Estadística I Examen extraordinario, 25 de Junio 2013.
Grados en ADE, DER-ADE, ADE-INF, FICO, ECO, ECO-DER.

REGLAS DEL EXAMEN: 1) Usar cuadernillos diferentes para cada problema. 2) Hacer los cálculos con al menos dos cifras decimales significativas. 3) No se puede salir del examen durante los primeros 30 minutos. 4) No se permite salir de la clase sin entregar el examen.

1. Se consideran las siguientes variables tomadas de 47 escuelas de negocios de renombre en los Estados Unidos en el 2004: salario inicial medio (expresado en dolares), porcentaje de nuevos empleados graduados al final de sus estudios (% empleados), puntuación dada por evaluadores externos a los estudios de la escuela (puntuación) y porcentaje de admisiones en la escuela de negocios (% aceptación).

   (a) (0.5 puntos) La siguiente tabla contiene un breve resumen estadístico de las cuatro variables:

	Salario inicial medio (en $)	% empleados	% aceptación	puntuación
Mínimo	66340	42	9.2	3
Máximo	107320	77.60	56.6	4.6
Percentil 10%	70245	50.08	15.12	3.2
Percentil 25%	75482	55.7	22.8	3.4
Percentil 50%	82357	62.2	31	3.6
Percentil 75%	92657	68	38.25	4.1
Percentil 90%	98611	70.74	46.3	4.42

   Obtener el rango intercuartílico de todas las variables.

   (b) (0.75 puntos) Dibujar diagramas de cajas para todas las variables. Basándose en estos, ¿cuáles son los rasgos más importantes de la forma de las variables? ¿Hay datos atípicos? Justificar las respuestas.

   (c) (0.75 puntos) La tabla siguiente contiene las frecuencias absolutas de las variables "salario inicial medio" en ciertas clases:

Salario inicial medio	Frecuencia absoluta
$[65000, 71000)$	6
$[71000, 77100)$	9
$[77100, 83200)$	10
$[83200, 89300)$	6
$[89300, 95400)$	6
$[95400, 101500)$	8
$[101500, 107600)$	2

   Dibujar un histograma de la variable utilizando las clases definidas en la tabla anterior. Describir las principales características del histograma.

   (d) (0.5 puntos) El conjunto de datos también contiene una variable categórica que clasifica cada escuela en tres categorías, "Normal", "Alta" and "Muy Alta" en función de la calidad general de la escuela. Seleccionamos al azar una escuela en cada categoría y observamos sus porcentajes de admisiones. Los valores obtenidos son 18.2 para "Muy Alta", 28 para "Alta" y 36.9 para "Normal". Decidir y justificar que escuela de las tres seleccionadas es mejor en relación con el porcentaje de admisiones dentro de su categoría considerando el resumen estadístico de la variable "% aceptación" de la tabla siguiente:

	Normal	Alta	Muy Alta
Media	37.67	26.93	17.3
Mediana	36.9	27.65	17.05
Desviación típica	8.87	8.21	5.31
Mínimo	24.3	12.8	9.2
Máximo	56.6	40	25.3
Rango	32.3	27.2	16.1

Solución.

(a) Los rangos intercuartílicos están dados por:

$$\begin{aligned} RIC &= Q_3 - Q_1 = 92657 - 75482 = 17175 \\ RIC &= Q_3 - Q_1 = 68 - 55.7 = 12.3 \\ RIC &= Q_3 - Q_1 = 38.25 - 22.8 = 15.45 \\ RIC &= Q_3 - Q_1 = 4.1 - 3.4 = 0.7 \end{aligned}$$

respectivamente.

(b) Los cinco puntos necesarios para dibujar el diagrama de cajas para salario medio inicial (en $) son:

$$\begin{aligned} x_{(\min)} &= 66340 \\ Q_1 &= 75482 \\ Q_2 &= 82357 \\ Q_3 &= 92657 \\ x_{(\max)} &= 107320 \end{aligned}$$

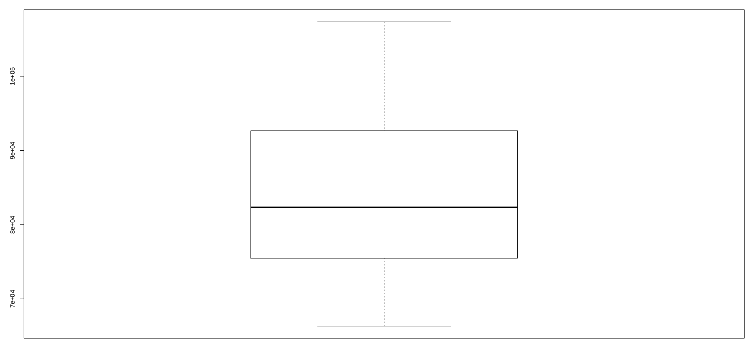


Figure 1: Diagrama de caja para salario medio inicial (en $)

La distancia entre la mediana y el primer quartil es menor que la distancia entre la mediana y el tercer cuartil. La variable tiene asimetría positiva. Más aún, como $x_{(\min)} = 66340$ y $x_{(\max)} = 107320$, entonces, no podemos decir que haya atípicos porque:

$$\begin{aligned} Q_1 - 1.5 \times RIC &= 49719.5 \\ Q_3 + 1.5 \times RIC &= 118419.5 \end{aligned}$$

Los cinco puntos necesarios para dibujar el diagrama de cajas para % empleados son:

$$\begin{aligned} x_{(\min)} &= 42 \\ Q_1 &= 55.7 \\ Q_2 &= 62.2 \\ Q_3 &= 68 \\ x_{(\max)} &= 77.6 \end{aligned}$$

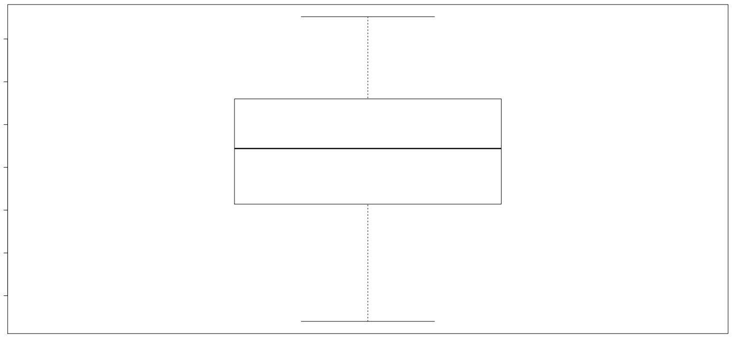


Figure 2: Diagrama de cajas para % empleados

La distancia entre la mediana y el primer cuartil es mayor que la distancia entre la mediana y el tercer cuartil. La variable tiene asimetría negativa. Más aún, como $x_{(\min)} = 42$ y $x_{(\max)} = 77.6$, entonces no hay atípicos en % empleados ya que:

$$\begin{aligned} Q_1 - 1.5 \times RIC &= 37.25 \\ Q_3 + 1.5 \times RIC &= 86.45 \end{aligned}$$

Los cinco puntos necesarios para dibujar el diagrama de cajas para % aceptación son:

$$\begin{aligned} x_{(\min)} &= 9.2 \\ Q_1 &= 22.8 \\ Q_2 &= 31 \\ Q_3 &= 38.25 \\ x_{(\max)} &= 56.6 \end{aligned}$$

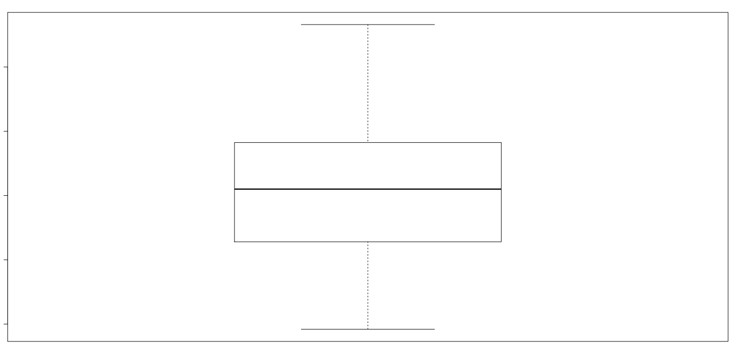


Figure 3: Diagrama de cajas para % aceptación

La distancia entre la mediana y el primer cuartil es ligeramente mayor que la distancia entre la mediana y el tercer cuartil. La variable tiene una ligera asimetría positiva. Más aún, como $x_{(\min)} = 9.2$ y $x_{(\max)} = 56.6$, entonces, no hay atípicos en % aceptación ya que:

$$\begin{aligned} Q_1 - 1.5 \times RIC &= -0.375 \\ Q_3 + 1.5 \times RIC &= 61.425 \end{aligned}$$

Los cinco puntos necesarios para dibujar el diagrama de cajas para puntuación son:

$$
\begin{aligned}
x_{(\min)} &= 3 \\
Q_1 &= 3.4 \\
Q_2 &= 3.6 \\
Q_3 &= 4.1 \\
x_{(\max)} &= 4.6
\end{aligned}
$$

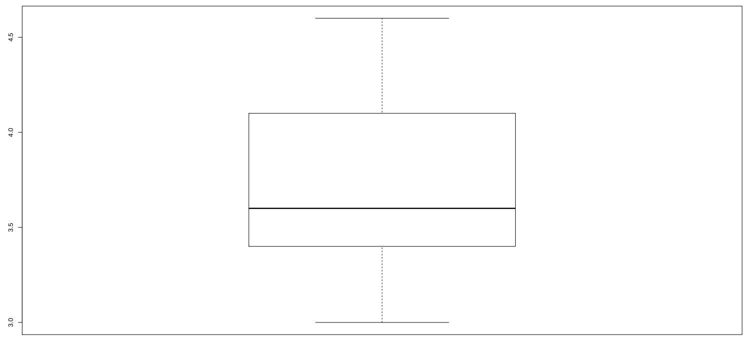


Figure 4: Diagrama de cajas para puntuación

La distancia entre la mediana y el primer cuartil es menor que la distancia entre la mediana y el tercer cuartil. La variable tiene asimetría positiva. Más aún, como $x_{(\min)} = 3$ y $x_{(\max)} = 4.6$, entonces, no hay atípicos en puntuación ya que:

$$
\begin{aligned}
Q_1 - 1.5 \times RIC &= 2.35 \\
Q_3 + 1.5 \times RIC &= 5.15
\end{aligned}
$$

(c) El histograma muestra una clara bimodalidad sugiriendo la presencia de dos poblaciones diferentes. Por un lado, "Normal" y "Alta" y por el otro "Muy Alta".

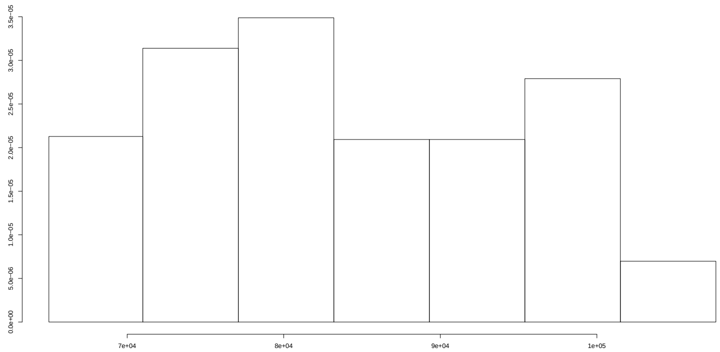


Figure 5: Histograma para salario inicial medio

(d) Primero, estandarizamos las tasas obtenidas como sigue:

$$z_{Normal} = \frac{36.9 - 37.67}{8.87} = -0.086, \; z_{Alta} = \frac{28 - 26.93}{8.21} = 0.131, \; z_{MuyAlta} = \frac{18.2 - 17.3}{5.31} = 0.169$$

El mayor valor corresponde a la universidad de la categoría "Muy Alta", con una tasa de aceptación de 0.169.

2. Se considera una variable aleatoria X con función de distribución acumulada:

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.25 & 0 \leq x < 2 \\ 0.75 & 2 \leq x < 7 \\ 1 & x \geq 7 \end{cases}$$

   (a) (0.5 puntos) Obtener la función de probabilidad de la variable aleatoria X.
   (b) (0.5 puntos) Obtener $P(0 < X < 3)$ y $P(3 \leq X < 7)$.
   (c) (0.5 puntos) Obtener la media y la varianza de la variable aleatoria X.
   (d) (0.5 puntos) Indicar VERDADERO o FALSO para la siguiente afirmación y justificar la respuesta:

$$P(X \leq 2.5) = P(X < 2.5).$$

   (e) (0.5 puntos) Sea $Y = 2X - 1$ una nueva variable. Obtener la media y la varianza de la nueva variable aleatoria.

Solución.

   (a) La función de probabilidad de la variable aleatoria X está dada por:

$$p(X) = \begin{cases} 0.25 & x = 0 \\ 0.5 & x = 2 \\ 0.25 & x = 7 \end{cases}$$

   (b) Por un lado

$$P(0 < X < 3) = P(X = 2) = 0.5$$

   Por el otro,

$$P(3 \leq X < 7) = 0$$

   (c) Por un lado,

$$E[X] = 0 \times 0.25 + 2 \times 0.5 + 7 \times 0.25 = 2.75$$

   Por el otro,

$$Var[X] = E\left[X^2\right] - E[X]^2 = 0^2 \times 0.25 + 2^2 \times 0.5 + 7^2 \times 0.25 - 2.75^2 = 14.25 - 7.5625 = 6.6875.$$

   (d) La afirmación es VERDADERA ya que X es una variable aleatoria discreta y no tiene peso en $x = 2.5$.
   (e) Por un lado:

$$E[Y] = E[2X - 1] = 2E[X] - 1 = 2 \times 2.75 - 1 = 4.5$$

   Por el otro:

$$V[Y] = V[2X - 1] = 2^2 V[X] = 4 \times 6.6875 = 26.75$$

3. Una cierta compañía distribuye teléfonos móviles que se fabrican en dos plantas de producción, A y B. Esta compañía empaqueta los teléfonos en paquetes de 5 unidades para luego distribuirlos por las tiendas.

   (a) (1 punto) Se sabe que el 2% de los teléfonos que proceden de la planta A y el 1% de los teléfonos de la planta B son defectuosos. Dado que el 20% de los teléfonos procede de la planta A y el 80% de los teléfonos procede de la planta B, obtener la probabilidad de que un teléfono cualquiera sea defectuoso.
   (b) (0.5 puntos) Sea X el número de teléfonos defectuosos en uno de los páquetes. Obtener la distribución de X e identificar los parámetros de dicha distribución.
   (c) (0.5 puntos) Una cierta tienda pide a la compañía 100 paquetes. Basándose en el apartado a), obtener la probabilidad (aproximada) de que más de 10 teléfonos móviles del total de los 100 paquetes sean defectuosos.
   (d) (0.5 puntos) Obtener la probabilidad (exacta) de que más de uno de los 500 teléfonos enviados a la tienda sean defectuosos. Aproximar la misma probabilidad mediante el Teorema Central del Límite. ¿ Es esta una buena aproximación? Justificar la respuesta.

Solución.

(a) Utilizando el Teorema de la Probabilidad Total,

$$p = 0.02 \times 0.2 + 0.01 \times 0.8 = 0.012$$

(b) Se tiene que $X \sim Bin\,(5, 0.012)$.

(c) Sea Y el número de teléfonos defectuosos en 100 paquetes. Entonces, $Y \sim Bin\,(500, 0.012)$ y

$$\Pr(Y > 10) = \Pr\left(\frac{Y-6}{2.4347} > \frac{10-6}{2.4347}\right) = \Pr(Z > 1.64) \simeq 0.05,$$

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

(d) Sea Y el númnero de teléfonos defectuosos en 100 paquetes. Entonces, $Y \sim Bin\,(500, 0.012)$ y

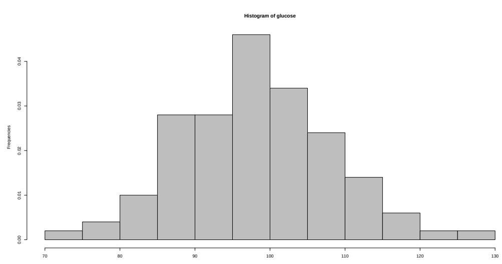
$$\Pr(Y > 1) = 1 - \Pr(Y \leq 1) = 1 - \Pr(Y = 0) - \Pr(Y = 1) =$$
$$= 1 - \binom{500}{0} 0.012^0 (1-0.012)^{500} - \binom{500}{1} 0.012^1 (1-0.012)^{499} =$$
$$= 1 - 0.0024 - 0.0169 = 0.9807.$$

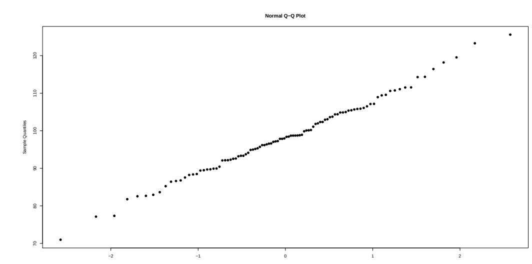
Por otro lado, esta probabilidad puede ser aproximada por el TCL como se hace en c). Entonces:

$$\Pr(Y > 1) = \Pr\left(\frac{Y-6}{2.4347} > \frac{1-6}{2.4347}\right) = \Pr(Z > -2.05) \simeq 0.9798.$$

La aproximación es buena ya que n es grande y ya que $500 \times 0.012 \geq 5$ and $500 \times 0.988 \geq 5$.

4. Una muestra aleatoria simple de 100 personas se selecciona y se mide su glucosa en sangre (en mg/dL).





(a) (0.5 puntos) ¿Puede describirse la glucosa mediante una distribución Normal? Justificar la respuesta.

(b) (1 punto) La muestra tiene media muestral 98.36 y varianza de 100.288. Calcular el intervalo de confianza al 95% para el valor medio de la glucosa en sangre.

(c) (0.5 puntos) Razonar sin hacer cálculos si la longitud del intervalo de confianza al 90% será menor, igual o mayor que el obtenido al nivel del 95%.

(d) (0.5 puntos) Según los expertos la glucosa en sangre tiene valor medio de 98 y desviación típica de 10. Obtener la probabilidad de que la suma de la glucosa en sangre de las 100 personas sea mayor de 9900 mg/dL.

Solución:

(a) Si. El histograma es aproximadamente simétrico y puede ajustar bien la densidad normal. Los cuantiles muestral del qq-plot se ajustan a la recta bastante bien.

(b) Tenemos $z_{.975} = 1.96$. Por lo tanto el intervalo de confianza es $(96.39, 100.32)$.

(c) El intervalo de confianza al 90% será más estrecho dado que el nivel de confianza también lo es.

(d) Si X_i es el nivel de glucosa en sangre para la persona i, entonces, tenemos que:

$$\Pr\left(\sum_{i=1}^{100} X_i > 9900\right) = \Pr\left(\overline{X} > 99\right) = \Pr(Z > 1) \simeq .8413$$

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