Examen Final de Estadística I, 27 de Mayo de 2013.
Grados en ADE, DER-ADE, ADE-INF, FICO, ECO, ECO-DER.

NORMAS: 1) Entregar cada problema en un cuadernillo distinto, aunque esté en blanco.
2) Realizar los cálculos con al menos dos cifras decimales significativas.
3) No se podrá abandonar el examen hasta transcurridos 30 minutos depués de haber empezado.
4) No está permitido salir del aula sin entregar el examen, aunque esté en blanco.

1. La siguiente tabla contiene un resumen estadístico sobre 47 facultades de ciencias empresariales de renombre en Estados Unidos en 2004. Las variables consideradas son: salario inicial medio, expresado en dólares, porcentaje de recién licenciados empleados al acabar sus estudios (% empleados), puntuación dada por evaluadores externos a los estudios (calificación) y porcentaje de candidatos aceptados en la facultad (% aceptación).

Salario Inicial Medio (en $)		*% empleados*		*% aceptación*		*calificación*	
Media	84045.89	Media	61.82	Media	30.59	Media	3.74
Error típico	1604.68	Error típico	1.24	Error típico	1.67	Error típico	0.06
Mediana	82357	Mediana	62.20	Mediana	31	Mediana	3.60
Moda	#N/A	Moda	56.20	Moda	25.3	Moda	3.50
Desviación estándar	11001.10	Desviación estándar	8.51	Desviación estándar	11.48	Desviación estándar	0.44
Varianza de la muestra	121024239.18	Varianza de la muestra	72.41	Varianza de la muestra	131.81	Varianza de la muestra	0.20
Curtosis	-0.89	Curtosis	-0.56	Curtosis	-0.39	Curtosis	-1.03
Coeficiente de asimetría	0.29	Coeficiente de asimetría	-0.08	Coeficiente de asimetría	0.21	Coeficiente de asimetría	0.38
Rango	40980	Rango	35.60	Rango	47.4	Rango	1.6
Mínimo	66340	Mínimo	42.00	Mínimo	9.2	Mínimo	3
Máximo	107320	Máximo	77.60	Máximo	56.6	Máximo	4.6
Suma	3950157	Suma	2905.60	Suma	1437.8	Suma	175.8
Cuenta	47	Cuenta	47	Cuenta	47	Cuenta	47
Percentiles		**Percentiles**		**Percentiles**		**Percentiles**	
10%	70245	10%	50.08	10%	15.12	10%	3.2
25%	75482	25%	55.7	25%	22.8	25%	3.4
50%	82357	50%	62.2	50%	31	50%	3.6
75%	92657	75%	68	75%	38.25	75%	4.1
90%	98611	90%	70.74	90%	46.3	90%	4.42

a) (0.75 puntos) Compara la dispersión de las cuatro variables consideradas. ¿Qué medida de dispersión es más adecuado utilizar?

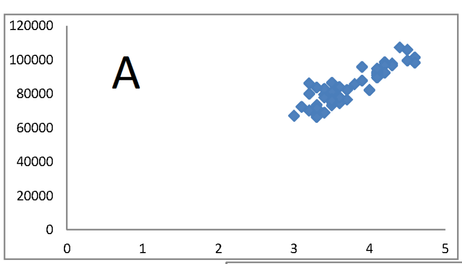
b) (0.75 puntos) Justifica si hay datos atípicos en la variable porcentaje de empleados.

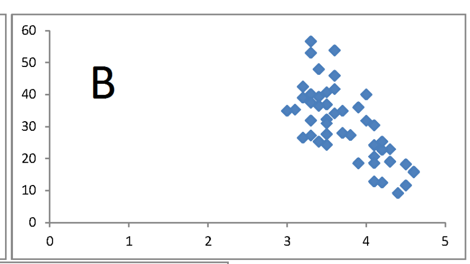
c) (1 punto) Dada la siguiente tabla:

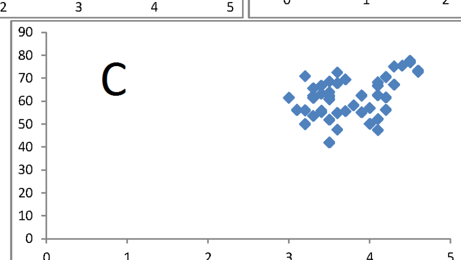
	calificación (categorizada)		
salario inicial medio	Normal	Alta	Muy Alta
$[65000, 71000)$	6	0	0
$[71000, 77100)$	8	1	0
$[77100, 83200)$	7	3	0
$[83200, 89300)$	4	2	0
$[89300, 95400)$	0	5	1
$[95400, 101500)$	0	1	7
$[101500, 107600)$	0	0	2

Calcula las frecuencias relativas condicionadas para la variable salario inicial medio cuando la variable calificación toma el valor "Muy Alta".
Calcula aproximadamente la media del salario inicial medio condicionado a que la calificación sea "Muy Alta".

d) (0.5 puntos) Los coeficientes de correlación entre la variable calificación y cada una de las otras tres variables incluidas en el estudio son -0.705, 0.887 y 0.414. Indica qué coeficiente corresponde a cada uno de los gráficos (A, B, C):







Solución.

a) (0.75 puntos) Para comparar la dispersión de las variables consideradas la medida descriptiva más adecuada es el *coeficiente de variación*. Vienen dados por:

$$c.v.(SIM) = \frac{11001.1}{84045.89} = 0.1309, \quad c.v.(empleados) = \frac{8.51}{61.82} = 0.1376$$

$$c.v.(aceptacion) = \frac{11.48}{30.59} = 0.3753, \quad c.v.(calificacion) = \frac{0.44}{3.74} = 0.1186$$

Por tanto, la variable más dispersa es *aceptación*.

b) (0.75 puntos)

$SIM \mid calificacion = Muyalta$	$[89300, 95400)$	$[95400, 101500)$	$[101500, 107600)$
f_i	0.1	0.7	0.2

Para calcular aproximadamente el salario medio condicionado a que la calificación obtenida por la Facultad sea “Muy Alta”, tenemos que obtener la marca de clase de cada una de las 3 clases anteriores, que son respectivamente: 92350, 98450 y 104550.

$$\overline{SIM} | (calificacion = Muyalta) = 92350 \cdot 0.1 + 98450 \cdot 0.7 + 104550 \cdot 0.2 = 99060 dólares$$

c) (1 punto) El rango inter-cuartílico de la variable *empleados* es $RIQ = Q_3 - Q_1 = 68 - 55,7 = 12.3$. Los límites inferior y superior son, respectivamente, $LI = Q_1 - 1.5 \cdot RIQ = 55.7 - 1.5 \cdot 12.3 = 37.25 < 42 = min$, $LS = Q_3 + 1.5 \cdot RIQ = 68 + 1.5 \cdot 12.3 = 86.45 > 77.6 = max$. Luego, no hay datos atípicos.

d) (0.5 puntos) El gráfico A se corresponde con $r_{xy} = 0.887$ y el C con $r_{xy} = 0.414$, ya que en ambos casos las variables muestran una relación directa, que es claramente más fuerte en el primer caso. El gráfico B se corresponde con $r_{xy} = -0.705$, puesto que las variables presentan una relación inversa.

2. Un examen tipo-test consta de cuatro preguntas con cuatro opciones de respuesta cada una. Cada pregunta tiene una única respuesta posible y se puntúa con 0.75 si la respuesta es correcta, -0.25 si la respuesta es incorrecta y 0 si no se contesta. Sea X el número de preguntas acertadas por un alumno que contesta todas las preguntas al azar y sea Y la nota obtenida por un alumno que contesta todas las preguntas al azar.

   a) (0.75 puntos) Calcula y representa gráficamente la función de distribución de X.

   b) (0.5 puntos) Calcula la esperanza y la varianza del número de preguntas acertadas por un alumno que contesta todas las preguntas al azar.

   c) (0.5 puntos) Determina el conjunto de posibles notas que puede obtener una persona en este examen tipo-test.

   d) (0.5 puntos) Calcula la esperanza de Y.

   e) (0.5 puntos) Para la siguiente afirmación elige una de las tres opciones y justifica la respuesta. El valor calculado en el apartado d) anterior corresponde a:

      i) La nota que obtiene un alumno en el examen si contesta todas las preguntas al azar.

      ii) Aproximadamente la nota media que obtendría un alumno sobre una serie de exámenes del mismo tipo si en todos contestara todas las preguntas al azar.

      iii) La nota por debajo de la cual se encontrarían las notas de la mitad de la clase si todos los estudiantes contestaran todas las preguntas al azar.

Solución.

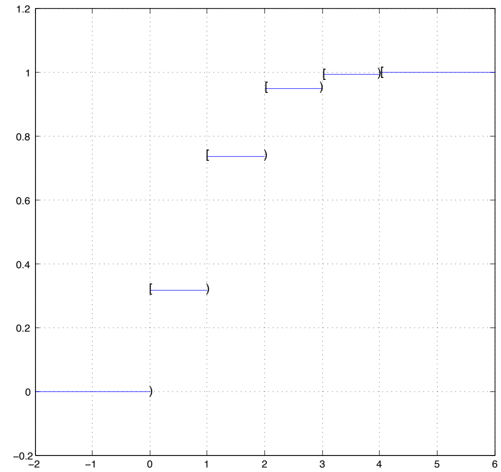
a) (0.75 puntos) X = número de preguntas acertadas por un alumno que contesta todas las preguntas al azar. $X \sim B(4, 1/4)$.
La función de probabilidad de X es:

$$P(X = x) = \binom{4}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{4-x}, \quad x = 0, 1, \ldots, 4.$$

La función de distribución de X es:

$$F_X(u) = P(X \leq x) = \sum_{x=0; x \leq u}^{4} \binom{4}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{4-x}, \quad u \in .$$

Se representa a continuación:



b) (0.5 puntos) $E[X] = 4 \cdot \frac{1}{4} = 1$. $Var[X] = 4 \cdot \frac{1}{4} \cdot \frac{3}{4} = 0.75$.

c) (0.5 puntos) $Y = 0.75 \cdot X - 0.25 \cdot (4 - X) = X - 1$, por lo tanto el soporte de Y será $\{-1, 1, 0, 2, 3\}$.

d) (0.5 puntos) $E[Y] = E[X - 1] = E[X] - 1 = 1 - 1 = 0$.

e) (0.5 puntos) II)

3. En un estudio de la red de metro se ha comprobado que, en una amplia franja horaria, tiene sentido suponer que el tiempo de espera de una persona en el andén hasta que llega el tren sigue una distribución uniforme. En la franja horaria de estudio, los trenes pasan con una frecuencia de cinco minutos, es decir, la persona esperará un mínimo de cero minutos y un máximo de cinco minutos.

a) (0.5 puntos) ¿Cuál es el tiempo medio de espera en el andén? ¿Y la varianza del tiempo de espera?

b) (1 punto) ¿Cuál es la probabilidad de que una persona espere en el andén entre de tres minutos y cuatro minutos?

c) (1 punto) La empresa demoscópica DemosCIII quiere medir el tiempo de espera de 36 pasajeros elegidos al azar en distintos momentos a lo largo de la franja horaria de estudio. La empresa sacará un informe con la media del tiempo de espera de esos 36 pasajeros. Calcula la probabilidad de que dicha media sea superior a tres minutos.

Solución:

a) (0.5 puntos) Sea X la v.a. tiempo de espera en el andén de un pasajero. Su esperanza es $(0+5)/2 = 2.5$ y su varianza es $(5-0)^2/12 = 2.0833$.

b) (1 punto)

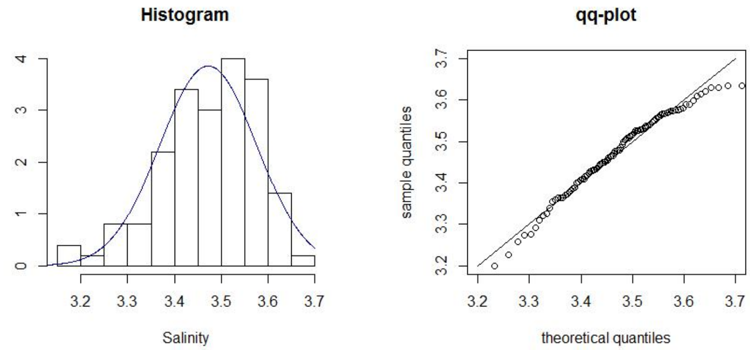
$$P(3 < X < 4) = \int_3^4 f(x)dx = \int_3^4 \frac{1}{5}dx = \frac{1}{5}x|_3^4 = \frac{4-3}{5} = 1/5.$$

c) (1 punto) Podemos considerar el tiempo de espera de los 36 pasajeros $X_1, X_2, \ldots, X_n$ como una m.a.s. de tamaño $n = 36$. Como el tamaño es suficientemente grande, la media muestral de dicha m.a.s. sigue aproximadamente una distribución normal (Teorema Central del Límite).
Su media es 2.5, que es la varianza de X.
La varianza de X es $(5-0)^2/12 = 2.0833$, por tanto la varianza de $\bar{X}$ es $2.0833/36 = 0.05787$ y su desviación típica vale 0.24. y su desviación típica es:

$$P(\bar{X} > 3) = P(\frac{\bar{X} - 2.5}{0.24} > \frac{3 - 2.5}{0.24}) = P(Z > 2.08) = 1 - 0.9812$$

4. Con el fin de medir la salinidad del Océano Pacífico se seleccionaron de forma aleatoria 100 probetas de agua y se midió su salinidad (en %).

a) (0.5 puntos) Atendiendo a los gráficos siguientes, ¿Es correcto describir la salinidad mediante una distribución Normal? Justifica la respuesta.



b) (0.75 puntos) Si para la muestra de tamaño 100 anterior se ha obtenido una media de 3.45, calcula el intervalo de confianza al 95% para la salinidad media suponiendo una varianza poblacional conocida de 0.04.

c) (0.5 puntos) El intervalo de confianza al 99% será ¿más estrecho o más ancho que el anterior? Justifica la respuesta sin calcular de nuevo el intervalo.

Solution:

a) (0.5 points) No. The histogram is skewed to the left and does not fit the normal density. The sample quantiles of the qq-plot are also not fitting the line.

b) (0.75 points) We have $z_{0.975} = 1.96$. Hence the confidence interval equals $[3.4108, 3.4892]$.

c) (0.5 points) The 99% confidence interval will be wider, since this confidence interval should include the true mean with higher probability than the 95% one.