Statistics I Final Exam, 27 May 2013.
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 table contains a statistical summary about 47 renowned Business Schools of the United States in 2004. The considered variables are: mean starting salary, expressed in dollars, percentage of new graduates employees at the end of their studies (% employees), score given by external evaluators to the universitary studies (score) and percentage of accepted applicants in the Business School (% acceptance).

Salario Inicial Medio (en $)		*% empleados al terminar*		*% aceptación en Facultad (2003)*		*Calificación Univ. (5 máx.)*	
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 ppoints) Compare the dispersion of the four considered variables. What dispersion measure do you think is better to use in this case?

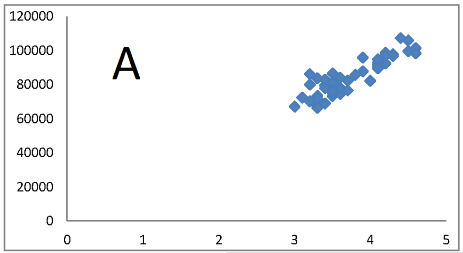
b) (0.75 points) Justify the presence/absence of outliers in the variable percentage of employees.

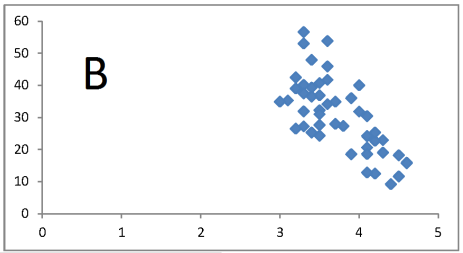
c) (1 point) From the following table:

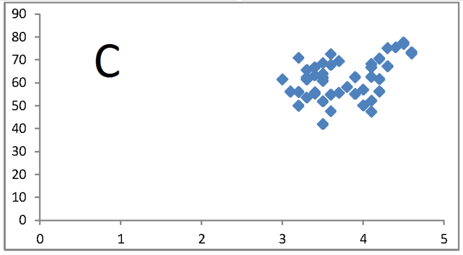
	score (categorized)		
mean starting salary	Normal	High	Very High
$[65000, 71000)$	6	0	0
$[71000, 77100)$	8	1	0
$[77100, 83200)$	7	3	0
$[83200, 89300)$	4	2	0
$[89300, 95400)$	0	5	1
$[95400, 101500)$	0	1	7
$[101500, 107600)$	0	0	2

Find the conditional relative frequencies for the variable mean starting salary given that the variable score is equal to “Very High”. Compute approximately the average of the mean starting salary given that the score is equal to “Very High”.

d) (0.5 points) The correlation coefficients between the variable score and each of the other three variables included in this study are -0.705, 0.887 y 0.414. Identify each correlation coefficient with each of the following plots (A, B, C):







Solución.

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

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

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

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

b) (0.75 points)

$SIM \vert calificacion = Muyalta$	$[89300, 95400)$	$[95400, 101500)$	$[101500, 107600)$
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 point) 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 points) 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. A multiple choice exam consists of four questions with four answer choices each. Each question has a single possible response and it is scored with 0.75 if the answer is correct, -0.25 if the answer is incorrect and 0 if it is not answered. Let X be the number of correct answers by a student who answers all the questions at random and let Y be the score obtained by a student who answers all the questions at random.

a) (0.75 points) Compute and plot the distribution function of X.

b) (0.5 points) Find the expectation and variance of the number of correct answers by a student who answers all the questions at random.

c) (0.5 points) Find the set of possible scores that a person can get in this multiple choice exam.

d) (0.5 points) Find the expectation of Y.

e) (0.5 points) Choose one of the following three choices and justify your answer. The value calculated in part d) above corresponds to:

i) The score that a student gets in the exam if he/she answers all the questions at random.

ii) Approximately the mean score that would be obtained by a student on a series of exams of the same type if in all of them he/she answered all the questions at random.

iii) The question mean score that a student gets in the exam if he/she answers all the questions at random.

Solución.

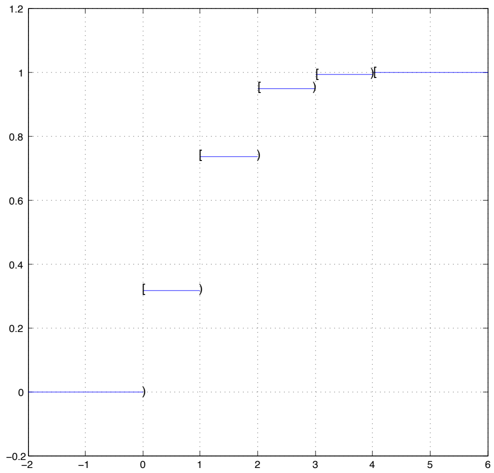
a) (0.75 points) $X =$ número de preguntas acertadas por un alumno que contesta todas las preguntas al azar. $X \sim B(4, 1/4)$.
La función de probabilidad de X es:

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

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 points) $E[X] = 4 \cdot 14 = 1$. $Var[X] = 4 \cdot 14 \cdot 34 = 0.75$.

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

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

e) (0.5 points) II)

3. In a metro network study, it has been found that, for a wide time zone, the waiting time until the train arrives can be considered uniformly distributed. In the time zone of study, the trains run with a frequency of five minutes, that is to say, a person can wait a minimum of zero minutes and a maximum of five minutes.

a) (0.5 points) Find the expected waiting time and the variance of the waiting time.

b) (1 point) Find the probability that a person waits between three and four minutes before the train arrives.

c) (1 point) The survey company DemosCIII wants to measure the waiting time of 36 randomly selected passengers at different times along the time zone of study. The company will release a report with the average waiting time for these 36 passengers. Calculate the probability that this average is greater than three minutes.

Solución:

a) (0.5 points) 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 point)

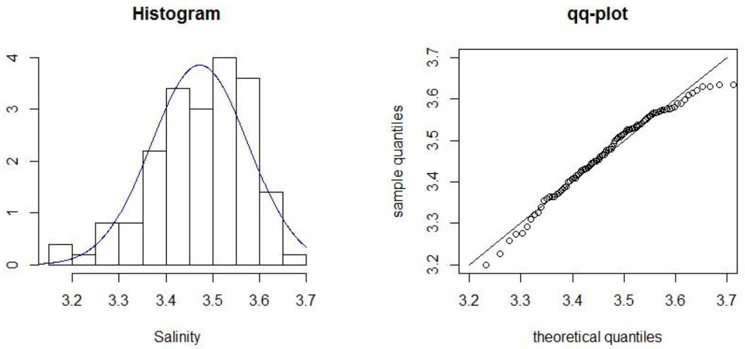
$$P(3 < X < 4) = \int_3^4 f(x)dx = \int_3^4 \frac{1}{5}dx = \frac{1}{5}x|_3^4 = \frac{4-3}{5} = 1/5.$$

c) (1 point) 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:

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

4. A simple random sample of 100 water samples was selected from the Pacific Ocean and its salinity (in %) was measured.

a) (0.5 points) According to the following plots, can the salinity be described by a Normal distribution? Justify your answer.



b) (0.75 points) The collected sample has a sample mean of 3.45. Calculate the 95% confidence interval for the mean salinity, assuming that the population variance is 0.04.

c) (0.5 points) Will the 99% confidence interval be wider or narrower than the 95% one? Justify your answer without computing the interval.

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.