

Topics

1. Introduction to Data, Types of Data
2. Levels of Measurement
3. Definition and Uses of Statistics
4. Types of Statistics – Descriptive, Inferential
5. Analyzing Individual Variables-
 - Measures of Central Tendency and Dispersion
 - Using graphs to Explore data
 - Preliminary Analysis: Outlier detection, Missing value treatment
 - Normal Distribution – Bell curve, Z score
 - Descriptive statistics using Excel
6. Analyzing Relationship among Variables
 - Correlation: Correlation coefficient, Correlation Matrix, 2D Scatter plot

What is Data?

- Data is often viewed as the lowest level of abstraction from which information and knowledge are derived.
- Data can be numbers, words, measurements, observations or even just descriptions of things. Also, data is a representation of a fact, figure and idea.
- Data on its own carries no meaning. In order for data to become information, it must be interpreted and take on a meaning.

An example of raw data table. It is just a collection of random info and data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)	Height (in)	Newspaper readership (times/wk)
2	1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30	2263	67	61	5
3	2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63	64	7
4	3	DOE01	JOE01	Elmira	New York	Male	Graduate	Math	US	26	2221	78	73	6
5	4	DOE02	JOE02	Lackawana	New York	Male	Graduate	Econ	US	33	1716	78	68	3
6	5	DOE03	JOE03	Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65	71	6
7	6	DOE04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69	67	5
8	7	DOE05	JOE05	Cimax	North Carolina	Male	Graduate	Politics	US	39	1577	96	70	5
9	8	DOE03	JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US	21	1842	87	62	5
10	9	DOE04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	18	1813	91	62	6
11	10	DOE05	JANE05	New York	New York	Female	Graduate	Math	US	33	2041	71	66	5
12	11	DOE06	JOE06	Hot Coffe	Mississippi	Male	Undergraduate	Econ	US	18	1787	82	67	3
13	12	DOE06	JANE06	Java	Virginia	Female	Graduate	Math	US	38	1513	79	59	5
14	13	DOE07	JOE07	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	30	1637	79	63	4
15	14	DOE08	JOE08	Moscow	Russia	Male	Graduate	Politics	Russia	30	1512	70	75	6
16	15	DOE07	JANE07	Drunkard Creek	New York	Female	Undergraduate	Math	US	21	1338	82	64	5
17	16	DOE08	JANE08	Mexican Hat	Utah	Female	Undergraduate	Econ	US	18	1821	80	63	3
18	17	DOE09	JANE09	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19	1494	75	60	3
19	18	DOE10	JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	31	2248	95	59	4
20	19	DOE11	JANE11	Caracas	Venezuela	Female	Undergraduate	Math	Venezuela	18	2252	92	68	5
21	20	DOE09	JOE09	San Juan	Puerto Rico	Male	Graduate	Politics	US	33	1923	95	63	7
22	21	DOE12	JANE12	Remote	Oregon	Female	Undergraduate	Econ	US	19	1727	67	62	7
23	22	DOE10	JOE10	New York	New York	Male	Undergraduate	Econ	US	21	1872	82	73	4
24	23	DOE13	JANE13	The X	Massachusetts	Female	Graduate	Politics	US	25	1767	89	68	6
25	24	DOE14	JANE14	Beijing	China	Female	Undergraduate	Math	China	18	1643	79	65	6
26	25	DOE11	JOE11	Stockholm	Sweden	Male	Undergraduate	Politics	Sweden	19	1919	88	64	4
27	26	DOE12	JOE12	Embarrass	Minnesota	Male	Graduate	Econ	US	28	1434	96	71	4
28	27	DOE13	JOE13	Intercourse	Pennsylvania	Male	Undergraduate	Math	US	20	2119	88	71	5
29	28	DOE15	JANE15	Loco	Oklahoma	Female	Undergraduate	Econ	US	20	2309	64	68	6
30	29	DOE14	JOE14	Buenos Aires	Argentina	Male	Graduate	Politics	Argentina	30	2279	85	72	3
31	30	DOE15	JOE15	Acme	Louisiana	Male	Undergraduate	Econ	US	19	1907	79	74	3

Exploring Data

Generally one of the first things to do with new data is to get to know it by asking some general questions like but not limited to the following:

- What variables are included? What information are we getting?
- What is the format of the variables: string, numeric, etc.?
- What type of variables: categorical, continuous, and discrete?
- Is this sample or population data?

After looking at the data you may want to know

- How many males/females?
- What is the average age?
- How many undergraduate/graduates students?
- What is the average SAT score? It is the same for graduates and undergraduates?
- Who reads the newspaper more frequently: men or women?

Types of Data / Variables

Categorical Data is the data that is non numeric.

e.g.. Favorite color, Place of Birth, Types of Car

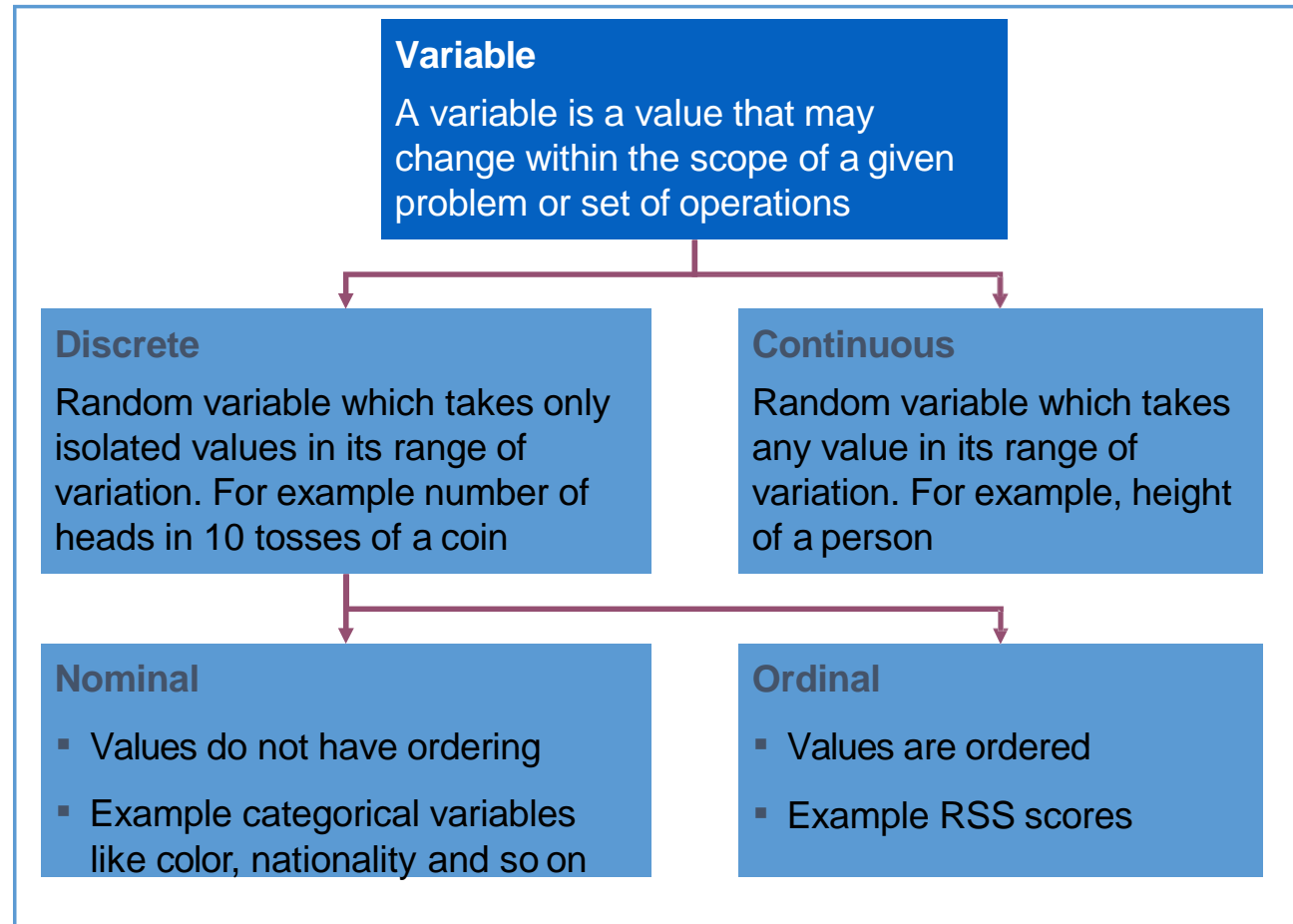
Quantitative Data is numerical. There are 2 types of quantitative data.

1. Discrete data can only take specific values;


e.g. shoe size, number of brothers, number of cars in a car park.

2. Continuous data can take any numerical value;

e.g. height, mass, length.



Examine the differences between Categorical and quantitative data.

Categorical Data	Quantitative Data
<p>Deals with descriptions.</p> <ol style="list-style-type: none"> 1. Data can be observed but not measured. 2. Colors, textures, smells, tastes, appearance, beauty, etc. 3. Categorical → Category 4. Ex: Oil Painting 	<ol style="list-style-type: none"> 1. Deals with numbers. 2. Data which can be measured. 3. Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc. <p>Quantitative → Quantity</p>
 <ol style="list-style-type: none"> 1. blue/green color, gold frame 2. smells old and musty 3. texture shows brush strokes of oil paint 4. peaceful scene of the country 5. masterful brush strokes 	<ol style="list-style-type: none"> 1. picture is 10" by 14" 2. with frame 14" by 18" 3. weighs 8.5 pounds 4. surface area of painting is 140 sq. in. 5. cost \$300

What is Statistics?

Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting numerical data to assist in making more effective decisions.

μ

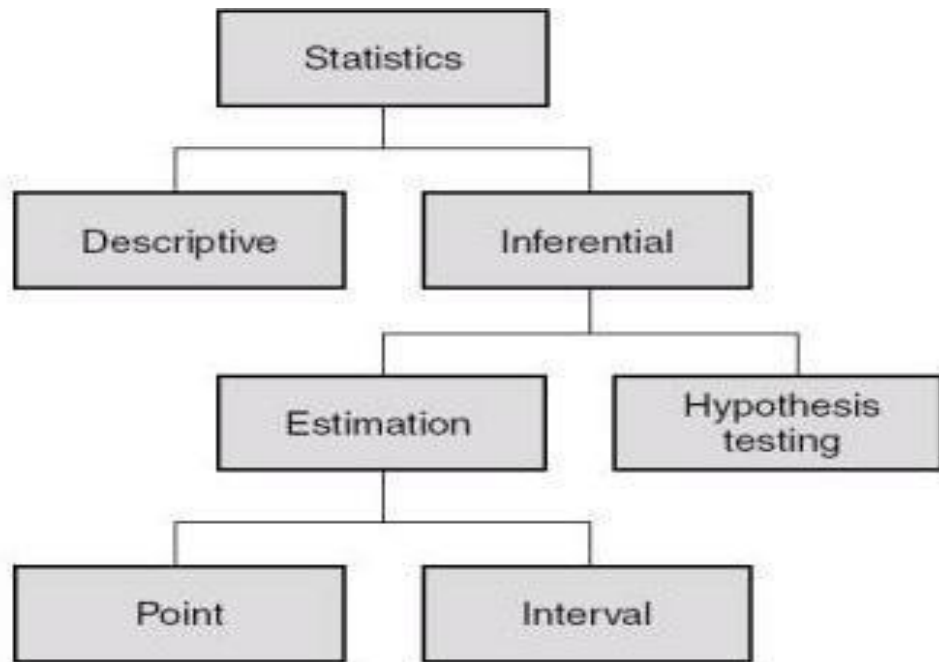
λ

Σ

σ

β

Types of Statistics



Descriptive Statistics

Study the basic features of the data that describe what is or what the data shows.

Statistical methods can be used to summarize or describe a collection of data.

involves the analysis of numeric data, pictures, graphs and figures.

Inferential Statistics

Study patterns, randomness and uncertainty in the data.

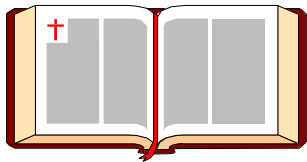
used to draw inferences about the process or population being studied .

used to make conclusions and future predictions by analysing numeric data.

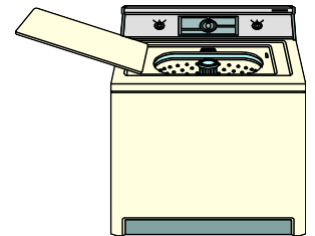
Descriptive Statistics

Descriptive Statistics: Methods of organizing, summarizing, and presenting data in an informative way.

EX 1: A Gallup poll found that 49% of the people in a survey knew the name of the first book of the Bible. The statistic 49 describes the number out of every 100 persons who knew the answer.



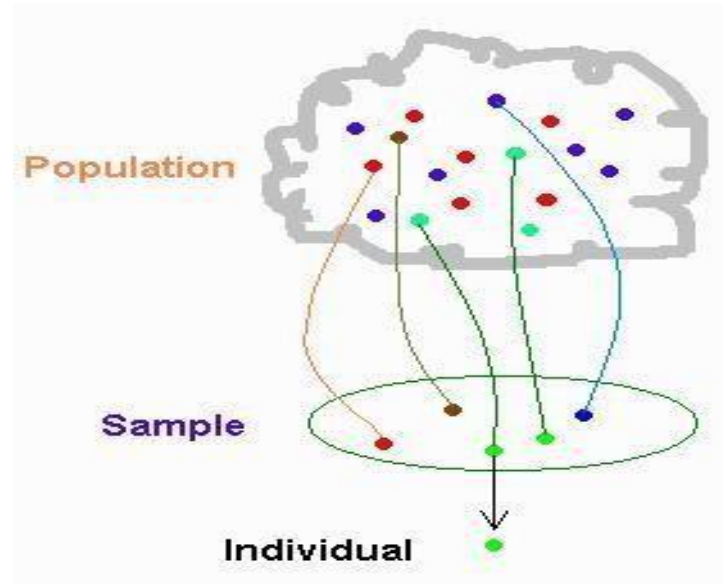
EX 2: According to Consumer Reports, General Electric washing machine owners reported 9 problems per 100 machines during 2001. The statistic 9 describes the number of problems out of every 100 machines.



Inferential Statistics

Inferential Statistics: A decision, estimate, prediction, or generalization about a population, based on a sample.

A **Population** is a **Collection** of all possible individuals, objects, or measurements of interest.



A **Sample** is a portion, or part, of the population of interest

Examples of inferential statistics

Example 1: TV networks constantly monitor the popularity of their programs by hiring Nielsen and other organizations to sample the preferences of TV viewers.



Example 2: Wine tasters sip a few drops of wine to make a decision with respect to all the wine waiting to be released for sale.



Example 3: The accounting department of a large firm will select a sample of the invoices to check for accuracy for all the invoices of the company.



"HOWEVER, BY USING AN ALTERNATE METHOD OF ACCOUNTING...."

The data type and Statistical StatisticalAnalysis



The type(s) of data collected in a study determine the type of statistical analysis used.



One of the primary purposes of classifying variables according to their level or scale of measurement is to facilitate the choice of a statistical analysis used to analyze the data.

There are certain statistical analyses which are only meaningful for data which are measured at certain measurement scales.

Statistical representation of data

For example ...

Categorical data are commonly summarized using ?Frequencies/percentages? (or ?proportions?).

11% of students have a tattoo

2%, 33%, 39%, and 26% of the students in class are, respectively, freshmen, sophomores, juniors, and seniors.

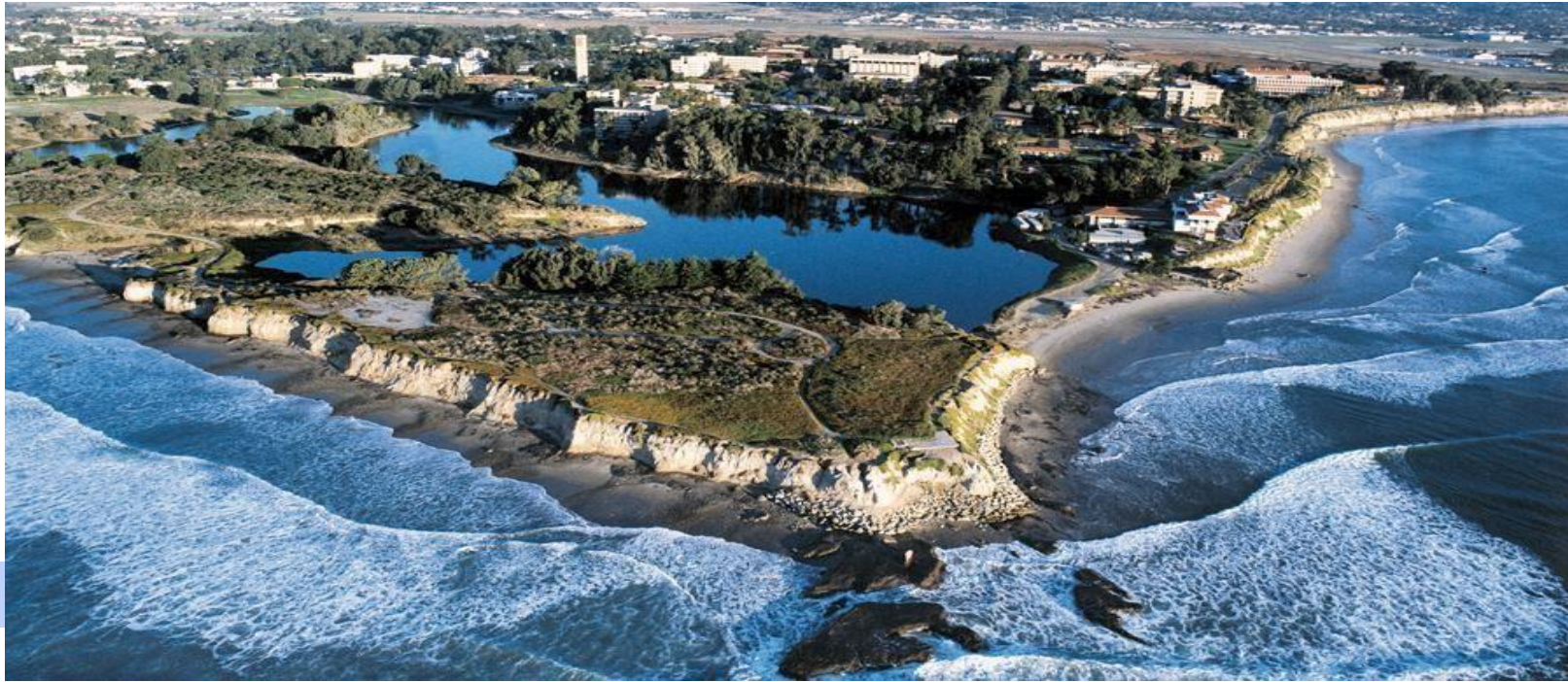
And for example ?

Measurement data are typically summarized using ?averages? (or ?means?).

Average number of siblings Fall 1998 Stat 250 students have is 1.9.

Average weight of male Fall 1998 Stat 250 students is 173 pounds.

Average weight of female Fall 1998 Stat 250 students is 138 pounds.



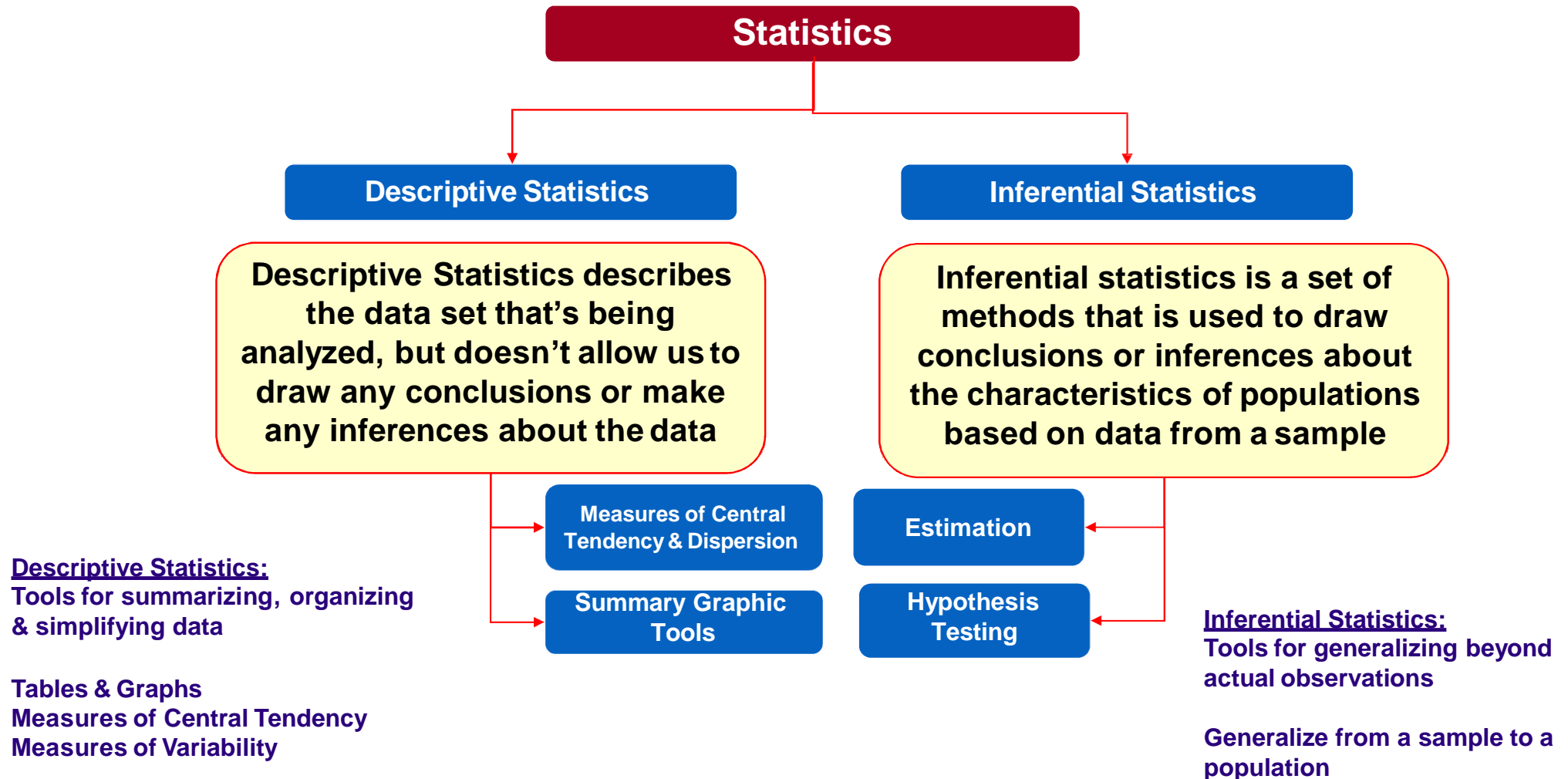
Analyzing Individual Variables- Univariate Independent Analysis

- Measures of Central Tendency
- Measures of Dispersion/Variability
- Using graphs to Explore data

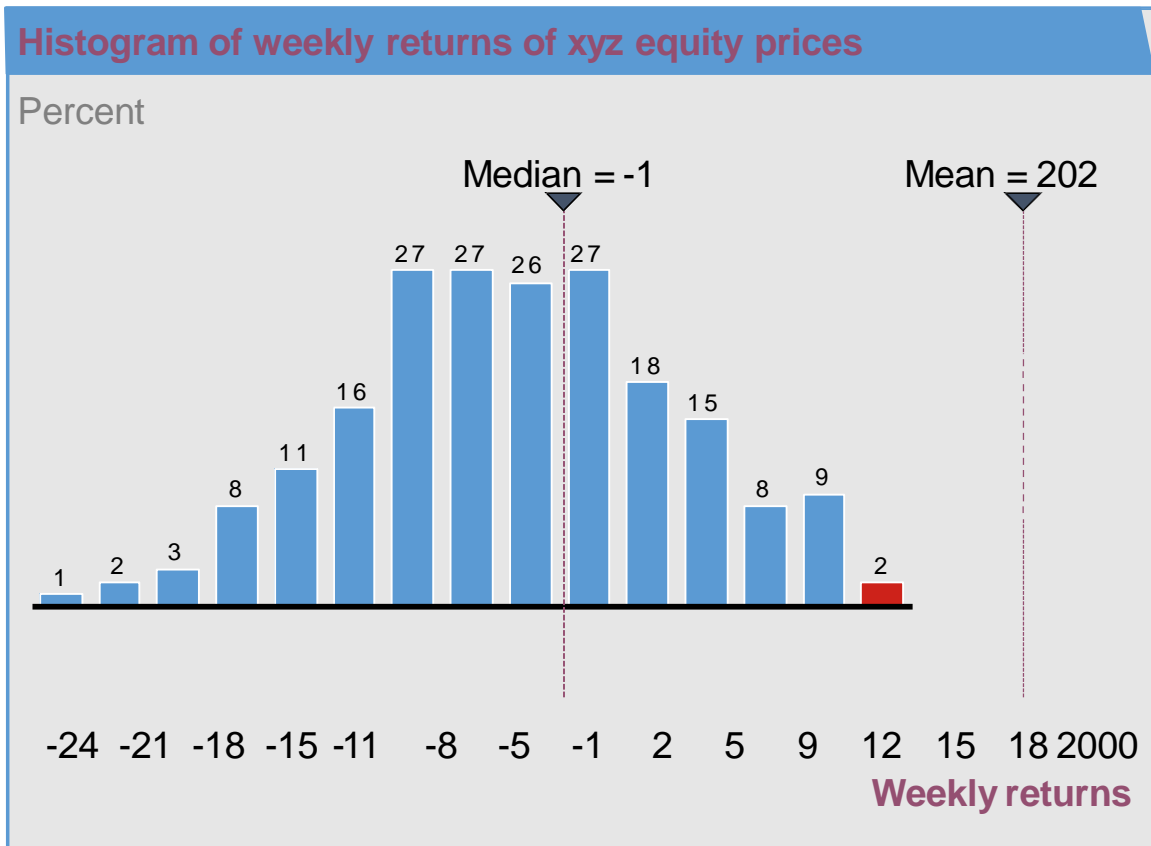
Preliminary Analysis:

- Missing data
- Outlier detection
- Normal Distribution

Recall the Branches of Statistics



Mean, median and mode are different measures of central tendency



Measure	
Mean	<div><div></div>It is the easiest metric to understand and communicate</div>
	<div><div></div>Mean is prone to presence of outliers</div>
Median	<div><div></div>Median is a more “robust” to presence of outliers</div>
	<div><div></div>It is more complicated to communicate</div>
Mode	<div><div></div>Not very practical since it is affected by skewness</div>
	<div><div></div>Most real life distributions are multimodal</div>

Other Measures of Central Tendencies...

Weighted Mean :

$$= 588/28$$

$$= 21$$

sum

Wt	Score	W x S
2	12	24
4	18	72
6	20	120
7	21	147
9	25	225
28		588

Geometric Mean: The geometric mean is the n th root of the product of the scores. (used for Logarithmic distributions)

$$(\prod X)^{\frac{1}{N}}$$

Harmonic Mean :

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \frac{1}{a_4} + \dots + \frac{1}{a_n}}$$

For the numbers 4 and 9,

$$\text{Harmonic Mean} = \frac{2}{\frac{1}{4} + \frac{1}{9}} = \frac{72}{13} = 5.54$$

Gurgaon to Delhi you travel at 40 miles per hour, **Delhi to Faridhabad** you travel at 60 miles per hour, then your average speed is given by the Harmonic Mean of 40 and 60, which is 48 miles per hour; that is; the total amount of time for the trip is the same as if you travelled the entire trip at 48 miles per hour.

The Central Tendencies Summary

Mean:

It's just the average of the data, computed as the sum of the data points divided by the number of points

Mode:

Mode is the most common value in the data set.

Tricky circumstances:

If no value occurs more than once, then there is no mode

If two values occur as frequently as each other and more frequently than any other, then there are two modes (in the same way, there could also be more than two modes).

Median:

Median is the value in the middle of the data set, when the data points are arranged from smallest to largest.

If there is an odd number of data points, then just arrange them and look for the middle value

Tricky circumstances:

If there is an even number of data points, you will need to take the average of the two middle values.

Appropriate Measures of Central Tendency

The selection should be based on level-of-measurement.

Tips for selecting

use the mode when...

- variables are measured at the nominal level
- you want a quick and easy measure for ordinal and interval-ratio variables
- you want to report the most common score

use the median when...

- variables are measured at the ordinal level
- variables measured at the interval-ratio level have highly skewed distributions
- you want to report the central score. The median always lies at the exact center of a distribution.

use the mean when...

- variables are measured at the interval-ratio level
- you want to report the typical score. The mean is "the fulcrum that exactly balances all of the scores."
- you anticipate additional statistical analysis.

Examples

- ▶ What is a typical student in the class doing? - Mean
- ▶ To compare performance of any single student against group - Median
- ▶ A parent wanting to know whether their child better or worse than typical child at his grade level - Mode

Are these sufficient?

e.g. x_1, x_2, x_3 Are the times taken to get to Delhi in different modes of transport

	Auto	Office Transport	Own Car
	7	9	1
	6	9	3
	3	9	5
	8	9	7
	12	9	9
	9	9	9
	9	9	9
	13	9	11
	13	9	13
	9	9	15
	10	9	17
Mean	9	9	9
Median	9	9	9
Mode	9	9	9

NO!!!

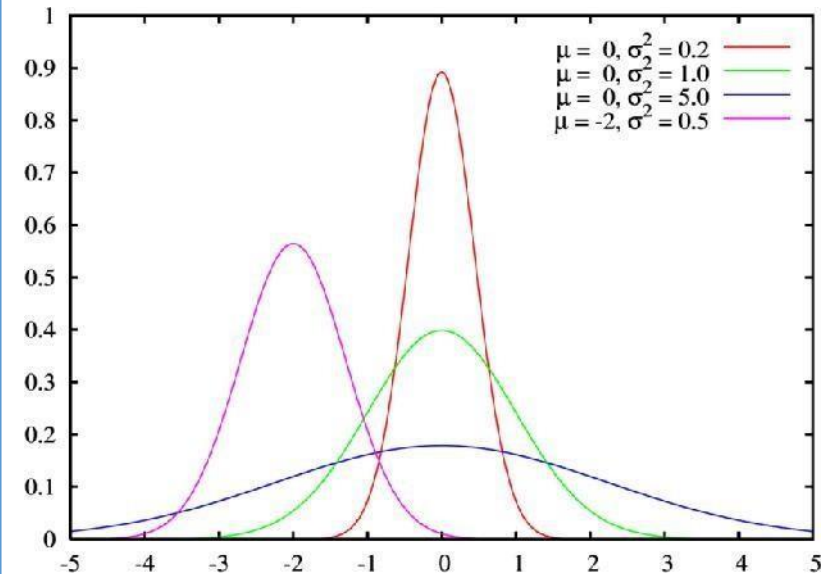
Measures of Dispersion(Variance)

Dispersion refers to the spread or variability in the data.

It determines how spread out are the scores around the mean.

The basic question being asked is how much do the scores deviate around the Mean? The more “bunched up” around the mean the better your ability to make accurate predictions.

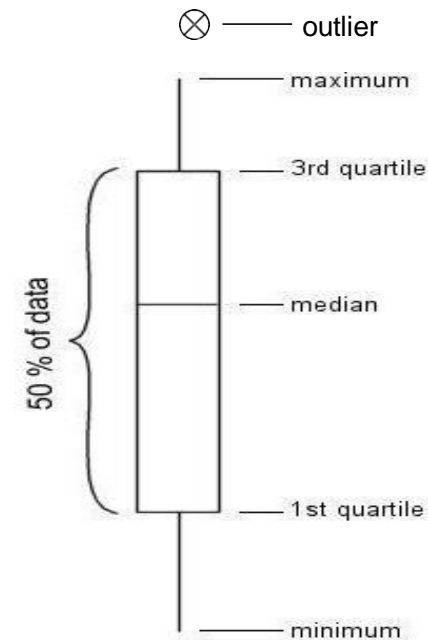
Distributions with different dispersions



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

Measures of Dispersion

- Range
 - Inter-Quartile Range
 - Mean Deviation
 - Standard Deviation
 - Variance
 - Percentiles/Quartiles
- Box-plot
 - Reveals the spread of the data
 - Outliers defined using the $Q1 - 1.5(Q3 - Q1)$ and $Q3 + 1.5(Q3 - Q1)$



Variance and Standard Deviation

Variance: the arithmetic mean of the squared deviations from the mean.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$= \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots}{N}$$

X is the value of an observation in the population

m is the arithmetic mean of the population

N is the number of observations in the population

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

Standard deviation: The square root of the variance.

Now try this...

	Auto	Office Transport	Own Car
	7	9	1
	6	9	3
	3	9	5
	8	9	7
	12	9	9
	9	9	9
	9	9	9
	13	9	11
	13	9	13
	9	9	15
	10	9	17
Mean	9	9	9
Median	9	9	9
Mode	9	9	9

Std Dev	3.0	0.0	4.9
Variance	9.2	0.0	24.0

Coefficient of Variation

The **coefficient of variation (CV)** is a normalized measure of dispersion of a probability distribution. It is defined as the ratio of the standard deviation to the mean :

- Measure of *relative* dispersion
- Always a %
- Shows variation relative to mean
- Used to compare 2 or more groups

$$CV = \frac{\sigma}{\mu} (100)$$

Which Cricketer do you like? Who is more consistent?

Dravid	150	150	130	125	145	110	100	152	120	50	128
Sehwag	230	240	150	50	173	23	20	300	45	1	128

	Dravid	Sehwag
Mean	123.636	123.636
Median	128	128
CV	24%	84%

Skewness

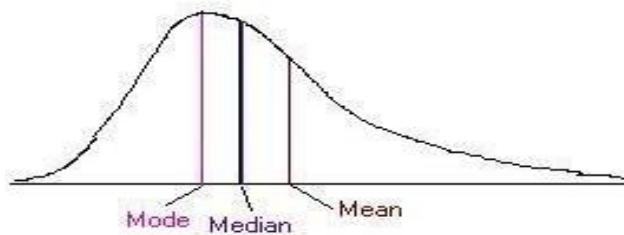
Lack of Symmetry

- A distribution is **skewed** if one of its tails is longer than the other.
- If the distribution of the data is symmetric then skewness is zero

Positive Skew

This means that the distribution has a long tail to the right

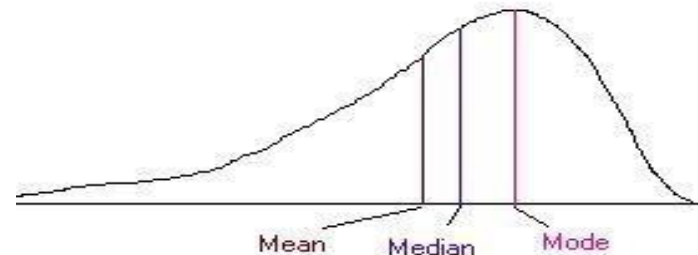
Mean > Median > Mode



Negative Skew

This means that the distribution has a long tail to the left

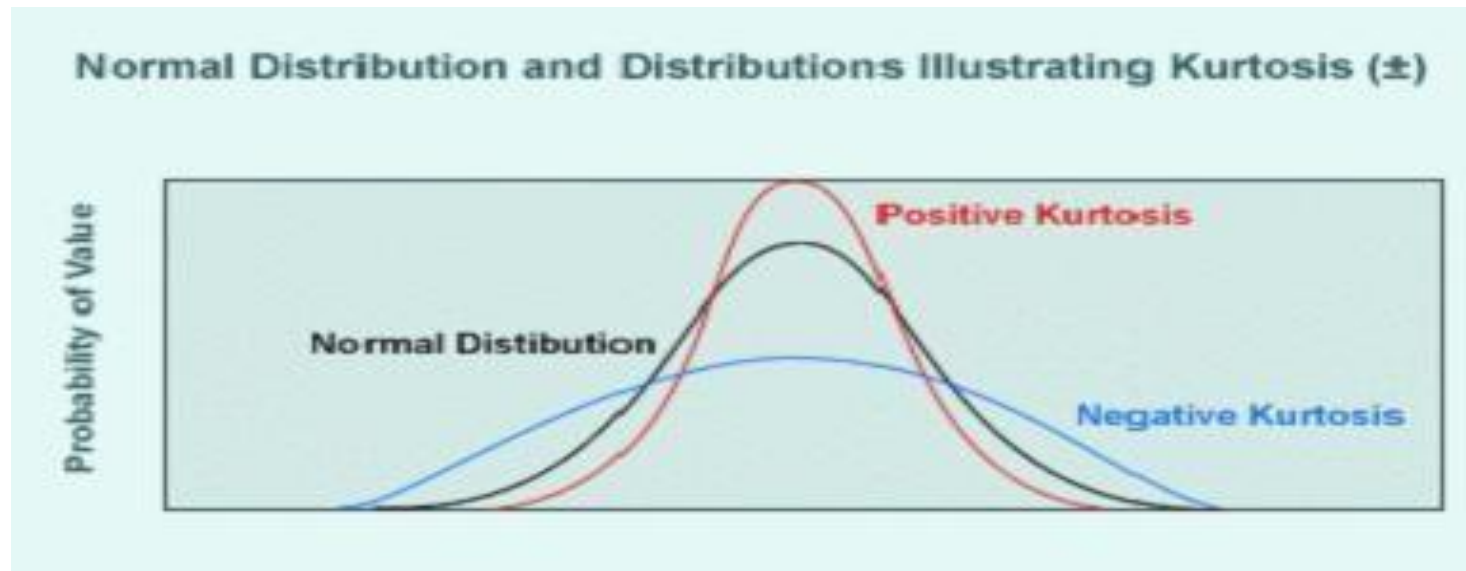
Mean < Median < Mode



Measure : Mean – Median or Mean – Mode

Kurtosis

- ▶ Kurtosis measures the "peakedness" of a distribution.
- ▶ Higher Kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations
- ▶ The Kurtosis of the Normal Distribution is 3.



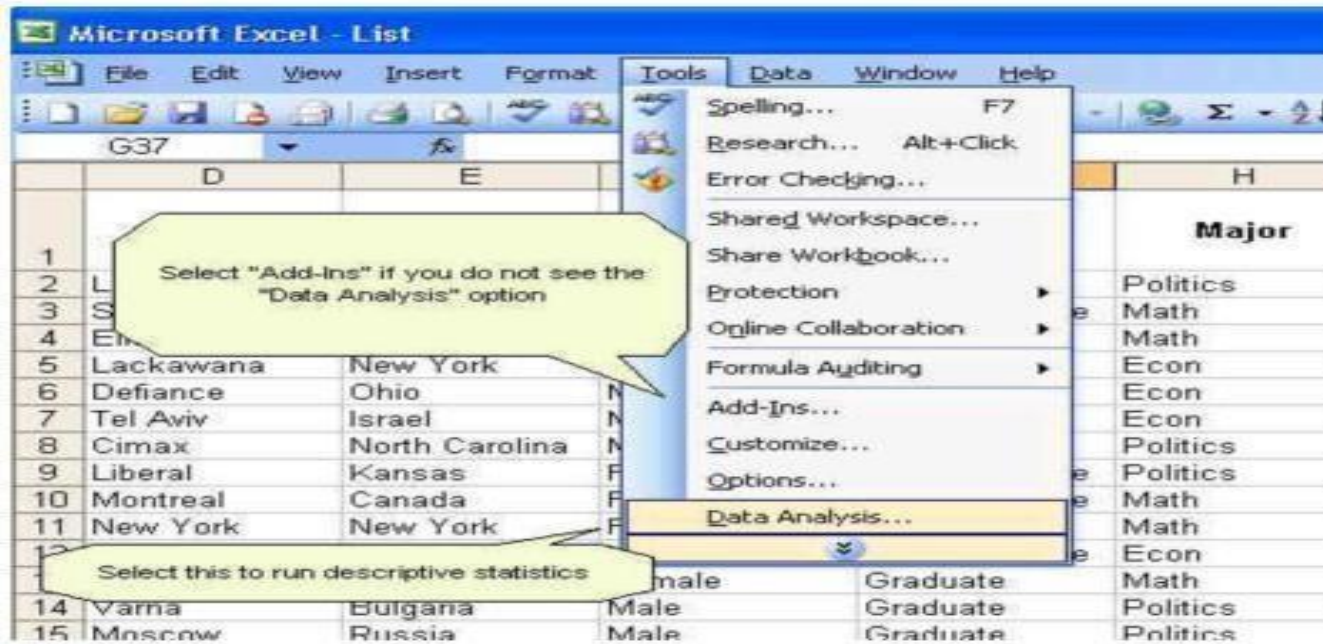
Leptokurtic

Mesokurtic

Platykurtic

Descriptive statistics (using excel's data analysis tool)

Let's get some descriptive statistics for this data. In excel go to Tools – Data Analysis. If you do not see “data analysis” option you need to install it, go to Tools– Add-Ins, a window will pop-up and check the “Analysis ToolPack” option, then press OK. Try running data analysis again.



Descriptive statistics

Microsoft Excel - List															
File Edit View Insert Format Tools Data Window Help															
O1 Age															
J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
Age	SAT	Average score (grade)	Height (in)	Newspaper readership (times/wk)	Age	SAT	Average score (grade)	Height (in)	Newspaper readership (times/wk)	Age	SAT	Average score (grade)	Height (in)	Newspaper readership (times/wk)	
30	2263	67	61	5	Mean	25.2	Mean	1848.9	Mean	80.40091	Mean	66.43333	Mean	4.866667	
19	2006	63	64	7	Standard Error	1.254326	Standard Error	50.22638	Standard Error	1.845084	Standard Error	0.850535	Standard Error	0.23358	
26	2221	78	73	6	Median	23	Median	1817	Median	79.74968	Median	66.5	Median	5	
33	1716	78	68	3	Mode	19	Mode	#N/A	Mode	67	Mode	68	Mode	5	
37	1701	65	71	6	Standard Deviation	6.870226	Standard Deviation	275.1122	Standard Deviation	10.10594	Standard Deviation	4.658573	Standard Deviation	1.279368	
25	1786	69	67	5	Sample Variance	47.2	Sample Variance	75686.71	Sample Variance	102.1301	Sample Variance	21.7023	Sample Variance	1.636782	
39	1577	96	70	5	Kurtosis	-1.04975	Kurtosis	-0.84663	Kurtosis	-0.99191	Kurtosis	-1.06683	Kurtosis	-0.97241	
21	1842	87	62	5	Skewness	0.557191	Skewness	0.155668	Skewness	-0.11236	Skewness	0.171893	Skewness	-0.05191	
18	1813	91	62	6	Range	21	Range	971	Range	32.88251	Range	16	Range	4	
33	2041	71	66	5	Minimum	18	Minimum	1338	Minimum	63	Minimum	59	Minimum	3	
18	1787	82	67	3	Maximum	39	Maximum	2309	Maximum	95.88251	Maximum	75	Maximum	7	
38	1513	79	59	5	Sum	756	Sum	55467	Sum	2412.027	Sum	1993	Sum	146	
30	1637	79	63	4	Count	30	Count	30	Count	30	Count	30	Count	30	
30	1512	70	75	6	<div>These are the descriptive statistics for the labels you selected</div>										
21	1338	82	64	5											
18	1821	80	63	3											
19	1494	75	60	3											
31	2248	95	59	4											
18	2252	92	68	5											
33	1923	95	63	7											
19	1727	67	62	7											
21	1872	82	73	4											
25	1767	89	68	6											
18	1643	79	65	6											
19	1919	88	64	4											
26	1434	96	71	4											
20	2119	88	71	5											
20	2309	64	68	6											
30	2279	85	72	3											
19	1907	79	74	3											

These are the descriptive statistics for the labels you selected

Now we know something about our data

Data analysis using Graphs

Tables, charts and graphs are convenient ways to clearly show your data.

Sample data

The cafeteria wanted to collect data on how much milk was sold in 1 week. The table below shows the results. We are going to take this data and display it in 3 different types of graphs.

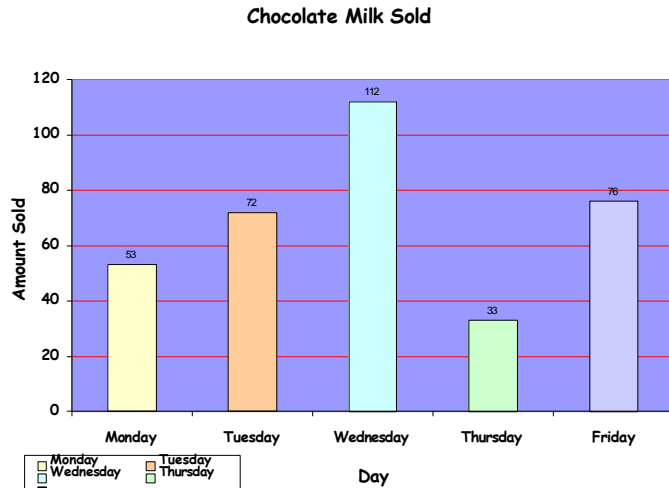
Day	Chocolate	Strawberry	White
Monday	53	78	126
Tuesday	72	97	87
Wednesday	112	73	86
Thursday	33	78	143
Friday	76	47	162

- ✓ Notice how each of the following examples are used to illustrate the data.
- ✓ Choose the best graph form to express your results.

Graphical Representation of variables

Bar Graph

- A bar graph is used to show relationships between groups.
- The two items being compared do not need to affect each other.
- It's a fast way to show big differences. Notice how easy it is to read a bar graph.

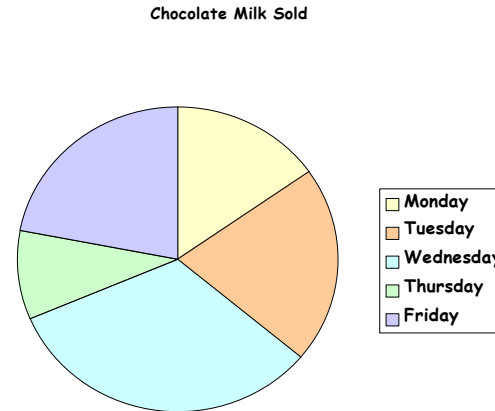


On what day did they sell the most chocolate milk?

a. Tuesday b. Friday c. Wednesday

Pie Graph

- A circle graph is used to show how a part of something relates to the whole.
- This kind of graph is needed to show percentages effectively.

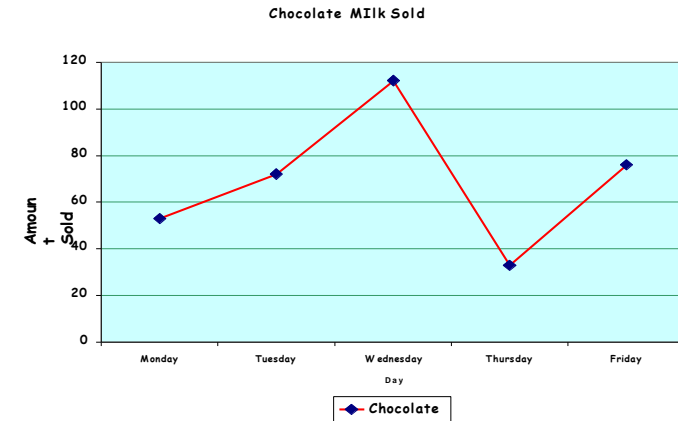


On what day was the least amount of chocolate milk sold?

a. Monday b. Tuesday c. Thursday

Line Graph

- A line graph is used to show continuing data; how one thing is affected by another.
- To see how things are going by the rises and falls a line graph.



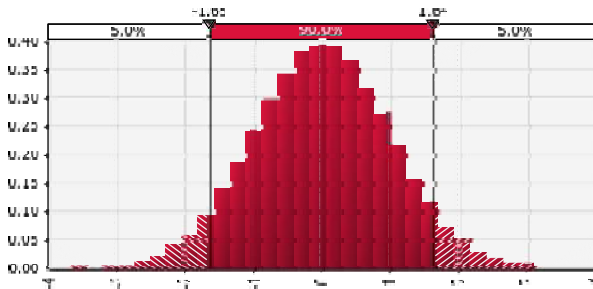
On what day did they have a drop in chocolate milk sales?

a. Thursday b. Tuesday c. Monday

Graphical Representation of variables

Histogram

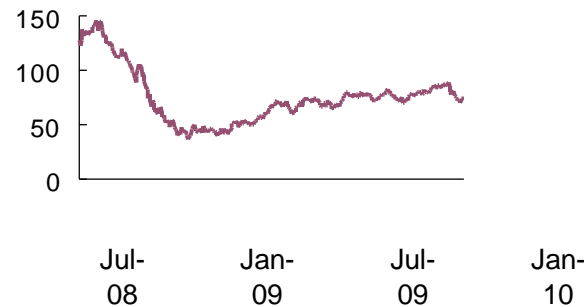
- A **histogram** is a special kind of bar chart which allows us to visualize the distribution of values of an ordinal/continuous variable
- Can be developed in Excel 2007 through Data>>>data analysis>>>histogram



Line charts

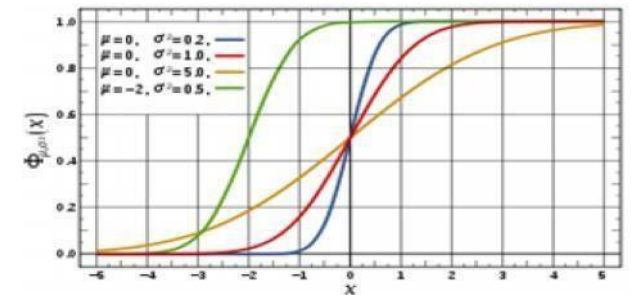
- A representation of data varying over time, eg. commodity prices
- It provides insights like trend of the data, seasonality or presence of outliers

Brent – 1 month forwards
\$/barrel



Ogives

- In statistics, an ogive is a graph showing the curve of a cumulative distribution function.
- It provides insights like distribution of population within a given range



1 Footnote

SOURCE: Wikipedia

Choosing the Right Graph

- Use a bar graph if you are not looking for trends (or patterns) over time; and the items (or categories) are not parts of a whole.
- Use a pie chart if you need to compare different parts of a whole, there is no time involved and there are not too many items (or categories).
- Use a line graph if you need to see how a quantity has changed over time. Line graphs enable us to find trends (or patterns) over time.

Common Chart Types

Chart Type	Typical Applications	Variants, Remarks
Area	Cumulated totals (numbers or percentages) over time	Percentage, Cumulative
Column/Bar	Observations over time or under different conditions; data sets must be small	Vertical (columns), horizontal (bars); multiple columns/bars, columns/bars centered at zero
Histogram	Discrete frequency distribution	Columns/bars without gaps
Line, Curve	Trends, functional relations	Data point connected by lines or higher order curves
Pie	Proportional relationships at a point in time	Segments may be pulled out of the the pie for emphasis (exploded pie chart)
Scatterplot	Distribution of data points along one or two dimensions	One-dimensional, two-dimensional
Map	Typically used for geographical data; can also be used for parts of devices, human or animal bodies	Useful, if an analog relation can be used for representing data

Outliers

- An **outlier** is an observation that is numerically distant from the rest of the data.
- An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.
- Outliers can occur by chance in any distribution, but they are often indicative *either* of measurement error or that the population has a heavy-tailed distribution.

Bill Gates makes \$500 million a year. He's in a room with 9 teachers, 4 of whom make \$40k, 3 make \$45k, and 2 make \$55k a year. What is the **mean** salary of everyone in the room? What would be the **mean** salary if Gates wasn't included?

Mean With Gates:

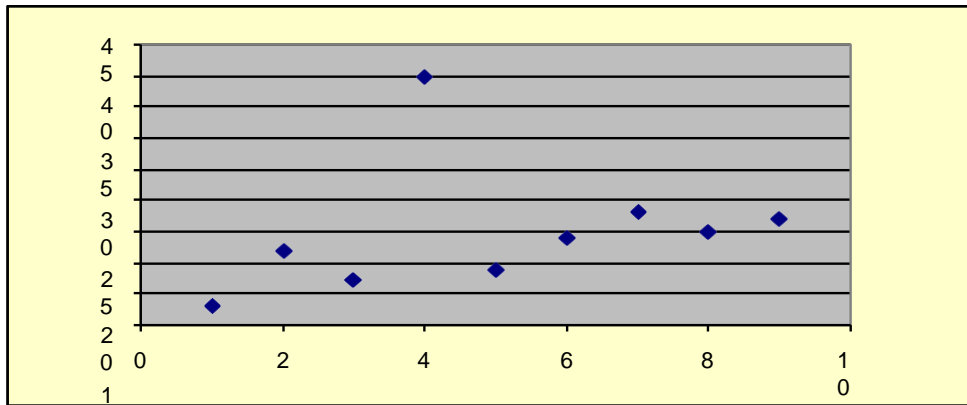
\$50,040,500

Mean Without Gates:

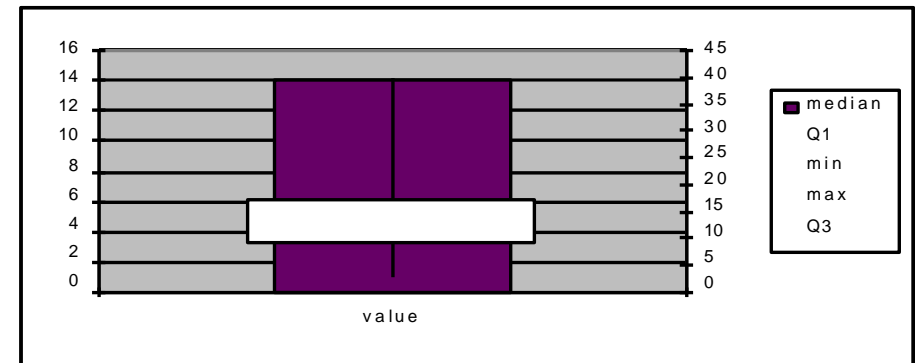
\$45,000

Plots for analyzing outliers

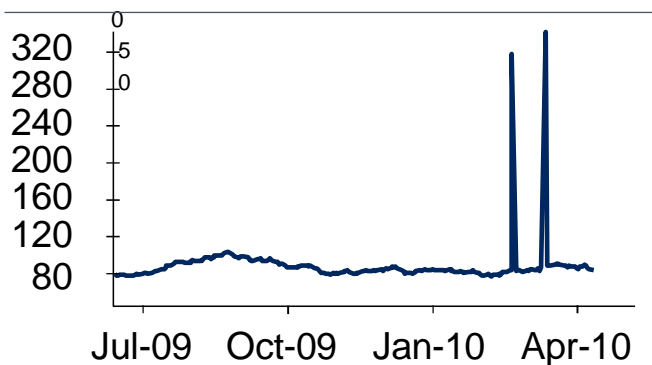
A **Scatterplot** is useful for "eyeballing" the presence of **outliers**.



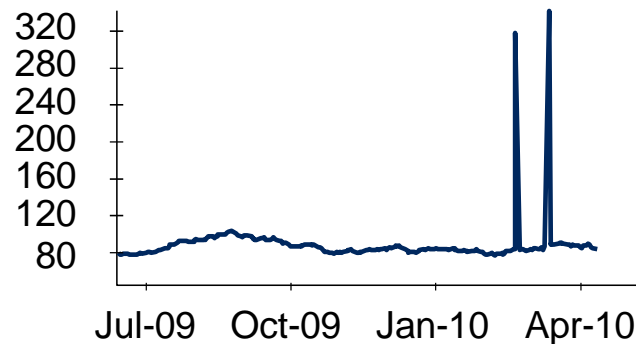
In a Box plot, a point beyond an inner fence on either side is considered a **mild outlier**. A point beyond an outer fence is considered an **extreme outlier**.



Stock Price of Peach Inc.



Hourly power prices



Missing Values

- In the ideal data collection project, complete data would exist for all variables across all experimental units (also called subjects, cases, or observations).
- Unfortunately, for a number of reasons it is inevitable that some values won't be collected, will become lost, or will be unusable.

There are a number of reasons why data become missing.

- sensor failures
- omitted entries in databases
- non-response in questionnaires.
- loss to follow up
- lack of overlap between linked data sets
- dropping out of school, graduation, etc.
- survey design: “skip patterns” between respondents

Imputation Methods

Some common imputation methods

- Mean (median, mode) imputation
- Pairwise deletion a.k.a. available case analysis
- Dummy variable adjustment
- List wise deletion a.k.a. complete case analysis
- Multiple imputation (MI)

Some facts about missing data

Why not just delete cases with missing values rather than impute values at all?

- a. Deletion can introduce substantial bias into the study. And, the loss in sample size can appreciably diminish the statistical power of the analysis.
- b. As a rule of thumb, if a variable has more than 5% missing values, cases are not deleted.

Should I use original data or imputed data when reporting results?

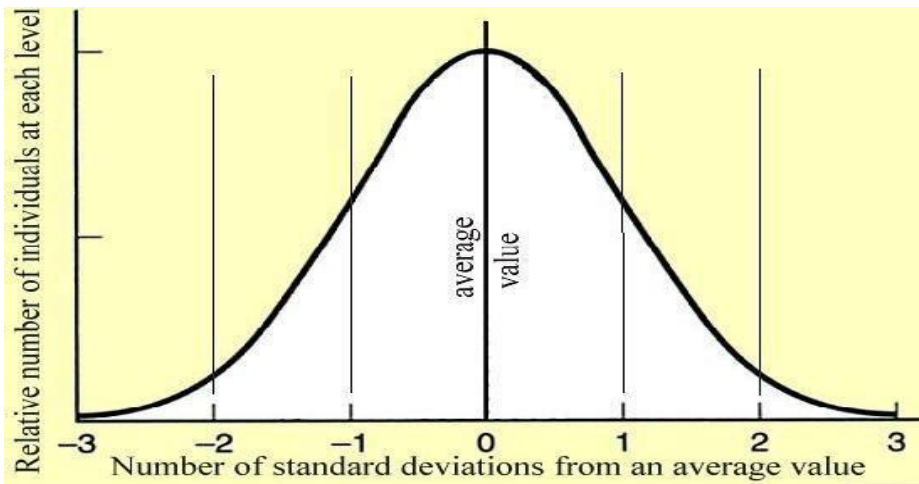
- a. The original dataset may be biased by a large number of non-random missing values.
- b. The imputed dataset is a "what-if" hypothetical dataset which relies on estimation, though it is a "best guess" attempt to present what choices respondents are likely to have made, given their responses on other items.
- c. It is preferable to run all analyses on both the original and imputed datasets, and discuss in the report where imputation would make a difference for the substantive interpretations.

Normal Distributions

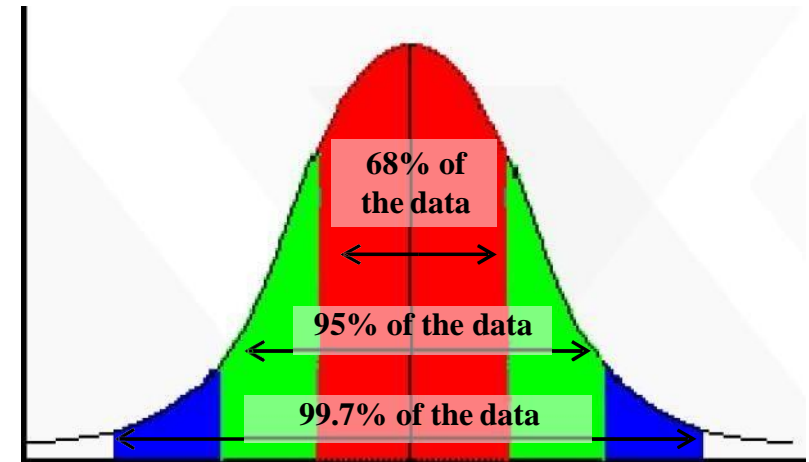
- The normal distribution is a pattern for the distribution of a set of data which follows a bell shaped curve. This also called the Gaussian distribution
- **Normal Distribution** has the mean, the median, and the mode all coinciding at its peak and with frequencies gradually decreasing at both ends of the curve.
- The *normal distribution* is a theoretical ideal distribution. Real-life empirical distributions never match this model perfectly. However, many things in life do approximate the normal distribution, and are said to be “normally distributed.”

The Bell Shaped Curve

- The bell shaped curve has the following characteristics:
 - The curve is concentrated in the center and decreases on either side.
 - The bell shaped curve is symmetric and Unimodal
 - The curve extends to $+$ / $-$ infinity
 - Area under the curve = 1



68-95-99.7 Rule



The empirical rule states that for a normal distribution:

- 68% of the data will fall within 1 SD of mean
- 95% of the data will fall within 2 SD's of the mean
- Almost all (99.7%) of the data will fall within 3 SD's of the mean

Are my data “normal”?

- Not all continuous random variables are normally distributed!!
- It is important to evaluate how well the data are approximated by a normal distribution

Are my data normally distributed?

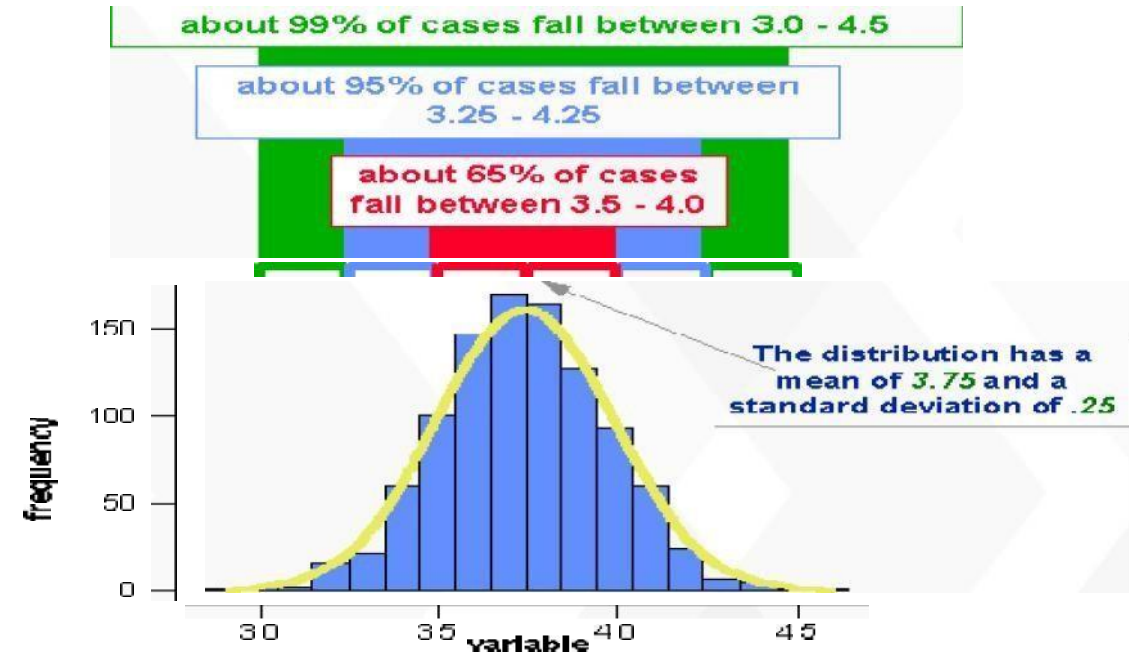
1. Look at the histogram! Does it appear bell shaped?
2. Compute descriptive summary measures—are mean, median, and mode similar?
3. Do 2/3 of observations lie within 1 std dev of the mean? Do 95% of observations lie within 2 std dev of the mean?
4. Look at a normal probability plot—is it approximately linear?
5. Run tests of normality (such as Kolmogorov-Smirnov). But, be cautious, highly influenced by sample size!

Standard (Z) Scores – standard normal variable

- A **standard score** (also called **Z score**) is the number of standard deviations that a given raw score is **above** or **below** the mean.

$$Z = \frac{X - \mu}{\sigma}$$

All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation:



How good is rule for real data?

Check some example data:

The mean of the weight of the women = 127.8

The standard deviation (SD) = 15.5

Practice problem

If birth weights in a population are normally distributed with a mean of 109 oz and a standard deviation of 13 oz

- a. What is the chance of obtaining a birth weight of 141 oz *or heavier* when sampling birth records at random?
- b. What is the chance of obtaining a birth weight of 120 *or lighter*?

Answer

- a. What is the chance of obtaining a birth weight of 141 oz *or heavier* when sampling birth records at random?

$$Z = \frac{141 - 109}{13} = 2.46$$

From the chart or SAS → Z of 2.46 corresponds to a right tail (greater than) area of:

$$P(Z \geq 2.46) = 1 - (.9931) = .0069 \text{ or } .69 \%$$

- b. What is the chance of obtaining a birth weight of 120 *or lighter*?

$$Z = \frac{120 - 109}{13} = .85$$

From the chart → Z of .85 corresponds to a left tail area of:

$$P(Z \leq .85) = .8023 = 80.23\%$$

Applications of Normal Distribution to Business Administration

- Modern portfolio theory assumes that the returns of diversified asset portfolio follow a normal distribution.
- In operations management, process variations often are normally distributed
- In human resource management, employee performance sometime is considered to be normally distributed.

Correlation

What is the relationship between two variables?

Relationship between hours studying (X) and grades on a midterm (Y)?

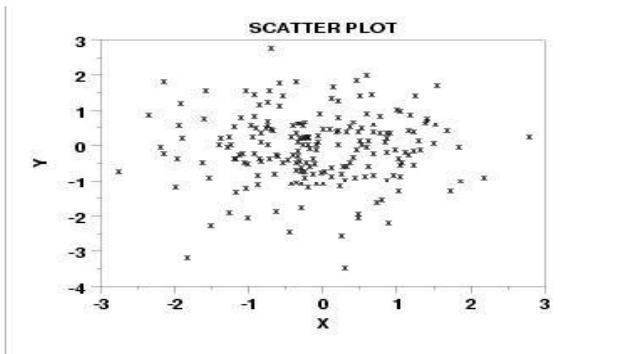
Relationship between self-esteem (X) and depression (Y)?

The relationship between two variables over a period, especially one that shows a close match between the variables' movements

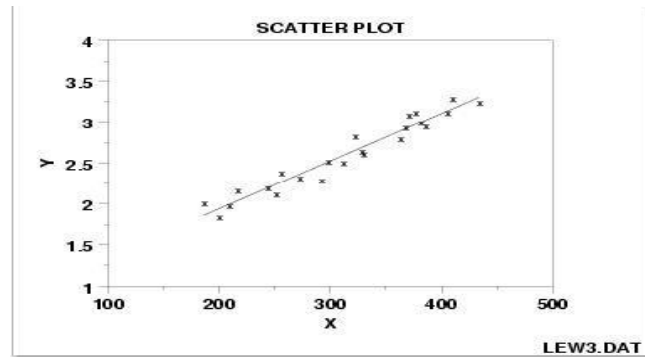
Direction and strength of relationship between two variables

Graphical representation of data in a bivariate setup

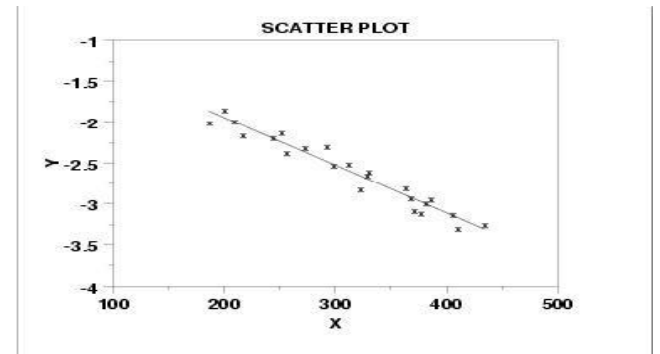
No association



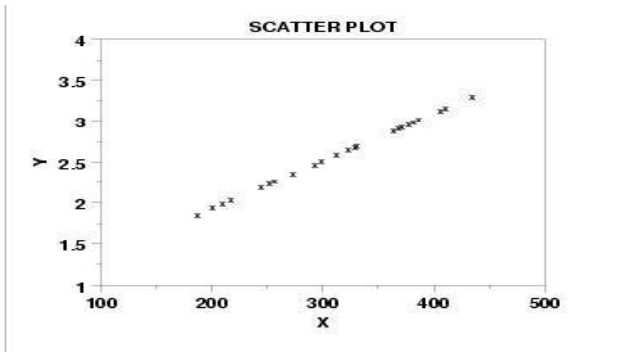
Strong linear relationship



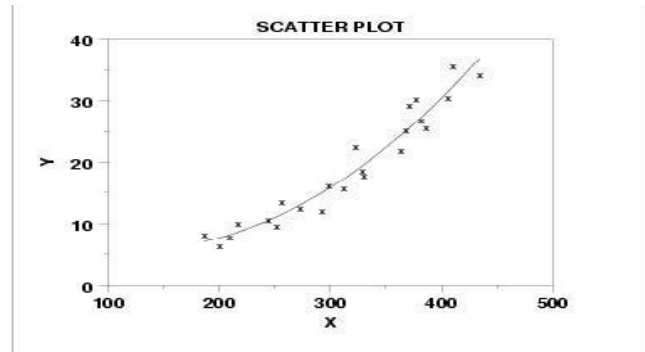
Strong linear relationship



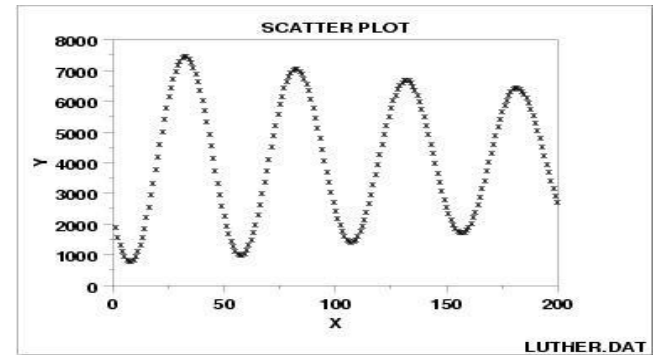
Exact linear relationship



Quadratic relationship

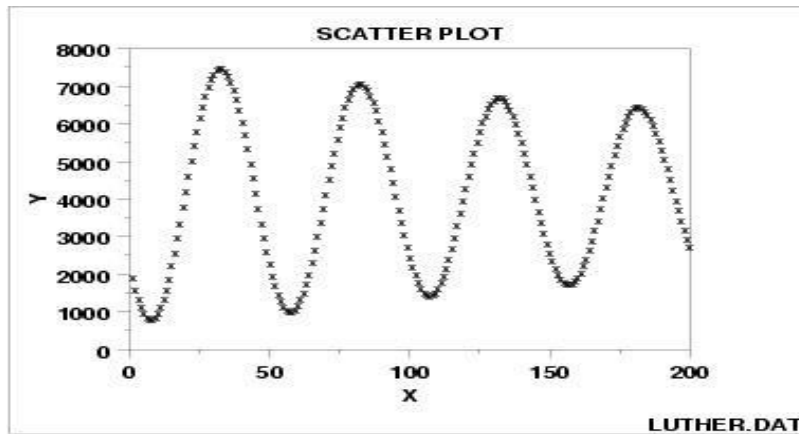


Sinusoidal relationship (damped)



Correlation measures may be misleading in certain scenarios

Correlation and independence



No correlations: **Does not imply no association**

Spurious correlation

A **spurious relationship** is a mathematical relationship in which two events or variables have no direct causal connection, yet it may be wrongly inferred that they do, due to either coincidence or the presence of a certain third, unseen factor (referred to as a "confounding factor" or "lurking variable")

Another popular example is a series of Dutch statistics showing a positive correlation between the number of storks nesting in a series of springs and the number of human babies born at that time. Of course there was no causal connection; they were correlated with each other only because they were correlated with the weather nine months before the observations



Examples

- Increase in height results in weight increase for children
- Attending lessons leads to improved grades
- Age of the car impact its stopping distances
- More the years of education higher the income

Business Examples

- Rising unemployment leads to a decrease in sales of taste the difference products
- Increase in demand of a product leads to increase in supply
- More efficient the workers higher the productivity