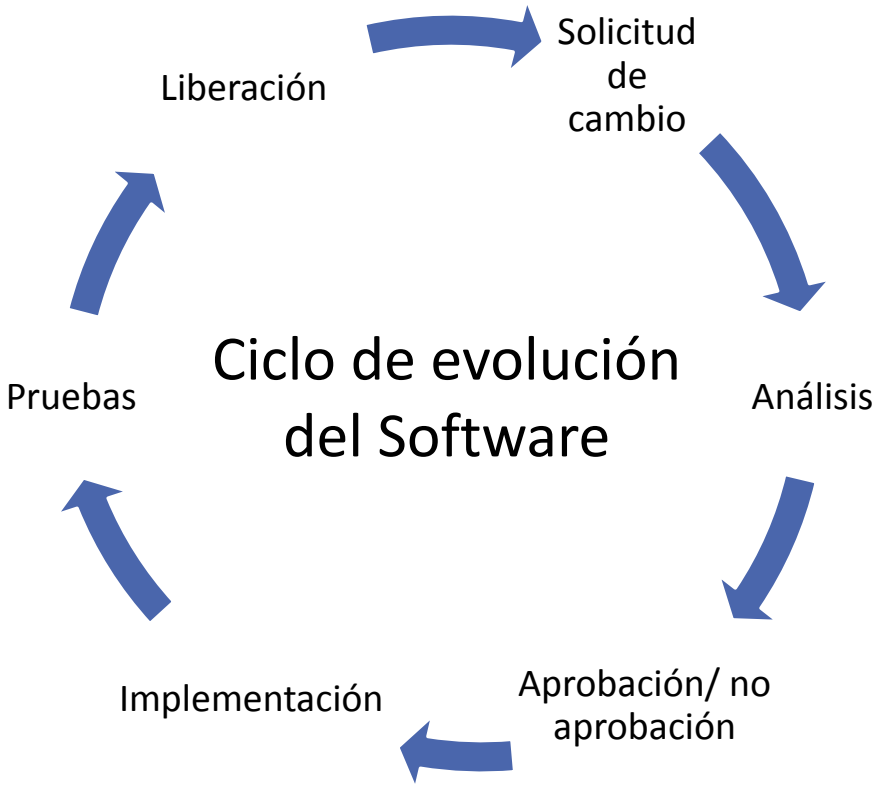
3. Elementos básicos de la investigación empírica

Evolución del Software



Evolución del Software

- Gran parte del costo en un proyecto de software es consumido por la realización de cambios, más que el desarrollo en sí mismo.
- La preocupación principal es mantener la flexibilidad y calidad del software
- Predecir:
  - Defectos
  - Cambios
  - Esfuerzo
  - Costo

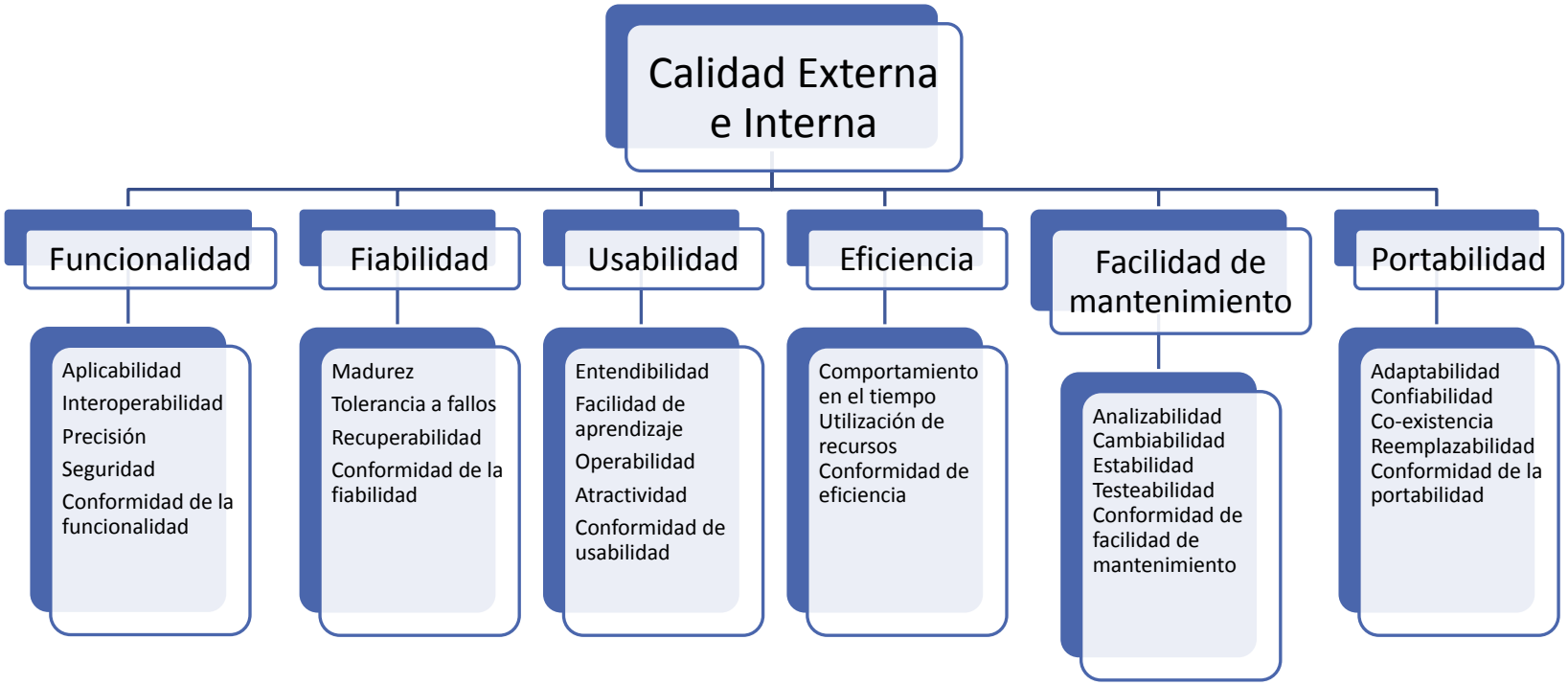
¿Por qué interesa poder precedir?

Calidad del Software

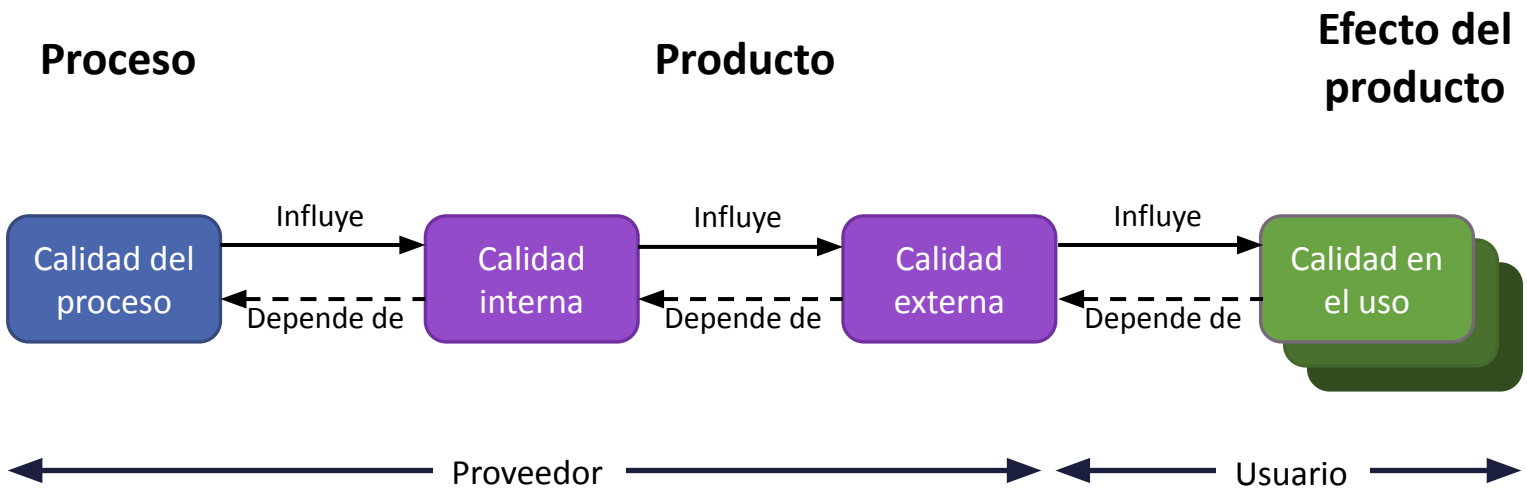
- ¿Qué es?

- Visión:
  - **Del Usuario** - adecuación al uso
  - **Del Productor** - adecuación a las especificaciones
  - **Del Producto** - características específicas
    - comportamiento externo (visible para todos)
      - características internas (normalmente sólo visibles al productor)

Atributos de calidad del software (producto)



Impacto de los diferentes aspectos de calidad

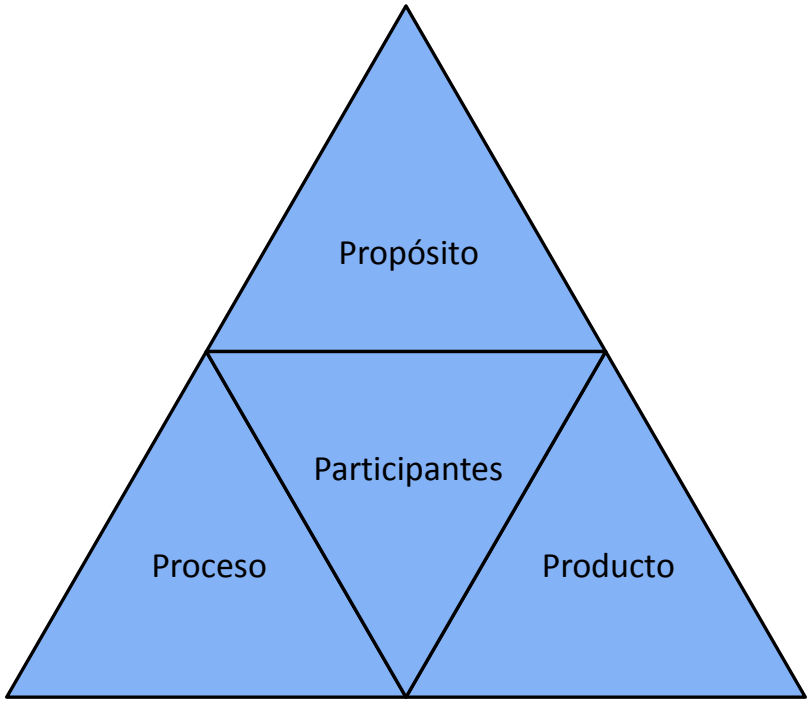


Propósito de investigación y aspectos de calidad del software

- Dentro de las preguntas de investigación que se trabajan en Ingeniería de Software, la gran mayoría involucra aspectos de calidad del software

- Objetivo: conocer, mejorar, predecir, controlar, etc.

Elementos básicos de la investigación empírica



Enfoques de investigación

- **Enfoque cualitativo**: Enfoque de investigación basado en recopilar y analizar datos no numéricos. Principalmente compuestos por:
  - Texto
  - Gráficas
  - Imágenes
- En general ayuda a organizar los datos en forma de "categorías"
  - ¿Qué hace que un sistema sea de buena calidad?
    - Amigabilidad de interfaz de usuario, tiempo de respuesta, confiabilidad, seguridad, recuperación ante fallas.
  - ¿Cuál es su grado de satisfacción con el sistema?
    - Excelente, muy bueno, bueno, intermedio, malo, muy malo.

Enfoques de investigación (cont)

- **Enfoque cuantitativo**: Busca encontrar una relación numérica entre dos o más grupos. Se basa en cuantificar una relación o comparar variables. Los datos son cuantitativos son valores numéricos (continuos o discretos)
- Ejemplos:
  - Líneas de código
  - Cantidad de defectos
  - Costo (horas trabajadas/dinero)
- Dependiendo del enfoque (cualitativo/cuantitativo) son los tipos de análisis que se pueden realizar

Tipos de investigación

- **Descriptiva/Exploratoria**: provee una descripción de los conceptos y los hechos tal cual son observados en una determinada realidad
- **Correlacional**: provee el grado de relación entre dos variables (o más)
- **Causa-efecto (explicativo)**: explica la relación entre las variables, en donde se encuentra un efecto entre una y otra.
  - Son más estructuradas que las dos anteriores

Ejemplo de investigación causa-efecto

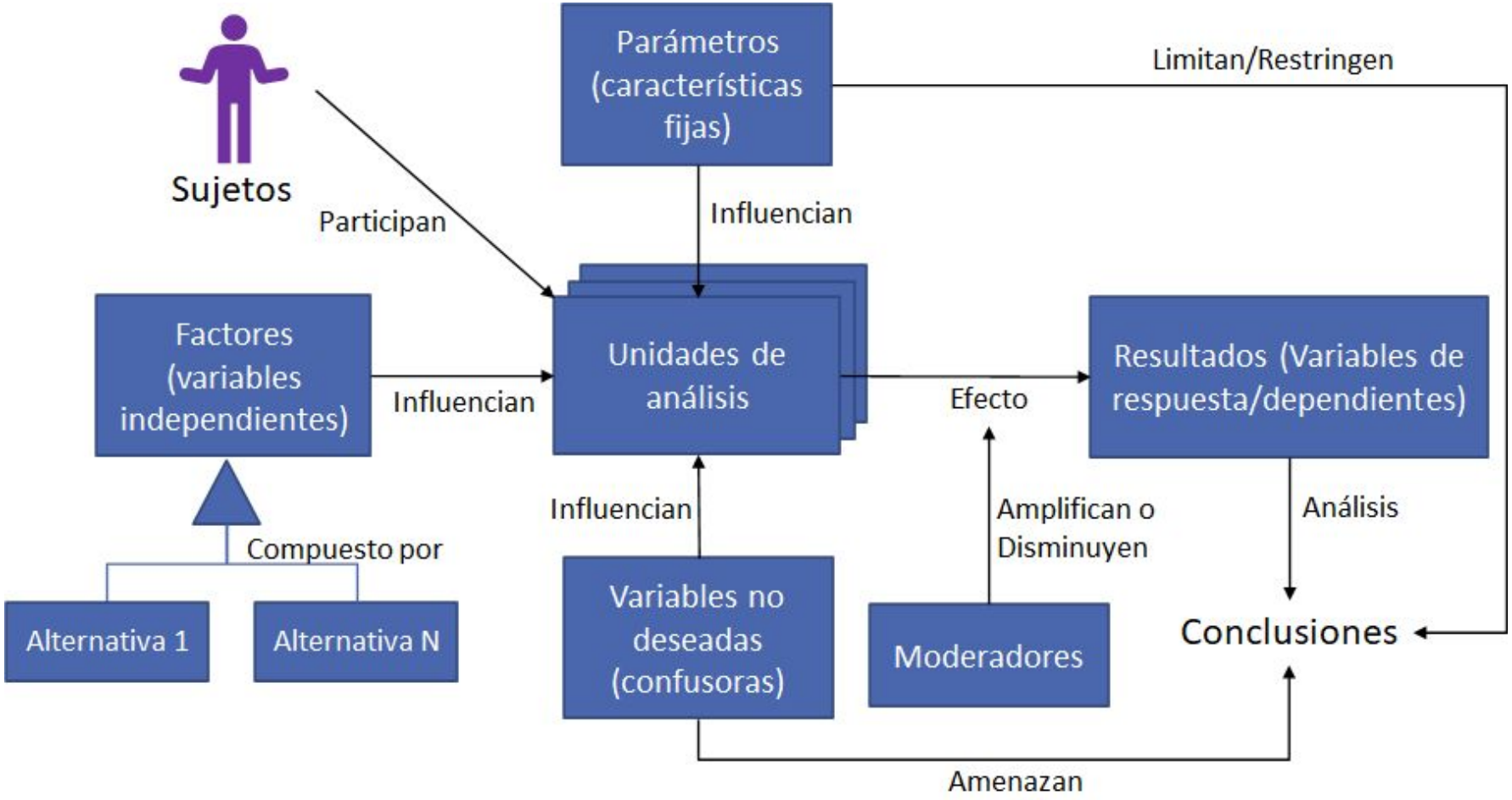
Correlación: ***"Si el volumen de un gas es constante, a un incremento en la temperatura, ha de seguir un incremento en la presión"***

- Variables involucradas: **volumen, temperatura, presión del gas**
- Si se conoce el volumen y la temperatura: ¿Podemos predecir qué va a pasar con la presión del gas?
- ¿Por qué aumenta la presión? (valor explicativo, relación causa-efecto)
  - La temperatura se incrementa y el volumen del gas se mantuvo constante

Ejemplo de investigación causa-efecto (cont)

- Explicación del fenómeno (más elaborada):
  1. "Un incremento de la temperatura aumenta la energía cinética de las moléculas del gas".
  2. "El incremento de la energía cinética causa un aumento en la velocidad del movimiento de las moléculas".
  3. "Puesto que las moléculas no pueden ir más allá del recipiente con volumen constante, éstas impactan con mayor frecuencia la superficie interior del recipiente. (Debido a que se desplazan más rápido, cubren más distancia y rebotan en el recipiente más frecuentemente.)".
  4. "En la medida en que las moléculas impactan los costados del recipiente con mayor frecuencia, la presión sobre las paredes del recipiente se incrementa".

Componentes de un estudio empírico



Nota: no todos los elementos aplican a todos los estudios empíricos

Variables dependientes e independientes

- Variables de estudio: dependientes e independientes
  - Variables independientes (predictoras): son variables **de entrada** las cuales son manipuladas o controladas por el investigador
  - Variables dependientes (de respuesta): son variables **de salida**, resultado del análisis del efecto de las variables independientes

Variables confusoras/intervinientes (no deseadas)

- Son variables que afectan a la variable de respuesta, pero no son manipuladas ni controladas por el investigador (no forman parte del estudio)
- Ejemplo:
  - Variable A: alimentación que se recibe en la infancia (variable independiente).
  - Variable B: nivel de inteligencia posterior de la persona (variable dependiente).
  - Variable C: nivel socio-económico (variable confusora que influye a A).
- Un estudio bien diseñado **intenta asegurar** (en muchos casos no es posible) que el efecto sobre la variable dependiente sólo puede atribuirse a la variable independiente (manipulada) y no a variables confusoras (o no controladas).
- Existen técnicas para bloquear o aislar el efecto que estas variables producen

Conceptos de estadística

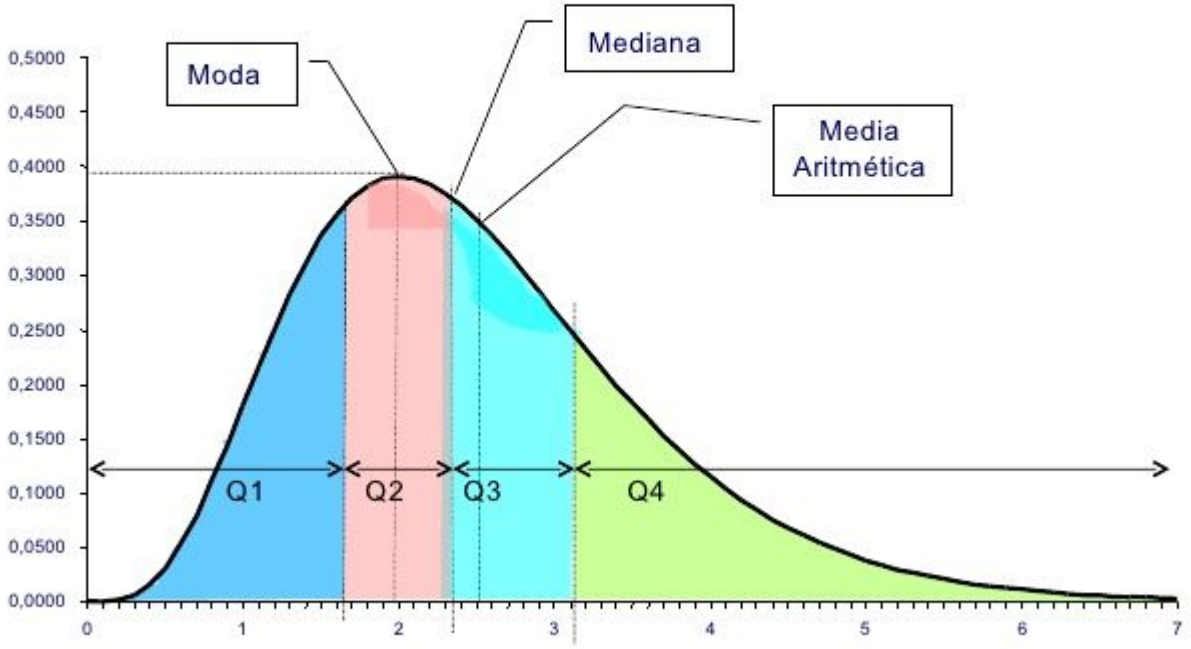
- Estadística descriptiva
  - Medidas de tendencia central
  - Medidas de dispersión
- Reducción del conjunto de datos
- Tipos de métodos de análisis estadísticos
  - Métodos paramétricos
  - Métodos no paramétricos

Estadística Descriptiva

- Analiza y **caracteriza** un conjunto de datos con el objetivo de describir las características y comportamientos de este conjunto mediante medidas de resumen, tablas o gráficos.
- Se utilizan para **validar la correctitud** de los datos recolectados antes de comenzar con el análisis estadístico
- Se deben depurar (o “reducir”) de tal forma que puedan ser leídos fácilmente y se puedan utilizar para el posterior análisis estadístico
- Tipos de medidas básicas:
  - **De tendencia central**
  - **Dispersión**

Medidas de tendencia central

- Indican el “medio” de un conjunto de datos



Medidas de tendencia central (cont.)

- **Media aritméti** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Es el promedio.
  - Se calcula sumando todas las muestras y dividiendo el total por el número de muestras
  - Es significativa para las escalas de *intervalo y ratio*

- **Mediana**: $\tilde{x}$
  - Representa el valor medio de un conjunto de datos, tal que el número de muestras que son mayores que la mediana es el mismo que el número de muestras que son menores que la mediana.
  - Se calcula ordenando las muestras en orden ascendente o descendente, y seleccionando la observación del medio.
  - Este cálculo está bien definido si n es impar. Si n es par, la mediana se define como la media aritmética de los dos valores medios.
  - Esta medida es significativa para las escalas *ordinal, de intervalo y ratio*

Medidas de tendencia central (cont.)

- **Moda:** representa la muestra más común
  - Se calcula contando el número de muestras para cada valor único y seleccionando el valor con más cantidad
  - La moda está bien definida si hay solo un valor más común que los otros. Si este no es el caso, se calcula como la mediana de las muestras más comunes
  - La moda es significativa para las escalas *nominal, ordinal, de intervalo y ratio*

Medidas de tendencia central (cont.)

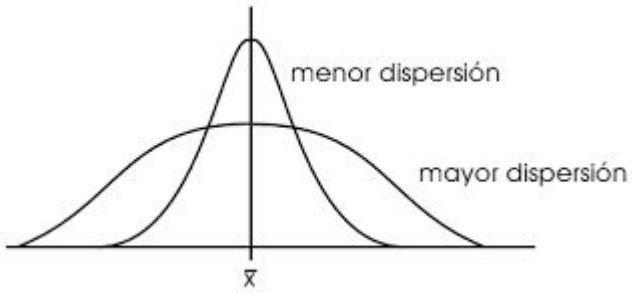
- **A tener en cuenta:**
  - La media aritmética y la mediana son iguales si la distribución de las muestras es simétrica.
  - Si la distribución es simétrica y tiene un único valor máximo, las tres medidas son iguales.
  - Las medidas de tendencia central no proveen información sobre la dispersión del conjunto de datos.
  - Cuanto mayor es la dispersión, más variables son las observaciones, cuanto menor es la dispersión, más homogéneas a la media son las observaciones.

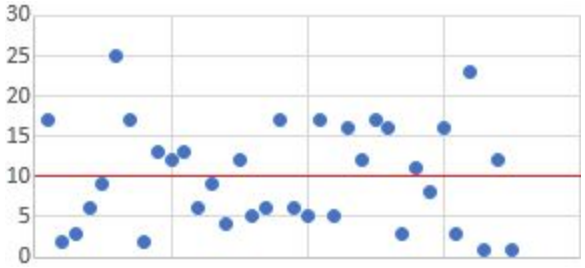
Ejercicio de medidas de tendencia central

- Media: 7 6,67
- Mediana 8 7
- Moda: 9 9

Medidas de dispersión

- Miden el nivel de desviación de la tendencia central, qué tan diseminados o concentrados están los datos respecto al valor central





Medidas de dispersión (cont.)

- **Varianza:** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Da una medida de qué tan alejadas de la media están las observaciones
  - Se calcula como la media de las desviaciones de las observaciones respecto a la media aritmética
  - Dado que la suma de las desviaciones es siempre cero, se toman las desviaciones al cuadrado
  - La varianza es significativa para las escalas de *intervalo y ratio*
- **Desviación estándar:** $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - Se prefiere sobre la varianza porque tiene las mismas dimensiones (unidad de medida) que los valores de las muestras.
  - La desviación estándar es significativa para las escalas de *intervalo y ratio*

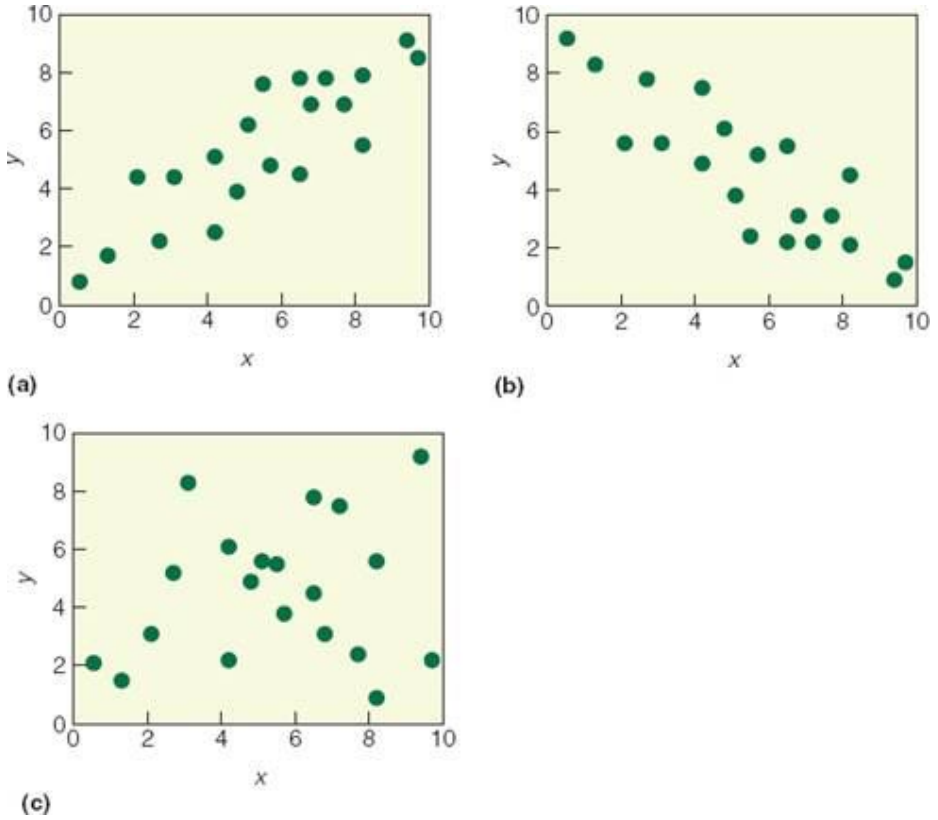
Medidas de dispersión (cont.)

- **Dispersión**: ***coeficiente de variación*** $100 \cdot \frac{s}{\bar{x}}$
  - Se puede expresar como un porcentaje de la media
  - Permite comparar la dispersión o variabilidad de dos o más grupos
  - Esta medida no tiene dimensión y es significativa para la escala *ratio*

- **Rango** $range = x_{max} - x_{min}$
  - Es la distancia entre el valor máximo y el mínimo
  - Es una medida significativa para las escalas de intervalo y ratio
  - Cuando el conjunto de datos consiste en muestras relacionadas de a pares (xi; yi) de dos variables X e Y, puede ser interesante examinar la dependencia entre estas variables
  - Regresión lineal
  - Covarianza
  - Coeficiente de correlación lineal

Ejercicio de dispersión

- ¿Qué muestra presenta mayor dispersión?



Reducción del conjunto de datos

- Las estadísticas descriptivas se ven fuertemente influenciadas por aquellas observaciones que su valor dista significativamente del resto de los valores recolectados *(outliers)*
  - Influencian las medidas de dispersión, aumentando la variabilidad de lo que se está midiendo
  - En algunos casos se decide quitarlos de los datos a analizar porque no son representativos de la población, ya que fueron causados por algún tipo de anomalía:
    - Errores de medición
    - Variaciones no deseadas en las características de los sujetos
- Cuidado al quitar outliers sin un análisis pormenorizado
  - Debido a esto se demoró en detectar del agujero de la capa de ozono