PREWORK
SESIÓN 2

Introducción

Un **diagrama de dispersión**, **gráfica de dispersión** o **gráfico de burbujas** es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos[1].

Objetivo

- Lograr una mayor comprensión de la situación o fenómeno con el cual se relacionan los datos mediante herramientas de visualización.

Temas

- Gráficos de dispersión
- Boxplots y outliers
- Histogramas
- Series de tiempo y descomposición

Gráficos de dispersión

Los diagramas de dispersión son útiles para reconocer tendencias en datos, cuando estos son graficados en puntos, se deben tomar dos variables de tipo cuantitativo, a continuación, se presenta una imagen donde se pueden observar dos tipos de correlación que generalmente son de interés. Cuando se trata de una alta dispersión se puede suponer que la correlación es cercana a cero, y cuando tenemos poca dispersión se puede deber a correlaciones cercanas a 1 o a -1, sin embargo puede existir poca dispersión si los datos se aglomeran en clusters de información, es decir los datos están muy “compactados”, la correlación puede obtenerse mediante el comando *cor*().

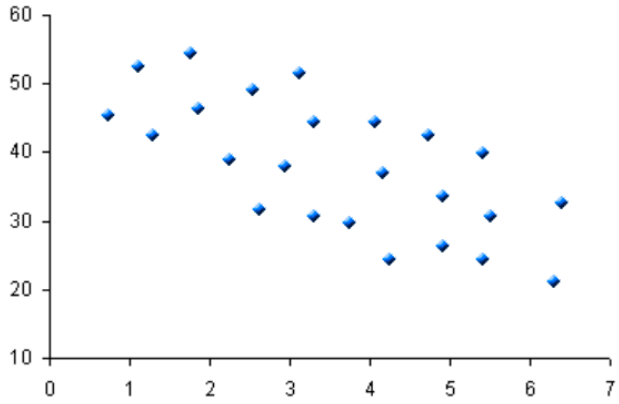


Ilustración 1Correlación negativa

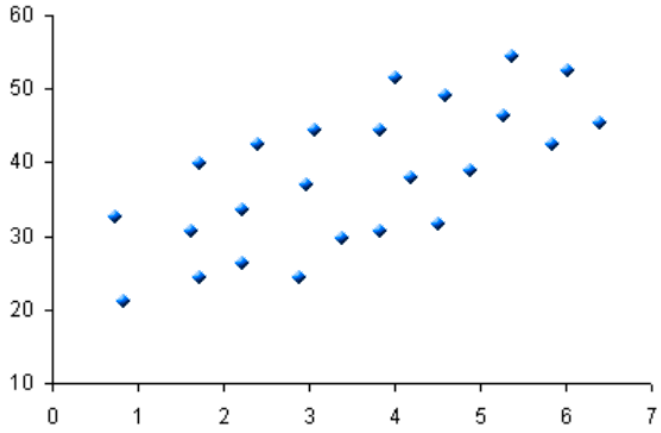


Ilustración 2 Correlación positiva

La ilustración 1 es una correlación lineal negativa, esto sucede cuando una variable crece (en este caso eje de las “x”) y la otra disminuye (eje de la “y”). Para el caso de la ilustración 2, se representa una correlación lineal positiva, cuando una variable crece (eje “x”), la otra también lo hace (eje “y”). Puede suceder que no se pueda reconocer un patrón específico, esto también es útil ya que indica que las variables no tienen una correlación o que esta no es tan fácil de determinar visualmente debido a que su coeficiente de correlación es muy bajo.

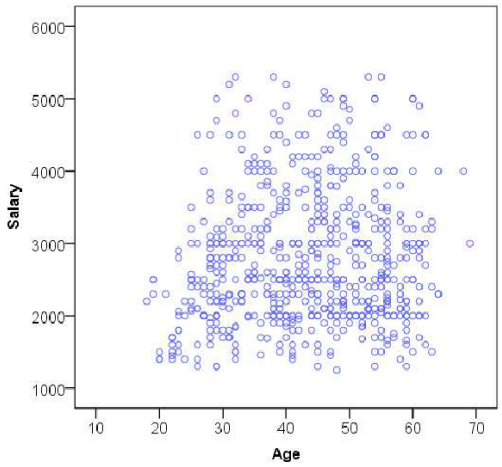


Ilustración 3 No hay correlación

El comando en R para realizar un gráfico de dispersión es: plot(var1, var2, ...)

Boxplots y outliers

También conocido como diagrama de caja y bigote, box plot, box-plot o boxplot. Es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles (1ero; 25%, 2do; 50%, 3ro; 75%) . El máximo interés del box-plot es visualizar la distribución de una variable numérica de la manera más simplificada posible. Sólo utiliza los valores de los cuartiles, los extremos (q1 - 1.5*IQR y q3 + 1.5*IQR) y valores raros o outliers. No depende de valores ponderados como la media. Simplemente se fija en las características de la posición. El diagrama siguiente será de mucha utilidad para comprenderlos. En R se utiliza el comando boxplot para graficarlos.

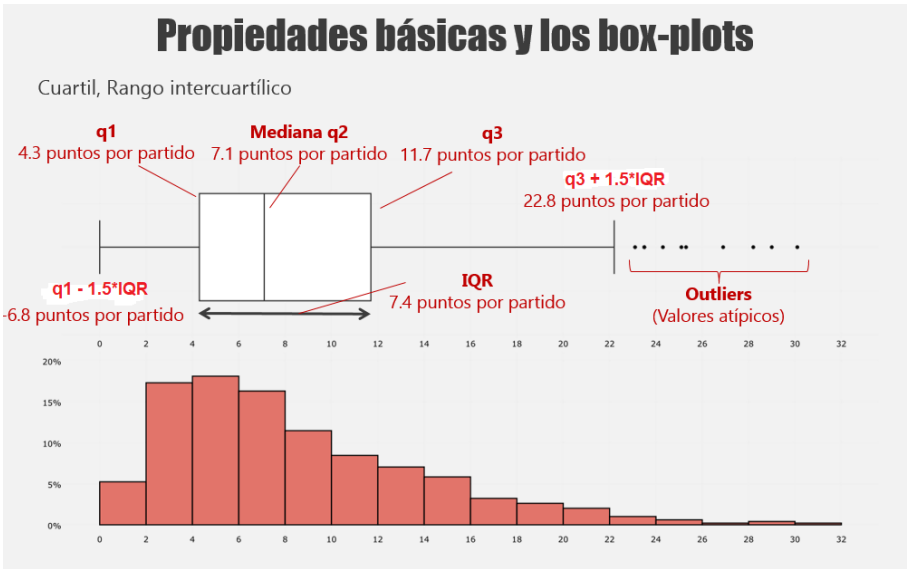
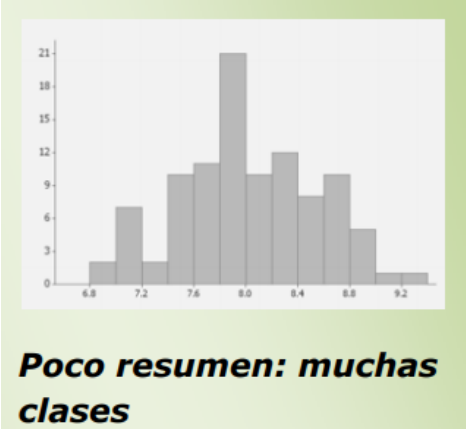
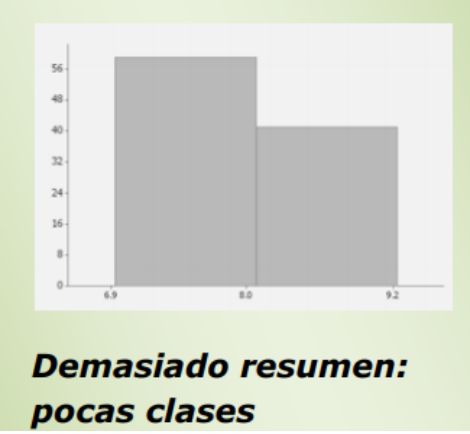


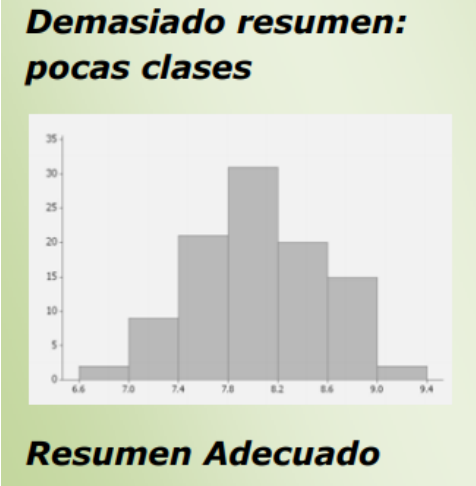
Ilustración 4 Características del Boxplot

Histogramas

Es una gráfica de la distribución de un conjunto de datos. Es un tipo especial de gráfica de barras, en la cual una barra va pegada a la otra, es decir no hay espacio entre las barras. Cada barra representa un subconjunto de los datos. Un histograma muestra la acumulación ó tendencia, la variabilidad o dispersión y la forma de la distribución. Un histograma es una gráfica adecuada para representar variables continuas, aunque también se puede usar para variables discretas. Es decir, mediante un histograma se puede mostrar gráficamente la distribución de una variable cuantitativa o numérica. Los datos se deben agrupar en intervalos de igual tamaño, llamados clases.







Se grafican en el eje de las X las clases y en el eje Y las frecuencias de nuestros datos entonces de ese modo obtenemos el histograma, que es la representación visual de la distribución de frecuencias.

Para realizar un histograma se utiliza el comando en R: hist(var)

Series de tiempo

Es un conjunto de valores observados durante una serie de periodos temporales, secuencialmente ordenada. Son variables estadísticas bidimensionales en donde el tiempo es la variable independiente, y la otra es la variable dependiente.

Se construyen modelos de series de tiempo para:

- Obtención del mecanismo
- Estudio de su evolución futura o predicción.

Se realiza:

- Analizando los componentes o factores que determinan los resultados de la información.

El método clásico para el análisis de series de tiempo identifica cuatro componentes: TENDENCIA (T).- El movimiento general a largo plazo de los valores de la serie de tiempo (Y) sobre un extenso periodo de años.

FLUCTUACIONES CÍCLICAS (C).- Movimientos ascendentes y descendentes recurrentes respecto de la tendencia, con una duración de varios años.

VARIACIONES ESTACIONALES (E).- Movimientos ascendentes y descendentes respecto de la tendencia que se consuman en el término de un año y se repiten anualmente, estas variaciones suelen identificarse con base en datos mensuales o trimestrales.

VARIACIONES IRREGULARES (I).- Las variaciones erráticas respecto de la tendencia que no puedan atribuirse a las influencias cíclicas o estacionales. A continuación, se muestran las partes de una serie de tiempo.

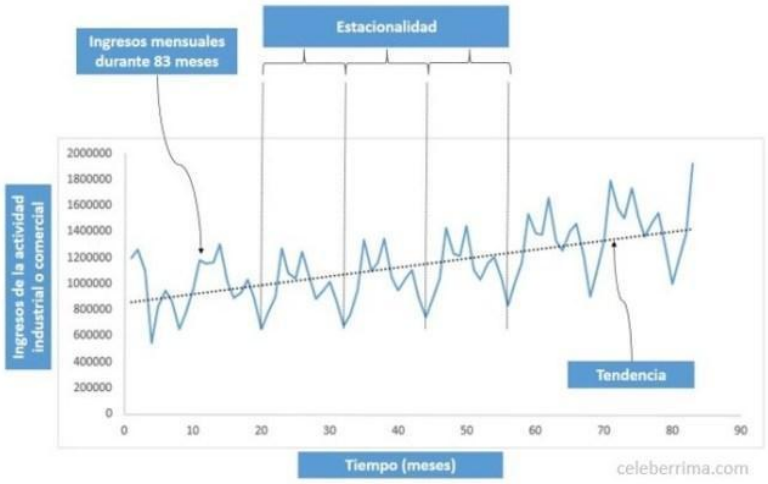


Ilustración 5 Ejemplo de una serie de tiempo y sus partes

Existe la descomposición aditiva y multiplicativa las cuales ayudan a entender el comportamiento de la serie de tiempo.

Para construir las series de tiempo se utilizará el comando ts()

Quiz.

1. Si tenemos un gráfico donde se presente una alta dispersión, esto quiere decir que:
    a. Existe una correlación positiva
    b. La correlación es cercana a 1
    c. La correlación es cercana a -1
    d. La correlación es cercana a cero

2. ¿Con cuál función se puede calcular la correlación de dos variables?
    a. cos()
    b. cor()
    c. com()
    d. correl()

3. Los diagramas de box-plot sirven para detectar outliers
    a. Falso
    b. Verdadero
4. Este tipo de gráfico se utiliza para determinar la variabilidad, dispersión o forma de los datos.
    a. Dispersión
    b. Box-plot
    c. Histograma
    d. Circular
5. Son características de una serie de tiempo
    a. Tendencia
    b. Estacionalidad
    c. Variable dependiente
    d. Correlación

BIBLIOGRAFÍA UTILIZADA

1. Jarrell, Stephen B. (1994). *Basic Statistics* (Special pre-publication edición). Dubuque, Iowa: Wm. C. Brown Pub. p. 492