Universidad Nacional de Colombia
Sede Medellín

Analítica descriptiva
- Explorando los datos -

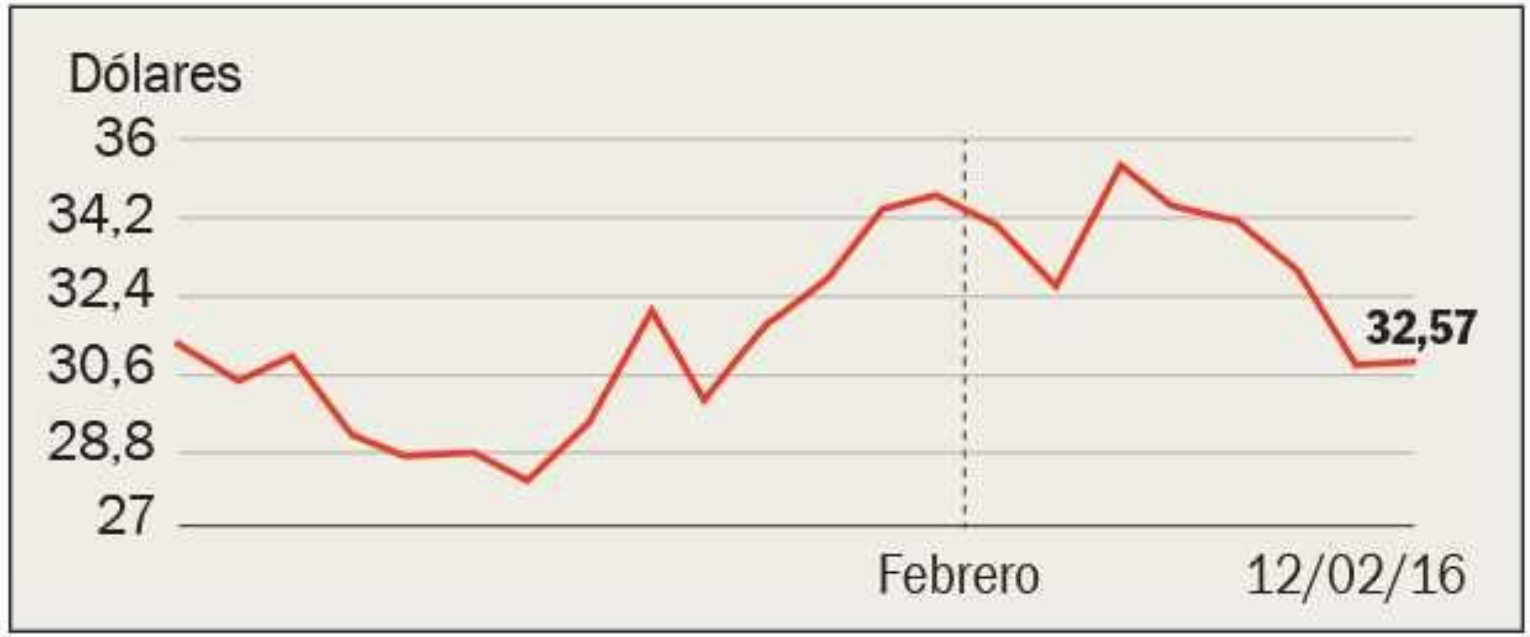
Profesora: Patricia Jaramillo A. Ph.D

Our culture encodes a strong bias either to neglect or ignore variation. We tend to focus instead on measures of central tendency, and as a result we make some terrible mistakes, often with considerable practical import.

Stephen Jay Gould, naturalist, 1941 – 2002

BARRIL SIN FONDO

El petróleo sigue bajando y se aproxima a los 20 dólares, lo que se veía imposible hace unos meses.



Fuente: revista Semana 2016

Explorando los datos

El propósito es que, datos recolectados, y que inicialmente puedan tener poco significado, se conviertan en **insumo útil** para la toma de decisiones:

- **Caracterizándolos** estadísticamente a partir de frecuencias, promedios, medidas de variabilidad, relaciones con otras variables, etc.
- **Graficando** los valores de la serie en histogramas, diagramas de dispersión (scatterplots), graficas de series de tiempo, etc. (Visualización)

Pero **lo más importante no es almacenar y recuperar datos** sino obtener conocimiento **útil** a partir de estos grandes volúmenes de datos.

Ojo: no son sólo técnicas matemáticas, **necesitan el ingenio del analista para su correcta interpretación y utilidad**

Existen muchos softwares para análisis de datos masivos

Es importante aplicar técnicas de Resumen, estadística, visualización, machine Learning, Data Mining, etc.

Pero lo más importante es entender lo que se puede y debe hacer y como se hace.

Para explorar los datos:

1. Lo mas importante es **identificar el problema** que requiere ser resuelto. Por ejemplo, *Qué pasa en el departamento donde se redujeron drásticamente las ventas? Qué pasa con las otras variables asociadas*?

2. **Recolectar** datos para ayudar a entender el problema, a través de encuestas, datos de la Web, etc.

3. **Analizar** los datos usando las herramientas que veremos **en este módulo (y otros que verán en otros cursos)**

4. Inferir posibles futuros, **Tomar decisiones**, cambiar políticas, etc.

Conceptos básicos

- **Población**: todas las entidades de interés
- **Muestra**: Subconjunto representativo de la población
- **Data set:** matriz de datos con variables en columnas y observaciones en filas
- **Variable:** característica observable que varía entre los diferentes individuos o eventos de una población. Puede ser cuantitativas o categóricas. Ejemplo, salario, genero, edad.
- **Observación (o registro):** es una lista de los valores de las variables de un miembro de la muestra.
- **Probabilidad de ocurrencia:** Se estima a partir de datos históricos para entender la frecuencias de valores de la variable en el pasado e inferir hacia el futuro..

Tipos de variables

- **Numéricas**: tiene un valor matemático: por ejemplo: salario
- **Categórica**: ordinal si existe un orden natural de sus valores. Si no, es nominal

	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	35	Male	Minnesota	1	$65,400	5
3	2	61	Female	Texas	2	$62,000	1
4	3	35	Male	Ohio	0	$63,200	3
5	4	37	Male	Florida	2	$52,000	5
6	5	32	Female	California	3	$81,400	1
7	6	33	Female	New York	3	$46,300	5
8	7	65	Female	Minnesota	2	$49,600	1
9	8	45	Male	New York	1	$45,900	5
10	9	40	Male	Texas	3	$47,700	4
11	10	32	Female	Texas	1	$59,900	4
12	11	57	Male	New York	1	$48,100	4
13	12	38	Female	Virginia	0	$58,100	3
14	13	37	Female	Illinois	2	$56,000	1
15	14	42	Female	Virginia	2	$53,400	1
16	15	38	Female	New York	2	$39,000	2
17	16	48	Male	Michigan	1	$61,500	2
18	17	40	Male	Ohio	0	$37,700	1
19	18	57	Female	Michigan	2	$36,700	4
20	19	44	Male	Florida	2	$45,200	3
21	20	40	Male	Michigan	0	$59,000	4
22	21	21	Female	Minnesota	2	$54,300	2
23	22	49	Male	New York	1	$62,100	4
24	23	34	Male	New York	0	$78,000	3
25	24	49	Male	Arizona	0	$43,200	5
26	25	40	Male	Arizona	1	$44,500	3
27	26	38	Male	Ohio	1	$43,300	1
28	27	27	Male	Illinois	3	$45,400	2
29	28	63	Male	Michigan	2	$53,900	1
30	29	52	Male	California	1	$44,100	3
31	30	48	Female	New York	2	$31,000	4

Ejemplo: encuesta sobre la gestión de un presidente de Estados Unidos

Variables Categóricas:
Genero (nominal)
Estado(nominal)
Opinión (ordinal) (1= fuertemente en desacuerdo, 2=desacuerdo, 3=neutral, 4= de acuerdo, 5=fuertemente de acuerdo)

Variables numéricas:
Edad
Hijos
Salario

Las variables numéricas pueden ser:

- **Discretas**: si se puede contar como por ejemplo, el número de hijos.
- **Continuas**: si esta en un espacio continuo como por ejemplo, salario.

- **Series de tiempo:** colección de datos a lo largo del tiempo.

Esta distinción es importante porque el tipo de análisis puede ser diferente en uno u otro caso.

Análisis estadístico de datos

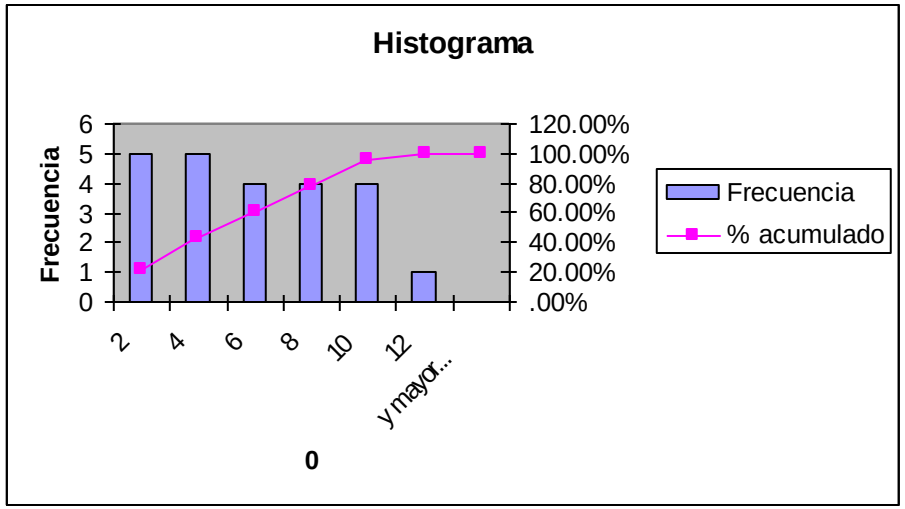
Estimar, media, varianza, histograma y demás estadísticos.

Ejemplo: edad de una población

edades
1
5
4
10
1
5
4
8
10
9
5
0
4
7
8
10
1
12
0
1
5
4
3
8

Media	5.21
Desviación estándar	3.56

Edad	*Frecuencia*	*% acumulado*
2	5	21.74%
4	5	43.48%
6	4	60.87%
8	4	78.26%
10	4	95.65%
12	1	100.00%
y mayor...	0	100.00%



Histograma

- Es un gráfico de barras que relaciona frecuencia con valores de la variable.

- Las variables numéricas se agrupan en intervalos. Cada intervalo (clase) es representado por su valor medio.

- Para construirlo se cuenta cuántas observaciones caen en cada intervalo o clase

- La selección del número de clases y su amplitud puede ser complicado: Un histograma con muy pocas clases agrupa demasiadas observaciones y uno con muchas deja muy pocas en cada clase.

Los pasos para la construcción de un histograma con los siguientes:

1. Calcule el número de intervalos k necesarios de acuerdo con el total de observaciones recolectadas n y haciendo uso de las siguientes expresiones y criterios (también importa su criterio propio):

Criterio

n	k
30 - 50	5 - 7
51 - 100	6 - 10
101 - 250	7 - 12
>250	10 - 25

Expresiones

$$k = 1 + 3{,}33 \log_{10}(n)$$

$$2^{k-1} < n < 2^k$$

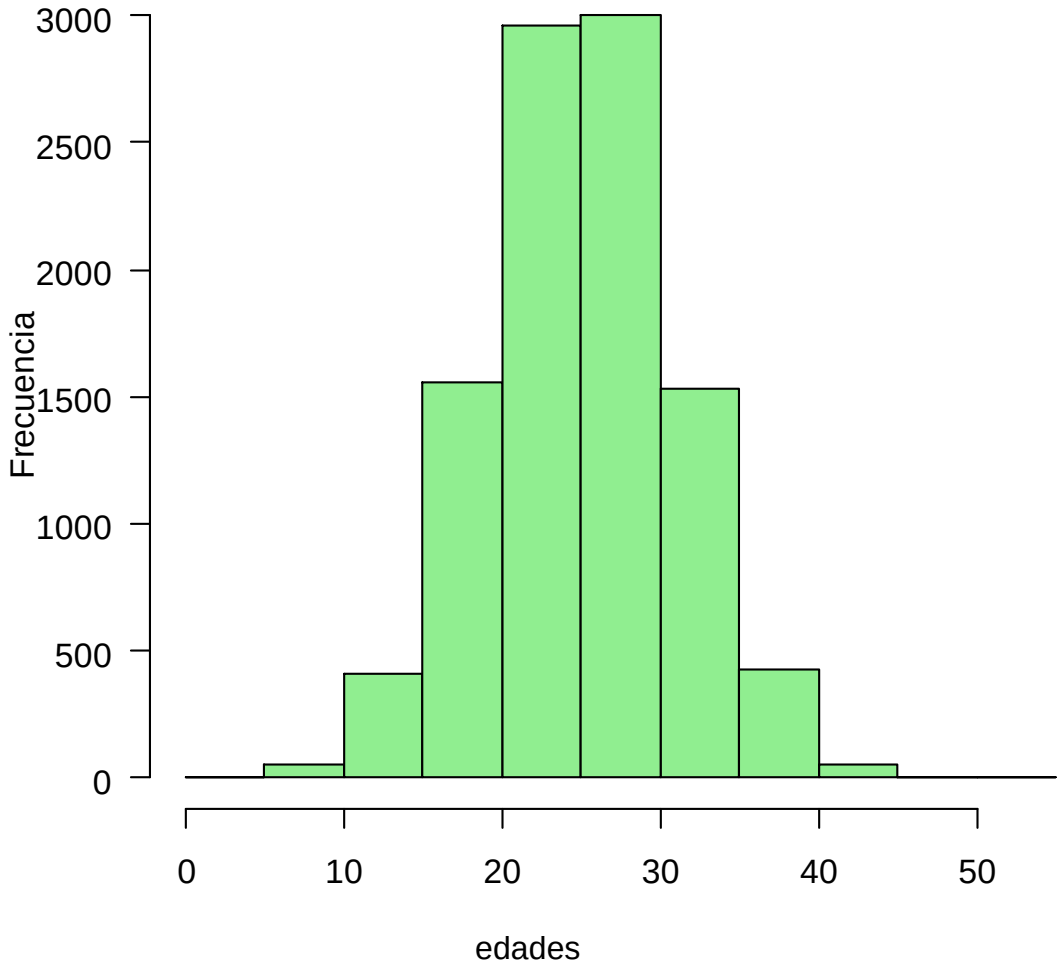
$$k = \sqrt[3]{2n}$$

$$k = \sqrt{n}$$

2. Determine la amplitud de cada uno de los intervalos de clase del histograma de acuerdo con k (preferiblemente intervalos del mismo tamaño)

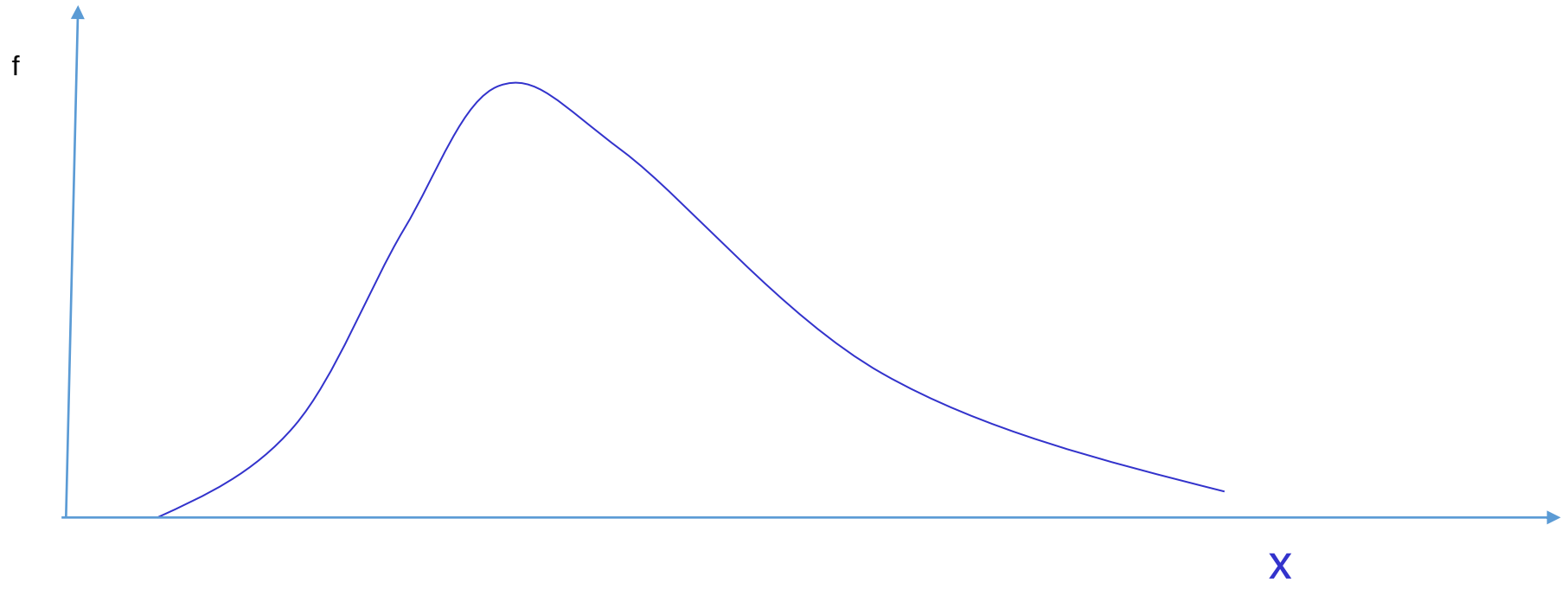
3. Cuente la cantidad de valores que caen en determinado intervalo (frecuencia absoluta de cada clase). Cada clase se convierte en el evento o valor X de la variable en cuestión.

Ejemplo histograma

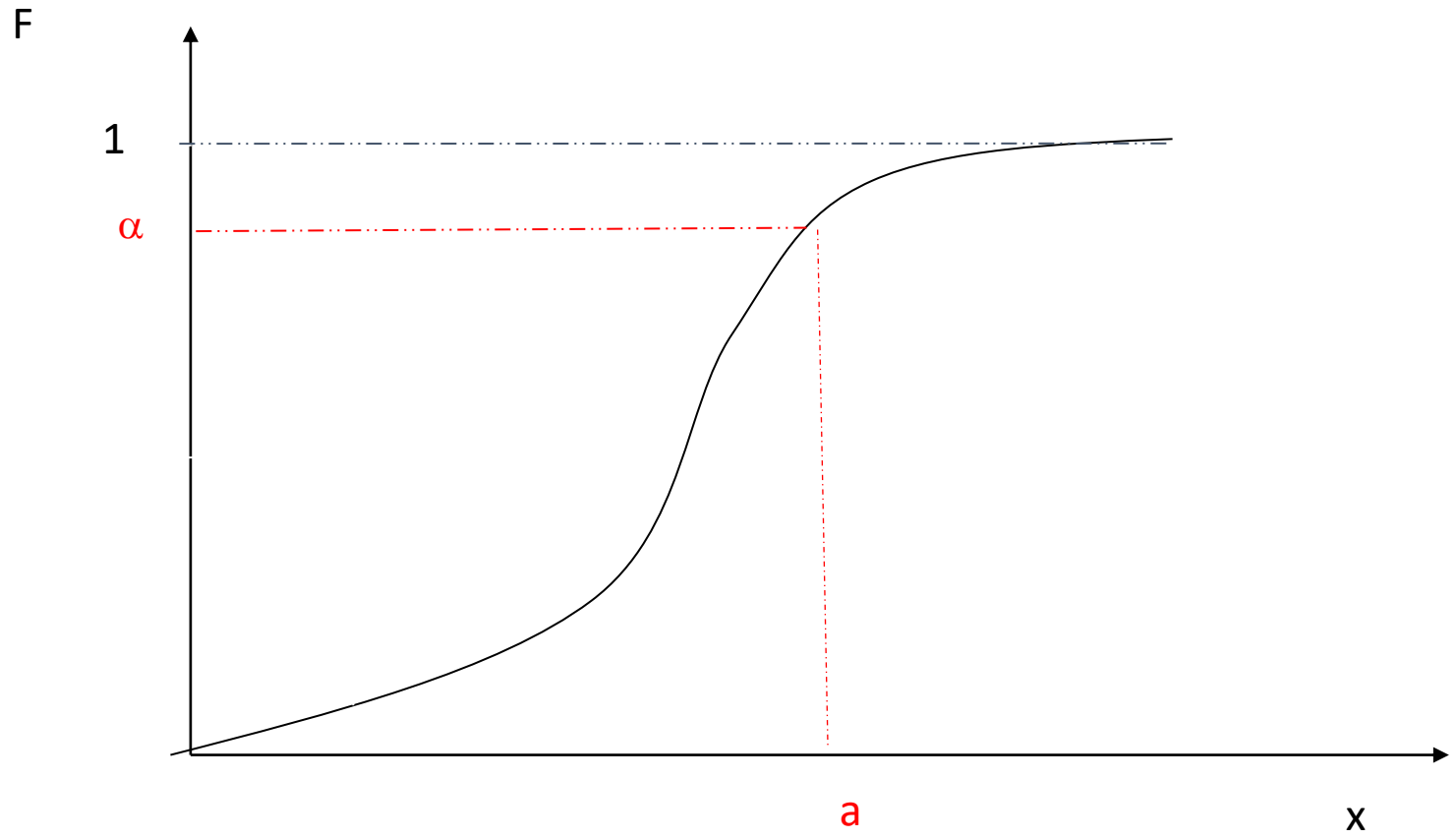


En un mundo continuo….

Para variables numéricas, cuando el tamaño de la clase es pequeñísimo, se convierte en **Función de distribución de probabilidad**



O si se elabora acumulada: Función de distribución acumulada



α es la probablidad de que x <= a

Medidas de Tendencia Central

- La media muestral es la suma de todos los valores de una variable dividida entre el número total de datos de los que se dispone.

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

- La media geométrica es la raíz n-ésima del producto de todas las observaciones disponibles.

$$g = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

- La mediana es el valor que se encuentra en la mitad de los datos ordenados.

- La moda es el valor de la variable que se presenta con mayor frecuencia, es decir, el valor que más se repite (esta es la única que tiene sentido con variables categóricas)

Medidas de Dispersión o variabilidad

- La Varianza Muestral (s^2) es el promedio de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética de los datos. La desventaja es que tiene como unidades de medida el cuadrado de las unidades de medida en que se mide la variable estudiada.

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

- La Desviación Típica o Estándar (s). Es la medida de dispersión más utilizada y conocida en Estadística. Representa la dispersión de los datos en las mismas unidades de medida de la variable.

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

Al aumentar el tamaño de la muestra, disminuye la varianza y la desviación típica.

Ejemplo interesante: importancia de la varianza frente a la media

Qué observa?

Proveedor A (peso de producto)			Proveedor B (peso de producto)	
	kg			kg
	102,61			103,21
	103,25			93,66
	96,34			120,87
	96,27			110,26
	103,77			117,31
	97,45			110,23
	98,22			70,54
	102,76			39,53
	101,56			133,22
	98,16			101,91
media	100,039		media	100,074
Varianza	9,109832222		Varianza	736,3652711
Desviación estándar	3,018249861		Desviación estándar	27,13605113

Coeficiente de variación: Corresponde a la desviación estándar como porcentaje de la media aritmética.

$$CV = \frac{s}{\bar{x}}$$

Cuando la media es próxima a 0, CV pierde significado, ya que puede dar valores muy grandes, que no necesariamente implican una alta dispersión de los datos.

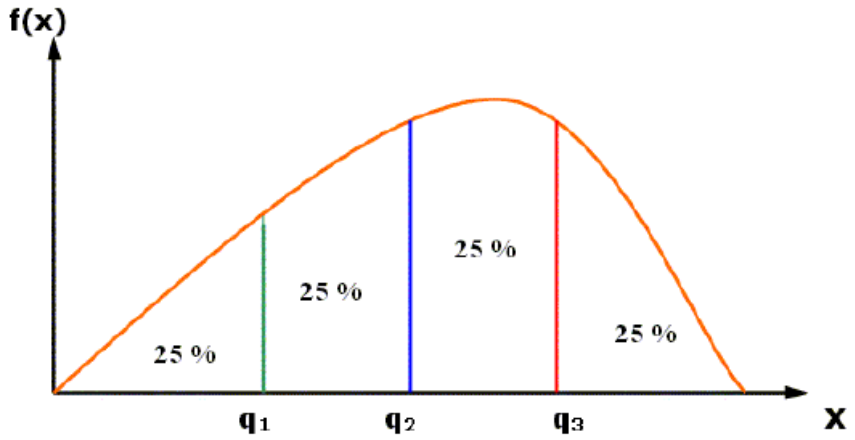
Otras Medidas de Dispersión o variabilidad

- **Rango:** el valor máximo menos el mínimo.

- **Rango Intercuartil (IQR).** El tercer cuartil menos el primer cuartil tal que, entre esos valores, está el 50% de la muestra. Es menos sensible que el rango a los outliers.

Medidas de Posición

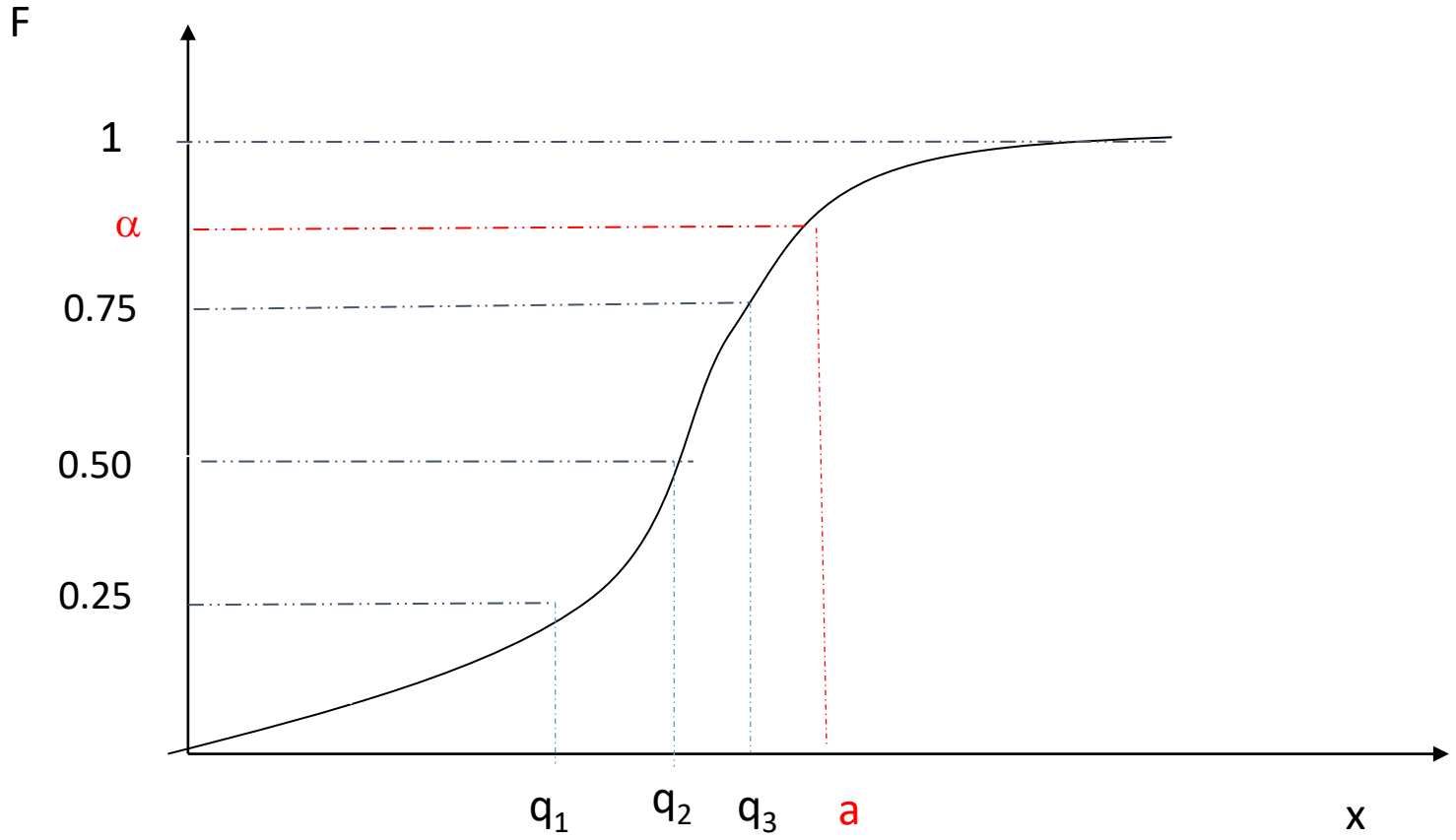
Los cuartiles son valores de la variable que dejan por debajo de sí, el 25%, 50% y el 75% de las observaciones.



El percentil es el valor de la variable que indica el porcentaje de observaciones iguales o menores a esa cifra.

Los decíles son los valores de X que son iguales o dejan por debajo al 10%, 20%, 30%, 40%,…, 100% de las observaciones.

Función de distribución acumulada



α es la probablidad de que $x \leq a$

Medidas de forma: Coeficiente de sesgo

Que tan sesgada hacia un lado esta la función de distribución. Puede estar:

- Simétrica
- Sesgada hacia la izquierda
- Sesgada hacia la derecha

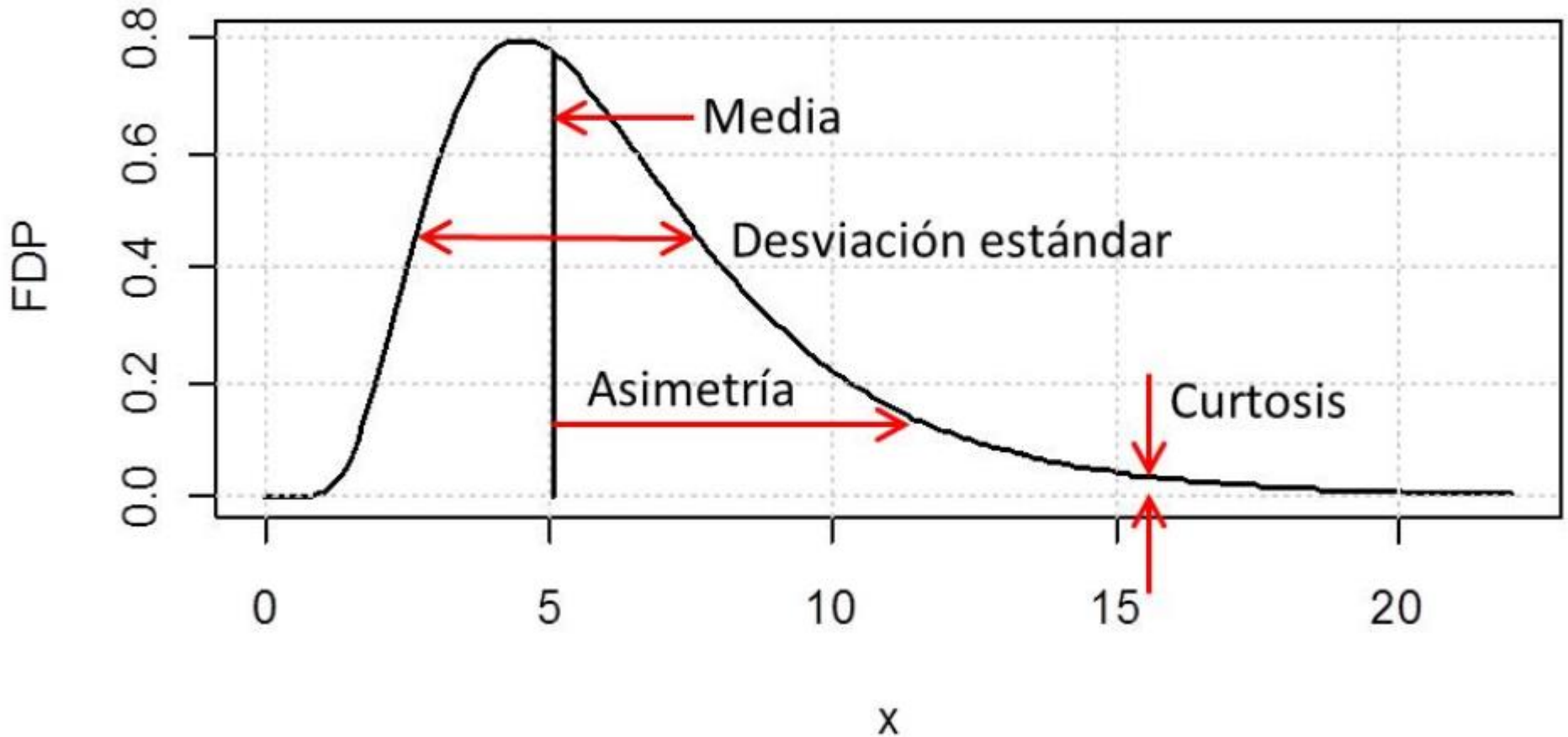
Cuando el coeficiente es mayor de 1 o menor de -1: función altamente sesgada

Cuando el coeficiente esté entre -0.5 y 0.5 la función es cercana a la simétrica

Medidas de forma: Kurtosis

Mide que tan achatada o empinada es la función de distribución. Se compara con la normal y puede ser

- **Mesocúrtica** (como la función normal), curtosis = 3
- **Platicúrtica** (relativamente aplanada cerca a la media), curtosis <3
- **Leptocúrtica** (relativamente levantada o con un pico cerca de la media y declina rapidamente), curtosis >3


FDP
0.0
0.2
0.4
0.6
0.8
Media
Desviación estándar
Asimetría
Curtosis
0
5
10
15
20
x

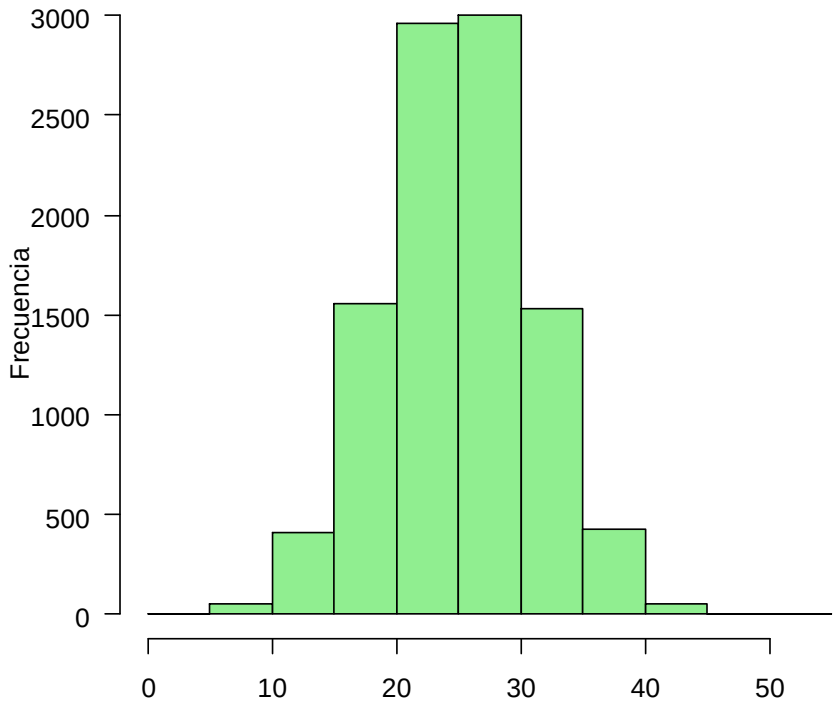
Para pensar…

Los modelos financieros antes de 2008, consideraban como buena aproximación de ajuste a las variables, a la normal, pero en realidad tenían **altas kurtosis**: creían que los valores extremos nunca podían ocurrir cuando en realidad tenían probabilidades más altas.

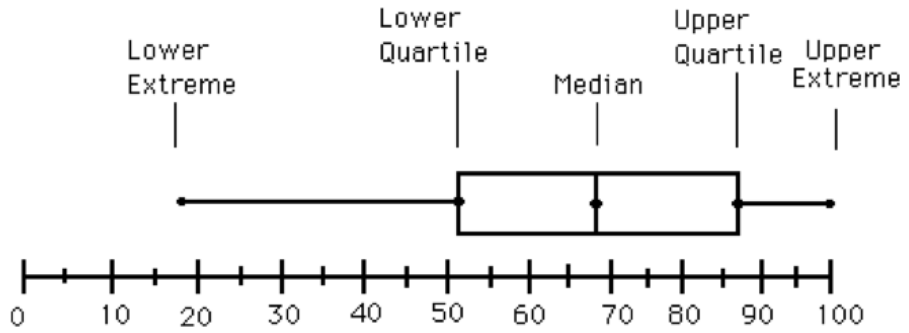
Desafortunadamente muchos valores extremos ocurrieron y eso hizo caer a la economía en una gran recesión.

Visualización: la importancia de las gráficas

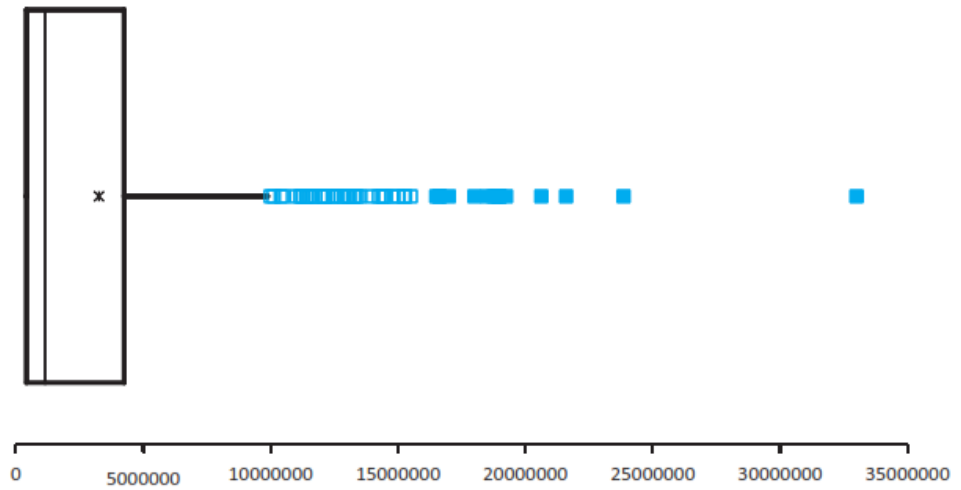
- **Histogramas:** muestran la distribución de una variable numérica. Dividen todo el rango en un numero discreto de categorías



Box and whisker Plots: muestra la mediana, los cuartiles y los valores **extremos** en una línea que permite ver la distribución de los datos. Compuesto por un "caja", y dos líneas a los lados, los "bigotes".

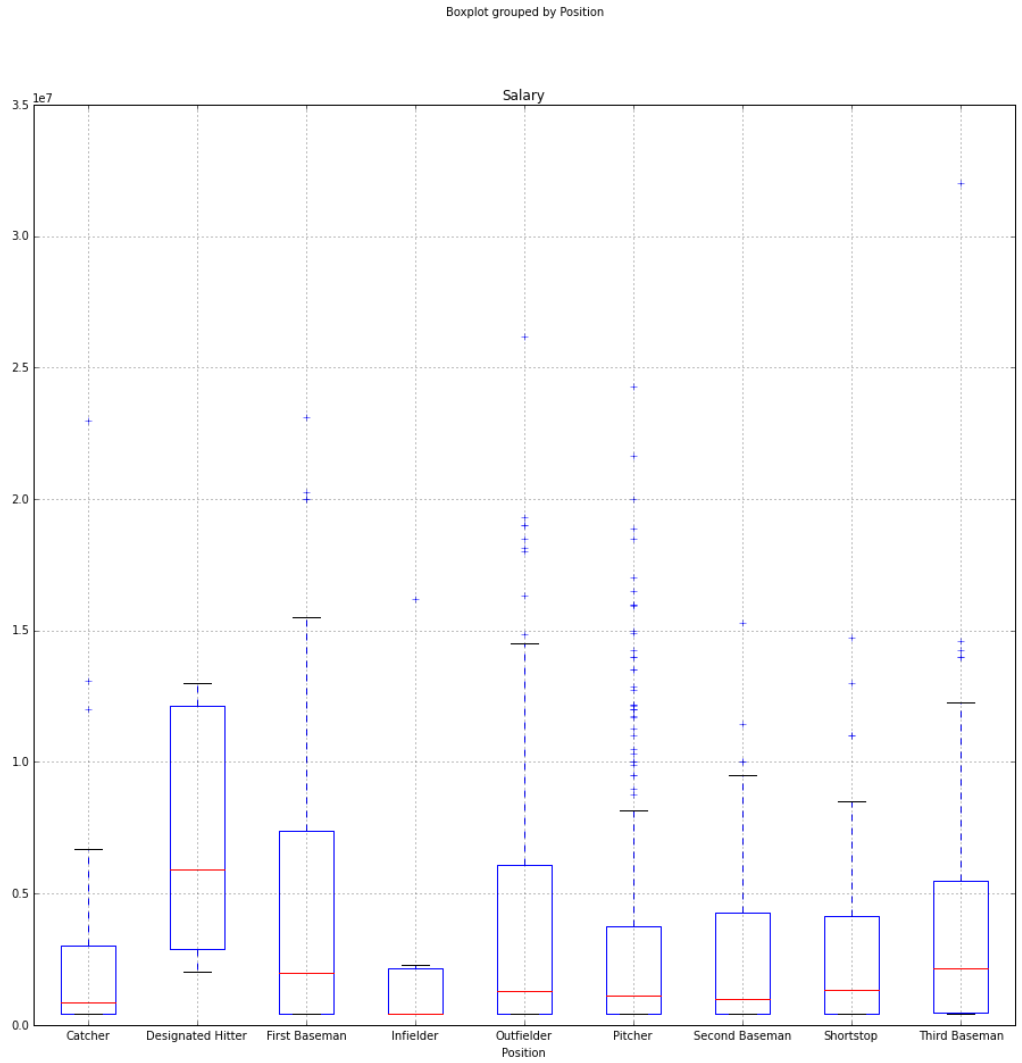


Ejemplo:

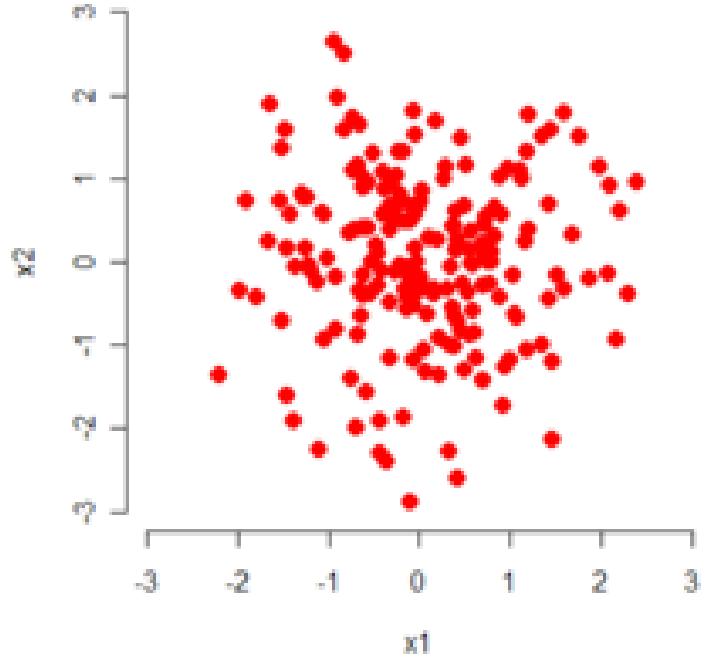


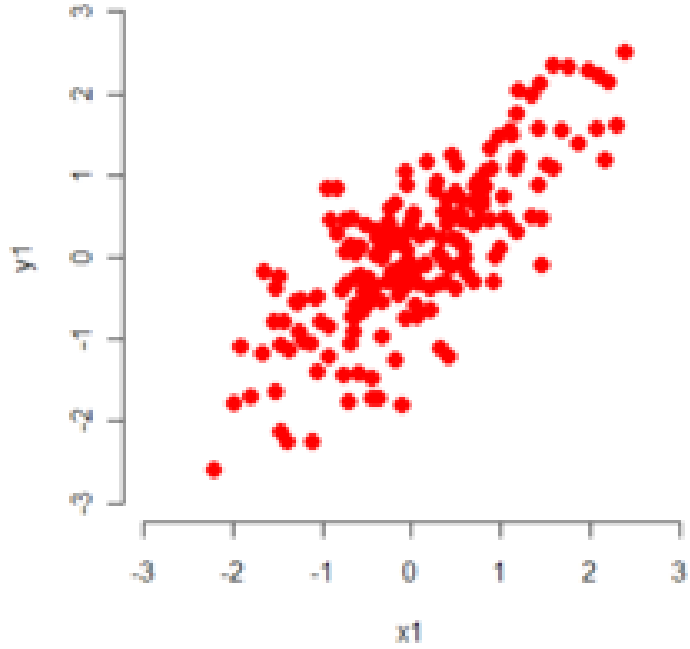
side-by-side box plots:

Útil para comparar un conjunto de variables



Scatter-plot: Permite comparar dos series de datos y valorar su interdependencia.

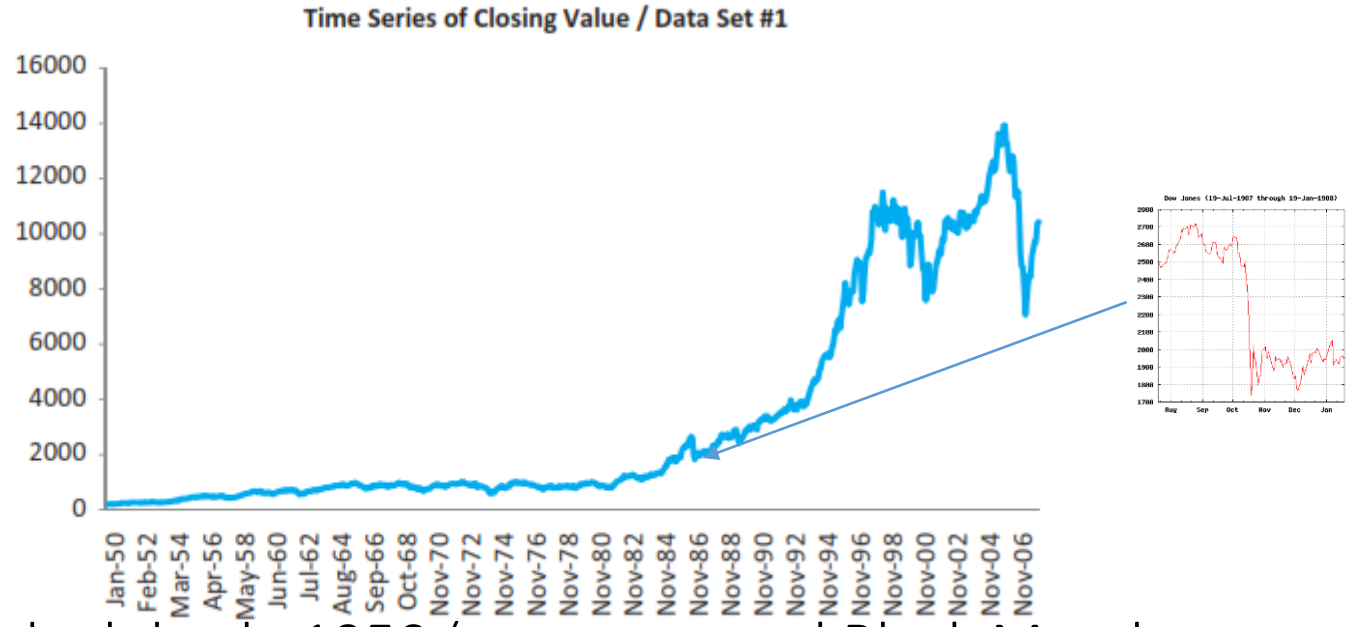




Gráfica de series de tiempo

Ejemplo: Serie de Dow Jones

One Variable Summary	Closing Value Data Set #1
Mean	3222.12
Std.Dev.	3840.15
Median	952.39
1st Quartile	755.23
3rd Quartile	3913.42



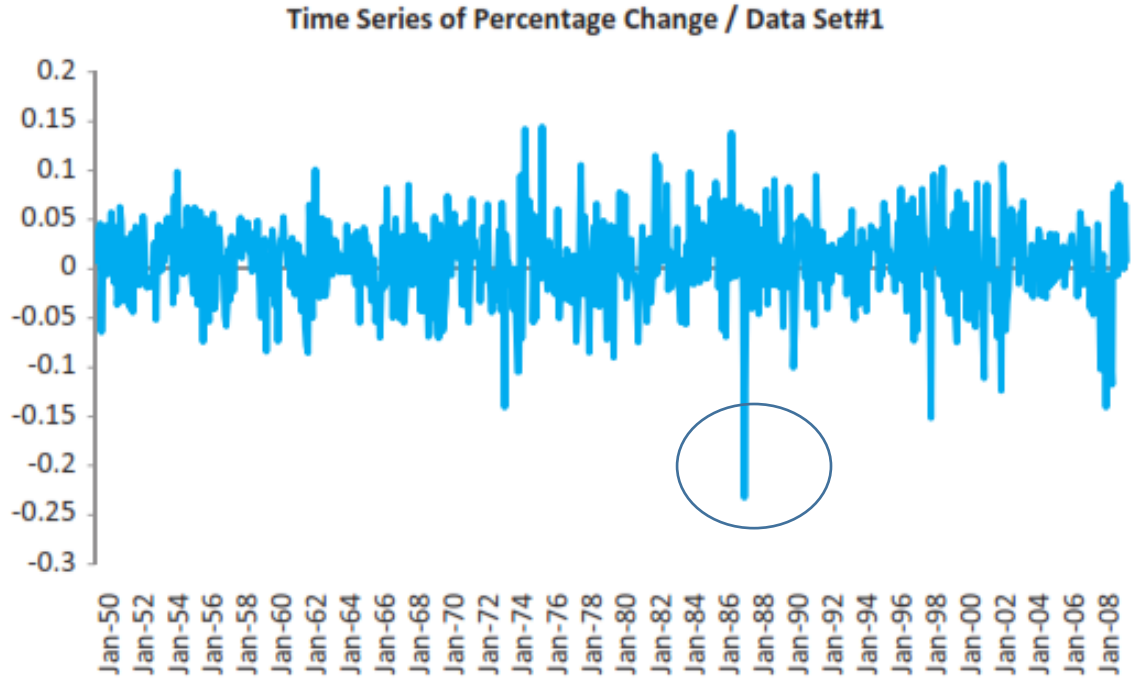
La gráfica muestra un incremento gradual desde 1950 (excepto por el Black Monday en 1987, **cuando los mercados de valores de todo el mundo se desplomaron rápidamente**) pero en la ultima década grandes oscilaciones.

Ojo con la diferencia entre la media y la mediana

En situaciones como esa, es útil mirar los porcentajes de cambio de un periodo a otro. La figura muestra que esos porcentajes han fluctuado alrededor de 0, con algunas excepciones (como el Black Monday)

Fuente: **Albright, S. C. and Winston, W. (2015). Business Analytics. *Data Analysis and decision making*. 5th ed. Cengage learning**

One Variable Summary	Percentage Change Data Set #1
Mean	0.00638
Std.Dev.	0.04175
Median	0.00851
1st Quartile	-0.01648
3rd Quartile	0.03289



Valores extremos (*Outliers*)

Son observaciones que se salen extremadamente de la moda:

- Algunos lo definen empíricamente como un valor que se sale del rango de +/-3 desviaciones desde la media.

- No es apropiado eliminarlos y al contrario pueden dar mucha información valiosa

- Una salida es hacer análisis sin eliminarlos y eliminándoles, y hacer conclusiones

Valores faltantes (*missing values*)

Esto es algo muy frecuente en la realidad. Y suelen ser ignorados aunque también se ha propuesto diferentes posibilidades para reemplazarlos:

- Usar el promedio de los demás valores de la variables, aunque en realidad no es una buena opción.
- Es mejor usar correlaciones: por ejemplo si una persona tiene 55 años, tiene un MBA de Harvard y ha sido gerente en una compañía petrolera pero no tenemos la información de su salario, no se espera que este sea el promedio. Es mas probable que exceda por ejemplo 20 millones mensuales.