FCNM
Facultad de Ciencias Naturales y Matemáticas

FCNM
Facultad de Ciencias Naturales y Matemáticas

ESPOL

Escuela Superior Politécnica del Litoral

Facultad de Ciencias Naturales y Matemáticas

Estadística

UNIDAD1: Estadística Descriptiva

PARTE 1: Introducción

Profesor: Mat. Fernando Sandoya, PhD.

Temáticas a cubrir

- Qué es la estadística
- Por qué y para qué usar la estadística
- El método científico y la estadística
- Clasificación de la estadística
- Conceptos básicos
- Tablas y distribuciones de frecuencias.
- Representación gráfica de datos
- Descripción de datos Bivariados

- La estadística es una ciencia relativamente joven, pero que tiene sus inicios en la antigüedad. Desde las primeras sociedades organizadas los gobernantes llevaban un registro de las posesiones y producción de los ciudadanos.

De esta manera, sabían qué impuestos debían pagar al Estado. De ahí, el nombre de estadística.

- Conforme avanza la civilización, las actividades se vuelven más complejas y es más difícil organizarlo todo. Sin embargo, las técnicas estadísticas desarrolladas tienen cada vez más ámbitos de aplicación.

De hecho, en nuestra época vivimos una nueva revolución donde es mucho más fácil adquirir, registrar, almacenar, analizar y encontrar patrones aparentemente ocultos en todo el mar de datos.

- Actualmente la estadística se ha fusionado con otros campos de la ciencia: la informática, la biología, la medicina para dar lugar a nuevas disciplinas científicas como:

  - ✓ Machine Learning
  - ✓ Bioestadística
  - ✓ Deep Learning
  - ✓ Otros

- La estadística se aplica, prácticamente, en todas las áreas profesionales, por ejemplo: ingeniería, ciencias químicas, ciencias de la salud, economía, políticas públicas, educación, servicios, etc. y está en auge gracias a que en la actualidad se ha hecho más fácil y económico la recopilación y registro de datos, indispensable en la era del Big Data.
- Big Data significa la existencia de grandes cantidades de datos.
- Por ejemplo, en Ciencias de la salud es de amplia aplicación, pues se utiliza para probar la seguridad y efectividad de los medicamentos en caso de una enfermedad en expansión para conocer su crecimiento y ubicar su origen. También se utiliza en los controles de rutina para poder prevenir la aparición de enfermedades.

- En la industria: se utiliza para llevar el control de calidad, para realizar las campañas de marketing, para prevenir los fallos en el proceso de producción, o para calcular el coste del servicio posventa en caso de defectos en su producto.
- En la macroeconomía permite describir a la población, calcula indicadores de todo tipo, como los de empleo, producción, y pone a disposición estos datos tanto para el propio gobierno y el sector´público como para los ciudadanos interesados.
- Incluso existe una industria estadística como las aseguradoras, las consultoras especializadas que dan servicio a empresas, o los centros de estudios de opinión que suelen encargarse de recopilar datos para hacer campañas de marketing, o, incluso, campañas políticas.

- En ocasiones, no queda tan claro la diferencia entre las matemáticas y la estadística. La principal diferencia entre las matemáticas y la estadística es que esta última trabaja en base a la incertidumbre. Por ejemplo, si son fiables los datos con los que vamos a trabajar, si los instrumentos estaban bien calibrados, o si algo afectó a nuestro sistema sin que nos diéramos cuenta. Aun así, es un primer paso para sistematizar las observaciones y encontrar las leyes matemáticas de los sistemas.
- Lo mejor de todo es que la estadística es una parte de la matemática muy aplicada y práctica, porque trabaja generalmente con datos reales.

¿Para qué sirve la estadística?

- La Estadística estudia la aleatoriedad (incertidumbre) asociada a toda clase de fenómenos.
- ALEATORIEDAD ≠ DETERMINISMO

DEFINICIÓN: Estadística es la ciencia que proporciona las herramientas para recolectar, transformar, interpretar y analizar datos para obtener información para la toma de decisiones

ESTADÍSTICA DESCRIPTIVA:

- Es la parte de la estadística que nos permite comprender un conjunto de datos. Con la estadística descriptiva conoceremos las tendencias y homogeneidad dentro de un conjunto de datos.

PROBABILIDAD:

- La probabilidad es un modelo de como pueden suceder los eventos con incertidumbre. Con la probabilidad buscaremos el orden dentro del azar de sucesos probables. , deducir las leyes que rigen esos fenómenos

ESTADÍSTICA INFERENCIAL:

- Se encarga de estudiar si analizando una muestra podemos proyectar (inferir) las propiedades de la población completa de donde procede.

Estos conceptos fundamentales son válidos para pocos, algunos cientos, para miles, o para miles de millones de datos.

Estadística descriptiva

Es el estudio de las técnicas para recopilar, organizar y presentar datos obtenidos en un estudio estadístico para facilitar su análisis y comprensión

Estadística descriptiva

¿Qué herramientas usa la estadística descriptiva?

- Tablas
- Medidas (números) que caracterizan a los datos
- Gráficos

Variables y datos.

La materia prima de la estadística son los **datos**.

- ¿Qué es un dato? Información concreta sobre hechos, elementos, etc., que permite estudiarlos, analizarlos o conocerlos.
- ¿Qué es una variable? Para distinguir entre los diferentes tipos de datos vamos a definir primero el concepto de **variable**.
- De acuerdo a su formato o a su naturaleza **LAS VARIABLES** pueden ser de diferentes tipos.

Variables.

Una variable es cualquier característica que se registra o mide sobre cualquier entidad (objeto, persona, organización, etc.)

Por ejemplo, cuando nos fijamos en los vehículos podemos diferenciarlos por: la marca, por el número de asientos, por su kilometraje, por su antigüedad, el cilindraje, la potencia del motor o por el combustible que utilizan.

Lo mismo, por ejemplo, si tenemos que mirar las características de los animales: Se pueden dividir por alimentación, por gestación, por su especie o por cualquier otra característica que se nos ocurra. Dichas características o atributos son lo que llamaremos **variables.**

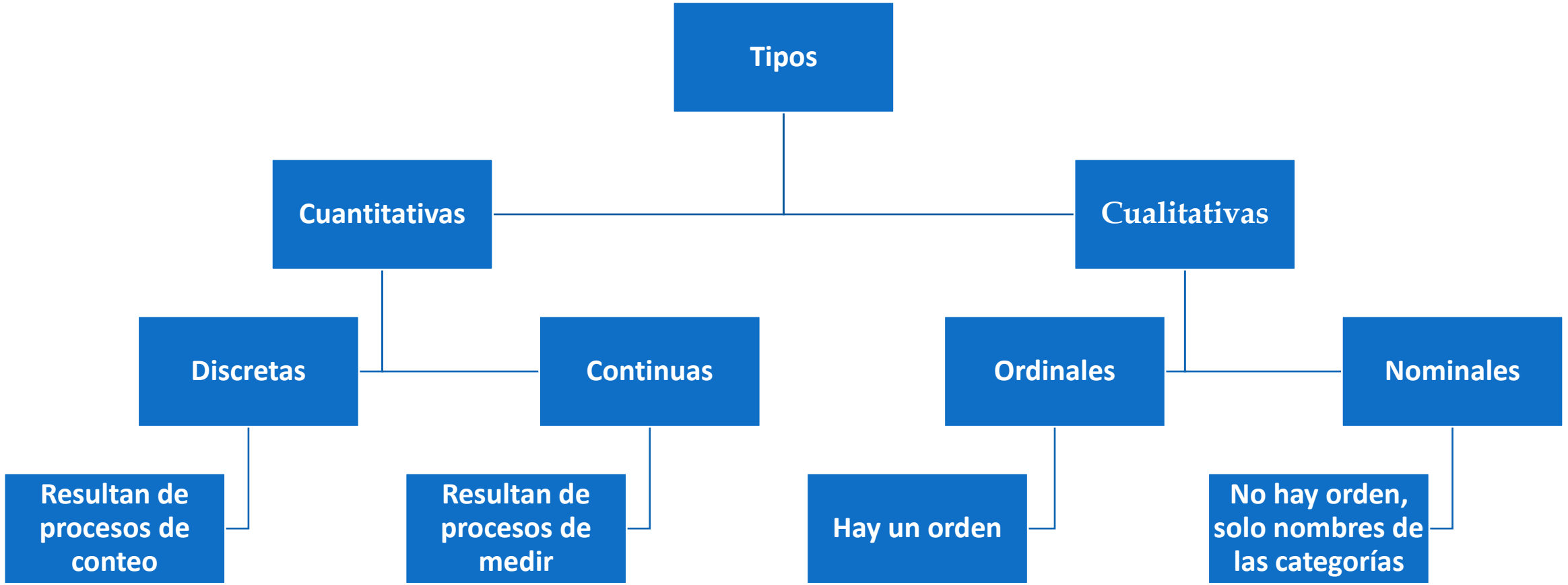
Los datos son los valores que toman las variables para los individuos específicos.

Tipos de variables según su valor.

A las variables las vamos a representar con un símbolo. Por ejemplo, la variable peso la representaremos con una *P*, la variable color con una *C*, etc. o podemos utilizar un nombre adecuado para la característica que estemos estudiando.

Las variables pueden ser de diferentes tipos según su naturaleza:

- **Cuantitativas**: y dentro de las cuantitativas nos vamos a encontrar las cuantitativas discretas y las cuantitativas continuas.
- **Cualitativas**: Las variables cualitativas no cuantifican a una cualidad, simplemente la representan de manera no numérica. Y dentro de las cualitativas nos vamos a encontrar con las ordinales y las nominales.


Tipos
Cuantitativas
Cualitativas
Discretas
Continuas
Ordinales
Nominales
Resultan de procesos de conteo
Resultan de procesos de medir
Hay un orden
No hay orden, solo nombres de las categorías

Escala de medición (2)

Ejemplos:

- Número de hijos en una familia.
- Grado de satisfacción con la prestación de un servicio.
- Litros de agua consumidos por persona al día.
- Barrio de residencia de los estudiantes
- Temperatura del cuerpo humano.
- Número de goles marcados por un jugador en un partido.
- Ubicación de una carretera respecto de un punto de referencia (Km 85, Ruta 5).
- Nivel de productividad.
- Nivel socioeconómico.
- Ingreso familiar mensual.
- Variables usadas en test de rendimiento.
- Grupo sanguíneo de los estudiantes.

Estadística descriptiva

Fuente de datos:

- Registros administrativos (Historias clínicas)
- Aplicación de formularios
- Realización de experimentos

Data Sources

Registros y datos.

- Toda la información que podemos obtener de un experimento, la podemos identificar en sus diferentes características. Cada una de ellas la podemos entender como una variable.
- Una vez que tenemos definidas nuestras variables, pasamos a registrarlas (a mano, en una hoja electrónica, en una base de datos). Cada uno de los registros será un dato. Por ejemplo, si nuestra variable es la superficie en metros cuadrados de construcción de la casa, en cada una de estas celdas vamos a registrar uno a uno los datos.

Registros y datos.

La forma estándar de registrar los datos y variables es en forma de una tabla (con filas y columnas) de la siguiente manera:

- En cada fila se ubica cada dato (los registros de cada entidad)
- En las columnas van las variables

	A	B	C	D	E	F	G	H	I	J	K
1	vehículo	Antigüedad	KM	Combustible	HP	Pintura resistente	Automatico	CC	Puertas	Peso	Precio
2	1	16	46986	Diesel	90	1	0	2000	3	1165	13500
3	2	16	72937	Diesel	90	1	0	2000	3	1165	13750
4	3	17	41711	Diesel	90	1	0	2000	3	1165	13950
5	4	19	48000	Diesel	90	0	0	2000	3	1165	14950
6	5	23	38500	Diesel	90	0	0	2000	3	1170	13750
7	6	25	61000	Diesel	90	0	0	2000	3	1170	12950
8	7	20	94612	Diesel	90	1	0	2000	3	1245	16900
9	8	23	75889	Diesel	90	1	0	2000	3	1245	18600
10	9	20	19700	Petrol	192	0	0	1800	3	1185	21500
11	10	16	71138	Diesel	69	0	0	1900	3	1105	12950
12	11	18	31461	Petrol	192	0	0	1800	3	1185	20950

Registros y datos.

Cuando registramos los datos en el formato estándar tabular debemos identificar el tipo de variables.

Por ejemplo, si tenemos datos de un censo:

- Si nos interesa el tipo de vivienda, (casas, departamentos, cuartos).
- Si lo que nos interesa es la superficie construida (pequeña, mediana o grande).
- Si lo que nos interesa es la población (urbana, rural).
- Por el precio de venta
- Una vez que tenemos definidas nuestras variables, pasamos a registrarlas. Cada uno de los registros será un dato. Por ejemplo, si nuestra variable es la superficie en metros cuadrados, en cada una de estas celdas vamos a registrar uno a uno los datos.

Registros y datos.

Existen diferentes plataformas para el registro electrónico de los datos, entre ellas:

- Excel.
- Google Sheets.
- CSV
- Bases de datos SQL, etc.

ACTIVIDAD: Recopilación de Datos

Estudiantes presentes

- ❑ Provincia de Nacimiento
- ❑ Edad
- ❑ Género
- ❑ Estatura
- ❑ Peso
- ❑ N° de Materias Aprobadas
- ❑ Tipo de Música Favorita
- ❑ Percepción de ESPOL
- ❑ ¿Tiene Laptop?

Acciones a realizar

1. Crear un repositorio de datos, usando diferentes formatos:
2. XLSX; CSV; TXT
3. Describir las variables y sugerir la codificación de las mismas
4. Determinar el tamaño de la población de estudio

PARTE 2

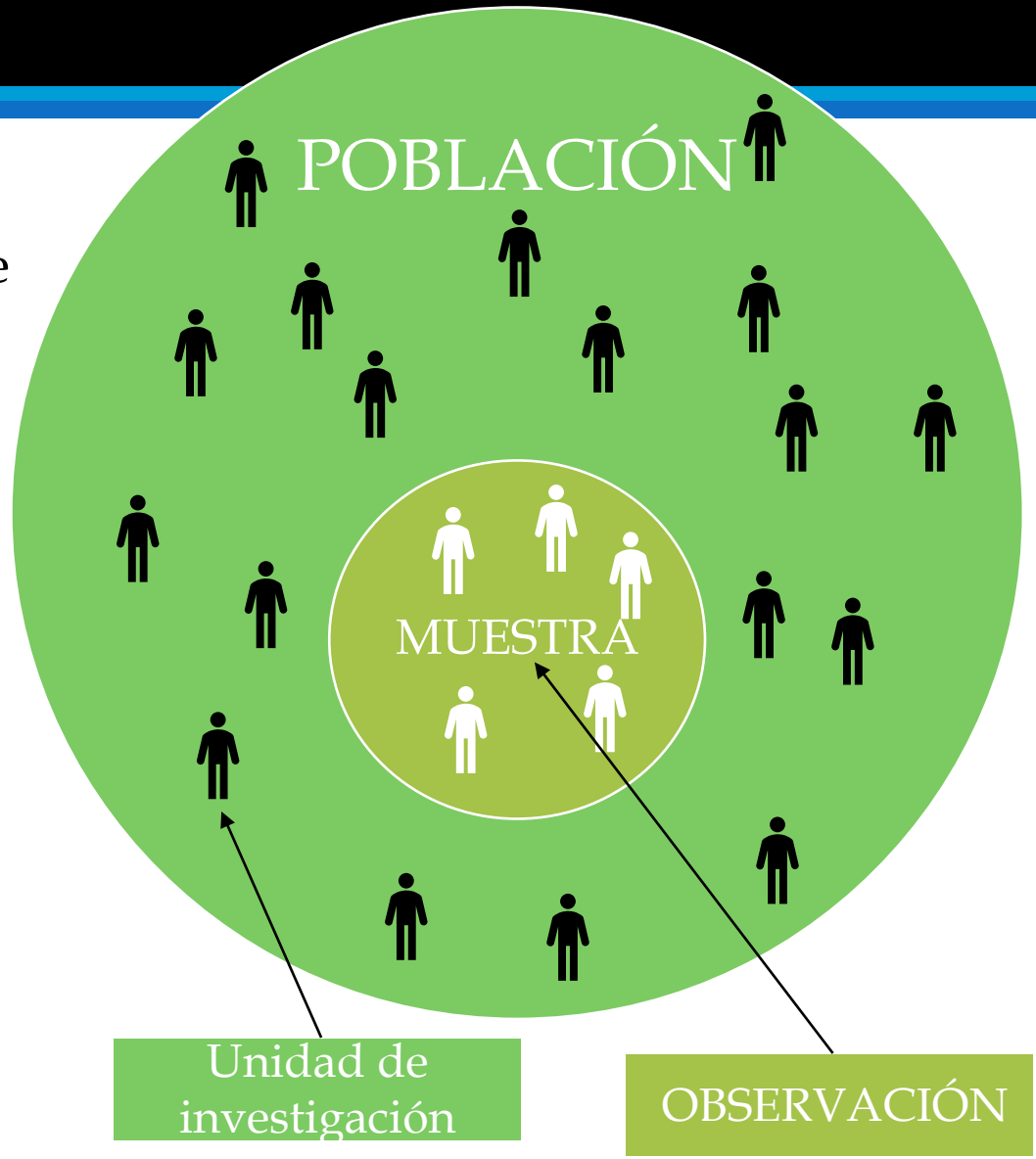
Conceptos básicos

Población objetivo
Conjunto bien definido de **N** elementos que son objeto de medición

Unidades de investigación
Elementos de la Población Objetivo, cuyas características son las variables (cuantitativas o cualitativas).

Muestra
Subconjunto de n unidades de investigación tomadas de la población objetivo. $n < N$

Observación muestral o simplemente observación
Cada uno de los valores incluidos en la Muestra.



Conceptos básicos

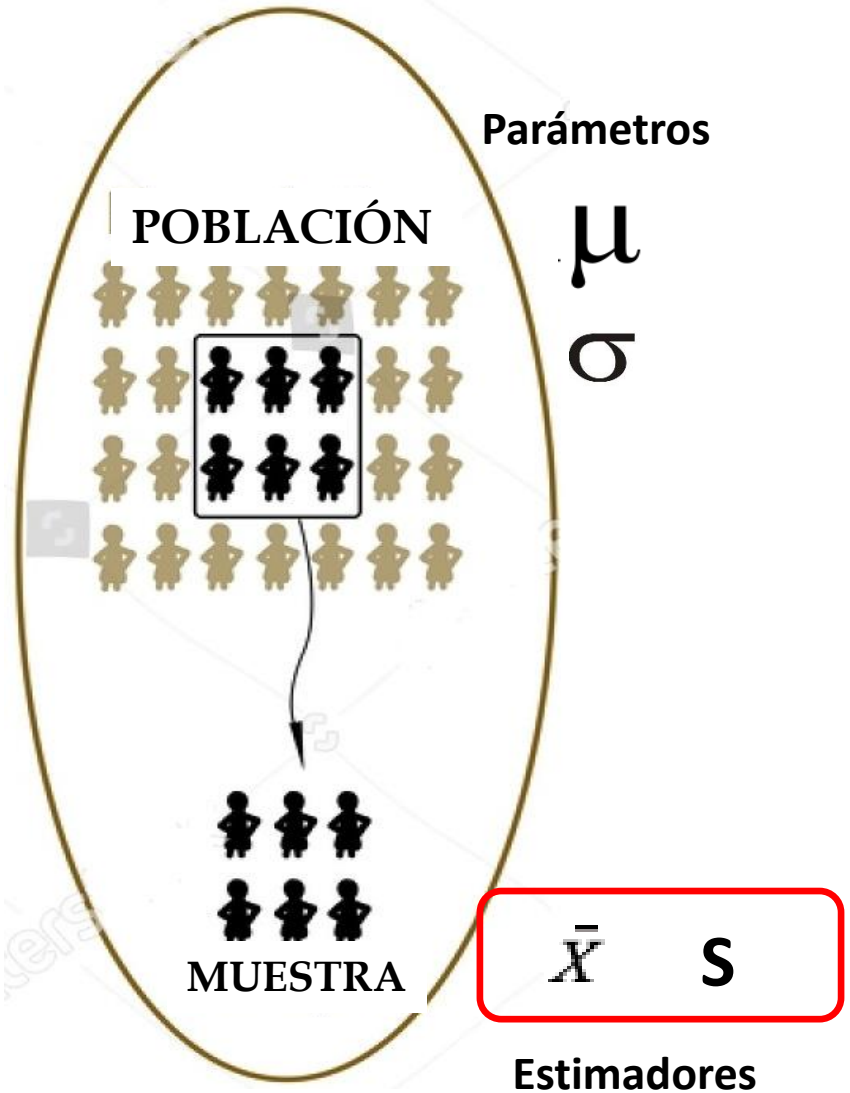
Parámetro

Es una cantidad numérica calculada a partir de los elementos de una **población.**

Estimador o estadístico:

Es una cantidad numérica calculada a partir de los elementos de una **muestra.**

Normalmente nos interesa conocer un parámetro, pero por la dificultad que conlleva estudiar a *TODA* la población, calculamos un estimador sobre una muestra y "confiamos" en que sean próximos. Más adelante veremos como elegir muestras para que el error sea "confiablemente" pequeño.



Tablas y frecuencias

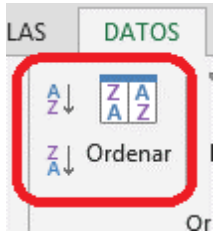
¿Qué contiene la muestra?

¿Qué debo hacer para obtener información?

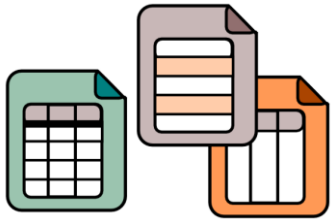
ESTADÍSTICA
Fernando Sandoya, PhD.
Facultad de Ciencias Naturales y Matemáticas

Tablas y frecuencias

1. Ordenar los datos

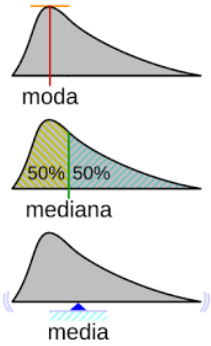


2.Tabular los datos ordenados



3. Graficar los datos ordenados

4. Calcular a partir de la muestra medidas de interés para la toma de decisiones.



Tablas y frecuencias

Tabla de frecuencias

Instrumento diseñado para la agrupación de datos y así facilitar su interpretación

Frecuencia Absoluta

Frecuencia Absoluta Acumulada

Frecuencia Relativa

Frecuencia Relativa Acumulada

Tablas y frecuencias

Partición de la Población:

Cuando analizamos una población (o una muestra), muchas veces es conveniente organizar los datos en clases o categorías. Estas deben ser tales que:

- Cada observación ***debe*** pertenecer a una, y sólo una clase o categoría.
- Todas las categorías o clases deben ser disjuntas.
- Se puede redefinir la longitud, el número de clases y los extremos de cada clase de tal manera que las clases tengan la misma longitud y los intervalos de cada clase incluyan a todos los datos, sean excluyentes y los valores de los extremos de cada clase sean simples

- Una vez definida una partición se pueden construir tablas de frecuencias o Gráficos de Barras, de pastel, etc.

Tablas y frecuencias

Cuando tenemos variables cualitativas las categorías son aquellas definidas por los valores que toma la variable.

Cuando tenemos variables cuantitativas las categorías o clases son definidas en base a intervalos en los cuales toman valores las variables.

	F. ABSOLUTA	F. ABSOLUTA ACUMULADA	F. RELATIVA	F. RELATIVA ACUMULADA
VARIABLES NOMINALES	X		X	
ORDINALES, CUANTITATIVAS	X	X	X	X

Tablas y frecuencias

Frecuencia Absoluta.

Se llama frecuencia absoluta de la clase c_i al número total de individuos u observaciones que pertenece a dicha clase y se denota por f_i.

Como las clases c_1, c_2, ... c_k son una partición de la muestra, es fácil verificar que

$$n = \sum_{i=1}^{k} f_i$$ ← número total de observaciones o tamaño de la muestra

Tablas y frecuencias

Frecuencia Absoluta Acumulada.

Se llama frecuencia absoluta acumulada de la clase c_i al número total de individuos u observaciones que pertenece desde el menor valor hasta dicha clase (es decir el número de individuos que se acumulan hasta ese valor) y se denota por F_i.

Esta frecuencia acumulada se puede calcular para variables cuantitativas o cualitativas ordinales.

Como las clases c_1, c_2, ... c_k son una partición de la muestra, es fácil verificar que

$F_k = n$ ← número total de observaciones o tamaño de la muestra

Tablas y frecuencias

Frecuencia Relativa.

Se llama frecuencia relativa de la clase c_i a la proporción de individuos que pertenecen a la clase sobre el total de individuos o tamaño de la muestra. Se de nota por f_i/n. Se puede verificar que:

$$\frac{f_i}{n} \qquad \text{nótese que ....} \quad \sum_{i=1}^{k} f_i / n = 1$$

Generalmente se representan como un porcentaje

Tablas y frecuencias

<u>Frecuencia Relativa Acumulada</u>.

Se llama frecuencia relativa acumulada de la clase c_i a la proporción de individuos que se acumulan hasta la clase i, es decir la suma de las frecuencias relativas hasta esa clase. Se denota por F_i. Se puede verificar que

$$F_k = 1$$

Se aplican sobre variables cuantitativas organizadas en categorías o para variables cualitativas ordinales.

Tablas y frecuencias

Cuando tenemos variables cualitativas las categorías son aquellas definidas por los valores que toma la variable.

Cuando tenemos variables cuantitativas las categorías o clases son definidas en base a intervalos en los cuales toman valores las variables.

	F. ABSOLUTA	F. ABSOLUTA ACUMULADA	F. RELATIVA	F. RELATIVA ACUMULADA
VARIABLES NOMINALES	X		X	
ORDINALES, CUANTITATIVAS	X	X	X	X

Tablas y frecuencias

Ejemplo: Se realiza un estudio sobre el origen de contagio de una enfermedad estacional. Un grupo de 25 pacientes se selecciona al azar y se investiga el origen del contagio, registrándose los siguientes resultados: HOG TRA TRA EXT HOG HOG HOG EXT EXT EXT TRA TRA HOG TRA EXT HOG TRA EXT EXT EXT HOG TRA TRA HOG HOG

Donde HOG = en el hogar, TRA= en su sitio de trabajo y Ext= en otros ambientes.

a. ¿Cuál es la unidad experimental?

b. ¿Cuál es la variable que se mide? ¿Es cualitativa o cuantitativa?

c. Hallar la tabla de frecuencias

Tablas y frecuencias

Cuando se tiene una variable **cuantitativa,** para representarla con una tabla de frecuencias esta se debe agrupar previamente en clases o categorías de clasificación. Por ejemplo, se podría medir los ingresos de personas.

Como regla práctica, el número de clases debe ser de 5 a 12, de igual ancho; cuantos más datos haya, más clases se requieren. Las clases deben ser escogidas para que cada una de las mediciones caiga en una clase y sólo en una.

Tablas y frecuencias

Para encontrar las clases o categorías se debe realizar lo siguiente:

k= Número de clases

MIN= mínimo de los valores

MAX= máximo de los valores

Rango = MAX-MIN

Ancho=rango/k

Clases= [Min, min+Ancho), [min+Ancho, min+2 Ancho)…

[Min+(k-1)Ancho, +∞).

ESTADÍSTICA
Fernando Sandoya, PhD.
Facultad de Ciencias Naturales y Matemáticas

Tablas y frecuencias

Ordinal	Clase	Marca de clase (mi)	Frecuencia Absoluta (fi)	Frecuencia Acumulada	Frecuencia relativa	Frecuencia relativa acumulada
1	$[a_1,a_2)$	$(a_1+a_2)/2$	f_1	$F_1=f1$	f_1/n	F_1/n
2	$[a_2,a_3)$	$(a_2+a_3)/2$	f_2	$F_2=f1+f2$	f_2/n	F_2/n
3	$[a_3,a_4)$	$(a_3+a_4)/2$	f_3	$F_3=f_1+f_2+f_3$	f_3/n	F_3/n
K	$[a_k,a_{k+1})$	$(a_k+a_{k+1})/2$	f_k	$F_k=n$	f_k/n	$F_k/n=1$

Número de observaciones = n

Cantidad de clases recomendadas (k): 5 – 12

Ancho de clase = Rango/ k

Tablas y frecuencias

Ejemplo: Se han registrado los pesos (en libras) de 30 bebés de gestación completa al momento de nacer:

7.2 7.8 6.8 6.2 8.2 8.0 8.2 5.6 8.6 7.1 8.2 7.7 7.5 7.2 7.7 5.8 6.8 6.8 8.5 7.5 6.1 7.9
9.4 9.0 7.8 8.5 9.0 7.7 6.7 7.7

Gráficos

Ejemplo: Se está estudiando la presencia de obesidad entre mujeres de clase media que se dedican al hogar, en una muestra de 40 hogares se encontraron los siguientes pesos:

112, 113, 117, 150, 152, 153, 119, 107, 108, 111, 160, 161, 163, 120, 123, 123, 124, 126, 128, 131, 132, 132, 134, 135, 136, 137, 138, 140, 141, 142, 143, 143, 145, 147, 148, 148, 153, 158,158, 160

Gráficos

K := N° Clases = 7
R := Rango = máx { x_i } - mín { x_i } = 163 - 107 = 56
A := Ancho = R / K = 55 / 7 = 9

Límites	Marca	Conteo	Frecuencias ABS - REL - REL. AC.
102,5-111,5	107	///	3
111,5-120,5	116	~~////~~	5
120,5-129,5	125	~~////~~	5
129,5-138,5	134	~~////~~ ///	8
138,5-147,5	143	~~////~~ //	7
147,5-156,5	152	~~////~~ /	6
156,5-165,5	161	~~////~~ /	6

Tablas y frecuencias

1. Realizar los ejemplos de las láminas anteriores en Excel.
2. Utilizar el libro de Excel con las órdenes de adquisiciones para el laboratorio de Química y establecer frecuencias según una variable cualitativa
3. Utilizar el libro de Excel con las órdenes de adquisiciones para el laboratorio de Química y establecer frecuencias según una variable cuantitativa

Gráficos

Una forma de representar a los datos de una muestra, cuando estos se han agrupado en diferentes clases o categorías son los gráficos.

Existen diferentes tipos de gráficos, que se usan según el contexto. En particular para datos agrupados en categorías son recomendables los siguientes gráficos:

Variables Nominales

- Gráficos de barras verticales
- Gráficos de pastel

Variables Ordinales

- Histogramas de frecuencias
- Gráficos de pastel
- Gráficos de Pareto
- Ojiva

Variables Ordinales

- Histogramas de frecuencias

Variables Cuantitativas

- Diagramas de dispersión, gráficos de líneas

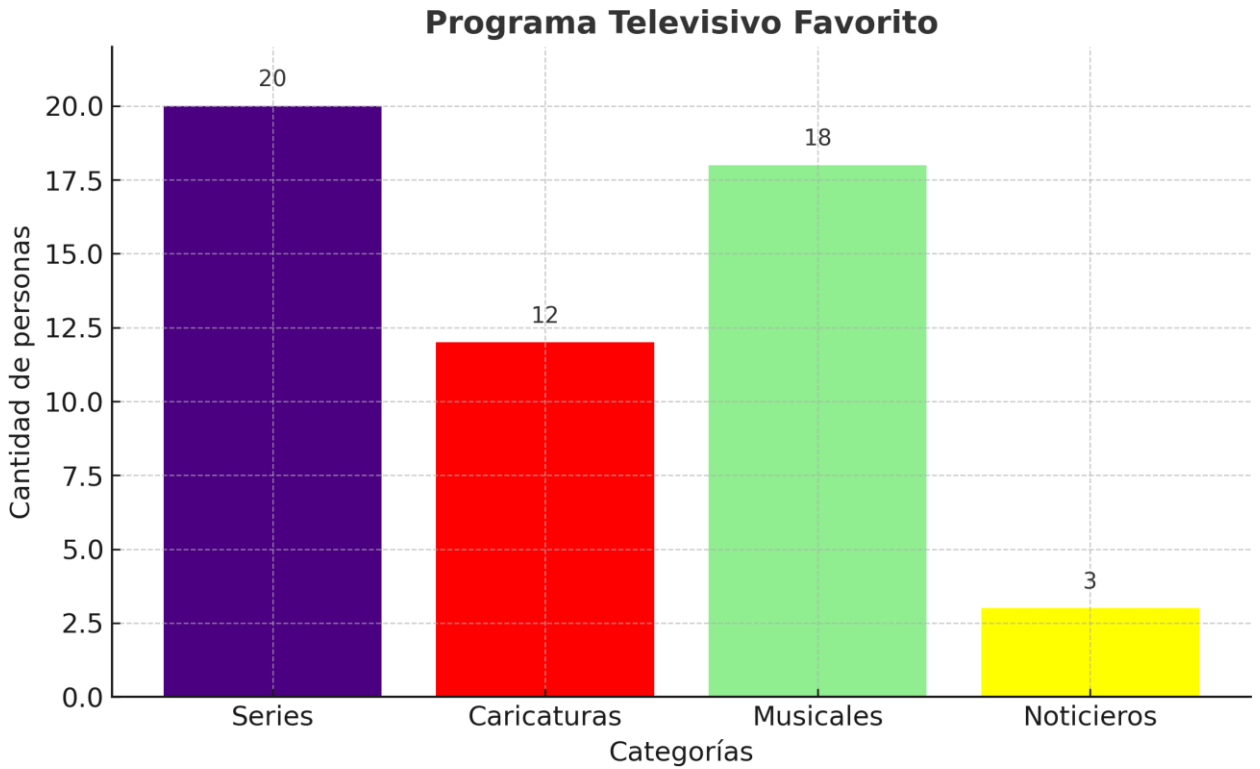
Estadística descriptiva
Tipos de gráficos

Nombre del gráfico	Muestra	Notas
Gráfico de dispersión	Relación entre dos variables numéricas	
Gráficos de linea	Relación entre dos variables numéricas	Se utiliza cuando existe un orden secuencial para la variable x, por ejemplo, el tiempo
Histograma	Distribución de una variable numérica	
Gráfico de cajas	Distribución de 1 variable numérica dividida por los valores de otra variable	
Gráfico de barras	Distribución de 1 variable categórica	

Estadística descriptiva
Tipos de gráficos

GRÁFICOS DE BARRAS

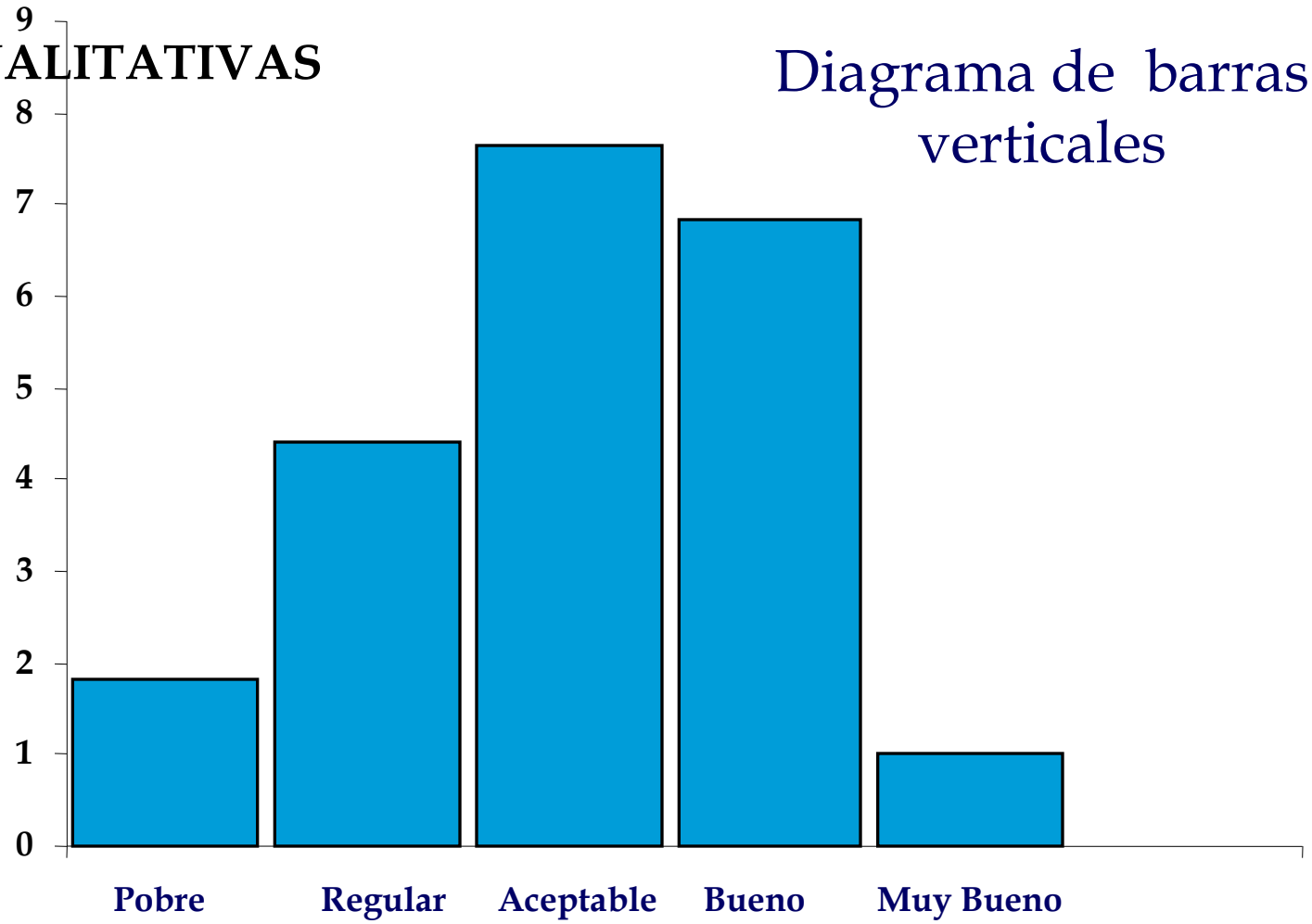
El gráfico de barras contiene en el eje de las X las categorías de la variable y en el eje Y las frecuencias absolutas.



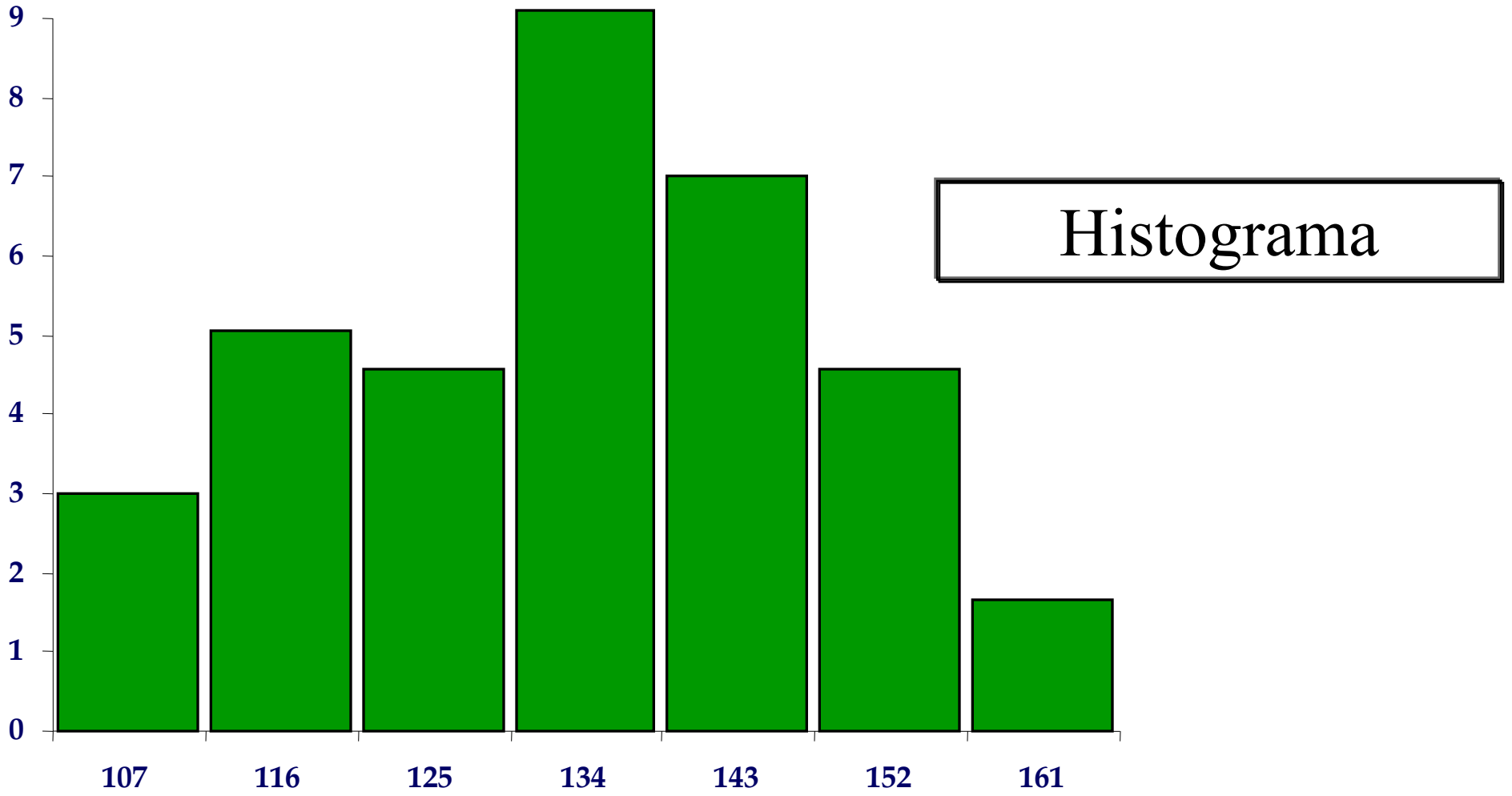
Estadística descriptiva
Tipos de gráficos

VARIABLES CUALITATIVAS

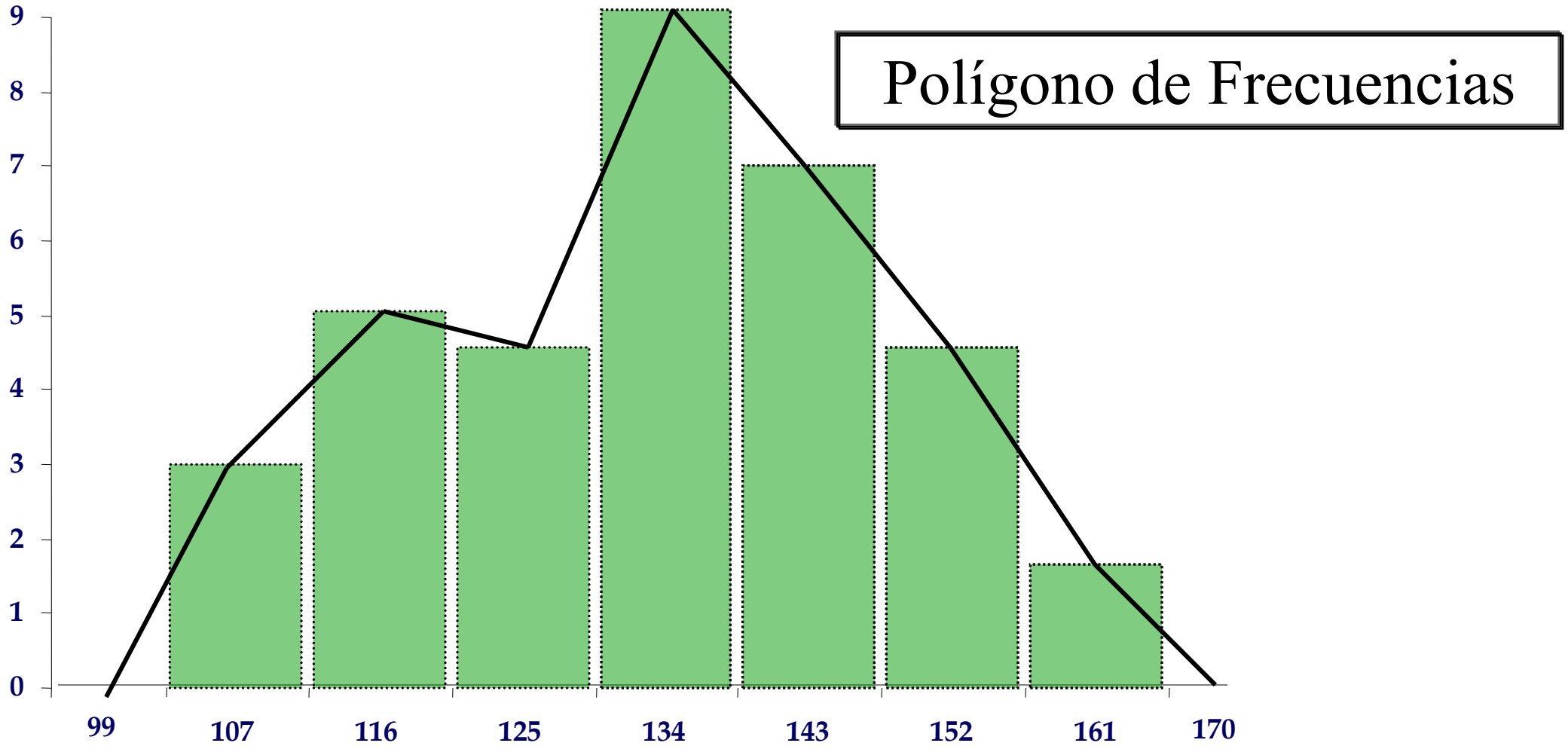
Diagrama de barras verticales



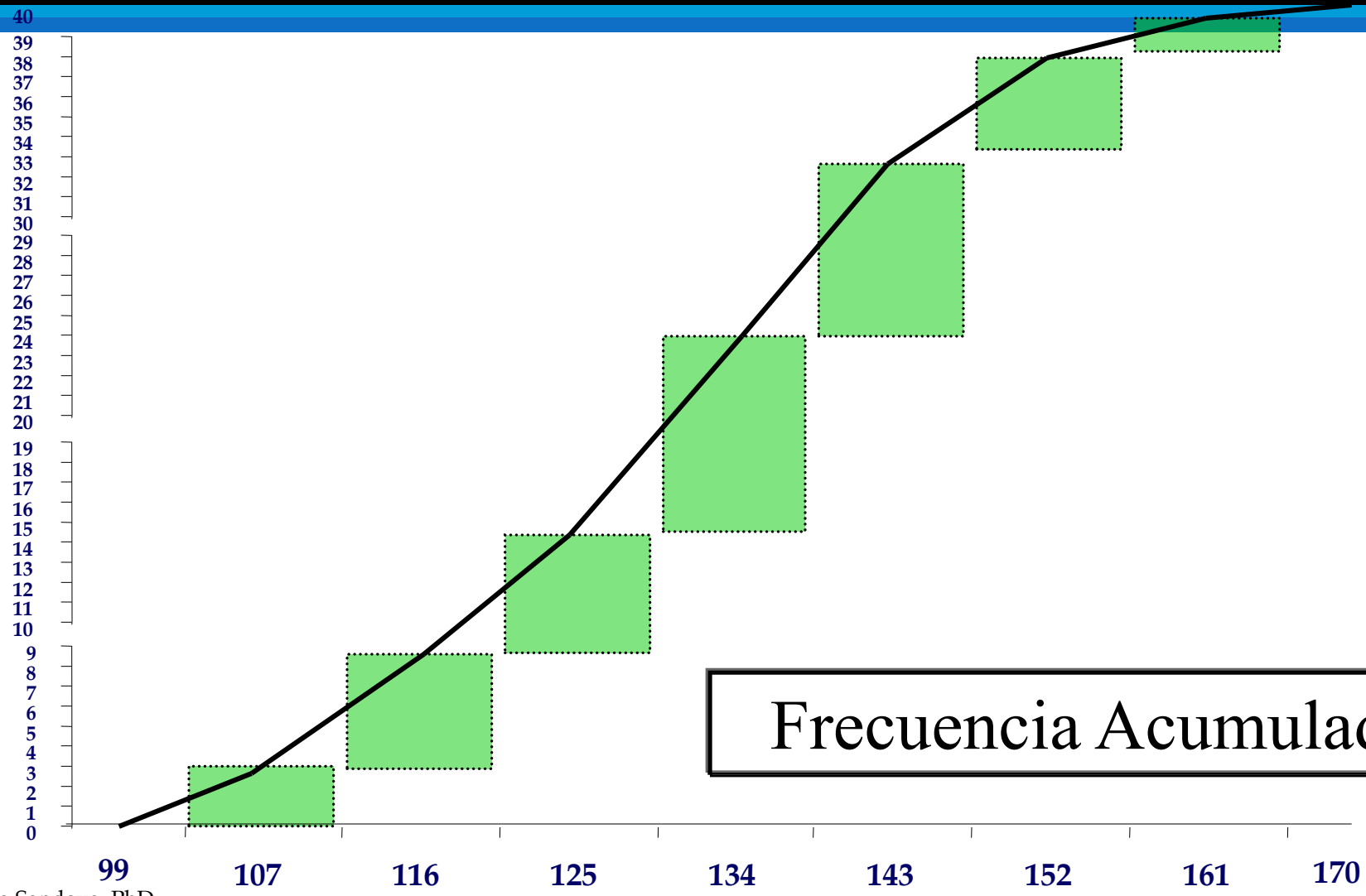
Estadística descriptiva
Tipos de gráficos



Estadística descriptiva
Tipos de gráficos



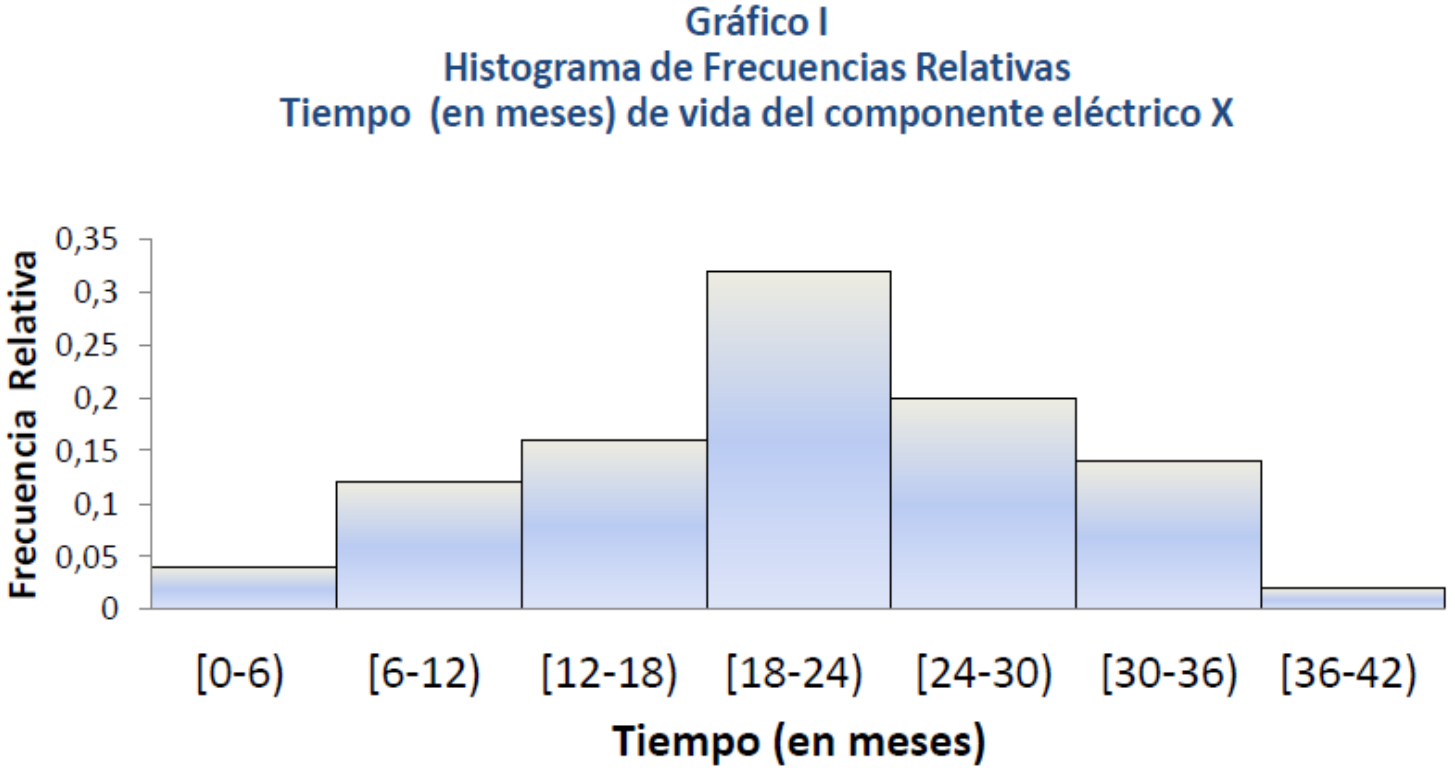
Estadística descriptiva
Tipos de gráficos



Frecuencia Acumulada: Ojiva

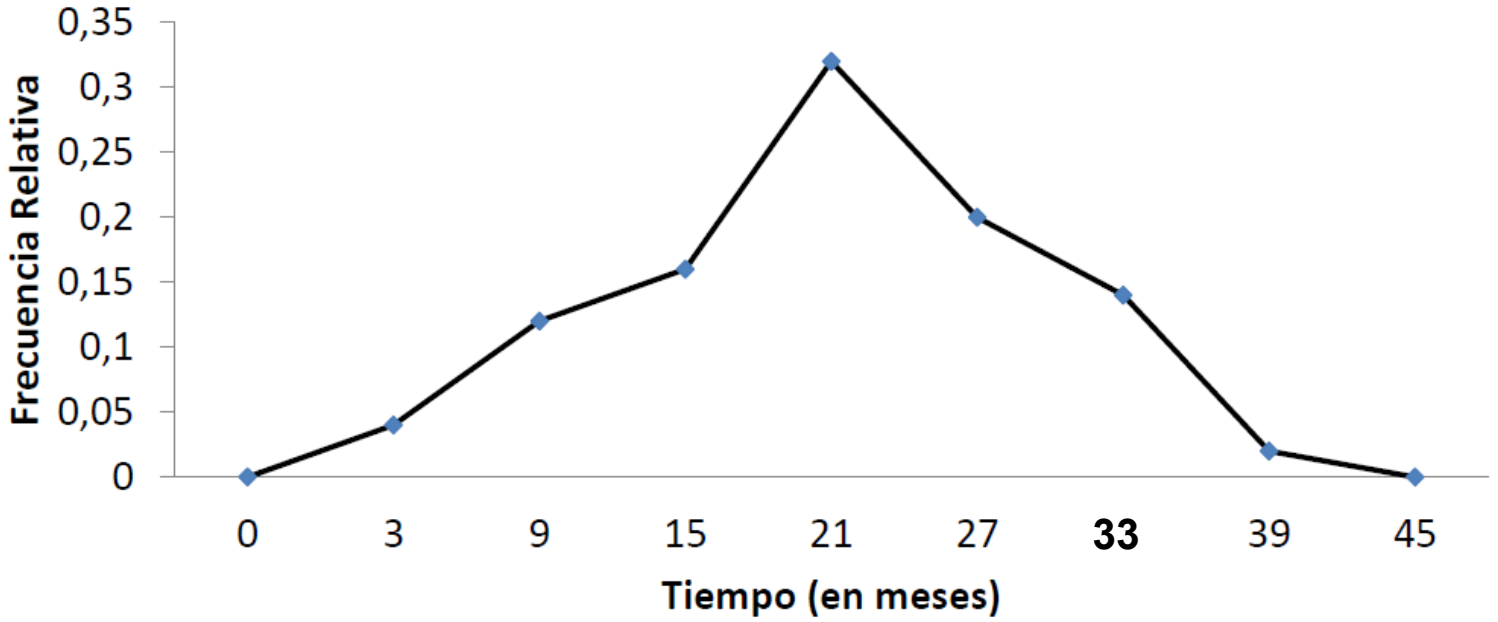
GRÁFICOS – HISTOGRAMA DE FRECUENCIAS

El histograma en un gráfico bidimensional en cuyo eje de las X se encuentran las clases y en el eje Y las frecuencias relativas o absolutas.



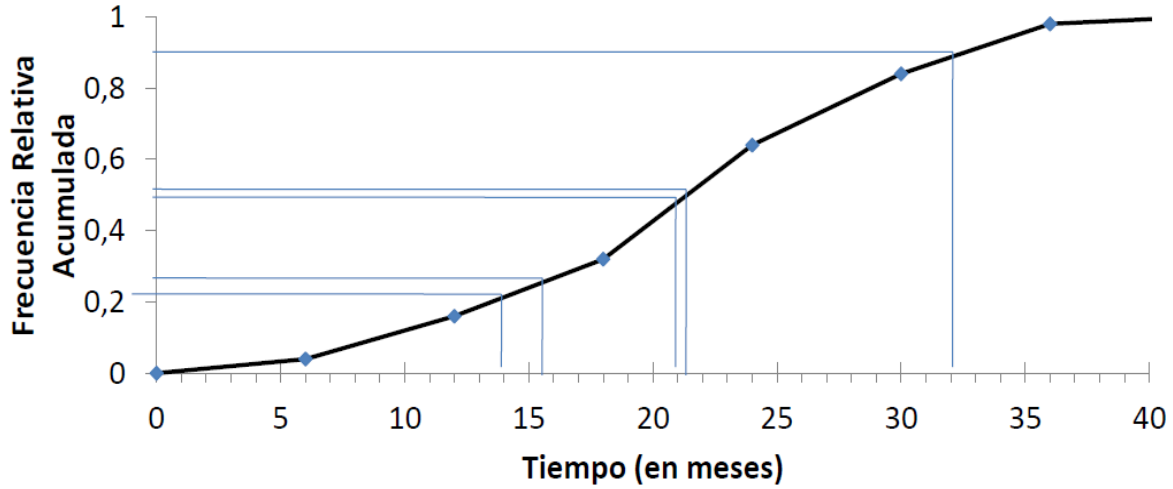
GRÁFICOS – POLÍGONO DE FRECUENCIAS

El Polígono es un gráfico bidimensional en cuyo eje **X** se encuentran las marcas de clase y en el eje **Y** las frecuencias relativas o absolutas.



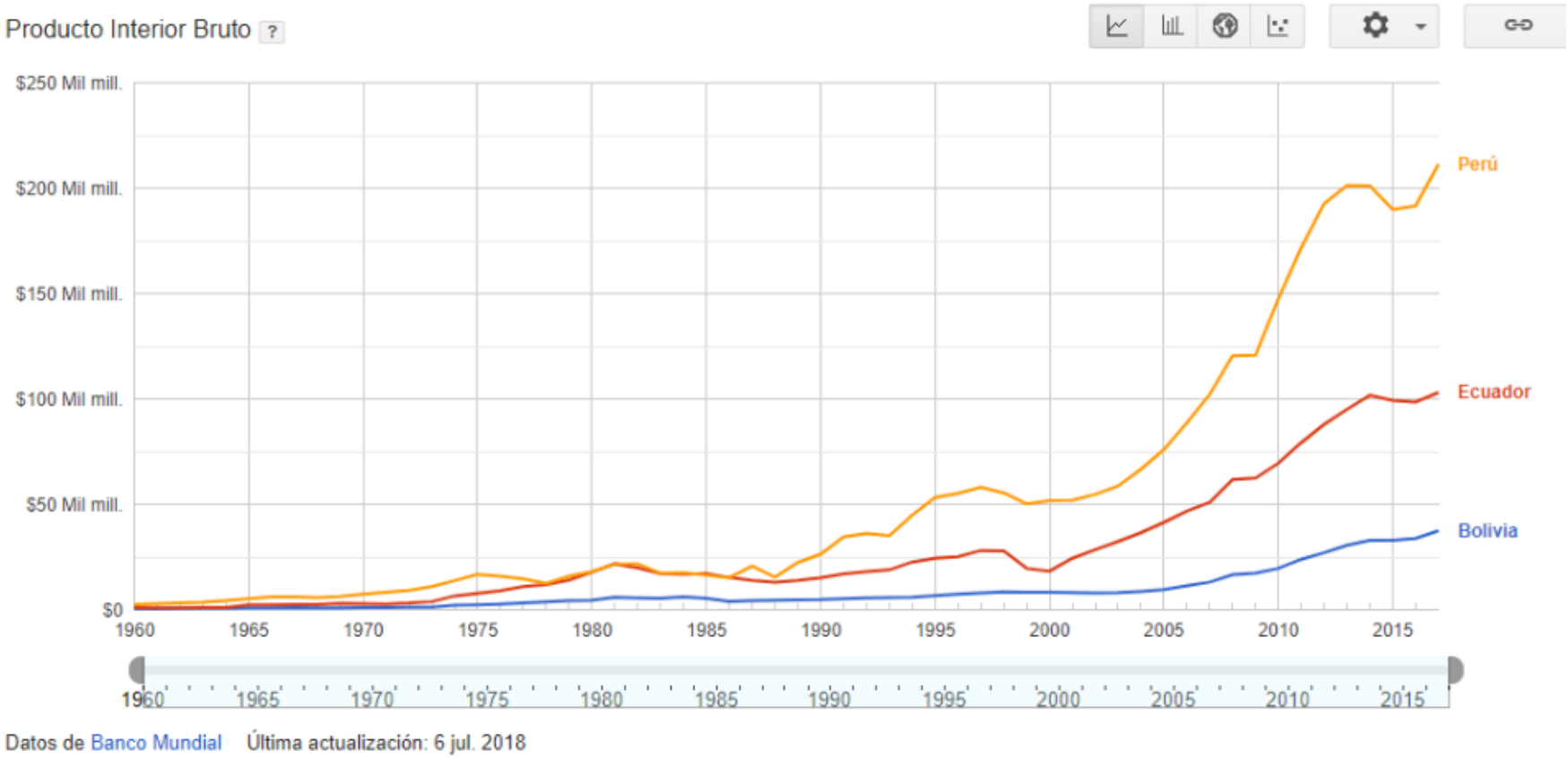
GRÁFICOS – OJIVAS

Es un gráfico que presenta en el eje horizontal las marcas de clase de la característica cuantitativa que se está investigando y en el eje vertical la **frecuencia relativa acumulada** .



Estadística descriptiva
Tipos de gráficos

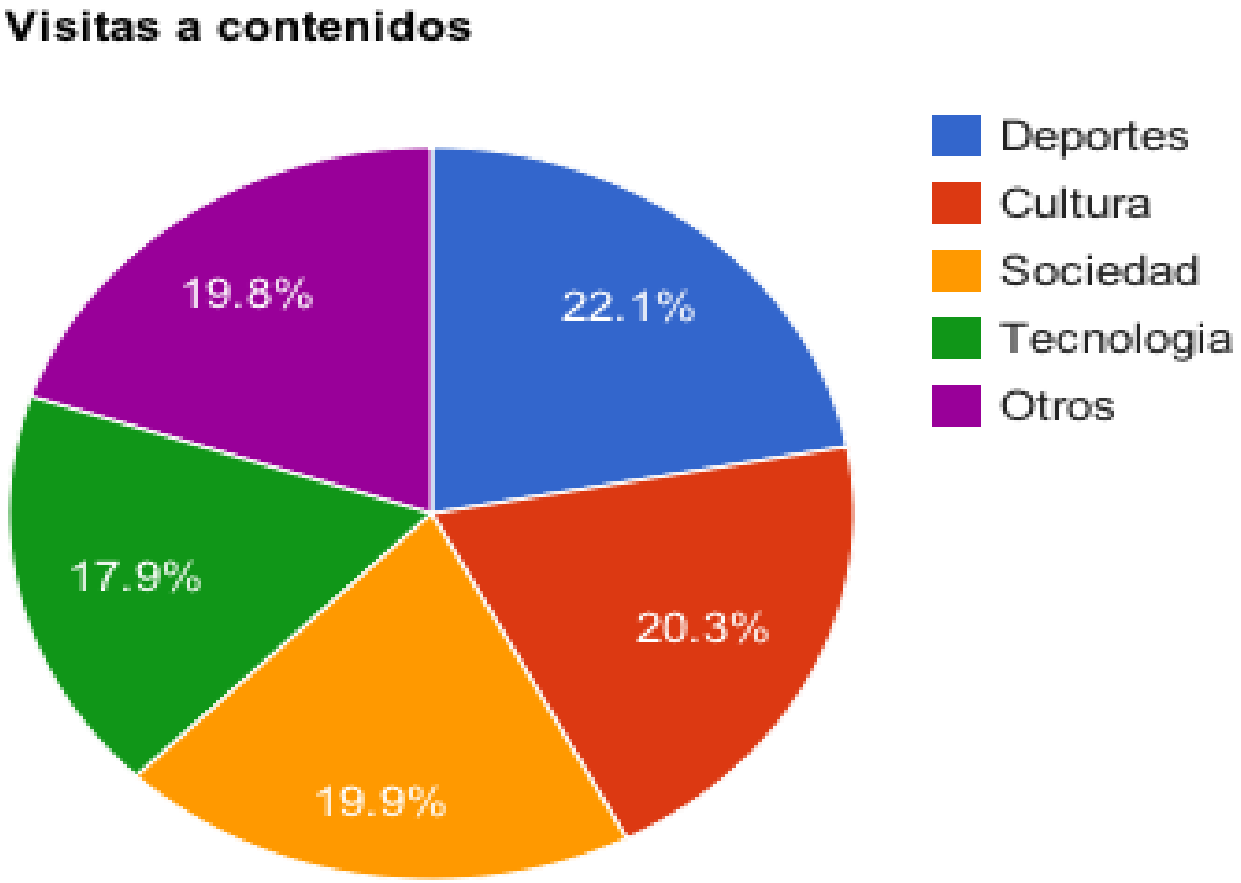
OTROS GRÁFICOS – SERIES TEMPORALES



Estadística descriptiva
Tipos de gráficos

GRÁFICOS – PASTEL

Gráfico circular cuya área se divide en sectores que representan los resultados porcentuales para una variable que generalmente puede ser cualitativa.

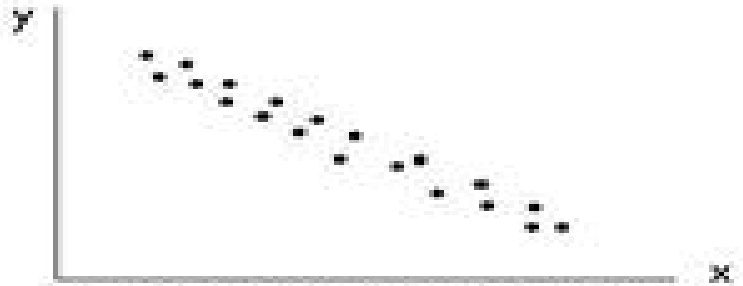


Estadística descriptiva
Tipos de gráficos

OTROS GRÁFICOS – DISPERSIÓN

Se realiza para visualizar la relación entre dos variables, dependiendo de la forma que toma la "nube" de datos.

Fuerte correlación lineal positiva
r cerca de 1



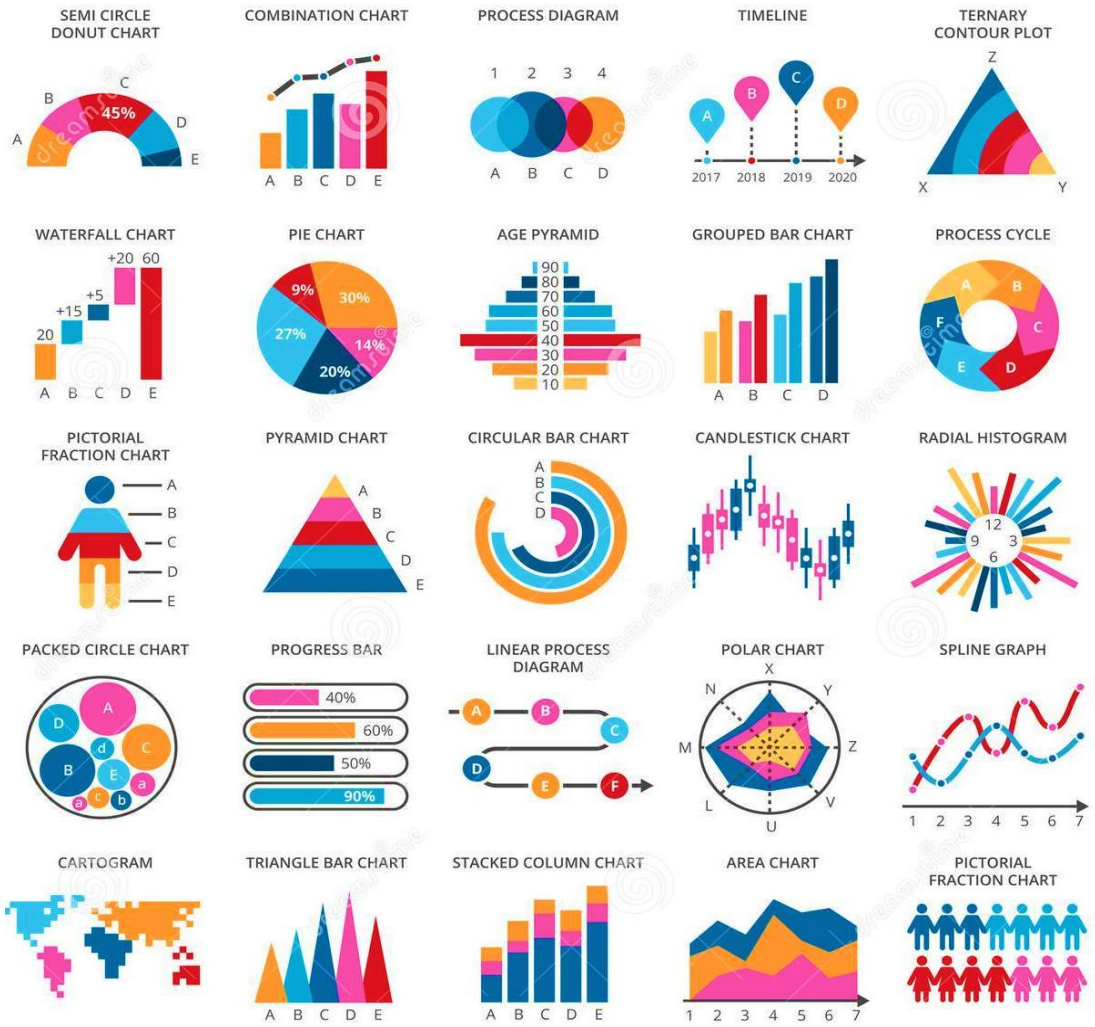
Fuerte correlación lineal negativa
r cerca de -1

Ninguna correlación lineal aparente
r cerca de cero

Correlación curvilínea
r cerca de cero

Estadística descriptiva
Tipos de gráficos

OTROS TIPOS DE GRÁFICOS ESTADÍSTICOS



Estadística descriptiva
Practica 1.1

Con el archivo *Adquisiciones laboratorio Química.xlsx* realizar lo siguiente:

1. Considerar la variable **estado_odc (estado de la orden de compra)**.
2. Determinar la tabla de frecuencias: absolutas y relativas.
3. Construir los gráficos de barras y de pastel, ¿qué puede comentar al respecto?
4. Hacer lo mismo respecto a la variable **cod_prov** (código del proveedor).
5. Realizar 4 solo considerando órdenes de compra con el estado concluido.

ESTADÍSTICA
Fernando Sandoya, PhD.
Facultad de Ciencias Naturales y Matemáticas

PARTE 2:

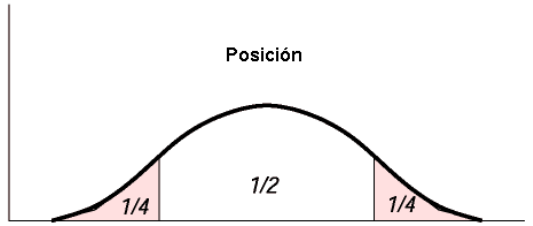
ANÁLISIS DE LA INFORMACIÓN: MEDIDAS DE TENDENCIA CENTRAL, VARIABILIDAD Y LOCALIZACIÓN

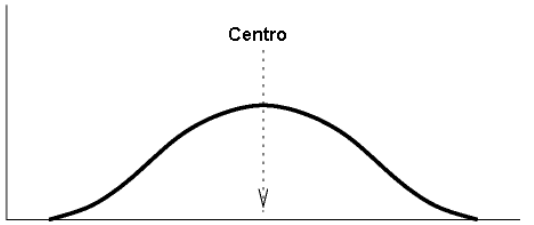
MEDIDAS

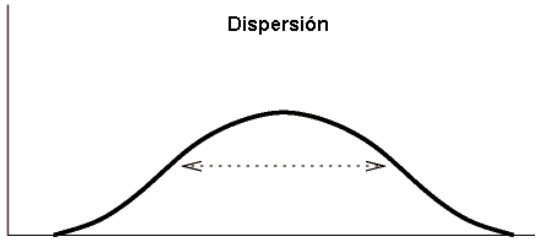
- Medidas de tendencia central
- Medidas de localización
- Medidas de dispersión/variabilidad
- Valores atípicos/outliers
- Gráficos de caja

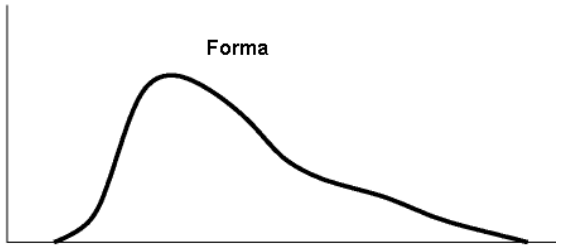
MEDIDAS

- Posición
  - Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos: Cuantiles, percentiles, cuartiles, deciles,...
- Centralización
  - Indican valores con respecto a los que los datos parecen agruparse: Media, mediana y moda
- Dispersión
  - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización: Desviación típica, coeficiente de variación, rango, varianza
- Forma
  - Asimetría
  - Apuntamiento o curtosis









MEDIDAS

Estadísticos de Orden

Ordenar la muestra $\boldsymbol{X_{(i)}}$

$$\boldsymbol{X_{(1)} = mín \{X_1, X_2, \dots X_n\}}$$

$$\boldsymbol{X_{(n)} = máx \{X_1, X_2, \dots X_n\}}$$

Medidas de Tendencia Central

- Estas medidas tienden a ubicarse en el centro del conjunto de datos.
- Proporcionan un valor simple y representativo, que resume un gran volumen de información.

- Media
- Moda
- Mediana

Medidas de Tendencia Central

Medidas de tendencia central

Media Aritmética - Promedio

Suma de los valores de todos los datos dividido entre la cantidad de datos

$$\bar{x} = \sum_{i}^{n} \frac{X_i}{n}$$

Mediana

Valor para el cual el 50% de los datos son menores o iguales al mismo; valor ubicado en el punto medio una vez que se han ordenado los datos.

Moda

Elemento observado que más se repite

Medidas de dispersión

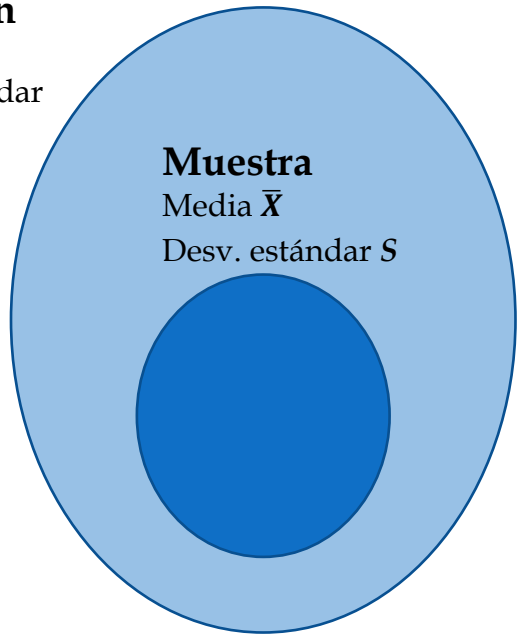
Son valores que proveen información adicional acerca del comportamiento de los datos describiendo numéricamente su dispersión o VARIABILIDAD.

Medidas de dispersión

Media › | Mediana › | Percentil › | Desviación estándar ›

Es una medida de **dispersión**. Cuantifica la distancia promedio de los valores de una muestra respecto a su media. Se calcula como la raíz cuadrada del sumatorio de diferencias entre cada valor y la media de la muestra elevado al cuadrado. Su fórmula matemática sería la siguiente:

Población
Media μ
Desv. estándar
σ



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

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}}$$

Donde

S = *Desviación estándar muestral*
$\bar{X}$ = *Media*
X_i = *Dato i de la muestra*
n = *Tamaño de la muestra*

σ = *Desviación estándar poblacional*
μ = *Media*
X_i = *Dato i de la población*
N = *Tamaño de la población*

MUESTRA

$$Varianza = S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

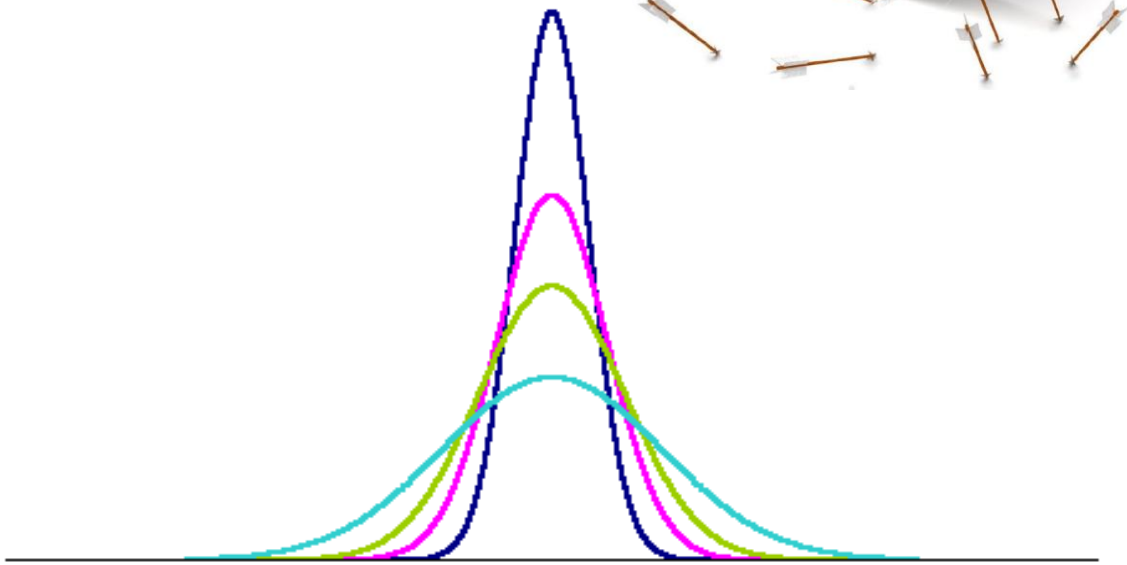
POBLACIÓN

$$Varianza = \sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}$$

Medidas de dispersión

Rango Muestral
Diferencia entre el máximo valor y el mínimo valor de la muestra

$$\boldsymbol{R = X_{(n)} - X_{(1)}}$$



Medidas de dispersión

Desviación típica

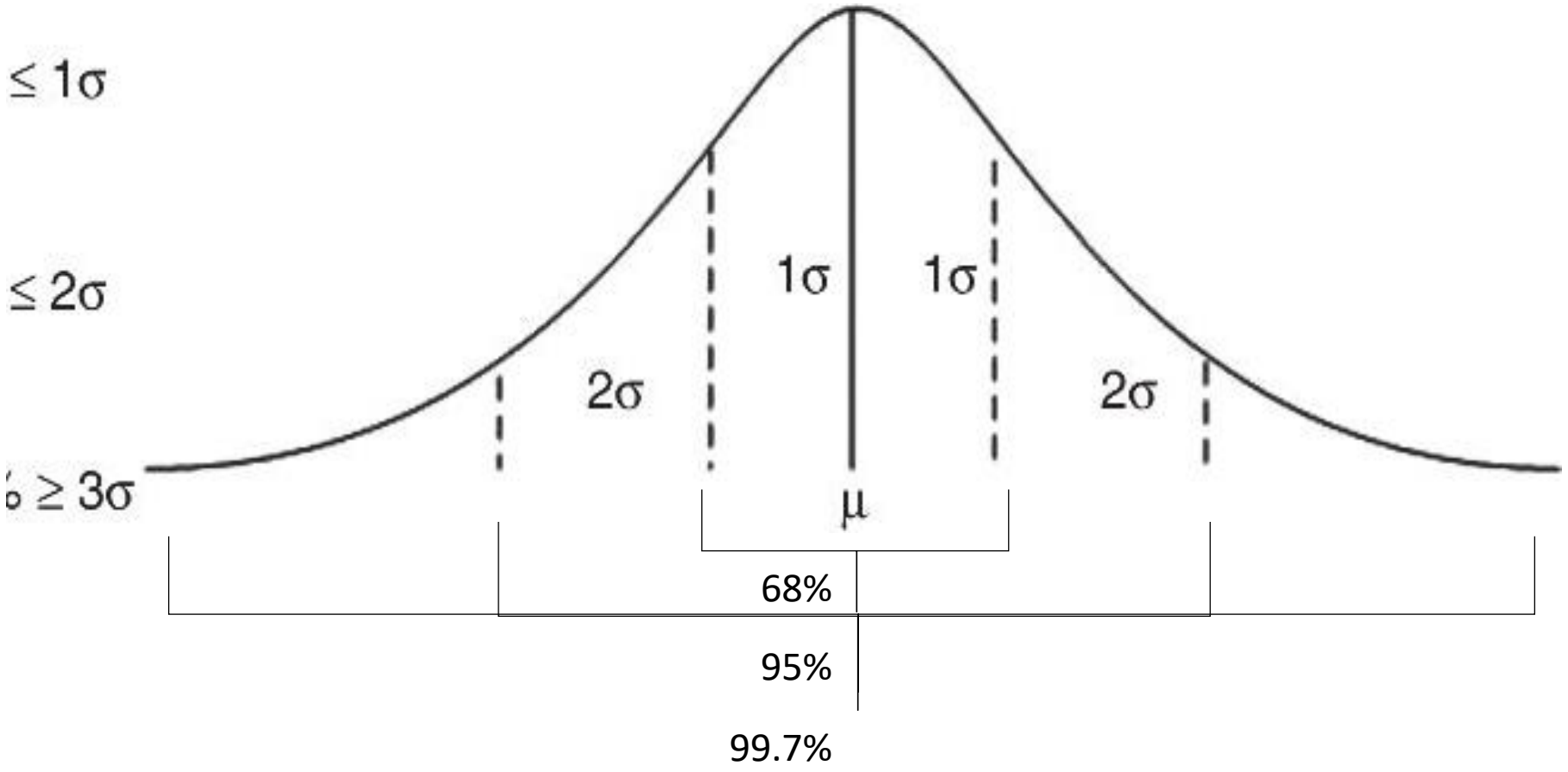
Por ejemplo, se ha registrado la edad de personas que han contraído una enfermedad contagiosa en ambientes de trabajo sometidos a mucho estrés, los resultados fueron las siguientes edades:

33 34 34 39 33 36 37 36 44 42 37 45 29 40 38 26 41 33 26 31 37 31 38 42 41 27 33 45 29 39 43 31 30 38 43 35 36 37 27 32

Determinar la desviación típica.

Medidas de dispersión

Regla Empírica



Medidas de posición

- Miden la “posición” de valores dentro del conjunto de datos respecto al total de los datos.

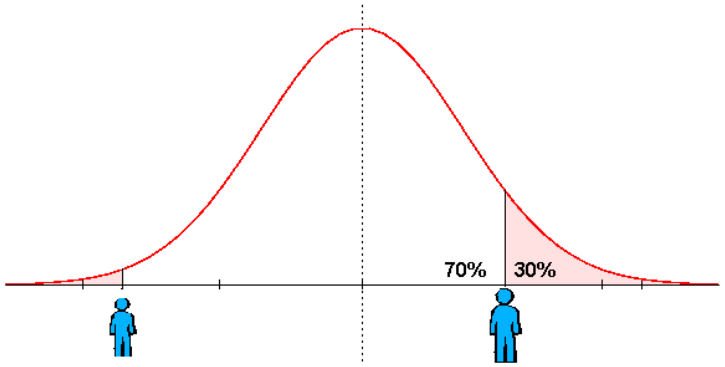
- Percentiles
- Cuartiles
- Deciles

Medidas de posición

- Se define el **cuantil** de orden α como un valor de la variable por debajo del cual se encuentra una frecuencia acumulada α.
- Casos particulares son los percentiles, cuartiles, deciles, quintiles,...

- **Percentil** de orden k = cuantil de orden k/100
  - La mediana es el percentil 50
  - El percentil de orden 15 deja por debajo al 15% de las observaciones. Por encima queda el 85%

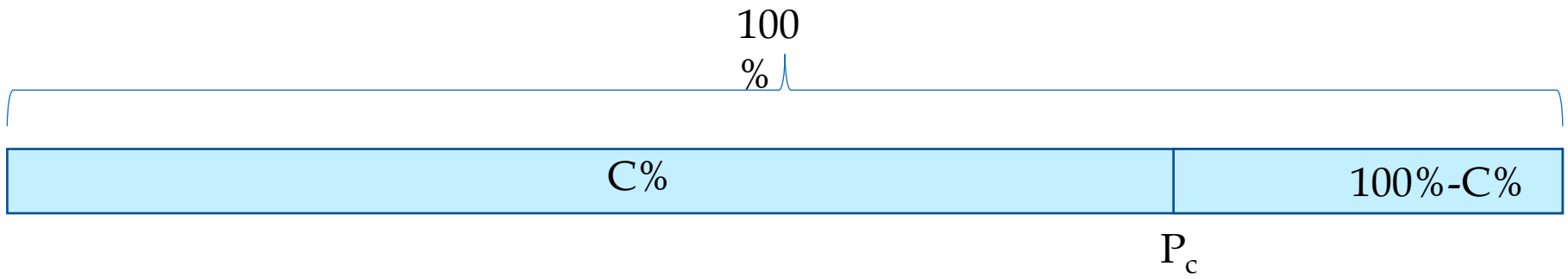
- **Cuartiles**: Dividen a la muestra en 4 grupos con frecuencias similares.
  - Primer cuartil = Percentil 25 = Cuantil 0,25
  - Segundo cuartil = Percentil 50 = Cuantil 0,5 = mediana
  - Tercer cuartil = Percentil 75 = cuantil 0,75



Medidas de posición

Percentiles

Los percentiles miden la posición relativa de un dato dentro de la secuencia ordenada de las observaciones. Un percentil **c** se representa con P_c, por ejemplo, el percentil 30 de una población (o muestra) se representa con P_{30}



Medidas de posición

Percentiles

Por ejemplo, se ha registrado la edad de personas que han contraído una enfermedad contagiosa en ambientes de trabajo sometidos a mucho estrés, los resultados fueron las siguientes edades:

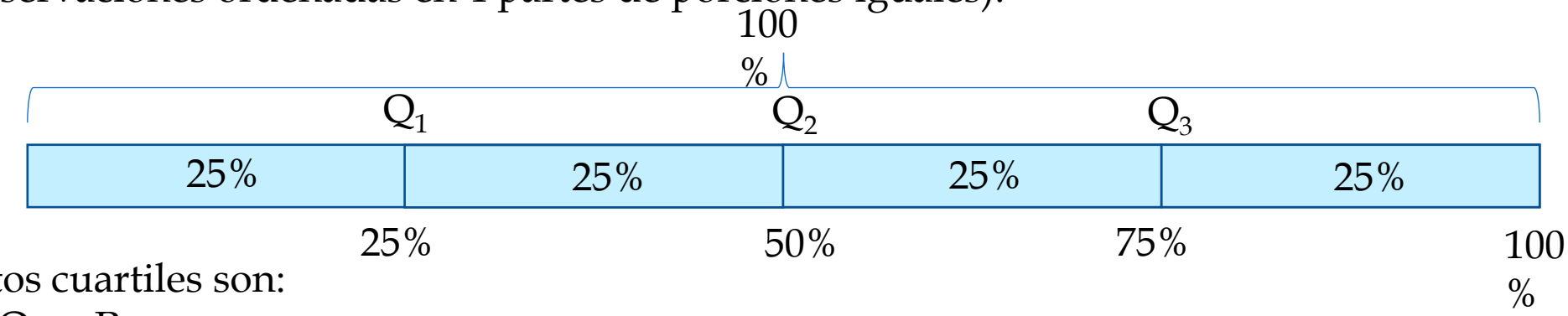
33 34 34 39 33 36 37 36 44 42 37 45 29 40 38 26 41 33 26 31 37 31 38 42 41 27 33 45 29 39 43 31 30 38 43 35 36 37 27 32

Determinar el percentil 90 de esas edades.

Medidas de posición

Cuartiles

Los cuartiles miden la posición relativa de un dato dentro de la secuencia ordenada de las observaciones en porcentajes de 25%, 50% y 75% (es decir dividen al total de las observaciones ordenadas en 4 partes de porciones iguales).



Estos cuartiles son:

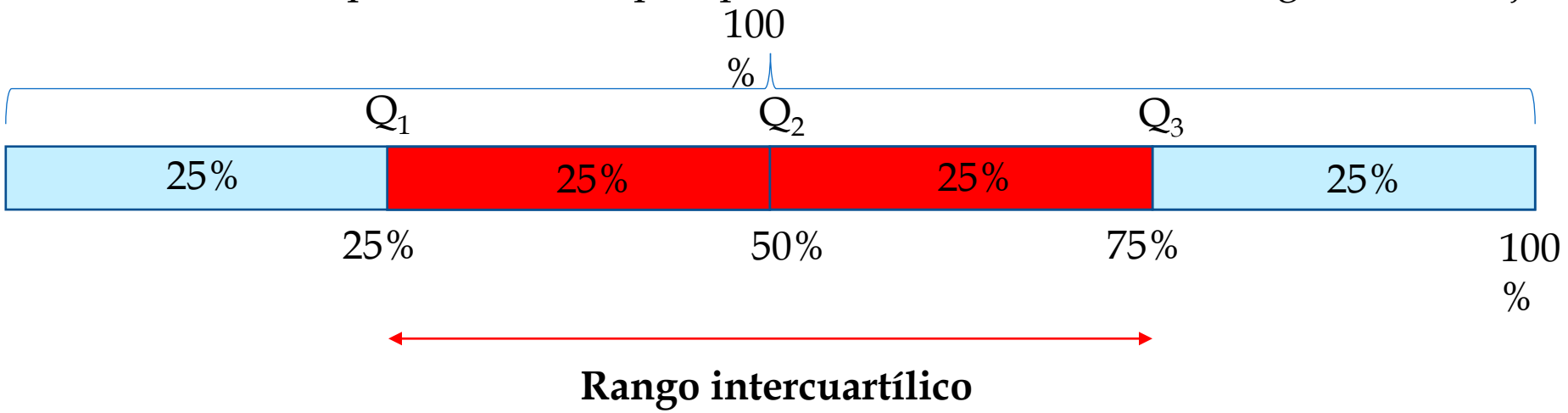
- $Q_1 = P_{25}$
- $Q_2 = P_{50}$
- $Q_3 = P_{75}$

En el ejemplo anterior: 33 34 34 39 33 36 37 36 44 42 37 45 29 40 38 26 41 33 26 31 37 31 38 42 41 27 33 45 29 39 43 31 30 38 43 35 36 37 27 32 Determinar los cuartiles de esas edades.

Medidas de posición

Cuartiles

La zona entre el primer y tercer cuartil es la zona más importante de los datos, se dice que son los datos más representativos, y la diferencia entre Q_3 y Q_1 se denomina RANGO INTERCUARTÍLICO, que es un valor que aparece en los denominados diagramas de caja



Estadística descriptiva
Medidas de Posición: Deciles

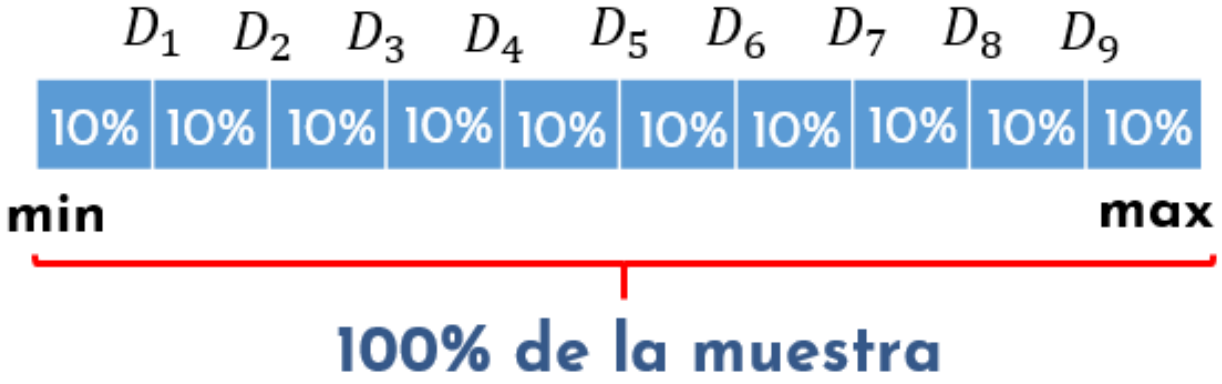
Son valores que dividen a los datos de la muestra en grupos de tamaño aproximado de 10%.

Primer Decil

- A la izquierda de D_1 están incluidos 10% de los datos (aproximadamente)
- A la derecha de D_1 están el 90% de los datos (aproximadamente)

Segundo Decil

- A la izquierda de D_2 están incluidos 20% de los datos (aproximadamente)
- A la derecha de D_2 están el 80% de los datos (aproximadamente), y así sucesivamente



Otras medidas de dispersión

Variables Cuantitativas:

Coeficiente o tasa de Variación, Varianza, desviación típica
Rango, Rango Intercuartílico (ICR)

Ejemplo: determinar estas medidas de dispersión para el ejemplo anterior

- RANGO = Max(Xi)-Min(Xi)
- RANGO INTERCUARTÍLICO = IQR = Q_3-Q_1

El coeficiente de variación es igual a la división de la desviación típica para la media y se expresa generalmente de manera porcentual. Permite la comparación de la variabilidad entre muestras de diferente magnitud

GRÁFICO DE CAJA (gráfico de caja y bigotes)

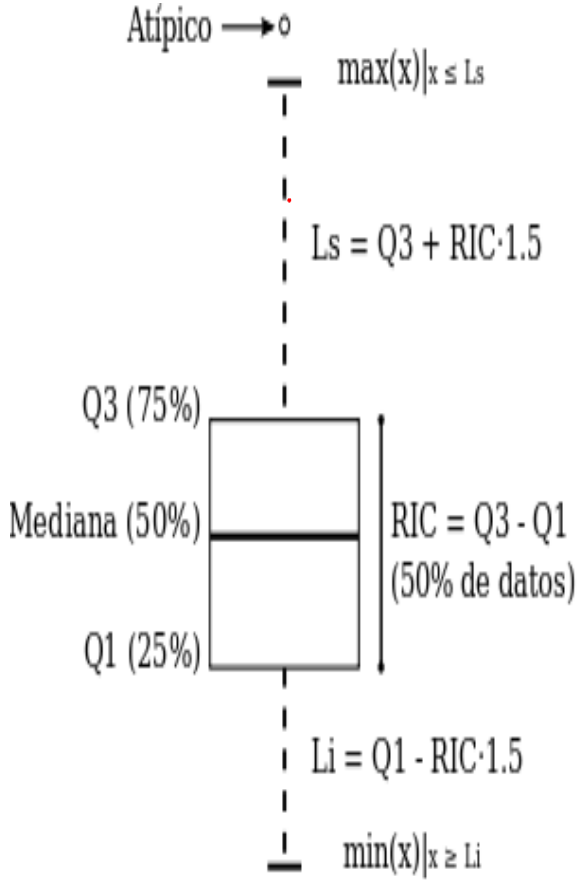
Primer Cuartil Q1
Valor de X tal que no más del 25% de los datos son menores al mismo

Segundo Cuartil Q2- Mediana
Valor de X tal que no más del 50% de los datos son menores al mismo

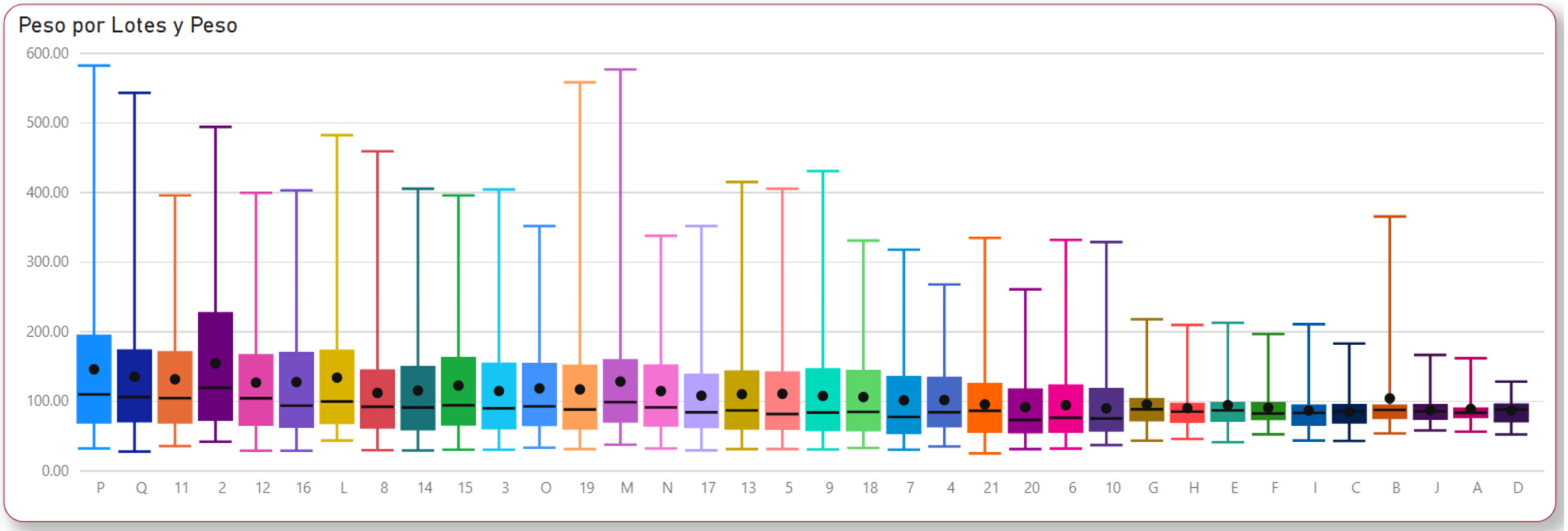
Tercer Cuartil Q3
Valor de X tal que no más del 75% de los datos son menores al mismo

Rango Intercuartil RI
Diferencia entre el tercer cuartil y el primero

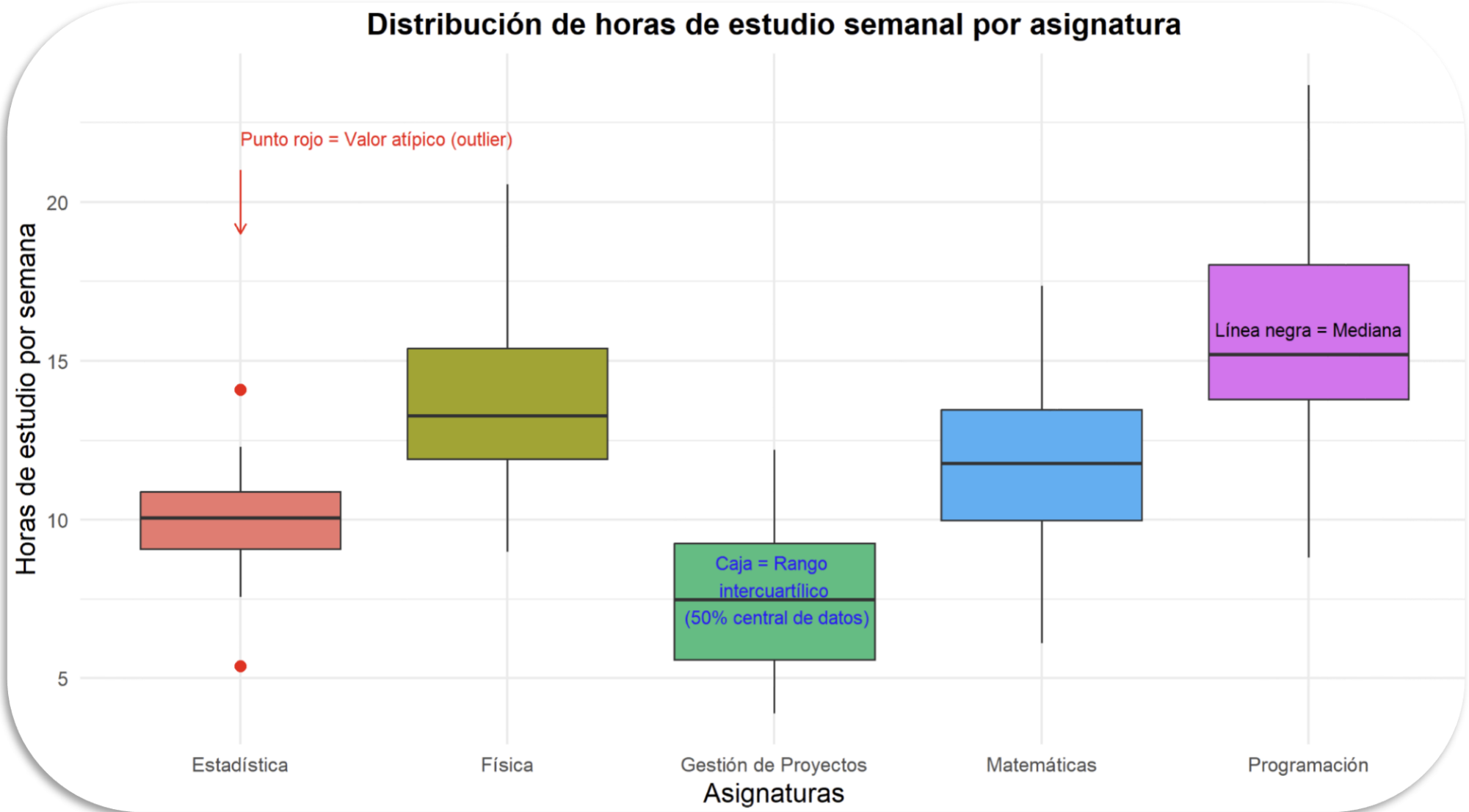
Valores Aberrantes (atípicos, anómalos, outliers)
Valores ubicados 1.5RI del Q1 y del Q3



GRAFICOS QUE CONJUGAN TENDENCIA CENTRAL Y DISPERSIÓN



GRAFICOS QUE CONJUGAN TENDENCIA CENTRAL Y DISPERSIÓN



Otras medidas de dispersión: Coeficiente de variación

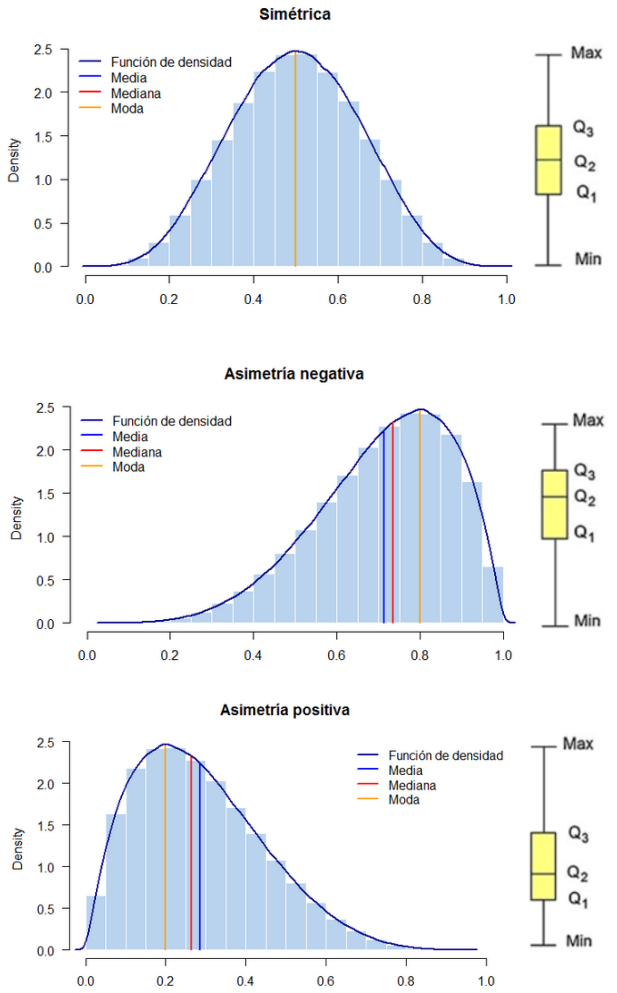
Es un valor que se usa para comparar la variabilidad de los datos de diferentes grupos.

$$V = \frac{S}{\bar{X}}$$

Asimetría o Sesgo

- Una distribución es simétrica si la mitad izquierda de su distribución es la imagen especular de su mitad derecha.
- En las distribuciones simétricas media y mediana coinciden. Si sólo hay una moda también coincide
- La asimetría es positiva o negativa en función de a qué lado se encuentra la cola de la distribución.
- La media tiende a desplazarse hacia las valores extremos (colas).
- Las discrepancias entre las medidas de centralización son indicación de asimetría.

$$A_K = \frac{3(\bar{X} - M_e)}{S}$$



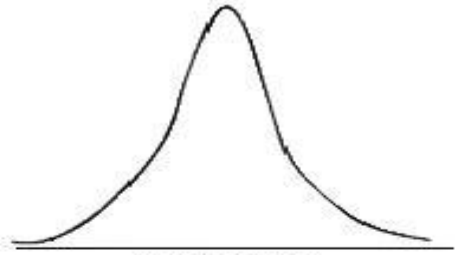
Otras medidas

KURTOSIS

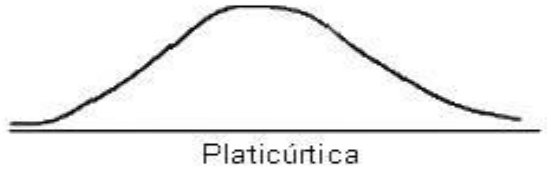
Evalúa el grado de apuntamiento de la distribución, el coeficiente es:

$$K_U = \frac{P_{75} - P_{25}}{2(P_{90} - P_{10})}$$

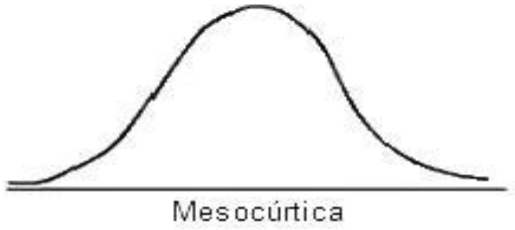
>3



<3



=3



Práctica

Con el archivo *datos vehículos.xlsx* realizar lo siguiente:

1. Hallar medidas de tendencia central, dispersión y posición para las variables cuantitativas.
2. Realizar diagramas de caja para los precios de vehículos automáticos y no automáticos.

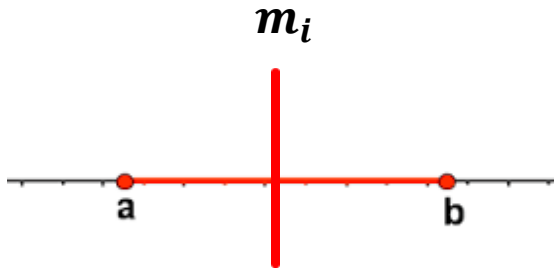
Con el archivo *Adquisiciones laboratorio Química.xlsx* realizar lo siguiente:

1. Hallar medidas de tendencia central, dispersión y posición para las variables cuantitativas.
2. Realizar diagramas de caja.

Media y Varianza muestral para datos agrupados

Calculando estadísticos de datos agrupados

Media $\bar{x} = \sum_{i}^{k} \frac{f_i m_i}{n}$



Varianza $s^2 = \sum_{i=1}^{k} \frac{f_i (m_i - \bar{x})^2}{n-1}$

Media y Varianza muestral para datos agrupados

Estadísticos de datos agrupados

Mediana

$$\widetilde{X} = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} * a_i$$

Li: Límite inferior del intervalo que contiene a la mediana
n: Cantidad de datos
Fi-1: Frecuencia acumulada del intervalo anterior
fi: Frecuencia absoluta del intervalo donde se encuentra la mediana
ai: ancho del intervalo

Moda

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} * a_i$$

Li: Límite inferior del intervalo que contiene a la moda
fi: Frecuencia absoluta del intervalo modal
ai: ancho del intervalo

Cuantiles

$$C_\alpha = L_i + \frac{\alpha \cdot n - F_{i-1}}{f_i}(a_i)$$

L_i: Límite inferior del intervalo que contiene al cuantil
n: Cantidad de datos
F_{i-1}: Frecuencia acumulada del intervalo anterior
f_i: Frecuencia absoluta del intervalo donde se encuentra el cuantil
a_i: ancho del intervalo

Calculando estadísticos de datos agrupados

Peso	M. Clase	Fr.	Fr. ac.
40 – 50	45	5	5
50 – 60	55	10	15
60 – 70	65	21	36
70 - 80	75	11	47
80 - 90	85	5	52
90 - 100	95	3	55
100 – 130	115	3	58
58			

$$\bar{x} = \frac{\sum_i x_i n_i}{n} = \frac{45 \cdot 5 + 55 \cdot 10 + \ldots + 115 \cdot 3}{58} = 69{,}3$$

$$Mediana = C_{0,5} = L_i + \frac{0{,}5 \cdot 58 - F_{i-1}}{n_i}(a_i)$$
$$= 60 + \frac{0{,}5 \cdot 58 - 15}{21}(10) = 66{,}\hat{6}$$

$$P_{75} = C_{0,75} = L_i + \frac{0{,}75 \cdot 58 - F_{i-1}}{f_i}(a_i) = 70 + \frac{43{,}5 - 36}{11}(10) = 76{,}8$$