Laboratorio 2 (Respuestas) - Inferencia Estadística

Estadística Descriptiva

Laboratorista: Héctor Lira Talancón

Ago-Dic 2017

1. La siguiente información corresponde a una muestra de 13 autos usados. Se registró el valor de las variables marca, años de uso y precio (x10,000) del auto.

Marca	B	C	C	B	A	B	A	D	B	A	A	C	C
Años uso (X)	11	12	8	7	12	8	11	10	8	9	6	7	6
Precio (Y)	3.6	4.0	5.0	6.0	6.0	7.6	8.0	8.0	8.0	8.1	8.6	10.0	12.0

Datos: $\sum_{i=1}^{13} x_i y_i = 800.9$, $\sum_{i=1}^{13} x_i = 115$, $\sum_{i=1}^{13} y_i = 94.9$, $\sum_{i=1}^{13} x_i^2 = 1073$, $\sum_{i=1}^{13} y_i^2 = 759.29$

a) Con los trece datos de Años de uso, llene la siguiente tabla de frecuencias:

Años de uso	f	f% (p_i)	F	F%
$[5, 7)$	2	$\frac{2}{13}$	2	$\frac{2}{13}$
$[7, 9)$	5	$\frac{5}{13}$	7	$\frac{7}{13}$
$[9, 11)$	2	$\frac{2}{13}$	9	$\frac{9}{13}$
$[11, 13)$	4	$\frac{4}{13}$	13	1

b) Usando exclusivamente la tabla anterior calcule la moda, mediana y media de los años de uso.

Definimos las marcas de clase, c_i, como $\frac{a_i+b_i}{2}, i = 1, ..., 4 \Rightarrow \{c_1, c_2, c_3, c_4\} = \{6, 8, 10, 12\}$

Media $= \bar{x} = \sum_{i=1}^{4} p_i c_i = \frac{2}{13} \cdot 6 + \frac{5}{13} \cdot 8 + \frac{2}{13} \cdot 10 + \frac{4}{13} \cdot 12 = 9.23$

Definimos k^* como la clase donde se acumula por primera vez más del 50% de las observaciones $\Rightarrow k^* = 2$, i.e., $c_{k^*} = [7, 9)$.

Mediana $= \tilde{x} = a_{k^*} + \frac{b_{k^*} - a_{k^*}}{p_{k^*}}(0.5 - p_{k^*-1}) = 7 + \frac{9-7}{5/13}(0.5 - \frac{2}{13}) = 8.8$

Moda $= c_{k'}$, donde $k' =$ es tal que $p_{k'} = max\{p_1, ..., p_4\} \Rightarrow$ moda $= 8$

c) Considere las variables marca, años de uso y precio. Para cada variable diga cuál es su escala de medición y si se trata de una variable cualitativa o cuantitativa.

Variable	escala	cual./cuant.
Marca	nominal	cual.
Años de uso	razón	cuant.
Precio	razón	cuant.

d) Construya un diagrama de caja y brazos con los datos de la variable precio e interprete el diagrama.

e) Calcule el grado de asociación lineal entre las variables años de uso (X) y precio (Y), interprete el resultado.

El grado de asociación lineal entre X e Y se denota como ρ_{XY}.

$\rho_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{1}{n-1}(\sum_{i=1}^{13} x_i y_i - n\bar{x}\bar{y})}{\sqrt{\frac{1}{n-1}(\sum_{i=1}^{13} x_i^2 - n\bar{x}^2)}\sqrt{\frac{1}{n-1}(\sum_{i=1}^{13} y_i^2 - n\bar{y}^2)}} = \frac{\frac{1}{n-1}}{\sqrt{\frac{1}{n-1}}\sqrt{\frac{1}{n-1}}} \frac{800.9 - 13 \cdot \frac{94.9}{13} \frac{115}{13}}{\sqrt{(1073 - 13 \cdot (\frac{115}{13})^2)}\sqrt{(759.29 - 13 \cdot (\frac{94.9}{13})^2)}} =$

$\frac{-38.6}{60.865} = -0.634$

2. Suponga que $x_1, ..., x_n$ son datos de una muestra de tamaño n y que $y_1, ..., y_n$ y $z_1, ..., z_n$ son transformaciones definidas por $y_i = \alpha x_i$ y $z_i = \beta x_i, i = 1, ...n$ con $\alpha, \beta \in \mathbb{R}$. Conteste las siguientes preguntas:

a) Calcule las medias de los datos transformados $\bar{y}$ y $\bar{z}$.

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \alpha x_i = \alpha \frac{1}{n}\sum_{i=1}^{n} x_i = \alpha \bar{x}$

$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \beta x_i = \beta \frac{1}{n}\sum_{i=1}^{n} x_i = \beta \bar{x}$

b) Calcule las varianzas de los datos transformados s_Y^2 y s_Z^2.

$s_Y^2 = \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2) = \frac{1}{n-1}(\sum_{i=1}^{n} (\alpha x_i)^2 - n(\alpha\bar{x})^2) = \frac{1}{n-1}(\sum_{i=1}^{n} \alpha^2 x_i^2 - \alpha^2 n\bar{x}) =$
$\alpha^2 \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = \alpha^2 s_X^2$

$s_Z^2 = \frac{1}{n-1}(\sum_{i=1}^{n} z_i^2 - n\bar{z}^2) = \frac{1}{n-1}(\sum_{i=1}^{n} (\beta x_i)^2 - n(\beta\bar{x})^2) = \frac{1}{n-1}(\sum_{i=1}^{n} \beta^2 x_i^2 - \beta^2 n\bar{x}) =$
$\beta^2 \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = \beta^2 s_X^2$

c) Calcule la covarianza de los datos transformados s_{YZ}^2.

$s_{YZ} = \frac{1}{n-1}(\sum_{i=1}^{n} y_i z_i - n\bar{y}\bar{z}) = \frac{1}{n-1}(\sum_{i=1}^{n} (\alpha x_i)(\beta x_i) - n(\alpha\bar{x})(\beta\bar{x})) = \frac{1}{n-1}(\sum_{i=1}^{n} \alpha\beta x_i^2 - \alpha\beta n\bar{x}^2) =$
$\alpha\beta \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = \alpha\beta s_X^2$

d) Calcule el coeficiente de correlación muestral ρ_{YZ}.

$\rho_{YZ} = \frac{s_{YZ}}{s_Y s_Z} = \frac{\alpha\beta s_X^2}{\sqrt{\alpha^2 s_X^2}\sqrt{\beta^2 s_X^2}} = \frac{\alpha\beta s_X^2}{\alpha s_X \beta s_X} = \frac{\alpha\beta s_X^2}{\alpha\beta s_X^2} = 1$

3. Marque con una 'x' las afirmaciones que sean falsas.

☐ Si la desviación media con respecto a la mediana es cero, todos los datos tienen el mismo valor.

Sabemos que $DM(\tilde{x}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}|$. Dado que es una suma de elementos $d_i \geq 0 \Rightarrow DM(\tilde{x}) = 0 \Leftrightarrow d_i = 0, \forall i \Rightarrow x_i = \tilde{x}, \forall i$. VERDADERO.

☐ En un diagrama de caja y brazos se dispone de un criterio para saber la dispersión de los datos.

El diagrama de caja y brazos muestra de forma gráfica conceptos como la amplitud y la amplitud intercuartílica. VERDADERO.

☐ La distribución de frecuencias es una tabla que organiza los datos en clases. Estas clases pueden o no, ser exhaustivas y mutuamente excluyentes.

Dos condiciones necesarias de las clases es que sean exhaustivas y mutuamente excluyentes. FALSO.

☐ La mediana de un conjunto de datos no necesariamente es igual a uno de los valores observados en el conjunto.

Si se tiene un número par de observaciones, la mediana se calcula como el promedio entre la observación i e $i+1$. VERDADERO.

4. Considere la muestra compuesta por 20 mediciones de tiempos de entrega en un restaurante de comida rápida medidos en minutos:

4.14	4.20	4.35	4.58	5.26	5.29	5.41	5.55	5.75	6.19
6.37	6.56	6.71	6.81	9.32	9.32	12.58	13.61	15.27	22.64

Datos: $\sum_{i=1}^{20} x_i = 159.91, \sum_{i=1}^{20} x_i^2 = 1,699.746, m_3 = 184.3635, m_4 = 2695.727$

a) Construya un histograma con las clases [4, 5], (5, 6], (6, 12], (12, 20], (20, 25]. ¿Cómo se calcula el porcentaje de observaciones que están por arriba de 20 a partir de este histograma?

Definimos:
H = número de clases
l_B = longitud más común,
l_k = longitud del intervalo k,
Factor de ajuste = $\text{FA}_k = \frac{l_B}{l_k}$,
Frecuencia relativa ajustada = $p_k^A = \text{FA}_k p_k$
Factor de corrección = $FC = \sum_{i=1}^{H} p_i^A$
Frecuencia relativa corregida = $p_k^C = \frac{p_k^A}{FC}$

Clase	Intervalo	longitud	f	f% (p_i)	Factor de ajuste (FA)	Frec. rel. ajustada	Frec. rel. corregida
1	$[4,5)$	1	4	$\frac{4}{20}$	$\frac{1}{1}$	$1 \cdot (\frac{4}{20})$	$\frac{\frac{1}{5}}{FC}$
2	$[5,6)$	1	5	$\frac{5}{20}$	$\frac{1}{1}$	$1 \cdot (\frac{5}{20})$	$\frac{\frac{1}{4}}{FC}$
3	$[6,12)$	6	7	$\frac{7}{20}$	$\frac{1}{6}$	$(\frac{1}{6}) \cdot (\frac{7}{20})$	$\frac{\frac{7}{120}}{FC}$
4	$[12,20)$	8	3	$\frac{3}{20}$	$\frac{1}{8}$	$(\frac{1}{8}) \cdot (\frac{3}{20})$	$\frac{\frac{3}{160}}{FC}$
5	$[20,25)$	5	1	$\frac{1}{20}$	$\frac{1}{5}$	$(\frac{1}{5}) \cdot (\frac{1}{20})$	$\frac{\frac{1}{100}}{FC}$

$FC = \frac{1}{5} + \frac{1}{4} + \frac{7}{120} + \frac{3}{160} + \frac{1}{100}$

El histograma se construye considerando las alturas dadas por la Frecuencia Relativa Corregida.

b) [Tendencia central] Calcule la media, mediana, moda, media podada y la media de Windsord, considerando $\alpha = 0.20$. (La media podada, $\bar{x}_\alpha$, se calcula eliminando el $100\alpha\%$ superior e inferior, mientras que la media de Windsord, $\bar{x}_{W_\alpha}$, se calcula sustituyendo las observaciones que se eliminan en la media podada por los valores menor y mayor de las observaciones consideradas). ¿Cómo se comparan entre ellas? ¿Qué se puede concluir acerca de la distribución?

Media = $\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = \frac{159.91}{20} = 7.9955$

Mediana = $\tilde{x} = \frac{x_{(\lfloor l \rfloor)} + x_{(\lfloor l \rfloor + 1)}}{2}$, donde $l = \alpha(n-1) + 1 = 0.5(20-1) + 1 = 10.5$

$\Rightarrow \tilde{x} = \frac{x_{(10)}+x_{(11)}}{2} = \frac{6.19+6.37}{2} = 6.28$

Moda $= mo = 9.32$

Media podada: del conjunto de observaciones quitamos el $100(0.20)\% = 20\%$ inferior y superior, i.e., las 4 observaciones inferiores y superiores.
$\Rightarrow$ Media podada $= \bar{x}_\alpha = \frac{1}{(20-8)} \sum_{i=5}^{16} x_i = \frac{78.54}{12} = 6.545$

Media de Windsord: en este caso, en lugar de eliminar las 4 observaciones inferiores y superiores, sustituimos los 4 valores inferiores por el menor valor considerado y los 4 valores superiores por el mayor valor considerado. Nuestro nuevo conjunto de datos es el siguiente:

5.26	5.26	5.26	5.26	5.26	5.29	5.41	5.55	5.75	6.19
6.37	6.56	6.71	6.81	9.32	9.32	9.32	9.32	9.32	9.32

$\Rightarrow$ Media de Windsord $= \bar{x}_{W_\alpha} = \frac{1}{20} \sum_{i=1}^{20} x_i^* = \frac{136.86}{20} = 6.843$

Comparamos las medidas de tendencia central:
$\tilde{x} < \bar{x}_\alpha < \bar{x}_{W_\alpha} < \bar{x} < mo$

Podemos concluir que la distribución tiene sesgo a la derecha pues $\tilde{x} < \bar{x}$.

c) [Localización] Calcule el resumen de los 5 números: $(x_{min}, q_1, q_2, q_3, x_{max})$

Solo nos falta calcular q_1 y q_3 pues $q_2 = \tilde{x}$:

Sabemos que $p_\alpha = fracc(l)x_{(\lfloor l \rfloor+1)} + [1 - fracc(l)]x_{(\lfloor l \rfloor)}$ donde $l = \alpha(n-1)+1$ y que $q_1 = p_{0.25}$ y $q_3 = p_{0.75}$. Por lo tanto, $l(0.25) = 0.25(20-1)+1 = 5.75$ y $l(0.75) = 0.75(20-1)+1 = 15.25$.

$\Rightarrow q_1 = (0.75)x_{(6)} + (0.25)x_{(5)} = (0.75)5.29 + (0.25)5.26 = 5.2825$, y

$\Rightarrow q_3 = (0.25)x_{(16)} + (0.75)x_{(15)} = (0.75)9.32 + (0.25)9.32 = 9.32$

Por lo tanto, $(x_{min}, q_1, q_2, q_3, x_{max}) = (4.14, , 5.28, 6.28, 9.32, 22.64)$

d) [Dispersión] Calcule la amplitud, amplitud intercuartílica, varianza muestral, desviación estándar muestral y el coeficiente de variación muestral.

Amplitud $= x_{max} - x_{min} = 22.64 - 4.14 = 18.5$

Amplitud intercuartílica $= q_3 - q_1 = 9.32 - 5.28 = 4.0375$

Varianza muestral $= s^2 = \frac{1}{20-1}(\sum_{i=1}^{20} x_i^2 - n\bar{x}^2) = \frac{1}{20-1}(1699.746 - 20 \cdot 7.9955^2) = 22.16766$

Desviación estándar muestral $= \sqrt{s^2} = \sqrt{22.16766} = 4.708254$

Coeficiente de variación muestral $= CV = \frac{s}{\bar{x}} = \frac{4.708254}{7.9955} = 0.588863$

e) [Forma] Calcule el coeficiente de asimetría muestral y el coeficiente de curtosis muestral. Mencione cómo es la simetría de la distribución, lo apuntada de la distribución y cómo son sus colas. ¿Es consistente el resultado de la simetría con el obtenido en el inciso b)?

Coeficiente de asimetría muestral $= \frac{m_3}{s^3} = \frac{184.36}{4.71^3} = 1.766425$

Coeficiente de curtosis muestral $= \frac{m_4}{s^4} - 3 = \frac{2695.727}{4.71^3} - 3 = 2.485755$

El coeficiente de asimetría nos dice que la distribución tiene sesgo a la derecha pues es > 0. El resultado es consistente con el obtenido en el inciso b).

El coeficiente de curtosis nos dice que la distribución es leptocúrtica. Esto quiere decir que la distribución presenta un pico pronunciado. Las colas son más anchas que las de una distribución normal.

5. Ahora considere los datos agrupados obtenidos en la pregunta 4. inciso a). Calcule media muestral, mediana, desviación media respecto a la mediana, varianza muestral y moda muestral. ¿La moda de los datos individuales coincide con la clase modal?

Media muestral $= \sum_{i=1}^{k} p_i c_i =$, considerando p_i la frecuencia relativa de la clase i, $k =$ número de clases y $c_i =$ las marcas de clase calculadas como $\frac{a_i+b_i}{2}$, con a_i, b_i los límites inferior y superior, respectivamente.
$\Rightarrow \bar{x} = 4.5 \cdot \frac{4}{20} + 5.5 \cdot \frac{5}{20} + 9 \cdot \frac{7}{20} + 16 \cdot \frac{3}{20} + 22.5 \cdot \frac{1}{20} = 8.95$

Mediana muestral $= a_{k^*} + \frac{b_{k^*} - a_{k^*}}{P_{k^*}}(0.5 - P_{k^*-1})$ donde k^* es la clase donde se acumula por primera vez más o igual al 50% de las observaciones y P_i la frecuencia relativa acumulada. En este caso, $k^* = 3$.
$\Rightarrow \tilde{x} = 6 + \frac{12-6}{16/20}(0.5 - \frac{9}{20}) = 6.375$

Desviación media respecto a la mediana muestral $= DM(\tilde{x}) = \sum_{i=1}^{k} p_i |c_i - \tilde{x}| = \frac{4}{20}|4.5 - 6.375| + \frac{5}{20}|5.5 - 6.375| + \frac{7}{20}|9 - 6.375| + \frac{3}{20}|16 - 6.375| + \frac{1}{20}|22.5 - 6.375| = 3.7625$

Varianza muestral $= s^2 = \frac{1}{n-1} \sum_{i=1}^{k} f_i (c_i - \bar{x})^2 = \frac{1}{19}(4 \cdot (4.5 - 8.95)^2 + 5 \cdot (5.5 - 8.95)^2 + 7 \cdot (9 - 8.95)^2 + 3 \cdot (16 - 8.95)^2 + 1 \cdot (22.5 - 8.95)^2) = 24.81316$

Moda muestral = marca de clase donde p_k es máximo
$\Rightarrow mo = 9$.

6. Ahora imagine que se le proporciona más información acerca de las entregas del restaurante de comida rápida. Información adicional: número de personas realizando el pedido, tipo de promoción que pidieron las personas y propina (como %).

tiempo (Y)	4.14	4.20	4.35	4.58	5.26	5.29	5.41	5.55	5.75	6.19
personas (X)	1	1	1	1	1	1	2	2	2	2
promoción (Z)	A	A	A	B	C	A	A	B	B	B
propina (W)	10.0	12.5	11.0	12.0	15	15	14.5	13.2	12.0	13.5

tiempo (Y)	6.37	6.56	6.71	6.81	9.32	9.32	12.58	13.61	15.27	22.64
personas (X)	2	2	2	2	3	3	3	3	3	3
promoción (Z)	B	B	B	B	C	C	B	B	C	C
propina (W)	13.0	12.5	11.5	11.0	10.5	11.0	10.5	9.2	9.5	8.5

Datos: $\sum_{i=1}^{20} w_i = 235.9$, $\sum_{i=1}^{20} w_i^2 = 2,849.13$, $\sum_{i=1}^{20} x_i = 40$, $\sum_{i=1}^{20} x_i^2 = 92$, $\sum_{i=1}^{20} x_i w_i = 455.5$.

a) Construye una tabla de contingencia considerando las variables número de personas y número de alimentos en el pedido.

	1	2	3	
A	4	1	0	5
B	1	7	2	10
C	1	0	4	5
	6	8	6	20

b) Construya la tabla de frecuencias relativas conjuntas, incluya las frecuencias marginales.

	1	2	3	
A	$\frac{4}{20}$	$\frac{1}{20}$	0	$\frac{5}{20}$
B	$\frac{1}{20}$	$\frac{7}{20}$	$\frac{2}{20}$	$\frac{10}{20}$
C	$\frac{1}{20}$	0	$\frac{4}{20}$	$\frac{5}{20}$
	$\frac{6}{20}$	$\frac{8}{20}$	$\frac{6}{20}$	1

c) Consturya un diagrama esquemático por tipo de promoción considerando la variable número de personas.

d) Calcule la covarianza y el coeficiente de correlación muestral para las variables tiempo de espera y propina si se sabe que $\sum_{i=1}^{20} y_i w_i = 1,771.302$. Interprete brevemente.

Covarianza $= s_{YW} = \frac{1}{n-1}(\sum_{i=1}^{n} y_i w_i - n\bar{y}\bar{w}) = \frac{1}{19}(1771.302 - 20 \cdot \frac{159.91}{20}\frac{235.9}{20}) = -6.044024$

El único elemento que nos hace falta para calcular el coeficiente de correlación muestral es s_W: $s_W^2 = \frac{1}{n-1}(\sum_{i=1}^{n} w_i^2 - n\bar{w}^2) = \frac{1}{19}(2849.13 - 20 \cdot (\frac{235.9}{20})^2) = 3.509974$

Coeficiente de correlación $= \rho_{YW} = \frac{s_{YW}}{s_Y s_W} = \frac{-6.044024}{\sqrt{22.17}\sqrt{3.51}} = -0.6851952$

El coeficiente de correlación muestral sugiere una relación lineal negativa fuerte.

e) ¿Cómo es la matriz de correlación considerando las variables Y, X y W?

$$\rho_{YXW} = \begin{bmatrix} 1 & 0.77 & -0.58 \\ 0.77 & 1 & -0.69 \\ -0.58 & -0.69 & 1 \end{bmatrix}$$

f) Defina $Z = \alpha Y + \beta W + \epsilon$, con $\alpha, \beta, \epsilon \in \mathbb{R}$. Calcule $\bar{z}$ y s_Z^2.

Media: $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{20}\sum_{i=1}^{20}(\alpha y_i + \beta w_i + \epsilon) = \frac{1}{20}(\sum_{i=1}^{20} \alpha y_i + \sum_{i=1}^{20} \beta w_i + \sum_{i=1}^{20} \epsilon) =$

$\frac{1}{20}(\alpha \sum_{i=1}^{20} y_i + \beta \sum_{i=1}^{20} w_i + 20 \cdot \epsilon) = \alpha\bar{y} + \beta\bar{w} + \epsilon = 7.9955\alpha + 11.795\beta + \epsilon$

Varianza: $s_Z^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 = ...$