UNIVERSIDAD NACIONAL DEL CHIMBORAZO

ANÁLISIS EXPLORATORIO DE DATOS

Patricia Hernández, PhD

Asignatura

Modelización Económica

RUTA A SEGUIR EN LA ASIGNATURA

01
ANALISIS EXPLORATORIO / CONFIRMATORIO DE DATOS

02
ANALISIS MULTIVARIANTE
(Modelos sectoriales)

- ANÁLISIS FACTORIAL
- TÉCNICA DE AGRUPACIÓN DE DATOS

02 MODELOS ECONOMÉTRICOS

- MODELOS MICRO
  Datos de panel
  Variable dependiente cualitativa (probit / logit)
  Modelos de frontera (DEA)
  Evaluación impacto
- MODELOS MACRO
  Análisis predictivo
  (Series de tiempo y ARIMA)
  Multiecuacionales (Vectores autorregresivos)


ANÁLISIS
EXPLORATORIO
DE DATOS (AED)

ANÁLISIS DE DATOS

ANÁLISIS DE DATOS

1 ANÁLISIS DESCRIPTIVO

- **¿Qué es?** Síntesis de la información que proporciona el conjunto de datos, extrayendo sus características más representativas.
- **¿Por qué es necesario?** Permite conocer los tipos de datos, descubrir patrones y preparar los datos para futuros análisis.
- **Tratamiento:** Aplicar funciones de estadística descriptiva para explorar la estructura del conjunto de datos, examinar los datos y las variables que presenta.

ANÁLISIS DE DATOS

2 RE-AJUSTE DE LOS TIPOS DE VARIABLES

- **¿Qué es?** Verificar que las variables se han almacenado con el tipo de valor correspondiente.
- **¿Por qué es necesario?** Una mala codificación de las variables puede influir negativamente en la agrupación de los datos o los resultados de los análisis.
- **Tratamiento:** Aplicar la codificación apropiada para cada una de las variables.

ANÁLISIS DE DATOS

3 DETECCIÓN Y TRATAMIENTO DE DATOS AUSENTES

- **¿Qué es?** Identificar la falta de algunos de los datos en la variable.
- **¿Por qué es necesario?** Los datos ausentes pueden generar problemas a la hora de aplicar técnicas de machine learning, elaborar modelos predictivos, realizar análisis estadísticos o generar representaciones gráficas.
- **Tratamiento:** Existen varias maneras de tratar los valores ausentes, como por ejemplo sustituirlos por la media o la mediana, o completar los valores faltantes con el valor anterior o posterior de la columna.

ANÁLISIS DE DATOS

4 DETECCIÓN Y TRATAMIENTO DE DATOS ATÍPICOS

- **¿Qué es?** Identificar datos con valores significativamente distintos a los que presenta la variable.
- **¿Por qué es necesario su tratamiento?** Pueden modificar los resultados y restar potencia a los análisis estadísticos o técnicas de machine learning aplicadas.
- **Tratamiento:** Disminuir su influencia en análisis posteriores o, en casos muy extremos, eliminarlos del conjunto de datos.

ANÁLISIS DE DATOS

5 ANÁLISIS DE CORRELACIÓN DE VARIABLES

- **¿Qué es?** Analizar la relación entre dos o más variables.
- **¿Por qué es necesario?** Entre otras razones, para descartar posibles variables que aporten información redundante en el conjunto de datos, ocasionando ruido en los análisis.
- **Tratamiento:** Calcular los coeficientes de correlación para las variables para detectar coeficientes cercanos a 1 o -1.

TIPOS DE VARIABLES

Variables Cuantitativas

CONTINUA

Peso; Estatura; Temperatura, etc.

DISCRETA

Número de Hijos, Número de empleados de una empresa, Número de asignaturas aprobadas en un semestre, etc.

Variables Cualitativas

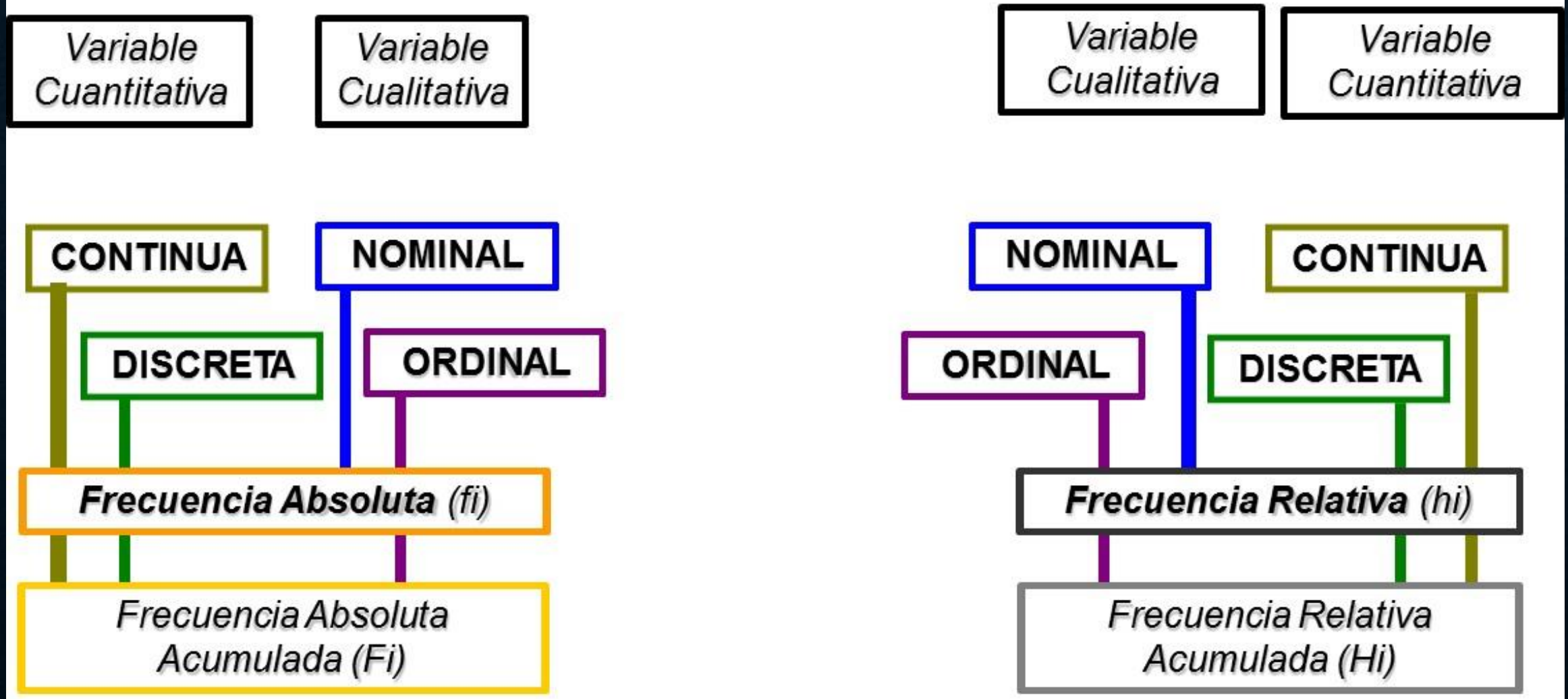
NOMINAL

Sexo, Deporte Favorito, profesiones

ORDINAL

calificaciones, estructura de cargos...

TIPOS DE FRECUENCIAS



ANÁLISIS DE DATOS

Dos o más variables		Componente a desarrollar	Técnicas o herramientas mínimas sugeridas
Tipo de variables	Cuantitativas	Descriptivo (cada variable)	Media, Med, Mo, s, CV, As, K
		Asociativo	Modelos de Regresión Lineal $Y = a + bX$
			Regresión Múltiple $Y = a + bX1 + cX2 + dX3\ldots$
			Correlación de Pearson
			Componentes Principales
			Anàlisis Factorial
		Contraste	Prueba de Hipótesis Media Muestral
			Prueba de Diferencias de Medias
			Pruebas para proporciones
			Estimaciones de series de tiempo
			Chi Cuadrado: Prueba Independencia

Nota: Tomado de Ramírez, Hernández y Morales (2022). La Investigación de Mercados. Orientaciones y aplicaciones en técnicas cuantitativa

ANÁLISIS DE DATOS

Dos o más variables		Componente a desarrollar	Técnicas o herramientas mínimas sugeridas
Tipo de variables	Cualitativas	Descriptivo (para cada una de las variables)	Frecuencias, Mo
		Asociativo	Tablas de Contingencia
			Coeficiente de Correlación Spearman (solo si la variable es ordinal)
			Coeficiente de Correlación de Contingencia
		Contraste	Pruebas para proporciones
			Chi Cuadrado como Prueba de Independencia
			Pruebas de Fisher

Nota: Tomado de Ramírez, Hernández y Morales (2022). La Investigación de Mercados. Orientaciones y aplicaciones en técnicas cuantitativa

ANÁLISIS DE DATOS

Dos o más variables		Componente a desarrollar	Técnicas o herramientas mínimas sugeridas
Tipo de variables	Combinadas	Descriptivo (cada variable)	Caracterizar cada variable de acuerdo a su tipo
		Asociativo	Modelos de Regresión
			Logit / Probit dependiendo de que la variable dependiente sea una variable dicotómica
			Series de tiempo
			Análisis Factorial
			Componentes Principales
			Tablas de Contingencia
			Coeficientes de Correlación de Contingencia / Spearman
		Contraste	Prueba de Fisher
			Chi Cuadrado como Prueba de Independencia
			Pruebas de Whitney, Wilcoxom
			Análisis envolvente de datos (DEA) …

ANÁLISIS UNIDIMENSIONAL VARIABLES CUANTITATIVAS

ORGANIZACIÓN DE LOS DATOS

Frecuencia:

a. Ordinaria:
Absoluta y relativa

b. Acumulativa:
Absoluta y relativa

c. Distribución de Frecuencia: Tabla donde se colocan los datos junto con sus frecuencia de manera organizada

ORGANIZACIÓN DE LOS DATOS

Tabla de Frecuencia

fi / N

Acumulado de hi (debe sumir 1)

Xi	fi	Fi	hi	Hi
N = Σ				

Valor que toma la variable

N° veces que toma ese valor

Acumulado de fi (debe sumar N)

ORGANIZACIÓN DE LOS DATOS

Xi	fi	Fi	hi	Hi
1	6			
2	7			
3	4			
4	2			
5	1			
N = Σ	**20**			

ORGANIZACIÓN DE LOS DATOS

semestre	Freq.	Percent	Cum.
Segundo	31	22,46	22,46
Tercero	57	41,30	63,77
Cuarto	32	23,19	86,96
Quinto	18	13,04	100,00
Total	138	100,00	

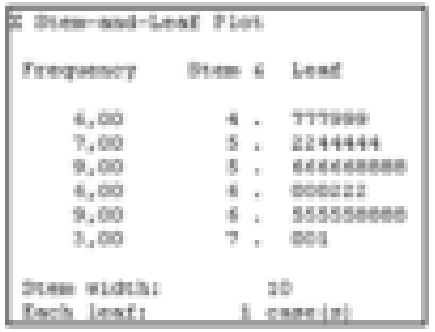
clas_provincia	Freq.	Percent	Cum.
Chimborazo	1	0,72	0,72
Cotopaxi	53	38,41	39,13
Pastaza	1	0,72	39,86
Pichincha	69	50,00	89,86
Santo Domingo	2	1,45	91,30
Tungurahua	12	8,70	100,00
Total	138	100,00	

TABULACIÓN Y REPRESENTACIÓN GRÁFICA

CUANTITATIVOS

EXPLORATORIO

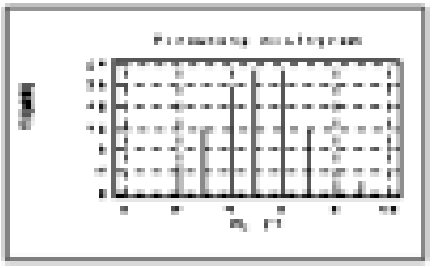
- STEM AND LEAF



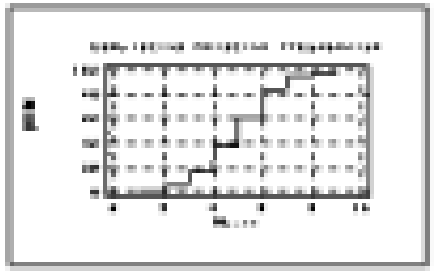
DISTR. SIMPLES

Xi	ni	fi	Ni	Fi
x1	n1	f1	N1	F1
x2	n2	f2	N2	F2
...	...	...	...	...
xi	ni	fi	Ni	Fi
...	...	...	...	...
xk	nk	fk	N	1
	N	1		

- BARRAS



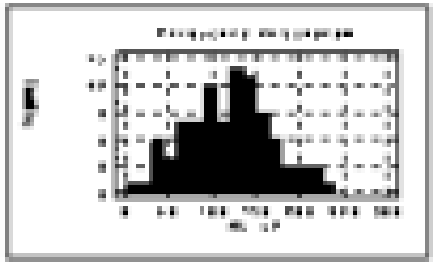
- DIAGRAMA EN ESCALERA



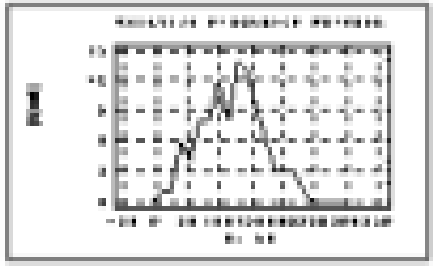
DISTR. AGRUPADAS

Li-1 Li	ci	ni	fi	Ni	Fi
L0 L1	c1	n1	f1	N1	F1
L1 L2	c2	n2	f2	N2	F2
...	...	...	...	...	...
Li-1 Li	ci	ni	fi	Ni	Fi
...	...	...	...	...	...
Lk-1 Lk	ck	nk	fk	N	1
		N	1		

- HISTOGRAMA



- POLÍGONO DE FREC SIMPLE

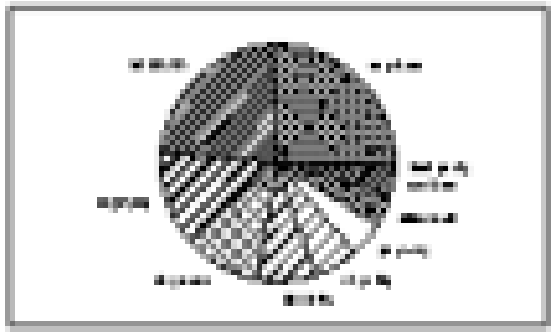


- POLÍGONO DE FREC ACUMULADAS

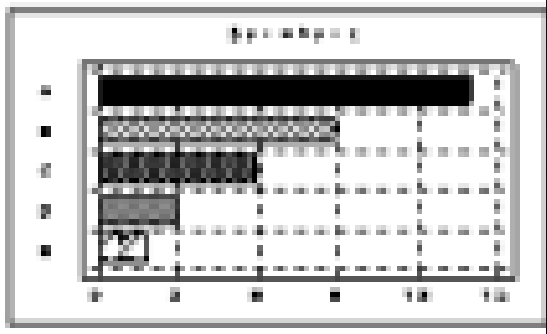
REPRESENTACIÓN GRÁFICA

CUALITATIVOS

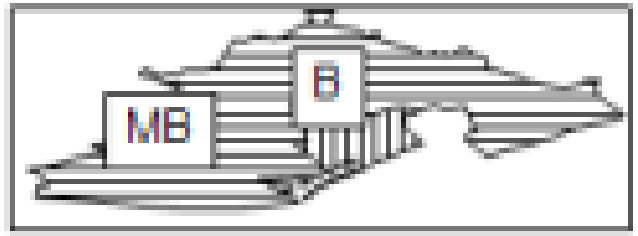
- DIAG. DE SECTORES



- DIAG. COLUMNAS O BARRAS



- CARTOGRAMAS
- PICTOGRAMAS



REPRESENTACIONES GRÁFICAS

1. **Diagrama de Barras**. (Variables discretas)
Se emplea la frecuencia absoluta o relativa

2. **Polígono de frecuencia**. (Variables discretas)
Se emplea la frecuencia absoluta o relativa. Se unen los extremos superiores de las barras

REPRESENTACIONES GRÁFICAS

Xi	fi	Fi	hi	Hi
68	10			
69	7			
70	9			
71	7			
72	8			
73	9			
N = Σ	**50**			

REPRESENTACIONES GRÁFICAS VARIABLES CUANTITAIVAS

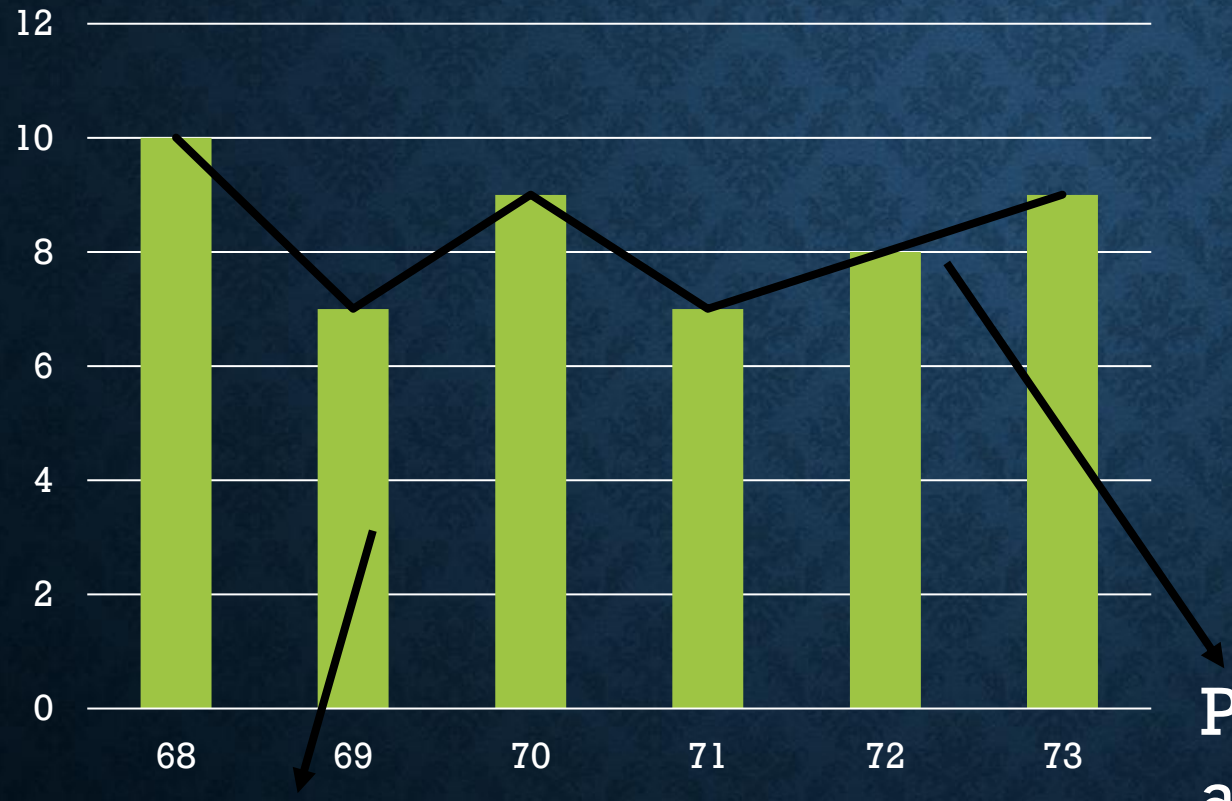


Diagrama barras frecuencia absoluta

Xi	fi	Fi	hi	Hi
68	10	10	0,20	0,20
69	7	17	0,14	0,34
70	9	26	0,18	0,52
71	7	33	0,14	0,66
72	8	41	0,16	0,82
73	9	**50**	0,18	**1**
N = Σ	**50**		**1**	

Polígono de frecuencia absoluta

REPRESENTACIONES GRÁFICAS VARIABLES CUANTITATIVAS

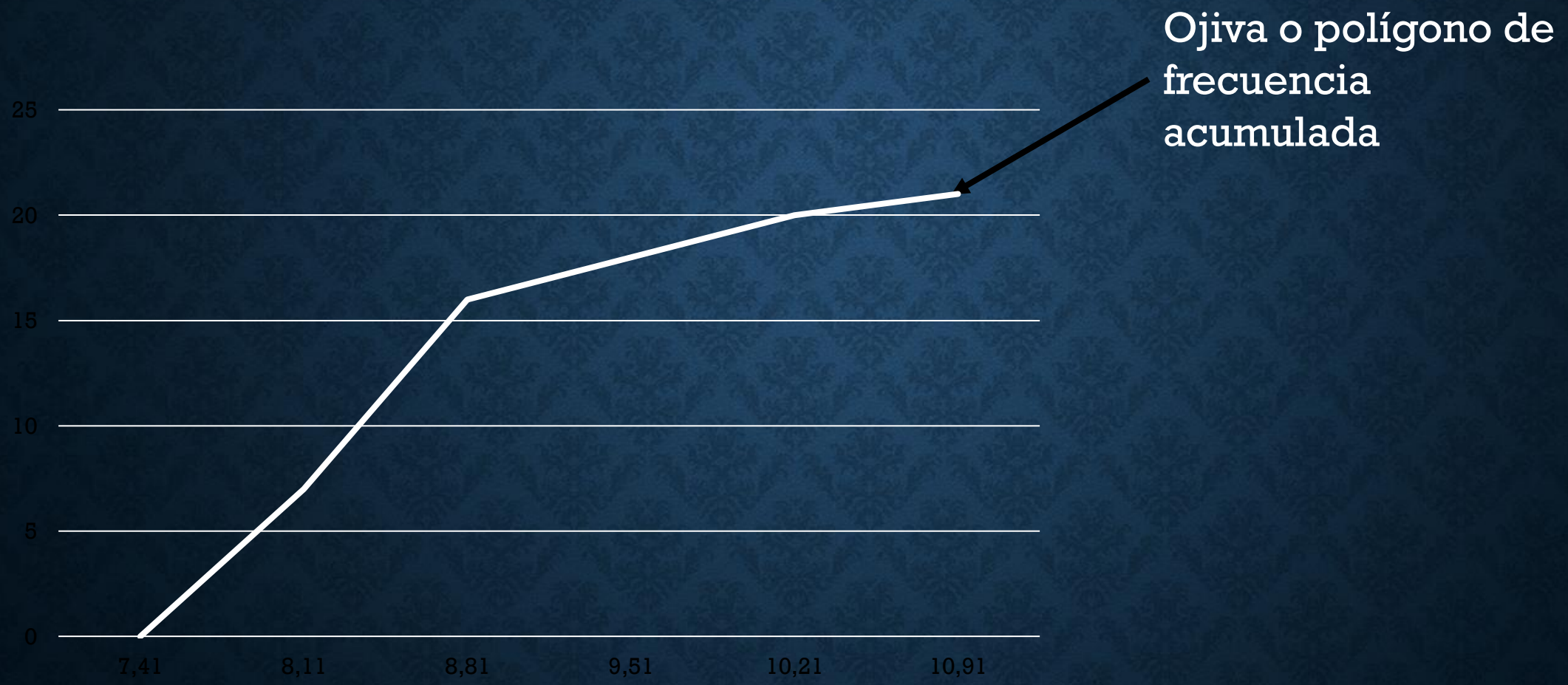


DIAGRAMA DE TALLO Y HOJAS VARIABLES CUANTITATIVAS

Una técnica útil para la observación inicial de los datos y la comprensión de lo que representan y cómo se distribuyen

Reproduce fielmente y consta de varias columnas, la primera, o tallo, representa en orden la mayor porción de los datos y las siguientes, separadas por una línea vertical, representan la menor porción de cada dato en particular.

DIAGRAMA DE TALLO Y HOJAS

Para construir este diagrama:

- **Se divide cada valor observado en dos partes:** hoja y tallo. Para ello, se fija la posición del dígito que se tomará como hoja y los tallos quedan determinados por los dígitos que quedan a la izquierda de dicha posición.
- **Se anotan en columna los tallos desde el menor hasta el mayor de forma sucesiva sin omisiones**. Los tallos deben ser consecutivos y abarcar todo el recorrido de la variable.

DIAGRAMA DE TALLO Y HOJAS

Para construir este diagrama:

- A la derecha de cada tallo se anotan de forma ordenada (de menor a mayor) sus hojas.
- Si el número de observaciones es excesivamente grande, es conveniente que cada hoja represente a un número determinado de elementos con el mismo tallo y hoja, debiéndose indicar en el diagrama.

DIAGRAMA DE TALLO Y HOJAS

Para construir este diagrama:

- Se puede completar el diagrama con las frecuencias simples o acumuladas anotándolas a la izquierda de los tallos que se obtiene sumando las hojas correspondientes a cada tallo
- En el encabezado se acostumbra a indicar el tamaño de la muestra, n, que es el número total de hojas.

DIAGRAMA DE TALLO Y HOJAS

Tabla 1.10. Fertilidad mundial en el 2010. Hijos promedio por mujer

Ubicación	Fer.	Ubicación	Fer.	Ubicación	Fer.	Ubicación	Fer.	Ubicación	Fer.
Bosnia	1,2	Russia	1,4	Korea N	1,9	S. Africa	2,5	Samoa	4,0
Korea S	1,2	Spain	1,4	Lebanon	1,9	Venezuela	2,5	Guatemala	4,1
Andorra	1,3	Barbados	1,5	Norway	1,9	Ecuador	2,6	Iraq	4,1
Germany	1,3	Cuba	1,5	Sweden	1,9	Panama	2,6	Sudan	4,2
Japan	1,3	Switzerland	1,5	Costa Rica	2,0	Peru	2,6	Congo	4,4
Malta	1,3	Canada	1,6	Ireland	2,0	Dominican	2,7	Cameroon	4,6
Poland	1,3	Serbia	1,6	Maldives	2,0	India	2,7	Kenya	4,9
Romania	1,3	Trinidad	1,6	Mongolia	2,0	Nicaragua	2,7	Senegal	5,0
Singapore	1,3	Armenia	1,7	N. Zealand	2,0	Israel	2,8	Liberia	5,1
Slovakia	1,3	Estonia	1,7	Dominica	2,1	Cambodia	2,9	S. Leone	5,2
Ukraine	1,3	Netherlands	1,7	Turkey	2,1	Egypt	2,9	Yemen	5,2
Austria	1,4	Australia	1,8	Uruguay	2,1	Nepal	2,9	Ethiopia	5,3
Bulgaria	1,4	Belgium	1,8	USA	2,1	Paraguay	3,1	Nigeria	5,3
Croatia	1,4	China	1,8	Argentina	2,2	Philippines	3,1	Angola	5,8
Czech	1,4	Denmark	1,8	Indonesia	2,2	S. Arabia	3,1	Zambia	5,8
Greece	1,4	Finland	1,8	Kuwait	2,2	Honduras	3,3	Congo	6,0
Hungary	1,4	Iran	1,8	Mexico	2,2	Syrian	3,3	Chad	6,2
Italy	1,4	Thailand	1,8	El Salvador	2,3	Namibia	3,4	Uganda	6,3
Latvia	1,4	UK	1,8	Sri Lanka	2,3	Zimbabwe	3,4	Somalia	6,4
Lithuania	1,4	Brazil	1,9	Colombia	2,4	Bolivia	3,5	Afghanistan	6,6
Macedonia	1,4	Chile	1,9	Jamaica	2,4	Haiti	3,5	Niger	7,1
Portugal	1,4	France	1,9	Qatar	2,4	Pakistan	4,0		

DIAGRAMA DE TALLO Y HOJAS

1,	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4
1,	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	8	8	9	9	9	9	9	9	9
2,	0	0	0	0	0	1	1	1	1	2	2	2	2	3	3	4	4	4						
2,	5	5	6	6	6	7	7	7	8	9	9	9												
3,	1	1	1	3	3	4	4																	
3,	5	5																						
4,	0	0	1	1	2	4																		
4,	6	9																						
5,	0	1	2	2	3	3																		
5,	8	8																						
6,	0	2	3	4																				
6,	6																							
7,	1																							

DIAGRAMA DE TALLO Y HOJAS

Supongamos que las edades de un colectivo formado por 45 trabajadores son los siguientes: 32, 32, 32, 34, 34, 35, 35, 35 ,35, 37 ,37, 37 ,37, 38 ,39, 40, 40, 41, 42, 42, 42, 42, 42, 42, 42, 43, 43, 43, 43, 43, 45, 45, 45, 45, 45, 45, 47, 47, 48, 49, 49, 50, 50, 51, 51, 51.

DIAGRAMA DE TALLO Y HOJAS

n_i	Tallo	Hojas
15	3	222445555777789
25	4	0012222222333335555577899
5	5	00111

Unidades de las hojas: 1 3|2 representa 32

n_i	Tallo	Hojas
5	3	22244
10	3	5555777789
15	4	001222222233333
10	4	5555577899
5	5	00111

REPRESENTACIONES GRÁFICAS VARIABLES CUALITATIVAS

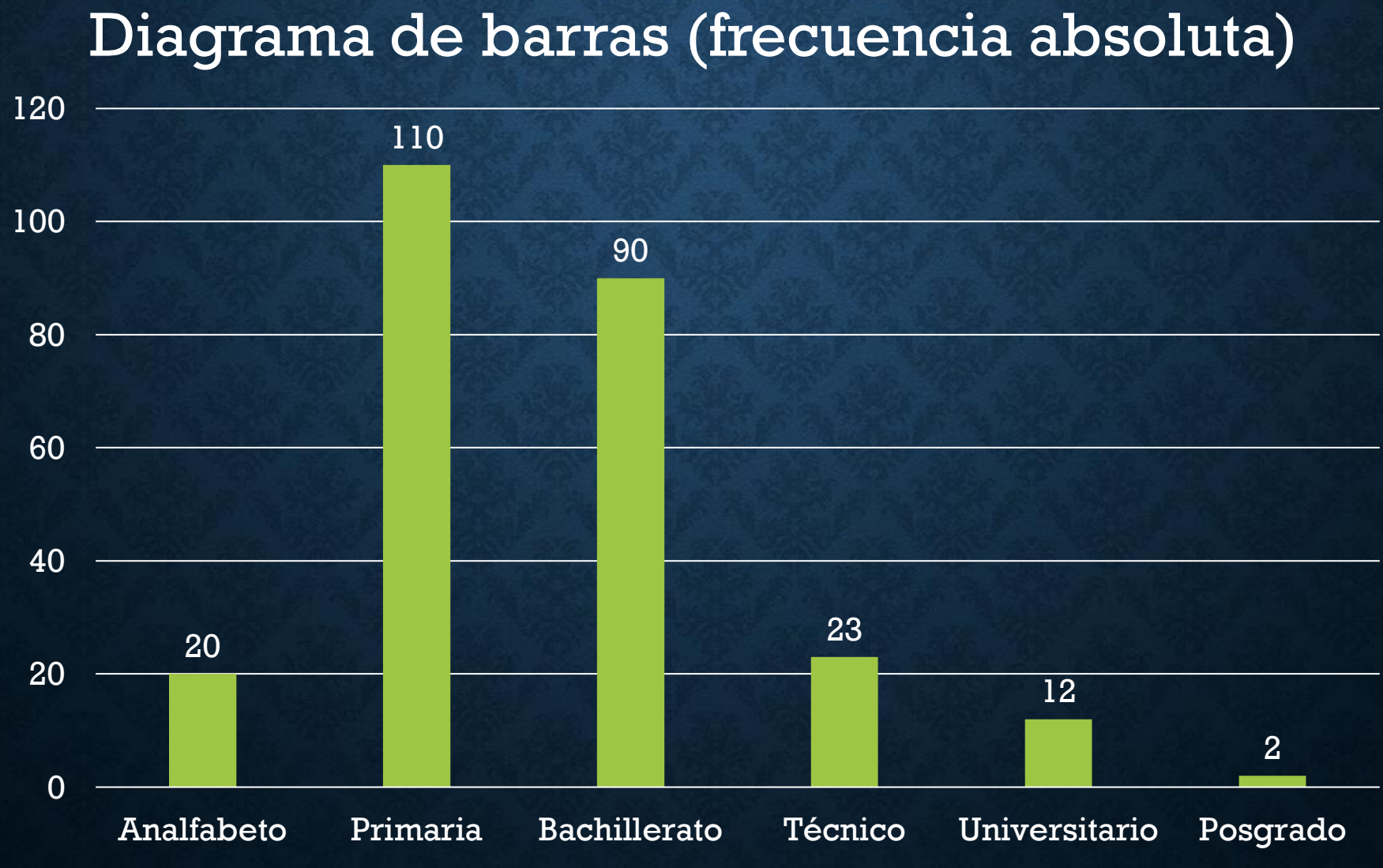
1.Histograma de barras
Eje “x” son las categorías y eje “y” la frecuencia

2. Diagrama sectores (torta – pastel).
Se unen las marcas de clase de cada intervalo

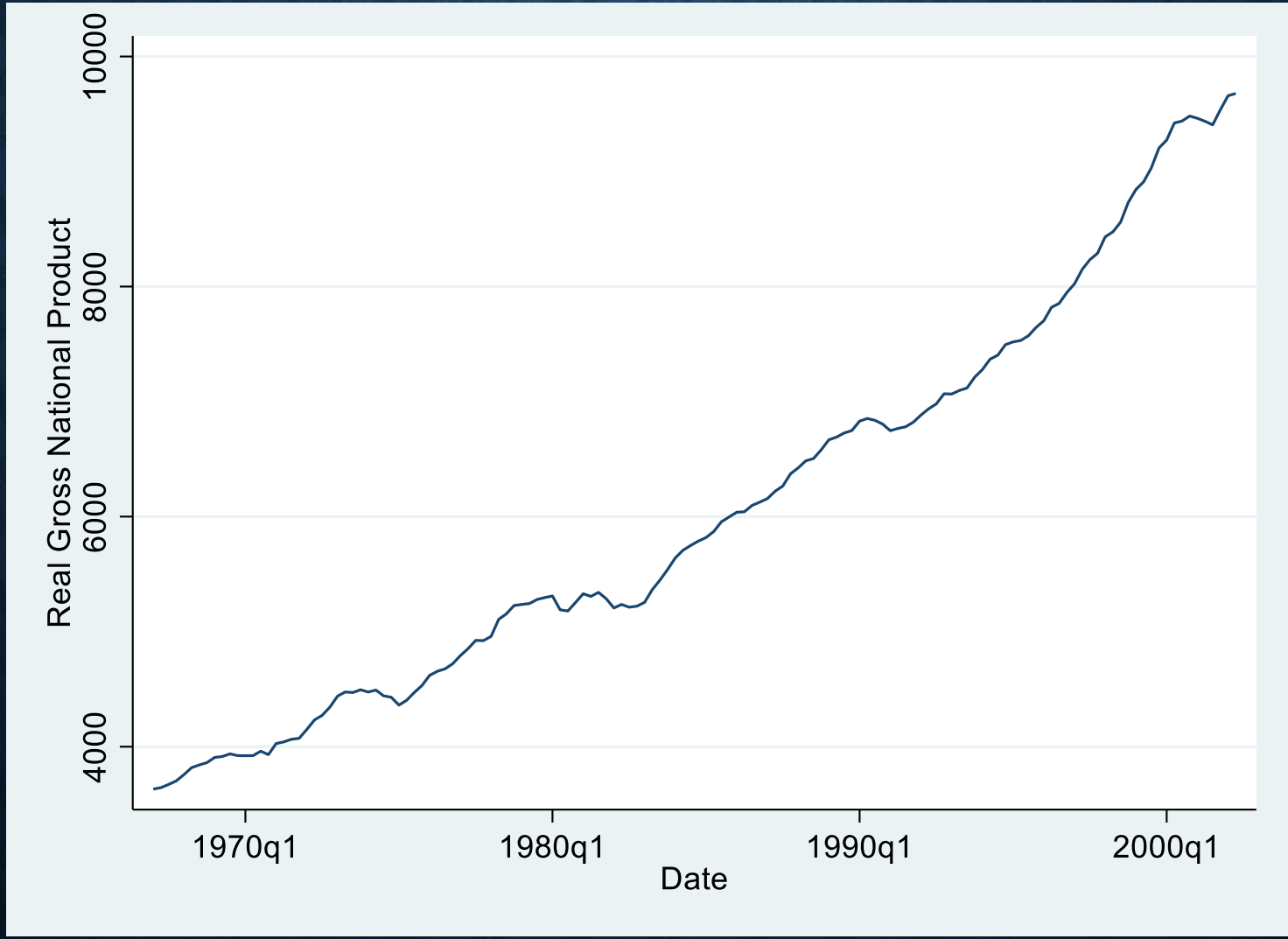
REPRESENTACIONES GRÁFICAS VARIABLES CUALITATIVAS

Nivel educativo	fi	Fi	hi	Hi
Analfabeto	20	20	0,08	0,08
Primaria	110	130	0,43	0,51
Bachillerato	90	220	0,35	0,86
Técnico	23	243	0,09	0,95
Universitario	12	255	0,05	0,99
Posgrado	2	257	0,01	1,00

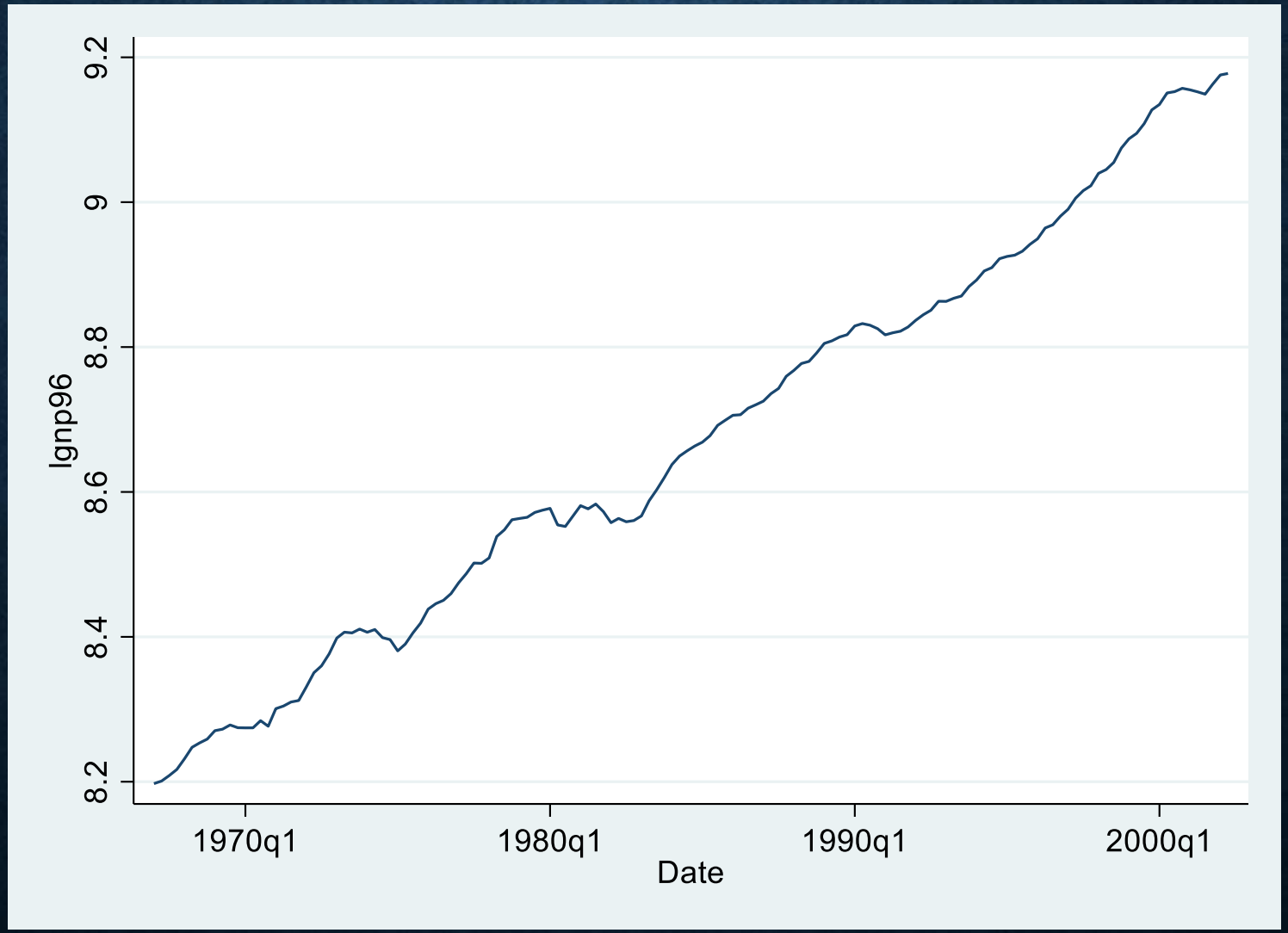
REPRESENTACIONES GRÁFICAS VARIABLES CUALITATIVAS



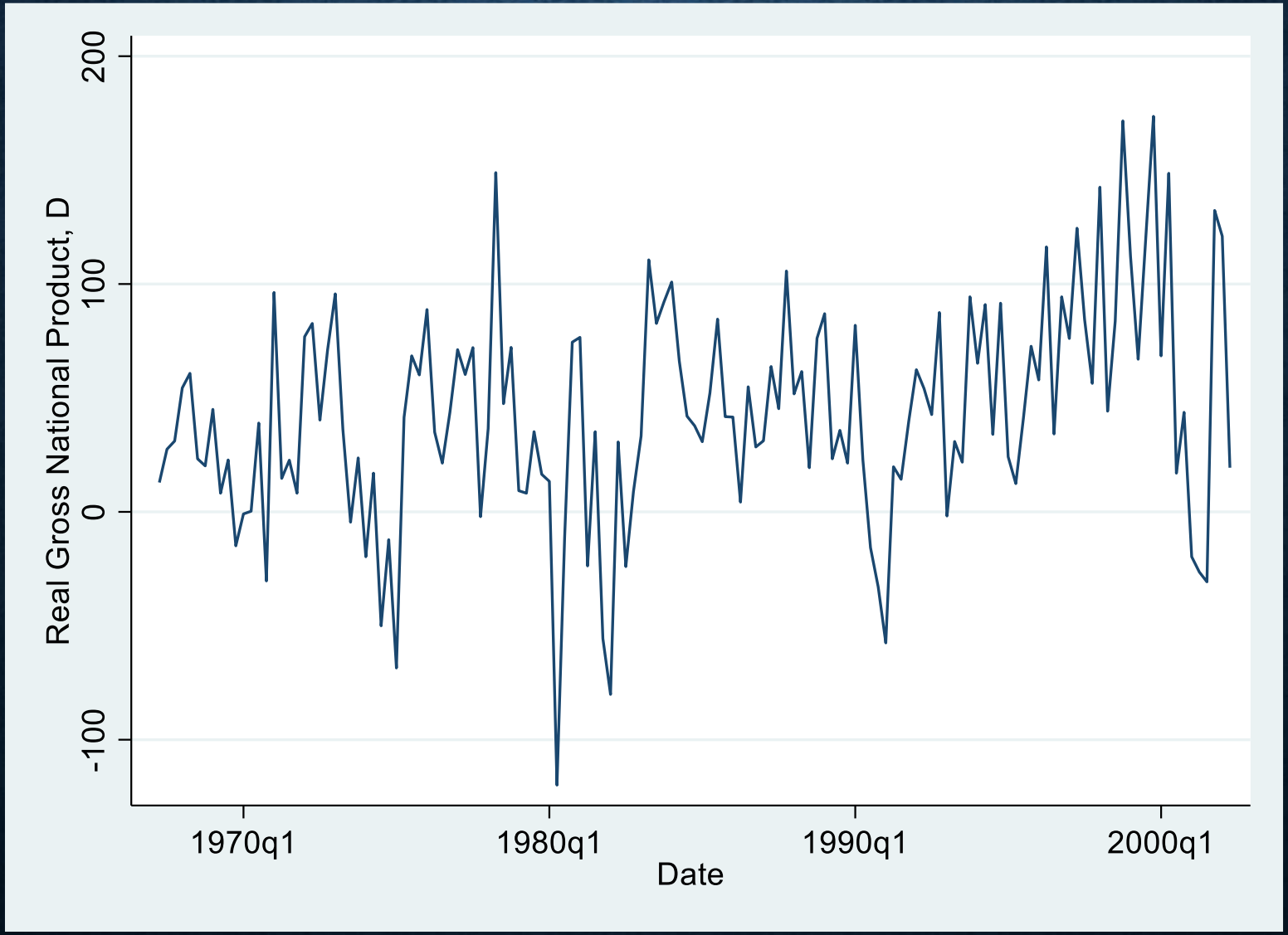
REPRESENTACIONES GRÁFICAS SERIES DE TIEMPO - NIVELES



REPRESENTACIONES GRÁFICAS SERIES DE TIEMPO - LOGARITMO



REPRESENTACIONES GRÁFICAS SERIES DE TIEMPO - DIFERENCIAS



ANÁLISIS UNIDIMENSIONAL
MEDIDAS DE TENDENCIA CENTRAL

Datos Cuantitativos

x

x_1

x_2

$\vdots$

x_n

Media Aritmética o Promedio

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$X = \frac{\sum_{i=1}^{n} (x_i . fi)}{n}$$

Datos Cuantitativos ordenados de menor a mayor

x

$x_{(1)}$

$x_{(2)}$

$\vdots$

$x_{(n)}$

Mediana

$M_E = x_{(k)}$ Si ***n*** es impar

$M_E = \frac{x_{(k)} + x_{(k+1)}}{2}$ Si ***n*** es par

$x_{(k)}$ = dato del centro

Datos

Cualitativos y Cuantitativos

Moda

M_o ="el dato que más se repite"

MEDIDAS TENDENCIA CENTRAL
MEDIA ARITMÉTICA (PROMEDIO)

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i f_i}{N} \qquad \bar{X} = \sum_{i=1}^{n} x_i h_i$$

MEDIDAS TENDENCIA CENTRAL
MEDIA ARITMÉTICA (PROMEDIO)

Xi	fi	Fi	hi	Hi	Xi * fi
1	6	6	0,30	0,30	
2	7	13	0,35	0,65	
3	4	17	0,20	0,85	
4	2	19	0,10	0,95	
5	1	20	0,05	1	
N = Σ	**20**		**1**		

MEDIA - PROPIEDADES

1. La media al ser un promedio de tendencia central, la suma de las desviaciones es CERO

$$\sum_{k=0}^{n}(\bar{X}-x_i)=\sum\bar{X}-\sum x_i=N\bar{X}-\sum x_i=0$$

Dado que: $\bar{X}=\sum\frac{x_i}{N}\Rightarrow N\bar{X}=\sum x_i$

MEDIA - PROPIEDADES

2. Transformación lineal de la variable

$$\bar{Y} = a + b\bar{X} \quad \text{si} \quad Y = a + bx$$

$$\bar{Y} = \sum \frac{y_i}{N} = \sum \frac{(a+bx_i)}{N} = \frac{aN+b\sum x_i}{N} = a + b\bar{X}$$

3. La media depende de los valores extremos

ANÁLISIS UNIDIMENSIONAL
MEDIDAS DE TENDENCIA CENTRAL
MEDIANA Y MODA

MEDIANA

La Mediana, a veces llamado media posicional:

- Es aquel valor de la variable que supera la mitad de las observaciones y a su vez es superado por la otra mitad de las observaciones.
- se le considera **como valor central**, ya que el promedio estará situado en el centro de la distribución.

MEDIANA - CARACTERÍSTICAS

a. Su aplicación es menos frecuente que la media aritmética
b. Presenta gran inestabilidad en el muestreo
c. Sus fórmulas son rígidas y no admiten tratamiento algebraico
d. Su medida no está afectada por los extremos
e. Para calcular la mediana se requiere un ordenamiento de los datos, de menor a mayor o viceversa
f. La mediana es utilizada con mayor frecuencia, cuando la distribución presenta el primero y el último intervalo abierto o no definido
g. El valor de este promedio depende del número de observaciones y no del valor de las mismas.

MEDIANA - CÁLCULO

Xi	fi	Fi	hi	Hi
1	6	6	0,30	0,30
2	7	13	0,35	0,65
3	4	17	0,20	0,85
4	2	19	0,10	0,95
5	1	20	0,05	1
N = Σ	20		1	

$$Posición = \frac{N}{2}$$

Clase medianal

$$Posición = \frac{20}{2} = 10$$

La mediana será el primer valor de xi con frecuencia absoluta acumulada > 10

Me= 2

MODA - CARACTERÍSTICAS

Es el valor de la variable o del atributo que presenta la mayor densidad, es decir, la mayor frecuencia. En otras palabras, la moda es el valor que más se repite en una serie de datos.

MODA - CARACTERÍSTICAS

1. Puede ser usada para datos cualitativos
2. No se ve afectada por los extremos y puede determinarse con extremos abiertos
3. Cuando las frecuencias son iguales no es útil este concepto, debido a que todos serían modas
4. Puede darse el caso que haya 2 modas (bimodal), 3 o más modas (multimodal), lo que hace difícil su interpretación
5. Su uso es bastante limitado
6. Al igual que la mediana, sus fórmulas no admiten tratamientos algebraicos.

MODA - CÁLCULO

Xi	fi	Fi	hi	Hi
1	6	6	0,30	0,30
2	7	13	0,35	0,65
3	4	17	0,20	0,85
4	2	19	0,10	0,95
5	1	20	0,05	1
N = Σ	20		1	

Clase modal Mo= 2

$$\bar{X} > Me = Mo$$

En este caso:
Media 2,25
Moda 2
Medina 2

ANÁLISIS UNIDIMENSIONAL
MEDIDAS DE DISPERSIÓN

MEDIDAS DE DISPERSIÓN

Son medidas que se emplean para determinar el grado de variabilidad o de dispersión de los datos con respecto a un promedio. Por lo general se les considera como promedio de las desviaciones respecto a algún valor central o medidas de posición.

En otras palabras, es el grado en que los datos numéricos tienden a esparcirse alrededor de un valor promedio.

MEDIDAS DE DISPERSIÓN

- **ABSOLUTAS:** Es imprescindible utilizarlos con un promedio. Tienen el inconveniente que no permiten comparaciones entre distribuciones de diferentes promedios
  Rango, varianza y desviación estándar

- **RELATIVAS:** Se obtiene por cocientes entre magnitudes de la misma dimensión, lo que permite comparaciones entre distribuciones heterogéneas.
  Coeficiente de variación (Pearson)

MEDIDAS DE DISPERSIÓN

MEDIDAS DE DISPERSIÓN

(ABSOLUTAS)

- -Rango
- -Varianza
- -Desviación Estándar

Datos Cuantitativos

x
x_1
x_2
$\vdots$
x_n

Rango

$$R = \max(x_i) - \min(x_i)$$

Desviación Típica o Estándar

$$s = \sqrt{s^2}$$

Varianza

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

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

MEDIDAS DE DISPERSIÓN

$$Recorrido\ o\ Rango = X_{máx} - X_{min}$$

-Cuanto mayor sea el recorrido, mayor será el campo de variación de la variable, por lo que no toma en cuenta las frecuencias.

-Su uso es bastante limitado y sólo se utiliza en aquellas ocasiones, en donde nos interesa tener una idea rápida de la variación en un grupo de datos.

- Es la más sencilla y proporciona menos información

-Sólo se toma en cuenta dos valores de la variable de toda la distribución

MEDIDAS DE DISPERSIÓN

La S^2 es la media aritmética de los cuadrados de las desviaciones respecto a la media aritmética, es decir, te da un aproximado sobre la cuantificación del grado de variabilidad en una distribución cualquiera.

$$S^2 = \frac{\sum(X_i - \overline{X})^2}{n}$$

La Desviación típica (S) es la raíz cuadrada de la varianza, o dicho de otro modo, es la raíz cuadrada de las desviaciones respecto a la media. Es la medida de dispersión más conocida y la más utilizada.

$$S = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n}}$$

MEDIDAS DE DISPERSIÓN

Propiedades varianza

1. La varianza de la suma (o resta) de una variable más (o menos) una constante, es igual a la varianza de la variable.

$$U_i = X_i \pm K \quad K: \text{Constante}$$

$$S_{U_i} = S_{X_i}$$

2. El valor de la varianza debe ser siempre +

$$S^2 \geq 0$$

MEDIDAS DE DISPERSIÓN

Propiedades varianza

3. La varianza de una constante por una variable, es igual al producto de la constante al cuadrado por la varianza de la variable.

$$d_i = K.X_i \quad K: Constante$$

$$S_{di}^2 = K^2 S_{xi}^2$$

MEDIDAS DE DISPERSIÓN

MEDIDAS DE DISPERSIÓN *(RELATIVAS)*

- Coeficiente de Variación
- Pearson

Coeficiente de Variación

$$CV = \frac{S}{\overline{X}} * 100$$

MEDIDAS DE DISPERSIÓN

Se obtiene dividiendo la desviación típica por su media aritmética, expresándose el resultado en términos porcentuales.

Este coeficiente de variación se emplea:

1. Cuando se desea comparar dos o más distribuciones, con el fin de determinar cuál tiene mayor o menor variabilidad relativa.
2. Cuando las distribuciones están dadas en unidades de medidas diferentes.
3. Cuando las distribuciones estén expresadas en la misma unidad, pero lo que importa es determinar la variación respecto a una base, por lo que debemos usar el CV.

Por lo tanto, cuanto menor sea el CV, menor será la dispersión relativa y, por tanto, mayor será la representatividad de la media aritmética.

MEDIDAS DE DISPERSIÓN

Xi	fi	Fi	hi	Hi		
1	6	6	0,30	0,30		
2	7	13	0,35	0,65		
3	4	17	0,20	0,85		
4	2	19	0,10	0,95		
5	1	**20**	0,05	**1**		
N = Σ	**20**		**1**			

ANÁLISIS UNIDIMENSIONAL
MEDIDAS DE POSICIÓN

MEDIDAS DE POSICIÓN

- **PERCENTILES**
- **CUARTILES**
- **DECILES**

MEDIDAS DE POSICIÓN

PERCENTIL

Si deseamos **dividir la distribución en cien partes con** igual número de observaciones, se tendrá 99 valores de la variable que separan a la frecuencia total de la distribución divididas en 100 partes iguales. Por lo tanto, existen 100 percentiles.

MEDIDAS DE POSICIÓN

Xi	fi	Fi	hi	Hi
1	6	6	0,30	0,30
2	7	13	0,35	0,65
3	4	17	0,20	0,85
4	2	19	0,10	0,95
5	1	20	0,05	1
N = Σ	20		1	

Clase percentil 20

Paso 1: Calcular N/100 $Posición = \frac{Nq}{100} = \frac{20 * 20}{100} = 4$

Paso 2: $P_{20} = 1$

MEDIDAS DE POSICIÓN

CUARTIL

Se divide la distribución en cuatro partes, de tal manera que cada una contenga igual número de observaciones, es decir, el 25% de las observaciones.

Se denomina cuartiles a los tres valores que separan a la frecuencia total de la distribución, dividida en cuatro partes iguales. El valor central es igual a la mediana y corresponde al segundo cuartil.

MEDIDAS DE POSICIÓN

Características:

1. Q_1 (Cuartil inferior): Es aquel valor de la variable que supera al 25% de las observaciones y a la vez es superado por el 75% restante.

2. Q_2 (2do cuartil): Es aquel valor de la variable que supera al 50% y a la vez es superado por el otro 50% de las observaciones.

3. Q_3 (3er cuartil): Es aquel valor de la variable que supera el 75% y es superado por el restante 25% de las observaciones.

MEDIDAS DE POSICIÓN

DATOS NO AGRUPADOS EN INTERVALOS (Q1= 25%)

Xi	fi	Fi	hi	Hi
1	6	6	0,30	0,30
2	7	13	0,35	0,65
3	4	17	0,20	0,85
4	2	19	0,10	0,95
5	1	20	0,05	1
N = Σ	20		1	

Clase cuartil 1 (25%)

Paso 1: Calcular N/4

$$Posición = \frac{N}{4} = \frac{20}{4} = 5$$

Paso 2: $Q_1 = 1$

MEDIDAS DE POSICIÓN

DECIL

Se divide la distribución en 10 partes. Se tendrá uno de los 9 valores que dividen la frecuencia total en diez partes iguales. El 1er decil es igual al valor que supera al 10% de las observaciones y a la vez es superado por el restante 90%.

$$Posición = \frac{N * d}{10}$$

MEDIDAS DE POSICIÓN

DATOS NO AGRUPADOS EN INTERVALOS (D1= 10%)

Xi	fi	Fi	hi	Hi
1	6	6	0,30	0,30
2	7	13	0,35	0,65
3	4	17	0,20	0,85
4	2	19	0,10	0,95
5	1	20	0,05	1
N = Σ	20		1	

Paso 1: Calcular N*1/10 $Posición = \frac{N}{10} = \frac{20}{10} = 2$

Paso 2: $D_1 = 1$

MEDIDAS DE POSICIÓN

$$P50 = D5 = Q2 = Me$$

DIAGRAMA DE CAJA

Permiten visualizar y comparar la distribución y la tendencia central de valores numéricos mediante sus cuartiles.

Los cuartiles son una forma de dividir valores numéricos en cuatro grupos iguales basados en cinco valores clave: mínimo, primer cuartil, mediana, tercer cuartil y máximo

La parte de la caja del gráfico ilustra el 50 por ciento medio de los valores de los datos, también conocido como rango intercuartílico o IQR.

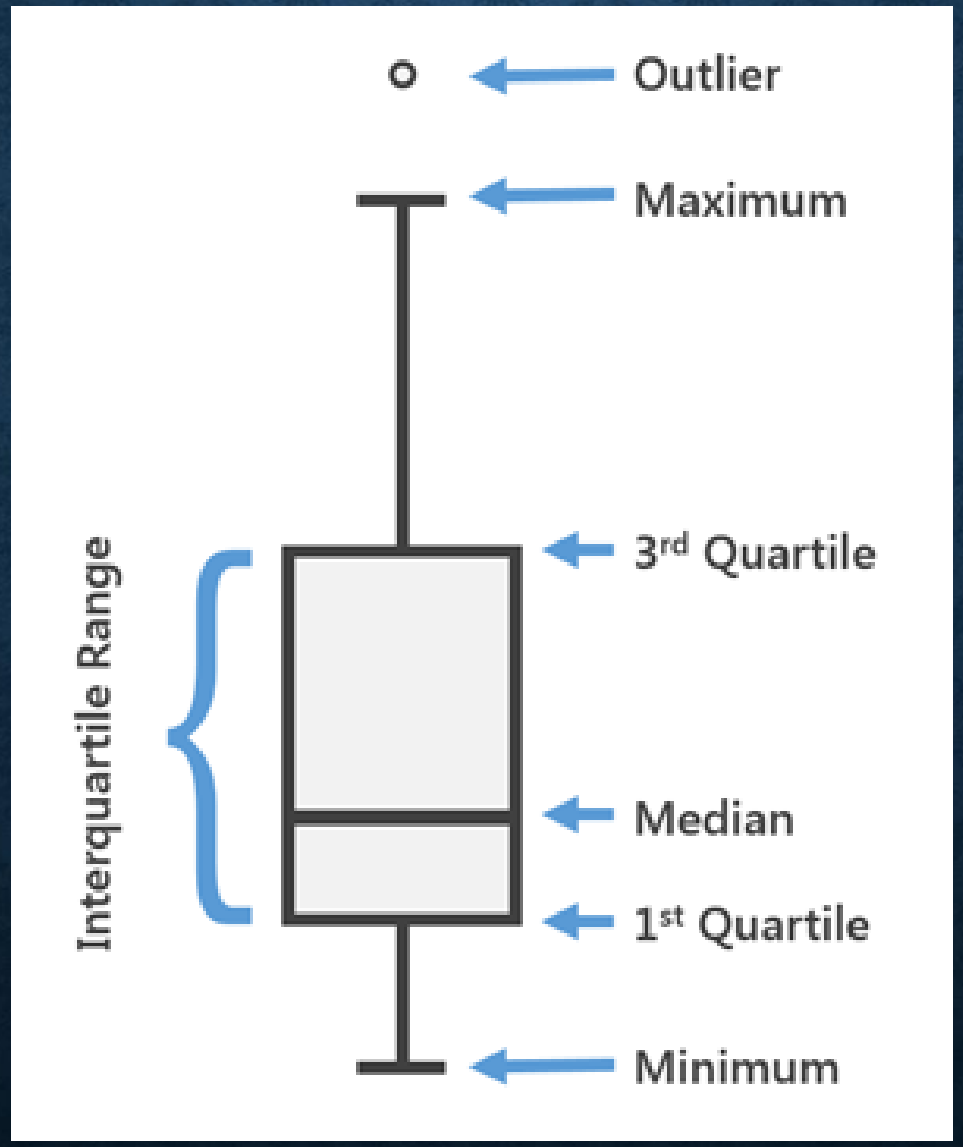
DIAGRAMA DE CAJA

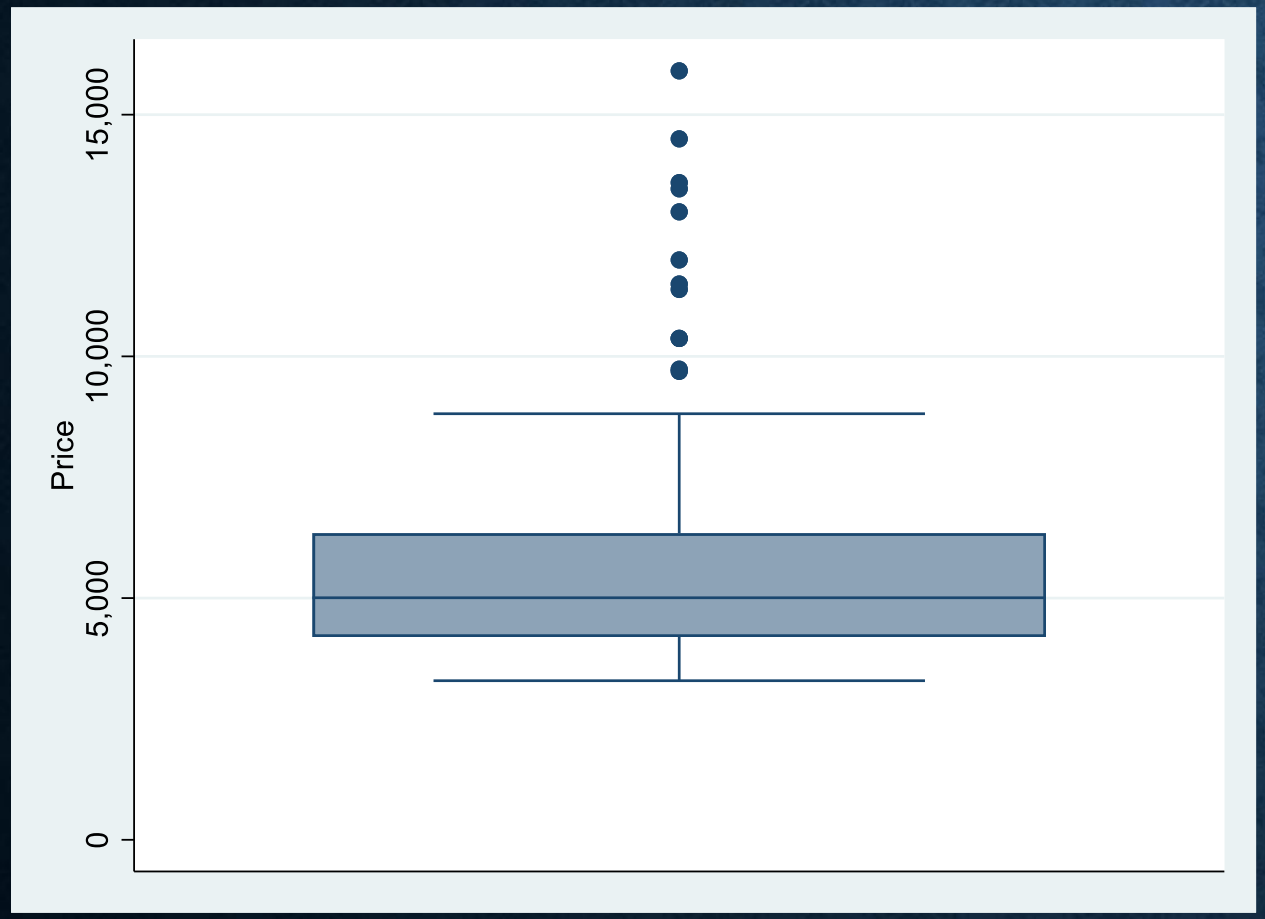
La media de los valores se representa como la línea que divide la caja por la mitad.

El IQR ilustra la variabilidad en un conjunto de valores. Un IQR grande indica una amplia dispersión de los valores, mientras que un IQR más pequeño indica que la mayoría de los valores quedan hacia el centro.

Los diagramas de caja también ilustran los valores mínimos y máximos de los datos mediante bigotes que se extienden desde la caja y, opcionalmente, valores atípicos como puntos que se extienden más allá de los bigotes.

DIAGRAMA DE CAJA
BOX PLOT






```
. summarize price, detail

                            Price
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3291            3291
 5%         3748            3299
10%         3895            3667
25%         4195            3748

50%       5006.5

                        Largest
75%         6342           13466
90%        11385           13594
95%        13466           14500
99%        15906           15906
```

ANÁLISIS UNIDIMENSIONAL
MEDIDAS DE FORMA

MEDIDAS DE FORMA

- La estructura de una distribución está determinada por la forma de su diagrama de barras o su histograma, ya que se puede ver, a través de estos gráficos, si las observaciones están o no muy concentradas en pocos valores de la variable o si la concentración se presenta en el centro o en uno de sus extremos.
- Las distribuciones en forma de campana, campaniformes, son las más habituales en la estadística práctica.

MEDIDAS DE FORMA

Estas distribuciones se caracterizan porque el mayor número de observaciones se agrupan en valores de la variable más o menos centrales, siendo poco común los valores extremos.

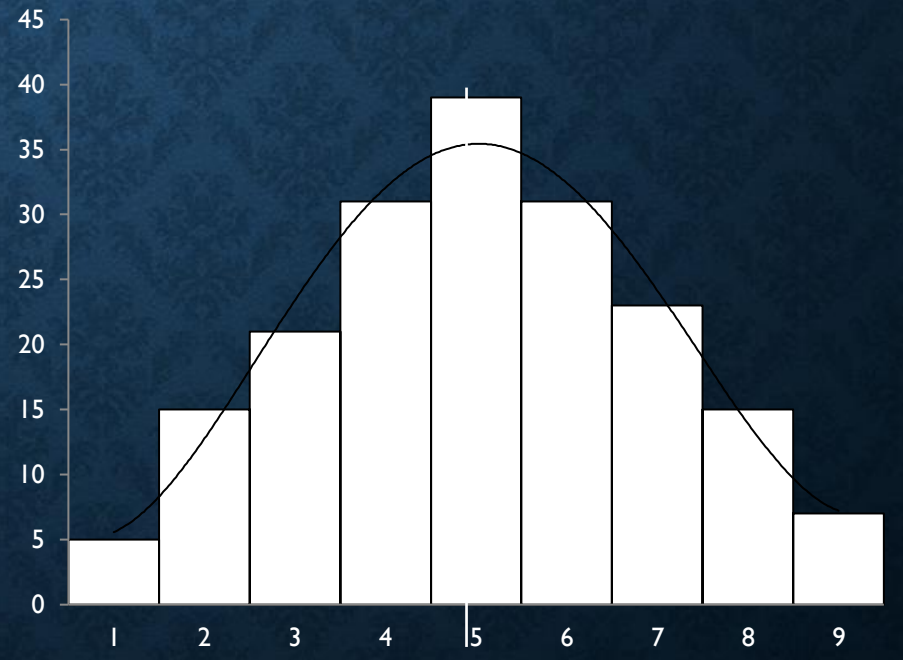
Las distribuciones pueden tener diferentes formas:

MEDIDAS DE FORMA

1. Simétrica

Si dividimos la distribución en dos, a ambos lados son iguales, por lo que la mediana, la moda y la media aritmética son iguales.

Las dos ramas son asíntotas del eje de las abscisas se le denomina Distribución de Gauss. Esta distribución se le conoce como **curva de errores o distribución normal** y es la más importante en estadística.



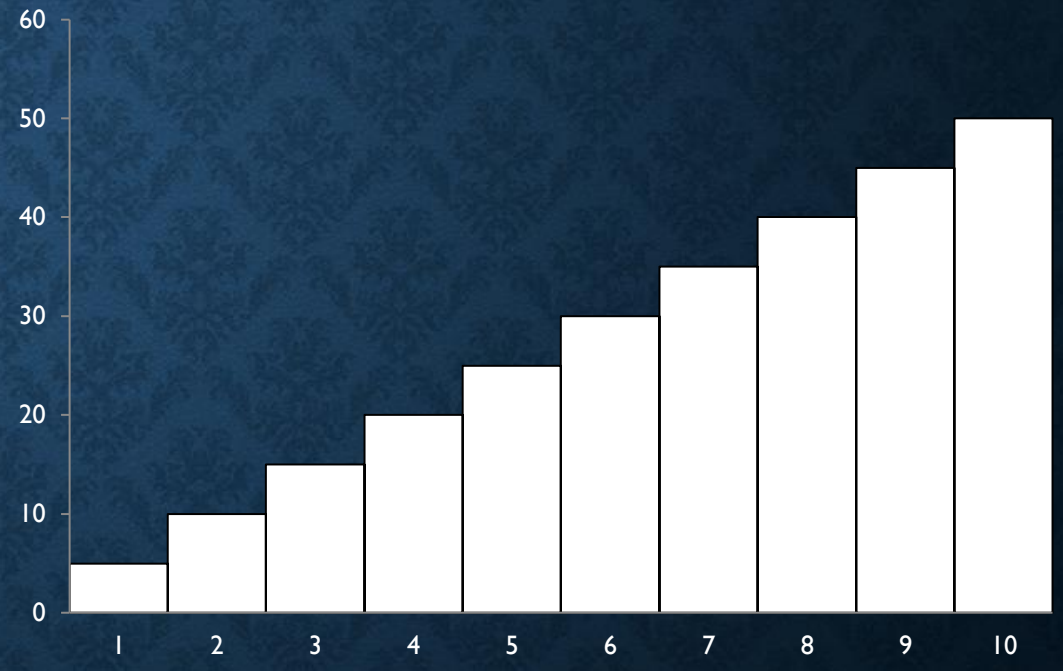
Media = mediana = moda

MEDIDAS DE FORMA

2. Asimétrica

Datos sesgados a la izquierda (sesgo negativo)

Poseen una cola izquierda más larga, en tanto que la media y la mediana se encuentran a la izquierda de la Moda. Suelen tener una media menor a la mediana.



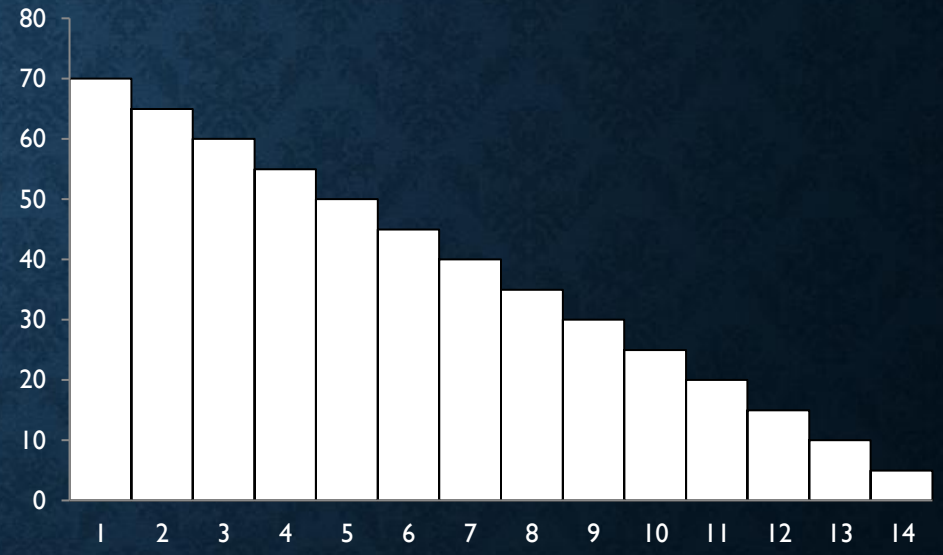
$$\overline{X} < Me < Mo$$

MEDIDAS DE FORMA

2. Asimétrica

Datos sesgados a la derecha (sesgo positivo)

Poseen una cola más larga a la derecha, en tanto que la media y la mediana se encuentran a la derecha de la Moda. Suelen tener una media mayor a la mediana.



$$Mo < Me < \overline{X}$$

MEDIDAS DE FORMA

Otras medidas o Coeficientes

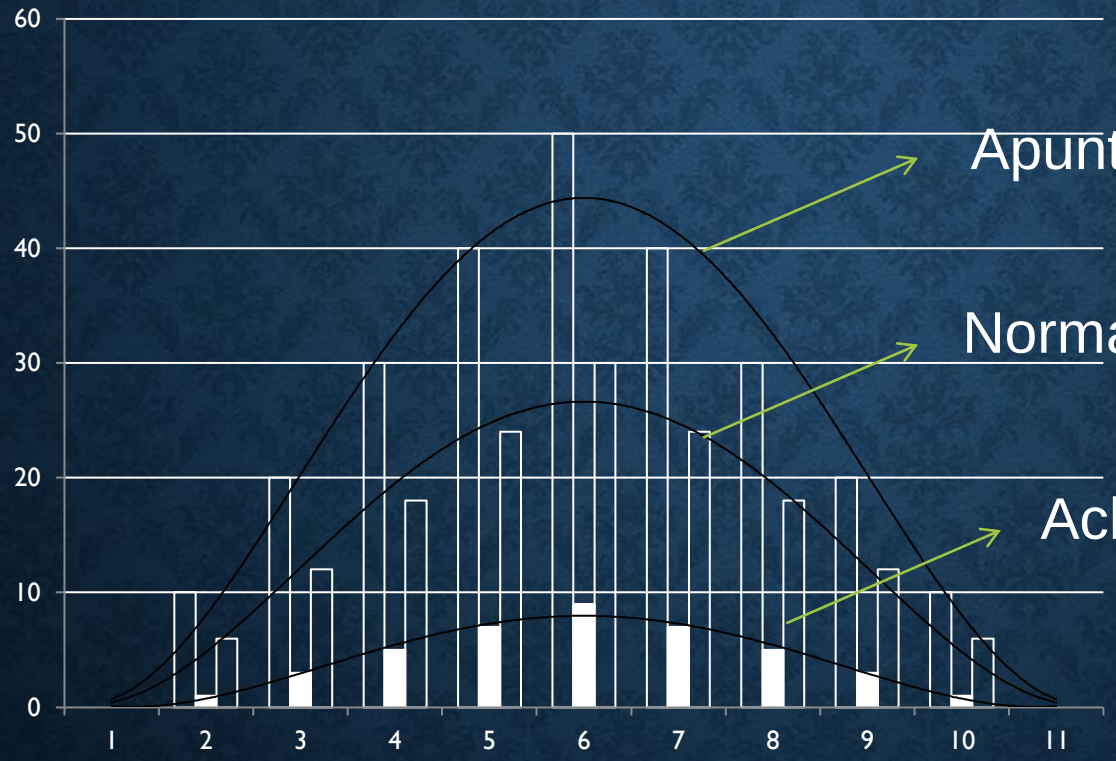
- -Asimetría
- -Kurtosis o Apuntamiento

Además de la **posición** y la **dispersión** de los datos, otra medida de interés en una distribución de frecuencias es la simetría y el apuntamiento o kurtosis.

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

Si *As* = 0 si la distribución es simétrica alrededor de la media.
Si *As* < 0 si la distribución es asimétrica a la izquierda
Si *As* > 0 si la distribución es asimétrica a la derecha

MEDIDAS DE FORMA



Apuntada / leptocúrtica

Normal / Mesocúrtica

Achatada / platicúrtica

MEDIDAS DE FORMA

Coeficiente de Curtosis

$$K = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n \cdot s^4}$$

-Si *K* = 3 la distribución se dice normal (similar a la distribución normal de Gauss) y recibe el nombre de ***mesocúrtica***.

-Si K > 3, la distribución es más puntiaguda que la anterior y se llama ***leptocúrtica***, (mayor concentración de los datos en torno a la media).

- Si K < 3 la distribución es más plana y se llama ***platicúrtica***.

MEDIDAS DE FORMA

Xi	**fi**	**Fi**	**hi**	**Hi**		
1	6	6	0,3	0,3		
2	7	13	0,35	0,65		
3	4	17	0,2	0,85		
4	2	19	0,1	0,95		
5	1	20	0,05	1		
$N = \Sigma$	20		1			