Análisis de datos

Dra. Amparo López Gaona
Fac. Ciencias, UNAM

Octubre 2015

Introducción

- El paso anterior al minado es conocer los datos con que se cuenta.
- Los datos típicamente son:
  - ruidosos,
  - de enorme volumen
  - provenientes de fuentes heterogéneas
- Interesa conocer:

Introducción

- El paso anterior al minado es conocer los datos con que se cuenta.
- Los datos típicamente son:
  - ruidosos,
  - de enorme volumen
  - provenientes de fuentes heterogéneas
- Interesa conocer:
  - tipos de datos (atributos, valores (rango), etc),
  - cómo se distribuyen (simétrica, normal, dispersa),
  - cómo medir similitudes,
  - cómo detectar outliers,
  - la calidad de los datos,
  - relaciones entre ellos, etc.

...Introducción

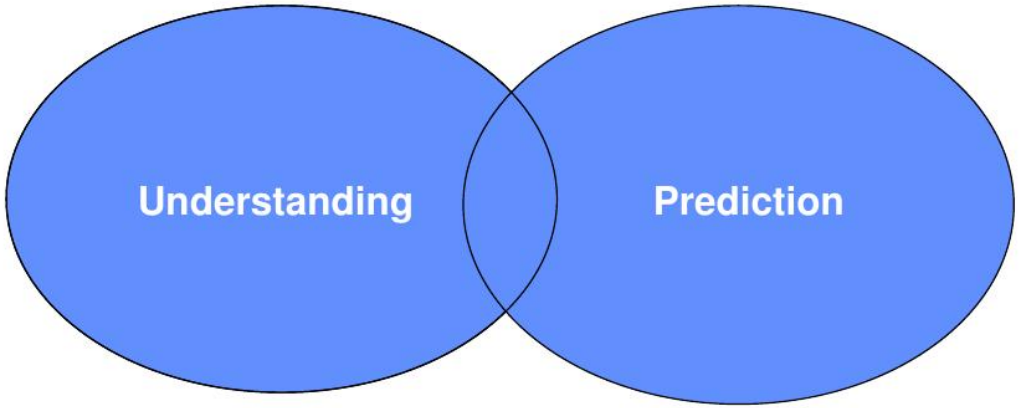
- Estadística tradicional
  - Primero se hace una hipótesis.
  - Se recaban los datos
  - Se analizan los datos.
- DM
  - Hipótesis leve o inexistente.
  - Los datos ya se tienen.
  - Análisis guiados por los datos no por hipótesis

...Introducción

- Estadística tradicional
  - Primero se hace una hipótesis.
  - Se recaban los datos
  - Se analizan los datos.
- DM
  - Hipótesis leve o inexistente.
  - Los datos ya se tienen.
  - Análisis guiados por los datos no por hipótesis
- Diferencias
  - Sí, en términos de cultura, motivación, sin embargo ...
  - Las ideas estadísticas son muy útiles en MD por ejemplo para validar la utilidad del conocimiento descubierto.

...Introducción

- Dos metas complementarias en la MD:



... Introducción (Ejemplo)

Suponer que recibes un mensaje de un investigador médico con el que vas a trabajar en un proyecto:

Hi,

I've attached the data file that I mentioned in my previous email. Each line contains the information for a single patient and consists of five fields. We want to predict the last field using the other fields. I don't have time to provide any more information about the data since I'm going out of town for a couple of days, but hopefully that won't slow you down too much. And if you don't mind, could we meet when I get back to discuss your preliminary results? I might invite a few other members of my team.

Thanks and see you in a couple of days.

... Introducción (Ejemplo)

Sabes que hay 100 registros. Los primeros que te dan son:

```
012   232   33.5   0    10.7
020   121   16.9   2   210.1
027   165   24.0   0   427.6
...
...
```

... Introducción (Ejemplo)

Sabes que hay 100 registros. Los primeros que te dan son:

```
012  232  33.5  0   10.7
020  121  16.9  2  210.1
027  165  24.0  0  427.6
...
...
```

Est: SO, you got the data for all the patientes?
Min: Yes, I haven't had much time for analysis, but I do have a few intersting results.
Est: Amazing. There were so many data issues with this set of patientes that I couldn't do much.
Min: Oh? I didn' t hear about any possible problems.
Est: Well, first there is field 5, the variable we want to predict. It's common knowledge among people who analyze this type of data that results are better if you work with the log of the values, but I didn't discover this until late. Was it mentioned to you?
Min: No.

Est: But surely you heard about what happened to field 4? It's supposed to be measured on a scale from 1 to 10, with 0 indicating a missing value, but because of a data entry eror, all 10's were changed into 0's. Unfortunately, since some of the patients have missing values for this field, it's impossibl to say whether a 0 in this field is a real 0 or a 10. Quite a few of the records have that problem.

Min: Interesting. Were there any other problems?

Est: Yes, fields 2 and 3 are basically the same, but I assume that you probably noticed that.

Min: Yes, but these fields were only weak predictiors of field 5.

Est: Anyway, giving all those problems. I'm surprised you were able to accomplish anything.

Min: True, but my results are really quite good. Field 1 is a very strong predictor of field 5. I'm surprised that this wasn't noticed before.

Est: What? Field 1 is just an identification number.

Min: Nonetheless, my results speak for themselves.

Est: Oh, no! I just remembered. We assingned ID numbers after we sorted the records based on field 5. There is a strong connec but it's meaningless. Sorry.

Entidades y atributos

- Los conjuntos de datos contienen "objetos" que representan una entidad.
- También se conocen como muestras, ejemplos, ejemplares, instancias.
- Ejemplos:
  - BD ventas: clientes, artículos, ventas.
  - BD médica: pacientes, tratamientos
  - BD universidad: alumnos, profesores, cursos.
- Las entidades se describen por atributos.
- Renglones en la BD $\rightarrow$

Entidades y atributos

- Los conjuntos de datos contienen "objetos" que representan una entidad.
- También se conocen como muestras, ejemplos, ejemplares, instancias.
- Ejemplos:
  - BD ventas: clientes, artículos, ventas.
  - BD médica: pacientes, tratamientos
  - BD universidad: alumnos, profesores, cursos.
- Las entidades se describen por atributos.
- Renglones en la BD $\rightarrow$ entidades.
- Columnas en la BD $\rightarrow$

Entidades y atributos

- Los conjuntos de datos contienen "objetos" que representan una entidad.
- También se conocen como muestras, ejemplos, ejemplares, instancias.
- Ejemplos:
  - BD ventas: clientes, artículos, ventas.
  - BD médica: pacientes, tratamientos
  - BD universidad: alumnos, profesores, cursos.
- Las entidades se describen por atributos.
- Renglones en la BD $\rightarrow$ entidades.
- Columnas en la BD $\rightarrow$ atributos.

Atributos

- Un atributo representa una característica o propiedad de una entidad y puede variar, de una entidad a otra o de un momento a otro. Ejemplo: rfc_cliente, nombre, dirección, color de ojos, temperatura, etc.
- Tipos:
  - Nominal:

Atributos

- Un atributo representa una característica o propiedad de una entidad y puede variar, de una entidad a otra o de un momento a otro. Ejemplo: rfc_cliente, nombre, dirección, color de ojos, temperatura, etc.
- Tipos:
  - Nominal: categorías, estados, o "nombres de cosas"
    - Color de cabello = {rubio, castaño, negro, blanco, rojo, gris}

Atributos

- Un atributo representa una característica o propiedad de una entidad y puede variar, de una entidad a otra o de un momento a otro. Ejemplo: rfc_cliente, nombre, dirección, color de ojos, temperatura, etc.
- Tipos:
  - Nominal: categorías, estados, o "nombres de cosas"
    - Color de cabello $=$ {rubio, castaño, negro, blanco, rojo, gris}
    - Estado civil, ocupación, rfc, código postal
  - Binario:

Atributos

- Un atributo representa una característica o propiedad de una entidad y puede variar, de una entidad a otra o de un momento a otro.
  Ejemplo: rfc_cliente, nombre, dirección, color de ojos, temperatura, etc.
- Tipos:
  - Nominal: categorías, estados, o "nombres de cosas"
    - Color de cabello = {rubio, castaño, negro, blanco, rojo, gris}
    - Estado civil, ocupación, rfc, código postal
  - Binario:
    - Atributo nominal con sólo dos estados.
    - Binario simétrico: Ambos valores son igualmente importantes. Ej: sexo
    - Binario asimétrico: Los valores no tienen la misma importancia. Ej. pruebas médicas (positivo vs. negativo)
      Por convención se asigna 1 al más importante

Atributos

- Un atributo representa una característica o propiedad de una entidad y puede variar, de una entidad a otra o de un momento a otro.
  Ejemplo: rfc_cliente, nombre, dirección, color de ojos, temperatura, etc.
- Tipos:
  - Nominal: categorías, estados, o "nombres de cosas"
    - Color de cabello $=$ {rubio, castaño, negro, blanco, rojo, gris}
    - Estado civil, ocupación, rfc, código postal
  - Binario:
    - Atributo nominal con sólo dos estados.
    - Binario simétrico: Ambos valores son igualmente importantes. Ej: sexo
    - Binario asimétrico: Los valores no tienen la misma importancia. Ej. pruebas médicas (positivo vs. negativo)
      Por convención se asigna 1 al más importante
  - Ordinal:

Atributos

- Un atributo representa una característica o propiedad de una entidad y puede variar, de una entidad a otra o de un momento a otro. Ejemplo: rfc_cliente, nombre, dirección, color de ojos, temperatura, etc.
- Tipos:
  - Nominal: categorías, estados, o "nombres de cosas"
    - Color de cabello = {rubio, castaño, negro, blanco, rojo, gris}
    - Estado civil, ocupación, rfc, código postal
  - Binario:
    - Atributo nominal con sólo dos estados.
    - Binario simétrico: Ambos valores son igualmente importantes. Ej: sexo
    - Binario asimétrico: Los valores no tienen la misma importancia. Ej. pruebas médicas (positivo vs. negativo)
      Por convención se asigna 1 al más importante
  - Ordinal:
    - Los valores tienen orden pero la magnitud entre los valores sucesivos es desconocida.
    - Eje:

Atributos

- Un atributo representa una característica o propiedad de una entidad y puede variar, de una entidad a otra o de un momento a otro.
  Ejemplo: rfc_cliente, nombre, dirección, color de ojos, temperatura, etc.
- Tipos:
  - Nominal: categorías, estados, o "nombres de cosas"
    - Color de cabello = {rubio, castaño, negro, blanco, rojo, gris}
    - Estado civil, ocupación, rfc, código postal
  - Binario:
    - Atributo nominal con sólo dos estados.
    - Binario simétrico: Ambos valores son igualmente importantes. Ej: sexo
    - Binario asimétrico: Los valores no tienen la misma importancia. Ej. pruebas médicas (positivo vs. negativo)
      Por convención se asigna 1 al más importante
  - Ordinal:
    - Los valores tienen orden pero la magnitud entre los valores sucesivos es desconocida.
    - Ejs: tamaño = {pequeño, mediano, grande}, las calificaciones, grados

Atributos numéricos

- Cuantitativo (enteros o reales)
- Intervalo
  - Medido sobre una escala de unidades de igual tamaño
  - Los valores tienen orden. Ej.

Atributos numéricos

- Cuantitativo (enteros o reales)
- Intervalo
  - Medido sobre una escala de unidades de igual tamaño
  - Los valores tienen orden. Ej. temperatura en ^oC o en ^oF, fechas, etc.
- Proporciones (Ratio)
  - Si una medida es de este tipo se puede hablar de un valor como múltiplo (proporción) de otro.
  - Ejemplos: longitudes, conteos (años de experiencia, cantidad de palabras), peso, altura, cantidades monetarias (10 veces más ricos...).

Atributos discretos vs. continuos

Otra forma de clasificar los atributos es en discretos y continuos.

Atributos discretos vs. continuos

Otra forma de clasificar los atributos es en discretos y continuos.

- Atributo discreto:
  - Tiene sólo un conjunto finito de valores.
  - Ej. color de cabello,, profesión, edad, palabras en un texto, fumador?
  - En ocasiones se representan como variables enteras.
- Atributo continuo:
  - Su valor es un número real.
  - Se representa, típicamente, con variables de punto flotante.
  - Ej. temperatura, altura, peso.

Análisis exploratorio de los datos

- Para que el pre-procesamiento de datos sea útil, es esencial tener un panorama general de los mismos.
- Una exploración preliminar de los datos permite conocer sus características:
  - Ayuda a seleccionar la herramienta correcta para el pre-procesamiento o análisis.
  - Permite usar las habilidades de las personas para reconocer patrones.
- Técnicas usadas en EDA
  - Resúmenes estadísticos.
  - Visualización.

Descripciones estadísticas básicas de los datos

- Las medidas estadísticas pueden utilizarse para identificar propiedades de los datos y dar luz, por ejemplo, de cuáles pueden ser ruido o outliers.
- Los resúmenes estadísticos son números que resumen propiedades de los datos:
  - Estas incluyen frecuencia, ubicación y dispersión.
    - Ubicación (medidas de tendencia central): media, mediana, moda, midrange.
    - Dispersión: rangos, cuantiles, varianza, desviación estándar, etc.
  - La mayoría de ellas se calculan en una sola pasada a los datos.

Medidas de tendencia central (Ubicación)

- Media. Sean $x_1, x_2, ..., x_n$ valores u observaciones de un atributo, entonces:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ejemplo:

Medidas de tendencia central (Ubicación)

- Media. Sean $x_1, x_2, ..., x_n$ valores u observaciones de un atributo, entonces:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ejemplo: `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} el promedio es 58, lo que indica que el salario medio es de $58,000.00

- Media aritmética con pesos: si a cada valor x_i se le asocia un peso w_i

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Medidas de tendencia central (Ubicación)

- Media. Sean $x_1, x_2, ..., x_n$ valores u observaciones de un atributo, entonces:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ejemplo: `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} el promedio es 58, lo que indica que el salario medio es de \$58,000.00

- Media aritmética con pesos: si a cada valor x_i se le asocia un peso w_i

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Es la medida más sencilla pero no la mejor forma de medir tendencia central.
- Media recortada (primeado): Se elimina el 2 % de los valores extremos.

... Medidas de tendencia central (Ubicación)

- Mediana: (Para datos asimétricos)
  Valor medio si hay un número impar de valores, o promedio de los dos valores centrales si es un número par de valores.
- Ejemplos:
  - `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la mediana es:

... Medidas de tendencia central (Ubicación)

- Mediana: (Para datos asimétricos)
  Valor medio si hay un número impar de valores, o promedio de los dos valores centrales si es un número par de valores.
- Ejemplos:
  - `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la mediana es: (52+56)/2 = 54. (54,000)

... Medidas de tendencia central (Ubicación)

- Mediana: (Para datos asimétricos)
  Valor medio si hay un número impar de valores, o promedio de los dos valores centrales si es un número par de valores.
- Ejemplos:
  - `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la mediana es: (52+56)/2 = 54. (54,000)
  - ¿si no existiera el 110?

... Medidas de tendencia central (Ubicación)

- Mediana: (Para datos asimétricos)
  Valor medio si hay un número impar de valores, o promedio de los dos valores centrales si es un número par de valores.
- Ejemplos:
  - `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la mediana es: (52+56)/2 = 54. (54,000)
  - ¿si no existiera el 110? mediana = 52,000
  - En rangos de valores:

Edad
1 – 5
6 – 15
16 – 20
21 – 50
51 – 80
81 – 110

Mediana:

... Medidas de tendencia central (Ubicación)

- Mediana: (Para datos asimétricos)
  Valor medio si hay un número impar de valores, o promedio de los dos valores centrales si es un número par de valores.
- Ejemplos:
  - `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la mediana es: (52+56)/2 = 54. (54,000)
  - ¿si no existiera el 110? mediana = 52,000
  - En rangos de valores:

Edad
1 – 5
6 – 15
16 – 20
21 – 50
51 – 80
81 – 110

Mediana: el intervalo 21 –50

... Medidas de tendencia central

- Frecuencia: Porcentaje de veces que el valor ocurre en un data set.
- Moda
  - Valor que aparece con más frecuencia en los datos.
  - Unimodal, bimodal, trimodal
  - Ejemplo:

... Medidas de tendencia central

- Frecuencia: Porcentaje de veces que el valor ocurre en un data set.
- Moda
  - Valor que aparece con más frecuencia en los datos.
  - Unimodal, bimodal, trimodal
  - Ejemplo: `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la moda es

... Medidas de tendencia central

- Frecuencia: Porcentaje de veces que el valor ocurre en un data set.
- Moda
  - Valor que aparece con más frecuencia en los datos.
  - Unimodal, bimodal, trimodal
  - Ejemplo: `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la moda es 52 y 70, es decir es bimodal.
  - ¿Cuál es la moda si todos los aparecen una sola vez?

... Medidas de tendencia central

- Frecuencia: Porcentaje de veces que el valor ocurre en un data set.
- Moda
  - Valor que aparece con más frecuencia en los datos.
  - Unimodal, bimodal, trimodal
  - Ejemplo: `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la moda es 52 y 70, es decir es bimodal.
  - ¿Cuál es la moda si todos los aparecen una sola vez? ninguna.
- La frecuencia y la moda se aplican sobre datos categóricos.
- Rango:

... Medidas de tendencia central

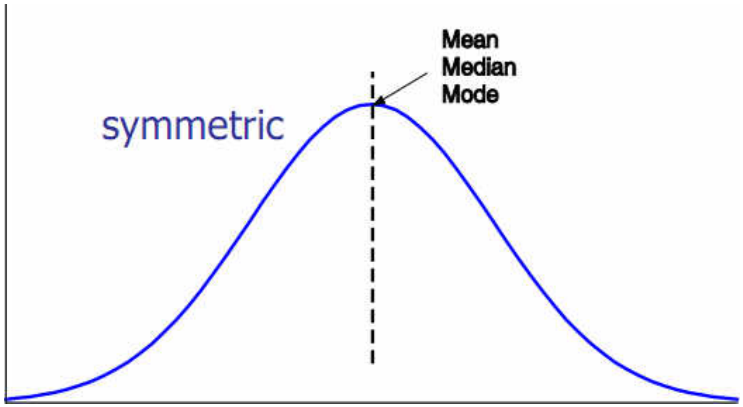
- Frecuencia: Porcentaje de veces que el valor ocurre en un data set.
- Moda
  - Valor que aparece con más frecuencia en los datos.
  - Unimodal, bimodal, trimodal
  - Ejemplo: `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la moda es 52 y 70, es decir es bimodal.
  - ¿Cuál es la moda si todos los aparecen una sola vez? ninguna.
- La frecuencia y la moda se aplican sobre datos categóricos.
- Rango: es la diferencia entre el valor mayor y el menor.
- *midrange* = promedio del mayor y menor valores en un conjunto.
  - Ejemplo: en el salario es

... Medidas de tendencia central

- Frecuencia: Porcentaje de veces que el valor ocurre en un data set.
- Moda
  - Valor que aparece con más frecuencia en los datos.
  - Unimodal, bimodal, trimodal
  - Ejemplo: `sueldo` = {30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110} la moda es 52 y 70, es decir es bimodal.
  - ¿Cuál es la moda si todos los aparecen una sola vez? ninguna.
- La frecuencia y la moda se aplican sobre datos categóricos.
- Rango: es la diferencia entre el valor mayor y el menor.
- *midrange* = promedio del mayor y menor valores en un conjunto.
  - Ejemplo: en el salario es $(30+110)/2 = 70$.

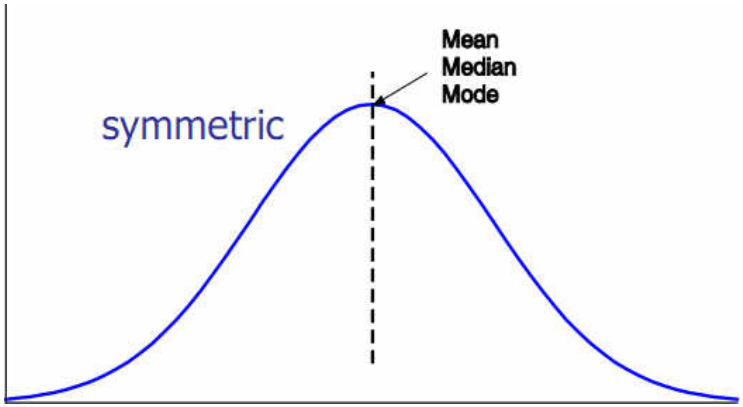
Desviación simétrica vs. sesgada de los datos

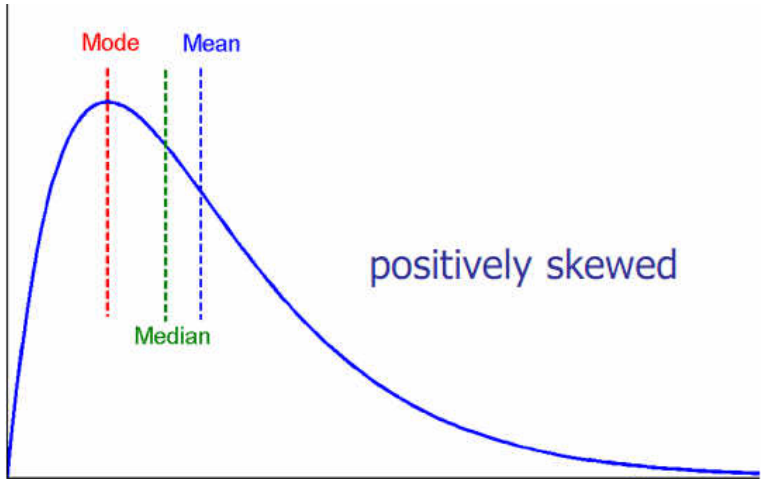
Mediana, media y moda de datos con sesgo simétrico, positivo y negativo.



Desviación simétrica vs. sesgada de los datos

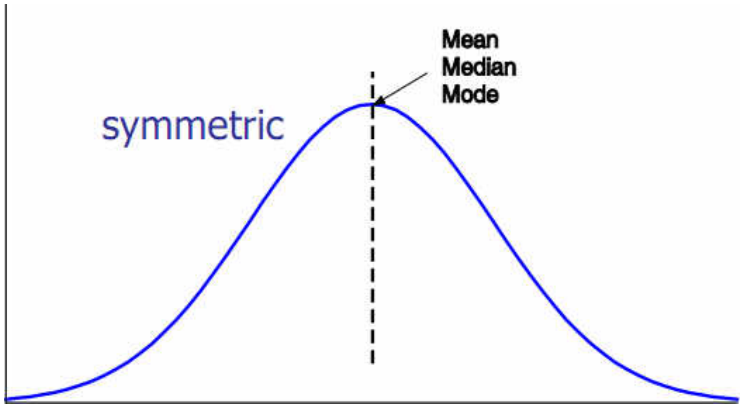
Mediana, media y moda de datos con sesgo simétrico, positivo y negativo.

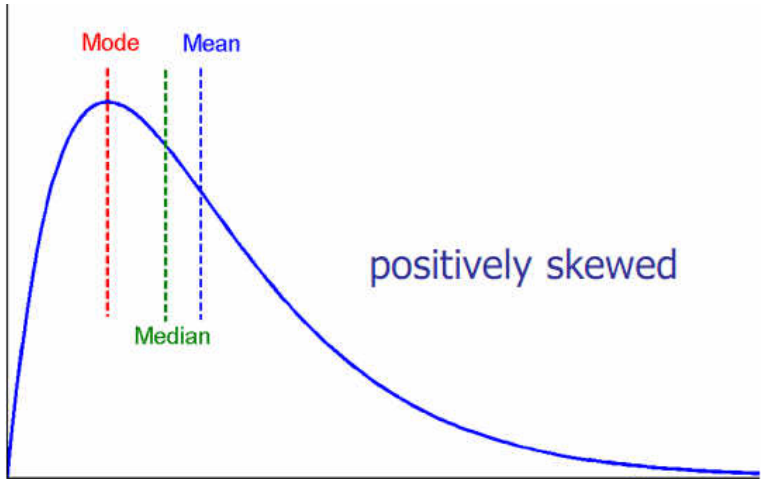


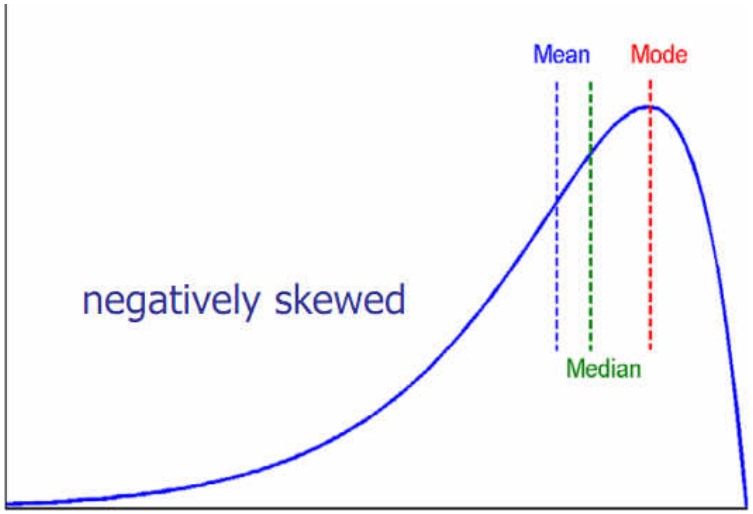


Desviación simétrica vs. sesgada de los datos

Mediana, media y moda de datos con sesgo simétrico, positivo y negativo.





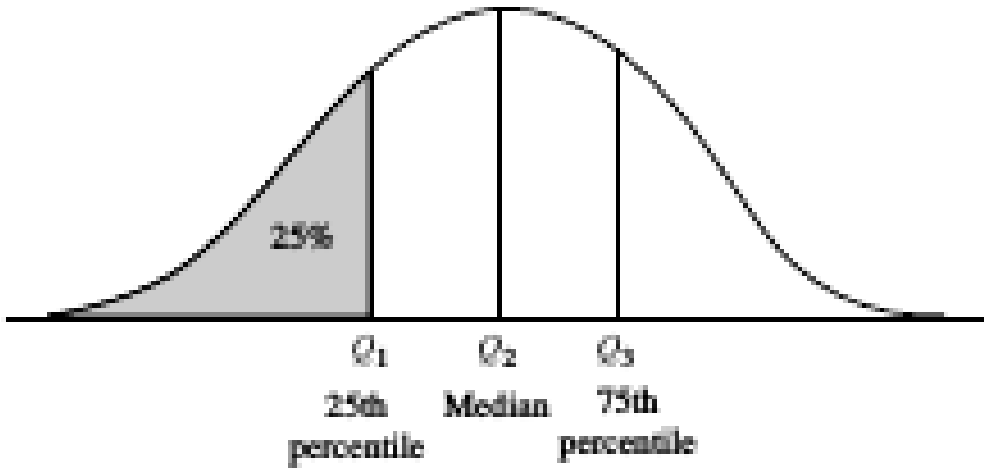


Medidas de dispersión de los datos

- Cuantiles son puntos tomados en intervalos regulares que dividen la muestra (ordenada) en conjuntos consecutivos de igual tamaño:
  - Percentiles: son 99 valores que dividen en cien partes iguales el conjunto de datos ordenados.
  - Deciles: son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales.
  - Cuartiles: son los tres valores que dividen al conjunto de datos ordenados en cuatro partes iguales:
    - El primer cuartil Q1 es el menor valor que es mayor que una cuarta parte de los datos
    - El segundo cuartil Q2 (la mediana), es el menor valor que es mayor que la mitad de los datos
    - El tercer cuartil Q3 es el menor valor que es mayor que tres cuartas partes de los datos

... Medidas de dispersión de datos (Cuartiles)

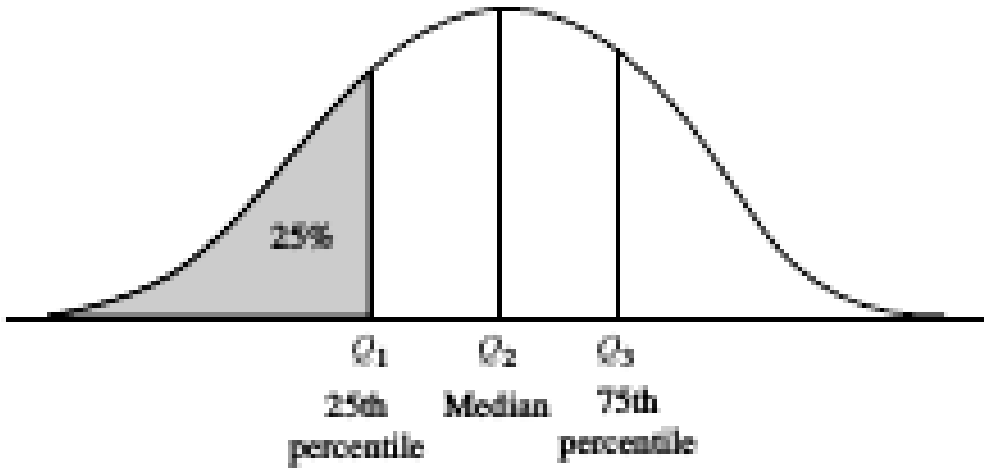
- Los cuartiles dividen la distribución en cuatro subconjuntos consecutivos de tamaño igual.
- El segundo cuartil corresponde a la mediana.



- La distancia entre el primer y el tercer cuartil se conoce como rango inter-cuartiles (IQR) $IQR = Q_3 - Q_1$.
- Ejemplo: $\{30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110\}$ se tiene que $Q_1 =$

... Medidas de dispersión de datos (Cuartiles)

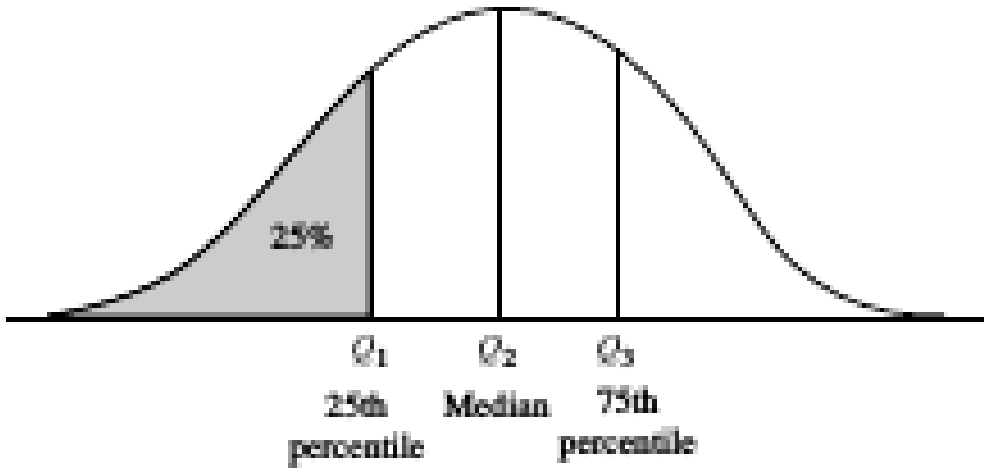
- Los cuartiles dividen la distribución en cuatro subconjuntos consecutivos de tamaño igual.
- El segundo cuartil corresponde a la mediana.



- La distancia entre el primer y el tercer cuartil se conoce como rango inter-cuartiles (IQR) $IQR = Q_3 - Q_1$.
- Ejemplo: $\{30,\ 36,\ 47,\ 50,\ 52,\ 52,\ 56,\ 60,\ 63,\ 70,\ 70,\ 110\}$ se tiene que $Q_1 = 47$, $Q_3 =$

... Medidas de dispersión de datos (Cuartiles)

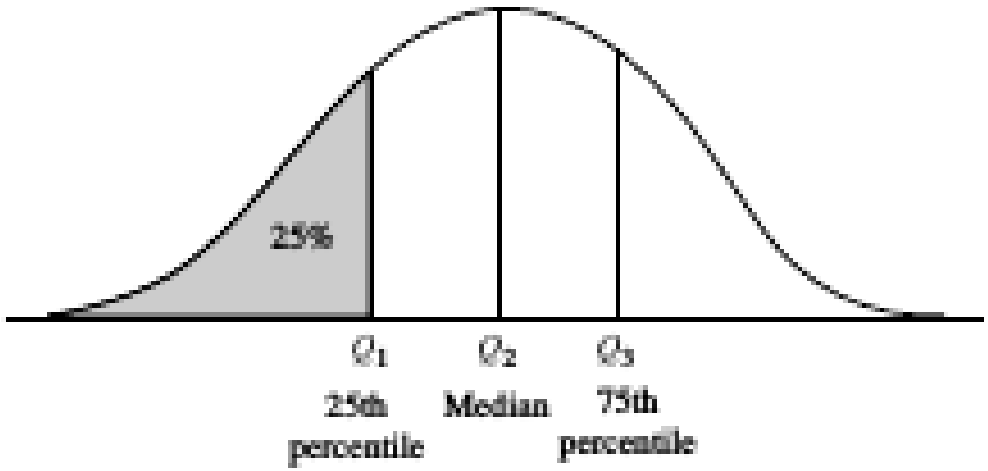
- Los cuartiles dividen la distribución en cuatro subconjuntos consecutivos de tamaño igual.
- El segundo cuartil corresponde a la mediana.



- La distancia entre el primer y el tercer cuartil se conoce como rango inter-cuartiles (IQR) $IQR = Q_3 - Q_1$.
- Ejemplo: $\{30,\ 36,\ 47,\ 50,\ 52,\ 52,\ 56,\ 60,\ 63,\ 70,\ 70,\ 110\}$ se tiene que $Q_1 = 47$, $Q_3 = 63$, $IQR =$

... Medidas de dispersión de datos (Cuartiles)

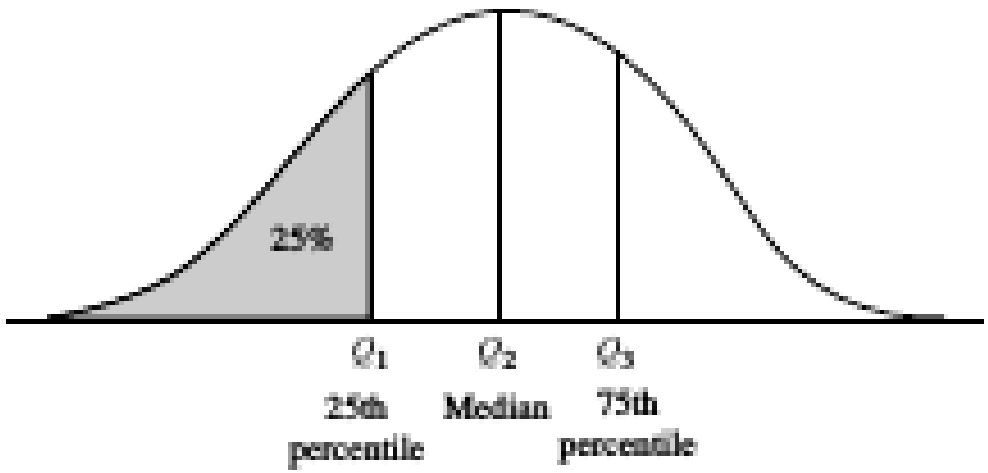
- Los cuartiles dividen la distribución en cuatro subconjuntos consecutivos de tamaño igual.
- El segundo cuartil corresponde a la mediana.



- La distancia entre el primer y el tercer cuartil se conoce como rango inter-cuartiles (IQR) $IQR = Q_3 - Q_1$.
- Ejemplo: $\{30,\ 36,\ 47,\ 50,\ 52,\ 52,\ 56,\ 60,\ 63,\ 70,\ 70,\ 110\}$ se tiene que $Q_1 = 47$, $Q_3 = 63$, $IQR = 63$ - $47 = 16$

... Medidas de dispersión de datos (Cuartiles)

- Los cuartiles dividen la distribución en cuatro subconjuntos consecutivos de tamaño igual.
- El segundo cuartil corresponde a la mediana.



- La distancia entre el primer y el tercer cuartil se conoce como rango inter-cuartiles (IQR) $IQR = Q_3 - Q_1$.
- Ejemplo: $\{30,\ 36,\ 47,\ 50,\ 52,\ 52,\ 56,\ 60,\ 63,\ 70,\ 70,\ 110\}$ se tiene que $Q_1 = 47$, $Q_3 = 63$, $IQR = 63$ - $47 = 16$ $Q_2 = 52$.

... Medidas de dispersión de datos

- Varianza para n valores del atributo X:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Desviación estándar σ es la raíz cuadrada de la varianza.
  - Permite conocer la desviación que presentan los datos en su distribución respecto de la media.
  - La σ de un grupo repetido de medidas da la precisión de éstas.
- Propiedades:
  - Siempre es ≥ 0. Si es cero entonces no hay dispersión es decir, todos los datos son iguales.
  - Una baja desviación estándar significa que los datos tienden a estar muy cerca de la media.
  - Una alta indica que los datos están dispersos sobre un rango grande de valores.

... Medidas de dispersión de datos

- Varianza para n valores del atributo X:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

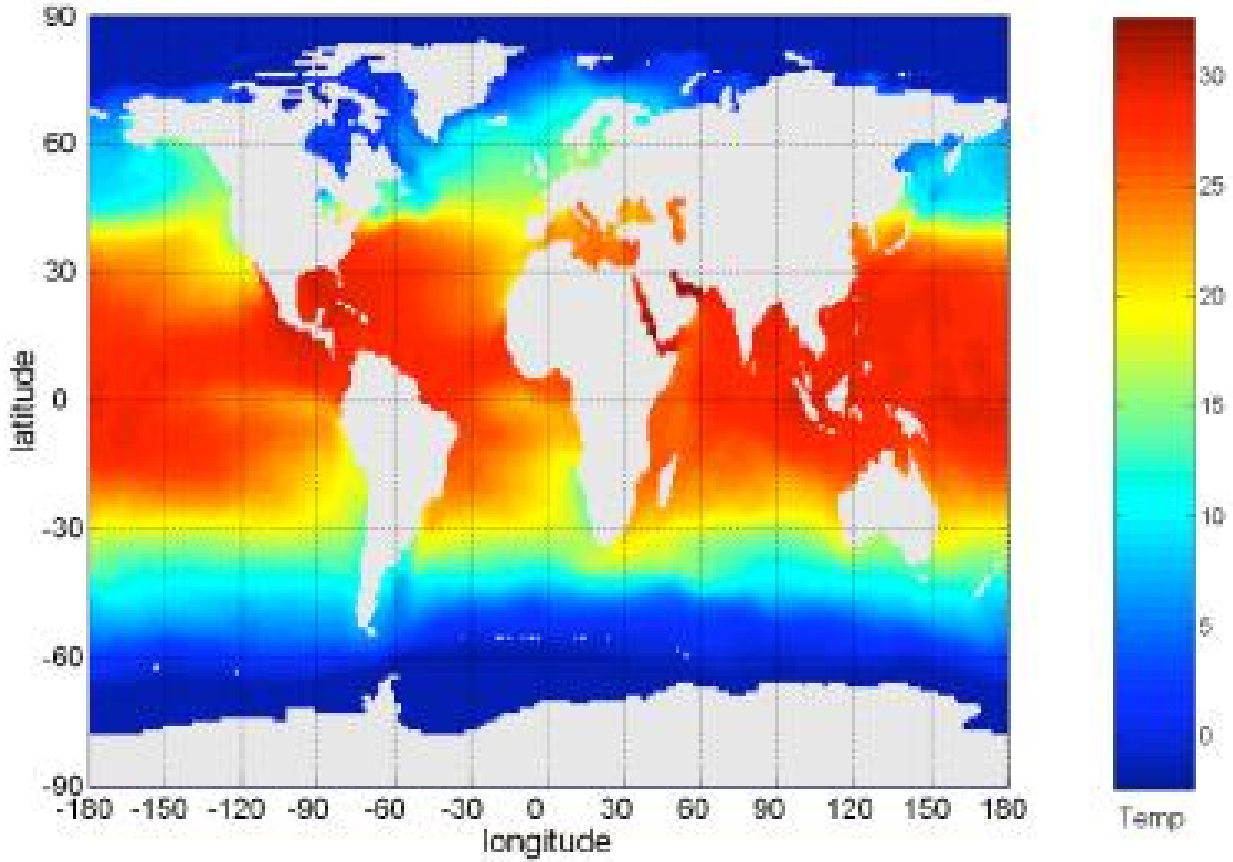
- Desviación estándar σ es la raíz cuadrada de la varianza.
  - Permite conocer la desviación que presentan los datos en su distribución respecto de la media.
  - La σ de un grupo repetido de medidas da la precisión de éstas.
- Propiedades:
  - Siempre es ≥ 0. Si es cero entonces no hay dispersión es decir, todos los datos son iguales.
  - Una baja desviación estándar significa que los datos tienden a estar muy cerca de la media.
  - Una alta indica que los datos están dispersos sobre un rango grande de valores.
- Ej. (0, 0, 14, 14), (0, 6, 8, 14) y (6, 6, 8, 8) cada muestra tiene media = 7. Las desviaciones estándar son 8.08, 5.77 y 1.15 respectivamente.

Visualización

- Visualización es la conversión de datos en un formato gráfico o tabular de tal manera que las características de los datos y sus relaciones puedan analizarse.
- Es una de las más poderosas y atractivas técnicas para exploración de datos.
  - Una gráfica dice más que mil palabras.
  - Los humanos tenemos habilidad para analizar grandes cantidades de información presentada visualmente.
  - Podemos detectar patrones generales y tendencias.
  - Podemos detectar patrones inusuales.

... Visualización

- Ejemplo: Temperatura de la Superficie Marina en Julio de 1982.



Principales técnicas de visualización

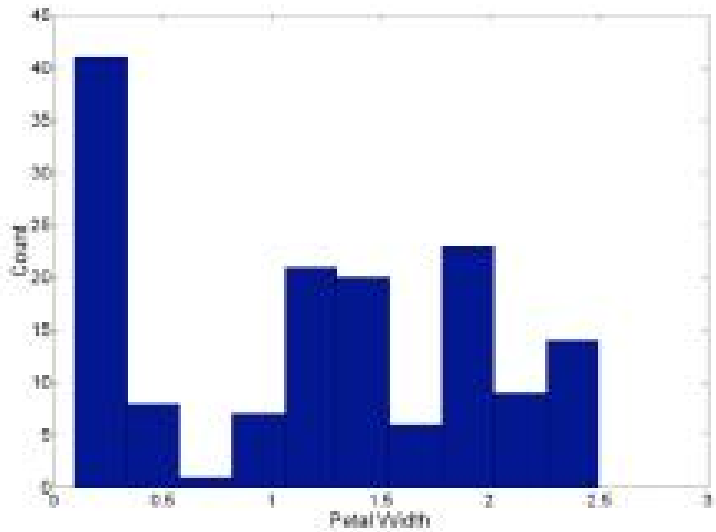
- Histograma: frecuencias.
- Cajas de bigotes: despliegue gráfico de los 5 números.
- Gráfica de cuantiles: cada valor X_i es emparejado con f_i, indicando que aproximadamente $100f_i\,\%$ de datos son $\leq x_i$.
- Gráfica cuantil-cuantil (q-q): gráfica los cuantiles de una distribución univariante contra los correspondientes cuantiles de otra.
- Gráfica de dispersión : cada par de valores es un par de coordenadas y se gráfica como puntos en el plano.

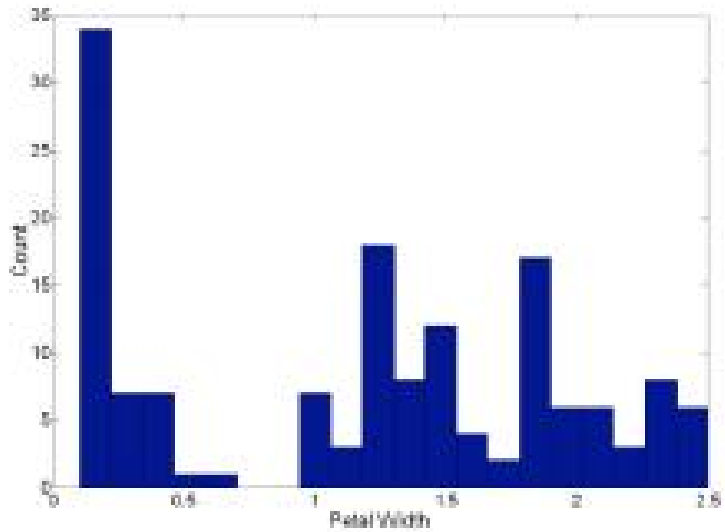
Histogramas

- Histograma. Gráfica que muestra las frecuencias de valores como barras.
- Usualmente muestra la distribución de valores de una sola variable.
- Divide los valores en cubetas y grafica una barra de la cantidad de objetos en cada cubeta. Típicamente del mismo ancho.
- La altura de cada barra indica el número de objetos.
- La figura del histograma depende del número de cubetas.

Histogramas

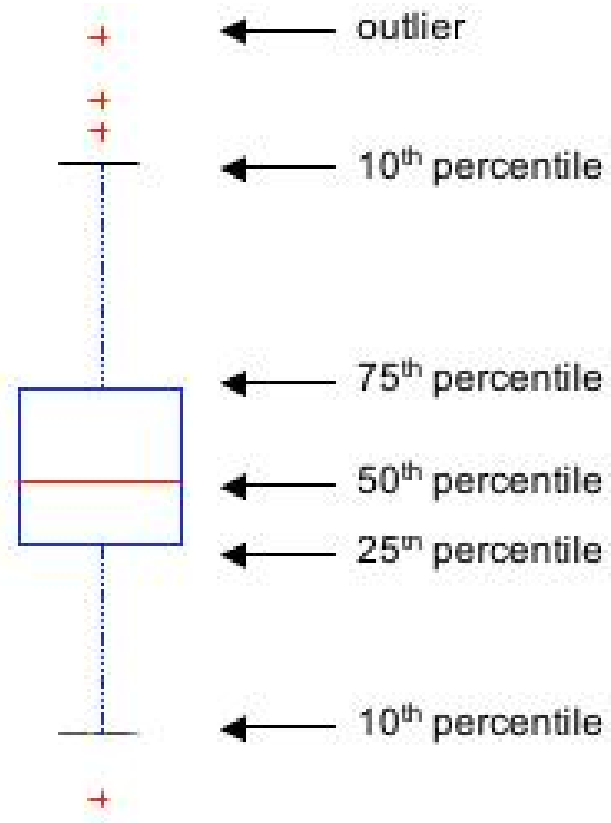
- Histograma. Gráfica que muestra las frecuencias de valores como barras.
- Usualmente muestra la distribución de valores de una sola variable.
- Divide los valores en cubetas y grafica una barra de la cantidad de objetos en cada cubeta. Típicamente del mismo ancho.
- La altura de cada barra indica el número de objetos.
- La figura del histograma depende del número de cubetas.
- Ejemplo: Ancho de pétalos de cierto lirio (10 y 20 cubetas).





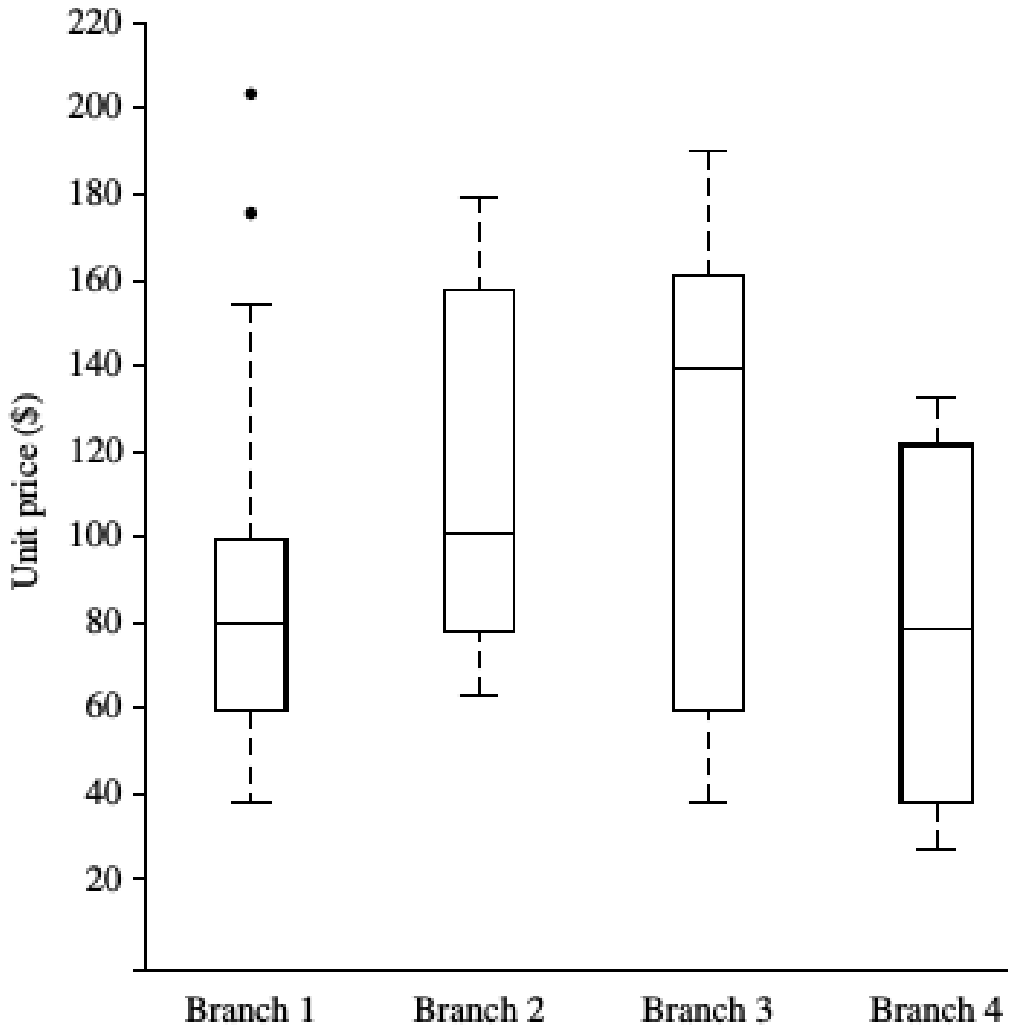
Visualización (Box plots)

- Inventadas por J. Tukey
- Una forma de desplegar la distribución de los datos.

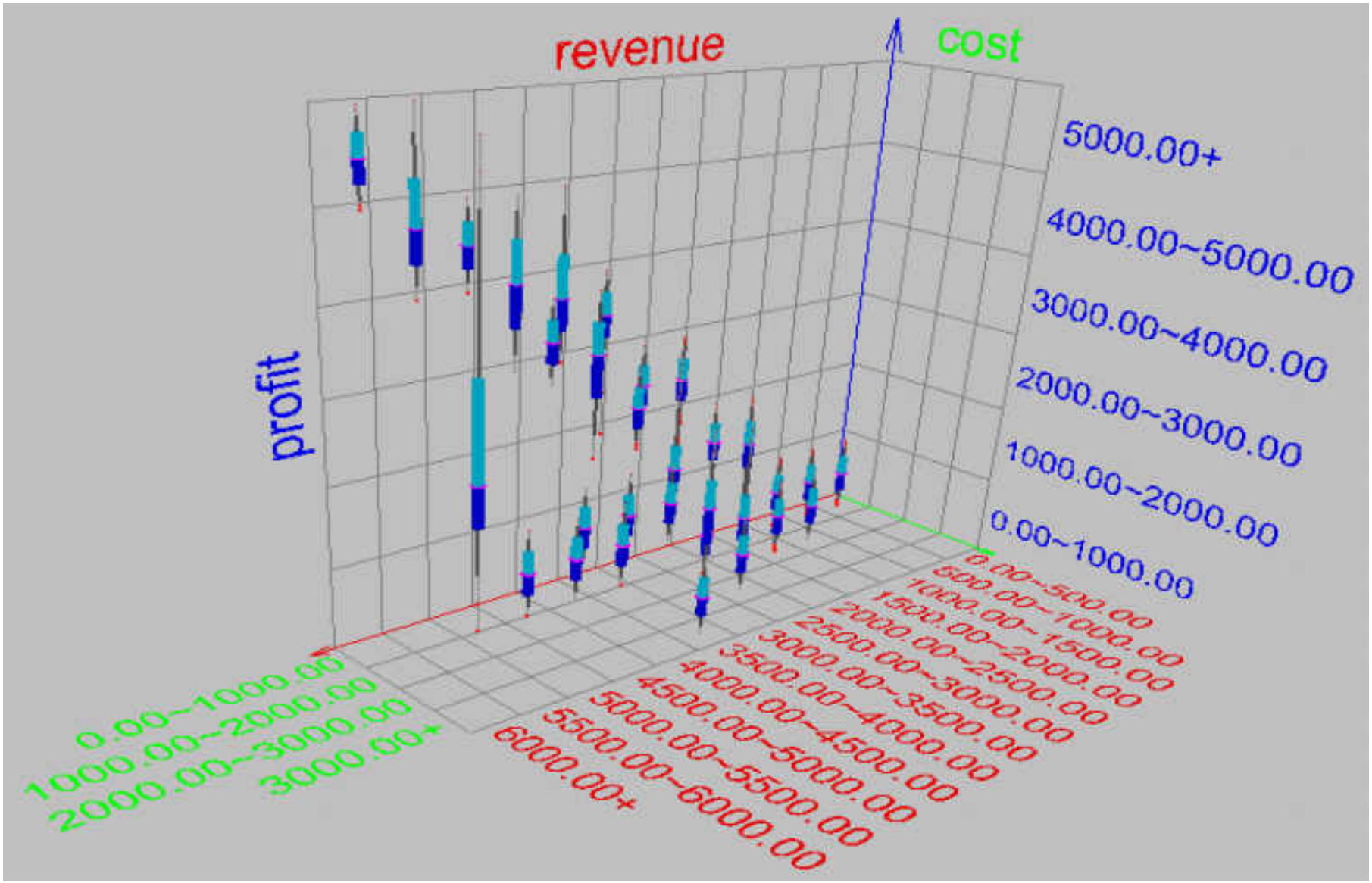


... Visualización (Boxplot)

Ejemplo:

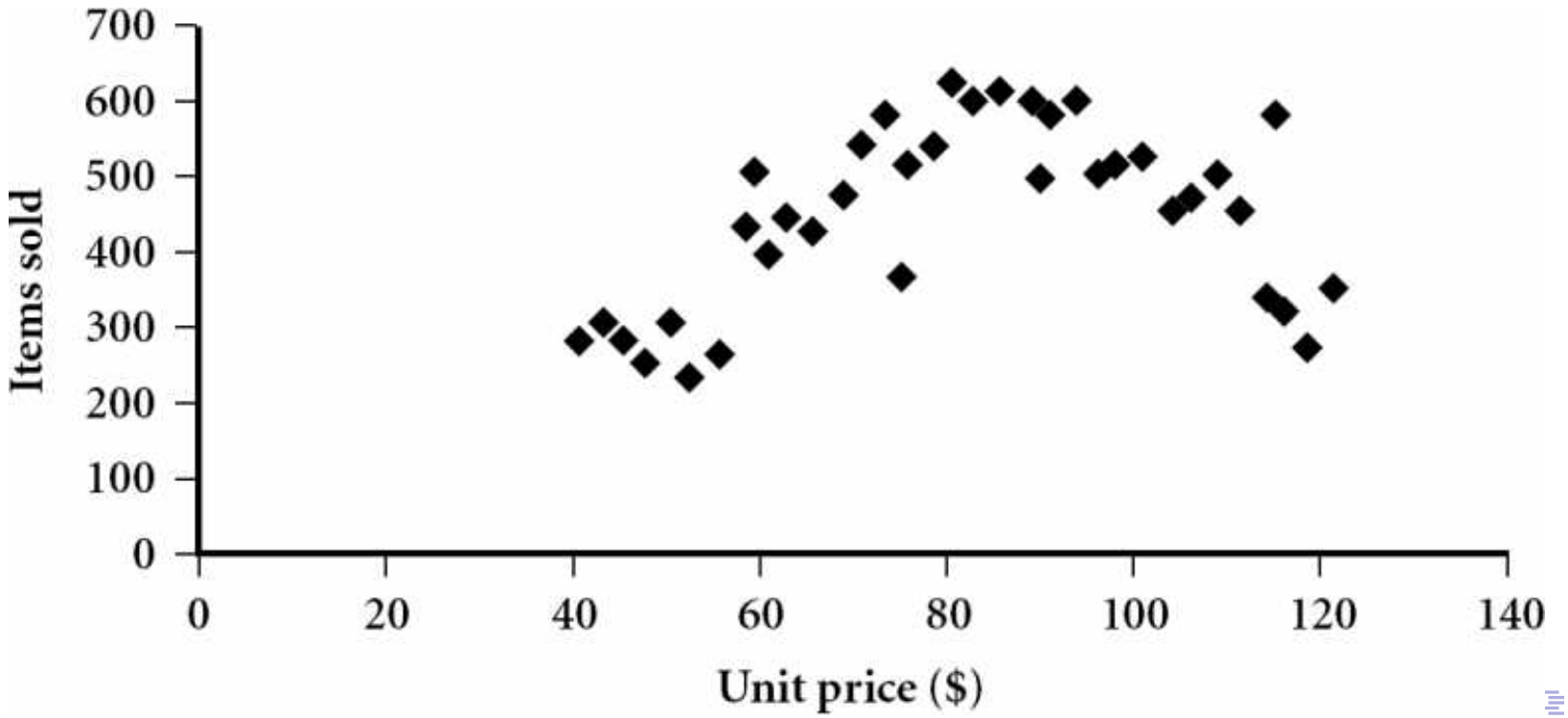


Visualización de dispersión de datos: cajas en 3D



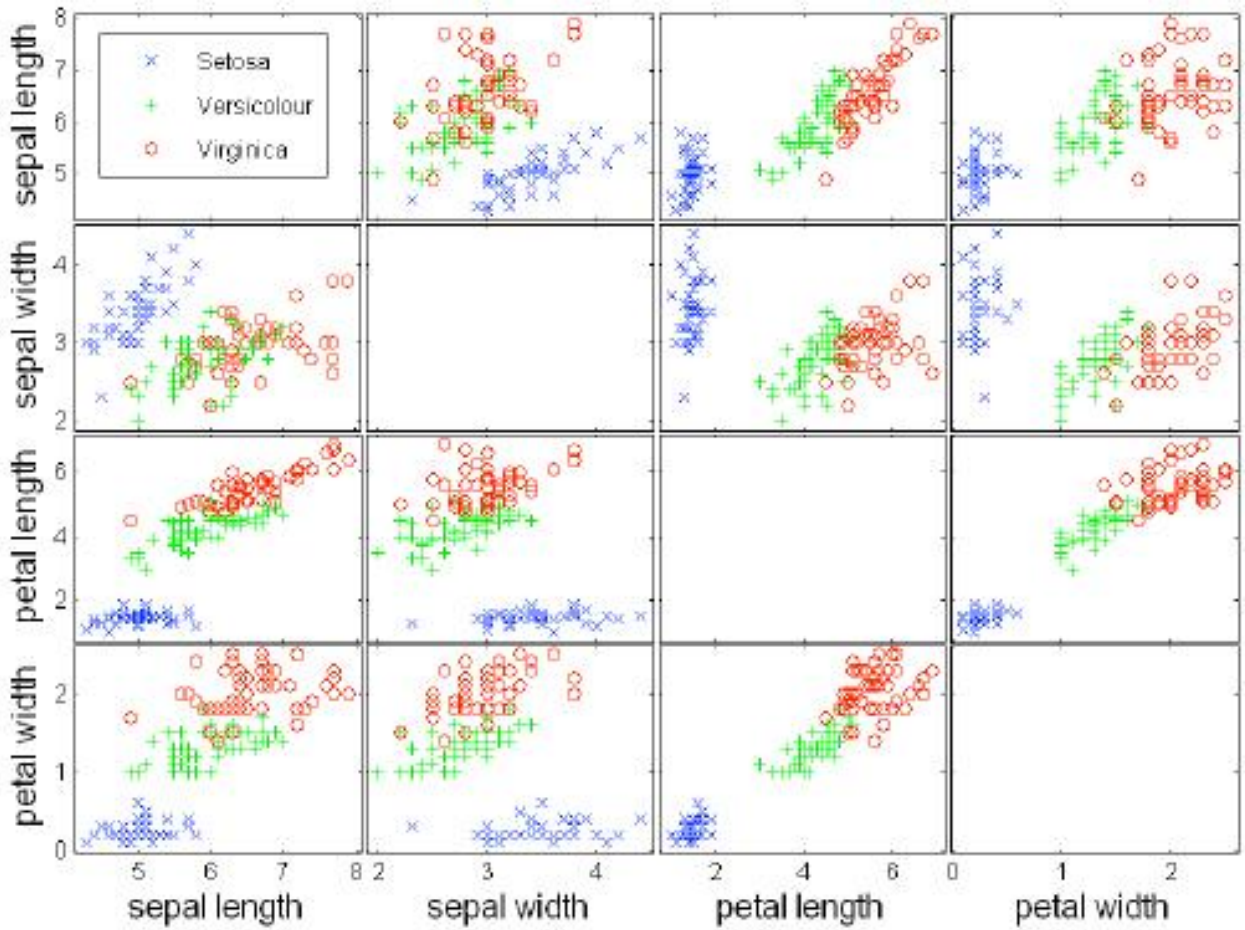
Gráfica de dispersión (scatter plot)

- Proporciona una primera mirada a datos bivariantes para ver agrupamientos de puntos, outliers, correlaciones, etc.
- Los valores de los atributos determinan la posición.
- Cada par de valores se trata como un par de coordenadas y se grafica como un punto en el plano.
- Los datos deben ser númericos.



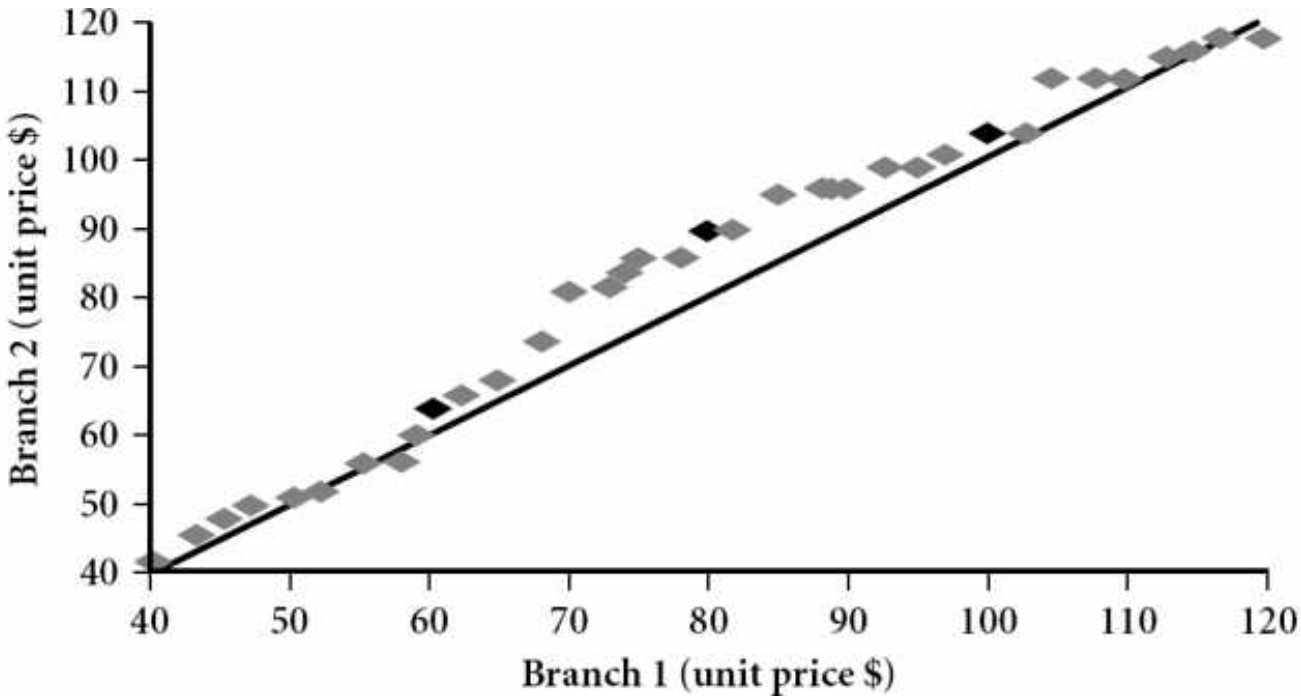
Gráfica de dispersión (scatter plot)

Es útil tener arreglos de estas gráficas debido a que pueden mostrar relaciones entre varios pares de atributos.

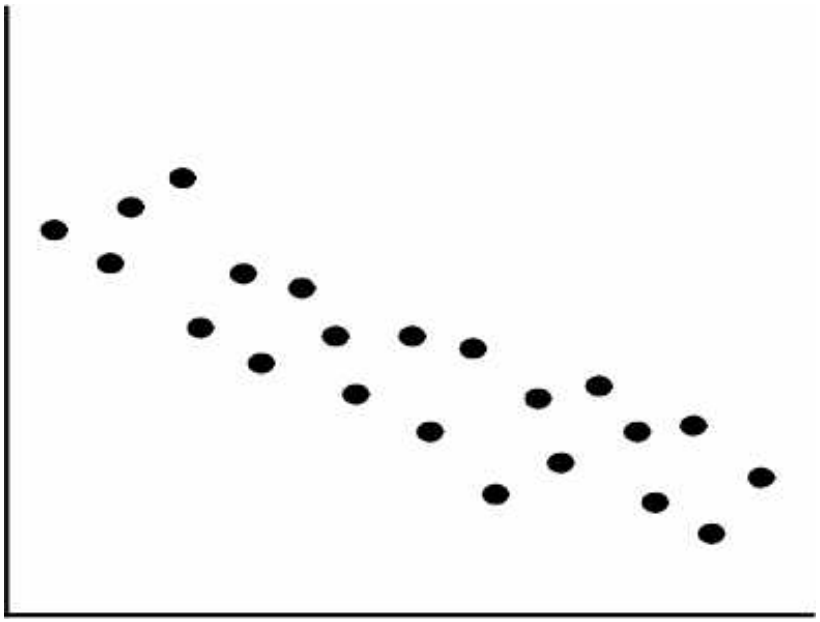


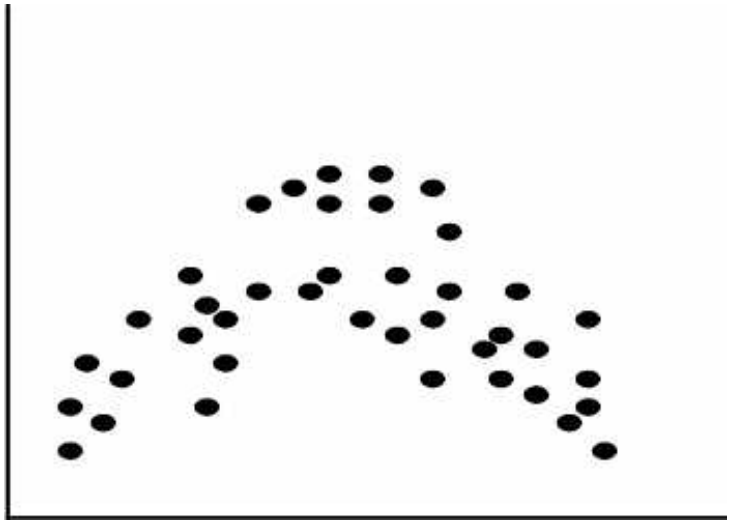
Gráfica cuantil-cuantil (q-q)

- Grafican los cuantiles de una distribución univariante contra los correspondientes cuantiles de otra.
- Permiten visualizar si hay un cambio de una distribución a otra.
- Ejemplo: mostrar el precio unitario de artículos vendidos en la sucursal 1 vs. la sucursal 2 para cada cuantil. El precio unitario de los artículos vendidos en la sucursal 1 tiende a ser menor que los de la 2.



Datos correlacionados





Datos no correlacionados