Estadística I Examen Extraordinario, 23 Junio 2015.
Grados en ADE, DER-ADE, ADE-INF, FICO, ECO, ECO-DER, TUR.

REGLAS DEL EXAMEN: 1) Usar cuadernillos separados para cada problema. 2) Hacer los cálculos con al menos dos decimales significativos. 3) No se puede abandonar el examen durante los primeros 30 minutos. 4) No está permitido salir de la clase sin entregar el examen.

1. En las siguientes tablas se recoge información acerca del *PIB* y de la *tasa de paro* de las Comunidades Autónomas españolas:

Tabla 1

CCAA	PIB	Tasa de paro
Andalucía	13595.8	18.6
Aragón	18766.0	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)	22065.1	5.7
País Vasco	22444.5	9.5
Rioja (La)	19624.4	6.0
Ceuta y Melilla	16213.8	9.2

Tabla 2

	PIB	Tasa de paro
media	17530.7	
mediana	16876.5	
cuasi-desviación	3251.1	
Q1	14834.7	9.3
Q3	20198.0	
Rango intercuartílico	5363.3	1.9
percentil 85	21360.3	12.0
percentil 15	14362.0	

Responder a las siguientes cuestiones, justificando cada una de ellas:

(a) (0.5 puntos) Completar los datos que faltan en la Tabla 2.

(b) (0.5 puntos) ¿Cuál de las dos variables tiene mayor variación?

(c) (0.5 puntos) Determinar el grupo de Comunidades Autónomas formado por el 15% de las que tienen mayor *PIB*.

(d) (0.5 puntos) Dibujar el diagrama de caja (box-plot) para la *tasa de paro*. ¿Qué forma presenta la distribución?

(e) (0.5 puntos) A partir del box-plot anterior, determinar si hay atípicos y/o atípicos extremos. Identificar de qué comunidades se trata.

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, $Q_3 = 11.2$, $P_{15} = 7.0$.

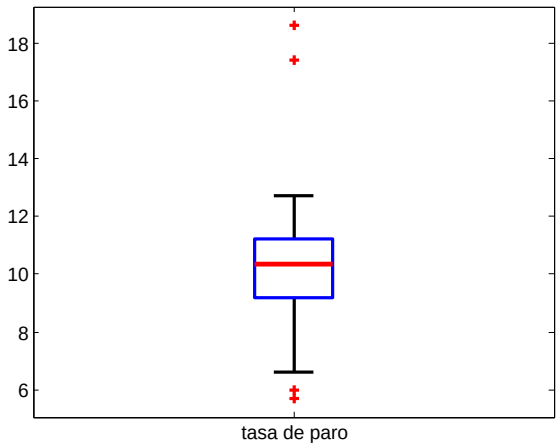
(b) Puesto que las unidades de medida y el rango de valores son muy distintos para el *PIB* y la *tasa de paro*, la cuasi-desviación típica no es un buen descriptivo para comparar sus variabilidades. Es mejor utilizar una medida adimensional, como el coeficiente de variación (CV). En este caso,

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

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. La duración (en minutos) de cierto proceso de fabricación es una variable aleatoria X con función de densidad

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

donde c es una constante adecuada.

(a) (0.5 puntos) Determinar el valor de la constante c para que $f(x)$ sea realmente una función de densidad. Dibujar la función de densidad.

(b) (0.75 puntos) ¿Cuál es la probabilidad de que el proceso de fabricación dure menos de 90 segundos?

(c) (0.75 puntos) Calcular la esperanza y la varianza de X.

(d) (0.5 puntos) Calcular exactamente la probabilidad $P(|X - \mathsf{E}(X)| \ge 0.3)$. Comparar este resultado con la cota superior que se obtendría mediante la desigualdad de Chebyschev. ¿Es contradictorio el resultado? ¿Cuando es adecuado utilizar la aproximación que porporciona la desigualdad de Chebyschev? (Indicación: Si no has podido calcular la esperanza y varianza de X, puedes considerar $\mathsf{E}(X) = 1.583$ y $\mathsf{var}(X) = 0.076$).

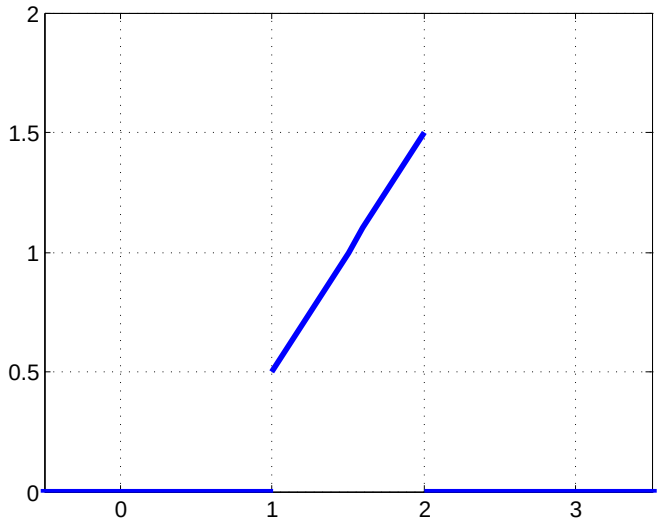
Solución.

(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. Una agencia de viajes ofrece tres tipos de destinos: regional, nacional e internacional. En general, los porcentajes de ventas suelen ser del 30% para regional, del 20% para nacional y del 50% para internacional y las reclamaciones que suele recibir son del 1% para regional, del 1% para nacional y del 1.5% para internacional.

(a) (0.5 puntos) Determinar el porcentaje de reclamaciones que recibe.

(b) (0.75 puntos) Determinar la probabilidad de que dada una reclamación, ésta sea debida a un destino regional.

(c) (0.5 puntos) Determinar la probabilidad de que un cliente contrate un destino internacional y no emita ninguna reclamación.

(d) (0.75 puntos) De un total de 10 clientes, determinar la probabilidad que 3 o más de ellos contraten un destino internacional y no emitan ninguna reclamación.

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. El salario semanal de trabajadores de producción no supervisada en la industria minera se supone que tiene media y desviación típica iguales a 630€y 35€, respectivamente. Se pide:

(a) (1 punto) Calcular la probabilidad de que el salario medio semanal de 100 de estos trabajadores esté entre 600€ y 660€.

(b) (0.5 puntos) Sin realizar ningún tipo de cálculo adicional, decidir si la probabilidad en (a) se incrementará o se reducirá si el número de trabajadores es 200 en lugar de 100. Justica tu respuesta.

(c) (0.5 puntos) Calcular la probabilidad de que la suma de salarios de 100 trabajadores sea mayor de 70000€.

(d) (0.5 puntos) El gobierno sospecha que algunas de las compañías mineras están pagando a los trabajadores de producción no supervisada un sueldo inferior al estipulado. Para comprobar esto, el gobierno obtiene una muestra del salario semanal de 50 trabajadores y obtienen una media muestral de 605€. Se pide obtener un intervalo de confianza al 90% para el salario medio suponiendo que dichos salarios tienen una distribución Normal con desviación típica 35. ¿Que nivel de confianza tienes en que la media verdadera sea realmente 630€?

Solución:

(a) Puesto que $n \geq 30$, por el CLT tenemos que $\overline{X} \sim N\left(630, \frac{35}{\sqrt{100}}\right) = N(630, 3.5)$. Entonces, la probabilidad requerida está dada por:

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

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

(b) Puesto que $200 > 100$, la varianza de $\overline{X}$ se espera que sea menor. Por lo tanto, la probabilidad de que $\overline{X}$ esté en el intervalo $(600, 660)$ debería aumentará (aunque ya es 1).

(c) Queremos calcular la probabilidad siguiente:

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

Esta cantidad es similar a:

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

Entonces, como en a), por el TCL tenemos que $\overline{X} \sim N(630, 3.5)$. Entonces, la probabilidad requerida está dada por:

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

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

(d) El intervalo de confianza para la media está dado por:

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

En consecuencia, estamos más de un 90% seguros de que el salario mensual verdadero medio es menor de 630€.