Solución del examen 7/02/06

(1) Llamamos X a la variable "nº de días de baja" y construimos la tabla auxiliar:

x_i	n_i	N_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
7	1	1	7	49
8	2	3	16	128
12	1	4	12	144
14	1	5	14	196
15	1	6	15	225
19	1	7	19	361
22	1	8	22	484
85	1	9	85	7225
total	9		190	8812

a) media aritmética: $\bar{x} = \frac{190}{9} = \underline{21.11}$

Mediana: $n = 9$ impar $\Rightarrow$ $Me = x_{\left(\frac{n+1}{2}\right)} = x_{(5)} = \underline{14}$

desviación típica:

$$s_n^2 = \overline{x^2} - \bar{x}^2 = \frac{8812}{9} - \left(\frac{190}{9}\right)^2 = 533.432$$

$$s_n = \sqrt{533.432} = 23.096 \simeq \underline{23.10}$$

b) Para dibujar el diagrama de caja necesitamos calcular Q_1, Q_3 y las barreras exteriores.

Para calcular Q_1 y Q_3 podemos

- utilizar la columna de frecuencias acumuladas:

$$\frac{n}{4} = 2.75 \Rightarrow Q_1 = 8$$

$$\frac{3n}{4} = 6.75 \simeq 7 \Rightarrow Q_3 = 19$$

- o bien, como hay pocos datos, calcularlos directamente:

7 | 8 8 | 12 | 14 | 15 | 19 22 | 85

$Q_1 = \frac{8+8}{2} = 8$ $\quad Q_2 = Me = 14$ $\quad Q_3 = \frac{19+22}{2} = 20.5$

Observación: Q_3 es distinto según el método usado. Se han dado por buenos ambos valores.

Para seguir con los cálculos voy a tomar $Q_3 = 20.5$

Rango intercuartílico: $RI = Q_3 - Q_1 = 12.5$

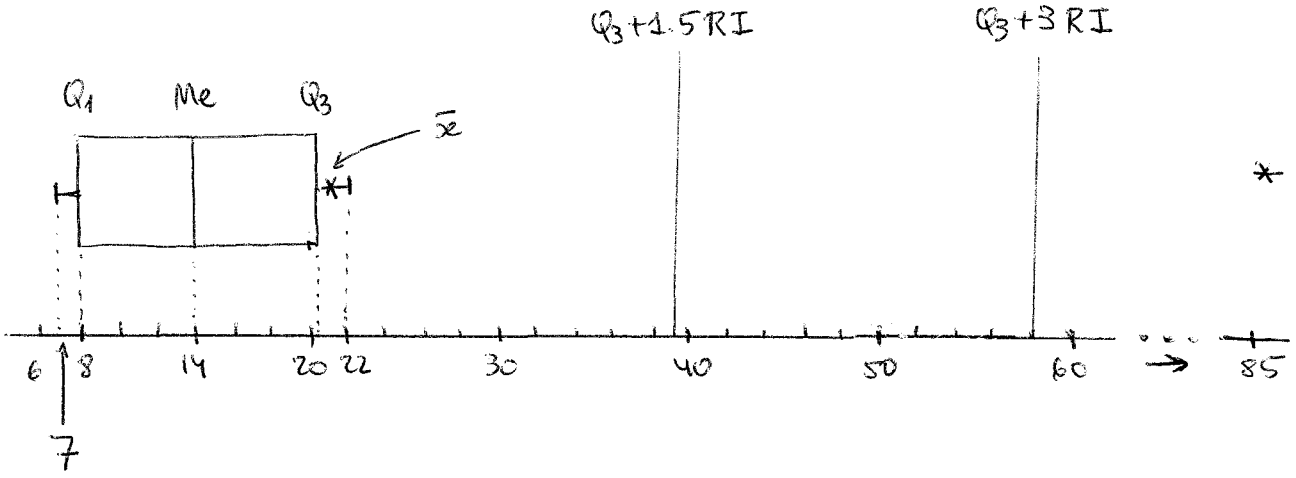
Barreras exteriores:

$Q_3 + 1.5\,RI = 39.25$ (primera B.E. superior)

$Q_3 + 3\,RI = 58$ (segunda B.E. superior)

$Q_1 - 1.5\,RI = -10.75$ ← está fuera del rango de los datos.

Con esta información dibujamos el diagrama de caja:



$x_i = 85$ es un valor atípico extremo, puesto que está fuera de la segunda B.E. superior.

Como medida de tendencia central sería preferible usar la mediana en lugar de la media, puesto que la mediana es robusta y, por tanto, no se ve afectada por la presencia de atípicos.

c) Se observa una asimetría hacia la derecha, puesto que $Me < \bar{x}$. Esto significa que hay más del 50% de los datos por debajo del punto medio de la distribución ($\bar{x}$). De hecho, mirando el diagrama de caja, vemos que hay más del 75% de los datos por debajo de $\bar{x}$. Esto es causado en parte por el valor atípico $x_i = 85$, que "tira de la media".

Puesto que la distribución no tiene forma de campana, calculamos el coeficiente de asimetría de Fisher:

$$As_F = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{s_n^3} = \frac{752100.69/9}{(23.096)^3} = 2.27 > 0$$

$\Downarrow$

asimetría positiva.

d) Para simetrizar estos datos, podemos utilizar las transformaciones siguientes: $\sqrt{x}$, $\log x$, $1/x$, ordenadas de menor efecto a mayor.
Las transformaciones del tipo x^p, $p>1$, no son adecuadas para este tipo de asimetría.

e) La relación entre los diagramas A, B, C, D y las transformaciones x^2, $\log x$, $\sqrt{x}$, $1/x$ es:

A $\longrightarrow$ $\log x$

B $\longrightarrow$ $\sqrt{x}$

C $\longrightarrow$ x^2

D $\longrightarrow$ $1/x$

(2) Llamemos $X =$ "nº de horas", $Y =$ "edad", $Z =$ "sexo".

a) X es cuantitativa continua, agrupada en intervalos.

Y es cuantitativa discreta (o continua), también agrupada en intervalos.

Z es categórica nominal.

Distribuciones marginales de las 3 variables:

Para X	x_i	$n_{i\cdot}$
$[0,20]$	10	33
$(20,30]$	25	24
$(30,40]$	35	43
		100

$x_i n_{i\cdot}$	$x_i^2 n_{i\cdot}$
330	3300
600	15000
1505	52675
2435	70975

tabla auxiliar para el apartado f)

Para Y	y_j	$n_{\cdot j}$
$[15,25]$	20	31
$(25,45]$	35	44
$(45,65]$	55	25
		100

$N_{\cdot j}$	$y_j n_{\cdot j}$	$y_j^2 \cdot n_{\cdot j}$
31	620	12400
75	1540	53900
100	1375	75625
		141925

tabla auxiliar para el apartado b) y para el apartado f)

Para Z:

z_k	n_k
Hombre	46
Mujer	54
	100

b) Mediana de edad:

$\frac{n}{2} = 50 \Rightarrow$ intervalo mediano $(25,45]$.

$$Me = 25 + (45-25)\frac{50-31}{75-31} = \underline{33.64}$$

c) $$\bar{x}_H = \frac{10\cdot 13 + 25\cdot 11 + 35\cdot 22}{13+11+22} = \frac{1175}{46} = 25.54$$

$$\bar{x}_M = \frac{10\cdot 20 + 25\cdot 13 + 35\cdot 21}{20+13+21} = \frac{1260}{54} = 23.33$$

$\bar{x}_H > \bar{x}_M \Rightarrow$ En media, los hombres miran más la TV.

d) El % de hombres que mira la TV más de 20 horas es $\frac{11+22}{46}\cdot 100\% = \underline{71.74\%}$

e) Nos piden la moda de las mujeres que miran la TV menos de 30 horas:

$[0,30]$

$Y =$ "edad"	n_i	n_i/L_i	
$[15,25]$	7	0.7	
$(25,45]$	24	1.2	← intervalo modal $(25,45]$
$(45,65]$	2	0.1	

$$Mo = 25 + (45-25)\frac{0.1}{0.7+0.1} = \underline{27.5}$$

f) Para medir la dispersión utilizaremos el coeficiente de variación, puesto que se trata de variables expresadas en distintas unidades de medida (una son horas y la otra años). Por tanto, no se pueden comparar sólo utilizando las desviaciones estándar o sólo las varianzas.

Edad: $\bar{y} = \frac{3535}{100} = 35.35$

$$s_y^2 = \overline{y^2} - \bar{y}^2 = \frac{141925}{100} - (35.35)^2 = 169.6275$$

Nº horas: $\bar{x} = \frac{2435}{100} = 24.35$

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{70975}{100} - (24.35)^2 = 116.8275$$

$$CV_X = \frac{s_x}{\bar{x}} = \frac{\sqrt{116.8275}}{24.35} = \underline{0.44}$$

$$CV_Y = \frac{s_y}{\bar{y}} = \frac{\sqrt{169.6275}}{35.35} = \underline{0.37}$$

$\Rightarrow$ X presenta más dispersión que Y.

g) Para ver si existe alguna relación entre la edad y el nº de horas que miran la TV podemos utilizar el coef. de correlación lineal de Pearson, puesto que las dos variable son cuantitativas. El coef. de correlación de Spearman no es adecuado porque las variables de nuestro estudio no son puntuaciones dadas por un observador externo.
Las medidas g2, g4 y g5 son para tablas de contingencia, no para tablas de doble entrada, como es nuestro caso.

	Nº horas (x_i)			
Edad (y_j)	10	25	35	$\left(\sum_i x_i n_{ij}\right) y_j$
20	2	10	19	18700
35	28	12	4	25200
55	3	2	20	42900
				86800

$$S_{xy} = \overline{xy} - \bar{x}\,\bar{y} = \frac{86800}{100} - (24.35)(35.35) = 7.2275$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{7.2275}{\sqrt{(169.6275)(116.8275)}} = \underline{\underline{0.05}}$$

r_{xy} es muy cercano a cero, por lo que deducimos que la relación lineal es extremadamente débil. De hecho, podríamos decir que no existe.
Aunque puede existir algún otro tipo de relación.

h) La variable "sexo" es categórica nominal, por tanto para estudiar la relación entre esta variable y el "nº de horas" deberemos utilizar medidas para tablas de contingencia, y en concreto, para variables nominales.
Quedan excluidas h1, h3.
Las medidas que se pueden utilizar son: h2, h4, h5.

El procedimiento a seguir es:

1º Realizamos un contraste de independencia basado en el estadístico χ^2 para saber si hay o no independencia. Si las variables son independientes, hemos terminado.
Si las variables son dependientes:

2º Medimos el grado de dependencia mediante el coef. de contingencia de Pearson o la V de Cramer. Ambas toman valores entre 0 y 1. Cuánto más cercanas a 1 mayor grado de asociación (dependencia) entre las variables.

(3) Llamamos $Y =$ "nº de televisores", $X =$ "ingresos".

$\bar{y} = 2$, $s_y^2 = 0.04$, $\bar{x} = 930$, $s_x^2 = 2410$, $r_{xy} = 0.9$.

a) Escogemos la recta de regresión de Y sobre X; nº de televisores en función de los ingresos mensuales.

$Y = a + bX$.

$$b = \frac{s_{xy}}{s_x^2} = \frac{r_{xy}\, s_x s_y}{s_x^2} = r_{xy}\frac{s_y}{s_x} = 0.9\sqrt{\frac{0.04}{2410}} = \underline{0.0037}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}; \quad s_{xy} = r_{xy}\, s_x s_y$$

$$a = \bar{y} - b\bar{x} = 2 - 0.0037(930) = \underline{-1.41}$$

b) Para valorar la bondad de ajuste del modelo utilizamos el coef. de determinación R^2, que en este caso coincide con r_{xy}^2.

$R^2 = r_{xy}^2 = 0.9^2 = 0.81 \longrightarrow$ 81% de la variabilidad de Y es explicada por el modelo.

Queda sin explicar el $100 - 81 = 19\%$ de la variabilidad.

El modelo es bastante fiable.

c) $x_i = 900$

$\hat{y}_i = -1.41 + (0.0037)900 = 1.92 \approx 2 \longrightarrow$ 2 televisores.

$x_i = 900$ es muy cercano a $\bar{x} = 930$.

(Recordemos que las mejores predicciones se obtienen para valores cercanos a la media)

d) $b > 0 \Rightarrow$ pendiente positiva $\longrightarrow$ descartamos D

$r_{xy} = 0.9 \Rightarrow$ relación lineal fuerte $\Rightarrow$ descartamos B

$R^2 = 0.81 \Rightarrow$ 81% de ajuste $\longrightarrow$ descartamos C.

El <u>gráfico A</u> es el que corresponde a nuestra nube de puntos.