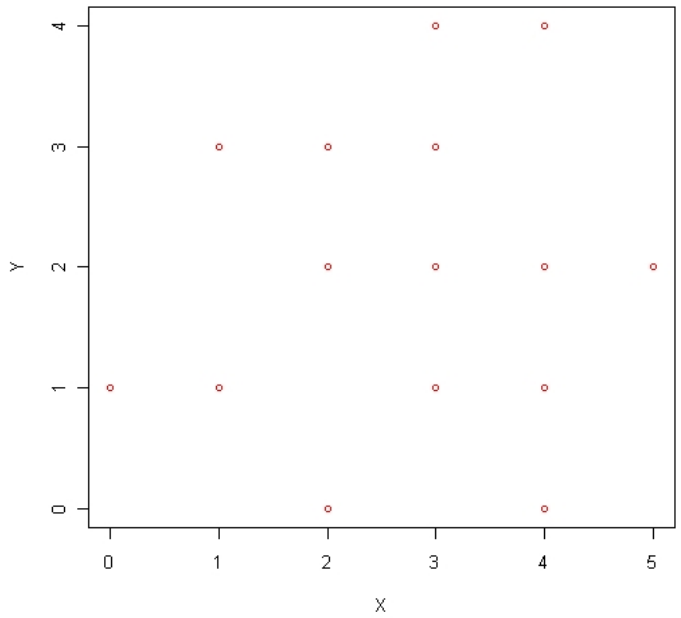
Media de esta distribución:

$$\overline{y}|x=3 = \sum_{j=1}^{5} y_j \cdot fr_{y_j|x=3} = 0 \cdot 0.037 + 1 \cdot 0.111 + 2 \cdot 0.222 + 3 \cdot 0.296 + 4 \cdot 0.296 = 2.627.$$

4. a) Tabla de doble entrada (distribución conjunta de frecuencias):

X: \ Y:	0	1	2	3	4	D. marg. de X
0	0	4	0	0	0	4
1	0	3	0	4	0	7
2	2	0	9	3	0	14
3	0	6	12	5	2	25
4	2	7	15	0	1	25
5	0	0	5	0	0	5
D. marg. de Y	4	20	41	12	3	80

Diagrama de dispersión:



b) Tanto la covarianza como el coeficiente de correlación han de ser positivos ya que las dos variables parecen tener una relación creciente. Además, sobre el valor del coeficiente de correlación, podemos decir que no estará próximo a 1, ya que la relación lineal entre las dos variables no parece muy fuerte.

c)

$$r_{(x,y)} = \frac{Cov(s,y)}{s_x \cdot s_y} \qquad Cov(x,y) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{xy}\right)$$

$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{80}(0\cdot 4 + 1\cdot 7 + 2\cdot 14 + 3\cdot 25 + 4\cdot 25 + 5\cdot 5) = 2.9375.$

$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{1}{80}(0\cdot 4 + 1\cdot 20 + 2\cdot 41 + 3\cdot 12 + 4\cdot 3) = 1.875.$

$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\overline{x}^2 = \frac{1}{80-1}(0^2\cdot 4 + 1^2\cdot 7 + 2^2\cdot 14 + 3^2\cdot 25 + 4^2\cdot 25 + 5^2\cdot 5) - \frac{80}{80-1}2.9375^2 = 1.553$

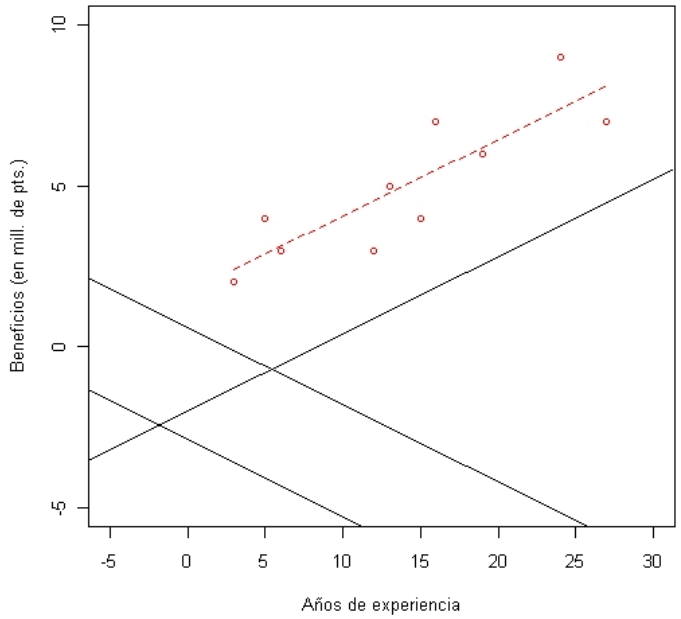
$s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n} y_j^2 - \frac{n}{n-1}\overline{y}^2 = \frac{1}{80-1}(0^2\cdot 4 + 1^2\cdot 20 + 2^2\cdot 41 + 3^2\cdot 12 + 4^2\cdot 3) - \frac{80}{80-1}1.875^2 = 0.74367$

$Cov(x,y) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i \cdot y_i - n\cdot\overline{x}\cdot\overline{y}\right) = \frac{1}{79}(0\cdot 0\cdot 0 + 0\cdot 1\cdot 4 + \cdots + 5\cdot 4\cdot 0 - 80\cdot 2.9375\cdot 1.5336) = 0.0174.$

$r_{(x,y)} = \frac{Cov(x,y)}{s_x\cdot s_y} = \frac{0.0174}{\sqrt{1.553}\cdot\sqrt{0.74367}} = 0.0162$

Como habíamos predicho, obtenemos valores positivos para la covarianza y el coeficiente de correlación. El valor del coeficiente de correlación es muy cercano a cero, lo que indica que prácticamente no hay relación lineal entre estas dos variables.

5. a) y b) Diagrama de dispersión y rectas:



c) La recta de regresión parece ser $y = 1.66 + 0.24x$, es decir, la que aparece punteada en el gráfico anterior.

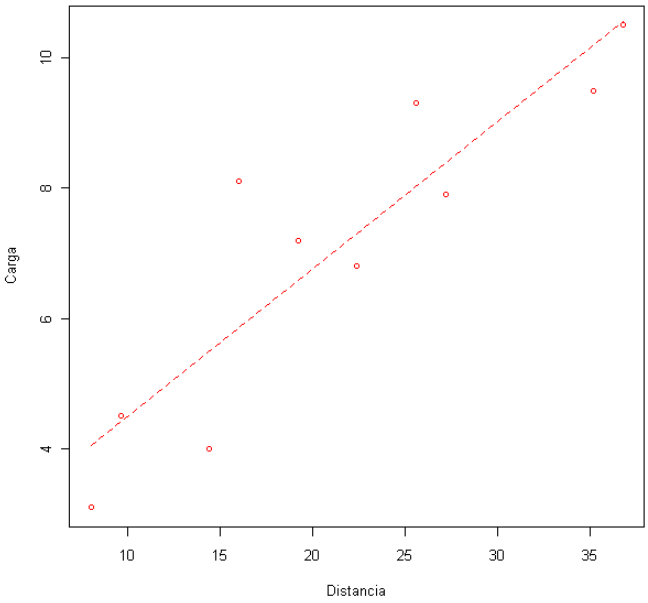
d) Coeficiente de correlación (los cálculos se harían como en el ejercicio 4): $r_{(x,y)} = 0.8587$.

e) La relación linea es positiva, es decir, a mayores valores de x, mayores valores de y, ya que el coeficiente de correlación es positivo. Además, como toma un valor alto (próximo a 1) podemos decir que la relación lineal es fuerte.

f) Recta de regresión de los años de experiencia (x) en función de los beneficios (y): $x = c + dy$ donde $d = \frac{Cov(x,y)}{s_y^2} = 3.0909$ y $c = \overline{x} - d\overline{y} = -1.4545$. La interpretación de la pendiente sería que un aumento de 1 millón de pesetas en los beneficios, se corresponde con un aumento de 3.0909 años en la experiencia de la empresa. (Parece que claro, que para este par de variables, la variable independiente debería ser los años de experiencia y la independiente los beneficios y no al revés).

La ordenada en el origen se interpretaría como los años de experiencia para una empresa que no obtuviese beneficios (0 millones). Evidentemente, el valor obtenido (-1.4545 años) no tiene sentido en este caso ya que 0 no está dentro del rango de valores de la variable beneficios utilizados para predecir la recta de regresión.

6. a) Diagrama de dispersión:



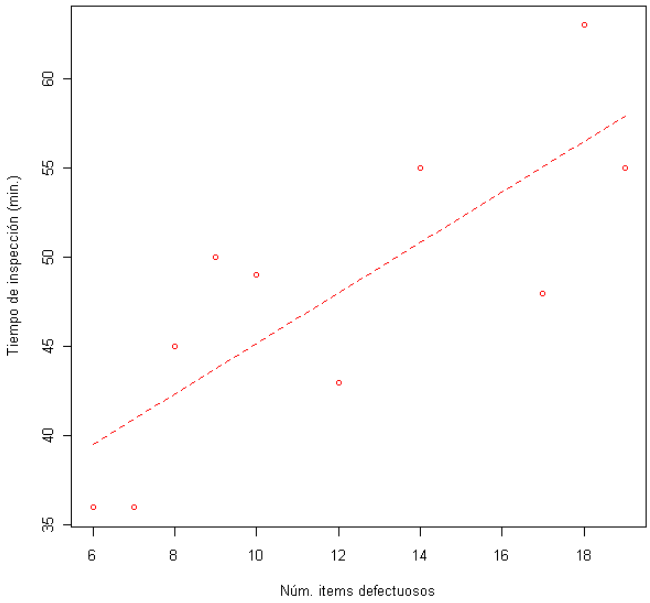
b) Recta de regresión por mínimos cuadrados: $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$. La ecuación de la recta es: $y = 2.2405 + 0.2261x$.

7. a) Recta de regresión por mínimos cuadrados: $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$. La ecuación de la recta es: $y = 7.9531 + 0.4408x$.

   b) Coeficiente de correlación (se calcula como en el ejercicio 4): $r_{(x,y)} = 0.9848$.

   c) El coeficiente de determinación es $r^2_{(x,y)} = 0.9699$, es decir, casi el 97 % de la variabilidad del tiempo de espera queda **explicada** por su dependencia lineal del número de pasajeros que llegan. Esto es, la relación lineal entre ambas variables es muy fuerte.

8. a) Diagrama de dispersión:

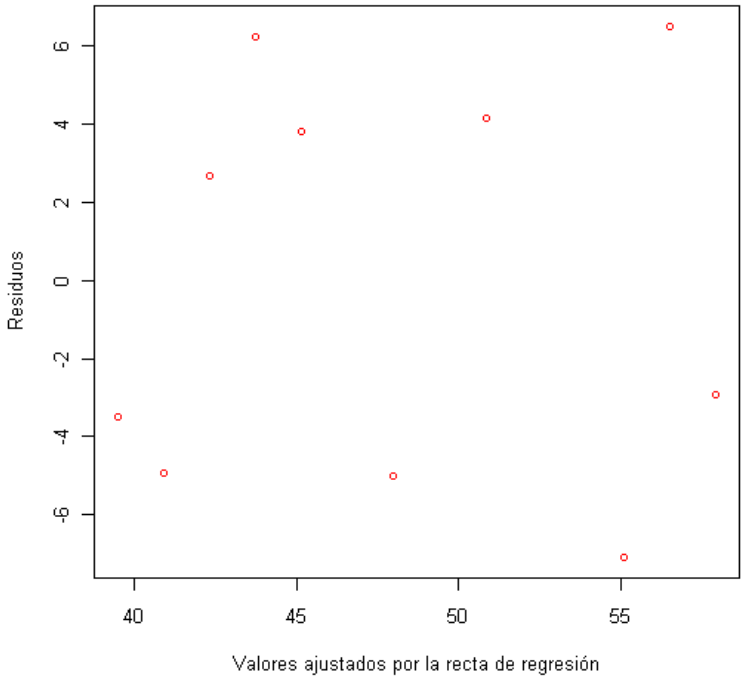


b) Recta de regresión por mínimos cuadrados: $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$. La ecuación de la recta es: $y = 31 + 1.4167x$.

c) Residuos:

# items def. (x_i)	17	9	12	7	8	10	14	18	19	6
t. inspección (y_i)	48	50	43	36	45	49	55	63	55	36
$\hat{y}_i = 31 + 1.4167x_i$	55.08	43.75	48.00	40.92	42.33	45.17	50.83	56.50	57.92	39.50
Res. ($e_i = y_i - \hat{y}_i$)	7.08	6.25	-5.00	-4.92	2.67	3.83	4.17	6.50	-2.92	-3.50

d) Gráfico de los residuos e_i frente a los valores predichos $\hat{y}_i$:



Los residuos se reparten de forma aleatoria en torno a la linea horizontal $y = 0$, y por tanto podemos decir que el ajuste de la recta de regresión es bueno.

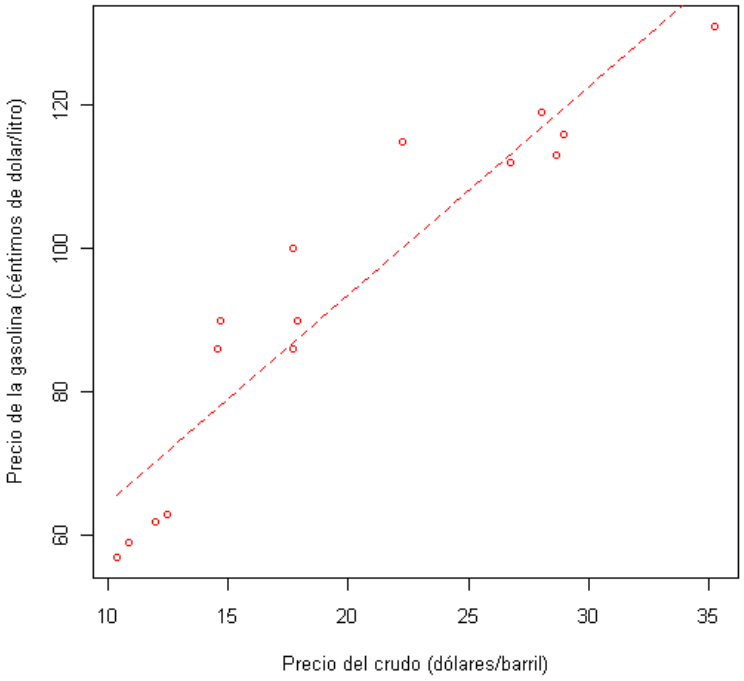
e) Coeficiente de determinación (cálculos como en el ejercicio 4): $r^2_{(x,y)} = 0.6299$. Esto quiere decir que el 70 % de la variabilidad del tiempo de inspección viene explicada por su dependencia lineal del número de items defectuosos.

9. El coeficiente de correlación es (cálculos como en el ejercicio 4): $r_{(x,y)} = 0.7911$. La relación lineal entre estas dos variables es positiva, es decir, a mayor tamaño de la familia mayor es el consumo de detergentes, ya que $r_{(x,y)}$ es positivo. Además, podemos decir que la relación lineal es fuerte ya que el valor del coeficiente de correlación es bastante alto (próximo a 1).

10. El coeficiente de correlación es (cálculos como en el ejercicio 4): $r_{(x,y)} = 0.7607$. La relación lineal entre estas dos variables es positiva, es decir, a mayor número de inventarios mayor es el porcentaje de ventas de estas compañías, ya que $r_{(x,y)}$ es positivo. Además, podemos decir que la relación lineal es fuerte ya que el valor del coeficiente de correlación es bastante alto (próximo a 1).

11. a) Recta de regresión para el precio de la gasolina (y) en función del precio del crudo (x): $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$. La ecuación de la recta es: $y = 35.51 + 2.91x$.

b) Diagrama de dispersión y la recta ajustada en el apartado anterior:



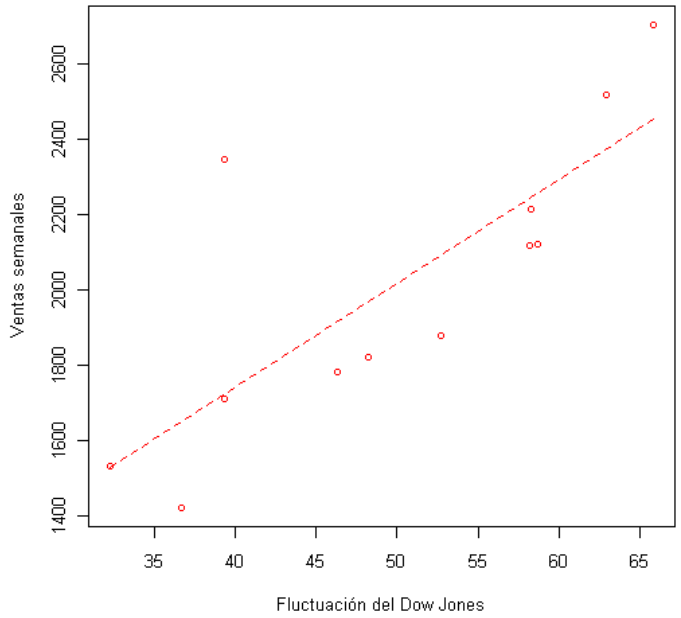
c) Si el precio del crudo cae a los 15\$, el precio estimado del litro de gasolina será $y(15) = 35.51 + 2.91 \cdot 15 = 79.16$ céntimos de dólar.

d) No tiene sentido hacerse la pregunta anterior para un precio del crudo de 0 dólares, ya que 0 no está dentro del rango de valores de x utilizados para calcular la recta de regresión.

e) Tampoco se puede emplear la recta de regresión obtenida en el apartado a) para predecir a futuro el precio del crudo a partir del precio de la gasolina, porque la relación a futuro entre los dos precios puede cambiar y dejar de tener el comportamiento descrito por la recta de regresión.

12. a) Recta de regresión para las ventas semanales (y) en función de la fluctuación del *Dow Jones* (x): $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$. La ecuación de la recta es: $y = 640.98 + 27.53x$.

Diagrama de dispersión y recta de regresión:



b) Parece haber cierta relación entre las dos variables, es decir, a mayores fluctuaciones en el *Dow Jones* se observan mayores ventas. En ese sentido se corroboraría la sospecha del dueño de la tienda. Sin embargo, podemos observar que el ajuste de la recta de regresión no es muy bueno. Se aprecia un dato atípico que "desplaza" la recta del centro de la nube. Y aún eliminando ese dato atípico, el resto de puntos tampoco parece seguir una tendencia lineal.

c) No necesariamente, ya que correlación no implica causalidad. En este caso, no parece razonable pensar que mayores fluctuaciones en el *Dow Jones* "provoquen" un aumento en las ventas. Lo que puede ocurrir es que haya variables subyacentes que tengan a la vez relación con las fluctuaciones del Dow Jones y las ventas de la tienda, y que hagan que cuando las primeras suban, las segundas suban también.

Estadística I
Ejercicios Tema 4
Curso 2009/10 – Soluciones

Observación: En todos los ejercicios de esta hoja usamos la notación $\overline{A}$ para referirnos al conjunto complementario del conjunto A.

1. En una urna hay 15 bolas numeradas de 2 al 16. Extraemos una bola al azar y observamos el número que tiene.

   a) Describe los sucesos, escribiendo todos sus elementos.

   A 'Obtener par' B 'Obtener impar' C 'Obtener primo'
   D 'Obtener impar menor que 9'

   b) ¿Qué relación hay entre A y B? ¿Y entre C y D?

   c) ¿Cuál es el suceso $A \bigcup B$? ¿y $C \bigcap D$?

   Solución.

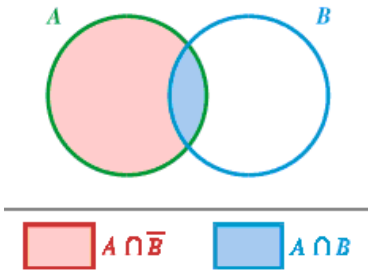
   a) $A = \{2,4,6,8,10,12,14,16\}$, $B = \{3,5,7,9,11,13,15\}$, $C = \{2,3,5,7,11,13\}$, $D = \{3,5,7\}$.

   b) $B = \overline{A}$ y $D \subset C$.

   c) $A \bigcup B = \Omega$ (Ω es el espacio muestral); $C \bigcap D = D$.

2. Sabiendo que $P[A \bigcap B] = 0.2$, que $P[\overline{B}] = 0.7$ y que $P[A \bigcap \overline{B}] = 0.5$, calcula $P[A \bigcup B]$ y $P[A]$.

   Solución.

   $$P[A] = P[A \bigcap \overline{B}] + P[A \bigcap B] = 0.5 + 0.2 = 0.7,$$
   $$P[B] = 1 - P[\overline{B}] = 1 - 0.7 = 0.3,$$
   $$P[A \bigcup B] = P[A] + P[B] - P[A \bigcap B] = 0.7 + 0.3 - 0.2 = 0.8.$$

3. Sabiendo que: $P[A] = 0.5$; $P[\overline{B}] = 0.6$; $P[\overline{A} \bigcap \overline{B}] = 0.25$,

   a) ¿son A y B sucesos independientes?

   b) Calcula $P[A \bigcup B]$ y $P[A|B]$.

   Solución.

   a)
   $$P[B] = 1 - P[\overline{B}] = 1 - 0.6 = 0.4,$$
   $$P[A \bigcup B] = 1 - P[\overline{A \bigcup B}] = 1 - P[\overline{A} \bigcap \overline{B}] = 1 - 0.25 = 0.75,$$
   $$P[A \bigcap B] = P[A] + P[B] - P[A \bigcap B] \quad \rightarrow$$
   $$\rightarrow \quad 0.75 = 0.5 + 0.4 - P[A \bigcap B] \quad \rightarrow \quad P[A \bigcap B] = 0.5 + 0.4 - 0.75 = 0.15.$$

   Por tanto: $P[A] \cdot P[B] = 0.5 \cdot 0.4 = 0.2$, mientras que $P[A \bigcap B] = 0.15$. Son distintos, luego los conjuntos A y B son independientes.

b) Hemos obtenido en el apartado anterior que: $P[A \bigcup B] = 0.75$. Por otra parte:

$$P[A|B] = \frac{P[A \bigcap B]}{P[B]} = \frac{0.15}{0.4} = 0.375.$$

4. En unas oposiciones, el temario consta de 85 temas. Se eligen tres temas al azar de entre los 85. Si un opositor sabe 35 de los 85 temas, ¿cuál es la probabilidad de que sepa al menos uno de los tres temas?

Solución. Tenemos que hallar la probabilidad de que ocurra el siguiente suceso:

A: 'el opositor conoce, al menos, uno de los tres temas'.

Para calcularla, utilizaremos el complementario, es decir: 'el opositor no conoce ninguno de los tres temas´. Si sabe 35 temas, hay 85 - 35 = 50 temas que no sabe; entonces:

$$P[A] = 1 - P[\overline{A}] = 1 - P[\text{'no sabe ninguno de los tres'}] = 1 - \frac{50}{85} \cdot \frac{49}{84} \cdot \frac{48}{83} = 1 - 0.198 = 0.802.$$

Por tanto, la probabilidad de que sepa al menos uno de los tres temas es de 0.802.

5. En una cadena de televisión se hizo una encuesta a 2.500 personas para saber la audiencia de un debate y de una película que se emitieron en horas distintas: 2.100 vieron la película, 1.500 vieron el debate y 350 no vieron ninguno de los dos programas. Si elegimos al azar a uno de los encuestados:

a) ¿Cuál es la probabilidad de que viera la película y el debate?

b) ¿Cuál es la probabilidad de que viera la película, sabiendo que vio el debate?

c) Sabiendo que vio la película, ¿cuál es la probabilidad de que viera el debate?

Solución. Organizamos la información en una tabla de doble entrada, completando los datos que faltan. Ver Cuadro 1.

	debate	no debate	
película	1450	650	2100
no película	50	350	400
	1500	1000	2500

Cuadro 1: Tabla películas.

Llamamos D = 'Vio el debate' y P = 'Vio la película'.

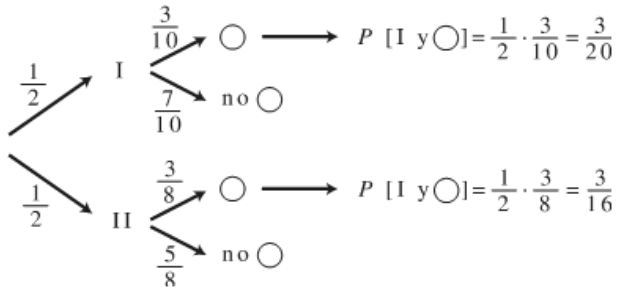
$$\begin{aligned} a)\ P\left[D \bigcap P\right] &= \frac{1450}{2500} = \frac{29}{50} = 0.58. \\ b)\ P[P|D] &= \frac{1450}{1500} = \frac{29}{30} = 0.97. \\ c)\ P[D|P] &= \frac{1450}{2100} = \frac{29}{42} = 0.69. \end{aligned}$$

6. Tenemos dos urnas: la primera tiene 3 bolas rojas, 3 blancas y 4 negras; la segunda tiene 4 bolas rojas, 3 blancas y 1 negra. Elegimos una urna al azar y extraemos una bola.

a) ¿Cuál es la probabilidad de que la bola extraída sea blanca?

b) Sabiendo que la bola extraída fue blanca, ¿cuál es la probabilidad de que fuera de la primera urna?

Solución. Hacemos un diagrama en árbol:



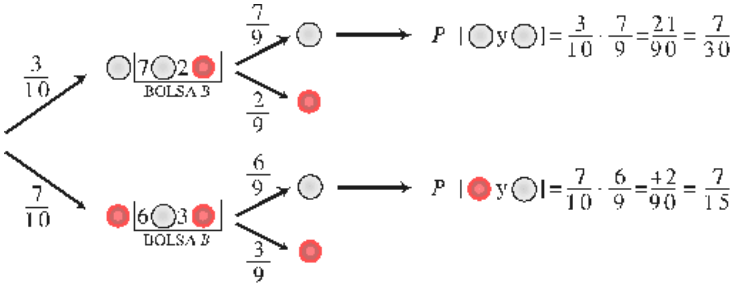
$$a)\ P[B] = \frac{3}{20} + \frac{3}{16} = \frac{27}{80}.$$
$$b)\ P[I|B] = \frac{I \bigcap B}{P[B]} = \frac{3/20}{27/80} = \frac{4}{9}.$$

7. Tenemos dos bolsas, A y B. En la bolsa A hay 3 bolas blancas y 7 rojas. En la bolsa B hay 6 bolas blancas y 2 rojas. Sacamos una bola de A y la pasamos a B. Después extraemos una bola de B.

   a) ¿Cuál es la probabilidad de que la bola extraída de B sea blanca?

   b) ¿Cuál es la probabilidad de que las dos bolas sean blancas?

   Solución. Hacemos un diagrama en árbol:



$$a)\ P[\text{'segunda bola blanca'}] = \frac{7}{30} + \frac{7}{15} = \frac{7}{30} \qquad b)\ P[\text{'}las dos blancas\text{'}] = \frac{7}{30}.$$

8. Lanzamos tres dados y anotamos el número de cincos que obtenemos.

   a) ¿Cuál es la distribución de probabilidad?

   b) Calcula la media y la desviación típica.

   Solución.

   a) Sea X la variable 'número de cincos obtenidos'. La variable X toma valores en el conjunto $\{0, 1, 2, 3\}$.
   $X = 0$ es que salga distinto de cinco en el primer dado, en el segundo y en el tercero.

$$P[X = 0] = P\,[\text{'no salga ningún 5'}] = \frac{5}{6}\frac{5}{6}\frac{5}{6} = \frac{125}{216} = 0.58.$$

   $X = 1$ es que se dé uno de los siguientes sucesos:

   - salga cinco en el primer dado, distinto de cinco en el segundo y distinto de cinco en el tercero,
   - o bien que salga distinto de cinco en el primer dado, cinco en el segundo y distinto de cinco en el tercero,
   - o bien que salga distinto de cinco en el primer dado, distinto de cinco en el segundo y cinco en el tercero.

$$P[X=1] = \frac{1}{6}\frac{5}{6}\frac{5}{6} + \frac{5}{6}\frac{1}{6}\frac{5}{6} + \frac{5}{6}\frac{5}{6}\frac{1}{6} = \frac{3\cdot 25}{216} = 0.35.$$

$X = 2$ es que se de uno de los siguientes sucesos:

- salga cinco en el primer dado, cinco en el segundo y distinto de cinco en el tercero,
- o bien que salga distinto de cinco en el primer dado, cinco en el segundo y cinco en el tercero,
- o bien que salga cinco en el primer dado, distinto de cinco en el segundo y cinco en el tercero.

$$P[X=2] = \frac{1}{6}\frac{1}{6}\frac{5}{6} + \frac{5}{6}\frac{1}{6}\frac{1}{6} + \frac{1}{6}\frac{5}{6}\frac{1}{6} = \frac{3\cdot 5}{216} = 0.07.$$

$X = 3$ es que salga cinco en el primer dado, en el segundo y en el tercero.

$$P[X=3] = P\,[\text{'salgan tres cincos'}] = \frac{1}{6}\frac{1}{6}\frac{1}{6} = \frac{1}{216} = 0.005.$$

Una vez tenemos esto, nos queda que la tabla de distribución de probabilidad es la siguiente:

x_i	0	1	2	3
p_i	0.58	0.35	0.07	0.005

NOTA: también se pueden calcular estas probabilidades usando que la variable 'salir cinco' sigue la distribución Bernoulli, y así X el número de veces que sale cinco, sigue la distribución Binomial con $n = 3$ y $p = \frac{1}{6}$.

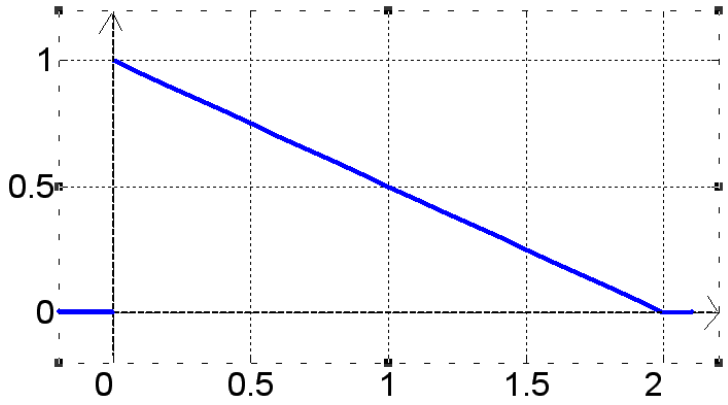
b) Ahora calculamos la media y la desviación típica de esta distribución.

$$\mu = \sum_{i=1}^{4} x_i p_i = 0\cdot 0.58 + 1\cdot 0.35 + 2\cdot 0.07 + 3\cdot 0.005 = 0.5,$$

$$\sigma^2 = \sum_{i=1}^{4} x_i^2 p_i - \mu^2 = 0^2\cdot 0.58 + 1^2\cdot 0.35 + 2^2\cdot 0.07 + 3^2\cdot 0.005 - 0.5^2 = 0.675 - 0.25 = 0.425,$$

$$\sigma = \sqrt{0.425} = 0.652.$$

9. La siguiente gráfica corresponde a la función de densidad de una variable continua X.



a) Calcula la probabilidad de que X sea menor que uno. Razónalo gráficamente.

b) Calcula la probabilidad de que X sea mayor que 0.5 y menor que 3/2. Razónalo analíticamente.

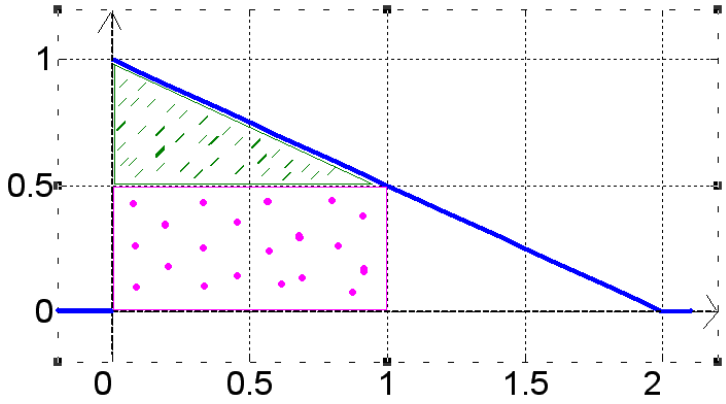
c) Calcula la media de la distribución.

d) Calcula la varianza de la distribución.

Solución. Primero vamos a escribir la función de densidad. Para ello nos fijamos que la recta dibujada es la recta $y = 1 - \frac{1}{2}x$. Entre 0 y 2 la función de densidad vale $1 - \frac{1}{2}x$, pero fuera de ese intervalo vale cero.

$$f(x) = \begin{cases} 1 - \frac{1}{2}x, & \text{si } x \in (0,2), \\ 0 & \text{si } x \notin (0,2). \end{cases}$$

a) Podemos calcular el área debajo de la curva en $x \in (-\infty, 1)$. Como la curva f sólo es positiva en $x \in (0, 2)$, entonces sólo tengo que calcular dicho área en $x \in (0, 1)$. Tengo dos trozos en ese área: un rectángulo rosa señalado con puntos y un triángulo verde señalado con rayas:



El área del rectángulo es $1 \cdot 0.5 = 0.5$ y el área del triángulo es $\frac{(1 \cdot 0.5)}{2} = 0.25$. Por tanto, $P[X < 1] = 0.5 + 0.25 = 0.75$.

b)

$$P[0.5 < X < 1.5] = \int_{0.5}^{1.5} \left(1 - \frac{1}{2}u\right) du = \left(u - \frac{1}{2}\frac{u^2}{2}\right)\Big|_{0.5}^{1.5} = 1.5 - \frac{1.5^2}{4} - 0.5 + \frac{0.5^2}{4} = \frac{1}{2}.$$

c)

$$\begin{aligned} \mu &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^2 x\left(1 - \tfrac{1}{2}x\right) dx = \int_0^2 \left(x - \tfrac{1}{2}x^2\right) dx = \left(\tfrac{x^2}{2} - \tfrac{x^3}{6}\right)\Big|_0^2 \\ &= \tfrac{4}{2} - \tfrac{8}{6} = \tfrac{12-8}{6} = \tfrac{2}{3} = 0.667. \end{aligned}$$

d)

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \int_0^2 x^2\left(1 - \tfrac{1}{2}x\right) dx - \mu^2 = \int_0^2 \left(x^2 - \tfrac{1}{2}x^3\right) dx - \mu^2 \\ &= \left(\tfrac{x^3}{3} - \tfrac{x^4}{8}\right)\Big|_0^2 - \left(\tfrac{2}{3}\right)^2 = \tfrac{8}{3} - \tfrac{16}{8} - \tfrac{4}{9} = \tfrac{2}{9} = 0.22. \end{aligned}$$

10. Un asesor financiero ha estimado que las ventas y los costes de algunos productos están relacionados con un índice I a través de las siguientes relaciones:

$$\text{Costes: } C = \tfrac{I+5}{7}, \qquad \text{Ventas: } V = \tfrac{25-I}{4}.$$

Si el índice I es una variable aleatoria X con función de densidad:

$$f_X(x) = \begin{cases} \frac{x}{108}, & \text{si } 3 \le x \le 15 \\ 0, & \text{en caso contrario.} \end{cases}$$

a) Calcular la función de distribución del índice I.

b) Calcular las medias y desviaciones típicas de los costes, las ventas y los beneficios.

c) Calcular la probabilidad de que el beneficio sea negativo.

Solución.

a) Calculamos la función de distribución de X. Para los valores entre 3 y 15:

$$F_X(x) = \int_{-\infty}^{x} \frac{u}{108} du = \int_3^x \frac{u}{108} du = \frac{u^2}{216}\Big|_3^x = \frac{x^2 - 9}{216}.$$

De manera que la función de distribución queda:

$$F_X(x) = \begin{cases} 0, & \text{si } x \le 3, \\ \frac{x^2-9}{216}, & \text{si } 3 \le x \le 15, \\ 1, & \text{si } x > 15. \end{cases}$$

b) Para esto primero calculamos la media de X y luego aplicamos sus propiedades para los costes, las ventas y los beneficios.

$$\mu_X = \int_{-\infty}^{+\infty} uf(u)du = \int_3^{15} u\frac{u}{108}du = \left.\frac{u^3}{324}\right|_3^{15} = \frac{31}{3},$$

$$\sigma_X^2 = \int_3^{15} u^2\frac{u}{108} - \left(\frac{31}{3}\right)^2 = \frac{92}{9} = 10.22,$$

$$\sigma = \sqrt{10.22} = 3.1972.$$

La variable índice I es en realidad la X. La variable beneficio B es igual a las ventas menos los costes, es decir:

$$B = V - C = \frac{25 - X}{4} - \frac{X + 5}{7} = \frac{115 - 11X}{28}.$$

Medias:

costes medios: $E[C] = \frac{E[X]+5}{7} = \frac{46}{21}$,

ventas medias: $E[V] = \frac{25-E[X]}{4} = \frac{11}{3}$,

beneficios medios: $E[V] = E[V] - E[C] = \frac{31}{21}$.

Desviaciones típicas:

d.t. del coste$d.t.[C] = \frac{1}{7}\sigma_X = 0.4567$,

d.t. de la venta$E[V] = \frac{1}{4}\sigma_X = 0.7993$,

d.t. del beneficio$E[V] = \frac{11}{28}\sigma_X = 1.256$.

c) El beneficio era $B = \frac{115-11X}{28}$, luego:

$$P[B < 0] = P[X > \frac{155}{11}] = \int_{\frac{155}{11}}^{15} \frac{x}{108}dx = 0.122.$$

11. Para cada una de las siguientes situaciones, indica si sigue una distribución binomial. En caso afirmativo, identifica en ella los valores de n y p :

    a) Lanzamos cien veces un dado y nos preguntamos por el número de unos que obtenemos.

    b) Extraemos una carta de una baraja y vemos si es un as o no. Sin devolverla al mazo, extraemos otra y también miramos si se trata de un as o no, ... y así sucesivamente hasta diez veces.

    Solución.

    a) Es una distribución binomial con $n = 100$, $p = 1/6$. Es decir, $\sim \mathcal{B}(100, 1/6)$.

    b) No es una binomial, pues la probabilidad de obtener as para la segunda carta es distinta que para la primera (al ser sin reemplazamiento las extracciones).

12. El 65 % de los alumnos de un cierto instituto cursan estudios universitarios al terminar el Bachillerato. En un grupo de ocho alumnos elegidos al azar, halla la probabilidad de que estudien una carrera:

    a) Alguno de ellos.

    b) Más de seis.

    c) Calcula la media y la desviación típica.

Solución. Si llamamos $X =$ 'número de alumnos, de un grupo de 8, que estudian carrera', se trata de una distribución binomial con $n = 8$, $p = 0.65$. Es decir, $\sim \mathcal{B}(8; 0.65)$.

a) $P[X > 0] = 1 - P[X = 0] = 1 - 0.35^8 = 0.9998$.

b) $P[X > 6] = P[X = 7] + P[X = 8] = \begin{pmatrix} 8 \\ 7 \end{pmatrix} 0.65^7 \cdot 0.35 + \begin{pmatrix} 8 \\ 8 \end{pmatrix} 0.65^8 = 0.169$.

c) Hallamos la media: $\mu = np = 8 \cdot 0.65 = 5.2$.

d) La desviación típica: $\sigma = \sqrt{npq} = \sqrt{8 \cdot 0.65 \cdot 0.35} = 1.35$.

13. En un sorteo que se realiza diariamente de lunes a viernes, la probabilidad de ganar es 0.1. Vamos a jugar los cinco días de la semana y estamos interesados en saber cuál es la probabilidad de ganar 0, 1, 2, 3, 4 ó 5 días.

a) Haz una tabla con las probabilidades.

b) Calcula la media y la desviación típica.

Solución.

a) Ver Cuadro 2

x_i	0	1	2	3	4	5
p_i	0.59049	0.32805	0.0729	0.081	0.0045	0.0001

Cuadro 2: Tabla sorteo

Observar que se trata de una $\mathcal{B}(5; 0.1)$ por ejemplo:

$$P[x_i = 0] = \begin{pmatrix} 5 \\ 0 \end{pmatrix} \cdot 0.1^0 \cdot 0.9^5 = 0.59049.$$

b) $\mu = np = 5 \cdot 0.1 = 0.5$, $\sigma = \sqrt{npq} = \sqrt{5 \cdot 0.1 \cdot 0.9} = 0.67$.

14. Explica para cada una de estas situaciones si se trata de una distribución binomial. En caso afirmativo, identifica los valores de n y p :

a) El 2 % de las naranjas que se empaquetan en un cierto lugar están estropeadas. Se empaquetan en bolsas de 10 naranjas cada una. Nos preguntamos por el número de naranjas estropeadas de una bolsa elegida al azar.

b) En una urna hay 2 bolas rojas, 3 blancas y 2 verdes. Sacamos una bola, anotamos su color y la devolvemos a la urna. Repetimos la experiencia 10 veces y estamos interesados en saber el número de bolas blancas que hemos extraído.

Solución.

a) Es una distribución binomial con $n = 10$, $p = 0.02$.

b) Es una distribución binomial con $n = 10$, $p = \frac{3}{7}$.

15. En cada una de estas situaciones, explica si se trata de una distribución binomial. En caso afirmativo, di cuáles son los valores de n y p:

a) El 3 % de las chinchetas que se hacen en una determinada fábrica salen defectuosas. Se empaquetan en cajas de 20 chinchetas. Estamos interesados en el número de chinchetas defectuosas de una caja elegida al azar.

b) En una urna hay 2 bolas rojas, 3 blancas y 2 verdes. Extraemos una bola, anotamos su color y la devolvemos a la urna. Repetimos la experiencia 10 veces y estamos interesados en saber el número de bolas de cada color que hemos obtenido.

Solución.

a) Es una distribución binomial con $n = 20$, $p = 0.003$.

b) No se trata de una distribución binomial ya que hay más de dos resultados posibles.

16. Una compañía telefónica recibe llamadas a razón de 5 por minuto. Si la distribución del número de llamadas es de Poisson, calcular la probabilidad de recibir menos de cuatro llamadas en un determinado minuto.

Solución. Sea X el número de llamadas por minuto que se reciben. Tenemos que X sigue una distribución de Poisson, con $\lambda = 5$. La distribución de probabilidad viene dada por:

$$P[X = x] = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Nos piden la probabilidad:

$P[X < 4] = P[X = 0]+P[X = 1]+P[X = 2]+P[X = 3] = 0.0067+0.0337+0.0842+0.1404 = 0.2650.$

17. El dueño de un criadero de árboles está especializado en la producción de abetos de Navidad. Estos crecen en filas de 300. Se sabe que por término medio 6 árboles no son aptos para su venta. Asume que la cantidad de árboles aptos para la venta por fila plantada sigue una distribución de Poisson.

a) Calcula la probabilidad de encontrar 2 árboles no vendibles en una fila de árboles.

b) Calcula la probabilidad de encontrar 2 árboles no vendibles en media fila de árboles.

Solución. Sea X el número de árboles no vendibles en una fila, tenemos que $X \sim \mathcal{P}(\lambda = 3)$. Sea Y el número de árboles no vendibles en media fila. El número medio de árboles no vendibles en media fila es 3. Tenemos que $Y \sim \mathcal{P}(\lambda = 3)$.

a)
$$P[X = 2] = \frac{6^2 \cdot e^{-6}}{2!} = 0.0446.$$

b)
$$P[Y = 2] = \frac{3^2 \cdot e^{-3}}{2!} = 0.2240.$$

18. Halla, en una distribución $\mathcal{N}(0, 1)$, las siguientes probabilidades:

a) $P[z > -0.2]$

b) $P[z > 1.27]$

c) $P[-0.52 < z < 1.03]$

Solución.

a) $P[z > -0.2] = P[z < 0.2] = 0.5793$

b) $P[z > 1.27] = 1 - P[z < 1.27] = 1 - 0.8980 = 0.1020$

c) $P[-0.52 < z < 1.03] = P[z < 1.03] - P[z < -0.52] = P[z < 1.03] - (1 - P[z > -0.52]) = P[z < 1.03] - (1 - P[z < 0.52]) = 0.8485 - (1 - 0.6985) = 0.5470$

19. El nivel de colesterol en una persona adulta sana sigue una distribución normal $\mathcal{N}(192, 12)$. Calcula la probabilidad de que una persona adulta sana tenga un nivel de colesterol:

a) Superior a 200 unidades.

b) Entre 180 y 220 unidades.

Solución.

a) Superior a 200 unidades.

$$P[X > 200] = P[\frac{x - 192}{12} > \frac{200 - 192}{12}] = P[z > 0.67] = 1 - P[z < 0.67] = 1 - 0.7486 = 0.2514$$

b) Entre 180 y 220 unidades.

$$\begin{aligned} P[180 < X < 220] &= P[\tfrac{180-192}{12} < \tfrac{x-192}{12} > \tfrac{220-192}{12}] = P[-1 < z < 2.33] \\ &= P[z < 2.33] - P[z < -1] = P[z < 2.33] - P[z > 1] \\ &= P[z < 2.33] - (1 - P[z < 1]) = 0.8314 \end{aligned}$$

20. El 7 % de los pantalones de una determinada marca salen con algún defecto. Se empaquetan en cajas de 80 para distribuirlos por diferentes tiendas. ¿Cuál es la probabilidad de que en una caja haya más de 10 pantalones defectuosos?

Solución. Si llamamos $X =$ 'número de pantalones defectuosos en una caja', entonces X es una binomial con $n = 80$ y $p = 0.07$. Hay que calcular $P[X > 10]$ La calculamos aproximando con una normal. La media de X es $np = 80 \cdot 0.07 = 5.6$. Su desviación típica es $\sigma = \sqrt{npq} = 2.28$. Así $X \sim \mathcal{B}(80; 0.07)$ se aproxima por $X' \sim \mathcal{N}(5.6; 2.28)$. Hay que tipificarla para tener $Z \sim \mathcal{N}(0, 1)$. *Atención:* No se ha aplicado la corrección por continuidad. El resultado sería más exacto si se aplicara.

$$P[X > 10] \approx P[X' > 10] = P\left[Z > \frac{10 - 5.6}{2.28}\right] = P[Z > 1.93] = 1 - P[Z < 1.93] = 1 - 0.9719 = 0.0281$$

21. Un examen de 100 preguntas admite como respuesta en cada una de ellas dos posibilidades, verdadero o falso. Si un alumno contesta al azar, calcula la probabilidad de que acierte más de 60 respuestas.

Solución. Si llamamos X al número de respuestas acertadas, entonces X sigue una distribución binomial con $n = 100$, $p = 1/2$. Tenemos que calcular $P[X > 60]$. La calculamos aproximando con una normal.
Primero calculamos la media de la binomial, y su desviación típica: $\mu = np = 50$ y $\sigma = \sqrt{npq} = 5$. Así consideramos las variables:

- $X \sim \mathcal{B}(100, 1/2)$ número de respuestas acertadas.
- $X' \sim \mathcal{N}(50, 5)$ la aproximación de X.
- $Z \sim \mathcal{N}(0, 1)$ es la normal estándard (se obtiene cuando tipificamos la X'.)

$$P[X > 60] \approx P[X' > 60] = P[Z > \frac{60 - 50}{5}] = P[Z > 2] = 1 - 0.9772 = 0.0228.$$

22. Una variable aleatoria X tiene la siguiente función de densidad

$$f(x) = \begin{cases} (1 + x^2)/12, & \text{si } x \in (0, 3), \\ 0, & \text{si } x \notin (0, 3). \end{cases}$$

Calcula:

a) la función de distribución de X,
b) las probabilidades $P(1 < X < 2)$ y $P(X < 1)$,
c) la esperanza y varianza de X,
d) la probabilidad $P(|X - E[X]| \geq 1)$ y compárala con la cota que se obtendría mediante la desigualdad de Chebychev.

Solución.

a) Calculamos la función de distribución de X:

$$F(x) = P(X \leq x) = P(X \in (-\infty, x]) = \int_{-\infty}^{x} f(t)\, dt,$$

es decir,

$$F(x) = \begin{cases} 0, & \text{si } x < 0, \\ (x + x^3/3)/12, & \text{si } 0 \leq x < 3, \\ 1, & \text{si } x \geq 3. \end{cases}$$

b) Calculamos las probabilidades $P(1 < X < 2)$ y $P(X < 1)$:

$$P(1 < X < 2) = \int_1^2 f(x)\,dx = \int_1^2 (1+x^2)/12\,dx = 0.278,$$

$$P(X < 1) = \int_{-\infty}^1 f(x)\,dx = \int_{-\infty}^0 0\,dx + \int_0^1 (1+x^2)/12\,dx = 0.111.$$

c) Calculamos la esperanza y varianza de X:

$$E[X] = \int_{-\infty}^{+\infty} x\,f(x)\,dx = \int_{-\infty}^0 x\,0\,dx + \int_0^3 x\,(1+x^2)/12\,dx + \int_3^{+\infty} x\,0\,dx = 2.0625.$$

$$E[X^2] = \int_{-\infty}^{+\infty} x^2\,f(x)\,dx = \int_{-\infty}^0 x^2\,0\,dx + \int_0^3 x^2\,(1+x^2)/12\,dx + \int_3^{+\infty} x^2\,0\,dx = 4.8,$$

y por tanto, $Var(X) = E[X^2] - (E[X])^2 = 4.8 - 2.0625^2 = 0.546$.

d) Primero calculamos exactamente la probabilidad $P(|X - E[X]| \geq 1)$, o bien utilizando la función de densidad de X o bien su función de distribución.

$$\begin{aligned} P(|X - E[X]| \geq 1) &= 1 - P(|X - E[X]| < 1) = 1 - P(-1 < X - E[X] < 1) \\ &= 1 - P(1.0625 < X < 3.0625) = 1 - P(1.0625 < X < 3) \\ &= 1 - [F(3) - F(1.0625)] = F(1.0625) = 0.1219, \end{aligned}$$

donde hemos tenido en cuenta que $E[X] = 2.0625$, y que X es una variable aleatoria continua con función de densidad diferente de cero en el intervalo $(0, 3)$.
En cambio, mediante la desigualdad de Chebychev obtenemos:

$$P(|X - E[X]| \geq 1) \leq \frac{Var(X)}{1^2} = 0.546,$$

que no es falso, pero tampoco es muy preciso. Recordad que esta desigualdad se utiliza como una aproximación de la probabilidad cuando no se dispone de la ley de probabilidad de la variable aleatoria.

23. Considerad la v.a. X que tiene ley uniforme discreta dada por la siguiente función de probabilidad:

$$P(X = x) = \begin{cases} 1/4, & x = 1, 2, 3, 4, \\ 0, & \text{en otro caso.} \end{cases}$$

Sean $X_1, \ldots, X_n$ son v.a. i.i.d. con la misma distribución que X, y considerad la v.a.

$$Y = \frac{1}{n}\sum_{i=1}^n X_i.$$

Calculad la probabilidad $P(2.4 < Y < 2.8)$ para $n = 36$.

Solución. Consideramos $X_1, X_2, \ldots, X_n$ v.a. i.i.d. con función de probabilidad $f(x)$. Calculamos la esperanza y varianza de una de estas v.a.:

$$E[X] = \frac{1}{4}(1+2+3+4) = \frac{5}{2}, \quad E[X^2] = \frac{1}{4}(1^2+2^2+3^2+4^2) = \frac{15}{2},$$

$$Var(X) = E[X^2] - (E[X])^2 = \frac{15}{2} - \frac{25}{4} = \frac{5}{4}.$$

Por tanto, según el T.C.L. la ley de Y, para $n = 36$, es:

$$Y \approx \mathcal{N}\left(\frac{5}{2}, \sqrt{\frac{5/4}{36}}\right) = \mathcal{N}(5/2, \sqrt{5/144}) = \mathcal{N}(2.5, 0.1871).$$

La probabilidad que nos piden es:

$$P(2.4 < Y < 2.8) = P\left(\frac{2.4 - 2.5}{0.1871} < Z < \frac{2.8 - 2.5}{0.1871}\right)$$

$$= P(-0.53 < Z < 1.60) = P(Z < 1.60) - P(Z < -0.53) = 0.64714.$$

Estadística I
Ejercicios Tema 5
Curso 2009/10 – Soluciones

---

1. La duración de un determinado tipo de pilas es una variable aleatoria con distribución normal de media de 50 horas y desviación típica de 5 horas. Empaquetamos las pilas en cajas de 16:

   a) ¿Cuál es la probabilidad de que la duración media de las pilas de una caja sea inferior a 48 horas?

   b) ¿Cuál es la probabilidad de que la duración de una de las pilas sea de entre 45 y 50 horas?

Solución. *X = "duración en horas de ese tipo de pilas". $X \sim N(50, 5)$. Tomamos una m.a.s. de la duración de 16 pilas: $X_1, \ldots, X_{16}$.*

a) Como la distribución de X es normal, tenemos que $\overline{X} \sim N(50, 5/\sqrt{16}) = N(50, 1.25)$. Entonces $Z = \frac{\overline{X}-50}{1.25} \sim N(0,1)$ y

$$\begin{aligned} P(\overline{X} < 48) &= P\left(\tfrac{\overline{X}-50}{1.25} < \tfrac{48-50}{1.25}\right) \\ &= P(Z < -1.6) = P(Z > 1.6) = 0.0548. \end{aligned}$$

b)

$$\begin{aligned} P(45 < X < 50) &= P\left(\tfrac{45-50}{5} < \tfrac{X-50}{5} < \tfrac{50-50}{5}\right) \\ &= P(-1 < X < 0) = P(X < 0) - P(X < -1) = 0.5 - 0.1587 = 0.3413. \end{aligned}$$

2. Las bolsas de azúcar envasadas por una cierta máquina tienen un peso medio de 500 gramos con una desviación típica de 35 gramos. Las bolsas se empaquetan en cajas de 100 unidades.

   a) Calcular la probabilidad de que el peso medio de las bolsas de una caja sea menor que 495 g.

   b) Calcular la probabilidad de que una caja pese más de 51 kg.

Solución. *X = "peso de las bolsas de azúcar en gramos". $E[X] = \mu = 500$ y $DT[X] = \sigma = 35$. Tomamos una m.a.s. del peso de 100 bolsas de azúcar: $X_1, \ldots, X_{100}$. No conocemos la distribución de X, pero como el tamaño de muestra es grande (> 30), podemos aplicar el Teorema central del límite, que dice que*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

cuando n es suficientemente grande.
Así pues, en este caso tendremos que $\overline{X} \sim N(500, 35/\sqrt{100}) = N(500, 3.5)$ aproximadamente.

a) Sea $Z = \frac{\overline{X}-500}{3.5} \sim N(0,1)$, entonces

$$\begin{aligned} P(\overline{X} < 495) &= P\left(\tfrac{\overline{X}-500}{3.5} \leq \tfrac{495-500}{3.5}\right) \\ &= P(Z \leq -1.43) = P(Z > 1.43) = 0.0764. \end{aligned}$$

b)

$$\begin{aligned} P(\textstyle\sum_{i=1}^{100} X_i > 51000) &= P(100 \cdot \overline{X} > 51000) = P(\overline{X} > 510) \\ &= P\left(\tfrac{\overline{X}-500}{3.5} > \tfrac{510-500}{3.5}\right) = P(Z > 2.86) = 0.0021. \end{aligned}$$

3. Para una muestra aleatoria simple $X_1, \ldots, X_4$ de una población de media μ y varianza $k\mu^2$, donde k es una constante desconocida, se consideran los siguientes estimadores de μ:

$$T_1 = \frac{X_1 + 4X_2}{5} \qquad T_2 = \frac{X_1 + X_2 + X_3 + X_4}{3}$$

a) Calcular el sesgo de T_1 y T_2.

b) Calcular el E.C.M. de T_1 y T_2.

c) ¿Para qué valores de k es el estimador T_2 mejor que T_1 de acuerdo al criterio del E.C.M.?

Solución. *Al tratarse de una muestra aleatoria simple tenemos que* $E[X_i] = \mu$ *y* $Var[X_i] = k\mu^2$, $i = 1, \ldots, 4$.

a)

$$E[T_1] = E\left[\frac{X_1 + 4X_2}{5}\right] = \frac{1}{5}E\left[X_1 + 4X_2\right] = \frac{1}{5}\left(E\left[X_1\right] + 4E\left[X_2\right]\right) = \frac{1}{5}\left(\mu + 4\mu\right) = \mu.$$

Puesto que T_1 *es insesgado (su esperanza coincide con el valor del parámetro), su sesgo es 0.*

$$E[T_2] = E\left[\frac{X_1 + X_2 + X_3 + X_4}{3}\right] = \frac{1}{3}E\left[X_1 + X_2 + X_3 + X_4\right] = \frac{1}{3}\left(E\left[X_1\right] + E\left[X_2\right] + E\left[X_3\right] + E\left[X_4\right]\right) =$$

$$\frac{1}{3}\left(\mu + \mu + \mu + \mu\right) = \frac{4}{3}\mu.$$

$$Sesgo(T_2) = E[T_2] - \mu = \frac{4}{3}\mu - \mu = \frac{1}{3}\mu.$$

b)

$$Var[T_1] = Var\left[\frac{X_1 + 4X_2}{5}\right] = \frac{1}{5^2}Var\left[X_1 + 4X_2\right] \overset{indep.}{=} \frac{1}{25}\left(Var\left[X_1\right] + 4^2 Var\left[X_2\right]\right) = \frac{1}{25}\left(k\mu^2 + 16k\mu^2\right) = \frac{17k\mu^2}{25}.$$

$$ECM(T_1) = Var[T_1] + Sesgo(T_1)^2 = Var[T_1] = \frac{17k\mu^2}{25}.$$

$$Var[T_2] = Var\left[\frac{X_1 + X_2 + X_3 + X_4}{3}\right] = \frac{1}{3^2}Var\left[X_1 + X_2 + X_3 + X_4\right] \overset{indep.}{=} \frac{1}{9}\left(Var\left[X_1\right] + Var\left[X_2\right] + Var\left[X_3\right] + \right.$$

$$\left. Var\left[X_4\right]\right) = \frac{1}{9}\left(k\mu^2 + k\mu^2 + k\mu^2 + k\mu^2\right) = \frac{4k\mu^2}{9}.$$

$$ECM(T_2) = Var[T_2] + Sesgo(T_2)^2 = \frac{4k\mu^2}{9} + \left(\frac{1}{3}\mu\right)^2 = \frac{(4k+1)\mu^2}{9}.$$

c)

$$ECM(T_2) \leq ECM(T_1) \Leftrightarrow \frac{(4k+1)\mu^2}{9} \leq \frac{17k\mu^2}{25} \Leftrightarrow 25\left(4k\mu^2 + \mu^2\right) \leq 9{\cdot}17k\mu^2 \Leftrightarrow 25\mu^2 \leq (153{-}100)k\mu^2$$

$$\overset{\mu \neq 0}{\Leftrightarrow} 25 \leq 53k \Leftrightarrow k \geq \frac{25}{53}.$$

Por tanto preferiremos T_2 *a* T_1*, de acuerdo al criterio del error cuadrático medio, cuando* k *sea mayor que* $25/53$*. (Si* $\mu = 0$ *ambos estimadores tendrían ECM igual a 0).*

4. Sea X la variable aleatoria cuya función de densidad es

$$f(x) = 0.5(1 + \theta x) - 1 \leq x \leq 1,$$

donde θ es un parámetro desconocido. Sea $X_1, \ldots, X_n$ una muestra aleatoria simple de tamaño n de X:

a) Demuestra que el estimador $\hat{\theta} = 3\overline{X}$ es un estimador insesgado de θ.

b) Si $n = 100$, calcula la probabilidad de que $\hat{\theta}$ sea mayor que θ.

Solución. *a) Vamos a calcular primero la esperanza de X:*

$$E[X] = \int_{-1}^{1} xf(x)dx = \int_{-1}^{1} x0.5(1+\theta x)dx = 0.5\left[\frac{x^2}{2} + \theta\frac{x^3}{3}\right]_{-1}^{1} = 0.5\left[\frac{1}{2} + \theta\frac{1}{3} - \frac{1}{2} + \theta\frac{1}{3}\right] = \frac{\theta}{3}.$$

Por tanto:

$$E[\hat{\theta}] = E[3\overline{X}] = 3E[\overline{X}] \overset{m.a.s.}{=} 3E[X] = 3\frac{\theta}{3} = \theta,$$

es decir, $\hat{\theta}$ es un estimador insesgado de θ.

b) Si $n = 100$, al tratarse de una m.a.s. podemos aplicar el teorema central del límite y tenemos que

$$\frac{\overline{X} - E[X]}{\sqrt{Var[X]/n}} \sim N(0,1) \Leftrightarrow \overline{X} \sim N(E[X], Var[X]/n)$$

y por tanto

$$\hat{\theta} = 3\overline{X} \sim N(3E[X], 9Var[X]/n) = N(\theta, 9Var[X]/n).$$

Por la simetría de la distribución normal, sabemos que $P(\hat{\theta} > \theta) = 0.5$.

5. Las notas de un test de aptitud siguen una distribucion normal con desviacion tipica 28.2. Una muestra aleatoria de 9 alumnos arroja los resultados siguientes:

$$\sum_{i=1}^{n} x_i = 1098 \qquad \sum_{i=1}^{n} x_i^2 = 138148$$

a) Hallar un intervalo de confianza al 90 % para la media poblacional.

b) Razonar sin hacer calculos si la longitud de un intervalo al 95 % será menor, mayor o igual que la del obtenido en el apartado anterior.

c) ¿Cuál será el tamaño de muestra mínimo necesario para obtener un intervalo al 90 % de nivel de confianza, con longitud 10? (longitud del intervalo = extremo superior-extremo inferior)

Solución.

$X =$ "notas del test de aptitud". $X \sim N(\mu, 28.2)$. Mediante muestreo aleatorio simple se toma una muestra donde

$$n = 9, \quad \overline{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = 122, \quad s = \sqrt{\frac{1}{8}\left(\sum_{i=1}^{9} x_i^2 - 9 \cdot \overline{x}^2\right)} = 21.58.$$

a) En este caso la cantidad pivotal es

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

y el intervalo de confianza para μ es

$$IC_{1-\alpha}(\mu) = \left[\bar{x} \pm z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right].$$

En nuestro caso

$$IC_{0.90}(\mu) = \left[122 \pm z_{0.05}\,\frac{28.2}{\sqrt{9}}\right] = [106.54\,,\,137.46]\,.$$

b) El intervalo al 95 % será mayor, puesto que a mayor nivel de confianza, mayor longitud del intervalo (a mayor α, mayor es el valor de $z_{\alpha/2}$).

c) La longitud del intervalo es $2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. Al nivel de confianza 0.95, si queremos un intervalo con longitud de a lo sumo 10:

$$2 \cdot z_{0.05} \frac{\sigma}{\sqrt{n}} \leq 10 \iff z_{0.05} \frac{\sigma}{\sqrt{n}} \leq 5 \iff z_{0.05} \frac{\sigma}{5} \leq \sqrt{n} \iff z_{0.05}^2 \frac{\sigma^2}{25} \leq n$$

$$\iff n \geq z_{0.05}^2 \frac{\sigma^2}{25} = 1.645^2 \frac{28.2^2}{25} = 86.08.$$

Por lo tanto el tamaño de muestra mínimo necesario será 87.

6. El gerente de operaciones de un periódico quiere determinar la proporción de periódicos impresos con defectos como demasiada tinta, configuración incorrecta de páginas, páginas duplicadas, etc. El gerente decide tomar una muestra aleatoria de 100 periódicos y encuentra que 35 contienen algún tipo de defecto.

   a) Si el gerente desea un 90 % de nivel de confianza al estimar la proporción verdadera de periódicos impresos con defectos, construye el intervalo de confianza.

   b) Utilizando la información muestral, determinar el tamaño de la muestra para que el error de estimación no sea superior al 5 %, con un nivel de confianza del 90 %.

   c) Si no se dispone de la información muestral, ni de información histórica fiable (caso más desfavorable), plantear el cálculo de n para el supuesto del apartado anterior.

Solución.

$X =$ "presencia de defectos en un periódico". $X \sim \mathcal{B}(p)$, donde p es la proporción de periódicos que se imprimen con defectos. Mediante muestreo aleatorio simple se toma una muestra donde

$$n = 100, \quad \hat{p} = \overline{x} = \frac{1}{100} \sum_{i=1}^{100} x_i = \frac{35}{100} = 0.35.$$

a) Tenemos una distribución de Bernouilli y un tamaño de muestra suficientemente grande para poder aplicar el Teorema Central del Límite, por lo tanto, el intervalo de confianza será:

$$IC_{1-\alpha}(p) = \left[\widehat{p} \pm z_{\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \right].$$

En nuestro caso,

$$IC_{0.90}(p) = \left[0.35 \pm z_{0.05} \sqrt{\frac{0.35 \cdot 0.65}{100}} \right] = [0.27\,,\, 0.43]\,.$$

b) El error de estimación es $z_{\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$. Si utilizamos la información muestral, es decir, suponemos que $\hat{p}$ va a valer aproximadamente 0.35 en cualquier muestra que tomemos, entonces a un nivel de confianza del 90 % tenemos que

$$z_{\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq 0.05 \iff z_{\alpha/2}^2 \frac{\widehat{p}(1-\widehat{p})}{n} \leq 0.05^2 \iff z_{\alpha/2}^2 \frac{\widehat{p}(1-\widehat{p})}{0.05^2} \leq n$$

$$\iff 1.645^2 \frac{0.35 \cdot 0.65}{0.05^2} \leq n \iff n \geq 1.645^2 \frac{0.35 \cdot 0.65}{0.05^2} = 246.25.$$

El tamaño de muestra mínimo necesario para obtener un error de estimación de a lo sumo el 5 % sería de 247.

c) En este caso no podemos suponer que $\hat{p}$ va a valer aproximadamente 0.35 en cualquier muestra, y por tanto como desconocemos $\hat{p}$ hemos de ponernos en el caso más desfavorable, es decir, cuando es igual a $1/2$. Entonces a un nivel de confianza del 90 % tenemos que

$$z_{\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq 0.05 \iff z_{\alpha/2}^2 \frac{\widehat{p}(1-\widehat{p})}{n} \leq 0.05^2 \iff z_{\alpha/2}^2 \frac{\widehat{p}(1-\widehat{p})}{0.05^2} \leq n$$

$$\iff 1.645^2 \frac{\widehat{p}(1-\widehat{p})}{0.05^2} \leq n \iff n \geq 1.645^2 \frac{0.25}{0.05^2} = 270.60.$$

El tamaño de muestra mínimo necesario para obtener un error de estimación de a lo sumo el 5 % sería en este caso de 271.

7. En la encuesta sobre intención de voto del CIS (febrero de 2008, link) de cara a las elecciones legislativas de 2008, aparece la siguiente información en la ficha técnica:

Error muestral:
Para un nivel de confianza del 95.5 % (dos sigmas), y P = Q, el error es de ±0.74 % para el conjunto de la muestra y en el supuesto de muestreo aleatorio simple.

¿Qué significa? ¿Cómo debemos interpretar los resultados de la encuesta?

Solución. *Cuando lo que queremos estimar es una proporción poblacional (en este caso, proporción de personas que votarán a un determinado partido), bajo las hipótesis del Teorema Central del Límite (m.a.s. y tamaño de muestra grande) sabemos que:*

$$\frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

de donde podemos obtener el siguiente intervalo de confianza para p:

$$IC_{1-\alpha}(p) = \left[\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right].$$

Pero puesto que desconocemos p, necesitamos sustituir $p(1-p)$ en la expresión del intervalo de confianza. Como $p \in [0,1]$, $p(1-p)$ en el intervalo $[0,1]$ es una parábola que alcanza su máximo en el punto $p = 0.5$, es decir, cuando $p = 1-p$ (lo que aparece expresado como P=Q en el enunciado). Entonces:

$$\forall p \in [0,1], p(1-p) \leq 0.5(1-0.5) = 0.25 \Rightarrow \forall p \in [0,1], \left[\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \subseteq \left[\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{0.25}{n}}\right].$$

El valor que se da como estimación de la proporción poblacional en las encuestas de intención de voto es $\widehat{p}$, pero el error muestral que se está cometiendo es $z_{\alpha/2}\sqrt{\frac{0.25}{n}}$ (la semiamplitud del intervalo, en este caso 0.0074) para un nivel de confianza de $(1-\alpha)$ % (en este caso 95.5 %, es decir, $\alpha = 0.045$).
Con estos datos podemos saber cuál ha sido el tamaño de muestra utilizado:

$$\alpha = 0.045 \Rightarrow z_{\alpha/2} \approx 2 \Rightarrow z_{\alpha/2}\sqrt{\frac{0.25}{n}} \approx \frac{1}{\sqrt{n}}.$$

(Obsérvese el uso de la expresión "dos sigmas", refiriéndose a que en la distribución normal se verifica que la probabilidad de que una variable $X \sim N(\mu, \sigma^2)$ tome valores en el intervalo $(\mu \pm 2\sigma)$ es 0.955).
Por tanto, el error muestral es 0.0074 si y sólo si $\frac{1}{\sqrt{n}} = 0.0074 \Leftrightarrow n \approx 18262$ (lo cual podemos comprobar en la primera página del documento del CIS, salvo errores de redondeo).