Examen Final de Estadística I, 27 de Mayo de 2015.
Grados en ADE, DER-ADE, ADE-INF, FICO, ECO, ECO-DER, TUR.

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. Se quiere determinar el impacto económico de los negocios de una gran compañia en las ventas totales de pequeñas empresas. Para ello se obtiene una muestra de 15 de esas pequeñas empresas y se toman los porcentajes del total de ventas anual de las 15 pequeñas empresas como resultado de las ventas a la compañía:

$$27 \quad 12 \quad 14.9 \quad 1.2 \quad 0.1 \quad 1 \quad 0.1 \quad 5.3 \quad 7.6 \quad 5 \quad 1 \quad 1 \quad 3.2 \quad 3 \quad 7$$

   (a) (0.5 puntos) Es la media muestral de los 15 porcentajes mayor que la mediana muestral? En caso afirmativo, ¿qué sugiere este resultado? Justifica tus respuestas.

   (b) (0.5 puntos) Calcular los tres cuartiles muestrales. Interpretarlos en términos de los porcentajes.

   (c) (0.5 puntos) Calcular la cuasi-varianza y el coeficiente de variación de la muestra de 15 porcentajes.

   (d) (1 punto) Dibujar un diagrama de caja de los datos e identificar los atípicos (si hubiere). Justifica tu respuesta.

   Solución.

   (a) La media muestral de los 15 porcentajes es $\overline{x} = 5.96$, mientras que la mediana muestral es $M = 3.2$. Entonces, la media muestral es mayor que la mediana muestral. Este resultado sugiere que la distribución de los datos es asimétrica a la derecha (asimetría positiva), es decir, hay un número reducido de pequeñas compañías para las que los porcentajes de ventas totales anuales a la gran compañía son notablemente mayores que para el resto.

   (b) Los cuartiles muestrales son $Q_1 = x_{(4)} = 1$, $Q_2 = x_{(8)} = 3.2$ y $Q_3 = x_{(12)} = 7.6$, respectivamente. Entonces, el 25% de los porcentajes son menores que el 1%, el 50% de los porcentajes son menores que el 3.2% y el 75% de los porcentajes son menores que el 7.6%. Consecuentemente, los tres cuartiles muestrales dividen la muestra en cuatro sub-muestras que contienen respectivamente el mismo número de porcentajes. En general, los porcentajes de ventas totales a la compañía para la mayoría de las pequeñas empresas representa menos del 7.6%.

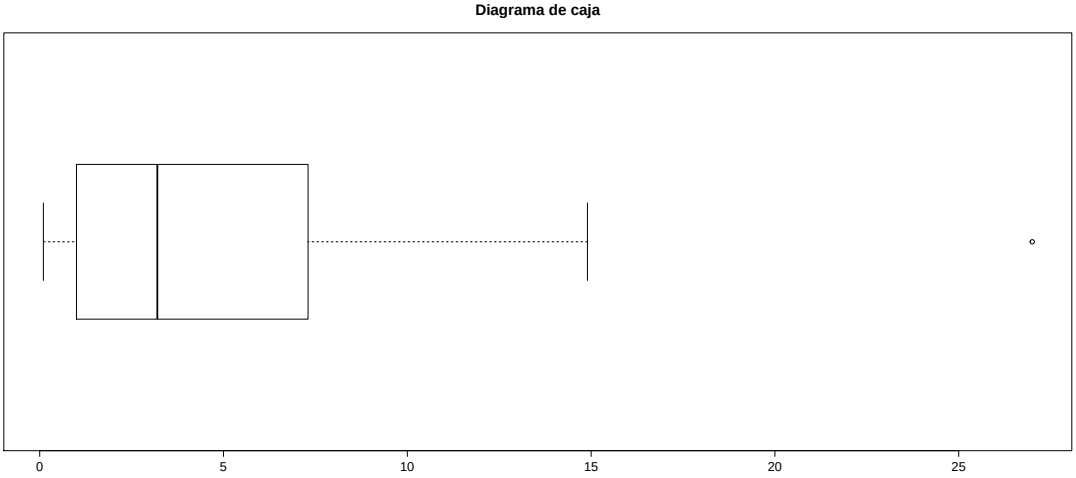
   (c) La cuasi-varianza muestral es $\widehat{\sigma}^2 = 53.2668$, mientras que el coeficiente de variación muestral es $CV = 1.2245$.

   (d) Para construir el diagrama de caja, necesitamos el rango intercuartílico que está dado por $IQR = Q_3 - Q_1 = 7.6 - 1 = 6.6$. Más aún, para construir las barras del diagrama y para detectar atípicos, si los hay, necesitamos los valores $Q_1 - 1.5IQR = 1 - 1.5 \times 6.6 = -8.9$ y $Q_3 + 1.5IQR = 7.6 + 1.5 \times 6.6 = 17.5$. Además, los valores máximo y mínimo en la muestra son 0.1 y 27, respectivamente. Entonces, hay un sólo dato atípico ya que $17.5 < 27$. El diagrama de caja aparece en la Figura 1.

2. Se realiza una encuesta a 500 personas estadounidenses acerca de su afiliación política y de su actitud respecto a un plan nacional de salud. Las respuestas se disponen en una tabla de contingencias de acuerdo a la afiliación política y la actitud como sigue:

	Actitud			
Afiliación	Favor	Indiferente	Contrario	Total
Demócrata	138	83	64	285
Republicano	64	67	84	215
Total	202	150	148	500

   Dada la información previa, responder a las siguientes preguntas:



(a) (1 punto) Obtener la distribución de frecuencias relativas conjunta de las dos variables y las distribuciones de frecuencias relativas marginales.

(b) (0.5 puntos) Obtener la distribución condicional de Actitud dada Afiliación. ¿Qué actitud es más frecuente entre Republicanos? Justifica tus respuestas.

(c) (0.5 puntos) Obtener la distribución condicional de Afiliación dada Actitud. ¿Qué afiliación política es más favorable al plan nacional de salud? Justifica tus respuestas.

(d) (0.5 puntos) ¿Sería correcto decir que entre la gente que no se opone al plan de salud propuesto, la frecuencia relativa de votantes Demócratas es mayor que la frecuencia relativa de votantes Republicanos? Justifica tu respuesta.

Solución.

(a) La distribución de frecuencias relativa conjunta de las dos variables y las distribuciones de frecuencias relativas marginales están dadas en la tabla siguiente:

	Actitud			
Afiliación	Favor	Indiferente	Contrario	Total
Demócrata	0.276	0.166	0.128	0.57
Republicano	0.128	0.134	0.168	0.43
Total	0.404	0.3	0.296	1

(b) La distribución condicional de Actitud dada Afiliación está dada por:

Actitud|Afiliación	Favor	Indiferente	Contrario
$f_{\cdot\|Democrata}$	0.4842105	0.2912281	0.2245614
$f_{\cdot\|Republicano}$	0.2976744	0.3116279	0.3906977

Por lo tanto, la Actitud más frecuente entre Republicanos es Contrario.

(c) La distribución condicional de Afiliación dada Actitud está dada por:

Afiliación|Actitud	$f_{\cdot\|Favor}$	$f_{\cdot\|Indiferente}$	$f_{\cdot\|Contrario}$
Demócrata	0.6831683	0.5533333	0.4324324
Republicano	0.3168317	0.4466667	0.5675676

Por lo tanto, la Afiliación política con mayor Actitud favorable al plan de salud es Demócrata.

(d) La respuesta es Si. Notar que,

Afiliación|Actitud	$f_{\cdot\|Nocontraria}$	$f_{\cdot\|Contrario}$
Demócrata	0.6278409	0.4324324
Republicano	0.3721591	0.5675676

Por lo tanto, entre la gente que no se opone al plan de salud propuesto, la frecuencia relativa de votantes Demócratas es mayor que la frecuencia relativa de votantes Republicanos.

3. El número de llamadas nocturnas a un servicio de emergencia de una empresa de calefacción y aire acondicionado tiene probabilidades 0.05, 0.1, 0.15, 0.35, 0.2 y 0.15 para 0, 1, 2, 3, 4 y 5 llamadas por noche, respectivamente.

   (a) (0.5 puntos) Obtener la función de probabilidad del número de llamadas al servicio en una noche. ¿Cuál es la probabilidad de que el número de llamadas en una noche sea mayor de 3?

   (b) (0.5 puntos) ¿Cuál es el número medio de llamadas por noche? ¿Y la varianza?

   (c) (0.5 puntos) Suponiendo que el número de llamadas en noches diferentes sean independientes, ¿cuál es el número medio de llamadas por semana? ¿y la varianza?

   (d) (1 punto) Obtener la probabilidad (o una aproximación) de que el número de llamadas en un año (365 días) sea mayor que 1300.

Solución.

(a) La función de probabilidad del número de llamadas de emergencia, X, está dada por:

$$\Pr(X = x) = \begin{cases} 0.05 & x = 0 \\ 0.1 & x = 1 \\ 0.15 & x = 2 \\ 0.35 & x = 3 \\ 0.2 & x = 4 \\ 0.15 & x = 5 \end{cases}$$

Por lo tanto, $\Pr(X > 3) = \Pr(X = 4) + \Pr(X = 5) = 0.2 + 0.15 = 0.35$.

(b) La media está dada por:

$$E[X] = 0 \times 0.05 + 1 \times 0.1 + 2 \times 0.15 + 3 \times 0.35 + 4 \times 0.2 + 5 \times 0.15 = 3$$

mientras que la varianza está dada por:

$$\begin{aligned} V[X] &= (0-3)^2 \times 0.05 + (1-3)^2 \times 0.1 + (2-3)^2 \times 0.15 + \\ &+ (3-3)^2 \times 0.35 + (4-3)^2 \times 0.2 + (5-3)^2 \times 0.15 = \\ &= 9 \times 0.05 + 4 \times 0.1 + 1 \times 0.15 + 0 \times 0.35 + 1 \times 0.2 + 4 \times 0.15 = 1.8 \end{aligned}$$

(c) La media y la varianza del número de llamadas de emergencia por semana, Y, coinciden con la media de $7X$ y con siete veces la varianza de X. Entonces,

$$E[Y] = 7 \times E[X] = 7 \times 3 = 21$$

mientras que

$$V[Y] = 7 \times V[X] = 7 \times 1.8 = 12.6$$

(d) La media y la varianza del número de llamadas de emergencia por año, M, coinciden con la media de $365X$ y con 365 veces la varianza de X. Entonces,

$$E[M] = 365 \times E[X] = 365 \times 3 = 1095$$

mientras que

$$V[M] = 365 \times V[X] = 365 \times 1.8 = 657.$$

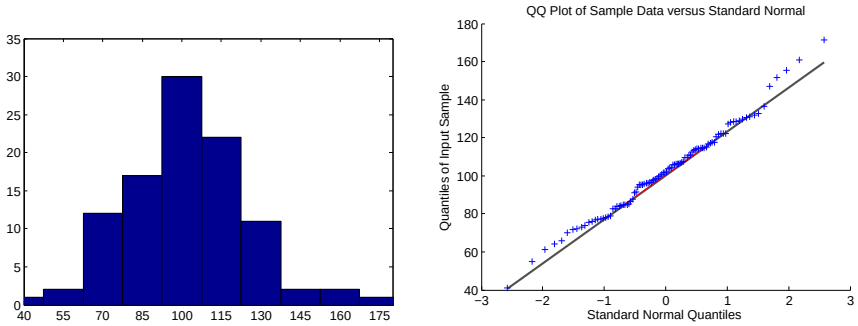
Por el TCL, tenemos que:

$$\Pr(M > 1300) = \Pr\left(\frac{M - 1095}{25.63201} > \frac{1300 - 1095}{25.63201}\right) = \Pr\left(\frac{M - 1095}{25.63201} > 7.997812\right) \simeq$$
$$\simeq \Pr(Z > 7.997812) = 1 - \Pr(Z \leq 7.997812) = 1 - 1 = 0,$$

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

4. El número de descargas por hora de cierta aplicación para móvil puede modelarse según una variable aleatoria de media 100 y desviación típica 20.

   (a) (0.5 puntos) Dada una muestra de tamaño 100 se obtienen los siguientes gráficos para el número de descargas por hora:



   Razonar si el número de descargas por hora puede modelarse según una ley Normal.

   (b) (0.75 puntos) Calcular la probabilidad (sin aproximar) de que en un día (24 horas) la media de descargas por hora esté comprendida entre 90 y 110, suponiendo que el número de descargas en cada hora es una muestra aleatoria simple. ¿Qué otra hipótesis es necesario asumir?

   (c) (0.75 puntos) Calcular la probabilidad aproximada de que en una semana (168 horas) el total de descargas llegue a superar las 17000, suponiendo que el número de descargas en cada hora es una muestra aleatoria simple. ¿Qué teorema es necesario aplicar para poder calcular esta probabilidad?

   (d) (0.5 puntos) Se han controlado las descargas de la aplicación durante 40 horas obteniéndose una media muestral de 99.5. Calcular el intervalo de confianza al 95% para la media de descargas por hora (suponiendo $\sigma = 20$ y que el número de descargas en cada hora es una muestra aleatoria simple).

Solución:

(a) Atendiendo al histograma y al qq-plot parece que los datos no se alejan demasiado de la ley Normal.

(b) Si se supone que la v.a. $X =$"número de descargas por hora" sigue una ley $N(100, 20^2)$, entonces dadas $n = 24$ v.a. independientes y con la misma ley que X, la media muestral $\bar{X}_n \sim N(100, 20^2/24)$. Por tanto,

$$P(90 < \bar{X}_n < 110) = P(-2.45 < Z < 2.45) = 2\,P(Z < 2.45) - 1 = 2 \cdot 0.9929 - 1 = 0.9858,$$

donde $Z \sim N(0,1)$. Para poder calcular esta probabilidad es necesario asumir que X tiene ley Normal.

(c) Dadas $X_1, \ldots, X_n$, $n = 168$ v.a. independientes y con la misma ley que $X =$"número de descargas por hora", consideramos la v.a. $\sum_{i=1}^n X_i = n\,\bar{X}_n$. Puesto que $n \geq 30$, el Teorema Central del Límite asegura que $\bar{X}_n \sim N(100, 20^2/168)$. Entonces,

$$P(\sum_{i=1}^{n} X_i > 17000) = P(n\,\bar{X}_n > 17000) = P(\bar{X}_n > 17000/168) \approx P(Z > 0.77) = 0.2206,$$

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

(d) A partir de las descargas observadas en 40 horas ($X_1, \ldots, X_n$ v.a. i.i.d. con $n = 40$) se obtuvo una media muestral de $\bar{x}_n = 99.5$. Suponiendo $\sigma = 20$, el intervalo de confianza al 95% para la media de descargas es

$$\bar{x}_n \mp z_{1-\alpha/2}\,\sigma/\sqrt{n} = 99.5 \mp 1.96 \cdot 20/\sqrt{40} = 99.5 \mp 6.20 = (93.3, 105.7).$$