RESOLUCIONES DE ALGUNOS EJERCICIOS DEL TRABAJO PRÁCTICO N°1- Estadística C-CA-D. 2020

1) Una importante empresa que vende productos cosméticos para adolescentes quiere investigar los perfiles de sus clientas. Como parte de este estudio desea conocer la edad promedio por clienta. Con este objetivo la empresa seleccionó de sus registros una muestra aleatoria de 200 clientas que compraron productos durante el último mes en los distintos puntos de ventas y para cada una de ellas se registró su edad. La edad promedio de las 200 clientas será utilizada para estimar la edad promedio de todas las clientas de la empresa.

a) ¿Cuál es la variable de interés? Clasificarla.
X= Edad por clienta de la empresa. Cuantitativa Continua.
b) ¿Cuál es la población?
Todas las clientas de la empresa.
c) ¿Cuál es la unidad observacional?
La clienta.
d) ¿Cuál es la muestra?
Las 200 clientas.
e) ¿Cuál es la población estadística?
El conjunto de todas las edades obtenidas de la población.

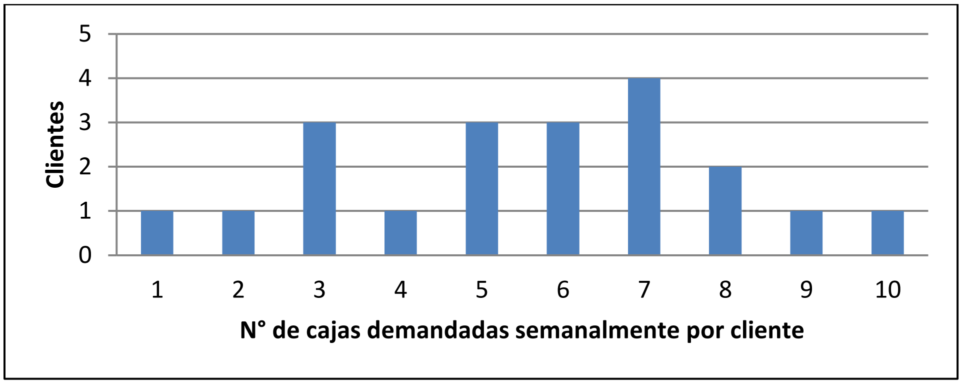
3) a) Nº de piezas en existencia de un artículo en el depósito de una empresa. Cuantitativa Discreta.
b) Cargo de un empleado en una empresa. Cualitativa.
c) Exportación mensual en toneladas de un producto en el mes de abril. Cuantitativa Continua
d) Exportación mensual en unidades de un producto en el mes de abril. Cuantitativa Discreta.
e) Cociente entre activo y pasivo circulante. Cuantitativa Continua
f) Plan de salud al que pertenece un empleado de una empresa. Cualitativa.
g) Formas de pago que ofrece un comercio. Cualitativa.
h) Número de alumnos que asisten a una clase. Cuantitativa. Discreta.
i) Venta mensual (en $) de un producto básico en un comercio. Cuantitativa. Continua
j) Respuestas favorables y desfavorables obtenidas en una encuesta de opinión. Cualitativa.
k) Tiempo de espera (en minutos) de un cliente en la atención de un servicio.
Cuantitativa.Continua
l) Número de lista de un partido político. Cualitativa.

4) a) X: N° de cajas demandadas semanalmente por cliente. Clasificación: cuantitativa discreta

b)

N° de cajas demandadas semanalmente por cliente	fi=N° clientes	fr	Fi	Fri
1	1	0,05	1	0,05
2	1	0,05	2	0,1
3	3	0,15	5	0,25
4	1	0,05	6	0,3
5	3	0,15	9	0,45
6	3	0,15	12	0,6
7	4	0,2	16	0,8
8	2	0,1	18	0,9
9	1	0,05	19	0,95
10	1	0,05	20	1
Total	20	1		

Gráfico de barras



c) Para X=6
Frecuencia absoluta fi=3. Hay 3 clientes de los encuestados que demandaron 6 cajas por semana.
Frecuencia acumulada Fi=12. Hay 12 clientes de los encuestados que demandaron a lo sumo 6 cajas por semana.

Para X=8
Frecuencia absoluta fi=2. Hay 2 clientes de los encuestados que demandaron 8 cajas por semana.
Frecuencia acumulada Fi=18. Hay 18 clientes de los encuestados que demandaron a lo sumo 8 cajas por semana.

d) Respuesta = 7/20=0,35.

e) El 60 % de los clientes encuestados compraron a lo sumo 6 cajas por semana. Mirar la frecuencia acumulada.

f) Respuesta= 38,8% Sale de hacer 7/18=0,388X100.

g)$\overline{X}$= 5,6 cajas (¡¡no se redondea el resultado final , sí podríamos hacerlo para la interpretación!!) Interpretación: el N° promedio de cajas demandadas semanalmente por cliente en la muestra fue de aproximadamente 6 cajas.
Moda= 7 cajas Interpretación: El N° de cajas demandadas semanalmente por cliente que se repitió con mayor frecuencia es 7 cajas.

h) Mediana (Q_2=cuartil de orden 2), el cuartil de orden 1(Q_1) y el cuartil de orden 3 (Q_3)

Orden de Q_1= (21/4)= 5,25. El cuartil de orden 1 es el promedio entre el dato 5 y el 6. Q1=_3,5_cajas
Interpretación: El 25% de los clientes encuestados demandó semanalmente a lo sumo 4 cajas (aproximadamente)

Orden de Q_2= (21/2)=10,5. El cuartil de orden 2 es Q2=6 cajas
Interpretación: (el 50% de los clientes encuestados demandó semanalmente a lo sumo 6 cajas)

Orden de Q_3= (3*21/4)=15,75. El cuartil de orden 3 es el promedio entre el dato 15 y el 16. Q3=7 cajas
Interpretación: (el 75% de los clientes encuestados demandó semanalmente a lo sumo 7 cajas)

i) Rango=9 (Valor máximo-Valor Mínimo=10-1) El N° de cajas demandadas semanalmente por los clientes encuestados osciló dentro de un rango de 9 cajas.

Desvío estándar S=2,39 cajas (Es el alejamiento promedio del N° de cajas demandadas semanalmente por los clientes encuestados con respecto N° promedio de cajas demandadas semanalmente por cliente)

j) Unidad de medida de la media, mediana, desvío estándar y los cuartiles: N° de cajas.
Unidad de medida de la varianza: N° de cajas al cuadrado.
Unidad de medida de la muestra: el cliente.

6) a) X: monto mensual del servicio de gas por casa de dos dormitorios durante el mes de julio. Cuantitativa continua.

Monto de gas ($) (Clases)	Fi =Nº casas	fr	Fi	Fri	Xi=marca de clase
[480-520)	5	0,08	5	0,08	500
[520-560)	7	0,12	12	0,20	540
[560-600)	8	0,13	20	0,33	580
[600-640)	11	0,19	31	0,52	620
[640-680)	8	0,13	39	0,65	660
[680-720)	9	0,15	48	0,80	700
[720-760)	8	0,13	56	0,93	740
[760-800)	4	0,07	60	1,00	780
Total	60	1			

c) Para la clase [560-600)
fi=8. Hay 8 casas de 2 dormitorios que gastaron durante el mes de julio pasado por lo menos $560 pero menos de $600.
Fi=20. Hay 20 casas de 2 dormitorios que gastaron durante el mes de julio pasado menos de $600.

Por ejemplo para la clase [640-680)
fi=8. Hay 8 casas de 2 dormitorios que gastaron durante el mes de julio pasado por lo menos $640 pero menos de $680.
Fi=39. Hay 39 casas de 2 dormitorios que gastaron durante el mes de julio pasado menos de $680.

d) y e) Para frecuencias absolutas:
Se puede realizar un Histograma con frecuencias absolutas o frecuencias relativas en el eje vertical, y el Polígono de frecuencias absolutas.

Para frecuencias acumuladas:
Se puede realizar Histograma de frecuencias acumuladas o frecuencias relativas acumuladas en el eje vertical, y la Ojiva.

f) Respuesta=43.33% sale de hacer (26/60)*100.

g) Respuesta=$680

h) Respuesta = 47.2% sale de (26/55)*100.

i) $\overline{X}$= $ 639,33 El monto promedio de gas ($) por vivienda de 2 dormitorios durante el mes de julio pasado en la muestra fue de $ 639,33

S= $ 81,84 Es el alejamiento promedio de los montos de gas ($) de las viviendas con respecto al monto promedio.

7) X_m: Monto a cobrar por cuenta durante el mes de marzo.

X_a: Monto a cobrar por cuenta durante el mes de abril.

Clasificación: cuantitativa continua.

Para evaluar la heterogeneidad de dos o más conjuntos de datos se deben calcular los coeficientes de variación. El **coeficiente de variación** de un conjunto de datos, ya sean muestrales o poblacionales, describe al desvío estándar en relación con la media. Para obtener ambos **coeficientes de variación** (CV) se divide cada desvío estándar por su media y se multiplica por 100.
En primer lugar, se calcula la media y el desvío estándar de cada muestra:

$$\overline{X_m} = \frac{\sum x_i f_i}{N} = \frac{11000 * 6 + 13000 * 13 \ldots}{50} = 14800$$

$$\overline{X_a} = \frac{\sum x_i f_i}{N} = \frac{11000 * 10 + 13000 * 14 \ldots}{50} = 14400$$

$$S_m = \sqrt{\frac{\sum (x_i - \overline{\overline{X_m}})^2 f_i}{N-1}} = \sqrt{\frac{(11000 - 14800)^2 * 6 + (13000 - 14800)^2 * 13 \ldots}{50-1}} = 2416$$

$$S_a = \sqrt{\frac{\sum (x_i - \overline{\overline{X_m}})^2 f_i}{N-1}} = \sqrt{\frac{(11000 - 14400)^2 * 10 + (13000 - 14400)^2 * 14 \ldots}{50-1}} = 2657$$

Luego se calcula el CV del monto a pagar por cuenta durante el mes de marzo y el CV del monto a pagar por cuenta del mes de abril:

$$CV_m = \frac{S_m}{\bar{X}_m} = \frac{2416}{14800} * 100 = 16\%$$

$$CV_a = \frac{S_a}{\bar{X}_a} = \frac{2657}{14400} * 100 = 18\%$$

Los montos fueron más heterogéneos en abril, pues el coeficiente de variación del conjunto de datos correspondiente al mes de abril es mayor que el correspondiente al mes de marzo.

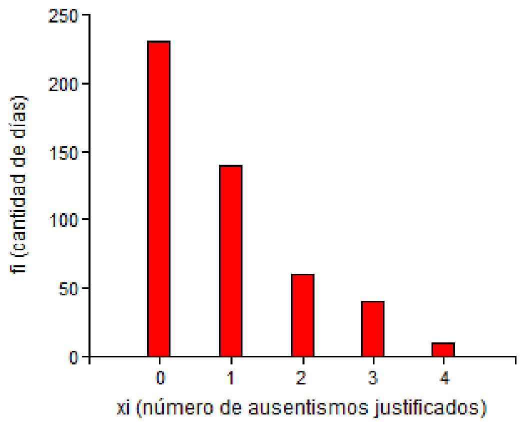
8) a) Se toma cada uno de los 480 días y se observa la característica que se desea medir. La unidad de observación es el día y la característica observada, es el número de ausentismos justificados en el día. Por lo tanto, la variable de interés es:

- Número de ausentismos justificados por día en una empresa.

b)

x_i (Número de ausentismos justificados por día)	f_i (días)	F_i	fr_i	Fr_i
0	230	230	230/480 = 0,48	230/480 = 0,48
1	140	370	140/480 = 0,29	370/480 = 0,77
2	60	430	60/480 = 0,13	430/480 = 0,90
3	40	470	40/480 = 0,08	470/480 = 0,98
4	10	480	10/480 = 0,02	480/480 = 1,00
Totales	∑ = 480		∑ = 1,00	

Como la variable es cuantitativa discreta, selecciono el gráfico de barras con frecuencias absolutas (también se podría haber representado con frecuencias relativas).



c) Cálculo de la media aritmética:

$$\bar{x} = \frac{\sum \ x_i . f_i}{n} = \frac{0.230 + 1.140 + 2.60 + 3.40 + 4.10}{480} = 0{,}875 \ ausentismos\ justificados$$

Interpretación: En promedio, la cantidad de ausentismos justificados por día en una empresa fue de 0,875.

Moda = 0 Interpretación: Lo más frecuente fue observar que no hubo ausentismos justificados en un día

d) el 50% de los días hace referencia a la mediana $\tilde{x}$:

el orden de la mediana es: $\frac{n+1}{2} = \frac{480+1}{2} = 240{,}50$ es la posición en la cual se ubica el dato central. Entonces, recurro a la columna de frecuencias absolutas acumuladas e imagino los 480 días ordenados de menor a mayor. El dato ubicado en la posición 240,5 está entre el dato 240 y el 241. Luego la característica que observo en ambos datos es el Número de ausentismos justificados

dato 240⟶ corresponde a 1 ausentismo justificado ese día.

dato 241⟶ corresponde a 1 ausentismo justificado ese día.

Por último, como estoy en el medio de ambos, promedio el resultado y el valor de la mediana resulta: $\frac{1+1}{2} = 1$ ausentismo justificado.

Finalmente, el 50% de los días se registró a lo sumo 1 ausentismo justificado.

e) el alejamiento promedio de la variable respecto de la media lo da el desvío estándar. En este caso calculamos el desvío estándar muestral. Por lo tanto:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{n-1}} =$$

$$\sqrt{\frac{(0-0{,}875)^2 * 230 + (1-0{,}875)^2 * 140 + (2-0{,}875)^2 * 60 + (3-0{,}875)^2 * 40 + (4-0{,}875)^2 * 10}{480-1}} = 1{,}05$$

Finalmente, el alejamiento promedio del número de ausentismos justificados con respecto a su media es de _1,05_ ausentismos justificados.

9) a) Para obtener el número promedio de camas por habitación, se suma el número de camas de todas las habitaciones y se lo divide por N=23. La unidad observacional es la habitación.

Entonces:

$$\bar{X} = \frac{\sum x_i}{N} = \frac{3+3+1\ldots}{23} = 2{,}5$$

En el hotel hay un promedio de 2,5 camas por habitación.

b) Para obtener el número promedio de camas por piso, se suma el número de camas de todos los pisos (coincide con el número de camas de todas las habitaciones) y se lo divide por N=3 pisos. La unidad observacional es el piso.

Entonces:

$$\bar{X} = \frac{\sum x_i}{N} = \frac{3+3+1\ldots}{3} = 19{,}3$$

En el hotel hay un promedio de 19,3 camas por piso.

c) Para obtener el número promedio de camas por habitación del tercer piso, se suma el número de camas de todas las habitaciones del tercer piso y se lo divide por N=6. La unidad observacional es el tercer piso.

Entonces:

$$\bar{X} = \frac{\sum x_i}{N} = \frac{2+1+3\ldots}{6} = 2{,}7$$

En el hotel hay un promedio de 2,7 camas por habitación del tercer piso.

10) a) Para evaluar la heterogeneidad u homogeneidad de dos conjuntos de datos, se utiliza el Coeficiente de Variación en términos porcentuales.

Si llamamos VF a lo referido al conjunto de Valuación Fiscal y TL al conjunto referido a Tamaño del Lote,

$CV_{VF} = \frac{s_{VF}}{\bar{x}_{VF}}.100\% = \frac{38{,}66}{192{,}45}.100\% = 20{,}09\%$ $CV_{TL} = \frac{s_{TL}}{\bar{x}_{TL}}.100\% = \frac{148{,}36}{333{,}25}.100\% = 44{,}52\%$

Por lo tanto, es más heterogénea la muestra del Tamaño del Lote.

b)

- el 50% de las casas muestreadas tienen una valuación fiscal inferior a… hace referencia a la mediana de VF que es $\tilde{x}_{VF}$ = 187,50 miles de pesos.
- el 25% de las casas muestreadas tienen una valuación fiscal inferior al cuartil de orden 1, justamente, porque divide al conjunto de datos en cuatro partes, y toma la primera; o sea Q_1 = 155,60 miles de pesos.
- Con el mismo criterio, dividimos el conjunto de datos en cuatro partes y tomamos las tres primeras partes (observen que el orden de Q_3 se obtiene multiplicando por ¾ a (n+1)). Entonces decimos: el 75% de las casas muestreadas tienen una valuación fiscal inferior a 215,85miles de pesos.

c)

- Dado que la mediana $\tilde{x}_{TL} = 355m^2$, sabemos que el 50% de las casas muestreadas tienen un tamaño de lote superior (o inferior, es indistinto porque divido a la mitad) a 355m².
- Se solicita el % de las casas muestreadas que tienen un tamaño de lote superior a 308,78m². Como la información corresponde al cuartil 1, y además observa por encima (dice "superior") de éste, no puedo referirme al primer 25% sino que hacer referencia al $\boxed{75\%}$ restante, que se ubica a la derecha (por encima) de este Q_1.
- Se solicita el % de las casas muestreadas que tienen un tamaño de lote superior a 390,65m². El concepto es el mismo al anterior. Sabemos que Q_3 = 390,65 deja a la derecha el 25% de los datos, y a la izquierda el 75% de los datos. Por lo tanto, el _25_% de las casas muestreadas tienen un tamaño de lote superior a 390,65m².

d) medir el alejamiento promedio de los valores de VF con respecto a su valor medio, hace referencia al desvío estándar, que ya está calculado y es: 38,66 mil pesos.

e) la máxima VF en la muestra es el Máximo que no está provisto. Sin embargo, el contar con el Rango que mide 185 y el mínimo que es 115, me indica que el conjunto de datos inicia en 115 mil pesos y recorre 185 miles de pesos más. Por lo tanto, el valor máximo es 115 + 185 = 300 mil pesos.

El mínimo tamaño de lote se calcula: 740 – 580 = 160 m².

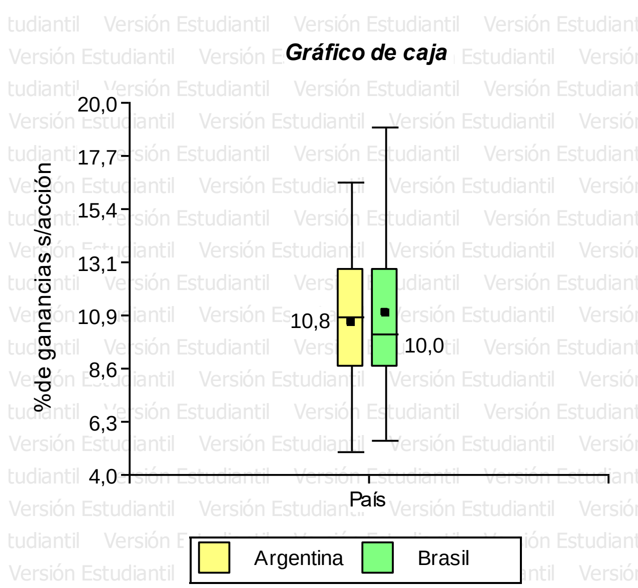
f) para obtener la VT total tendríamos que sumar todos los valores de la muestra, pero no tenemos todos los datos. Sin embargo, sabemos que el promedio de las 20 casas es 192,45 mil pesos. Entonces,

$$\bar{x} = \frac{\sum \ x_i . f_i}{n} =$$

$$192{,}45 = \frac{\sum \ x_i . f_i}{20}, o\ sea: 192{,}45\ x\ 20 = 3949\ miles\ de\ pesos\ \ \equiv 3.849.000\ \$$$

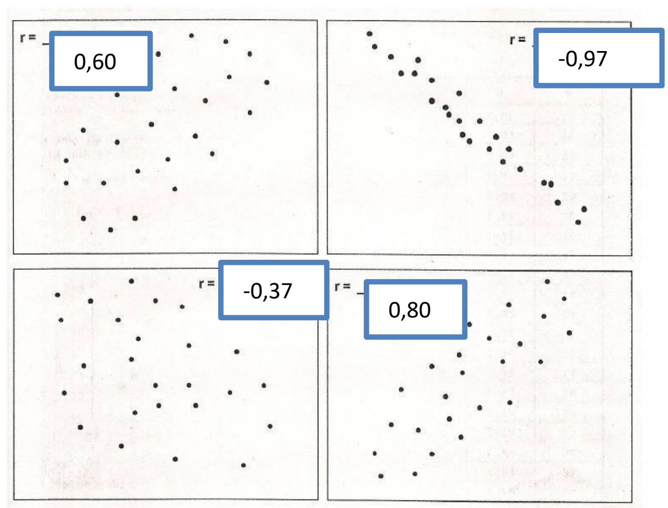
g) $\bar{x}$ =333,25 m².

11)

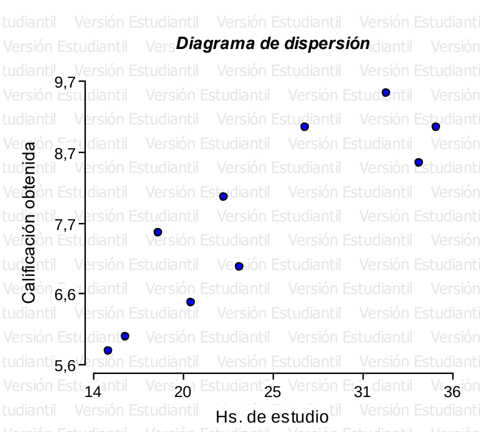


Este resumen de 5 números permite reconocer y comparar la forma de ambos conjuntos de datos. A partir de ambos gráficos de caja se puede observar que el comportamiento de los porcentajes de ganancias sobre acciones de las 15 empresas de Argentina es menos disperso con respecto a las de Brasil. Para el caso de Brasil puede observase que el comportamiento del 50% de los datos centrales presentan una asimetría a derecha. El bigote del extremo derecho de la caja tiene una mayor longitud lo que significa que hay una mayor dispersión en el 25% superior de la distribución de datos. Ambas características marcan la tendencia a una asimetría con sesgo positivo en el comportamiento de los porcentajes de ganancias sobre acciones de las 15 empresas de Brasil. Para el caso de Argentina puede observarse una distribución simétrica para el 50% central de los datos (la mediana cae en el centro de la caja). El 25% de los valores de los extremos presentan una variabilidad semejante dado que la longitud de ambos bigotes es casi la misma. El punto negro en ambas cajas representa el valor de la media, lo que permite comparar su valor con el valor de la mediana. Cuando un conjunto de datos es simétrico, o casi simétrico, las medidas de tendencia central son muy semejantes, esto es lo que puede observarse en el caso de Argentina donde el valor de la media es muy próximo al valor de la mediana. En cambio, para el caso de Brasil, dada la tendencia a una asimetría con sesgo positivo, se puede observar que el valor de la mediana es menor al de la media.

13)



14)



r = 0.87. Existe una correlación lineal fuerte positiva entre el número de horas dedicadas al estudio y la calificación obtenida en el correspondiente examen.