Statistics I Final Exam, 23 June 2015.
Degrees in ADE, DER-ADE, ADE-INF, FICO, ECO, ECO-DER, TUR.

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 contain information about the *GDP* and the *unemployment rate* of the Spanish Autonomous Regions:

Table 1

	GDP	Unemployment rate
Andalucía	13595.8	18.6
Aragón	18766	6.6
Asturias (Principado de)	15287.7	11.2
Balears (Illes)	20389.2	9.7
Canarias	16832.7	11.4
Cantabria	17425.7	10.6
Castilla y León	16920.4	11.1
Castilla-La Mancha	13978.5	10.1
Cataluña	20783.7	10.1
Comunidad Valenciana	16656.1	11.2
Extremadura	12240.6	17.4
Galicia	14683.7	12.7
Madrid (Comunidad de)	22968.7	7.4
Murcia (Región de)	14675.9	10.7
Navarra (Comunidad Foral de)	22065.1	5.7
País Vasco	22444.5	9.5
Rioja (La)	19624.4	6
Ceuta y Melilla	16213.8	9.2

Table 2

	GDP	Unemployment rate
mean	17530.7	
median	16876.6	
quasi-deviation	3251.1	
Q1	14834.7	9.3
Q3	20198.0	
Interquartile Range	5363.3	1.9
percentile 85	21360.3	12.0
percentile 15	14362.1	

Answer and justify the following questions:

(a) (0.5 points) Fill in the gaps in Table 2.

(b) (0.5 points) Which of the two variables is more disperse?

(c) (0.5 points) Determine the group of Autonomous Regions formed by the 15% with higher *GDP*.

(d) (0.5 points) Draw the box-plot of the *unemployment rate*. What can you tell about the shape of the distribution?

(e) (0.5 points) From the previous box-plot, decide if there are outliers and/our extreme outliers in the data. Identify the Autonomous Regions that can be considered outliers and/or extreme outliers.

Solución.

(a) En la Tabla 2 faltan algunos estadísticos descriptivos para la variable *tasa de paro*: media 10.5, mediana 10.3, cuasi-desviación típica 3.4, $Q_3 = 11.2$, $P_{15} = 7.0$.

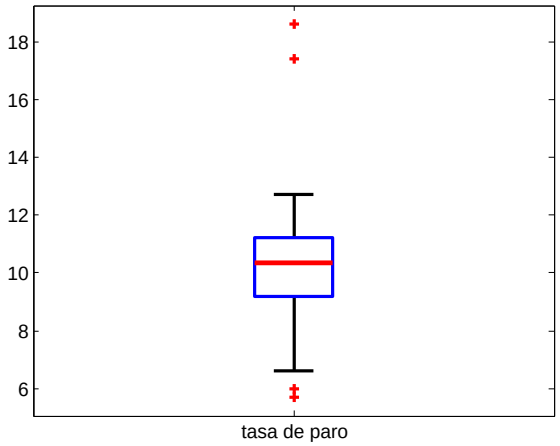
(b) Puesto que las unidades de medida y el rango de valores son muy distintos para el *PIB* y la *tasa de paro*, la cuasi-desviación típica no es un buen descriptivo para comparar sus variabilidades. Es mejor utilizar una medida adimensional, como el coeficiente de variación (CV). En este caso,

$$CV(PIB) = \frac{3251.1}{17530.7} = 0.19, \quad CV(t.paro) = \frac{3.4}{10.5} = 0.32,$$

luego la variación de la *tasa de paro* es mayor.

(c) El grupo de CCAA formado por el 15% con mayor *PIB* son aquellas cuyo *PIB* sea superior al percentil 85, es decir superior a 21360.3. Hay tres CCAA que cumplen esta condición: Navarra, País Vasco y Madrid.

(d) Diagrama de caja para la *tasa de paro*



La distribución presenta una ligera asimetría hacia la derecha, puesto que la media es ligerante superior a la mediana. La posible causa de esta asimetría son dos atípicos extremos con tasas de paro superiores a 17.

(e) Para la variable *tasa de paro* será un atíco cualquier observación superior a $Q_3 + 1.5 \times RI = 14.1$ o inferior a $Q_1 - 1.5 \times RI = 6.4$. Además, serán atípicos extremos aquellas observaciones superiores a $Q_3 + 3 \times RI = 16.9$ o inferiores a $Q_1 - 3 \times RI = 3.6$. Luego Navarra y La Rioja son atípicos porque sus tasas de paro son inferiores a 6.4 y Extremadura y Andalucía son atípicos extremos porque sus tasas de paro son superiores a 16.9.

2. The duration (in minutes) of certain manufacturing process is a random variable X with probability density function given by

$$f(x) = \begin{cases} \frac{1}{2} + c\,(1-x), & 1 \le x \le 2, \\ 0, & \text{elsewhere,} \end{cases}$$

where c is an appropriate constant.

(a) (0.5 points) Find the value of c so that $f(x)$ can be considered a probability density function. Draw the probability density function.

(b) (0.75 points) Calculate the probability that the manufacturing process lasts less than 90 seconds.

(c) (0.75 points) Calculate the expectation and variance of X.

(d) (0.5 points) Give the exact value of the probability $P(|X - \mathsf{E}(X)| \ge 0.3)$. Compare the previous result with the upper bound given by Chebyschev's inequality. Are both results contradictory? When is recommended to use Chebyschev's inequality? (Hint: If you were not able to compute the expectation and variance of X, you can use $\mathsf{E}(X) = 1.583$ and $\mathsf{var}(X) = 0.076$).

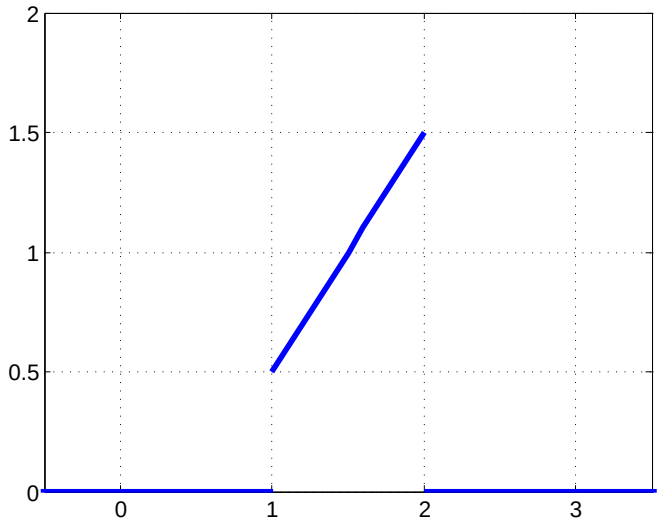
Solución.

(a) Para que $f(x)$ sea una función de densidad debe integrar 1 sobre su soporte, es decir,

$$\int_1^2 \left(\frac{1}{2} + c\,(1-x)\right) dx = 1 \Rightarrow \left(\frac{1}{2}\,x + c\left(x - \frac{x^2}{2}\right)\right]_1^2 = 1 \Rightarrow \frac{1}{2} + c\left(-\frac{1}{2}\right) = 1 \Rightarrow c = -1,$$

de manera que la función de densidad es

$$f(x) = \begin{cases} x - \frac{1}{2}, & 1 \le x \le 2, \\ 0, & \text{en caso contrario.} \end{cases}$$



(b) La probabilidad de que el proceso de fabricación dure menos de 90 segundos, es decir, menos de 1.5 minutos es

$$P(X < 1.5) = \int_1^{1.5} \left(x - \frac{1}{2}\right) dx = \left(\frac{x^2}{2} - \frac{x}{2}\right]_1^{1.5} = 0.375.$$

(c) La esperanza y varianza de X son:

$$\mathsf{E}(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx = \int_1^2 \left(x^2 - \frac{x}{2}\right) dx = \left(\frac{x^3}{3} - \frac{x^2}{4}\right]_1^2 = \frac{19}{12} = 1.5833.$$

Para calcular la varianza de X, utilizamos que $\mathsf{var}(X) = \mathsf{E}(X^2) - (\mathsf{E}(X))^2$, donde

$$\mathsf{E}(X^2) = \int_1^2 \left(x^3 - \frac{x^2}{2}\right) dx = \left(\frac{x^4}{4} - \frac{x^3}{6}\right]_1^2 = \frac{31}{12} = 2.5833,$$

luego $\mathsf{var}(X) = 31/12 - (19/12)^2 = 11/144 = 0.0764$.

(d) Para calcular $P(|X - \mathsf{E}(X)| \geq 0.3)$ de forma exacta se usa de nuevo la función de densidad de X:

$$\begin{aligned} P(|X - \mathsf{E}(X)| \geq 0.3) &= 1 - P(|X - \mathsf{E}(X)| < 0.3) = 1 - P(-0.3 < X - \mathsf{E}(X) < 0.3) \\ &= 1 - P(1.2833 < X < 1.8833) = 1 - \int_{1.2833}^{1.8833} \left(x - \frac{1}{2}\right) dx \\ &= 1 - \left(\frac{x^2}{2} - \frac{x}{2}\right]_{1.2833}^{1.8833} = 0.35002, \end{aligned}$$

donde se ha utilizado que $\mathsf{E}(X) = 1.5833$.

La aproximación que proporciona la desigualdad de Chebyschev es:

$$P(|X - \mathsf{E}(X)| \geq 0.3) \leq \frac{\mathsf{var}(X)}{0.3^2} = \frac{0.0764}{0.3^2} = 0.8489,$$

donde se observa que esta cota superior (0.8489) es mucho mayor que el valor exacto (0.35002). Este resultado no es contradictorio, puesto que Chebyschev proporciona una *cota superior*, es decir, que la probabilidad exacta siempre será menor que la aproximación mediante Chebyschev. Por este motivo, se recomienda usar la desigualdad de Chevyschev para aproximar probabilidades de una v.a. solamente cuando se desconozca la ley de probabilidad de dicha v.a.

3. A travel agency offers three types of destinations: regional, national and international. In general, the percentage of sales use to be 30% for regional destinations, 20% for national, 50% for international and the percentage of claims tend to be 1% for regional, 1% for national, 1.5% for international.

(a) (0.5 points) Compute the percentage of claims.

(b) (0.75 points) Given a claim, obtain the probability that comes from a regional destination.

(c) (0.5 points) Compute the probability that a client hire an international destination and does not claim.

(d) (0.75 points) Compute the probability that 3 or more of 10 clients hire international destinations and do not claim.

Solución. Consideramos los siguientes sucesos Reg =“destino regional”, Nac =“destino nacional” e $Inter$ =“destino internacional”, con probabilidades $P(Reg) = 0.30$, $P(Nac) = 0.20$, $P(Inter) = 0.50$), respectivamente. Además, consideramos los sucesos R =“el cliente emite reclamación” y $\bar{R}$ =“el cliente no emite reclamación”.

(a) El porcentaje de reclamaciones se obtiene aplicando el teorema de la probabilidad total:

$$\begin{aligned} P(R) &= P(R|Reg)\cdot P(Reg) + P(R|Nac)\cdot P(Nac) + P(R|Inter)\cdot P(Inter) \\ &= 0.01\cdot 0.30 + 0.01\cdot 0.20 + 0.015\cdot 0.50 = 0.0125, \end{aligned}$$

luego el porcentaje de reclamaciones es del 1.25%.

(b) Para obtener la probabilidad $P(Reg|R)$ es necesario aplicar el teorema de Bayes:

$$P(Reg|R) = \frac{P(R|Reg)\cdot P(Reg)}{P(R)} = \frac{0.01\cdot 0.30}{0.0125} = 0.24.$$

(c) La probabilidad de que un cliente contrate un destino internacional y no emita ninguna reclamación se obtiene como la intersección de los sucesos $Inter$ y $\bar{R}$, es decir:

$$P(Inter\cap \bar{R}) = \underbrace{P(\bar{R}|Inter)}_{1-P(R|Inter)}\cdot P(Inter) = 0.985\cdot 0.5 = 0.4925.$$

(d) Conisderemos la v.a. X =“número de clientes de un total de 10 que contratan destino internacional y no emiten reclamación”, que puede modelarse según una ley Binomial $Bin(10, 0.4925)$. Luego, la probabilidad de que 3 o más clientes contraten un destino internacional y no emitan ninguna reclamación es:

$$\begin{aligned} & P(X\geq 3) = 1 - P(X<3) = 1 - (P(X=0) + P(X=1) + P(X=2)) \\ = \;& 1 - \left[\binom{10}{0} 0.4925^0(1-0.4925)^{10} + \binom{10}{1} 0.4925^1(1-0.4925)^9 + \binom{10}{2} 0.4925^2(1-0.4925)^8\right] \\ = \;& 1 - \left[0.5075^{10} + 10\cdot 0.4925\cdot 0.5075^9 + 45\cdot 0.4925^2\cdot 0.5075^8\right] = 1 - 0.0602 = 0.9398. \end{aligned}$$

4. The weekly earnings of nonsupervisory production workers in the mining industry is assumed to have expectation and standard deviation equal to 630€ and 35€, respectively.

(a) (1 point) Compute the probability that the average weekly earnings of 100 workers is between 600€ and 660€.

(b) (0.5 points) Without doing any additional calculations, decide if the probability in (a) will increase or decrease if the number of workers is 200 instead of 100. Justify your answer.

(c) (0.5 points) Compute that probability that the sum of weekly earnings of 100 workers is larger than 70000€.

(d) (0.5 points) The government suspects that some of the mining companies are paying less money to nonsupervisory production workers. To check this, the government collects an independent sample of weekly earnings of 50 workers and obtain a sample mean of 605€. Obtain a 90% confidence interval for the mean earnings assuming that the earnings follows a Normal distribution with standard deviation 35. How confident are you that the true mean is in fact 630€?

Solution:

(a) Since $n \geq 30$, by CLT we have that $\overline{X} \sim N\left(630, \frac{35}{\sqrt{100}}\right) = N(630, 35)$. Then, the required probability is given by:

$$\Pr\left(600 < \overline{X} < 660\right) = \Pr\left(\frac{-30}{3.5} < \frac{\overline{X}}{3.5} < \frac{30}{3.5}\right) \simeq \Pr\left(-8.57 < Z < 8.57\right) =$$
$$= \Pr\left(Z < 8.57\right) - \Pr\left(Z < -8.57\right) = 1 - 0 = 1,$$

where $Z \sim N(0, 1)$.

(b) As $200 > 100$, the variance of $\overline{X}$ will decrease. Therefore, the probability for $\overline{X}$ to be in the interval $(600, 660)$ should increase. However, it is 1 already with $n = 100$.

(c) We want to compute the following probability:

$$\Pr\left(\sum_{i=1}^{100} X_i > 70000\right)$$

This is similar to:

$$\Pr\left(\overline{X} > \frac{70000}{100}\right) = \Pr\left(\overline{X} > 700\right)$$

Then, as in a), by CLT we have that $\overline{X} \sim N(630, 3.5)$. Therefore, the required probability is given by:

$$\Pr\left(\sum_{i=1}^{100} X_i > 70000\right) = \Pr\left(\overline{X} > 700\right) = \Pr\left(\frac{\overline{X} - 630}{3.5} > \frac{700 - 630}{3.5}\right) \simeq$$
$$\simeq \Pr\left(Z > 20\right) = 1 - \Pr\left(Z \leq 20\right) = 1 - 1 = 0$$

where $Z \sim N(0, 1)$.

(d) The confidence interval for the mean is given by:

$$\left(605 - 1.65\frac{35}{\sqrt{50}}, 605 + 1.65\frac{35}{\sqrt{50}}\right) = (596.8329, 613.1671)$$

Consequently, we are more than 90% confident that the true weekly earnings is smaller than 630€.