Statistics I Final Exam, 26 May 2014.
Degrees in ADE, DER-ADE, ADE-INF, FICO, ECO, ECO-DER.

EXAM RULES: 1) Use separate booklets for each problem. 2) Perform the calculations with at least two significant decimal places. 3) You may not leave the exam during the first 30 minutes. 4) You are not allowed to leave the classroom without handing in the exam.

1. The following tables and figures contain information about emergency calls to 112 (Civil Protection emergency telephone number) involving injured people in non-domestic accidents in Pamplona during the last 180 days. For each day, the variables observed are: the average response time (in minutes) of the ambulance service and the number of attended accidents according to their typology (traffic accident, labor accident and other).

Table 1: Average response time (in minutes)

(a) Global		(b) Working days		(c) Weekend	
Mean	13.9189854	Mean	12.65007263	Mean	20.26354953
Median	7.72569374	Median	7.43135166	Median	16.45645042
Standard Deviation	14.1249749	Standard Deviation	12.75266538	Standard Deviation	18.59931717
Sample Variance	199.514915	Sample Variance	162.6304743	Sample Variance	345.9345992
Kurtosis	2.00834811	Kurtosis	2.762768312	Kurtosis	-0.193607802
Skewness	1.52135364	Skewness	1.650832286	Skewness	0.858543921
Range	65.1620149	Range	61.09465423	Range	64.99579226
Minimum	0.02300493	Minimum	0.023004926	Minimum	0.189227592
Maximum	65.1850198	Maximum	61.11765915	Maximum	65.18501985
Sum	2505.41738	Sum	1897.510894	Sum	607.9064858
Sample size	180	Sample size	150	Sample size	30

Table 2: Number of attended accidents

	Type of accident		
	traffic	labor	other
Working days	111	9	30
Weekend	26	1	3

Figure 1: Average response time (global)

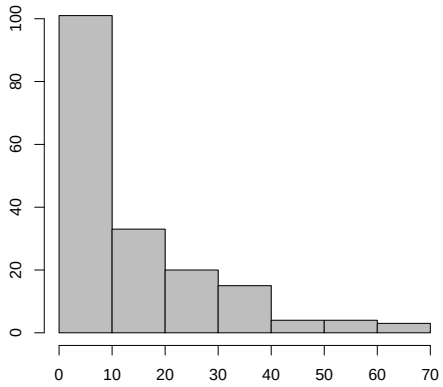
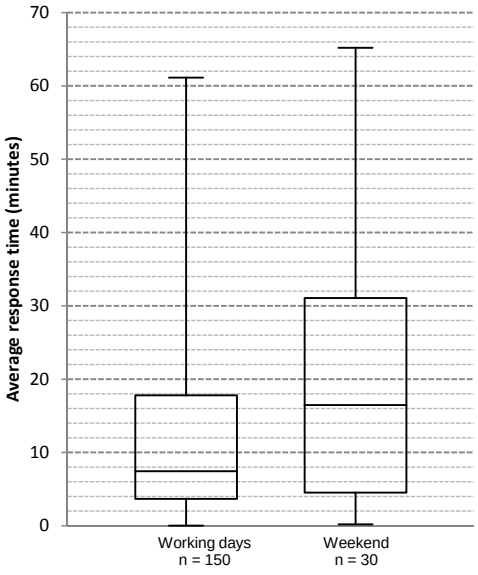


Figure 2: Average response time according to day typology



Answer the following questions and indicate, as clearly as possible, the tables and/or figures that help to justify your answer:

a) (0.5 points) Describe the form of the distribution of the average response time (global) and indicate if there are any outliers. Which measure of centrality is more appropriate in this case? Justify your answer.

b) (0.75 points) Decide whether the following sentences are true or false: **[1]** In more than half of the cases, the average response time of the ambulance service during the weekend is more than twice higher than that of the working days. **[2]** In 75% of the cases, the average response time of the ambulance service is greater than 30 minutes at weekends.

c) (0.5 points) Provide a maximum response time under which it is possible to ensure that 75% of the emergencies have been attended during the working days. Justify your answer.

d) (0.75 points) Calculate the distribution of the variable *type of accident* conditioned to the *type of day* and give the day type in which traffic accidents are the most frequent. Justify your answer.

Solución.

a) El Gráfico 1 es el histograma de la distribución de frecuencias absolutas de la variable *tiempo medio de respuesta (en minutos)* (sin distinguir por tipo de día). Se observa claramente que la distribución es asimétrica (con asimetría a la derecha, positiva).

Al tratarse de una distribución asimétrica es mejor dar la Mediana como medida de centralización. En este caso, el tiempo mediano de respuesta (ver panel (a) de la Tabla 1) es de 7.7257 minutos.

De forma aproximada (a partir del Gráfico 1) puede llegar a intuirse la presencia de datos atípicos. El razonamiento sería el siguiente: El tercer cuartil está en la posición 135, que corresponde a un valor situado entre 10 y 20, es decir, $Q_3 \in (10, 20)$. Por tanto, se deduce que $Q_3 < 20$, al igual que $RIC < 20$. Así, $Q_3 + 1.5\,RIC < 50$, con lo que cualquier dato superior a 50 puede ser considerado como atípico.

b) La afirmación **[1]** es cierta, puesto que la mediana del tiempo medio de respuesta de las ambulacias durante los fines de semana es de 16.46, mientras que de lunes a viernes vale 7.43 (ver paneles (b) y (c) de la Tabla 1 y también diagramas de caja del Gráfico 2). Sin embargo, la afirmación **[2]** es falsa, puesto que en el diagrama de caja del Gráfico 2 se observa que, para los fines de semana, el tercer cuartil del tiempo medio de respuesta es 31. Por tanto, el tiempo medio de repuesta es inferior a 31 minutos en el 75% de los casos.

c) Nos están pidiendo el tercer cuartil de la muestra: Q_3. En el diagrama de cajas correspondiente se aprecia que $Q_3 \approx 18$ minutos.

d) Para poder compararlas es más conveniente obtener las distribuciones condicionadas relativas (obtenidas a partir de la Tabla 2), que son:

Tipo de accidente más frecuente / tipo de día

	trafico	laboral	otros	
Lunes a Viernes	0,74	0,06	0,2	1
Sábado-Domingo	0,8666667	0,0333333	0,1	1

De ella se observa que los accidentes de tráfico son más frecuentes durante el fin de semana, con un porcentaje del 86.67%.

2. In a certain company, each visit of the technical service to fix a breakdown in the computer system has a cost of 350 euros, plus a fixed monthly fee of 175 euros regardless of the usage of the service. The monthly average number of breakdowns is 9.5 with a standard deviation of 2.

a) (0.5 points) Obtain the expectation and variance of the monthly cost of reparations (including the monthly fee).

b) (0.75 points) By using Chebyshev's inequality, provide a bound for the probability that in a given month the reparations cost is lower than or equal to 2000 euros or greater than or equal to 5000 euros.

c) (0.75 points) If we now assume that the monthly cost of reparations is continuous uniformly distributed with the expectation and variance obtained in *a)*, compute again the probability of the previous part.

d) (0.5 points) How can you explain the differences between the results obtained in parts *b)* and *c)*?

Solución.

a) X = número de averías al mes. $E[X] = 9.5$ y $Var[X] = 2^2 = 4$.
Sea C = coste mensual de las reparaciones. Como $C = 350X+175$, entonces $E[C] = 350 \cdot 9.5+175 = 3500$ y $Var[C] = 350^2 \cdot 4 = 490000$.

b)

$$P(C \leq 2000 \text{ ó } C \geq 5000) = P(C - 3500 \leq -1500 \text{ ó } C - 3500 \geq 1500) = P(|C - 3500| \geq 1500)$$

$$= P(|C - E[C]| \geq 1500) \overset{Des.Cheb.}{\leq} \frac{Var[C]}{1500^2} = \frac{490000}{1500^2} = 0.22.$$

c) Si $C \sim U(a, b)$, entonces $E[C] = \frac{a+b}{2} = 3500$ y $Var[C] = \frac{(b-a)^2}{12} = 490000$. Por tanto

$$\left.\begin{array}{rcl} \frac{a+b}{2} & = & 3500 \\ \frac{(b-a)^2}{12} & = & 490000 \end{array}\right\} \Leftrightarrow \left.\begin{array}{rcl} a & = & 7000 - b \\ b - a & = & \sqrt{12 \cdot 490000} \end{array}\right\} \Leftrightarrow \left.\begin{array}{rcl} a & = & \frac{7000-\sqrt{12 \cdot 490000}}{2} = 2287.56 \\ b & = & \sqrt{12 \cdot 490000} + 2287.56 = 4712.44 \end{array}\right\}$$

Es decir, $C \sim U(2287.56, 4712.44)$ por lo que $P(C \leq 2000 \;\; C \geq 5000) = 0$.

d) En el primer caso sólo contamos con los valores de la esperanza y la varianza, y la desigualdad de Chebyshev proporciona una cota muy gruesa de la probabilidad que se quiere calcular. En el segundo caso, al disponer de la distribución de la variable aleatoria podemos calcular la probabilidad de manera exacta.

3. The lifetime (in days) of a certain product has the following density function:

$$f(x) = \begin{cases} k\,(2 - x), & \text{si } 0 < x < 2, \\ 0, & \text{otherwise,} \end{cases}$$

for a given value of k.

a) (0.5 points) Obtain the value of the constant k.

b) (0.5 points) We know that this sort of product can only be sold between 12 and 48 hours after it has been manufactured. Calculate the probability that such a product can be sold. (Assume that one day has 24 hours).

c) (0.75 points) A certain company is able to manufacture 4 products of this kind per day. What is the probability that all the products manufactured during the last day cannot be sold?

d) (0.75 points) The company will break if all the products manufactured in at least 320 days of the year cannot be sold. Calculate the probability of bankruptcy of the company. (Assume that one year has 365 days).

Solución:

a) $\int_0^2 k\,(2 - x)\,dx = 1 \quad \Rightarrow \quad k\left(2x - \frac{x^2}{2}\right)\Big|_0^2 = 1 \quad \Rightarrow \quad k = \frac{1}{2}$.

b) Sea X la v.a. que representa la duración en días de cierto producto. Si el producto solamente está en condiciones óptimas para su venta entre 12 horas y 48 horas después de haber sido producido, entonces $P(0.5 < X < 2) = 0.5625$.

c) Sea Y la v.a. que cuenta el número de productos óptimos para la venta de un total de 4. $Y \sim B(4, 0.5625)$ y entonces, $P(Y < 4) = 1 - P(Y = 4) = 0.899$.

d) Sea T la v.a. que cuenta el número de días de un total de 365 en que no se venden todos los productos fabricados. $T \sim B(365, 0.899)$ que puede aproximarse por el TCL como $N(328.46, 5.983)$ y, por tanto, $P(\text{“}quiebra\text{''}) = P(T > 320) = P(Z > -1.41) = 1 - 0.0808 = 0.9192$, donde $Z \sim N(0, 1)$.

4. The points scored by a basketball team over a game can be modeled as a random variable with mean 90 and standard deviation 30.

   (a) (0.5 points) Given a sample of 35 games, compute the probability that the average score of the team is between 80 and 100 points.

   (b) (0.75 points) Given a sample of 35 games, calculate the 95% confidence interval for the average score per game. Assume that the sample mean is 85. Interpret the result.

   (c) (0.75 points) We are interested in reducing in a 10% the length of the confidence interval computed above. Obtain the sample size (number of games) that fulfilled such requirement.

   (d) (0.5 points) Now imagine that the team hires a new coach which is expected to reduce the standard deviation up to 15 (without changing the mean). Given that the league consists of 50 games, (approximately) plot the density function of the average score of a league for both coaches and compare both curves. Justify the results.

Solution:

a) Puesto que $n > 30$, puede aplicarse el TLC, que asegura que $\overline{X}$ tiene una ley aproximadamente normal. En concreto: $P\left(80 < \overline{X} < 100\right) = P\left(\frac{80-90}{30/\sqrt{35}} < Z < \frac{100-90}{30/\sqrt{35}}\right) = P\left(-1.97 < Z < 1.97\right) = 0.9512$, donde $Z = \frac{\bar{X}-90}{30/\sqrt{35}}$ sigue aproximadamente una distribucin $N(0,1)$.

b) En virtud del TLC, el intervalo de confianza es $I.C._{95\%}(\mu) = \overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 85 \pm 1.96\frac{30}{\sqrt{35}} = [75.06, 94.94]$. No sería correcto afirmar que con un 95% de probabilidad, el verdadero valor de la media está entre $[75.06, 94.94]$. Sin embargo, sí podríamos afirmar que el 95% de las veces que construimos un I.C. del modo anterior, encontraremos el verdadero valor de la media dentro de los mismos.

c) La longitud del intervalo es $2 \times 1.96 \times \frac{\sigma}{\sqrt{n_1}}$. Por tanto, queremos un n_2 tal que la nueva longitud del intervalo sea $2 \times 1.96 \times \frac{\sigma}{\sqrt{n_1}} \times 0.9$. Por tanto, tenemos la siguiente relación:

$$2 \times 1.96 \times \frac{\sigma}{\sqrt{n_1}} \times 0.9 = 2 \times 1.96 \times \frac{\sigma}{\sqrt{n_2}} \Rightarrow n_2 = \frac{n_1}{0.9^2} = \frac{35}{0.9^2} = 43.2 \rightarrow n_2 = 44.$$

d) Si dibujamos las funciones de densidad para el puntuaje promedio de un partido en una temporada, i.e. para $\overline{X}$, vemos que para ambos entrenadores la función está centrada en la misma media. Sin embargo, el segundo entrenador tiene un promedio de puntuaje más estable, por lo que dará resultados más homogéneos a lo largo de la temporada.