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

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

1

Profesora: *Dra. Fabiola Ocampo Botello*

"Los **datos** son hechos/informaciones y cifras que se recogen, analizan y resumen para su presentación e interpretación. A todos los datos reunidos para un determinado estudio se les llama **conjunto de datos** para el estudio" (Anderson, Sweeney & Williams, 2008:5).

"**Elementos** son las entidades de las que se obtienen los datos, los nombres de los elementos aparecen en la primera columna. **Una variable** es una característica de los elementos que es de interés. Los valores encontrados para cada variable en cada uno de los elementos constituyen los datos. Al conjunto de mediciones obtenidas para un determinado elemento se le llama **observación**" (Anderson, Sweeney & Williams, 2008:6).

2

Ejemplo:

Alumno	No. Boleta	Sexo	Semestre
Diana	2020001	Mujer	7
Yiya	2020005	Mujer	5
Kevin	2020007	Hombre	6

Tipos de datos

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.

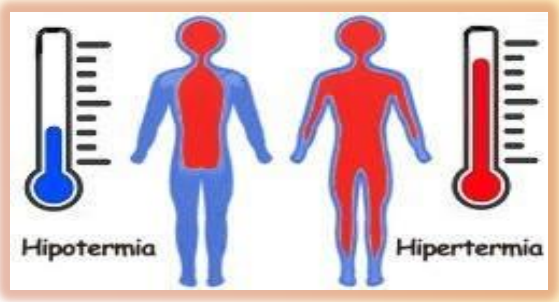
Fuente de la imagen: Capacitación on line
https://wwwcapacitacionononline.blogspot.com/2008/07/

3

Fuente de la imagen: Capacitación on line
https://wwwcapacitacionononline.blogspot.com/2012/03/grafica-variable-ordinal.html

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.*



Fuente de la imagen: Fuente saludable
https://www.fuentesaludable.com/porque-es-importante-para-el-organismo-mantener-una-temperatura-constante/

4

Fuente de la imagen: Estadística unidimensional
https://www.victormat.es/4ESOAC/Tema11-Estadistica/grficas.html

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.

También existen los datos **cualitativos y cuantitativos**

"Los **datos cualitativos** comprenden etiquetas o nombres que se usan para identificar un atributo de cada elemento. Los datos cualitativos emplean la escala nominal o la ordinal y pueden ser numéricos o no" (Anderson, Sweeney & Williams, 2008:7).

Fuente de la imagen: Ciencia sin seso…locura doble
https://www.cienciasinseso.com/graficos-de-variables-cualitativas/

5

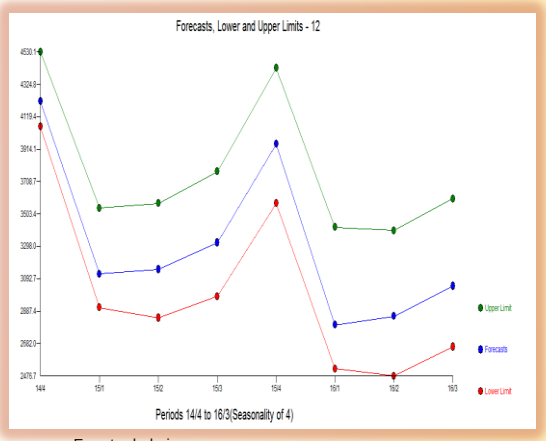
Fuente de la imagen: Universo Fórmulas
https://www.universoformulas.com/estadistica/descriptiva/poblacion-estadistica/

"Los **datos cuantitativos** requieren valores numéricos que indiquen cuánto o cuántos. Los datos cuantitativos se obtienen usando las escalas de medición de intervalo o de razón" (Anderson, Sweeney & Williams, 2008:7).

Datos de sección transversal y de series de tiempo

"**Datos de sección transversal** son los obtenidos en el mismo o aproximadamente el mismo momento (punto en el tiempo). Los **datos de series de tiempo** son datos obtenidos a lo largo de varios periodos." (Anderson, Sweeney & Williams, 2008:7).

6

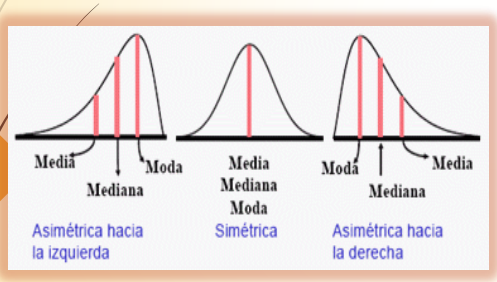


Fuente de la imagen:
https://stats.stackexchange.com/questions/247978/choosing-between-additive-and-multiplicative-model

Estadística y Estadísticas

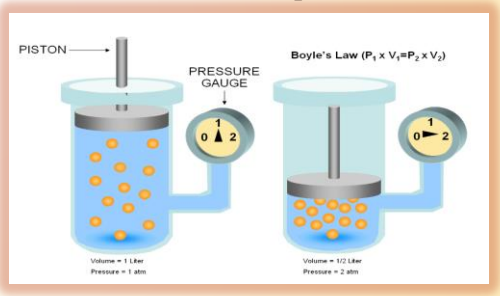
Bennet, Briggs y Triola (2011) establecen que **estadística** (en singular) es la ciencia que recolecta, organiza e interpreta datos y **estadísticas** (en plural) son los datos (números y otras partes de información) que describen o resumen algo. Castillo Morales (2013) establece que un estadístico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

Estadísticos de localización.



Fuente de la imagen: Enlázate con los números
https://patenlazandote.blogspot.com/2013/10/media-aritmetica-y-mediana.html

Estadísticos de dispersión.



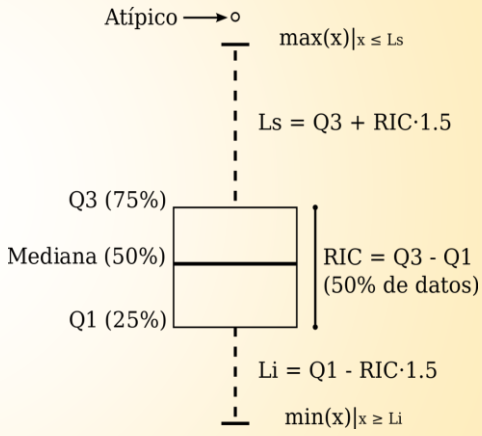
Fuente de la imagen: El vuelo de la gran Avetuarda
https://greatbustardsflight.blogspot.com/2015/01/la-atmosfera-estandar.html

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.

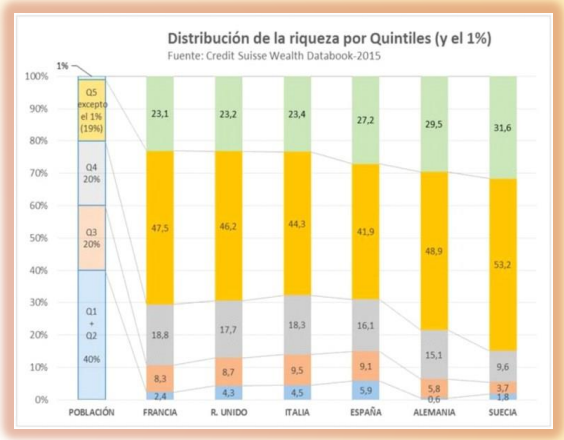
Los valores que dividen un conjunto de datos en partes iguales son: cuartiles, deciles, centiles y percentiles.



Fuente de la imagen: Diagrama de Caja. Wikipedia.
https://es.wikipedia.org/wiki/Diagrama_de_caja

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.



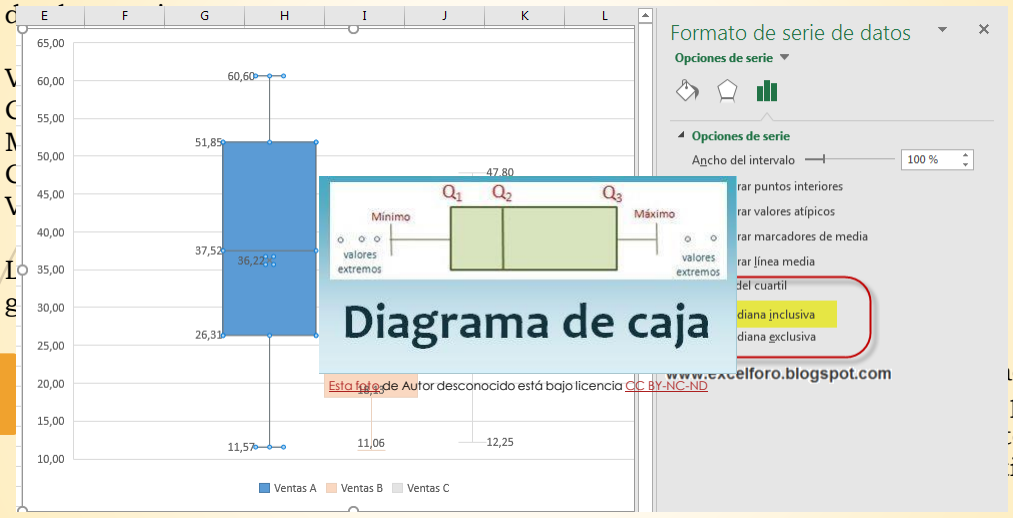
Fuente de la imagen: La Haine.
https://www.lahaine.org/mm_ss_mundo.php/la-brutal-desigualdad-de-suecia

9

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



ación
pero
o de
iles.

10

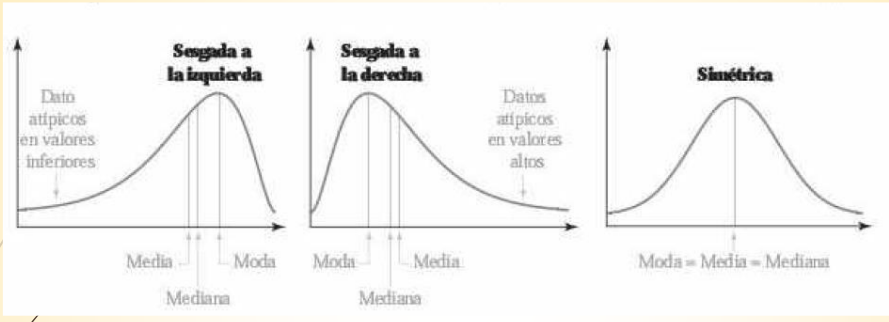
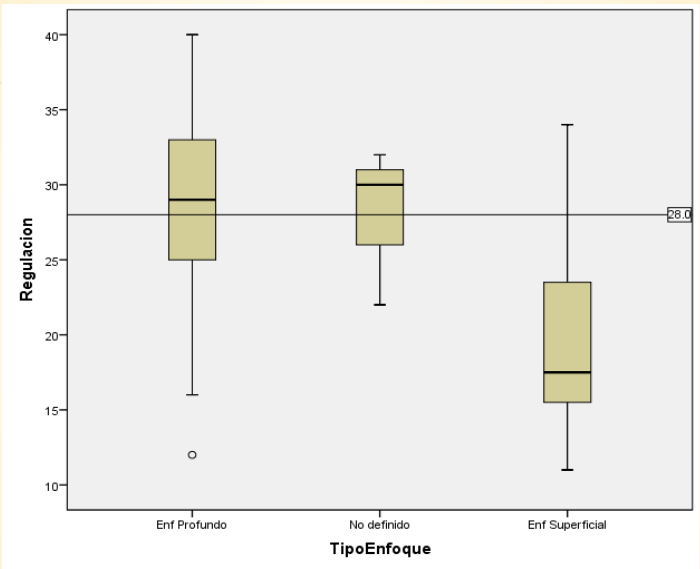


Figura tomada de: Bennet, Briggs & Triola (2011:159)

11

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 atípicos .

12



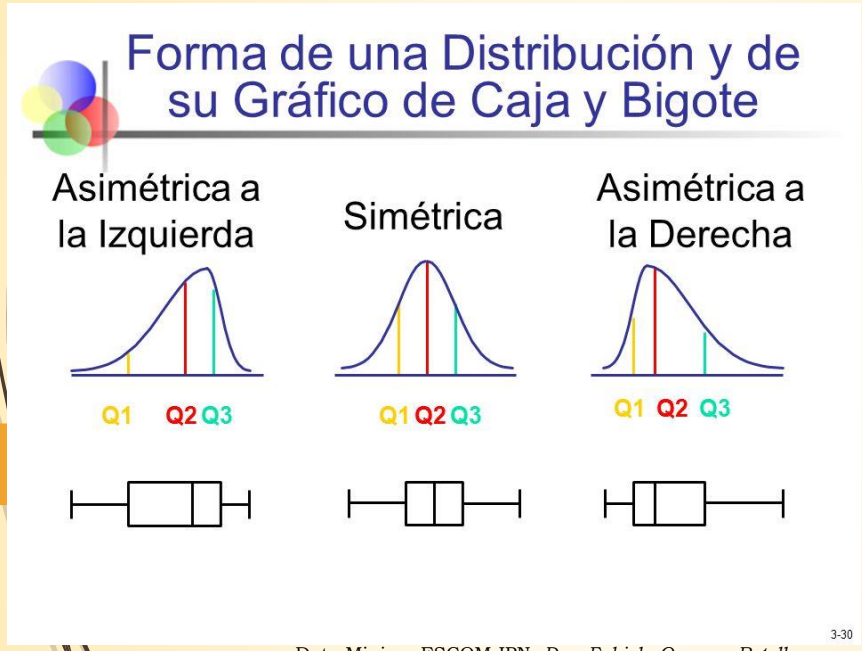


Figura tomada de:
Aranda Gómez, Juan. (2016). Describiendo datos, usando medidas numéricas. Recurso de la Web. Disponible en: https://slideplayer.es/slide/10266389/

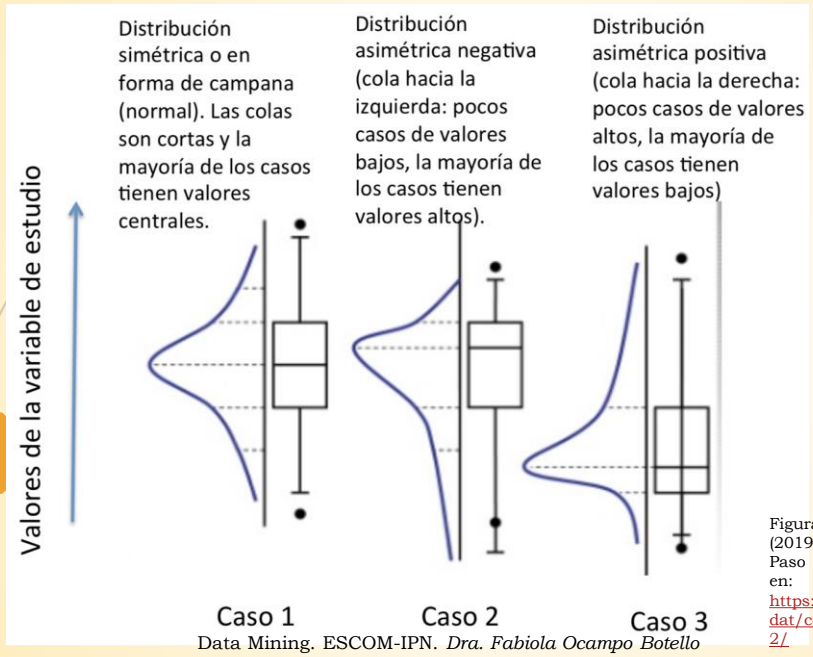
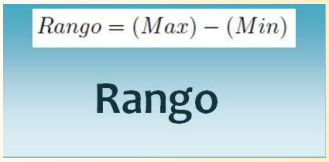


Figura tomada de: Ferrero, Rosana. (2019). Cómo describir tus pasos en R: Paso 2. Máxima Formación. Disponible en: https://www.maximaformacion.es/blog-dat/como-describir-tus-datos-en-r-paso-2/

Estadísticos de dispersión

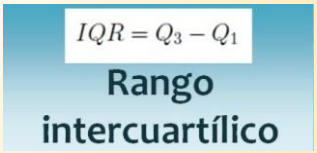
Se entiende por dispersión la separación que presentan los puntos entre sí o con respecto al centro de la gráfica. Si todos los datos tienen el mismo valor, no hay dispersión y esta vale cero (Castillo Morales, 2013). Los estadísticos de dispersión son:

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

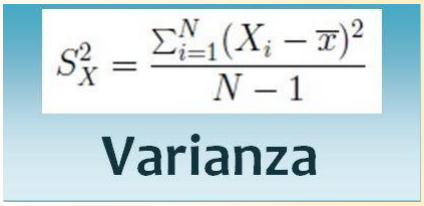


15

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

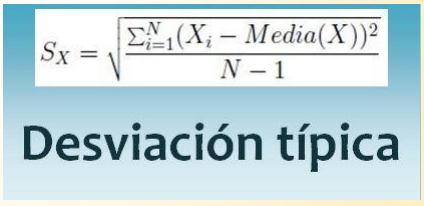


- Varianza. En una muestra se refiere a la diferencia entre el dato y la media elevadas al cuadrado.



16

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



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.

17

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.

Representación de datos

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

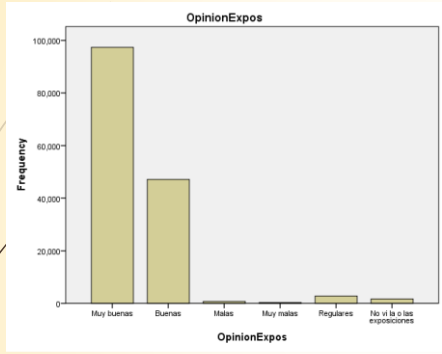
Tablas de frecuencias y de frecuencia acumulada

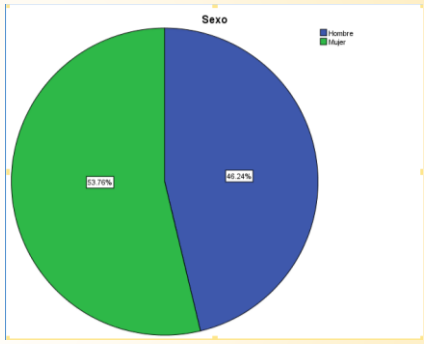
18

Eval_Gral

		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	

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



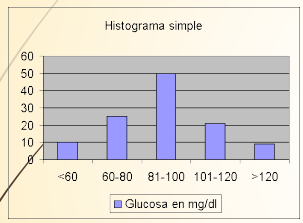


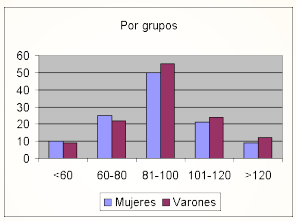
19

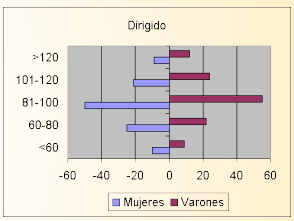
<u>Las gráficas de barras</u> se utilizan para variables categóricas, cualitativas, nominales.

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.







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

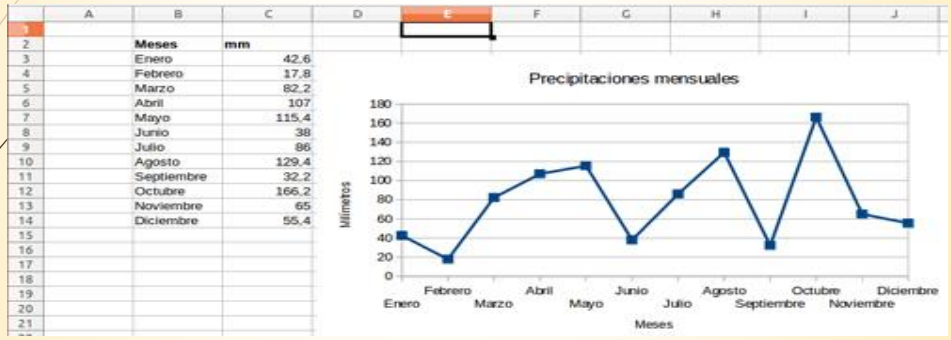
20

<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.

21



Meses	mm
Enero	42,6
Febrero	17,8
Marzo	82,2
Abril	107
Mayo	115,4
Junio	38
Julio	86
Agosto	129,4
Septiembre	32,2
Octubre	166,2
Noviembre	65
Diciembre	55,4

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

22

La gráfica de dispersión permite visualizar valores de dos variables.

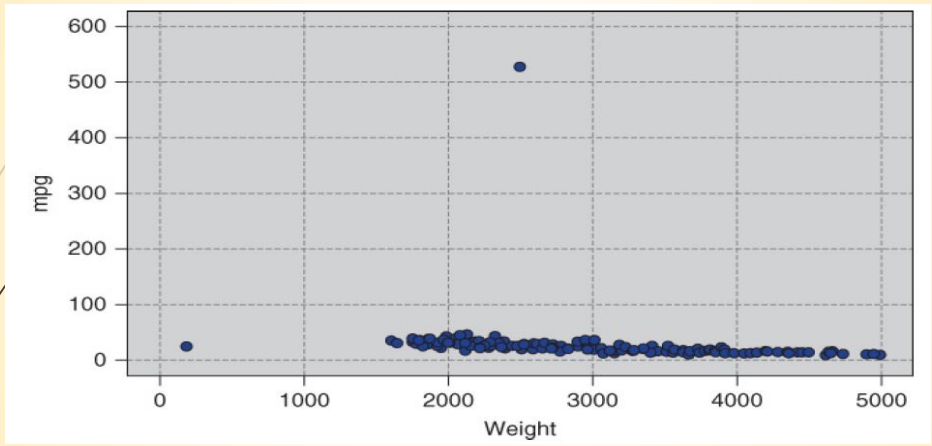
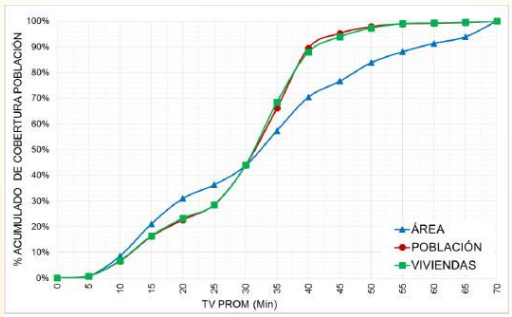


Figura 2.6 tomada de Larose & Larose (2015:Sección 2.5).

23

La gráfica de Ojiva

La gráfica de una distribución acumulada, llamada ojiva, es una gráfica que muestra los valores de los datos en el eje horizontal y las frecuencias acumuladas, las frecuencias relativas acumuladas o las frecuencias porcentuales acumuladas en el eje vertical (Anderson, Sweeney & Williams, 2008:39).



Data Mining. ESCOM-IPN. *Dra. Fabiola Ocampo Botello*

Representaciones de datos Cualitativos

24

Distribución de frecuencia. Una distribución de frecuencia es un resumen tabular de datos que muestra el número (frecuencia) de elementos en cada una de las diferentes clases disyuntas (que no se sobreponen). La **frecuencia relativa** de una clase es igual a la parte o proporción de los elementos que pertenecen a cada clase. **La frecuencia porcentual** de una clase es la frecuencia relativa multiplicada por 100 (Anderson, Sweeney & Williams, 2008:29).

Las gráficas que se utilizan con los datos cualitativos son:

- Gráfica de barras
- Gráfica de pastel

X_i	Frecuencia absoluta (n_i)	Frecuencia relativa (fi = ni/N)	Frecuencia relativa (fi = ni/N) en %
3	2	0,07	7%
4	4	0,13	13%
5	6	0,20	20%
6	7	0,23	23%
7	5	0,17	17%
8	3	0,10	10%
9	2	0,07	7%
10	1	0,03	3%
Total	30	1	100%

Data Mining. ESCOM-IPN. *Dra. Fabiola Ocampo Botello*

25

¿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.

26

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

Anderson, Sweeney & Williams. (2008). Estadística para administración y economía, 10ª edición. Cengage Learning.

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.