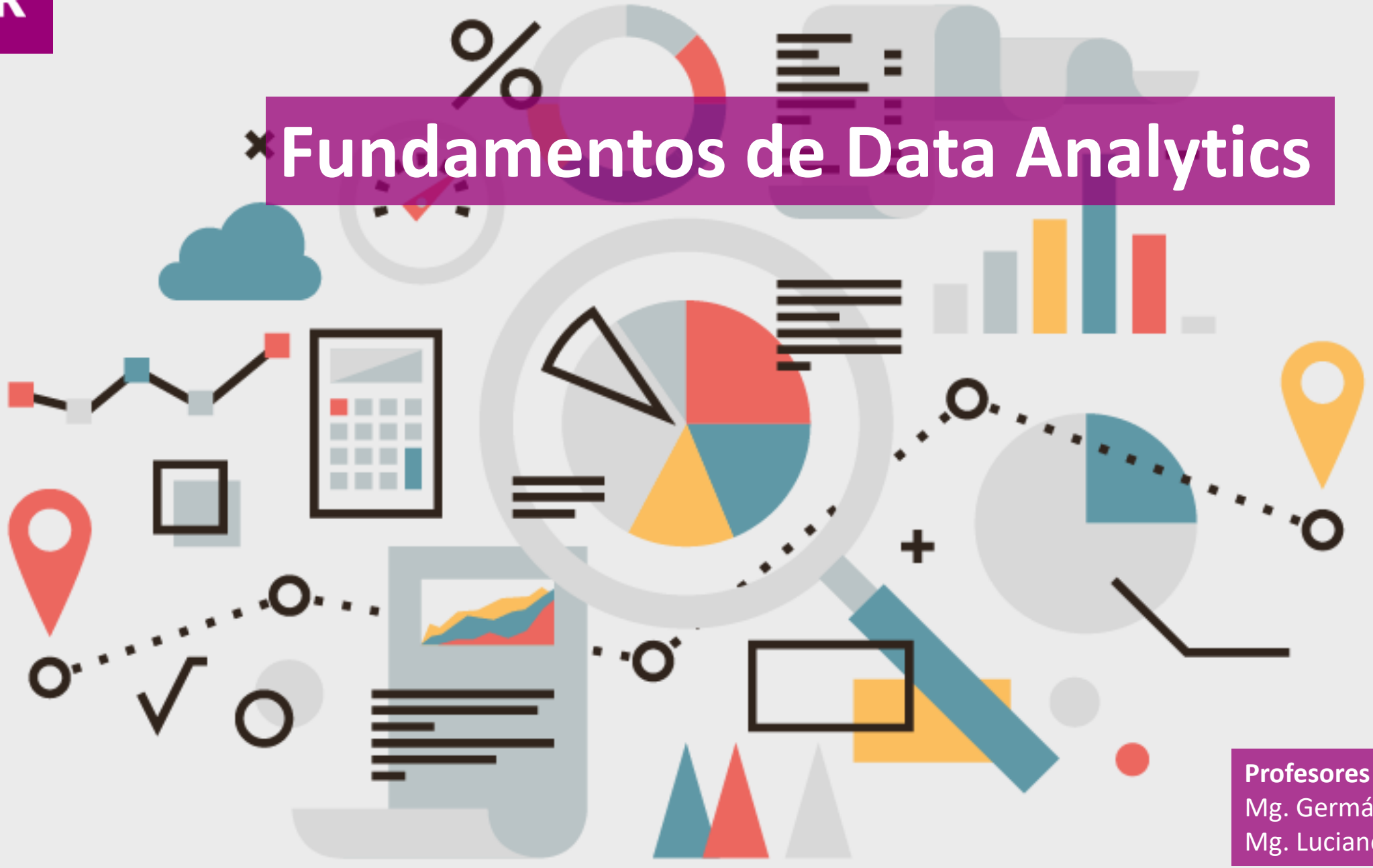
oficios
UNR

Fundamentos de Data Analytics



Profesores
Mg. Germán Tessmer
Mg. Luciano Jara Musuruana

oficios UNR

Introducción a la estadística

- Algunas definiciones necesarias
- Estadística descriptiva

- Esta en todos lados
- Es necesaria para la toma de decisiones
- Presenta de manera resumida la información
- Pensamiento estadístico y crítico

oficios UNR

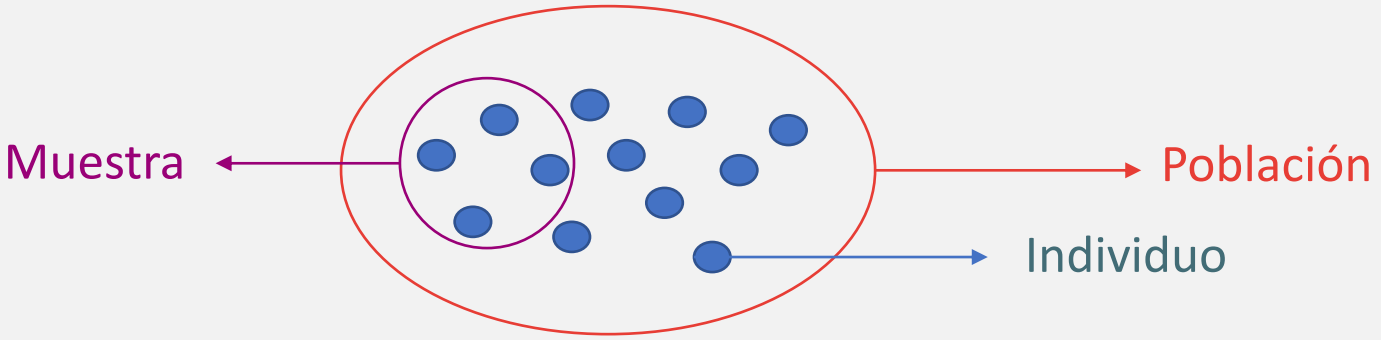
Primeras definiciones…necesarias

- Variable: Es una característica de interés del objeto o de los individuos que no se mantiene constante

- Dato: Es el valor que toma la observación de una variable en un determinado individuo u objeto

- Población: Es el conjunto completo de todos los individuos, o los objetos sobre los que se quiere investigar con respecto a una particularidad dada. **Parámetro**

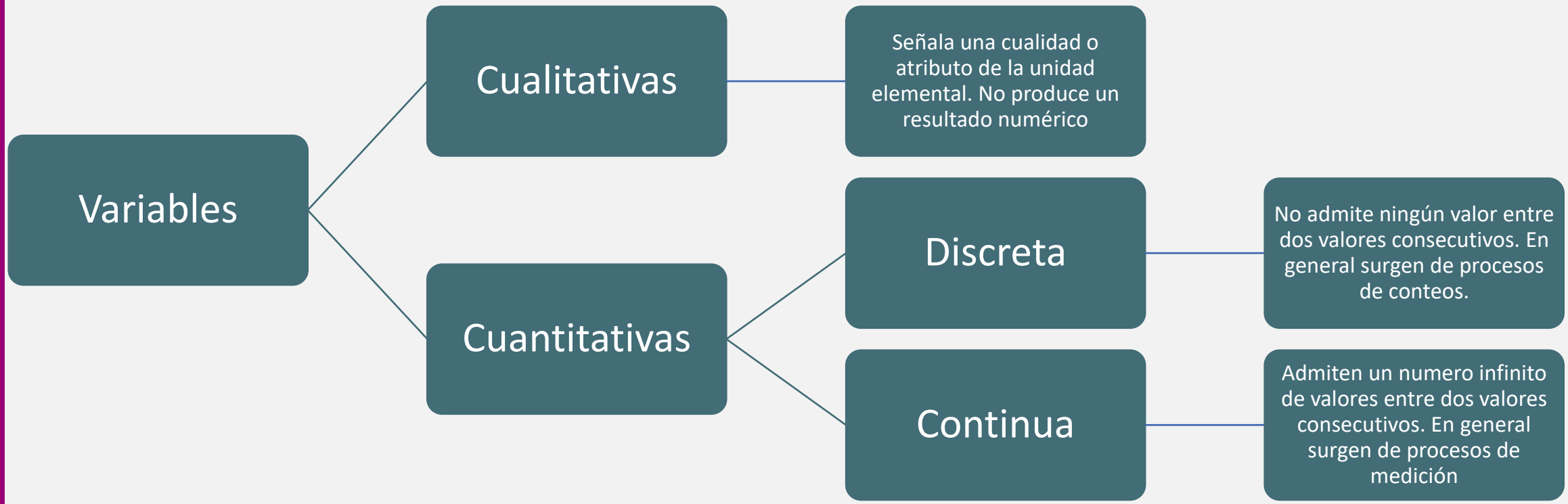
- Muestra: Es un subconjunto de miembros seleccionados de una población para el análisis. **Estadístico**



Ojo con los sesgos

oficios UNR

Tipos de variables



oficios UNR

Medidas descriptivas

De posición o ubicación

- Promedio o media aritmética
- Mediana
- Rango
- Cuartiles, deciles y percentiles

De dispersión

- Rango y rango intercuartílico
- Varianza
- Desvío estándar
- Coeficiente de variación

De forma

- Coeficiente de asimetría

oficios UNR

Promedio o media aritmética

La media de un conjunto de datos es la medida de tendencia central que se encuentra al sumar todos los valores de los datos y dividir el total por el número de datos.

Propiedades

1. Todo conjunto de datos de intervalos o razón posee una media
2. Es un valor único para cada conjunto de datos.
3. En su cálculo intervienen todas las observaciones
4. Se ve afectada por valores extremos (atípicos)

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

oficios UNR

Tipos de promedios

1. Ponderado $$\bar{X} = \frac{\sum_{i=1}^{n} w_i x_i}{n} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{n}$$

2. Condicional $$\bar{X}|(x_i > c) = \frac{\sum_{i=1}^{n} w_i x_i}{n} \quad \wedge \quad (x_i > c)$$

3. Podada

Se ordenan los datos en forma creciente y luego se elimina un cierto porcentaje de datos (redondear si no da entero) en cada extremo de la distribución, finalmente se promedian los valores restantes.

4. Geométrica $$\bar{X}_G = (x_1 * x_2 * \cdots * x_n)^{1/n}$$

oficios UNR

Mediana

La mediana de un conjunto de datos es la medida de tendencia central que indica el valor intermedio, cuando los datos originales se presentan en orden de magnitud creciente (o decreciente).

Propiedades

1. Todo conjunto de datos de nivel ordinal posee una mediana.
2. Es un valor único para cada conjunto de datos.
3. En su cálculo no intervienen todas las observaciones
4. No se ve afectada por valores extremos (atípicos)

Luego de ordenar los datos:

1. Si la cantidad de valores es impar, la mediana es el número ubicado en el intermedio exacto de la lista ordenada.

2. Si cantidad de valores es par, la mediana se obtiene calculando la media de los dos números intermedios de la lista ordenada.

oficios UNR

Moda

La moda de un conjunto de datos es el (los) valor(es) que ocurre(n) con mayor frecuencia.

Propiedades

1. No se ve afectada por valores extremos (atípicos)
2. Se puede trabajar con datos cualitativos
3. Puede haber múltiples modas o ninguna

Busque la variable que más se repita

oficios UNR

Cuartiles, deciles, percentiles

Estas medidas, dividen un conjunto de observaciones en partes iguales. Los cuartiles dividen a un conjunto de observaciones en cuatro partes iguales, los deciles dividen un conjunto de observaciones en 10 partes iguales y los percentiles en 100 partes iguales.

Propiedades

1. En su cálculo intervienen todas las observaciones
2. Se obtienen tantos resultados como conjuntos planteados

$$L_P{}^\circ = \frac{(n+1)P}{100}$$

Si sólo se toma en cuenta las medidas de posición de un conjunto de datos o si compara varios conjuntos de datos utilizando valores centrales, se llegará a una conclusión incorrecta.
Además de las medidas de posición, se debe tomar en consideración la dispersión.

- Lind, Marchal, y Wathen 2012

oficios UNR

Rango y rango intercuartílico

El rango de un conjunto de valores de datos es la diferencia entre el valor máximo de datos y el valor mínimo de datos. El rango intercuartílico mide la amplitud del 50% central de la distribución.

Propiedades

1. El rango es muy sensible a los valores extremos. El intercuartílico no se ve afectado por los valores extremos
2. Sólo los valores máximo y mínimo, no toma en cuenta todos los valores

$$R = X_{max} - X_{min}$$

$$RI = Q_3 - Q_1$$

oficios UNR

Desvío estándar y varianza

Son las medidas de dispersión más utilizadas, acompañan a la media aritmética. La variancia es el promedio de las sumas de los cuadrados de los desvíos de los valores de la variable, respecto a la media. Es difícil de interpretar porque está medida en unidades al cuadrado, por lo tanto, se interpreta el desvío estándar que tiene la misma unidad de medida que la variable bajo estudio.

Propiedades

1. La desviación estándar es una medida de cuánto se desvían los valores de datos de la media.
2. Nunca es negativo, es cero sólo cuando todos los valores de datos son exactamente iguales.
3. Las unidades de la desviación estándar son las mismas que las unidades de los valores de datos originales

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

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

oficios UNR

Coeficiente de variación

Las medidas de dispersión vistas hasta aquí son absolutas, en cambio el coeficiente de variación es una medida de dispersión relativa que se usa para comparar la dispersión entre dos o más distribuciones de variables con distinta unidad de medida, con distinto valor promedio o con distinto desvío.

Propiedades

1. El conjunto de datos que posee un coeficiente de variación menor es más homogéneo

$$CV = \frac{S}{\bar{x}} * 100$$

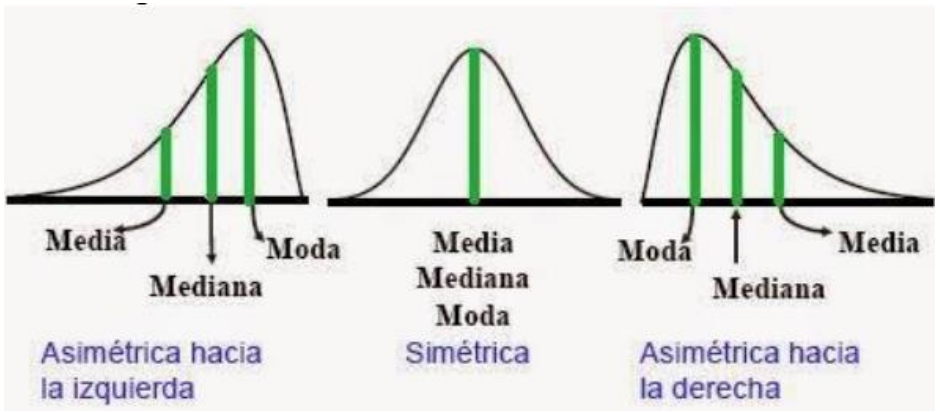
En “The Median Is not the Message” el autor, luego de diagnosticado de cáncer y que le dieran ocho meses de vida. Mostro que la distribución de tiempos de supervivencia se encuentra drásticamente sesgada a la derecha y que no sólo 50% de pacientes de cáncer similar sobreviven más de 8 meses, sino que el tiempo de supervivencia podía ser de años, no de meses.

- Stephen Gould 2013

oficios UNR

Coeficiente de asimetría

A modo intuitivo compararemos media, mediana y moda. Si los tres coinciden aproximadamente, la distribución es simétrica. Si la media es mayor que la mediana y la moda, la distribución es asimétrica hacia la derecha. Si la media es menor que la moda y la mediana, la distribución es asimétrica hacia la izquierda. En la práctica utilizaremos el coeficiente de asimetría de manera más directa



$$CA = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n * s^3}$$