Problemas de Estadística

I Una variable

Resolver con GeoGebra las preguntas que sean posibles [1]

1. **Saldo de las cuentas de ahorro**
   Un banco de la ciudad desea información respecto del saldo de las cuentas de ahorro de sus clientes. El resultado de una muestra aleatoria indica los siguientes balances (en cientos de pesos).

8	**9**	**19**	**5**	**13**	**14**	**5**	**6**	**12**
13	**3**	**6**	**5**	**11**	**6**	**7**	**4**	**11**
23	**5**	**32**	**25**	**21**	**6**	**7**	**11**	**12**
8	**15**	**16**	**17**	**9**	**10**	**8**	**22**	**18**

   a) Hacer un diagrama de tallo y hojas. Elegir la mejor opción.
   b) Hacer un histograma y una ojiva agrupando en intervalos de clase comenzando por [3; 8). Describir el sesgo de estos datos.
   c) Hallar los valores límites de *z* correspondientes al intervalo entre el percentil 82 y el percentil 32. Calcular la mediana y los percentiles de dos formas, agrupando con los intervalos de clase. y luego sin agrupar. Discutir la conveniencia de cada método.
   d) ¿Qué proporción de esta muestra se encuentra como máximo a una desviación estándar del valor medio(es decir entre $x_1 = \bar{x} - s = 4.92$ y $x_2 = \bar{x} + s = 18.52$)? Realizar el cálculo de dos formas, agrupando con los intervalos de clase. y luego sin agrupar. Observar la conveniencia del primer método. Obtener las ecuaciones matemáticas que definen a cada una de las rectas de los intervalos que se usen en la interpolación.
   e) dibujar el diagrama de caja.
   R: GeoGebra c) $-0.645<z<0.96$.
2. **Edad de cada cliente**
   Una muestra aleatoria de 30 clientes en la inauguración de una tienda de indumentaria aportó la siguiente información respecto a la edad de cada cliente:

43	**33**	**18**	**23**	**19**	**16**
51	**54**	**18**	**26**	**25**	**21**
17	**30**	**28**	**27**	**27**	**17**
32	**21**	**35**	**40**	**39**	**36**
48	**47**	**38**	**41**	**50**	**19**

   a) Hacer un diagrama de tallo y hojas. Elegir la mejor opción.
   b) Hacer un histograma y una ojiva agrupando en intervalos de clase, comenzando por [16; 23). Describir el sesgo de estos datos
   c) Realizar el diagrama de caja (mediana/ cuartiles/ amplitud). Calcular los cuartiles de dos formas, agrupando con los intervalos de clase. y luego sin agrupar. Discutir la conveniencia de cada método.

[1] Tener en cuenta que GeoGebra calcula las estadísticas sin agrupar por intervalos de clases para no tener así errores de agrupamiento.

d) ¿Cuál es la probabilidad (frecuencia relativa) de que un cliente seleccionado al azar tenga una edad comprendida entre $z_1 = -1.15$ ($x_1 = 17.92$) y $z_2 = 1.60$ ($x_2 = 49.90$)? Realizar el cálculo de dos formas, agrupando con los intervalos de clase. y luego sin agrupar. Discutir la conveniencia de cada método.
e) dibujar el diagrama de caja
R: GeoGebra. d) 78.4%.

3. **Edad de alumnos**
El director de una escuela para adultos desea saber la distribución de las edades de los alumnos en la clase, obteniendo la siguiente información (expresada en años):

63	**41**	**28**	**22**	**45**	**30**	**43**	**46**	**35**	**56**	**27**	**18**
29	**20**	**31**	**19**	**22**	**32**	**55**	**54**	**31**	**63**	**51**	**39**
29	**19**	**27**	**53**	**51**	**33**	**58**	**29**	**65**	**53**	**56**	**19**

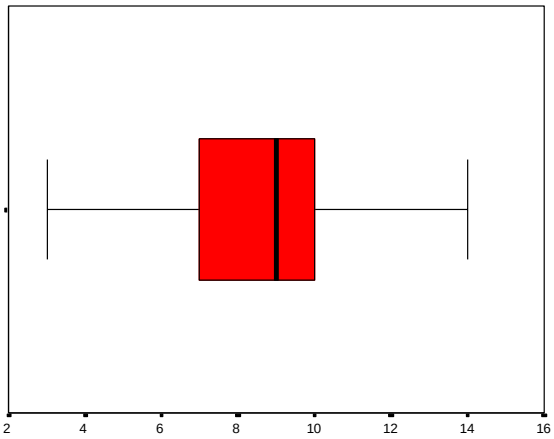
a) Hacer un diagrama de tallo y hojas. Elegir la mejor opción.
Agrupando en intervalos de clase [.,), siendo las dos primeras marcas 22.5 y 32.5,
b) ¿qué cantidad de alumnos se encuentra por debajo de $z = -1.64$ ($x = 21.7$)?
c) expresar en valores z los límites del intervalo comprendido entre los percentiles 19 y 81.
d) Dibujar el diagrama de caja y describir el sesgo de estos datos.
e) Dibujar el histograma y la ojiva. Colocar en cada uno de ellos las ecuaciones matemáticas que definen a cada una de las rectas. Verificar el cumplimiento de la relación derivada–integral para cada recta
R: GeoGebra b) 0 alumnos, c) -1.03;1.09

4. **Cola de un banco**
El diagrama de tallo y hojas dibujado, corresponde al tiempo de permanencia (en minutos) de los clientes en la cola de un banco.

Frec	**Tallo**	**Hojas**
2	**0 .**	**33**
6	**0 .**	**444555**
8	**0 .**	**66777777**
9	**0 .**	**888899999**
5	**1 .**	**00111**
3	**1 .**	**223**
2	**1 .**	**44**
Ancho Tallo:	**10.00**	
Cada hoja:	**1 caso(s)**	

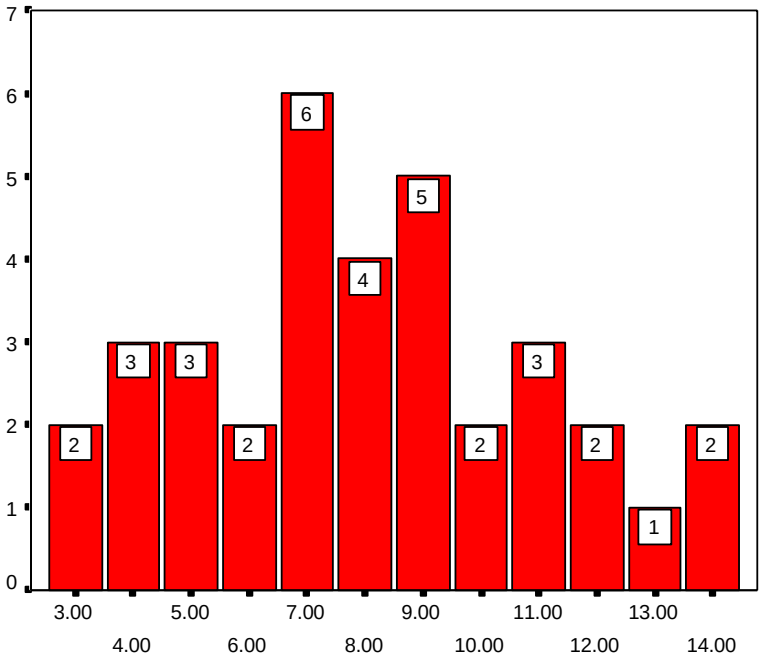
Ingresar los datos a GeoGebra (Ancho Tallo significa como leer el valor 1.0)
A partir de esta información responder las siguientes preguntas.
a) ¿Qué cantidad de clientes se encuentra por encima de z = 1.33?
b) Expresar en valores z los límites del intervalo comprendido entre los percentiles 22 y 80.
R: a) 3=8.57%; -0.95;0.96

5. **Precios de empresas de la bolsa de comercio**
El siguiente diagrama de caja se realizó con los datos de una muestra de los precios de cierre de 35 empresas de la bolsa de comercio de Buenos Aires. Del mismo se desprende que $x_I = 3$, $x_S = 14$, $Q_1 = 7$, $Q_2 = 9$ y $Q_3 = 10$.



Suponiendo que los datos se distribuyen linealmente en cada uno de los 4 intervalos, ingresar los datos a GeoGebra y hallar:
a) ¿qué cantidad de empresas se encuentra a más de una desviación estándar de la media?
b) expresar en valores z los límites del intervalo comprendido entre los percentiles 28 y 85
c) dibujar la forma que tiene el histograma. ¿El sesgo parece ser positivo, negativo o cero? Justificar la respuesta.
R: a) 12.79empresas=36.5%,
b) -.0538;1.15

6. **Peso de equipaje de mano**
Una compañía aérea desea saber cuánto pesa el equipaje de mano de los pasajeros. Un empleado toma una muestra aleatoria de 35 pasajeros que retornan de Chile y presenta los pesos (en kg) en el siguiente diagrama de barras. Si bien no confeccionó un histograma, lo cual hubiera sido más correcto, ¿Qué relación guarda este gráfico con un diagrama de tallo y hojas?.



Ingresar los datos a GeoGebra.
a) Confeccionar un histograma de 6 intervalos de base 2 kg y cuya primera marca sea 4 kg. En base al mismo hallar los valores de z correspondientes a los límites del intervalo comprendido entre los percentiles 29 y 75.
b) ¿qué proporción de la muestra se encuentra a no más de 1.5 desviaciones estándares de la media?
c) ¿En qué unidades se miden $\bar{x}, s, , z, \sigma$?.
R: a) Intervalos [,): – 0.506kg;0.75kg, Intervalos (,]: -0.70kg, 0.65kg.
b) Intervalos [,): 87.1%, Intervalos (,]: 87.8%.

7. **Peso de equipaje de mano**

Una empresa de transporte público de larga distancia desea establecer el peso del equipaje de mano que llevan sus pasajeros. Para ello toma una muestra aleatoria que se adjunta en la siguiente tabla (pesos en kg).

30	**27**	**12**	**42**	**35**	**47**	**38**	**36**	**27**	**35**
22	**17**	**29**	**3**	**21**	**0**	**38**	**32**	**41**	**33**
26	**45**	**18**	**43**	**18**	**32**	**31**	**32**	**19**	**21**
33	**31**	**28**	**29**	**51**	**12**	**32**	**18**	**21**	**26**

Ingresar los datos a GeoGebra.

a) Hacer un diagrama de tallo y hojas ¿Parece una distribución asesgada? ¿En caso contrario con que signo?

b) Calcular el valor z de x = 32. Puede usarse el diagrama de tallo y hojas como si fuera un histograma de datos agrupados para acelerar el cálculo.

c) Obtener el diagrama de caja calculando los estadísticos necesarios con los datos sin agrupar. Se considera que los datos que se encuentren a más de $1.5 \times AIC$ contados a partir de los extremos de la caja son valores extremos o atípicos que requieren un análisis especial. Individualizar los valores extremos de esta distribución, si es que los tiene. Justificar.

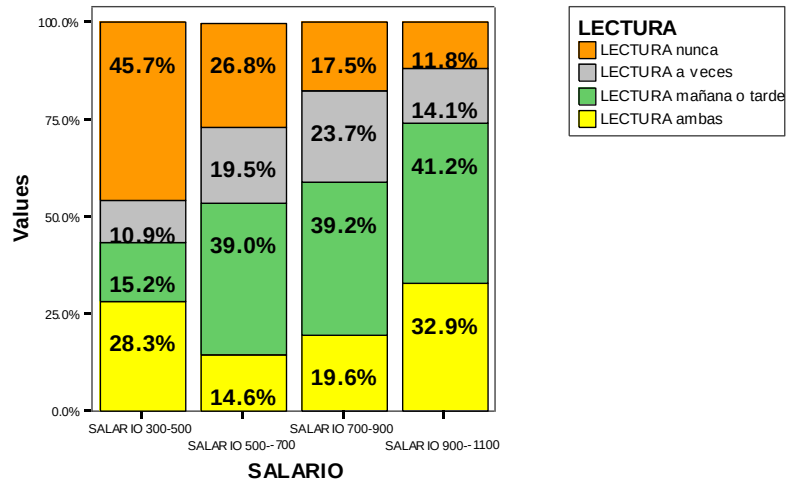
R: b) 0.547,

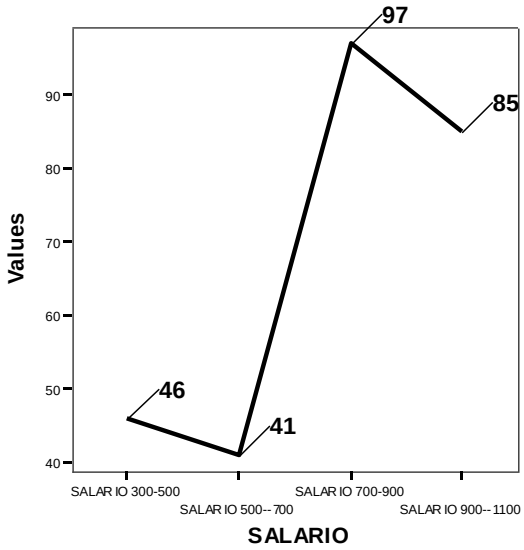
c) Q_1=21kg, Q_2=29.5kg, Q_3=35kg.

II Dos variables

Asociación

8. **Salario vs lectura de diarios**
Un editor de periódicos se pregunta si la costumbre de la gente de leer diarios está relacionada con el salario de los lectores. Se aplica una encuesta obteniéndose, entre otros, los siguientes gráficos.

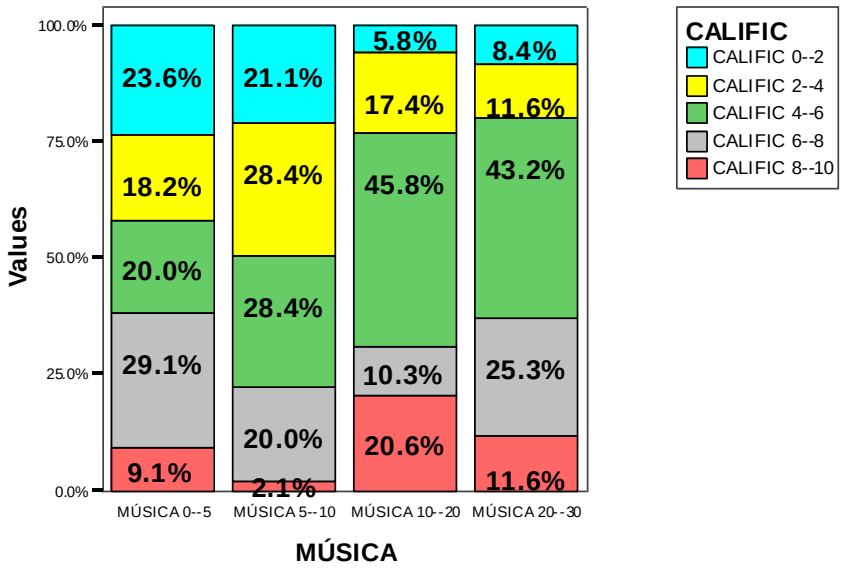




Ingresar los datos a GeoGebra y resolver.
Completar los espacios en blanco en el siguiente párrafo. Adjuntar el proceso de cálculo.
Para la variable de escala Salario, la media es _____, la mediana es _____, la desviación estándar es _____ y el valor z para un salario de 700$ es: _____. El sesgo de estos datos es _____ Dibujar el diagrama de barras y el circular.
En la muestra de _____personas, se tienen _____casos que nunca leen los diarios, de los cuales el _____% es decir_____, tienen salarios entre $900 y $1100. De los que tienen salarios entre $700 y $900, el _____% es decir _____, leen mañana o tarde. Entre los que leen ambas ediciones, la mayor proporción se presenta en los _____con un _____% y un total de _____personas. El grupo más numeroso se presenta en el cruce de _____ _____con _____. Son _____personas que representan el _____% del total. El cálculo del coeficiente de asociación adecuado a las variables en estudio da exactamente

_____ lo cual podría ser una indicación_____(Si/No) de que existe relación entre las variables.
R: 764.3$, 798$, 212.11$, 0.303, izquierdo,
269, 59, 16.9, 10, 39.2, 38, 900-1100, 42.4, 28, 700-900, mañana o tarde / 700-900, 38, 14.1, 0.330, No.

9. **Calificaciones vs tiempo escuchando música**
Un educador tiene la opinión de que las calificaciones que obtienen los alumnos depende del tiempo que se pasan escuchando música. Se entrega un cuestionario a los estudiantes con dos preguntas: ¿Cuántas horas por semana escuchas música? ¿Qué promedio de calificaciones tienes? Del procesamiento resultaron los siguientes gráficos (entre otros).



Count

		Total
MÚSICA	0--5	55
	5--10	95
	10--20	155
	20--30	95
Total		400

Ingresar los datos a GeoGebra.
Completar los espacios en blanco en el siguiente párrafo. Adjuntar el proceso de cálculo.
Para la variable de escala Calificaciones, la media es _____, la mediana es _____, la desviación estándar es _____ y el valor z para una calificación de 7.5 es: _____. El sesgo de estos datos es _____. Dibujar el diagrama de barras y el circular.
En la distribución marginal de la variable Música, la media es _____, la desviación estándar es _____ y el porcentaje de personas que se encuentran en el intervalo $-1.2 < z < 1.1$ es _____. El P_{58} es _____ y su valor z es _____
En la muestra de _____personas, se tienen _____casos que tienen calificación 8 a 10, de los cuales el _____% es decir_____, escuchan música entre 5 y 10 horas. De los que escuchan entre 10 y 20 horas, el _____% es decir _____, tienen calificación 0 a 2. Entre los que tienen calificación 4 a 6, la mayor proporción se presenta en los _____con un _____% y un total de _____personas. El grupo más numeroso se presenta en el cruce de _____ _____con _____. Son _____personas que representan el _____% del total. El cálculo del coeficiente de contingencia da exactamente _____ lo cual podría ser

una indicación_____(Si/No) de que existe relación entre las variables.
Resolver con GeoGebra.
R: 5puntos, 5puntos, 2.34puntos, 1.06, izquierdo, 13.87, 7.58, 69.1%, 15.3. 0.19.
400, 50, 4, 2, 5.8, 9, 10-20, 47.3, 71, 10-20, 4-6, 71, 17.8, 0.371, No.

10. **Presión sanguínea luego del entrenamiento**
La siguiente tabla bidimensional presenta las frecuencias conjuntas de datos de presión sanguínea diastólica de un grupo de deportistas luego del entrenamiento. La variable T, es el tiempo en minutos transcurrido desde el comienzo del descanso y la variable P es la presión.

P T	0	5	10	15	20
66	1	1	0	1	2
68	3	2	1	0	1
70	0	1	9	1	2
72	1	2	1	2	1
74	3	1	2	1	2

Ingresar los datos a GeoGebra.
a) Hallar manualmente y con GeoGebra las distribuciones marginales de P y T y los perfiles fila y columna.
b) Hallar manualmente y con GeoGebra las esperanzas, medianas y varianzas de las 2 distribuciones marginales y de las 5 distribuciones de P condicionadas a T. Describir el sesgo de cada distribución.
c) Hallar el coeficiente de correlación de Pearson (luego de estudiar regresión lineal), el coeficiente ϕ y el coeficiente de contingencia entre P y T.

11. **Economistas, ingenieros y abogados**
Se pregunta a 50 economistas, 40 ingenieros y 10 abogados si creen que la bolsa bajará, subirá o permanecerá igual en el próximo mes. El 20 % de los economistas opina que subirá, mientras que el 40 % de ellos piensa que bajará. El 50 % de los ingenieros se inclina que permanecerá igual y tan solo el 5 % cree que bajará. Por último, la mitad de los abogados cree que subirá y la otra mitad cree que bajará.
Ingresar los datos a GeoGebra.
a) Resumir los datos en una tabla de contingencias que cruce la Profesión con el Pronóstico. Hacer un gráfico de barras que equivalga a la tabla.
b) Realizar el gráfico de barras para la variable Profesión y luego para la variable Pronóstico (distribuciones marginales).
c) Hallar las tablas de las distribuciones condicionales y presentarlas junto con un diagrama de sectores.
R: GeoGebra.

12. **Votantes**
Una muestra de 200 votantes reveló la siguiente información sobre 3 candidatos A, B y C.

28 votaron a favor de A y B
98 a favor de A o B pero no de C
42 a favor de B pero no de A o C
122 a favor de B o C pero no de A
64 a favor de C pero no de A o B
14 a favor de A y C pero no de B

Ingresar los datos a GeoGebra.
a) Formar una tabla de contingencias. Tener en cuenta que los niveles de las variables deben ser mutuamente excluyentes y colectivamente exhaustivos. Cargarla en GeoGebra.
b) ¿Cuántos votantes estaban a favor de solo uno de los candidatos? ¿De los tres? ¿De B independientemente de A o C? ¿De A y B pero no de C? ¿De C independientemente de A o B?
R: 142votantes, 8votantes, 42votantes, 20votantes, 64votantes.
c) Obtener la tabla y un diagrama de barras de los porcentajes dentro de B. ¿Son independientes A y B? ¿A y C? ¿B y C?

Regresión simple

13. **Bebés y Mamás**
La tendencia actual es que los bebés no sean demasiado gordos. Investigaciones médicas indican que existe una correlación entre el peso de un bebé de 1 año de edad (variable x) y el peso de la madre de 30 años de edad (variable y). Una muestra aleatoria de 30 mamás produce la siguiente información:

x	9.50	11.30	10.40	10.90	9.10	6.80	11.30	9.50	7.70	10.90
y	56.60	56.60	54.40	56.60	58.90	54.40	65.70	58.90	58.90	58.90

x	11.80	10.00	8.20	9.50	11.30	10.40	10.90	9.10	6.80	11.80
y	63.40	49.80	52.10	56.60	56.60	54.40	56.60	58.90	54.40	63.00

x	7.20	7.90	6.90	7.30	7.20	7.80	9.10	11.00	11.50	10.80
y	47.00	52.00	51.00	49.00	50.00	64.00	71.00	64.00	70.00	70.10

Ingresar los datos a GeoGebra.
Datos sin agrupar
Completar los espacios en blanco en el siguiente párrafo. Adjuntar el proceso de cálculo.
Para la variable Bebé, la media es _____, la mediana es _____, la desviación estándar es _____, el valor z para un Peso de 10kg es: _____ y el _____% de bebés tienen menos de 9.5kg. Hacer un diagrama de tallo y hojas adecuado.
El cálculo del coeficiente de asociación adecuado a las variables en estudio da exactamente _____, lo cual podría ser una indicación de que _____ (Sí/No) existe una relación entre las variables.
R: 9.463 kg, 7.775 kg, 1.698 kg, 0.316, 45%
0.541, Si, $y = 38.7x + 2.02$, 58.9kg
Datos agrupados
Confeccionar una tabla de contingencias que cruce ambas variables (x en horizontal) agrupándolas en intervalos de clase. Para la variable x se desea que las dos primeras marcas sean 7 y 9. Para la variable y, se desea que el primer intervalo sea [45;54).
Ingresar la tabla de contingencias a GeoGebra.
Completar los espacios en blanco en el siguiente párrafo. Adjuntar el proceso de cálculo.
Para la variable Bebé, la media es _____, la mediana es _____, la desviación estándar es _____, el valor z para un Peso de 10kg es: _____ y el _____% de bebés tienen menos de 9.5kg. Dibujar un histograma y una ojiva.
En la muestra se tienen _____casos con el peso de la mamá entre 54 y 63 kg, de los cuales el _____% es decir_____, tienen bebés con pesos entre 8 y 10 kg. De los bebés con pesos entre 10 y 12 kg, el _____% es decir _____, tienen mamás con pesos entre 45 y 54 kg. Entre los bebés con pesos entre 10 y 12 kg, la mayor proporción se presenta en los _____con un _____% y un total de _____casos. El grupo más numeroso se presenta en el cruce de _____ _____con _____. Son _____casos que representan el _____% del total. El cálculo del coeficiente de contingencias da exactamente _____, lo cual podría ser una indicación de que _____ (Sí/No) existe una relación entre las variables.
R: 9.33 kg, 6.93 kg, 1.75 kg, 0.383, 42.8%
15, 33.3%, 5, 7.14%, 1, [63, 72), 42.8%, 6, [54, 63] y [10, 12), 6 20%. 0.492, Sí

14. Para el problema anterior **Calificaciones vs tiempo escuchando música**,
a) obtener el coeficiente de correlación de Pearson. ¿Coincide con la interpretación de independencia realizada en dicho problema? ¿Cuál coeficiente de asociación es más adecuado, el de correlación para variables cualitativas o el de contingencia para variables cuantitativas?
b) obtener la recta de regresión para Calificaciones=f(Tiempo escuchando música) y predecir la

calificación para Tiempo escuchando música igual a 7 horas.
R: a) 0.182. Sí. el de correlación
b) $Cal = 0.56Mus + 4.21$, 8.14 puntos.

15. **Coeficiente de inteligencia**
Un grupo de investigadores desea estudiar si los estudiantes con alto coeficiente de inteligencia, CI, tienen también altas calificaciones en la escuela. Se sabe que esto es parcialmente cierto pues otros factores afectan el comportamiento académico. Se toma una muestra de 12 estudiantes y se obtienen los datos de la tabla siguiente.

CI (x)	**117**	**92**	**102**	**115**	**87**	**76**	**107**	**108**	**121**	**91**	**113**	**98**
Califc(y)	**3.7**	**2.6**	**3.3**	**2.2**	**2.4**	**1.8**	**2.8**	**3.2**	**3.8**	**3.0**	**4.0**	**3.5**

Ingresar los datos a GeoGebra.
a) Hacer un diagrama de tallo y hojas adecuado para cada variable
b) Obtener el diagrama de caja y el percentil 43 de cada una de las variables.
c) Graficar el diagrama de dispersión y hallar el coeficiente de correlación. ¿Se puede estimar que se tendrá un buen ajuste por una recta de regresión?
d) Obtener las ecuaciones de las rectas de regresión de *y sobre x,* y de *x sobre y*. Graficarlas en el diagrama de dispersión obtenido en el punto a).
d) Utilizando la recta de regresión de *y sobre x*, predecir la calificación para un CI = 95.
Utilizando la recta de regresión de *x sobre y,* predecir el CI para una calificación de 3.1.
R: GeoGebra e) 2.775puntos, 103.2CI.

16. **Productos de artesanía**
En una fábrica de productos de artesanía, algunos trabajadores producen muchas piezas por día mientras que otros producen muy pocas. Los que producen menos, se justifican alegando que obtienen artículos de mejor calidad que los que producen más. El dueño de la fábrica desea probarlo y solicita que se tome una muestra de 15 trabajadores en donde figure para cada trabajador, el número de artículos producidos por día (x) y un índice de calidad promedio (y) para los artículos inspeccionados (los valores más altos significan alta calidad).

x	**12**	**15**	**5**	**1**	**20**	**17**	**19**	**46**	**20**	**25**	**39**	**25**	**30**	**27**	**29**
y	**7.7**	**8.1**	**6.9**	**8.2**	**8.6**	**8.3**	**9.4**	**7.8**	**8.3**	**5.2**	**6.4**	**7.9**	**8.0**	**6.1**	**8.6**

Ingresar los datos a GeoGebra.
a) Hacer un diagrama de tallo y hojas adecuado para cada variable
b) Obtener el diagrama de caja y el percentil 82 de cada una de las variables.
c) Obtener el diagrama de dispersión entre ambas variables.
d) Correlacionar ambos datos por el coeficiente de Spearman (si hay empates asignarle el promedio de los rangos respectivos).
e) Hallar la recta de regresión de: *y sobre x* y graficarla en el diagrama de dispersión
R: GeoGebra.

17. **Tiempo de estadía vs Gasto efectuado**
En un hotel se toma una muestra aleatoria de 10 turistas obteniéndose la tabla adjunta:

Turista	**t (dias)**	**G ($)**
1	**3.0**	**550**
2	**2.0**	**390**
3	**2.0**	**180**
4	**1.5**	**310**
5	**4**	**1350**
6	**1.5**	**390**
7	**2.5**	**560**
8	**1.5**	**365**
9	**2.0**	**362**
10	**3.0**	**720**

Ingresar los datos a GeoGebra.
a) Hacer un diagrama de tallo y hojas adecuado para cada variable
ab) Determinar si existe correlación lineal entre las variables: Tiempo de estadía (t) y Gasto efectuado (G).
c) En caso afirmativo, hallar la mejor recta de ajuste según el método de mínimos cuadrados.

d) ¿Es posible estimar el gasto que realizaría un turista que se quedara 7.5 días? ¿Y para 3.5 días?
R: GeoGebra. d) no, si, 941$.