Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

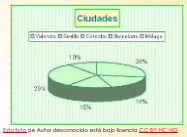
Minería de datos (*Data Mining*)
Tratamiento de datos. Parte 1

1

Profesora: *Dra. Fabiola Ocampo Botello*

Babbie (1988) y Hernández y otros (2003) establecen los siguientes niveles de medición:

- **Nominales**. Se utilizan para distinguir categorías comprendidas en una variable determinada, son mutuamente excluyentes entre sí. Existen dos o más categorías que no tienen orden ni jerarquía, por ejemplo: sexo (hombre o mujer), afiliación religiosa o política. Los números asignados a cada categoría son simplemente con fines de clasificación.



2

- **Ordinales**. Reflejan un orden de rango entre las categorías que forman una variable. Existen varias categorías que mantienen un orden y existe una jerarquía. Los números asignados reflejan tal jerarquía y los intervalos no necesariamente son iguales. Por ejemplo: clase social (alta, media, baja), categoría ocupacional en un empleo.

- **Intervalo**. Es similar al anterior, pero en este tipo de dato los intervalos entre las categorías son iguales en la medición, también se conoce como intervalos iguales para resaltar la característica que la distingue de una escala ordinal. El cero es arbitrario. *Por ejemplo si desea expresar la temperatura ambiental en categorías de 5 en 5 grados, el cero no indica la ausencia de temperatura.*

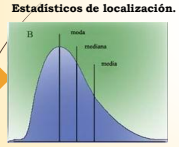
3

- **Razón**. Tiene las mismas características que las medidas de intervalo, pero el cero no es arbitrario, es real. Por ejemplo: las horas a la semana que una persona ve la televisión, el número de hijos, las ventas de un producto en un periodo de tiempo, la edad en años. Una distancia de 10 km está al doble de una de 5 km.

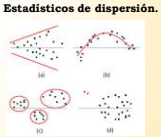
Tipos de datos

Bennet, Briggs y Triola (2011) establecen que **estadística** (en singular) es la ciencia que recolecta, organiza e interpreta datos y **estadísticas** (en plural) son los datos (números y otras partes de información) que describen o resumen algo. Castillo Morales (2013) establece que un estadístico es un procedimiento de cálculo que usa datos y constantes conocidas. Menciona que existen dos tipos de estadísticos.

Estadísticos de localización.



Estadísticos de dispersión.



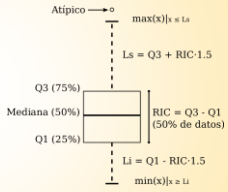
4

Estadísticos de localización

Los estadísticos de localización son: mínimo, máximo, semirrango, mediana, percentil 25% o primer cuartil, percentil 75% o tercer cuartil, percentil 95%, media y moda.

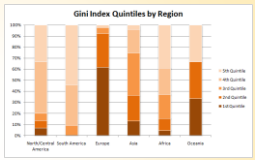
Semirrango: es el valor intermedio entre el máximo y el mínimo.

Los valores que dividen un conjunto de datos en partes iguales son: cuartiles, deciles, centiles y percentiles.



La **mediana** indica el centro de los datos.

Cuartiles. Dividen el conjunto de observaciones en cuatro partes iguales.
Q_1 es el primer cuartil, es el valor abajo del cual se encuentra el 25% de las observaciones.
Q_2 es la mediana
Q_3 es el valor por abajo del cual se encuentra el 75% de las observaciones.



Deciles dividen el conjunto de datos en 10 partes iguales.

Centiles dividen el conjunto de datos en 100 partes iguales.

Los datos más representativos de una serie de observaciones son:

Valor mínimo
Cuartil 1
Mediana
Cuartil 3
Valor máximo

Los cuales son los expresados en una gráfica de caja y bigotes.

Mason, Lind y Marchal (2000) establecen que la varianza y la desviación estándar son las medidas de dispersión más ampliamente usadas, pero existen otros medios para describir los valores que dividen un conjunto de datos en partes iguales. Estos son: cuartiles, deciles, centiles y percentiles.

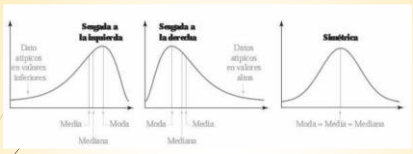
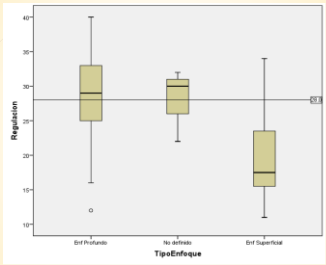


Figura tomada de: Bennet, Briggs & Triola (2011:159)

Una gráfica que no es simétrica tiende a desplegarse más hacia un lado que hacia otro.
La gráfica (a) tiene un sesgo a la izquierda, lo que indica que tiene valores bajos atipicos .

9



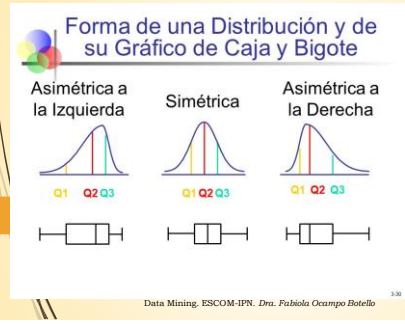


Figura tomada de: Aranda Gómez, Juan. (2016). Describiendo datos, usando medidas numéricas. Recurso de la Web. Disponible en: https://slideplayer.es/slide/10266389/

11

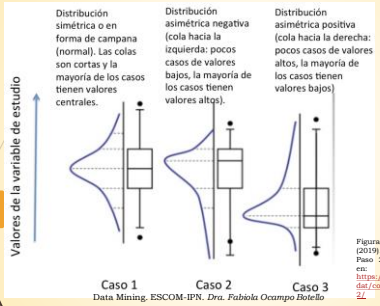


Figura tomada de: Ferrero, Rosana. (2019). Cómo describir tus pasos en R: Paso 2. Máxima Formación. Disponible en: https://www.maximaformacion.es/blog-dat/como-describir-tus-datos-en-r-paso-2/

12

Estadísticos de dispersión

Se entiende por dispersión la separación que presentan los puntos entre sí o con respecto al centro de la gráfica. Si todos los datos tienen el mismo valor, no hay dispersión y esta vale cero (Castillo Morales, 2013). Los estadísticos de dispersión son:

- Rango. se obtiene restando el valor mínimo al valor máximo y da la longitud del intervalo en donde se encuentra la muestra.

$$Rango = (Max) - (Min)$$

Rango

- Rango intercuartílico. Es la diferencia entre los percentiles 75% y 25%, entre éstos se encuentra el 50% de los valores intermedios de la muestra.

$$IQR = Q_3 - Q_1$$

Rango intercuartílico

- Varianza. En una muestra se refiere a la diferencia entre el dato y la media elevadas al cuadrado.

$$S_X^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{x})^2}{N-1}$$

Varianza

- Desviación estándar. Es la raiz cuadrada de la varianza.

$$S_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i - Media(X))^2}{N-1}}$$

Desviación típica

13

- La medida de tendencia central más utilizada es la media.
- Si se tiene un grupo muy heterogéneo en alguna puntuación, por ejemplo en el promedio, su varianza será muy grande con respecto a uno que es muy homogéneo.
- La varianza es un estadístico muy utilizada en la comparación de grupos, en el análisis de hipótesis, entre otros más.
- La medida de variabilidad más utilizada es la varianza. También llamada cuadrado medio. Nos dice qué tan dispersos están los valores con respecto a la media.

14

Representación de datos

La representación de datos se puede hacer mediante tablas o gráficas. Algunas de las utilizadas son:

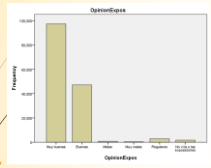
Tablas de frecuencias y de frecuencia acumulada

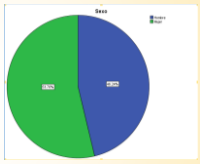
Eval_Giral

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	99	.1	.1	.1
	1	57	.0	.0	.1
	2	44	.0	.0	.1
	3	67	.0	.0	.2
	4	132	.1	.1	.3
	5	448	.3	.3	.6
	6	813	.5	.5	1.1
	7	2991	2.0	2.0	3.1
	8	14847	9.9	9.9	13.0
	9	33738	22.5	22.5	35.5
	10	96772	64.5	64.5	100.0
	Total	150008	100.0	100.0	

15

Gráficas de barras (datos cualitativos) y gráficas de pastel (datos cualitativos).



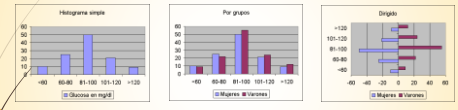


Las gráficas de barras se utilizan para variables categóricas, cualitativas, nominales.

16

Histograma

Un histograma es una gráfica de barras que muestra una distribución para datos cuantitativos (de intervalo o de razón de medida); las barras tienen un orden natural y las anchuras de las barras tienen un significado específico.



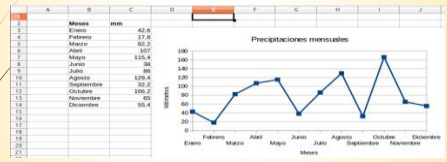
17

Figuras tomadas de: Ejemplos de tipos de representación gráfica. "s/f". http://www.hrc.es/bioest/Ejemplos_histo.html

<u>Los histogramas</u> se utilizan para representar frecuencias de variables continuas, cuantitativas, en donde cada barra representa la frecuencia de un intervalo de valores.

Gráfica de líneas

Muestra una distribución de datos cuantitativos, conecta una serie de puntos. Los puntos van donde iría la parte superior de la barra en el histograma. La posición horizontal de los puntos corresponde al centro de la clase.



18

Figura tomada de: Crear un gráfico de líneas ("S/f").
En: https://ordenadorpractico.es/mod/assign/view.php?id=274

19

La gráfica de dispersión permite visualizar valores de dos variables.

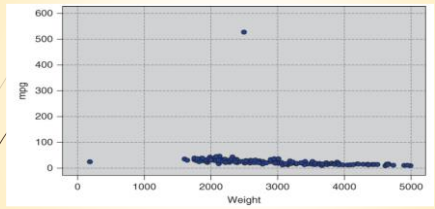


Figura 2.6 tomada de Larose & Larose (2015:Sección 2.5).

20

¿Qué hacemos con los datos sucios?

Han, Kamber & Pei (2012) presentan diversas rutinas de limpieza de datos para tratar los valores faltantes, suavizar el ruido al encontrar valores atípicos y corregir inconsistencias.

Valores faltantes:

1. Ignorar la tupla. Esto se hace por ejemplo cuando falta la etiqueta de la clase y lo que se realiza es la clasificación. Este método no es muy efectivo, a menos que a la tupla le falten varios valores.

2. Completar manualmente el valor faltante. Este enfoque consume mucho tiempo y puede no ser factible considerando un gran conjunto de datos con muchos datos faltantes.

3. Uso de una constante global para completar los valores faltantes. Se puede utilizar una etiqueta como "Desconocido". Pero, hay que tener cuidado por que el algoritmo de minería de datos puede detectar erróneamente que es un concepto interesante ya que existen muchos datos con ese valor.

21

4. Uso de una medida de tendencia central para el atributo. Puede ser la media o la mediana.

5. Usar la media o la mediana para todos las muestras que pertenecen a la misma clase. Por ejemplo, si se clasifica a los clientes de acuerdo al riesgo de crédito, se puede reemplazar el valor faltante con el promedio del valor de ingreso de los clientes de esa categoría.

6. Usar el valor más probable para completar el valor faltante. Este valor se puede determinar mediante una regresión.

Referencias bibliográficas

Babbie R. Earl. (1988). *Métodos de investigación por encuesta*. Fondo de Cultura Económica. México.
Bennet, Briggs & Triola (2011). *Razonamiento estadístico*. Pearson. México.
Castillo M., A. (2013). *Estadística aplicada*. México, ed. Trillas.
Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). *Data Mining: concepts and techniques*. Third edition. Morgan Kaufman Series.
Hernández Sampieri, R.; Fernández Collado, C; Baptista Lucio, P. (2003). *Metodología de la Investigación*. Tercera Edición. Editorial Mc. Graw Hill. D. F. México.
Larose, T. Daniel & Larose, D. Chantal. (2015). *Data Mining and Predictive Analytics*. Second Edition. Wiley.
Mason, Lind & Marchal. (2000). *Estadística para administración y Economía*. Alfaomega. México.

22