



# Statistics

Dr. Jyoti Thanvi

BeingDatum.com  
contact@beingdatum.com



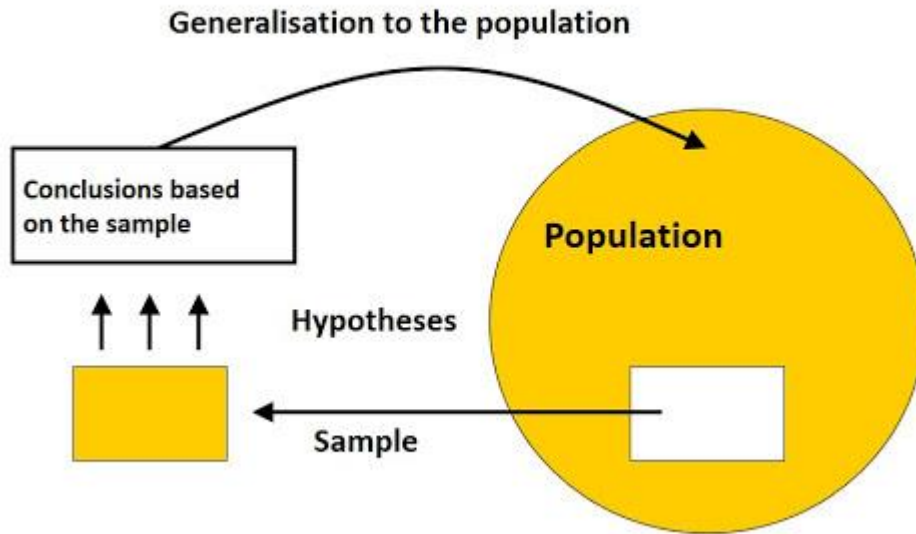
# Obvious Questions

- Why Statistics
- What is Statistics

# Before answering these questions, Let us understand “Inference”.

(a conclusion reached on the basis of evidence and reasoning.)

## The idea of statistical inference



- Deductive(is the process of reasoning from one or more statements to reach a logically certain conclusion)(We go from the general — the theory — to the specific — the observations)( General to particular)
- Inductive(Inductive reasoning makes broad generalizations from specific observations)(We make many observations, discern a pattern, make a generalization, and infer an explanation or a theory) ( Particular to general)



# Inductive Inference and Role of Statistics

When we want to make conclusions about the larger group on the basis of a sample the role of statistics comes into the picture. As “Statistics is the discipline that concerns the collection, compilation, analysis and interpretation of data.”

# Data & Types of Data

- The word *data* is the Latin plural of the word *datum*, which itself is the past participle of the verb *dare*(DAH-reh), meaning “to give.” So, it literally means “things given.” The lexicon meaning of data is facts or information.
- Some data are readily available as “things given,” whereas some data need to be diligently discovered and collected with ethical considerations, depending upon the research problem, need, and the researcher.

## Types of Data

The universe is filled with a huge amount of data, so it needs to be categorized broadly or the systematic conduct of research and synthesis of research outcomes. All of the available data may be classified broadly into **primary** and **secondary** data, and each of these in turn may be categorized into **quantitative** and **qualitative** data. Primary data are collected directly from the field by observing, interviewing, or administering a questionnaire. Secondary data are collected from already available sources. Data that cannot be measured by assigning a value or by ordering them in ascending or descending order are generally considered qualitative data, and data that can be subjected to some kind of quantification or measurement are generally considered quantitative data.

**Primary:** Field observations, Age, income, Educational level.

**Secondary:** Letters, diaries Census, annual Reports.

# Data Collection Methods

Researchers often employ a specific method or several data collection methods to collect data, such as observation, case study, questionnaire, interview, focus groups, rapid rural appraisal, and secondary data. Some of these methods overlap with others, and some are more popular than the others. Many of these data collection methods have different variations within them. It is crucial to note a few points on them.

**First**, researchers need to carefully select a method or a combination of data collection methods in such a way that they capture reality appropriately and accurately in order to answer the research questions and achieve the research objectives. Inappropriate or incorrect selection of data collection methods results in incorrect and misleading outcomes that distort the reality.

**Second**, after having selected the most appropriate data collection method(s), researchers need to develop adequate knowledge and

**Third**, it is important to be aware of the strengths and limitations of various data collection methods and where and when they can be best used.

**Fourth**, we should be aware of and effectively use in moderation data collection means with which we are all gifted. These are our five sense perceptions: eyes (seeing/observing), ears (hearing/listening), nose (smell), tongue (taste), and skin (touch).

# Data Collection Methods

Finally, while collecting data through the chosen method(s), researchers should ponder the following questions to keep the data collection process on track.

- What am I trying to discover?
- Why have I chosen the methods (research, sampling, data collecting) I have chosen?
- Do these methods help or hinder my efforts toward understanding reality?
- Are there any alternative methods to understand the phenomenon I am trying to understand?
- Do these categories of methods make any sense in understanding the reality?

# Ethical Considerations

Researchers need to collect data according to the set ethical standards, which are often based on certain values and principles: honesty, truthfulness, privacy and confidentiality, self-determination and voluntary involvement, zero physical and psychological harm, dignity and worth of human beings, accountability, right to know on the part of respondents, fairness and impartiality on the part of researchers, and informed consent. On the other hand, researchers should avoid breach of confidence and agreements, absence of informed consent or self-determination/autonomy of respondents, deception, risk of harm or offense, acts involving conflict of interest, and any unethical act.



# Impediments in Data Collection

Data collection is a planned, purposeful, and systematic activity. Despite choosing appropriate data collection methods; meticulously developing data collection instruments; planning adequate resources, including time; and meeting ethical standards, researchers may encounter several impediments in the data collection process. One probable reason for these impediments is

that the nature of the setting, the research problem, the researcher, the researched, the time of research, and the prevailing social conditions vary every time. Thus, the data collection impediments may be analyzed by looking at three “R” factors: the researcher; the research problem; and the researched, or a combination of these factors.

# Some important terms

- Population
- Census Survey
- Sample and Sample Survey
- Need of Sampling
- Various Sampling Techniques
- Design of Experiment

# Types of Data Received

- Categorical/ Nominal/Ordinal/Likert Scale
- discrete
- continuous

# Types of Scales

- Nominal- objects or people are categorized according to some criterion (gender, job category)
- Ordinal- Categories which are ranked according to characteristics (income- low, moderate, high)
- Interval- contain equal distance between units of measure- but no zero (calendar years, temperature)
- Ratio- has an absolute zero and consistent intervals (distance, weight)

# Types of Data

## Numerical (quantitative)

They are numerical values, sensible to add, subtract, take averages, etc.

1. **Continuous.** A number within a range of values, usually measured, such as height (within the range of human heights).
2. **Discrete.** Only take certain values (can't be decimal), usually counted, such as the count of students in a class

## Categorical (qualitative)

Take a number of distinct categories, but it wouldn't be sensible to do arithmetic operations. Sometimes we encode the categories to numerical values, act as place holders for the levels of the category.

1. **Ordinal.** Values that have inherent ordering levels, such as (high, medium, low).

2. **Categorical(Nominal).** If they don't have inherent ordering, such as gender

# Categorical

## Nominal

- Values represent discrete units
- Changing the order of units does not change their value

Like,

Gender: Male/Female

Eye Color

## Ordinal

- Values represent discrete units
- Let say, economic status: high/medium/low. These can be ordered as low, medium, high or vice versa.
- Let say, college level: primary, secondary, tertiary, as we can order these values, these are considered as ordinal variables

# Numerical

An numerical variable is similar to an ordinal variable, except that the intervals between the values of the numerical variable are equally spaced.

For example, suppose you have a variable such as annual income that is measured in dollars, and we have three people who make \$10,000, \$15,000 and \$20,000. The second person makes \$5,000 more than the first person and \$5,000 less than the third person, and the size of these intervals is the same. If there were two other people who make \$90,000 and \$95,000, the size of that interval between these two people is also the same (\$5,000).

# Reporting Count Data

- Frequency Tables
- Proportion/ Percentages
- Diagrams
- Cross Tabs
- Chi-Squares



# Summary Statistics

In descriptive statistics, **summary statistics** are used to summarize a set of observation, in order to communicate the largest amount of information as simply as possible.

Statisticians commonly try to describe the observations in a measure of location, or central tendency, such as the arithmetic mean a measure of statistical dispersion like the standard deviation a measure of the shape of the distribution like skewness or kurtosis.

- Average/Central Value
- Spread/ Variability
- Shape
- Order Statistics and Five Number Summary
- (The sample minimum (smallest observation), the lower quartile or *first quartile*, the median (the middle value), the upper quartile or *third quartile*, the sample maximum (largest observation))

# Average/Central Value/ Measures of Central Tendency

Measures of central tendency are measures of the location of the center or middle of a distribution. However, the definition of “center” or “middle” is deliberately left broad, such that the term *central tendency* can refer to a wide variety of measures. The three most common measures of central tendency are the mode, the mean, and the median.

## Measures of Central Tendency

**Median** = Middle value in the series of numbers

3   4   6   11   22

If number string has an even number of elements, the median is the arithmetic mean of two middle numbers

3   4   6   8   11   22

Median = 7

Remember, sample median is an estimate of population median

# Mode

- The mode for a collection of data values is the data value that occurs most frequently.
- If two values occur the same number of times and more often than the others, then the data set is said to be bimodal. The data set is multimodal if there are more than two values that occur with the same greatest frequency. The mode is applicable to qualitative as well as quantitative data.

**MEAN**

The "mean" is the "average". To find the mean, you add up all the numbers and then divide by the number of numbers.

TO FIND THE MEAN FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13  
average the set of numbers:

$$(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) \div 9 = 15$$

Note that the mean isn't a value from the original list; This is a common result. DO NOT assume that the mean will be one of the original numbers.

**MEDIAN**

The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in **numerical order**, so you may have sort the list first.

FOR AN ODD NUMBER OF VALUES: 1,5,2,8,7  
**Sort the numbers 1, 2, 5, 7, 8**

FOR AN EVEN NUMBER OF VALUES: 1,5,2,10,8,7  
**Sort the numbers: 1, 2, 5, 7, 8, 10.**

TAKE THE AVERAGE OF THE TWO MEAN NUMBERS:  $(5+7)/2 = 6$

**MODE**

The "mode" is the value that occurs most often. If no number is repeated, then there is no mode for the list.

TO FIND THE MODE FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13  
**Sort the numbers: 13, 13, 13, 13, 14, 14, 16, 18, 21**

©2007 KIDS.COM

# Mean

## *(Arithmetic Mean)*

The *arithmetic mean, or average, is the most common measure of central tendency*. Given a collection of data values, the mean of these data is simply the arithmetic average of these data values. That is, the mean is the sum of observations divided by the number of observations.

The arithmetic mean is not the only “mean” available. Indeed, there is another kind of mean that is called the geometric mean, harmonic mean. However, the arithmetic mean is by far the most commonly used. Consequently, when the term *mean is used, one* can assume that it is the arithmetic mean.

# Median

The median is the midpoint of a distribution such that the same number of scores is above the median as below it. In other words, the median is the 50<sup>th</sup> percentile. More specifically, the median for a collection of data values is the number that is exactly in the middle position of the list when the data are ranked (i.e., arranged in increasing order of magnitude).

# Comparisons of Measures of Central Tendency

To some extent, selection of the most appropriate measure of central tendency is dependent on the scale of measurement of the variable. Specifically, if the data are nominal, then only the mode is appropriate. If the data are ordinal, either the mode or the median may be appropriate. If the data are interval or ratio, the mode, median, or mean may be appropriate. For distributions that are symmetrical and unimodal, the three major measures of central tendency (i.e., mean, median, mode) are all the same. When the distribution is symmetrical and bimodal, the mean and the median coincide, but two modes are present. The less symmetrical the distribution, the greater the differential between the mean, the median, and the mode. For skewed distributions, they can differ markedly. Specifically, in positively skewed distributions, the mean is higher than the median, whereas in negatively skewed distributions, the mean is lower than the median.

Thus,  
comparing the mean and median can provide useful information about the level of skewness inherent in the distribution.

# Comparisons of Measures of Central Tendency.....

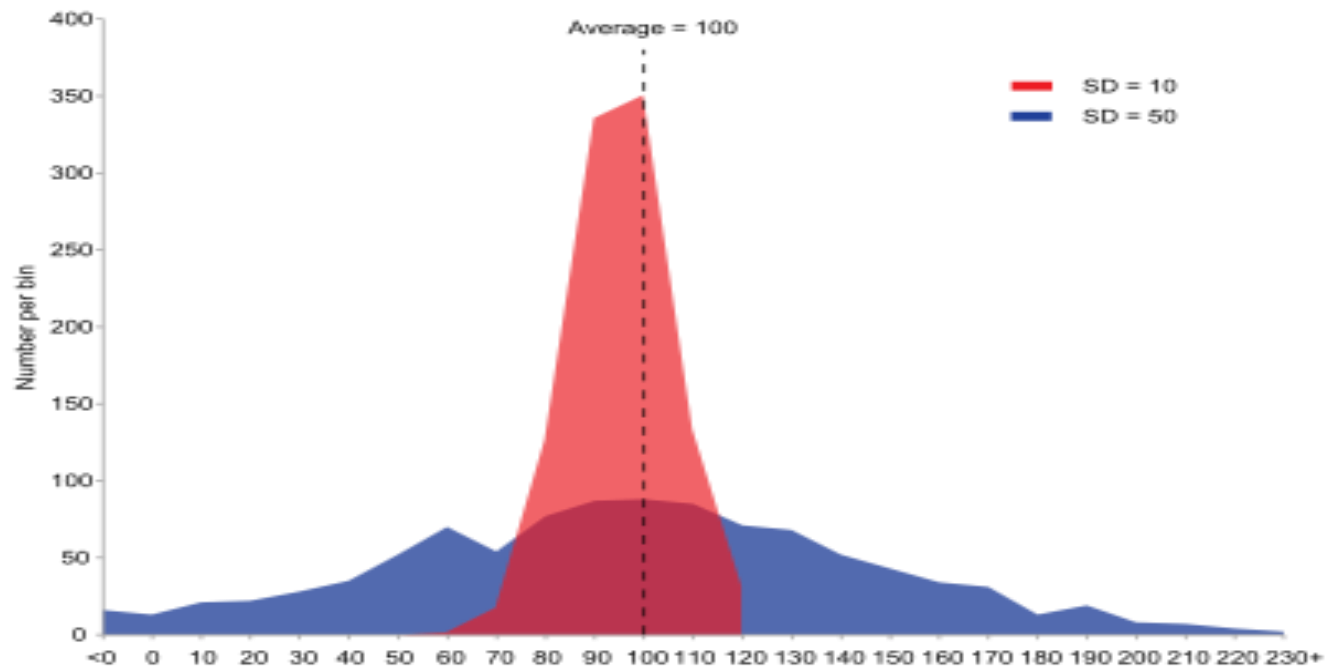
Of the all measures of central tendency discussed, the mean is by far the most widely used because it takes every score into account, is the most efficient measure of central tendency for approximately symmetric (normal) distributions, and uses a simple formula. Also, because the mean requires that the differences between the various levels of the categories on any part of the distribution represent equal differences in the characteristic or trait measured (i.e., equal unit or interval/ratio scale), it can be manipulated mathematically in ways not appropriate to the median and mode.

However, the mean does have several disadvantages. In particular, the mean is sensitive to skewed data. It is also sensitive to outliers. Thus, the mean often is misleading in highly skewed distributions and is less efficient than other measures of central tendency when extreme scores are possible.

# Measure of Dispersion

- Dispersion in statistics is a way of describing how spread out a set of data is.
- A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.
- The spread of a data set can be described by a range of descriptive statistics including range, variance/SD, and interquartile range. Spread can also be shown in graphs: dot plots, box plots.

Example of samples from two populations with the same mean but different dispersion. The blue population is much more dispersed than the red population.



# Range

Among all the measures of variability, the range is the most general and is an overall picture of how much variability there is in a group of scores. It provides an impression of how far apart scores are from one another and is computed by simply subtracting the lowest score in a distribution from the highest score in the distribution.

$$R = L - S$$



# Quartiles

- Quartiles are the values that divide a list of numbers into quarters:
- Put the list of numbers **in order**
- Then cut the list into **four equal parts**
- The Quartiles are at the "cuts"

**Example: 5, 7, 4, 4, 6, 2, 8**

- Put them in order: 2, 4, 4, 5, 6, 7, 8
- Cut the list into quarters:
- And the result is:
- Quartile 1 (Q1) = **4** ( $8 \times 1/4 = 2^{\text{nd}}$  observation)
- Quartile 2 (Q2), which is also the [Median](#), = **5** ( $8 \times 2/4 = 4^{\text{th}}$  observation)
- Quartile 3 (Q3) = **7** ( $8 \times 3/4 = 6^{\text{th}}$  observation)

**IQR = Q3 - Q1 = 7 - 4 = 3**

**Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8 what are Q1, Q2, Q3?**

**( Go and refer Box and Whisker Plots)**

# IQR/ Semi IQR

The **interquartile range (IQR)** is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.

$$\text{IQR} = Q_3 - Q_1.$$

The semi-interquartile range is a measure of, the dispersion or spread of a variable; it is the distance between the 1st quartile and the 3rd quartile, halved. It is common to describe a variable using a measure of central tendency, or average, most commonly the mean or median. However, in order to make sense of a measure of average, we need to have a measure of dispersion. When the mean is used as a measure of average, the standard deviation is usually used as the measure of dispersion. When the median is used, it is more appropriate to use the semi-interquartile range. The semi-interquartile range is preferred over the range because it is not affected by extreme scores. The range is calculated using only two data points: the highest and the lowest. If one of these values were to change, the range would change dramatically.

# Average Deviation

The *average deviation (AD)* is used as a measure of dispersion may be referred to as the *average absolute deviation* or *mean deviation*. The average deviation is often defined in one of two ways: by deviations from the mean ( $AD_M$ ) or by deviations from the median ( $AD_{Md}$ ). The average deviation is calculated by taking the difference between each score and the mean (or median), summing the absolute values of these deviations, and then dividing the sum by the number of deviations. As a measure of dispersion, the larger the AD, the greater is the variability in a distribution of scores.

- It gives equal weight to the deviation of every value from the mean or median.
- The average deviation from the median has the property of being the point at which the sum of the absolute deviations is minimal compared with any other point in the distribution of scores.
- Given that the AD is based on every value in the distribution of scores, it provides a better description of the dispersion than does the range or quartile deviation.
- In comparison with the standard deviation, the AD is less affected by extreme values and easier to understand.

# Average Deviation or Mean Deviation

- Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If  $x_1, x_2, \dots, x_n$  are the set of observation, then the mean deviation of  $x$  about the average  $A$  (mean, median, or mode) is
- Mean deviation from average  $A = 1/n [\sum_i |x_i - A|]$

For a grouped frequency, it is calculated as:

- Mean deviation from average  $A = 1/N [\sum_i f_i |x_i - A|]$ ,  $N = \sum f_i$

Here,  $x_i$  and  $f_i$  are respectively the mid value and the frequency of the  $i^{\text{th}}$  class interval.

Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment.

# Standard Deviation

The standard deviation

(abbreviated as *s* or *SD*) *represents* the average amount of variability in a set of scores as the average distance from the mean. The larger the standard deviation, the larger the average distance each data point is from the mean of the distribution.

- The standard deviation is computed as the average distance from the mean. The larger the standard deviation, the more spread out the values are, and the more different they are from one another.
- Just like the mean, the standard deviation is sensitive to extreme scores.
- If  $SD = 0$ , *there is absolutely no variability in the set of scores*, and they are essentially identical in value. This will rarely happen.

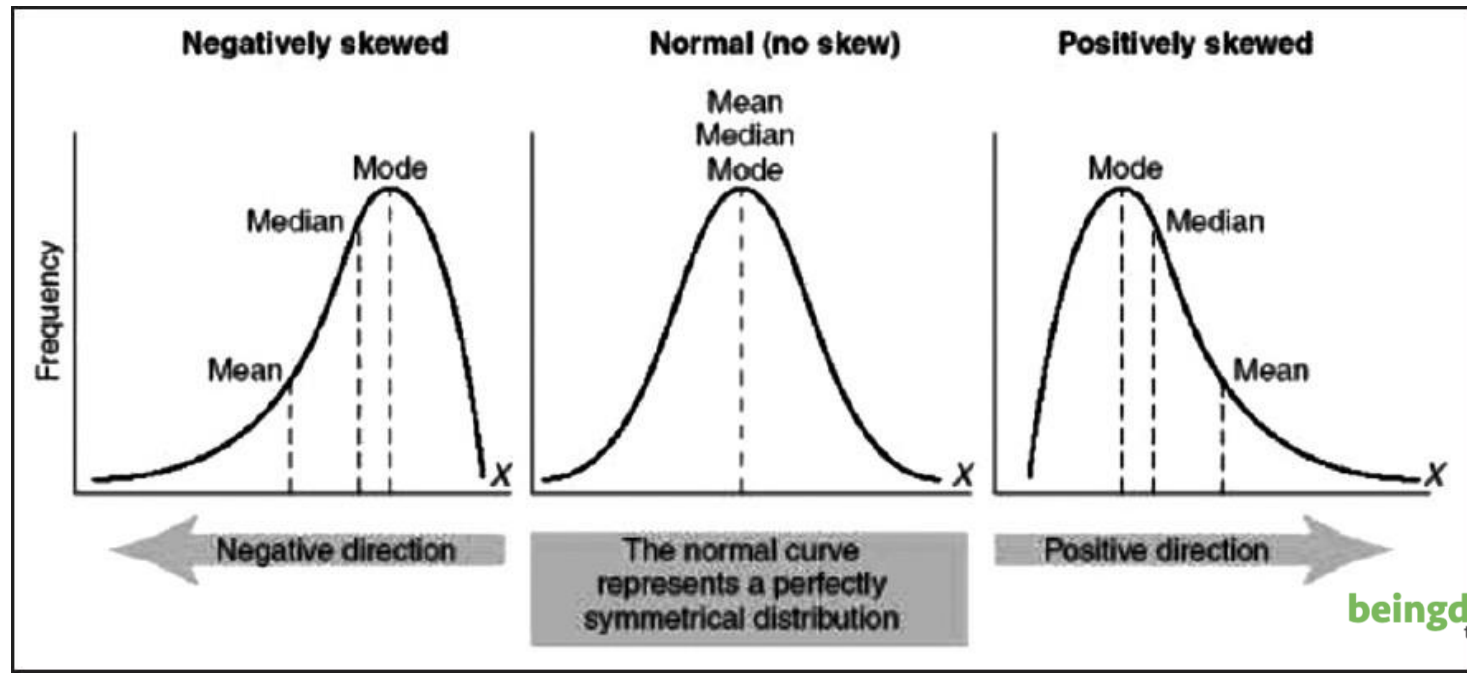
# Standard Deviation

- A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma,  $\sigma$ . It is also referred to as root mean square deviation. The standard deviation is given as
- $\sigma = [(\sum_i (y_i - \bar{y})^2 / n)]^{1/2} = [(\sum_i y_i^2 / n) - \bar{y}^2]^{1/2}$
- For a grouped frequency distribution, it is
- $\sigma = [(\sum_i f_i (y_i - \bar{y})^2 / N)]^{1/2} = [(\sum_i f_i y_i^2 / n) - \bar{y}^2]^{1/2}$
- The square of the standard deviation is the **variance**. It is also a measure of dispersion.

# Skewness

Skewness is a measure of the lack of symmetry, or the lopsidedness, a distribution has. In other words, one tail of the distribution is longer than another. A positively skewed distribution has a longer right tail than left, corresponding to a smaller number of occurrences at the high end of the distribution. This might be the case when you have a test that is very difficult: Few people get scores that are very high, and many more get scores that are relatively low.

A negatively skewed distribution has a shorter right tail than left, corresponding to a larger number of occurrences at the high end of the distribution. This would be the case for an easy test (lots of high scores and relatively few low scores).



# Kurtosis

Kurtosis is the quality of a distribution such that it is flat or peaked.

Kurtosis is commonly thought of as a measure of the “pointiness” of a frequency distribution. This is because kurtosis is the degree to which scores cluster in the tails of a frequency distribution:

A **platykurtic** distribution has many scores in the tails (often called a heavy-tailed distribution) and so is typically quite flat, whereas a leptokurtic distribution is relatively thin in the tails and so looks quite pointy.

Figure 1 shows both leptokurtic and platykurtic distributions.

The **leptokurtic** distribution is pointier than a normal distribution; conversely, the platykurtic distribution is flatter than a normal distribution. Kurtosis is typically measured using a scale that is centered on zero (the value of kurtosis in a normal distribution).

Negative values of kurtosis represent platykurtic distributions, and positive values indicate leptokurtic distributions. If a frequency distribution has positive or negative values of kurtosis, this tells you that this distribution deviates somewhat from a normal distribution.



# Moments

- Moments are popularly used to describe the characteristic of a distribution. They represent a convenient and unifying method for summarizing many of the most commonly used statistical measures such as measures of tendency, variation, skewness and kurtosis. Moments are statistical measures that give certain characteristics of the distribution. Moments can be raw moments, central moments and moments about any arbitrary point.

Three types of moments are:

1. Moments about arbitrary point, ( Raw)
2. Moments about mean, and ( Central)
3. Moments about origin ( Raw)

# Moments Calculation

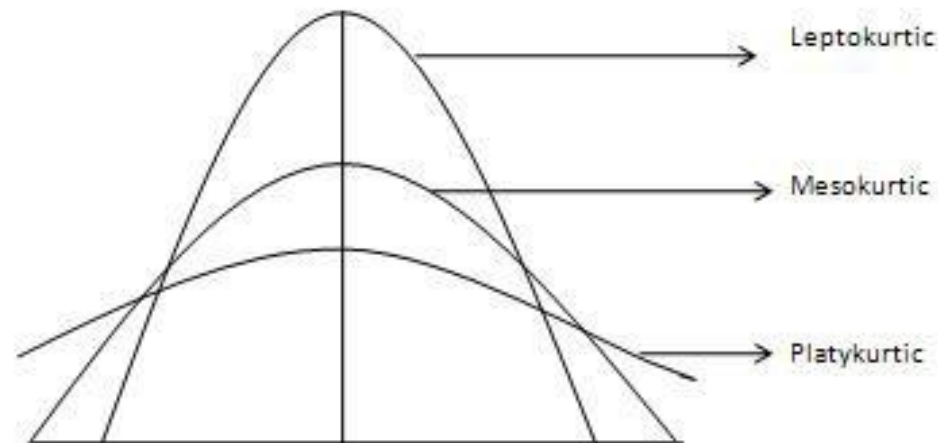
- Rth Raw Moment:  $\mu_r' = 1/N \sum f_i (X_i - A)^r$  (  $A = 0$  means raw moment with respect to origin)
- Rth Central Moment:  $\mu_r = 1/N \sum f_i (X_i - AM)^r$

Look at the formulas can we say that the first raw moment gives mean and the second central moment gives variance?

- Coefficient of skewness based upon moments is given by  $\beta_1 = (\mu_3)^2 / (\mu_2)^3$
- 0/positive/negative
- Coefficient of kurtosis based upon moments is given by  $\beta_2 = (\mu_4) / (\mu_2)^2$
- 3/less than 3/ more than 3

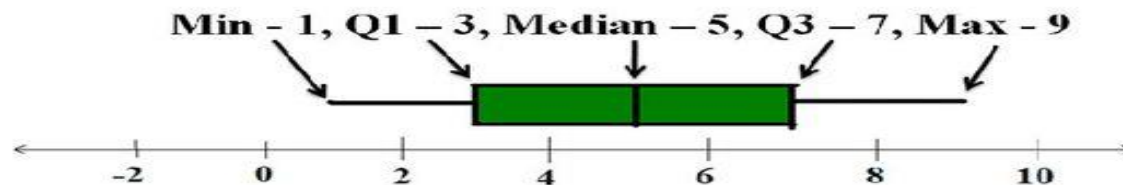
Finally, it should be noted that there are more precise ways to answer the queries that one can answer by looking at a histogram. By putting gathered scores into any of several existing formulas,

it is possible to provide numerical measures of the average score (e.g., by computing the arithmetic mean or median), the amount of variability (e.g., by computing the standard deviation or interquartile range), and the distribution's shape (e.g., by computing indices of skewness and kurtosis).



# Five Number Summary

The **five-number summary** of a data set consists of the **five numbers** determined by computing the minimum,  $Q_1$ , median,  $Q_3$ , and maximum of the data set.



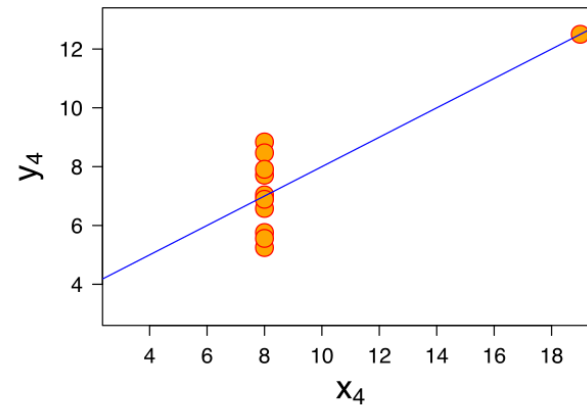
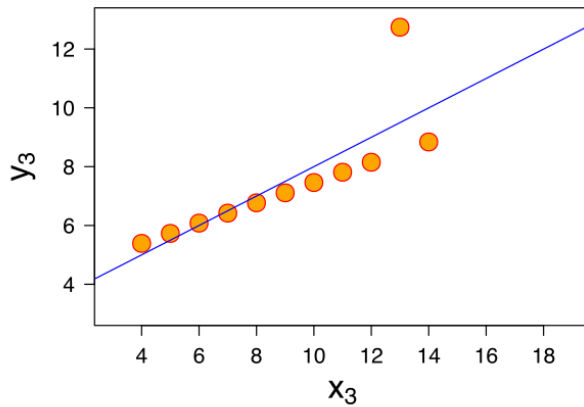
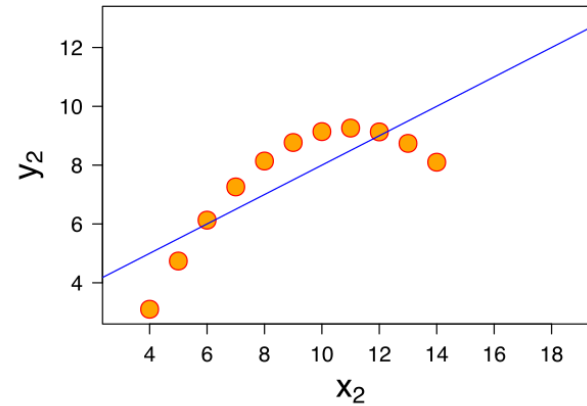
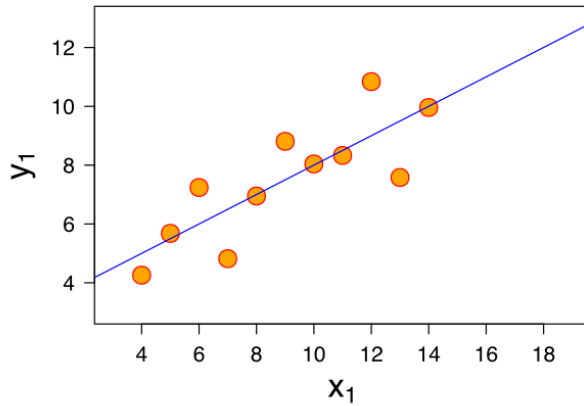
# Limitations of Summary

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

## For all four datasets:

Property	Value	Accuracy
Mean of $x$	9	exact
Sample variance of $x$	11	exact
Mean of $y$	7.50	to 2 decimal places
Sample variance of $y$	4.125	$\pm 0.003$
Correlation between $x$ and $y$	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

# Let Us Look Plots



# Explanations

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one outlier is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.
- The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.



# What Else is Missing







Concept of Bivariate/ Multivariate and cross tabs

# Bivariate Data

In statistics, **bivariate data** is **data** on each of two variables, where each value of one of the variables is paired with a value of the other variable. Typically it would be of interest to investigate the possible association between the two variables( It can be with all different types of variables) .

1. bi


2. examples


  
  



Bivariate Data


2 variable data

	Hours Studied	Test Score
student 1	3	90
student 2	1	86
student 3	5	84
student 4	4	92
student 5	3	91
student 6	5	100
student 7	0	76
student 8	1	82
student 9	2	85

bicycle 

binoculars 

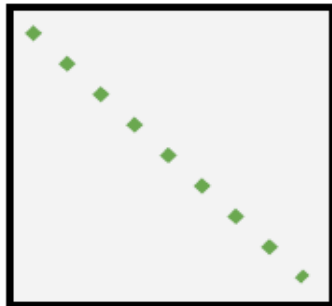
bilingual 

bivariate 

# Covariance

Covariance measures the directional relationship between the two variables. A positive covariance means that they move together while a negative covariance means they move inversely.

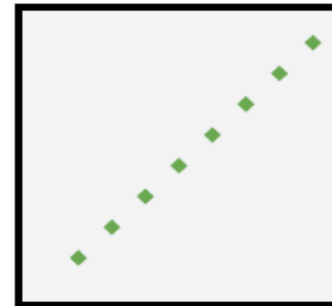
## COVARIANCE



Large Negative  
Covariance



Nearly Zero  
Covariance

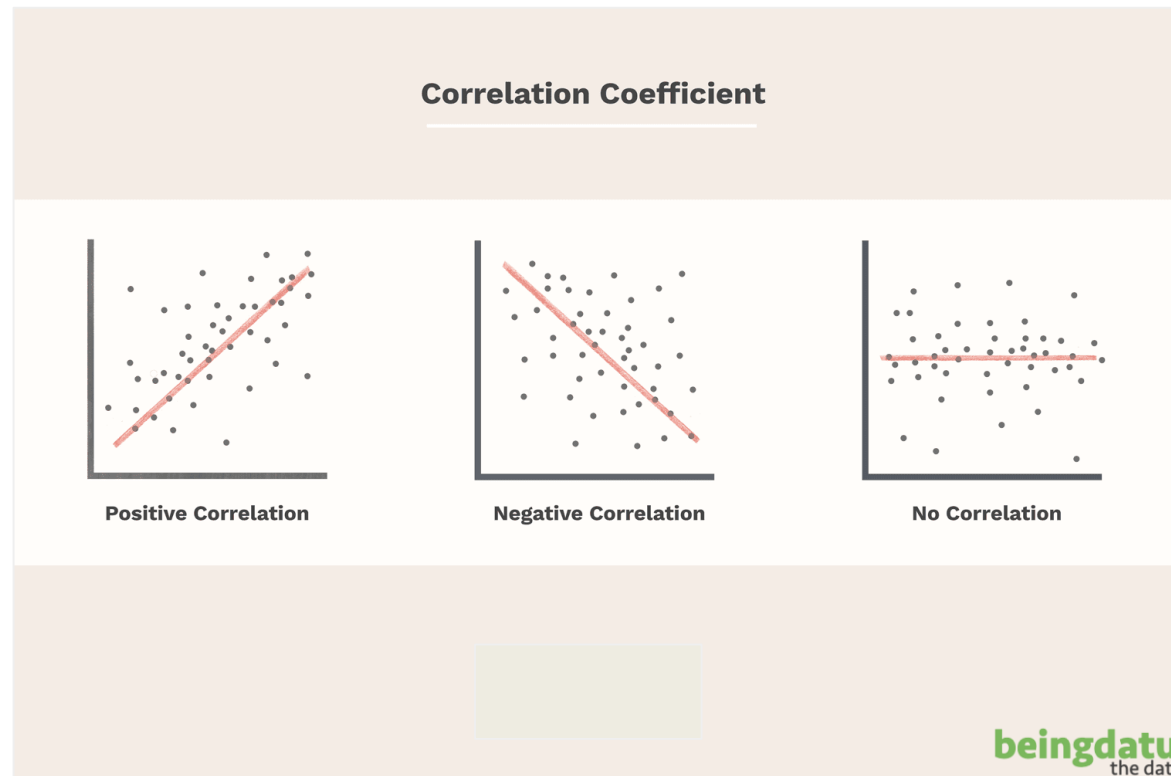


Large Positive  
Covariance

# Pearson Product-moment Correlation Coefficient

Pearson's product-moment correlation coefficient is by far the most common index of the relationship between two variables, or bivariate relationship. Pearson's product-moment correlation coefficient measures the degree to which the points in the scatter plot tend to cluster about a straight line. In other words, the product-moment correlation coefficient *measures the degree of linear relationship between two variables*. If we label the two variables of interest as  $x$  and  $y$ , then Pearson's product-moment correlation coefficient, denoted by  $r$ , is given by the following formula:

$$r = \text{Cov}(X, Y) / SD(X) \cdot SD(Y)$$

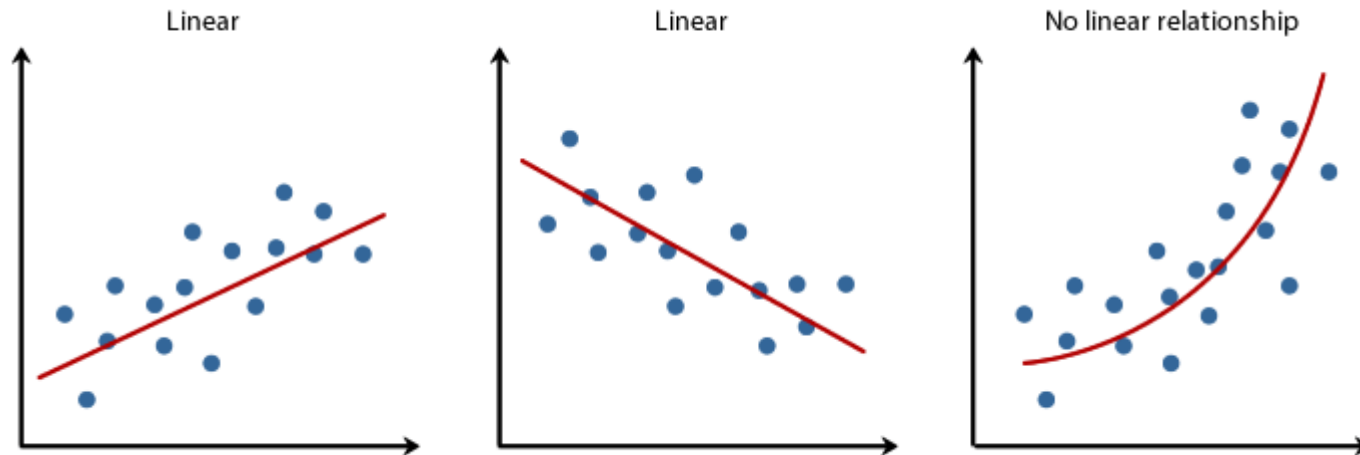


# Assumptions

A key assumption pertaining to Pearson  $r$  is *that* the relationship between the two underlying variables is linear. An effective way of assessing the linearity of relationships is via bivariate scatter plots.

# Regression Analysis

Regression analysis is the name for a family of techniques that attempts to predict one variable (an outcome or dependent variable) from another variable, or set of variables (the predictor or independent variables).



Copyright 2014. Laerd Statistics.

# Causation

- “[It] indicates that one event is the result of the occurrence of the other event; i.e., there is a causal relationship between the two events. This is also referred to as cause and effect.”
- In other words, does one variable *actually* impact the other?
- Causality vs. correlation is also a topic Michael Molnar examined [in a recent article for Forbes](#). Molnar warns that: “Confusing correlation with causation is not an unknown issue but it is becoming increasingly problematic as data increases and computers get more powerful... It gets to the heart of what we know - or think we know - about how the world works.”

# Getting it right

- Causality is an area that is frequently misunderstood, and it can be notoriously difficult to infer causation between two variables without doing a randomized controlled experience.
- Furthermore, correlation can be a useful measure. However, it has limitations as it is usually associated with measuring a linear relationship.
- Understanding that correlation does not imply causation, knowing the difference between the two, and being more skeptical before making bold claims, is critical in today's data-driven world.



# EDA

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

**OR**

The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

# What it Does

- That employs a variety of techniques (mostly graphical) to maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop **parsimonious models**; and
- determine optimal factor settings.

“Long before worrying about how to convince others, you first have to understand what’s happening yourself”. (Gelman and Hill, Data Analysis Using Regression and Multilevel Hierarchical Modeling, p.551)

# How it helps formal statistical modeling

- Suggest hypotheses about the cause of observed phenomena.
- Assess assumptions on which statistical inference will be based.
- Support the selection of appropriate statistical tools and techniques

# Probability

- Uncertainty and randomness occur in many aspects of our daily life and having a good knowledge of probability helps us make sense of these uncertainties. Learning about probability helps us make informed judgments on what is likely to happen, based on a pattern of data collected previously or an estimate.
- Classical Definition of Probability and important terms
- Random Experiment/ sample space/ mutually exclusive outcome/ equally likely outcome/ exhaustive outcome
- How to calculate the probability in simple cases?( What are the prerequisite to learn this? Basic set theory and concept of permutations and combinations.)
- Example ( two coins/ three coins/one dice/ two dice)
- Conditional Probability( Concept)
- Conditional Probability( Formula)
- Or  $P(A/B) = P(A \cap B)/P(B)$ , concept independence and how to check ?

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

# Probability.....

Example 1:

Toss three coins what is the probability of getting more than one heads?

Sol'n:  $S = (HHH, HHT, HTH, THH, HTT, THT, TTH, TTT)$  event  $A = (HHH, HHT, HTH, THH)$

$$P(A) = 4/8 = 1/2$$

Example 2:

Throw 2 dice, (a) what is the probability of getting even number on the first die?

(b) what is the probability of getting odd number on the second die?

(c) Are these two events independent?

(d) What is the probability that sum of the numbers on both the dice is 11?

Sol'n:  $S = (1,1) (1,2) \dots (1,6)$

$(2,1) \dots (2,6)$

$\dots$

$(6,1) \dots (6,6)$  total 36 points in the sample space

# Probability.....

(a) what is the probability of getting even number on the first die? =  $18/36 = 1/2$

(b) what is the probability of getting odd number on the second die? =  $18/36 = 1/2$

(c) Are these two events independent?

$P(A) = \frac{1}{2}$ ,  $P(B) = \frac{1}{2}$  and  $P(A \cap B) = \frac{9}{36} = \frac{1}{4}$  ( how),  $P(A) \times P(B) = P(A \cap B) = \frac{1}{4}$  independent

(d) What is the probability that sum of the numbers on both the dice is 11?

Sum can be any number from 2 to 12 but 11 can occur in following ways:

(5,6) and (6,5) so probability of getting 11 is =  $2/36$

Let me know what is the probability of getting sum 9?

# Random Variable

- Need of this tool in probability?
- First what is function in mathematics?
- **Function**, in mathematics, an expression, rule, or law that defines a relationship between one variable (the independent variable) and another variable (the dependent variable).

OR

A unique relation between two sets A and B such that every element of A has a unique image in B. Set A is known as domain, Set B is known as co-domain and subset of B which has an inverse image in A is known as range of the function

- Random Variable: It is a simple function defined from sample space to the real line.
- Notation: Generally random variables are represented by capital letters X,Y,Z and the range of the same random variable is given by the corresponding small letter x,y,z ( Function of rv's are also rv's explain)

# Random Variable.....

**EXAMPLE :** In the experiment of tossing a fair coin three times , *the sample space  $S$ , consists of* eight equally likely sample points  $S, = (HHH, \dots, TTT)$ . *If  $X$  is the r.v. giving the number of heads obtained, find*

(a)  $P(X = 2)$ ; (b)  $P(X < 2)$ .

(a) Let  $A \subset S$ , be the event defined by  $X = 2$ . *Then, from Prob. example, we have*

- $A = (X = 2) = \{C: X(C) = 2\} = \{HHT, HTH, THH\}$

Since the sample points are equally likely, we have

- $P(X = 2) = P(A) = 3/8$
- Let  $B \subset S$ , be the event defined by  $X < 2$ . Then
- $B = (X < 2) = \{c: X(c) < 2\} = (HTT, THT, TTH, TTT)$
- and  $P(X < 2) = P(B) = 4/8$



# Random Variable.....

Types of random variable: Discrete and continuous( Example and explanation)

- X is a discrete r.v. only if its range contains a finite or countably infinite number of points.
- X is a continuous r.v. only if its range contains an interval (either finite or infinite) of real numbers. Thus, if X is a. continuous r.v., then  $P(X=x) = 0$  ( WHY)

# Random Variable.....

- The *probability distribution* for a random variable describes how the probabilities are distributed over the values of the *random variable*.
- For a *discrete random variable*,  $x$ , the probability distribution is defined by a **probability mass function**, denoted by  $f(x)$ . This function provides the probability for each value of the random variable.
- For a *continuous random variable*, since there is an infinite number of values in any interval, the probability that a continuous random variable will lie within a given interval is considered. So here, the probability distribution is defined by **probability density function**, also denoted by  $f(x)$ .
- Both probability functions must satisfy two requirements:: (1)  $f(x)$  must be non-negative for each value of the random variable, and (2) the sum of the probabilities for each value (or integral over all values) of the random variable must equal one.

# Random Variable.....

Example1 : if we toss three coins write down the sample space, define on rv on this and write its pmf. Also calculate  $P(X \leq 2)$

Sol'n:  $S = (HHH, HHT, HTH, THH, HTT, THT, TTH, TTT)$ , let us define rv  $Y$  as number of tails then  $y = 0, 1, 2, 3$  and corresponding pmf is given by:

x	0	1	2	3	Total
f(x)	1/8	3/8	3/8	1/8	1

# Random Variable.....

Example2: If X is a continuous rv with following pdf:

$$f(x) = 6x(1-x) \quad 0 < x < 1$$
$$= 0 \text{ other wise}$$

- (a) Test if the given function is a pdf or not {integrate  $f(x)$  over 0 to 1}
- (b) Find  $P(X < \frac{1}{2})$
- (c) Can you conclude that the given pdf is symmetric?

# Random Variable.....

- Expected value of a rv is given by  $E(X) = \sum x.P(X=x)$  or  $\int xf(x)dx$
- Can you see any kind of similarity of this expression with the formula for mean?

Which is mean =  $1/N \sum f_i X_i$ .

- So expected value of a rv is nothing but the average value.
- Similarly we can extend this concept to variance and other moments.
- $V(X) = E[X - E(X)]^2 = \sum [X - E(X)]^2 P(X=x)$  or  $\int [X - E(X)]^2 f(x)dx$
- Z- score or standard score of any given rv variable can be given by:

$$Z = [X - E(X)] / \sqrt{V(X)}$$

or 
$$Z = (RV - \text{mean}) / SD$$

- Why we call it standard?

# Binomial Distribution

- Binomial distribution: The distribution of the number of 'successes',  $X$ , in a series of  $n$  independent **Bernoulli trials** where the probability of success at each trial is  $p$  and the probability of failure is  $q = 1-p$ .

$$P(X=x) = {}^nC_x p^x q^{(n-x)}$$

The mean, variance, skewness and kurtosis of the distribution are as follows:

- mean =  $np$
- variance =  $npq$
- skewness =  $(q-p)/\sqrt{npq}$
- kurtosis =  $3 - 6/n + 1/npq$

Let us learn the term parameters of the distribution.

What is the range of Binomial distribution?

In which kind of situation this can be used as probability models?

# Poisson Distribution

- We can derive or get Poisson from binomial with some conditions.
- The probability distribution of the number of occurrences,  $X$ , of some random event, in an interval of time or space.

$$P(X=x) = e^{-\lambda} \cdot \lambda^x / x! \quad x = 0, 1, 2, \dots$$

- The mean and variances of the distribution are both  $\lambda$ . The skewness of the distribution is  $1/\sqrt{\lambda}$  and its kurtosis is  $3 + 1/\lambda$
- Its range, its parameter
- In which situations this can be used a probability models?

# Normal Distribution

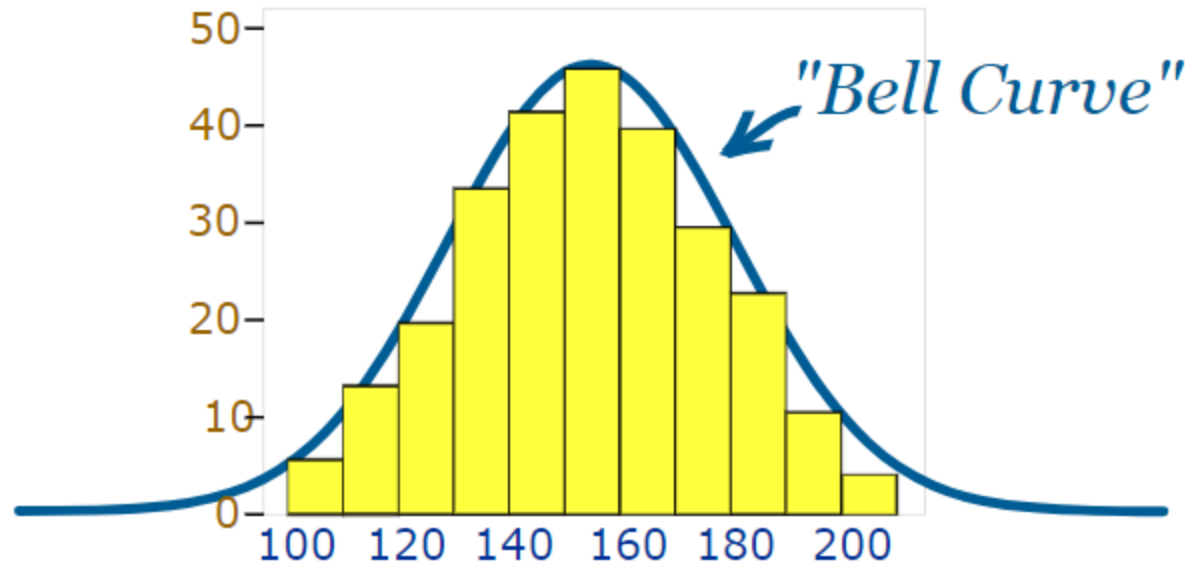
- What is normal distribution?
- Why it is important in statistics? ( CLT to many assumptions)
- Why it has been used as bench mark in statistical theory?
- What are its important features?



# Normal Distribution.....

- The **normal distribution**, also known as the **Gaussian distribution**, is a *probability distribution* that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. It has following properties:
- The normal curve is symmetrical about the mean  $\mu$ ;
- The mean is at the middle and divides the area into halves;
- The total area under the curve is equal to 1;
- It is completely determined by its mean and standard deviation  $\sigma$

# Normal Distribution.....



A Normal Distribution

# Statistical Inference

- Estimation( Interval or Point)
- Basics of estimation-Point
- Basics of estimation-Interval
- Testing
- Basics of Testing ( Types of hypothesis, level of significance, types of error and.....)
- Variety of problems we face in testing.