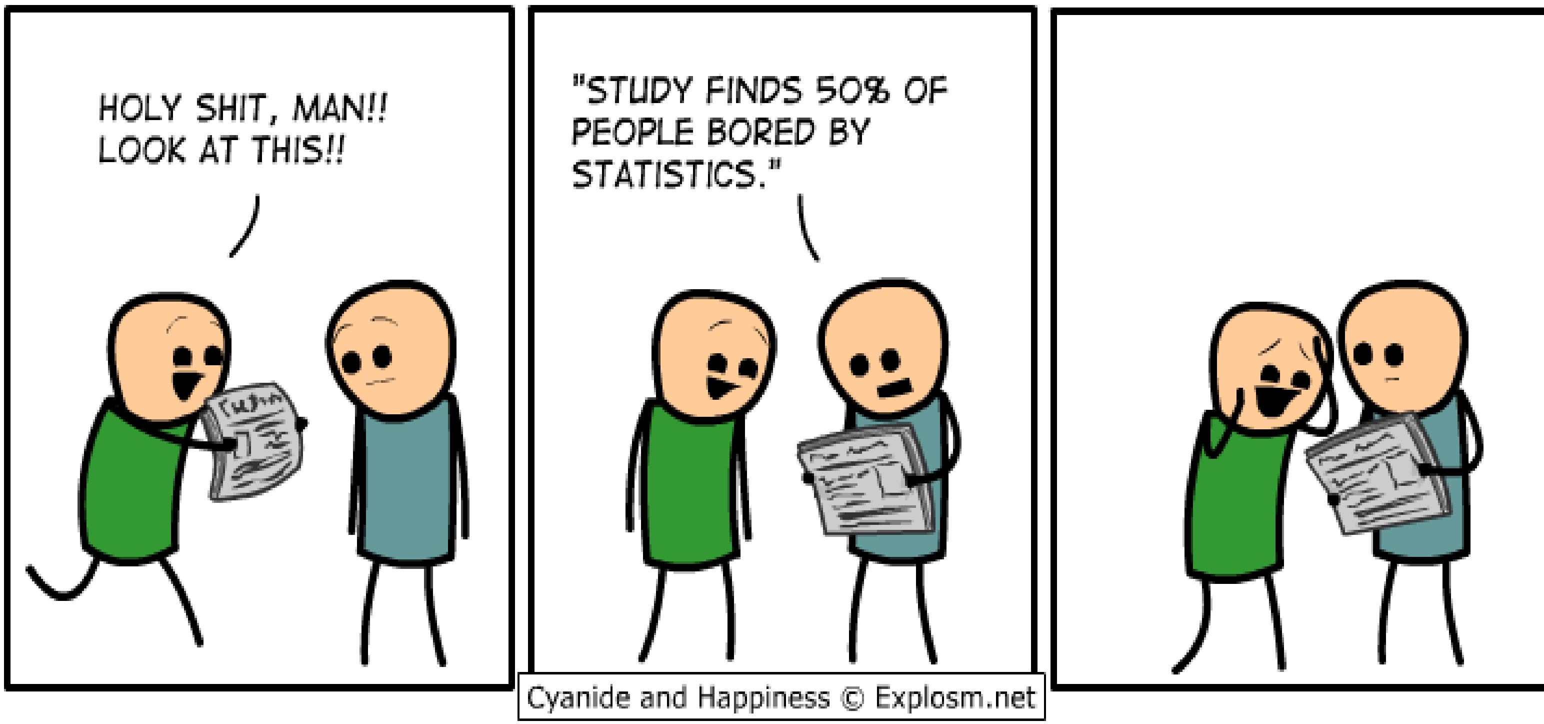
Basic Statistics



https://www.pinterest.com/pin/250723904226784778/

What is Statistics?

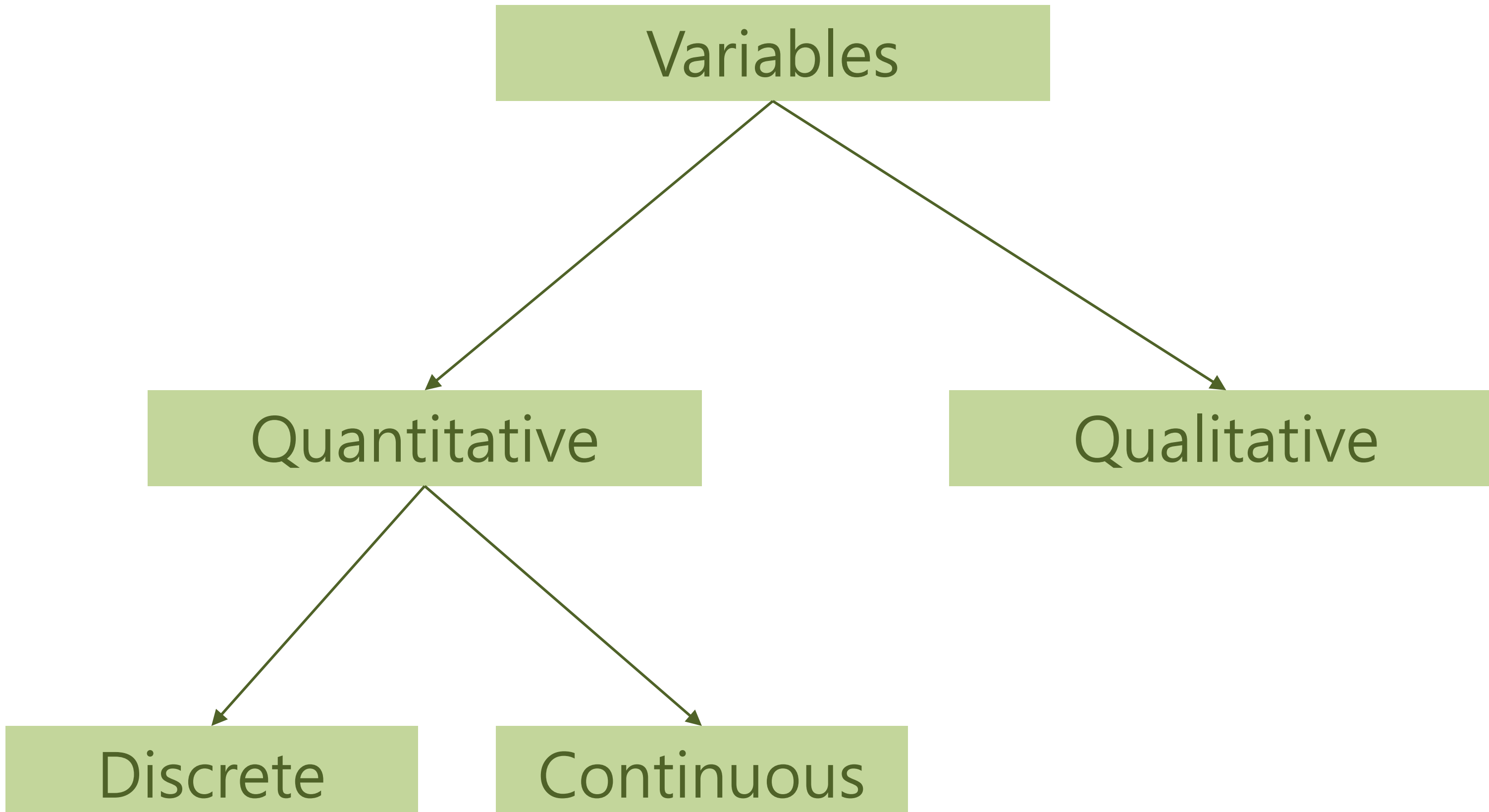
- Statistics is the science of
  - Collecting
  - Organizing
  - Summarizing
  - Analyzing and
  - Interpreting data
- The goal of statistics is to:
  - Infer facts
  - Make predictions about future
  - Help make better decisions

Basic Definitions

Variables

- Properties or characteristics of some event, object, or person that can take different values
- An attribute that describes a person, place, thing or ides
- Examples:
  - Height of a person
  - Hair color of a person
  - Inflation percentage in a country
  - Number of votes scored by a candidate in an election
  - Number of persons in a household

Types of Variables



Dependent & Independent Variables

- In statistics, the dependent variable is the event studied and expected to change whenever the independent variable is changed or altered
- E.g. Final marks obtained by students vs. time spent by the students
- The variable final mark is dependent on the independent variable time spent
- Independent variable represent the input or causes and is also known as predictor variable
- Dependent variable represent the output or effect and is also known as output variable

Quantitative & Qualitative Variables

- Quantitative variables take on values that are numeric for which arithmetic operations make sense
- E.g. Height of a person, GDP of a country etc.
- Qualitative variables take on values that are names or labels
- E.g. Hair color, breed of dog etc.
- Qualitative variables are numeric and hence arithmetic operations on Qualitative variables do not make sense

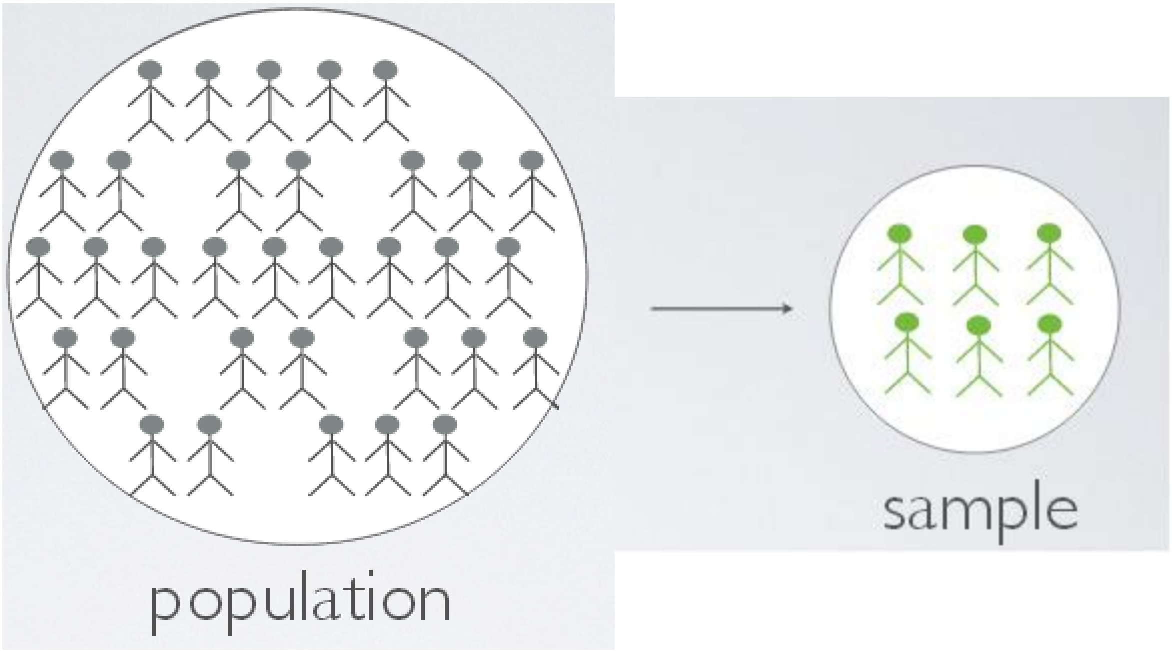
Discrete & Continuous Variables

- Discrete variables can take only certain values
- E.g. Number of persons in a household, outcome of rolling a six sided die
- Continuous variables can take any value between its maximum and minimum value
- E.g. Time taken by students to complete a 3 hour test, Height of students in a class

Classify the variables

- Breed of dog
- Blood sugar level of a person
- Final result in a examination
- Number of matches played by a player
- Batting average of a player
- Favourite colour
- State which you belong to
- Marks scored in a subject
- Average marks in all subjects

Population & Sample



http://www.bindichen.co.uk/post/Notes/data-analysis-and-statistical-inference/experimental-design-and-random-sample-assignment.html

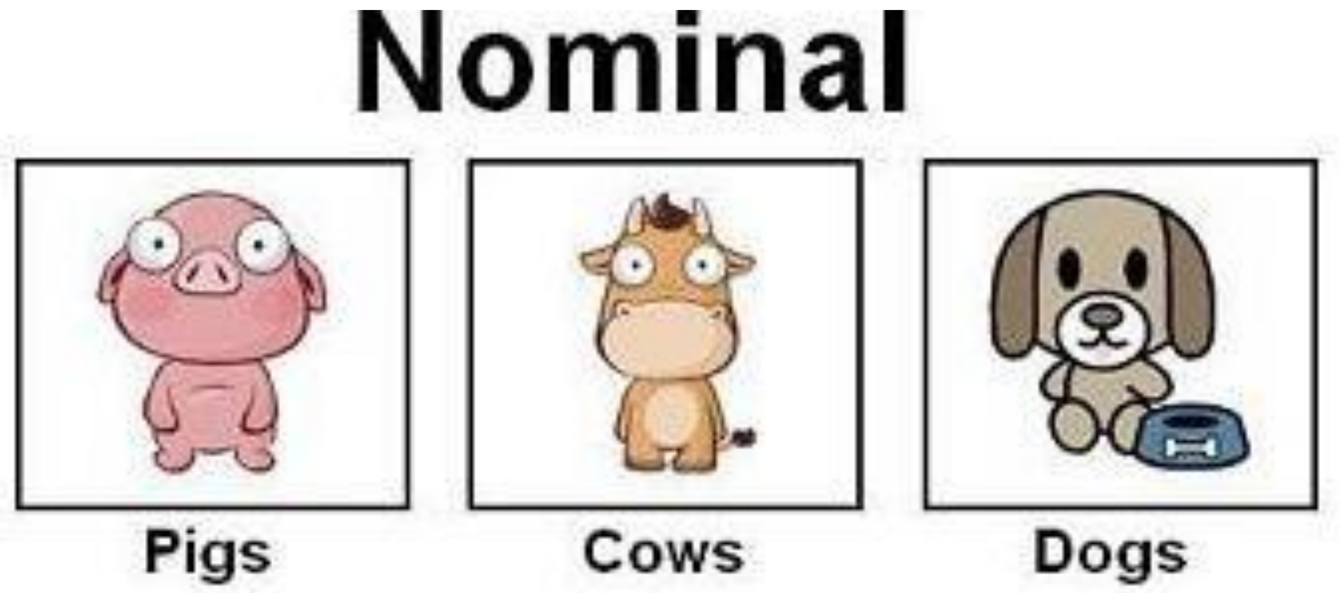
Population & Sample (contd.)

- Population is the whole set of values or individuals you are interested in
- The number of items or elements in a population is called the population size, denoted by N
- Sample is a subset of the population
- Sample size (number of observations in the sample) is denoted by n
- Randomization schemes help to build samples that are truly representative of the population

Scales of Measurement

- Measurement scales are used to categorize and/or quantify variables
- Four different scales are commonly used in statistics:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nominal Scale



www.suggestkeyword.com

Nominal Scale (contd.)

- Most basic level of measurement
- In Nominal Scale, data is neither measured nor ordered
- Subjects are merely allocated to distinct categories
- Also known as categorical or qualitative
- Examples:
  - Sex
  - Color preference
  - Religion
- Values can be stored as text or a numerical code
- However numbers do not imply order

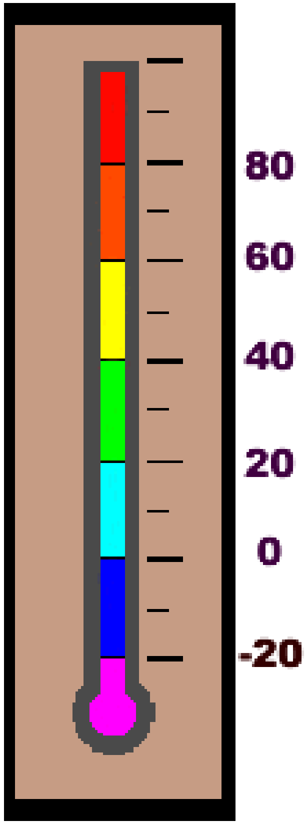
Ordinal Scale

http://www.ssbinterviewtips.in/2013/02/rank-structure-and-pay-scales-of-army.html

Ordinal Scale (contd.)

- Next level of measurement is Ordinal
- Data is ordered
- But difference between two levels may not be the same as the difference between another two levels
- Comparison is however possible
- Examples:
  - Military Rank
  - Consumer Satisfaction Ratings
  - Rankings in a class

Interval Scale



http://www.wpclipart.com/weather/weather_instruments/thermometer/thermometer.png.html

Interval Scale (contd.)

- Order is meaningful
- Intervals are equal
- Things that can be measured are expressed in interval scale
- But data has no zero point and hence ratio is of no real meaning
- Examples:
  - Time of a day
  - Temperature in Celsius

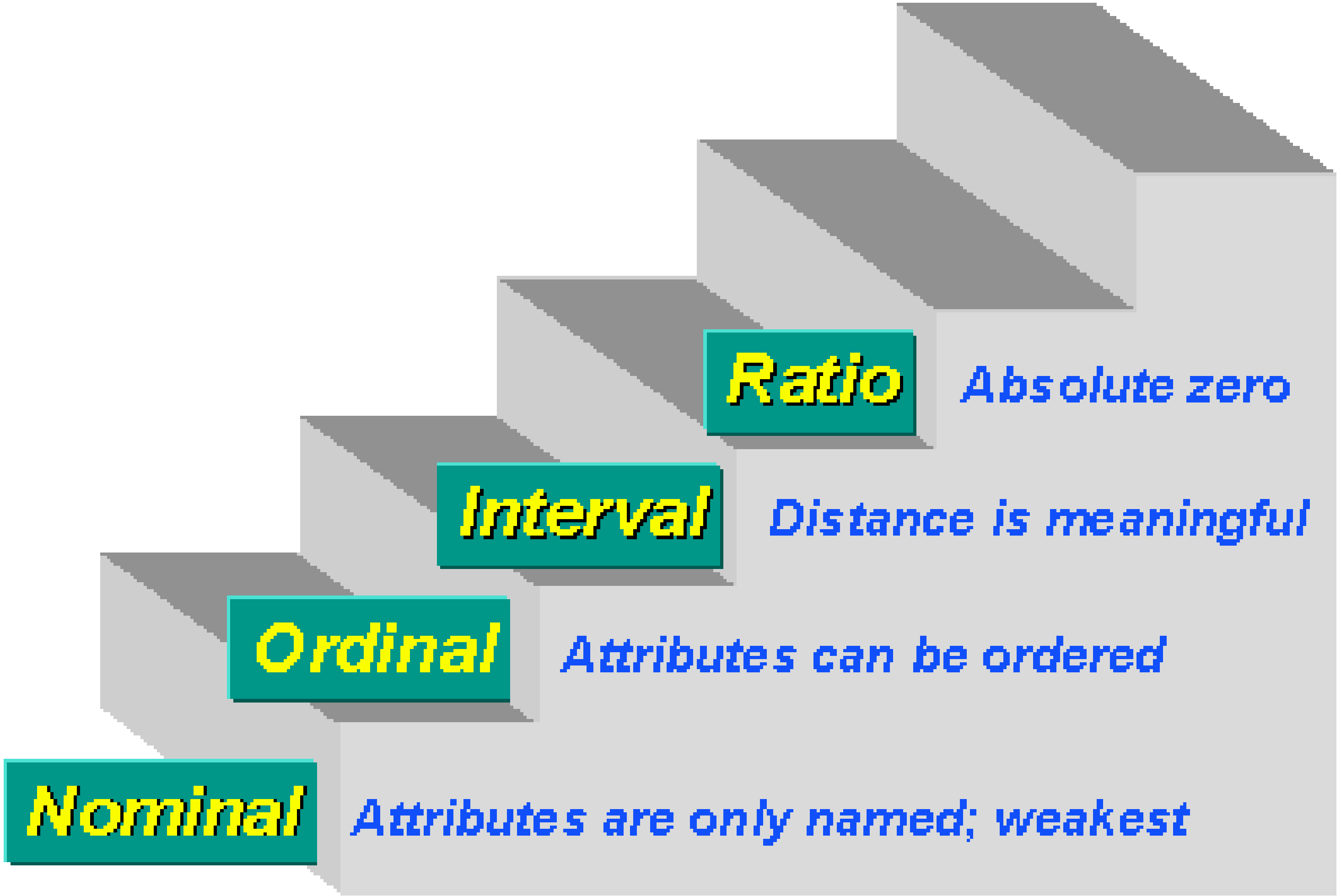
Ratio Scale

http://www.etc.edu.cn/eet/articles/measurescales/start.htm

Ratio Scale (contd.)

- Highest and most informative scale
- Has the qualities of nominal, ordinal and interval scales with the addition of an absolute zero point
- Examples:
  - Years of experience
  - Amount of money
  - Number of children in a household

Scales of Measurement (contd.)



http://www.socialresearchmethods.net/kb/measlevl.php

Scales of Measurement (contd.)

Let us consider a set of candidates taking an exam. Identify which scale of measurement to be used to measure the following variables:

- Sex of the candidate
- State which the candidate is from
- Optional subjects chosen by the candidate
- Marks obtained by the candidate
- BMI Classification of a candidate (Underweight, Ideal, Overweight, Obese)
- Final grade obtained by the candidate
- Confidence level of candidate before taking exam on a scale of 1 to 10

Summarizing Data

Central Tendency

Arithmetic Mean

- Sum of collection of numbers divided by the number of numbers
- Commonly known as average
- Measure of central tendency
- Commonly denoted by the symbol $\bar{x}$
- Arithmetic Mean = (sum of observations) / (No. of observations)

$$\bar{x} = \frac{(x_1 + x_2 + \ldots\ldots\ldots + x_n)}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

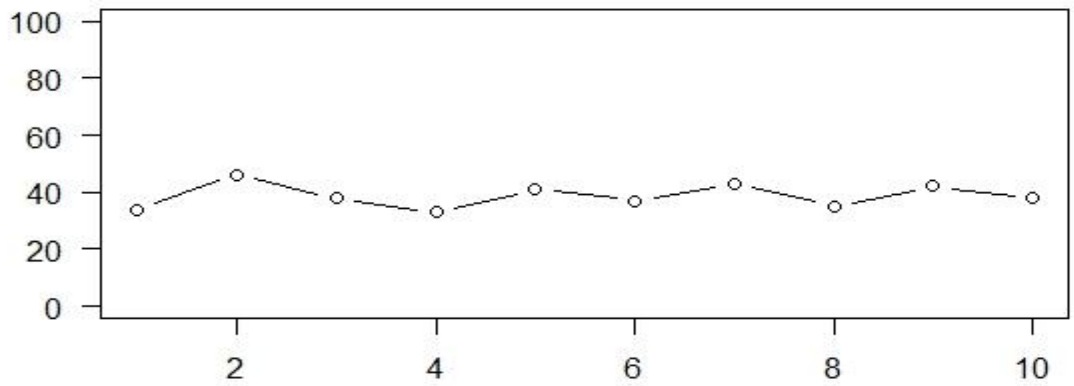
Median

- The number separating the higher half of the data sample from the lower half
- In other words it is the middle number between the smallest number and the largest number
- Can be found by arranging the observations from lowest value to the highest value and picking the middle one
- When the number of observations is odd: Median is $(n+1)/2^{th}$ observation
- When the number of observations is even: Median is mean of $n/2^{th}$ and $(n/2)+1^{th}$ observation

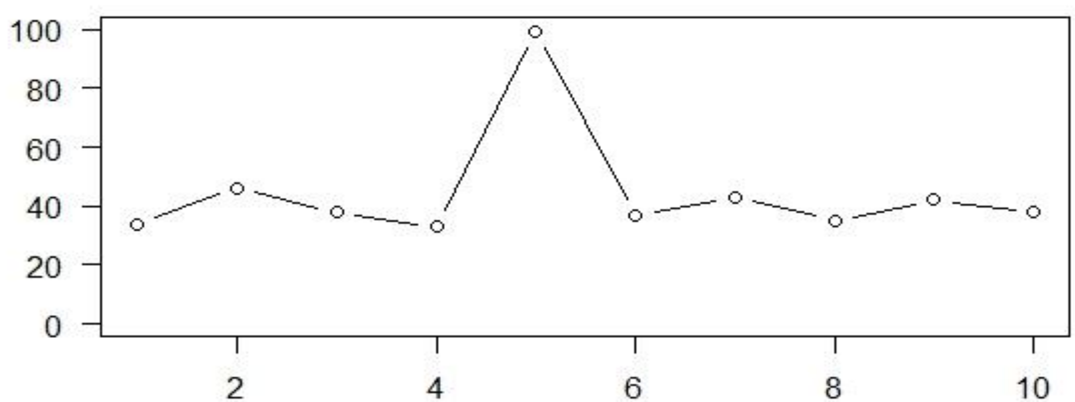
Trimmed Mean

- Consider the following set of numbers:
- 34, 46, 38, 33, 41, 37, 43, 35, 42, 38
- Find the mean and median
- Mean is 38.7 and Median is 38
- Now consider this new set of numbers:
- 34, 46, 38, 33, 99, 37, 43, 35, 42, 38
- Find the mean and median
- Mean is 44.5 and Median is 38

Trimmed Mean (contd.)



First Set



Second Set

Trimmed Mean (contd.)

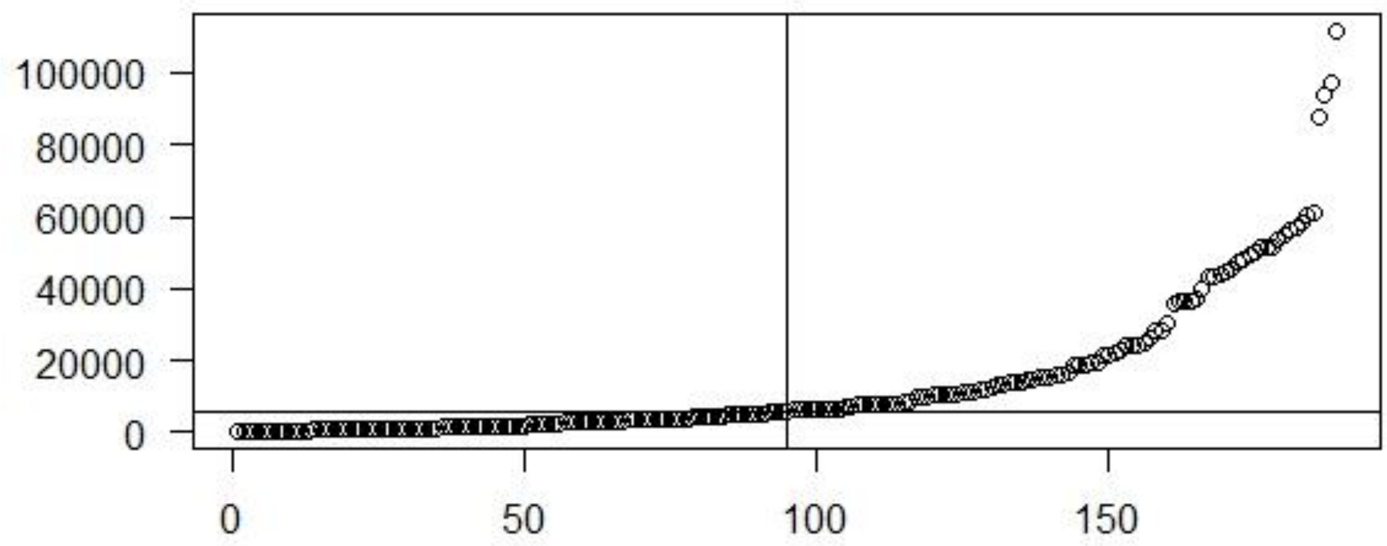
- Consider the following set of numbers:
- 34, 46, 38, 33, 41, 37, 43, 35, 42, 38
- Find the mean and median
- Mean is 38.7 and Median is 38
- Now consider this new set of numbers:
- 34, 46, 38, 33, 99, 37, 43, 35, 42, 38
- Find the mean and median
- Mean is 44.5 and Median is 38
- The difference is because there is an outlier in the data '99' which is vastly different from the rest of the data
- Outlier influences the mean

Trimmed Mean (contd.)

- Trimmed Mean is normal Mean except that a certain percentage of the extremes are omitted while calculating the mean
- This effectively removes the outliers from the observations
- The 10% trimmed mean of the observation is 39.125
- The Trimmed Mean 39.125 is more closer to the Median 38 than the true Mean 44.5

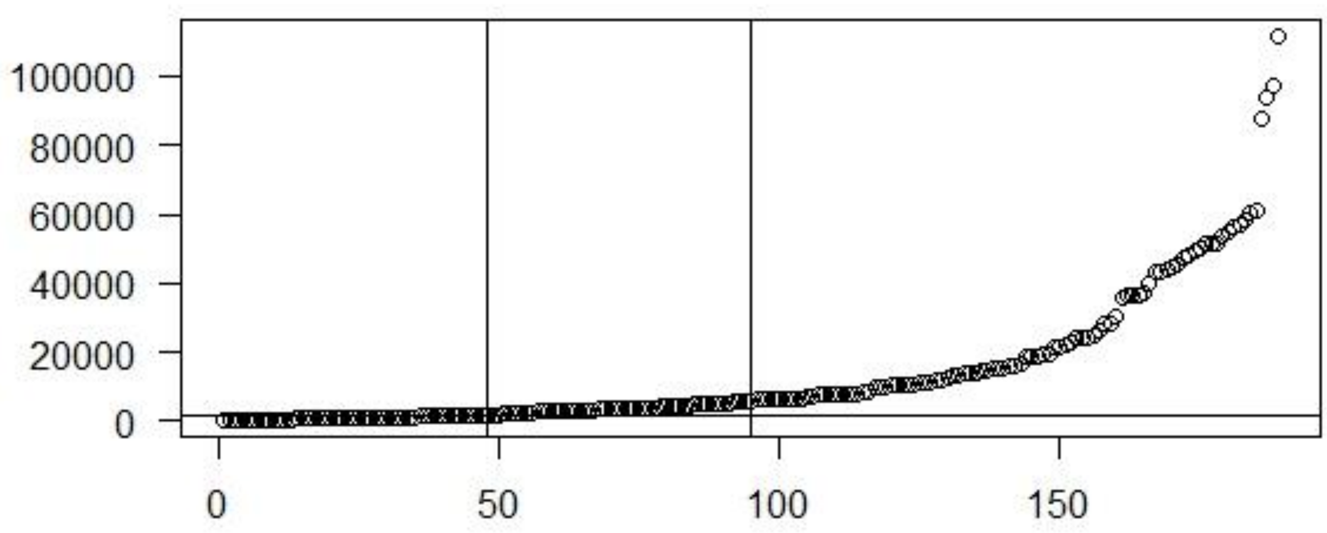
Quartile

- Median divides the data into two equal halves



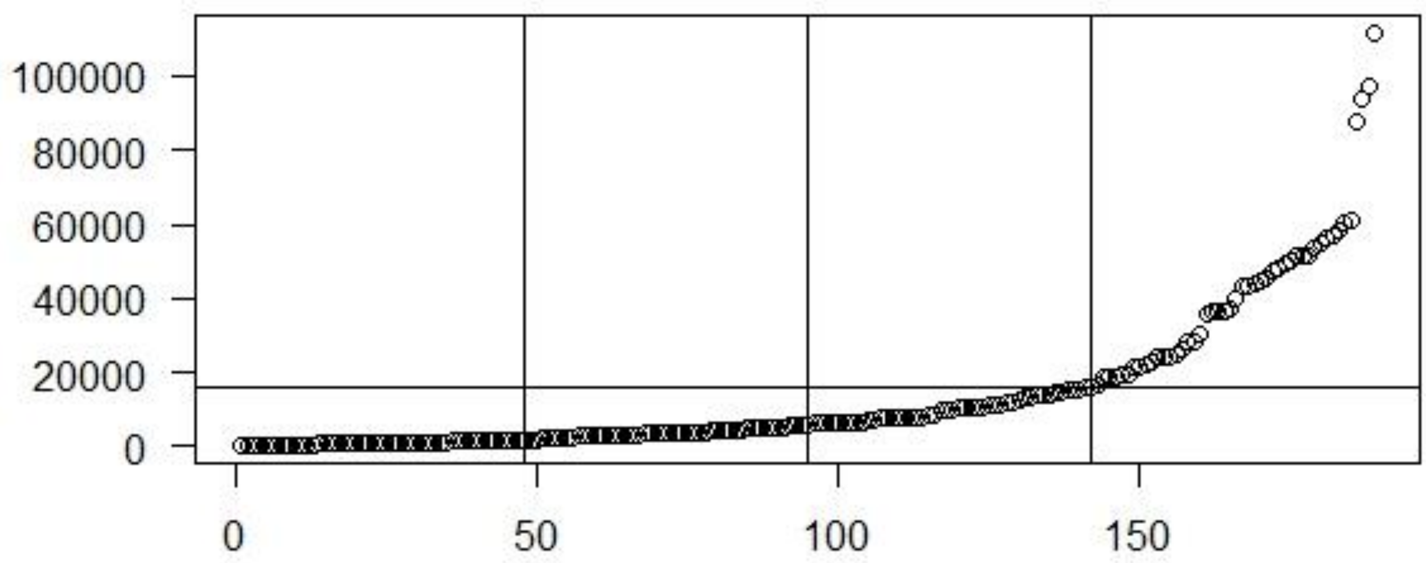
Quartile (contd.)

- Lets consider the lower half of the data



- The value 1781 is called the 1st Quartile
- 1/4th of the observations are below 1781

Quartile (contd.)



- The value 16199 is called the 3rd Quartile
- 3/4th of the observations are below 16199
- Median is the 2nd Quartile

Mode

- Mode is the value that appears most often in a set of data
- E.g. mode of 1,1,2,4,5,4,6,3,7,1 is 1 since 1 is appearing three times
- A data may have more than one mode
- E.g. the dataset 1,1,3,1,4,5,5,6,5 has two modes 1 and 5
- A data may not have a mode
- E.g. the dataset 1,2,3,4,5 doesn't have a mode since all the numbers appear only once
- There is no standard library function to find mode in R

Mode (contd.)

- Write a function to calculate mode
- Steps:
  - Select all the unique values of the object
  - Count the number of times each unique value appears in the object
  - Store the count values in a vector
  - Find out the maximum value of the count using ‘which’ function
  - Find out the unique values which correspond to the maximum value
  - Use ‘if’ condition to return an error message if all unique values have same count

Summarizing Data

Spread of the data

Range

- Range is the difference between the lowest and the highest values
- In {4, 5, 9, 3, 8} the lowest value is 3, and the highest is 9, so the range is $9 - 3 = 6$
- Most useful in representing the dispersion of small data sets

Inter Quartile Range

- Difference between the 1st quartile and the 3rd quartile
- Middle half of the data fits into the Inter Quartile Range (IQR)
- Also known as mid-spread or middle fifty

Variance & Standard Deviation

- Measures how spread out our data is with reference to the mean
- Variance is always positive
- Small variance means data are close to each other
- Large variance means data are spread out widely
- Standard deviation is the square root of variance

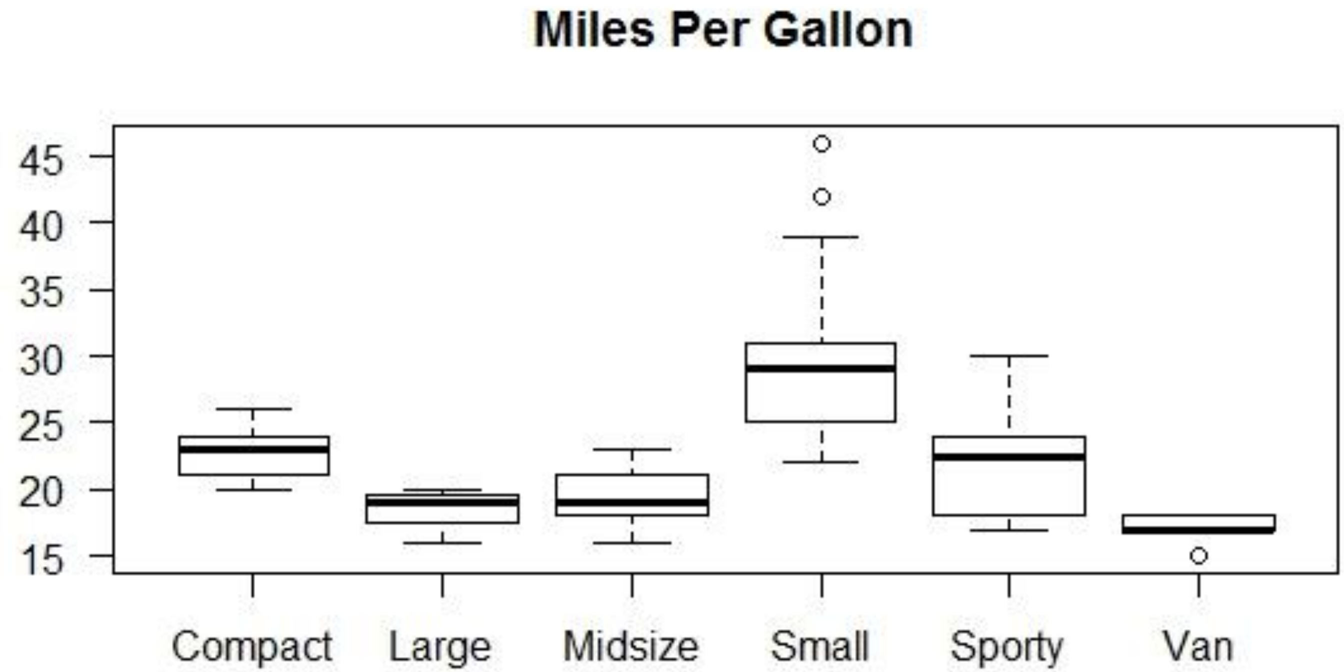
$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

Variance & Standard Deviation (contd.)

- The value in the denominator of the formula is called as the Degrees of Freedom
- Degrees of Freedom is the number of values in the final calculation of a statistic that are free to vary
- Standard deviation is usually denoted by the symbol sigma σ or S^2
- Empirical Rule:
  - ~ 68% of data lies within the range (mean – 1σ) and (mean + 1 σ)
  - ~ 95% of data lies within the range (mean – 2σ) and (mean + 2σ)
  - ~ 99.7% of data lies within the range (mean – 3σ) and (mean + 3σ)

Data Graphs

Box Plot

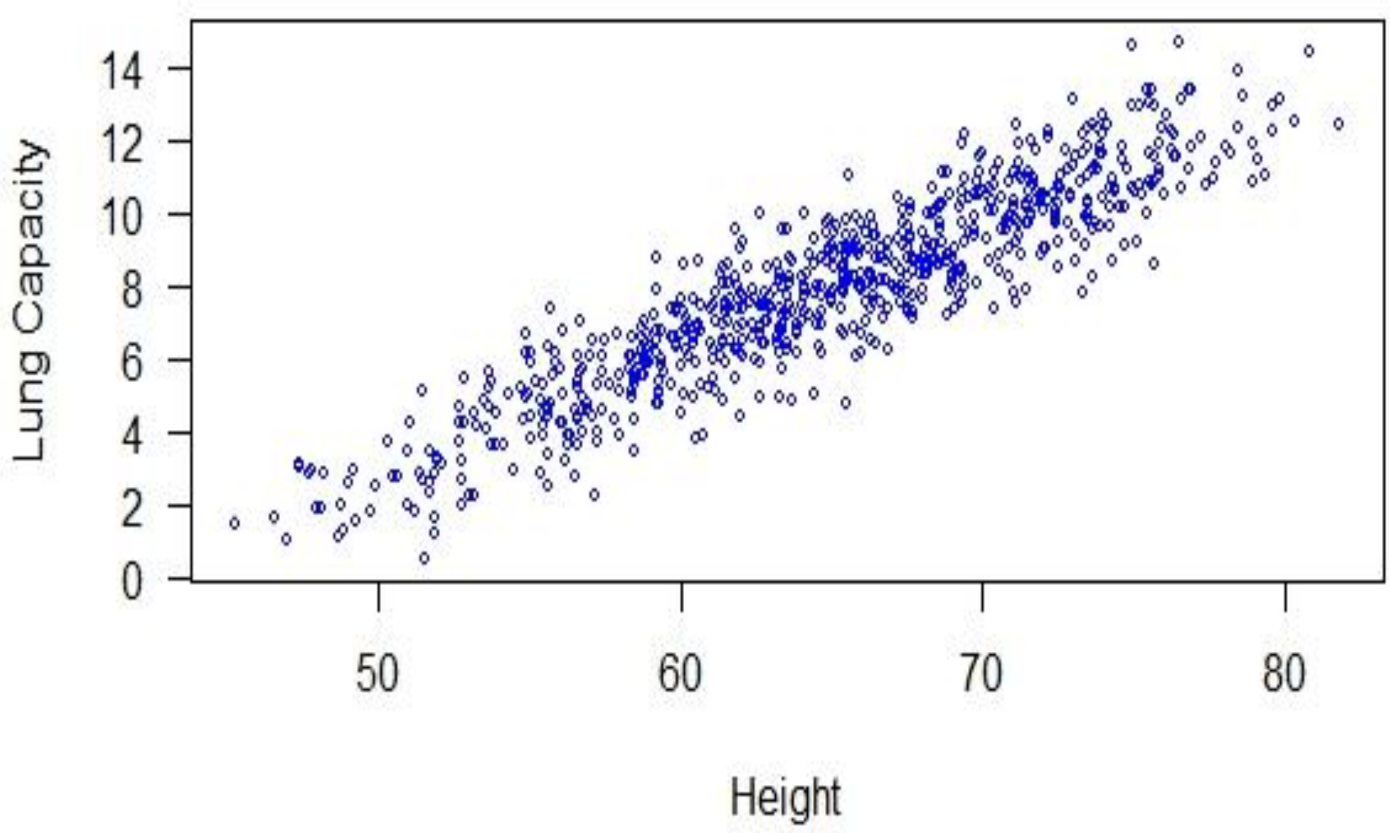


Miles Per Gallon
45
40
35
30
25
20
15
Compact
Large
Midsize
Small
Sporty
Van

Box Plot (contd.)

- Graphically summarize numerical variables
- Central line is the median
- The lower end of the box is the 25th percentile
- The upper end of the box is the 50th percentile
- Inter Quartile Range (IQR) is the difference between the 25th and 50th percentile
- The whiskers are the lower inner fence and the upper inner fence
- Lower inner fence = 1st Quartile – 1.5 * IQR
- Upper inner fence = 3rd Quartile + 1.5 * IQR
- Outliers are shown outside the whiskers
- Summary command in R will give you all the values needed to create a box plot

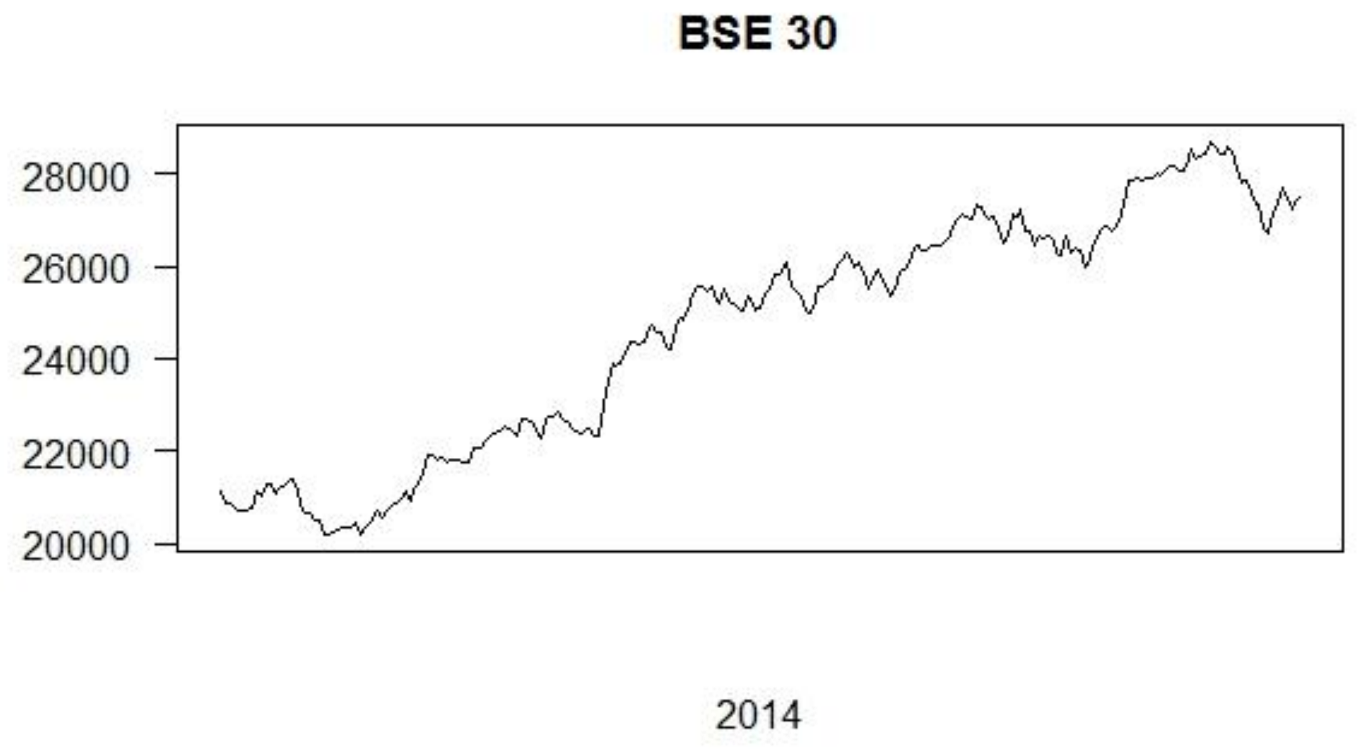
Scatter Plot



Scatter Plot (contd.)

- Displays two variables for a set of data in x and y axes
- It shows the relationship between two variables i.e. how much one variable is affected by the other
- The relationship between the two variables is called their correlation

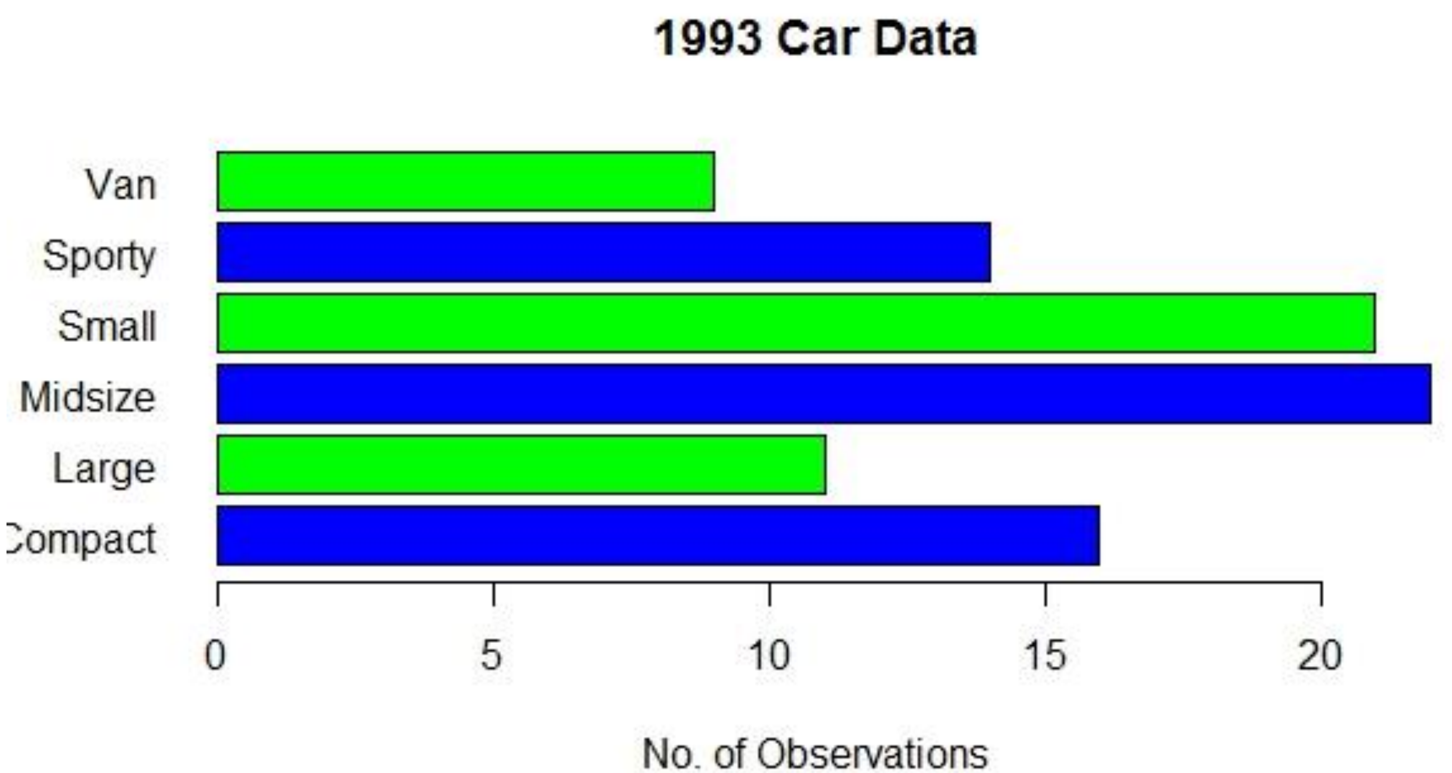
Line Graph



Line Graph (contd.)

- Displays information as a series of data points connected by straight line segments
- Useful in displaying data or information that changes continuously over time
- Also known as Line Chart

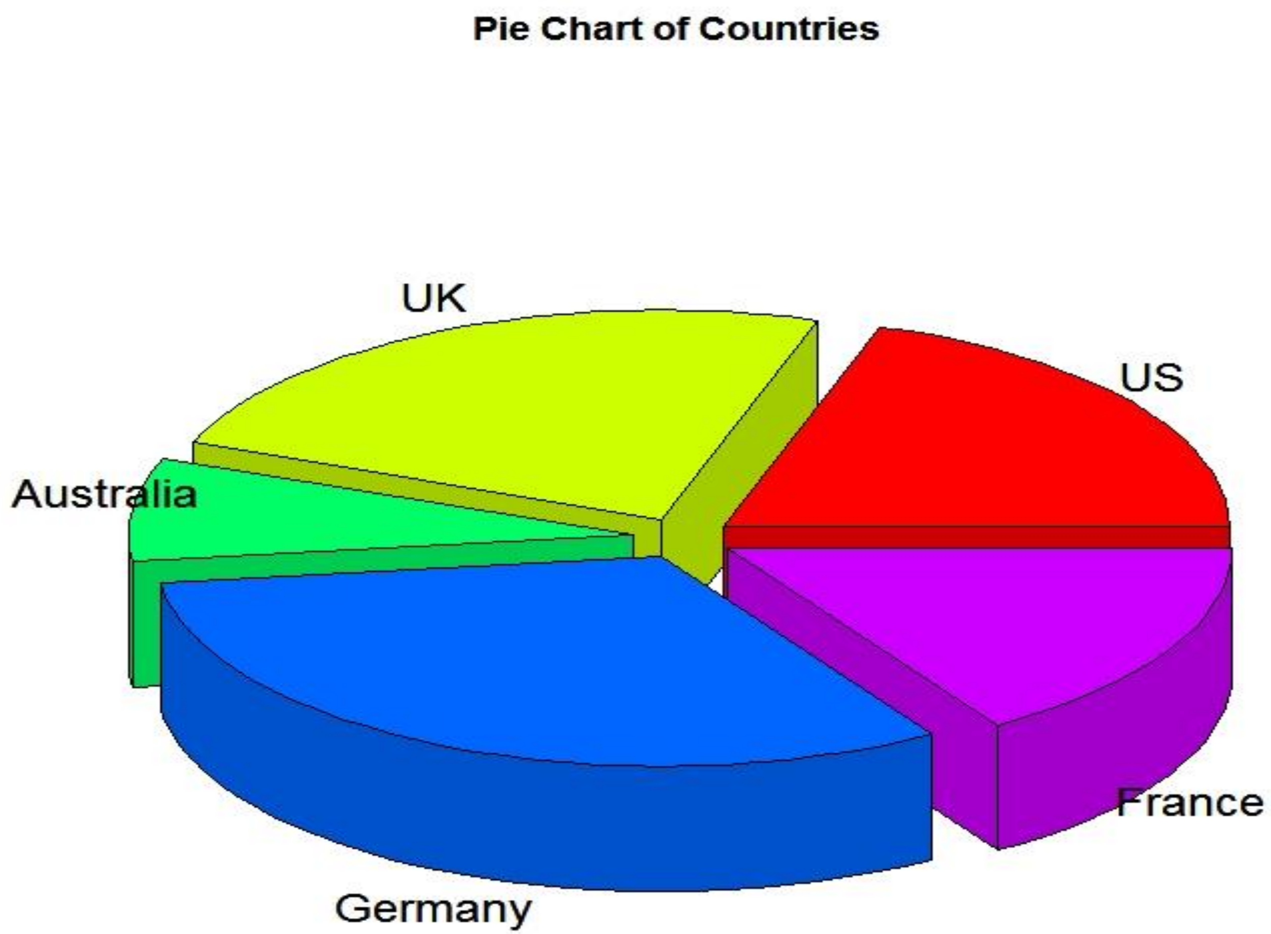
Bar Graph


1993 Car Data
Van
Sporty
Small
Midsize
Large
Compact
0
5
10
15
20
No. of Observations

Bar Graph (contd.)

- Visual display of frequency or relative frequency (%) for each category of a categorical variable
- Also known as ‘Bar Chart’ or ‘Column Bar Chart’
- To construct a Bar Graph:
  - Compute frequency for each category
  - Plot frequency or relative frequency of each category as a bar
  - Height of each bar should correspond to the frequency of each category

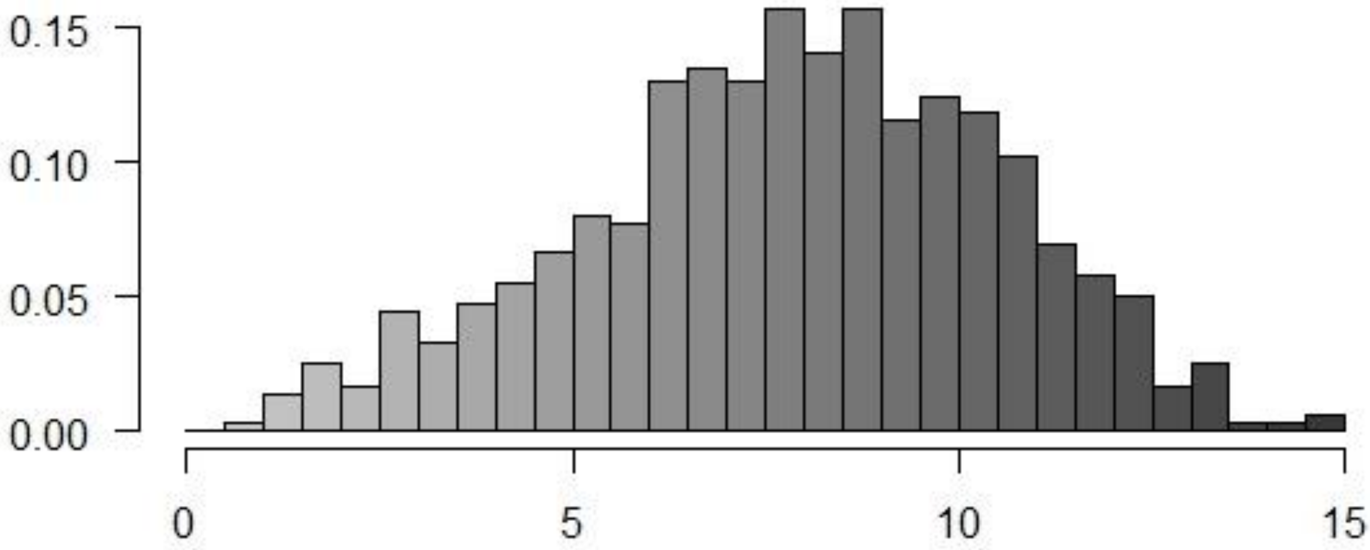
Pie Chart



Pie Chart (contd.)

- Circular graphic divided into slices
- The size of each slice (or the angle or length of the arc) is proportional to the quantity it represents
- Pie Chart is widely criticized by statisticians in recent times because they are difficult to interpret

Histogram



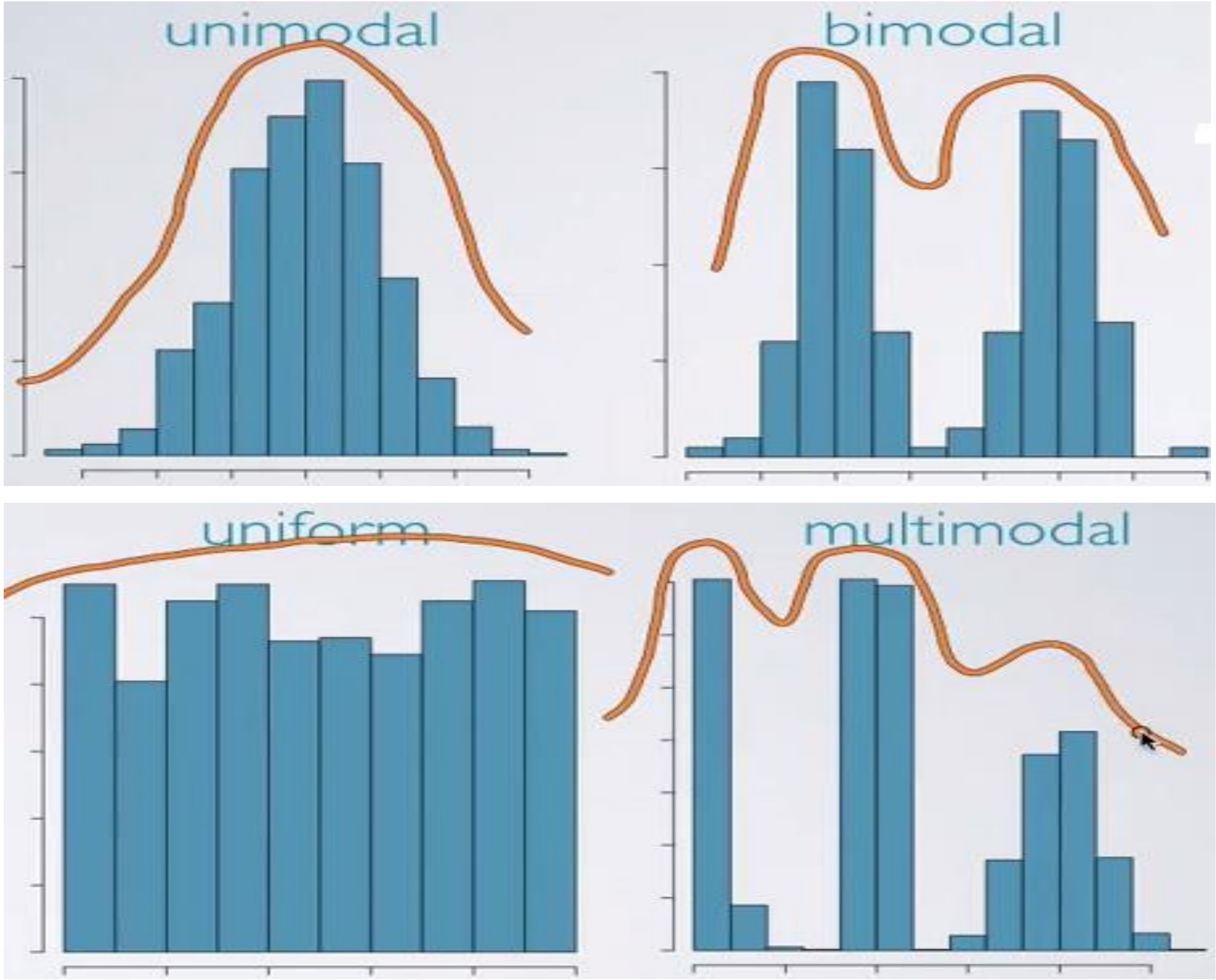
Histogram (contd.)

- Graphical representation of the distribution of numerical data
- Shows how often each different value in a set of data occurs
- To construct a Histogram:
  - Group numeric data into different bins of equal width
  - Place the bins on the horizontal axis
  - Place the frequencies (no. of occurrenc es) of each bin in the vertical axis usin g a bar extending across each bin

Histogram (contd.)

- We can get a lot of information about the data using a histogram
- The peak of the distribution is the mode of the data
- A histogram can be unimodal (single peak), bimodal (two peaks) or multimodal (multiple peaks) or uniform (no prominent peaks)

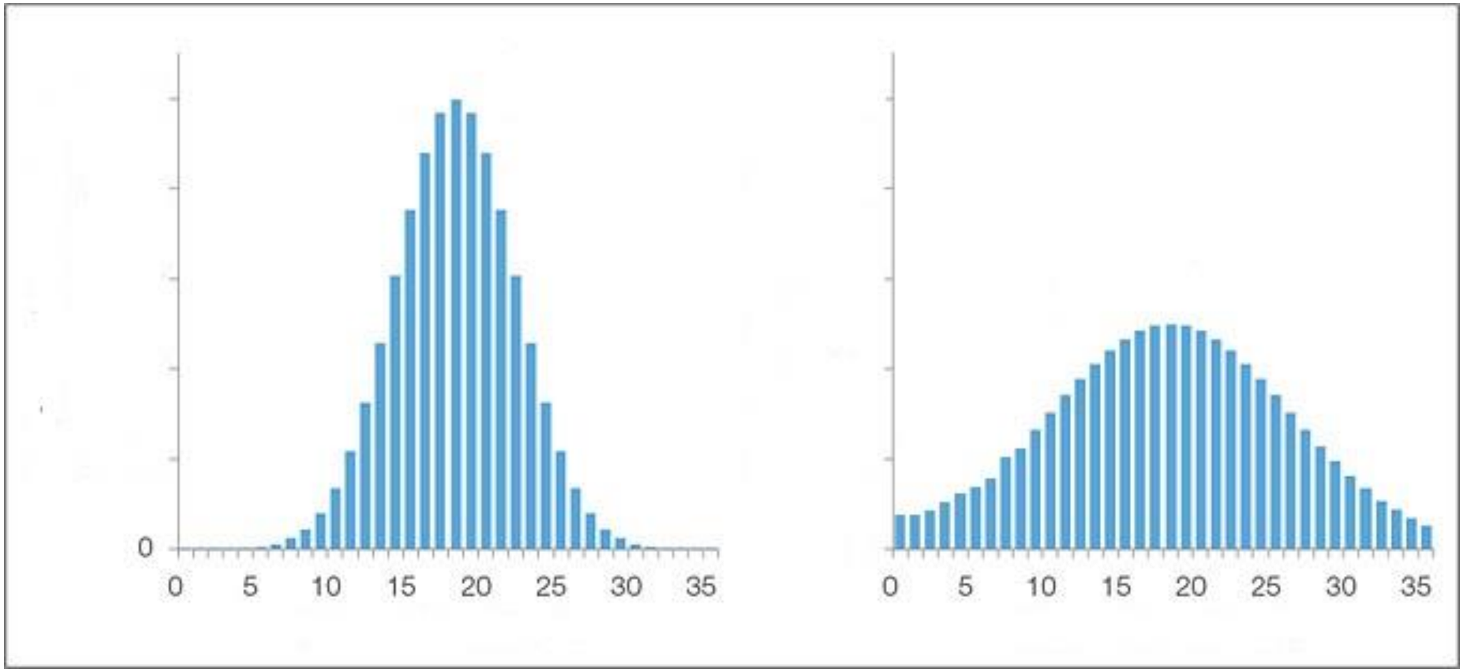
Histogram (contd.)



http://www.bindichen.co.uk/post/Notes/data-analysis-and-statistical-inference/visualizing-numerical-data.html

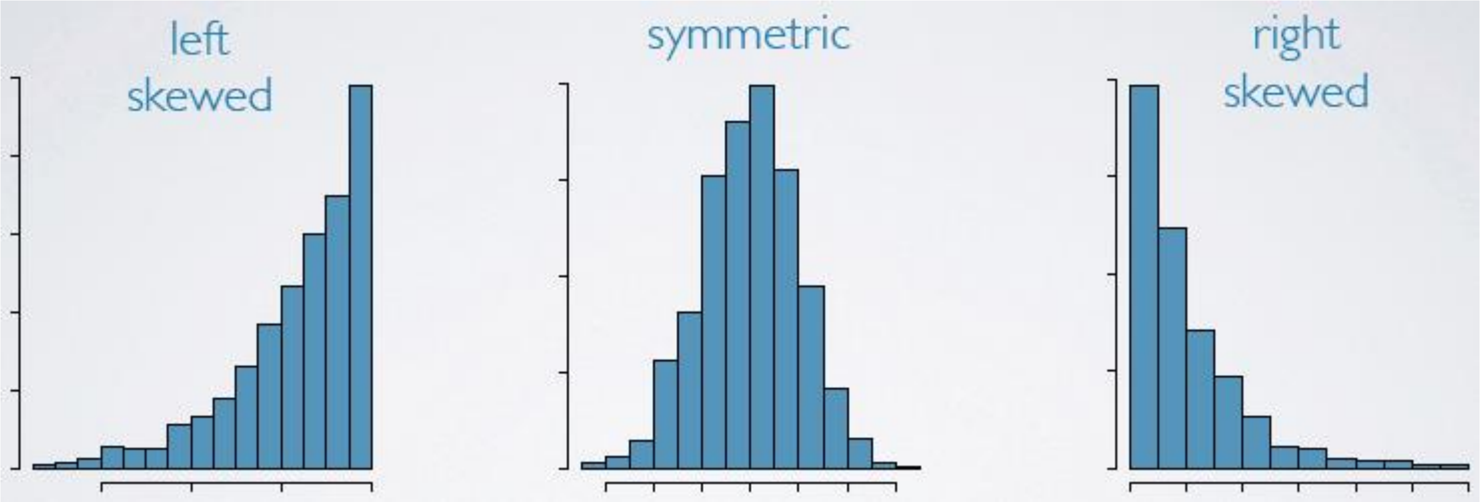
Histogram (contd.)

- Histogram also gives us an idea about the extent of spread of data



Histogram (contd.)

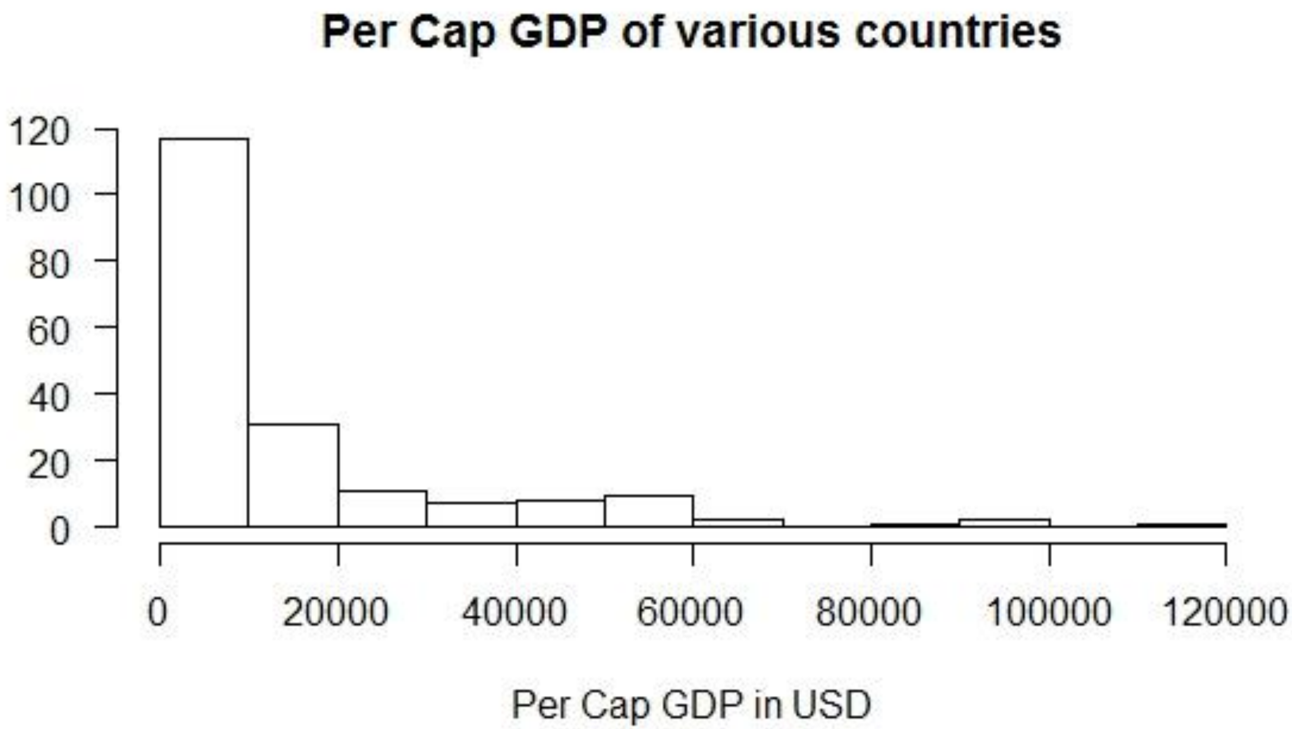
- We can also infer about the symmetry of the data from Histograms
- Left skewed is also known as negative skewed and right skewed is also known as positive skewed



http://www.bindichen.co.uk/post/Notes/data-analysis-and-statistical-inference/visualizing-numerical-data.html

Histogram (contd.)

- Lets have a look at the histogram of Per Cap GDP data



Introduction to Probability

Definitions

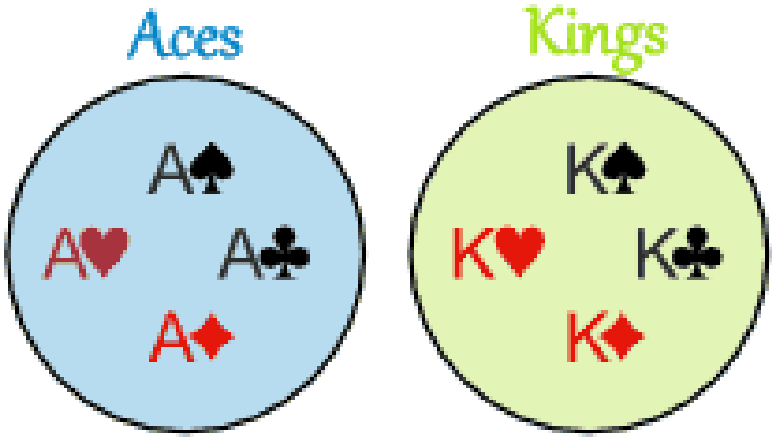
- **Probability** – The chance or likelihood that a particular uncertain event will occur
- **Sample Space** – Collection of all possible events
- **Event** – An event is a subset of a the sample space
  - For e.g. let us consider an experiment where a coin is tossed twice
  - The sample space is the all possible outcomes which in our case is {HH, HT, TH, TT}
  - We are interested in the case where at least one head occurs
  - So our event is {HH, HT, TH}

Definitions (contd.)

- **Complement of an Event** – All outcomes that are not part of an event
  - For e.g. the complement of the event at least one head occurs is no head occurs and is given by the subset {TT}
  - Complement of event A is denoted by A^C
- **Intersection of events** – The probability that events A and B both occur
  - Intersection of events A and B is denoted by $P(A \cap B)$
- **Union of events** – The probability that event A or B occurs
  - Union of events A and B is denoted by $P(A \cup B)$

Definitions (contd.)

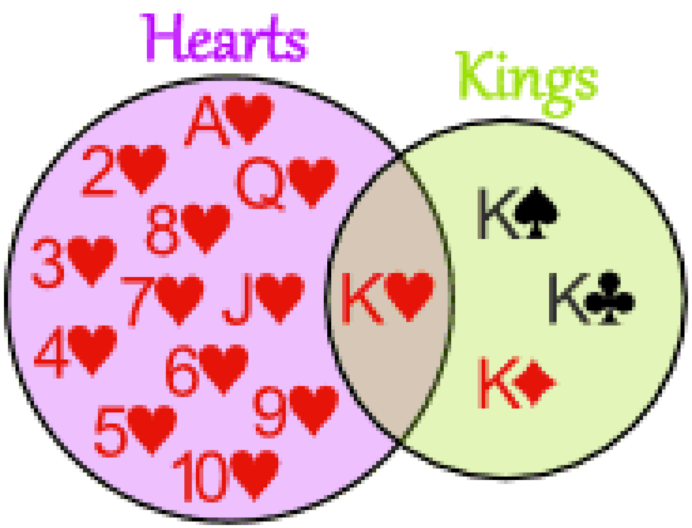
- **Mutually exclusive events** – Events that can't happen together
  - For e.g. When we draw a card from a deck of cards, the event of getting an ace and the event of getting a king are mutually exclusive



https://www.mathsisfun.com/data/probability-events-mutually-exclusive.html

Definitions (contd.)

- **Mutually exclusive events** – Events that can't happen together
  - But the event of getting a heart or the event of getting a king are not mutually exclusive since both the events have common elements



https://www.mathsisfun.com/data/probability-events-mutually-exclusive.html

Definitions (contd.)

- **Collectively exhaustive events** – Events that cover the entire sample space
  - For e.g. When a single coin is tossed, the event A of getting a tail and the event B of getting a head are collectively exhaustive
  - One of the events must occur

Definitions (contd.)

$$P(A) = \frac{\text{Number of Times 'A' Occurs}}{\text{Total Number of Possible Outcomes}}$$

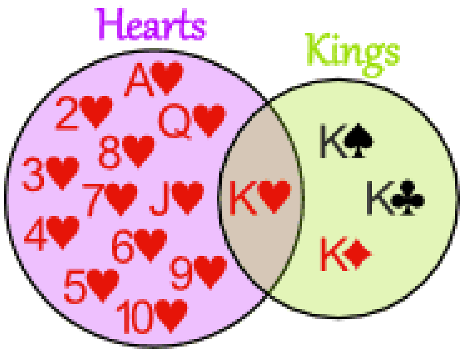
- Let us again consider the event at least one head occurs in a two coin toss
- The sample space is {HH, HT, TH, TT}
- Total number of possible outcomes in the event is the number of elements in the sample pace which is 4
- The subset of at least one head occurs is given by {HH, HT, TH}
- So the number of times out event can occur is 3
- So the required probability is 3/4
- Probability is always between 0 and 1 (0-100%)

Rules of Probability

- **Rule of subtraction** – The complement of any outcome is equal to one minus the outcome
- $P(A^C)=1-P(A)$
  - The probability of getting two heads in a two coin toss is ¼
  - The complement of the event is not getting any heads which is given by ¾
  - Now probability of getting two heads can be found by subtracting the probability of its complement from 1
  - 1-3/4 = 1/4

Rules of Probability (contd.)

- **Rule of addition** – The probability union of events A and B is equal to the probability that Event A plus the probability that Event B occurs minus the probability of intersection of Events A and B occur
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



- Calculate the probability of getting a king or getting an ace while drawing a card from a deck of cards

https://www.mathsisfun.com/data/probability-events-mutually-exclusive.html

Rules of Probability (contd.)

- Two events A and B are said to be independent of each other if the probability of one event occurring is unaffected by the occurrence or non-occurrence of the other event
  - For e.g. consider two events, a coin toss and a roll of a dice. These are independent of each other as outcome of one does not affect the other
- **Rule of Multiplication** – If two events A and B are independent, then
- P(A and B) = P(A).P(B)

Rules of Probability (contd.)

- P(A and B) = P(A).P(B)
  - Let us consider two events a coin toss and roll of a 6 faced dice
  - The sample space is given by {H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6}
  - Let event A be getting a head in the toss and event B be getting 1 or 2 in the roll of dice
  - Sub space of A and B is given by {H1, H2}
  - So P(A and B) is 2/12 = 1/6
  - P(A) = 1/2 and P(B) = 2/6 = 1/3
  - P(A).P(B) = 1/2 * 1/3 = 1/6

Conditional Probability

- Conditional Probability is the probability of an event A given that another event B has occurred
- Probability of A given B has occurred is denoted by P(A|B)

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

  - Let us consider the previous example
  - We want to find the probability of head in the toss and 1 or 2 in the roll of dice
  - Now we are given additional information that event B has occurred i.e. the outcome of the dice roll is 1 or 2

Conditional Probability (contd.)

- Probability of A given B has occurred is denoted by $P(A|B)$
  - We need to find $P(A \cap B)$
  - $A \cap B = \{H1, H2\}$
  - $P(A \cap B) = 2/12 = 1/6$
  - $P(A|B) = P(A \cap B) / P(B) = 1/2$
  - Let us verify: our new sample space is given by {H1, H2, T1, T2}
  - Total number of outcomes in the sample space has reduced to 4
  - Out of this the favourable outcomes are {H1, H2}
  - $P(A|B) = 2/4 = 1/2$

Bayes' Theorem

- Bayes' Theorem helps us to find P(A|B) if we already know P(B|A)

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

Odds

- Odds of an event is defined as the probability of that event occurring / probability of that event not occurring

$$\text{Odds} = \frac{\text{Probability of the event}}{1 - \text{Probability of event}} = \frac{P}{1-P}$$

- For example, consider a toss of a fair coin
- The odds of heads = p(Heads) / (1-p(Heads))
- 0.5/0.5 = 1 (or) 1:1
- In the roll of a fair die, the odds of getting 5 or 6
- 0.33/0.66 = ½ (or) 1:2

http://www.near-death.com/paranormal/synchronicity/the-odds-of-our-adcs-occurring-by-chance.html

Odds Ratio

- Is a ratio of two odds

$$OR = \frac{\frac{p_1}{1 - p_1}}{\frac{p_0}{1 - p_0}}$$

- In the roll of a fair die, the odds of getting 5 or 6
- 0.33/0.66 = ½ (or) 1:2
- Odds of getting 1 = 0.1666/0.83333 = 1/5
- Odds ratio = (1/2) / (1/5) = 2.5
- Odds of getting 5 or 6 is 2.5 times greater than the odds of getting 1

http://www.near-death.com/paranormal/synchronicity/the-odds-of-our-adcs-occurring-by-chance.html

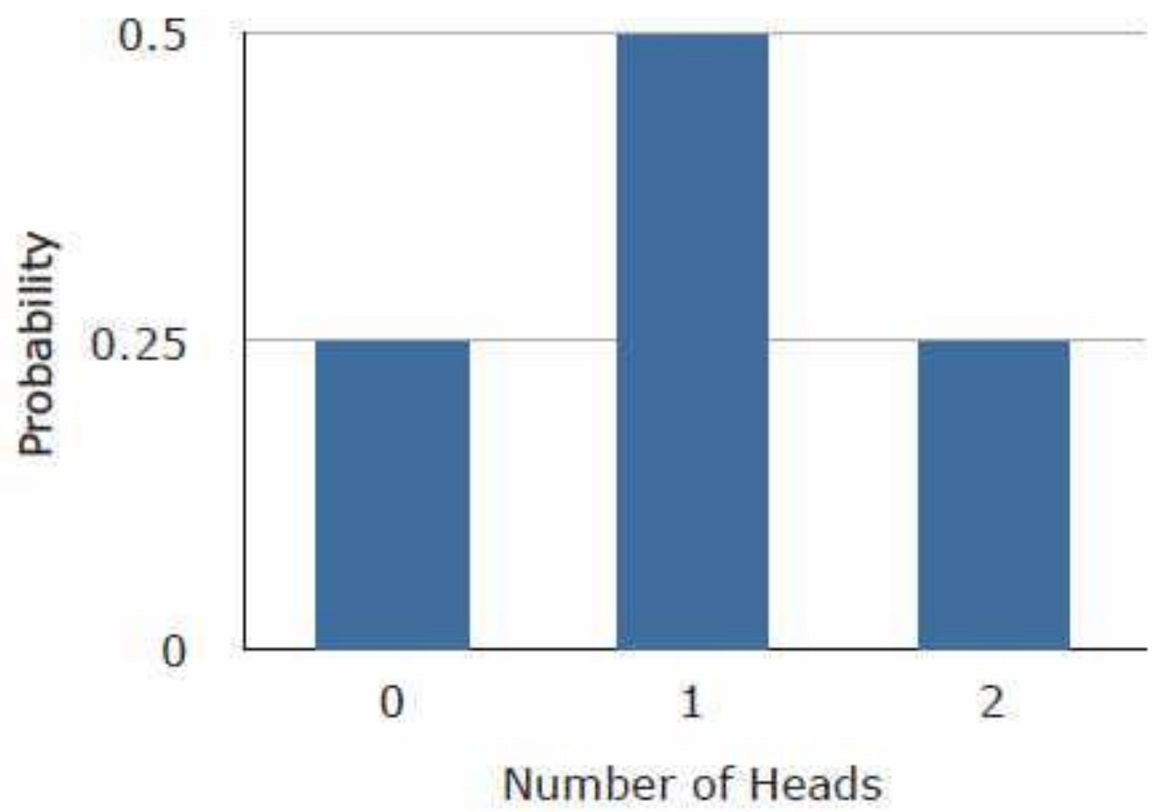
Probability Distributions

Binomial Distributions

- Consider an experiment where one coin is tossed 2 times
- Let Y denote the number of heads
- We know that:
  - $P(Y=0) = 1/4$
  - $P(Y=1) = 1/2$
  - $P(Y=2) = 1/4$

Binomial Distributions (contd.)

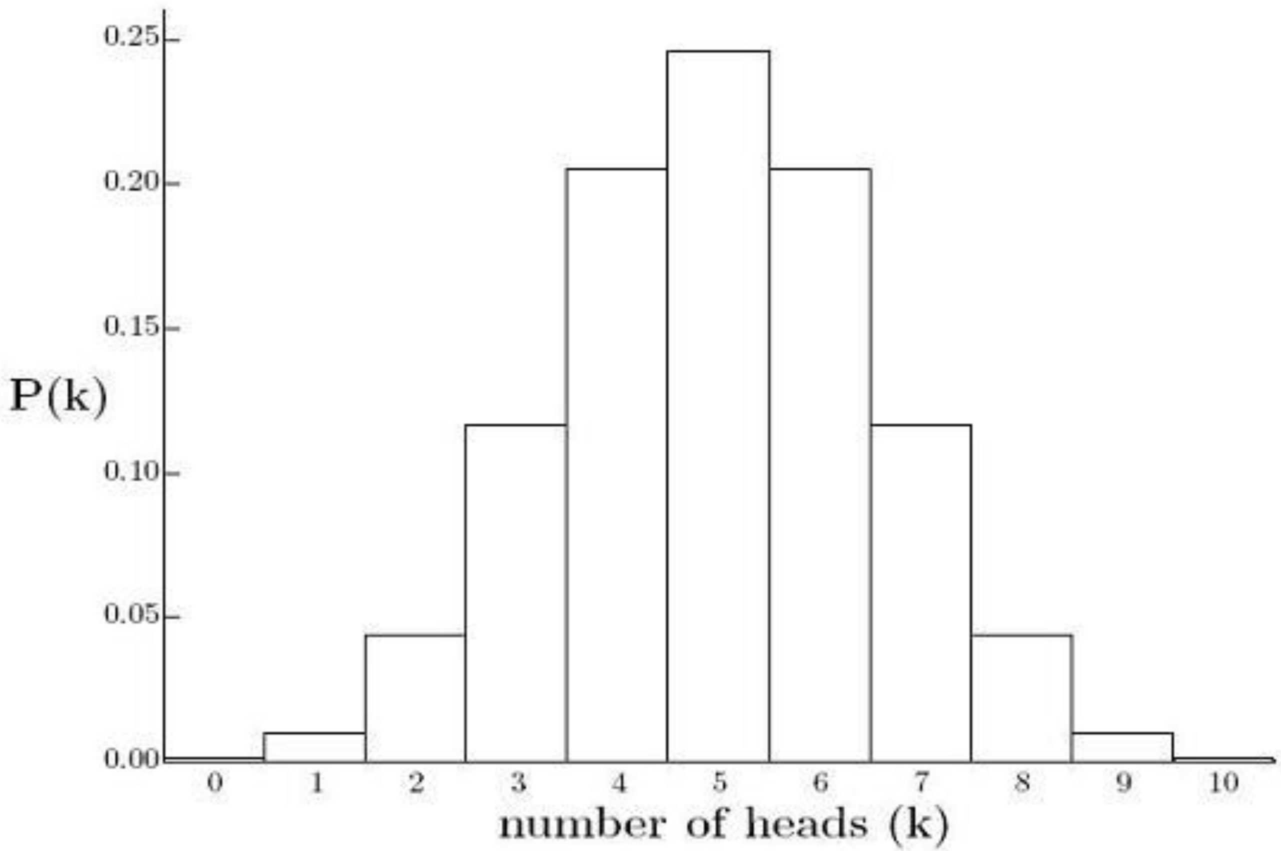
- Let us plot the probabilities
- This is called a Binomial Distribution



http://onlinestatbook.com/2/probability/binomial.html

Binomial Distributions (contd.)

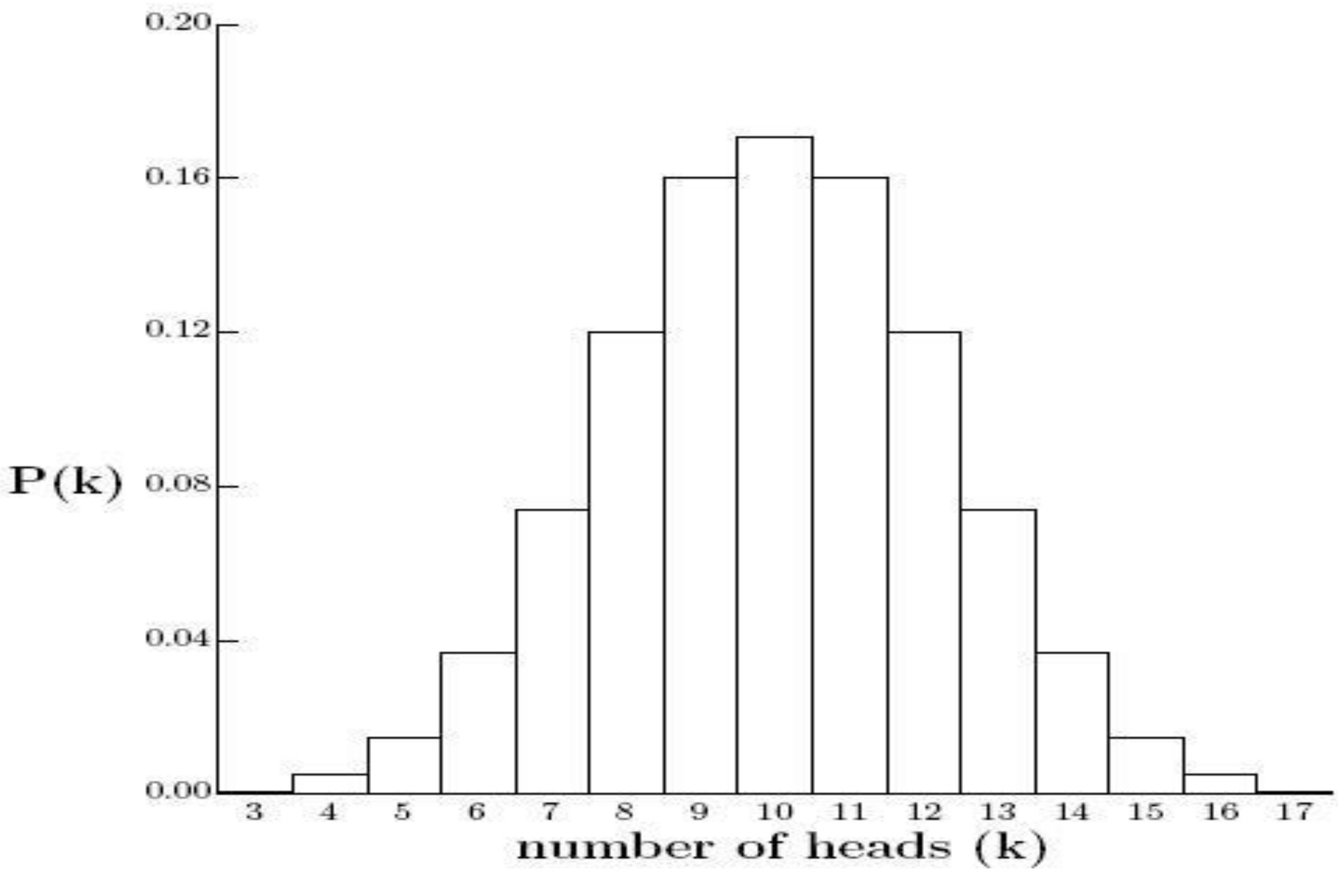
- Let us increase the number of tosses to 10



http://www.mathamazement.com/Lessons/Everyday-Math/05_Miscellaneous/05_05_Advanced-Math-Topics/normal-distribution.html

Binomial Distributions (contd.)

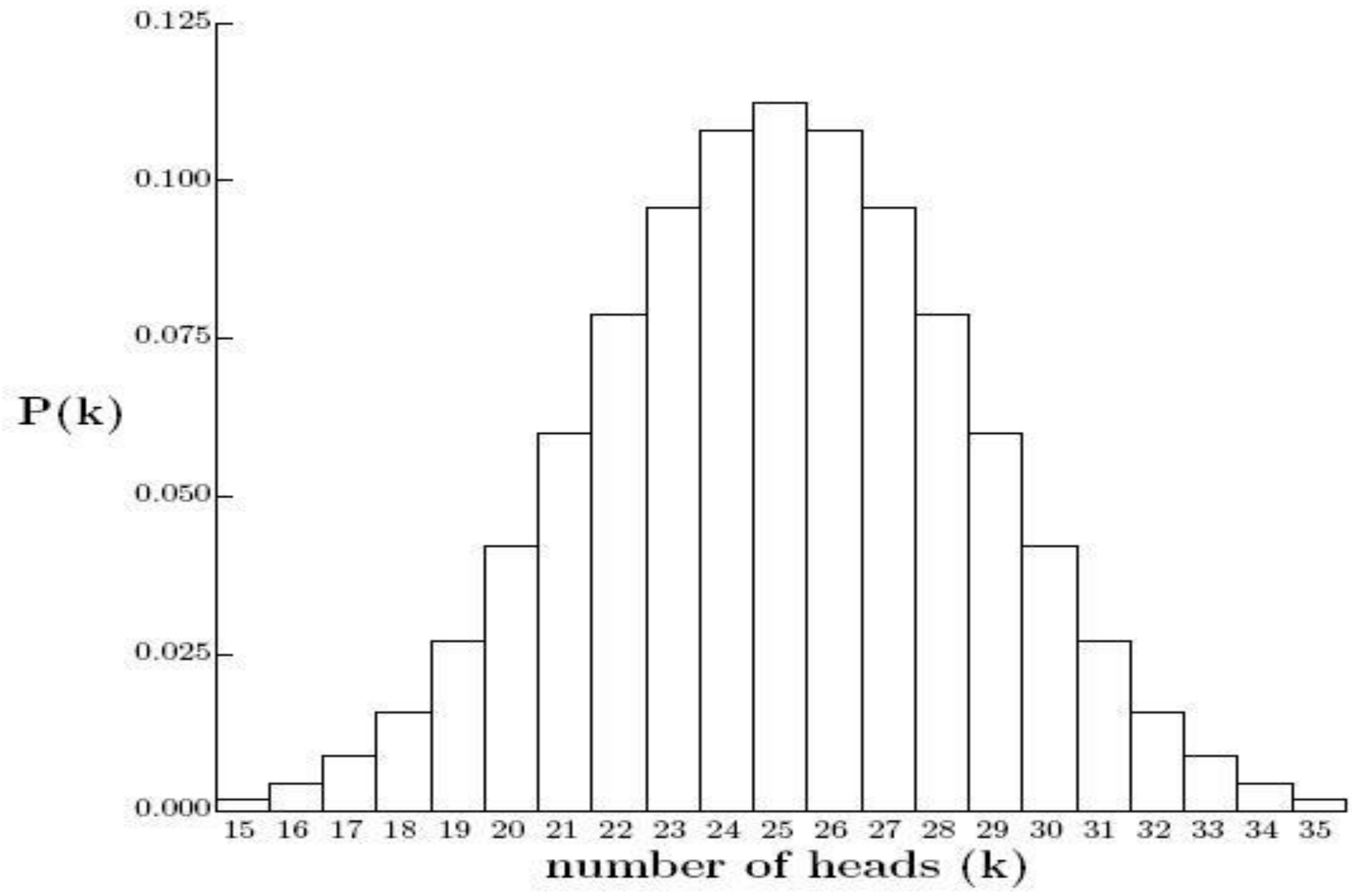
- Let us repeat the experiment with 20 tosses



http://www.mathamazement.com/Lessons/Everyday-Math/05_Miscellaneous/05_05_Advanced-Math-Topics/normal-distribution.html

Binomial Distributions (contd.)

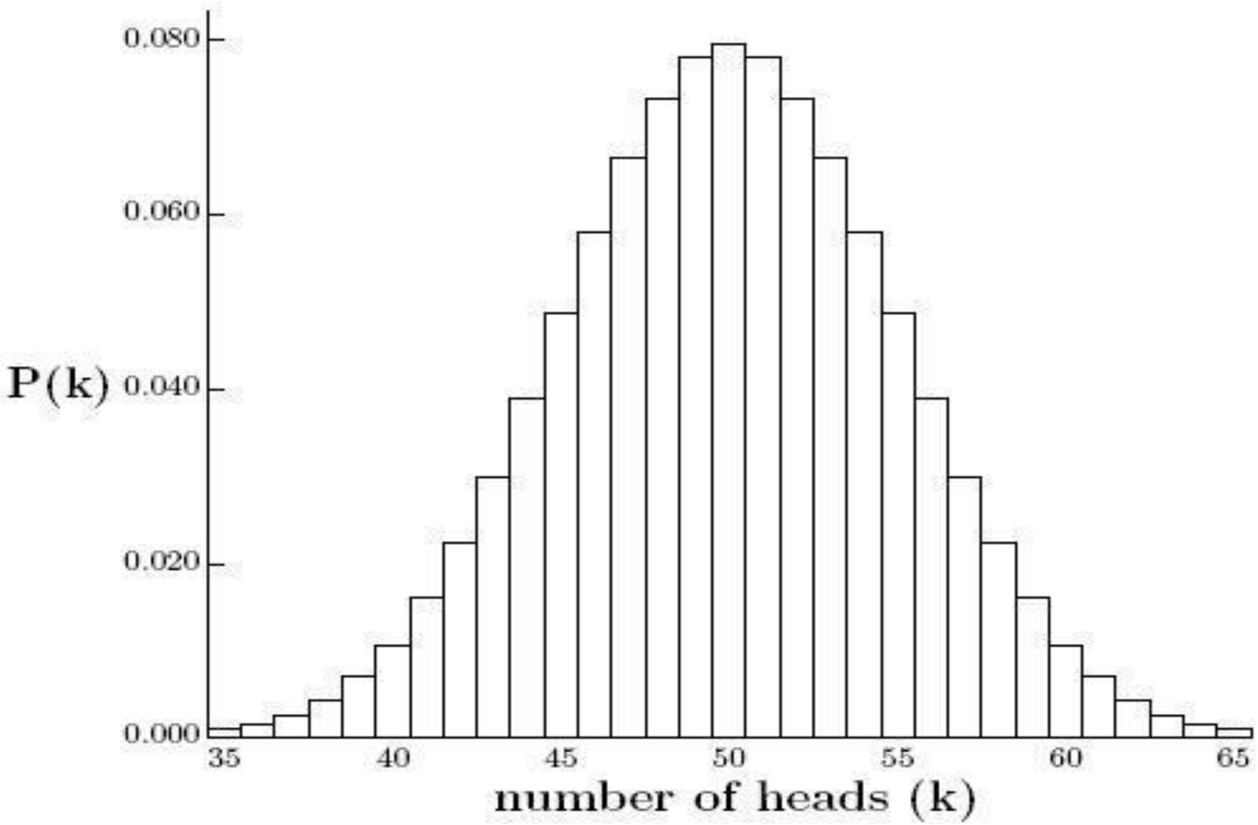
- 50 tosses



http://www.mathamazement.com/Lessons/Everyday-Math/05_Miscellaneous/05_05_Advanced-Math-Topics/normal-distribution.html

Binomial Distributions (contd.)

- 100 tosses

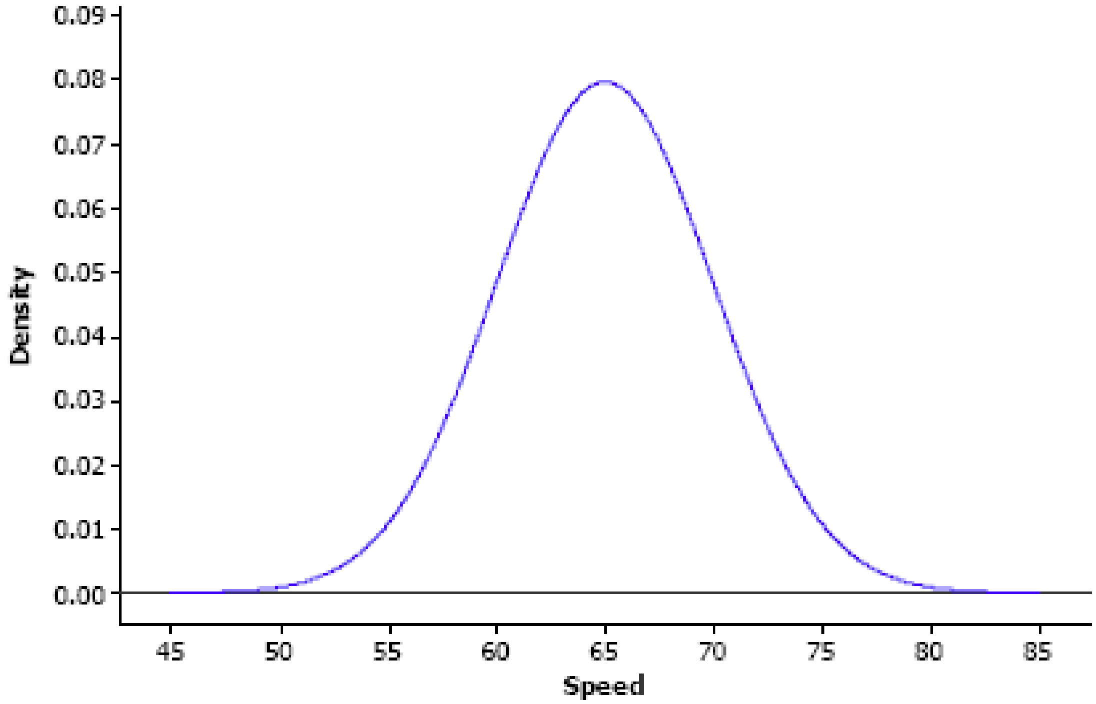


http://www.mathamazement.com/Lessons/Everyday-Math/05_Miscellaneous/05_05_Advanced-Math-Topics/normal-distribution.html

Continuous Distributions

- So far we have seen probability distributions of discrete variables
- Number of heads in our previous experiments can only be whole numbers
- Let us think of some continuous variables instead of a discrete variable like number of heads in a series of coin tosses
- Let us consider the average speed of vehicles at a point in a highway
- Let us graph the probability of speed of a random vehicle

Continuous Distributions (contd.)

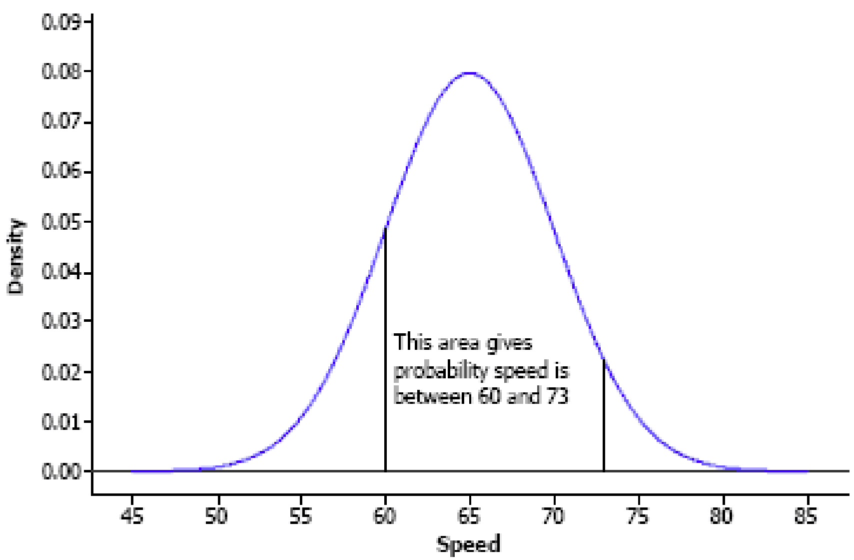


- This graph is called probability density function

http://sites.stat.psu.edu/~ajw13/stat200/mos/05_probdistr/05_probdistr_print.html

Continuous Distributions (contd.)

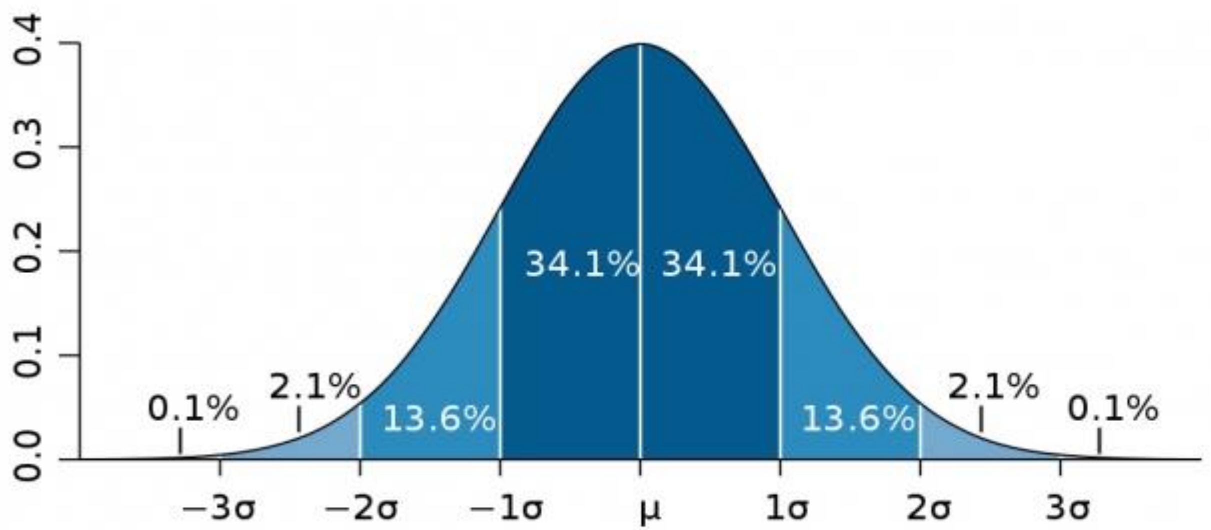
- Unlike discrete distribution, we can't directly find the probability of a point
- In continuous distributions probability can be found only for intervals
- Probability for an interval = Area under the curve in that interval



http://sites.stat.psu.edu/~ajw13/stat200/mos/05_probdistr/05_probdistr_print.html

Normal Distribution

- Popularly known as bell-shaped curve
- Normal distribution has a mean of μ and a standard distribution of σ



http://news.mit.edu/2012/explained-sigma-0209

Normal Distribution (contd.)

- Features of a Normal Distribution:
  - Mean=Median=Mode
  - Symmetric about the centre
  - Defined by two parameters μ and σ
- Many natural occurring phenomenon follow normal distribution
  - Height of a large population
  - Blood pressure

Standard Normal Distribution

- Standard Normal Distribution is a special case of Normal Distribution where mean, $\mu = 0$ and standard deviation, $\sigma = 1$