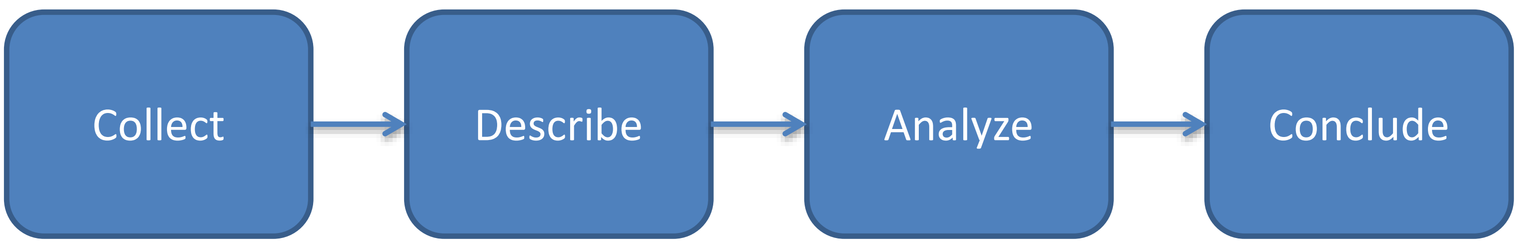
Basic Statistics

QuantInsti

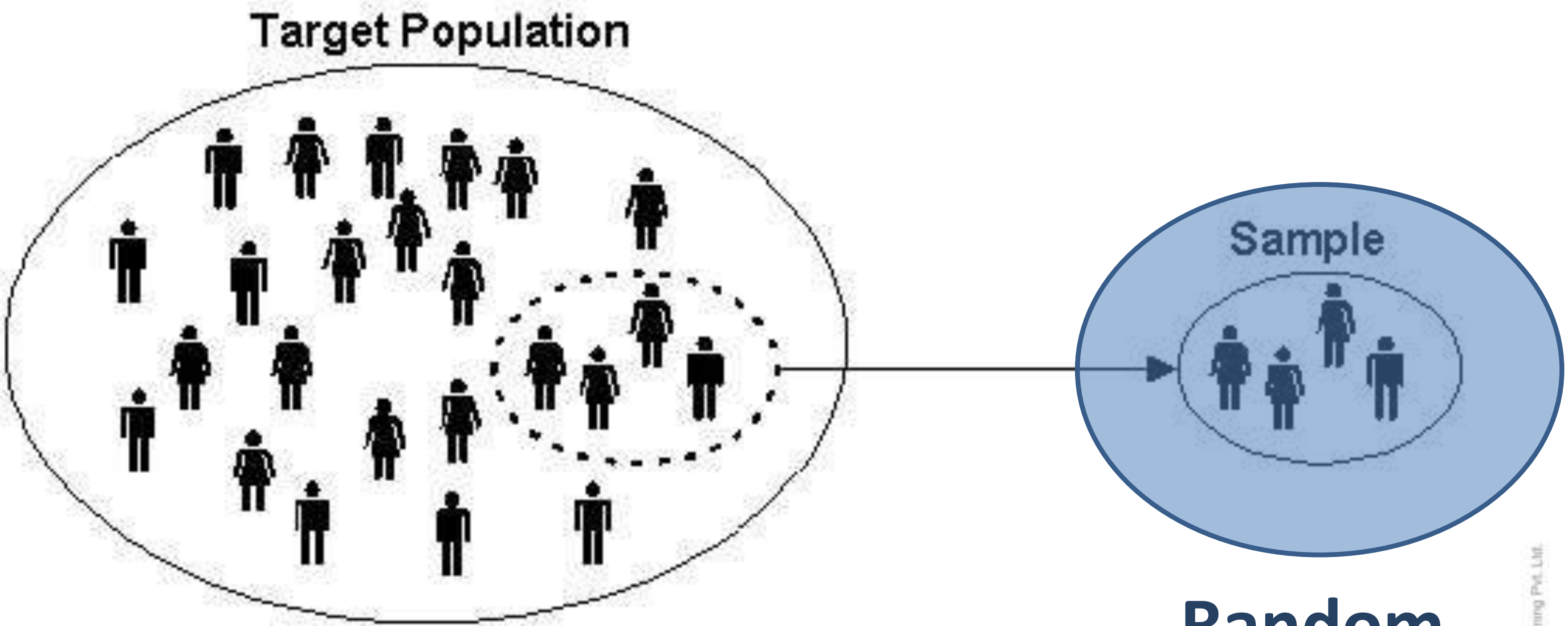
11 January 2014

What is Statistics

- To learn about something, you must first collect data
- Statistics is the art of learning from data

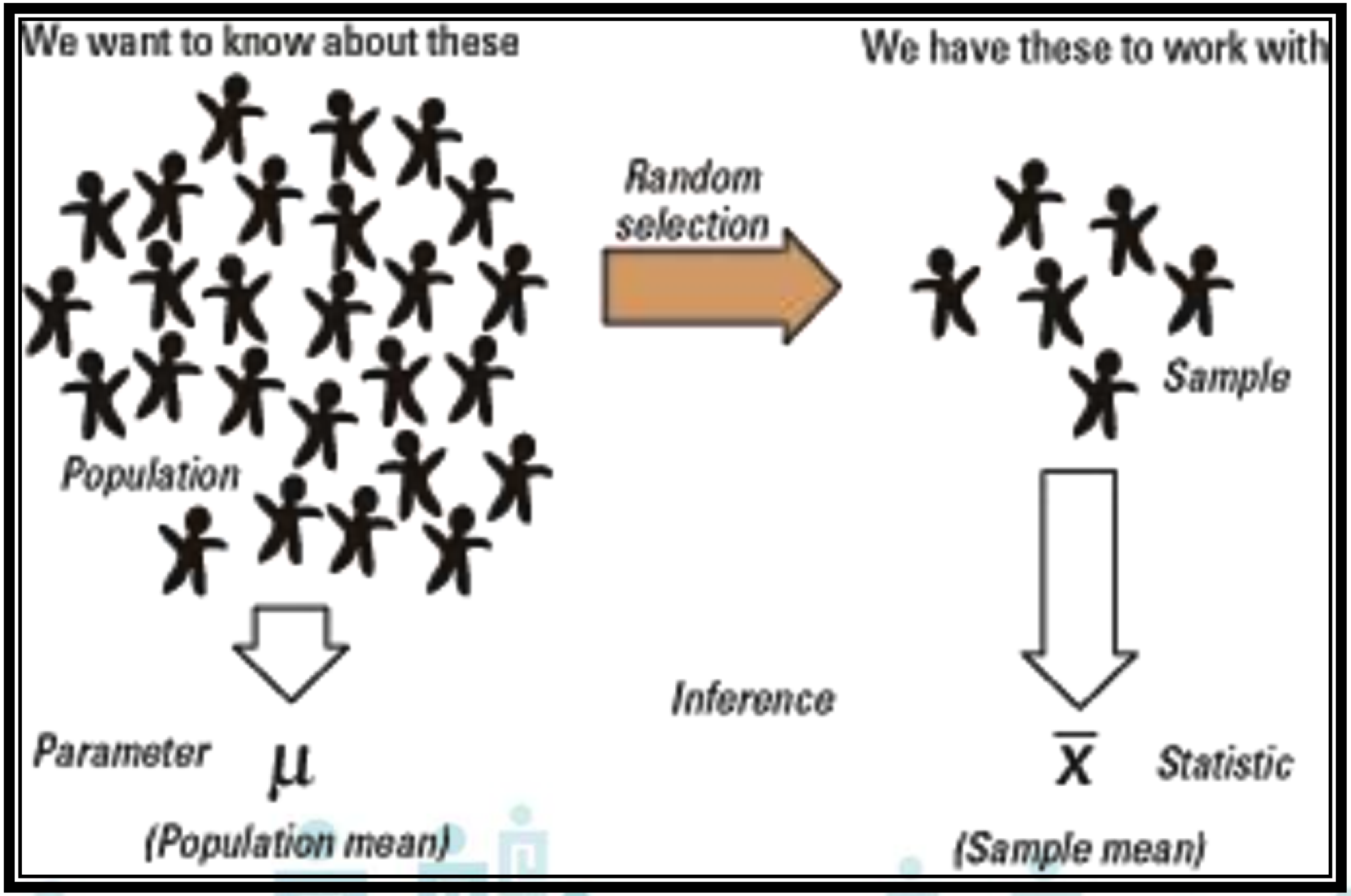


Population and Sample



Random

Parameter and Statitic

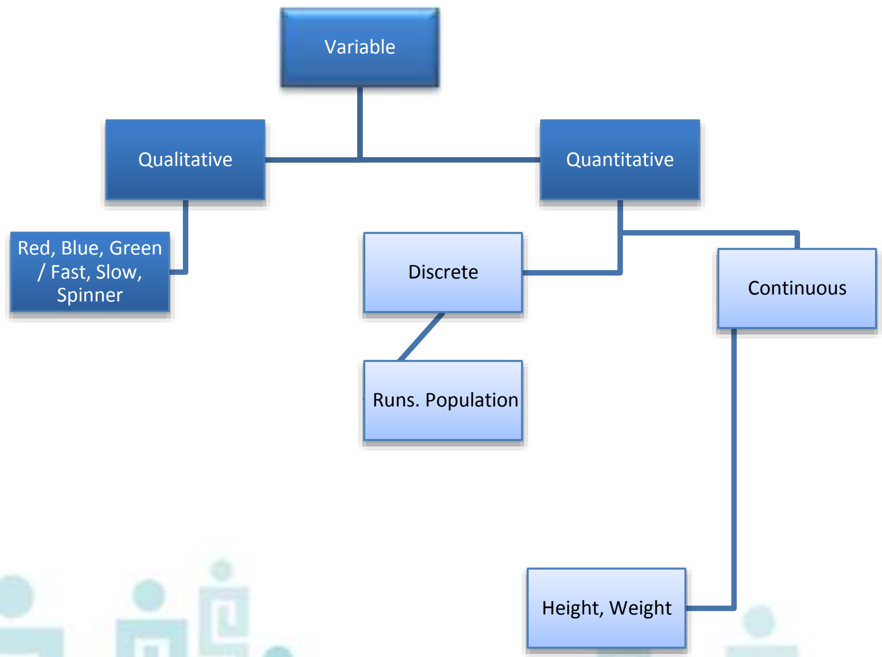


Branches of Statistics

- **Descriptive statistics:**
  - Organization, summarization, and display of data.
- **Inferential statistics:**
  - Draw conclusions about a population
  - Probability is a basic tool

Variables

- Any characteristics, number, or quantity
- Can be measured or counted
- Value can 'vary'



Random Variable

- unique numerical value with every outcome
- value will vary from trial to trial
  - E.g. outcome of a coin toss, H/T

Classification of Data

According to number of variables

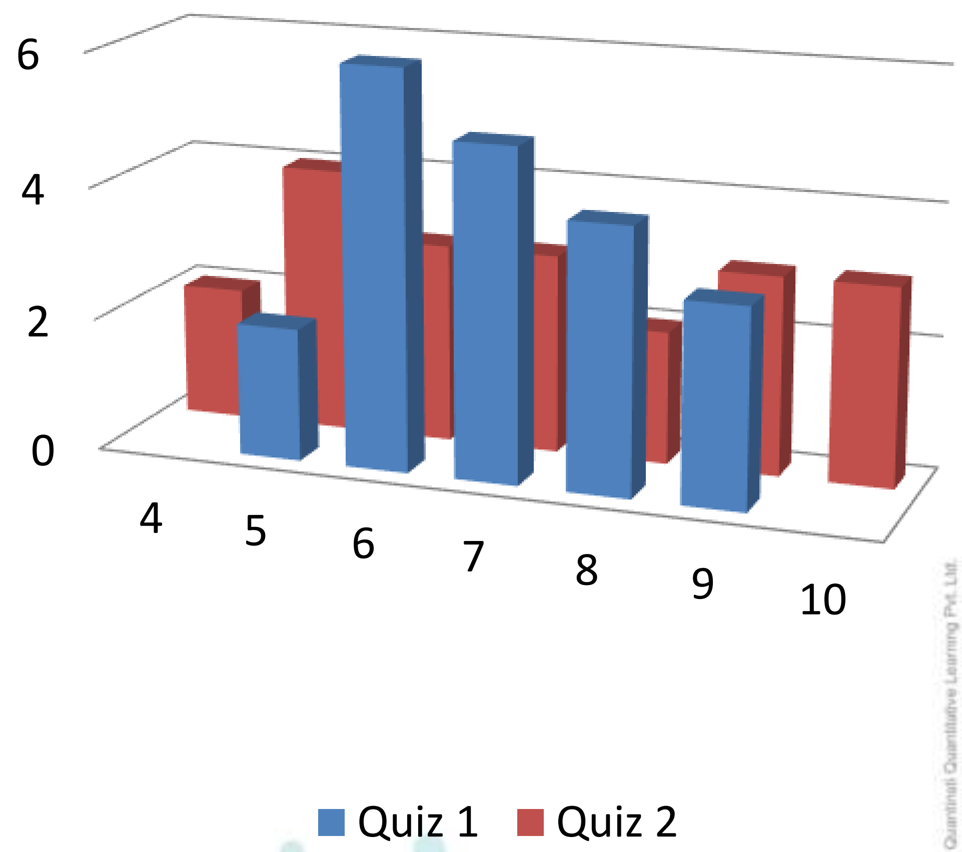
- Univariate
  - only one variable, *e.g. average weight*
- Bivariate
  - two variables, *e.g. relationship between the height and weight*

Central Tendency

- Mean
  - Sum of observations / number of observations
- Median
  - Middle value of observations
- Mode
  - Most frequently occurring value

Variability

- How 'spread out' the data is

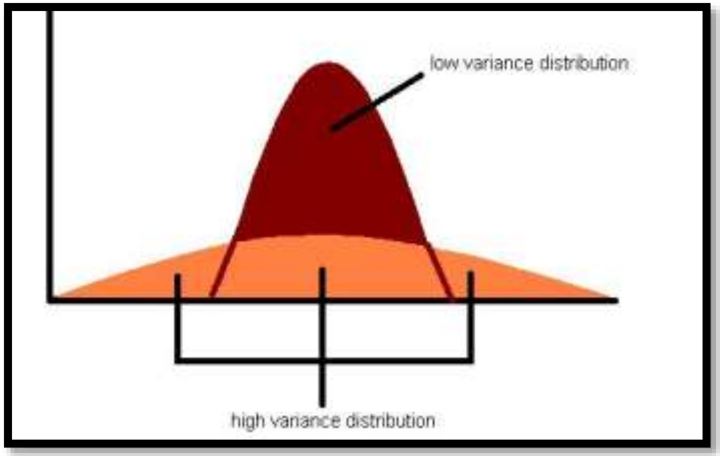


Range and Quartile

- Range
  - Difference between largest and smallest value
- Quartile
  - Divide a rank-ordered data into four equal parts

Variance

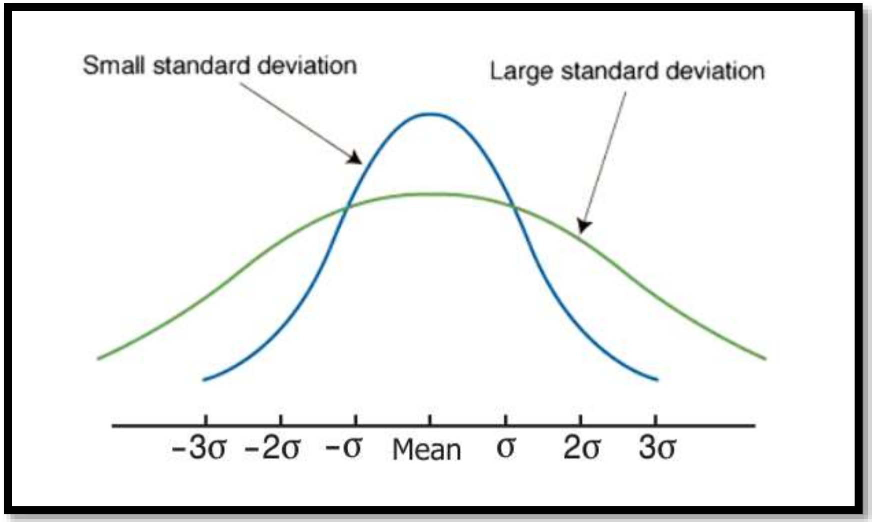
- Measure of variation or diversity in a distribution



- For a population,
  - $\sigma^2 = \Sigma (X_i - \mu)^2 / N$

- For a sample,
  - $s^2 = \Sigma (x_i - x)^2 / (n - 1)$

Standard Deviation

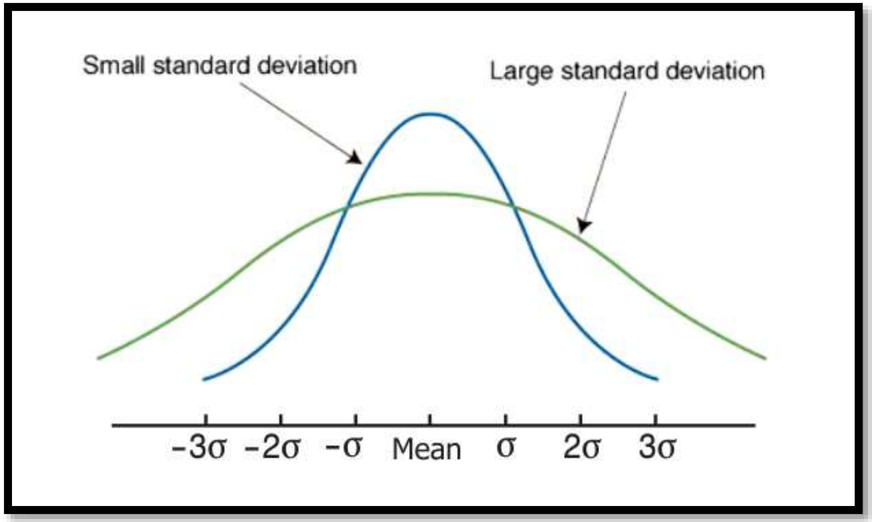


- For a population
  - $\sigma \;=\; \text{sqrt}[\,\sigma^2\,] \;=\; \text{sqrt}\,[\,\Sigma\,(\,X_i - \mu\,)^2 / N\,]$

- For a sample
  - $s \;=\; \text{sqrt}[\,s^2\,] \;=\; \text{sqrt}\,[\,\Sigma\,(\,x_i - x\,)^2 / (\,n - 1\,)\,]$

Standard Scores (z-Scores)

- how many standard deviations from the mean
  - $z = (X - \mu) / \sigma$



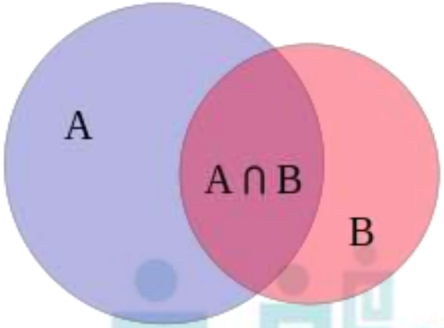
Data Patterns

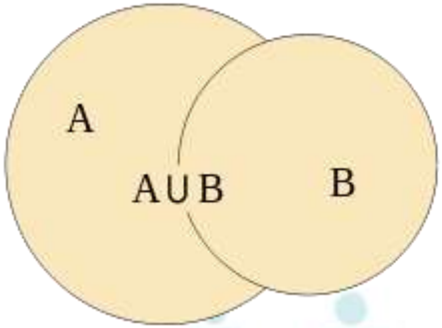
- Spread
- Shape
  - symmetry, number of peaks, skewness, and uniform

- Kurtosis

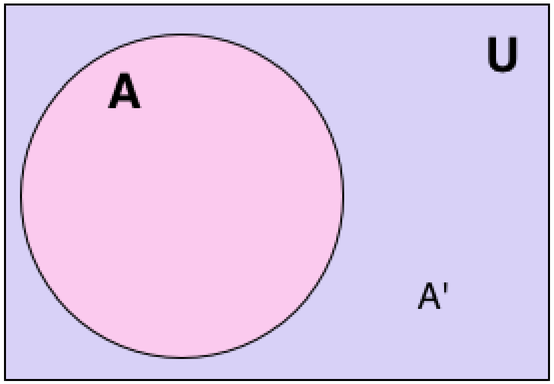
Basics of Set Theory - I

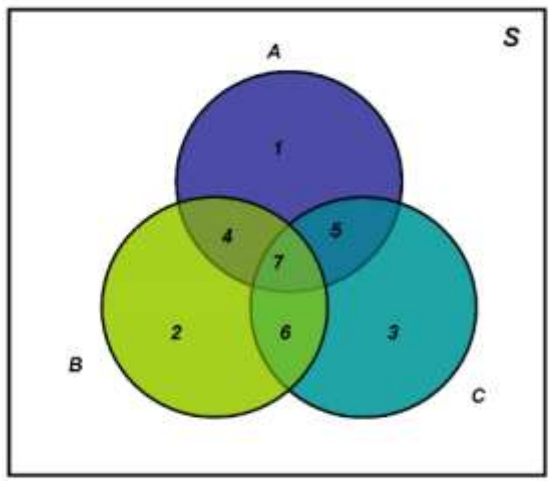
- Set
  - a well-defined collection of objects
    - E.g outcomes of rolling a dice, D={1,2,3,4,5,6}
    - Let A be rolls of even no, A={2,4,6};
      - Subset A is subset of D
    - Let B be rolls greater than 3; B = {4,5,6}
      - Union of sets, A ∪ B = {2, 4, 5, 6}
      - Intersection of sets, A ∩ B = {4, 6}





Basics of Set Theory - II





- $A \cap B$ = regions 4 and 7.
- $B \cap C$ = regions 6 and 7.
- $A \cup C$ = regions 1, 3, 4, 5, 6 and 7.
- $B' \cap A$ = regions 1 and 5.
- $A \cap B \cap C$ = regions 7.
- $(A \cup B) \cap C'$ = regions 1, 2 and 4.

Concepts of Probability

- Many events can't be predicted with total certainty
  - How **likely** they are to happen
  - the concept of **probability**

Tossing a coin

Probability of
coin landing Head is ½
coin landing Tail is ½

Throwing Dice

Probability of
any of {1,2,3,4,5,6} is 1/6

Some Definitions

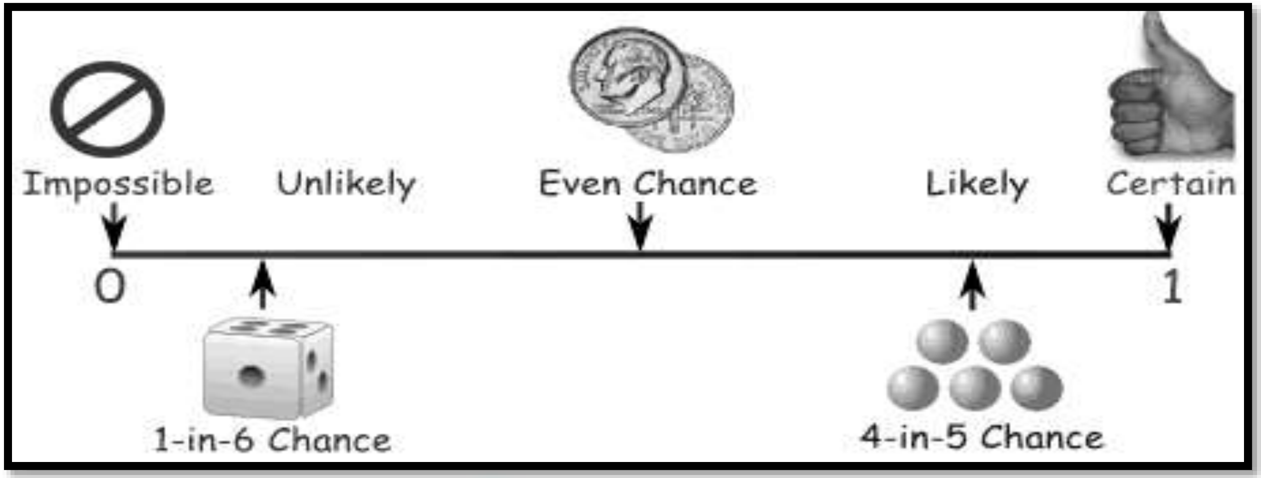
- Sample Space (S)
  - all possible outcomes of a statistical experiment
  - e.g. {H,T}, {1,2,3,4,5,6}
- Events (E)
  - sets, or collections, of outcomes
  - e.g. roll and even number {2,4,6}
  - Mutually exclusive or not
- Sample Point
  - One possible outcome

Probability Model

- Probability function P
  - Assigns a number to each outcome
  - Cannot be negative, i.e. P≥0
  - Sum of all probabilities in sample space = 1, i.e. ∑P =1
  - For event A,
    - $P(A) = \frac{\text{Number of outcomes favourable to A}}{\text{Total number of outcomes for the experiment}}$
  - E.g. A coin is tossed twice. What is the probability that at least on head occurs?

Rules of probability

- 0≤P(x) ≤1.
- ∑ P(x) = 1
- the probability of an event E is the sum of the probabilities of the outcomes in E:
  - P (E) = P x ∈ E P(x)
- P (A∪B) =P (A) +P (B) if A∩B=∅



Conditional Probability

- The probability of an event B occurring when it is known that some event A has occurred is called a conditional probability
- P (B | A)
  - **P (B|A) =P (A∩B) / P (A)**
  - *Example*
    - Roll a dice. What is the chance that you'd get a 6, given that you've gotten an even number?
    - *Solution*: Let A be the event of even numbers, and B of 6.
    - A ={ 2;4;6}; P(A) =1/2; (1)
    - B = { 6}; P(B) =1/6 ; (2)
    - A∩B={6}; P(A∩B) =1/6; (3)
    - P(B|A) =P(A∩B)/P(A)=1/3

Rules of Probability

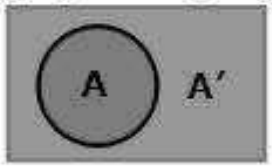
- Addition
  - Event A **or** Event B occurs
    - **$P(A \cup B) = P(A) + P(B) - P(A \cap B))$**
      - Mutually exclusive
- Subtraction
  - Event A will **not** occur
    - **$P(A) = 1 - P(A')$**
- Multiplication
  - Event A **and** Event B
    - **$P(A \cap B) = P(A) * P(B|A)$**
      - Independent events

Venn Diagrams

Venn Diagrams

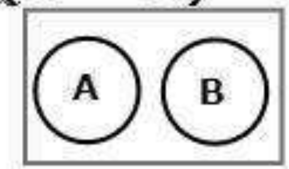
Complement

$P(A) + P(A') = 1$



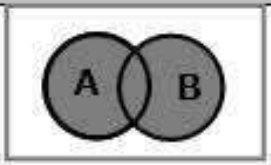
Mutually Exclusive

$P(A \cap B) = 0$



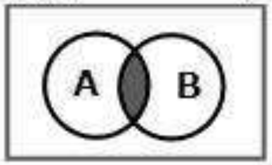
Union

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Intersection

$P(A \cap B)$



Independent

$P(A) \times P(B) = P(A \cap B)$

Bayes' Theorem

- $P(A_k|B) = \dfrac{P(Ak \cap B)}{P(A_1 \cap B) + P(A_2 \cap B) + ... + P(An \cap B)}$
  - Useful for calculating conditional probabilities
    - $A_1, A_2, ... , A_n$ mutually exclusive, form S
    - B is even from sample space, P(B)>0
  - Also written as
- $P(A_k|B) = \dfrac{P(Ak)P(B|Ak)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+...+P(An)P(B|An)}$

When to Apply Bayes' Theorem

- Set of mutually exclusive events
  - $\{ A_1, A_2, . . . , A_n \}$.
- There exists an event B
  - $P(B) > 0$.
- Want to compute a conditional probability
  - $P(A_k \mid B)$.
- You know either
  - $P(A_k \cap B)$ for each A_k
    - or
  - $P(A_k)$ and $P(B \mid A_k)$ for each A_k

Example from Handout

- Formulate the problem:
  - Mutually exclusive events:
    - A_1 : It rains
    - A_2 : It does not rain
  - Event B
    - B: Weatherman predicts rain
  - Goal:
    - Probability of rain, given weatherman predicts rain, i.e.
    - $\mathbf{P(A_1 \mid B)}$

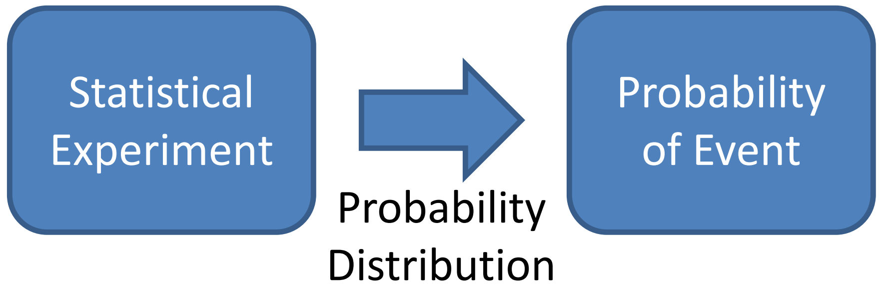
Solution

- We use second form of Bayes' theorem,

- $P(A_1|B) = \dfrac{P(A_1)\ P(B|A_1)}{P(A_1)\ P(B|A_1) + P(A_2)\ P(B|A_2)}$

  – P(A1) = 5/365 =0.0136985
  – P(A2) = 360/365 = 0.9863014
  – P(B | A1) = 0.9
  – P(B | A2) = 0.1

Probability Distributions



- maps outcome of a statistical experiment with probability of occurrence
- random variable X,
  - P(X) = 1 => probability that X is 1
- E.g. coin flipped twice
  - Outcomes: {HH, HT, TH, TT}
  - X = number of heads

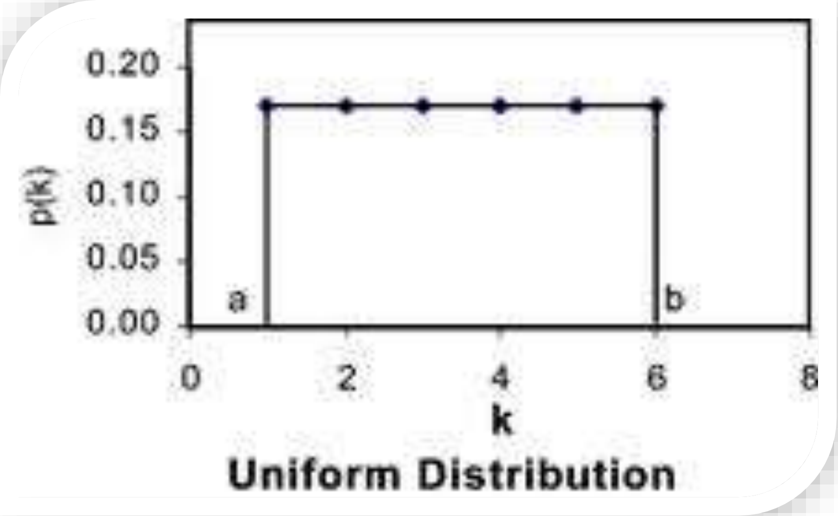
Number of Heads	Probability
0	0.25
1	0.50
2	0.25

Cumulative Probability

Number of Heads	Probability P(X = x)	Cumulative Probability: P(X ≤ x)
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00

Uniform Probability Distribution

- all values occur with equal probability
  - $P(X = x_k) = 1/k$



Discrete Probability Distributions

- random variable is a discrete variable
  - Binomial probability distribution
    - Each trial results in two possible outcomes
    - Example, flip a coin *n* number of times
  - Poisson probability distribution
    - Outcomes can be classified as successes or failures
    - Average number of successes is known
    - Example, average number of homes sold by a Realty Company

Binomial Distribution

- **x**: The number of successes that result from the binomial experiment.
- **n**: The number of trials in the binomial experiment.
- **P**: The probability of success on an individual trial.
- **Q:** The probability of failure on an individual trial. (This is equal to 1 - P.)
- **b(x, n, P):** Binomial probability - the probability that an n-trial binomial experiment results in exactly x successes, when the probability of success on an individual trial is P.
- **$^{n}C_{r}$**: The number of combinations of n things, taken r at a time.

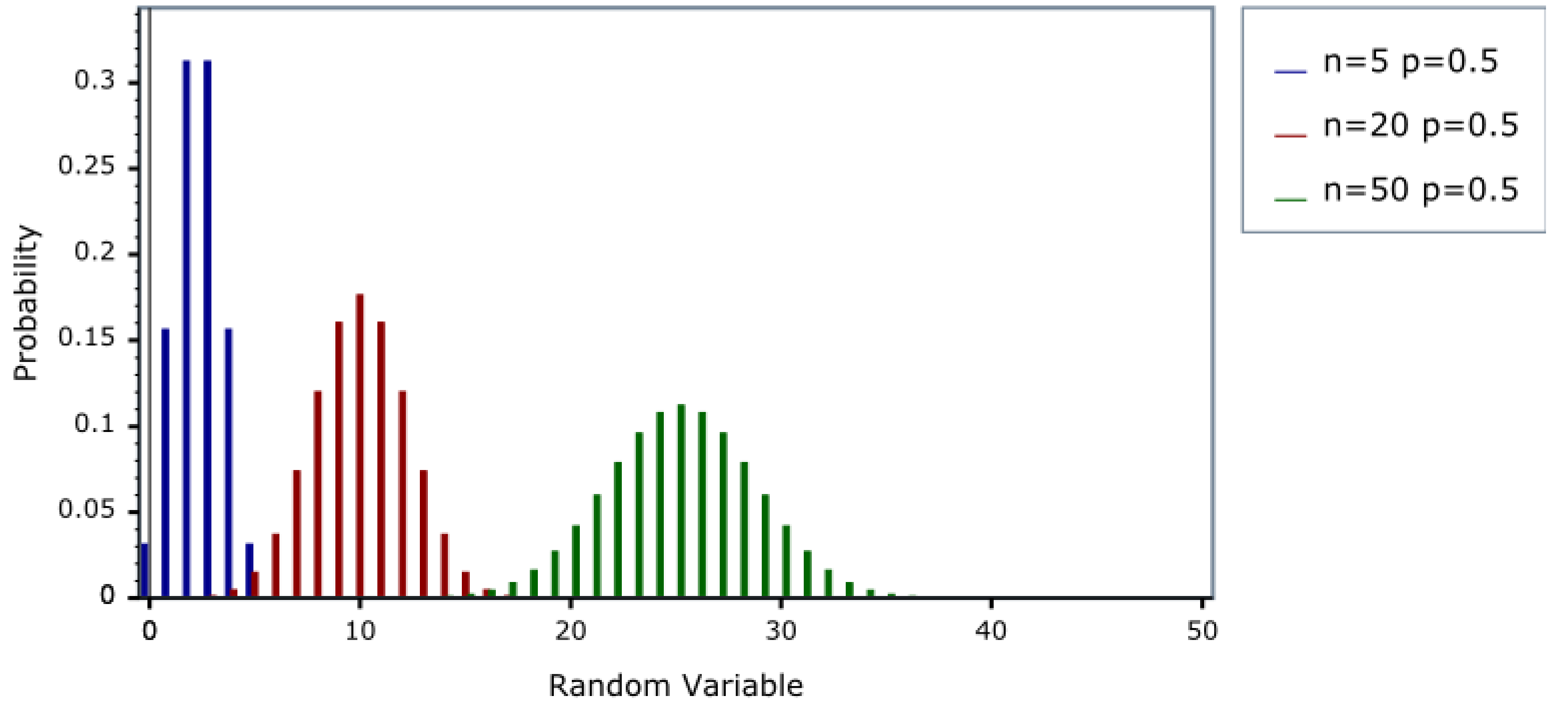
$$b(x, n, P) = {}^{n}C_{x} * P^{x} * (1 - P)^{n-x}$$

Binomial Distribution

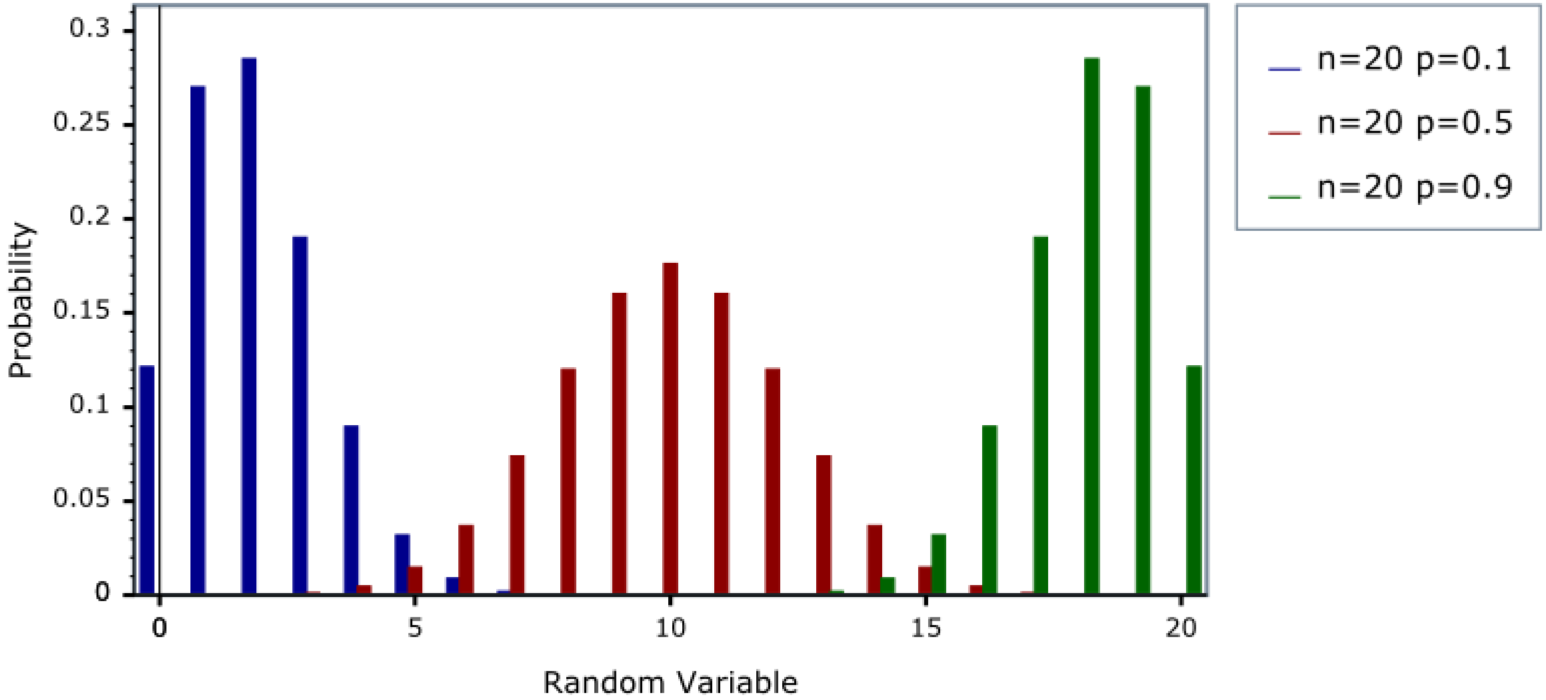
Number of heads	Probability
0	0.25
1	0.50
2	0.25

The mean of the distribution (μ_x) is equal to n * P .
The variance (σ^2_x) is n * P * (1 - P)
The standard deviation (σ_x) is sqrt[n * P * (1 - P)].

Binomial Distribution, P=0.5



Binomial Distribution, n = 20



Cumulative Binomial Probability

- probability that the binomial random variable falls within a specified range
- example, the cumulative binomial probability of obtaining 45 or fewer heads in 100 tosses of a coin

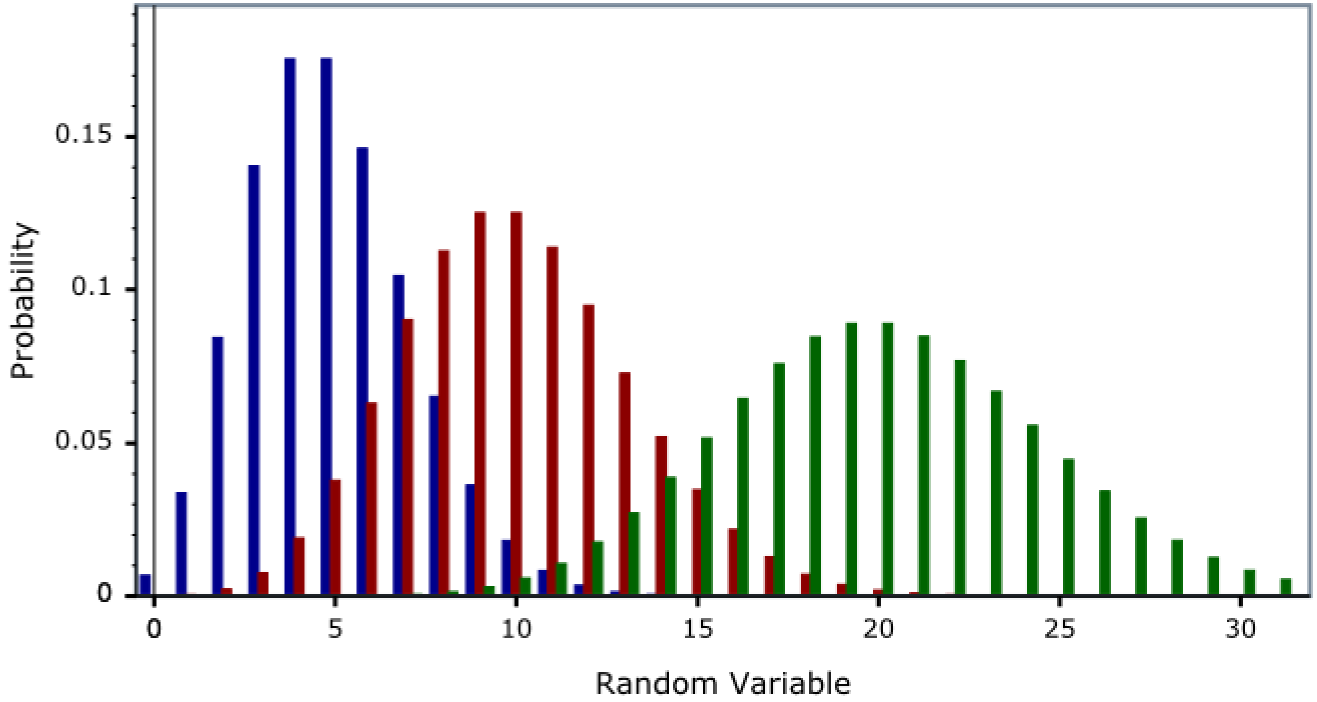
Poisson Distribution

- Outcomes that can be classified as successes or failures.
- Average number of successes in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
- The probability that a success will occur in an extremely small region is virtually zero

- **e:** A constant equal to approximately 2.71828.
- **λ or μ**: The mean number of successes that occur in a specified region.
- **x:** The actual number of successes that occur in a specified region.
- **P(x; λ or μ):** The Poisson probability that exactly x successes occur in a Poisson experiment, when the mean number of successes is μ.

$$P(x; \lambda) = (e^{-\lambda}) \lambda^{x}) / x!$$

Poisson Distribution



Poisson Distribution

Example

- The average number of homes sold by the Realty Company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

- *Solution:*

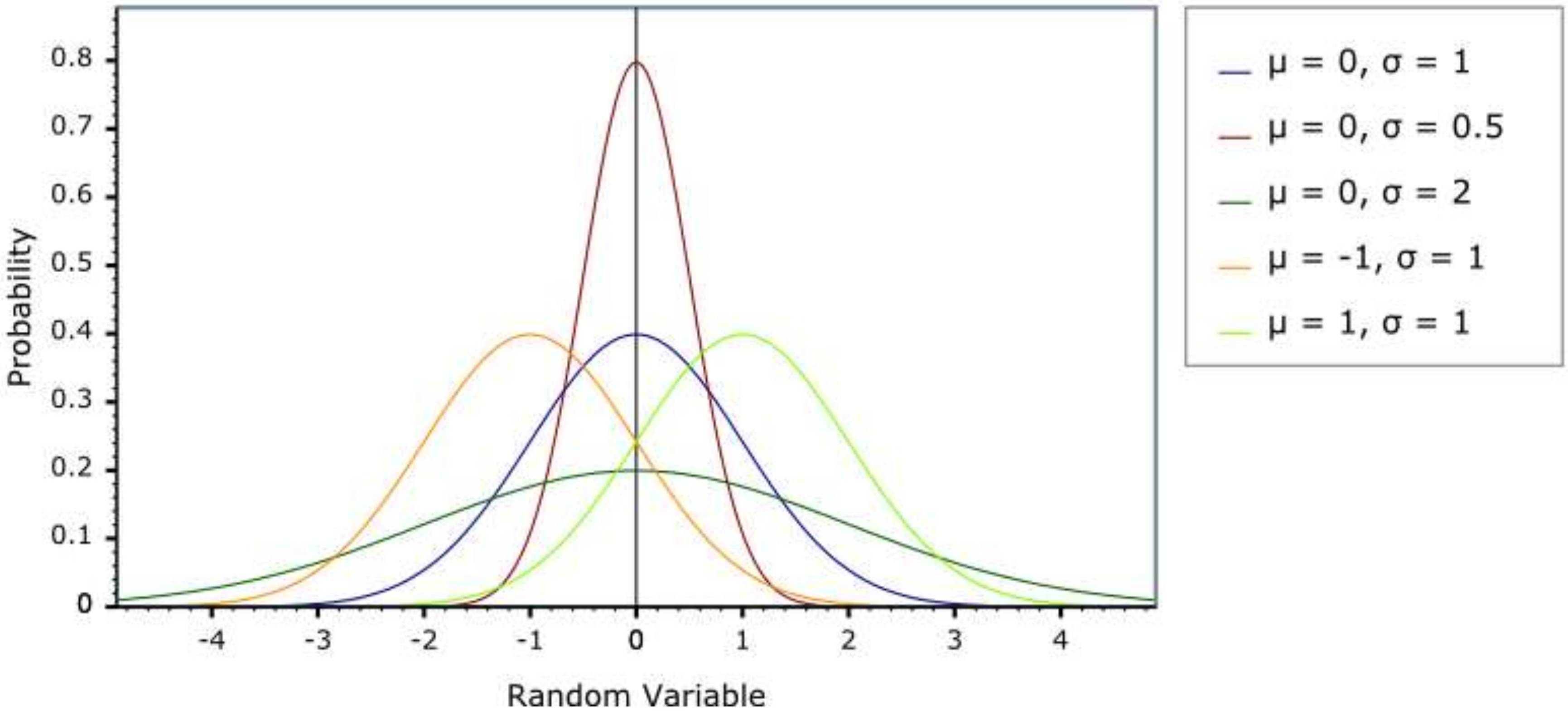
- $\mu = 2$;
- $x = 3$;
- $e = 2.71828$;

- $P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$

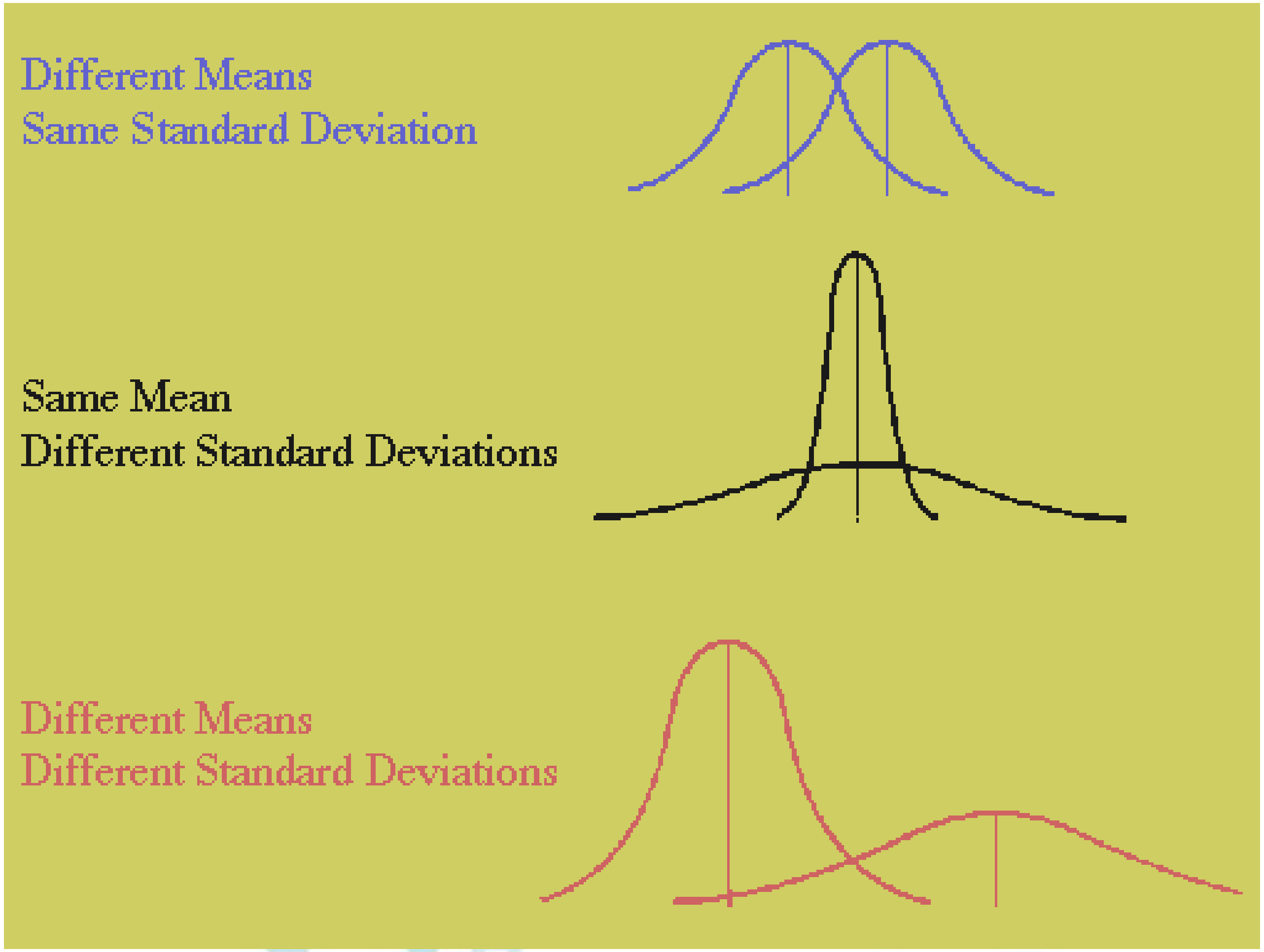
- $P(3; 2) = (2.71828^{-2}) (2^{3}) / 3!$
- $P(3; 2) = (0.13534) (8) / 6$
- $P(3; 2) = 0.180$

Normal Distribution

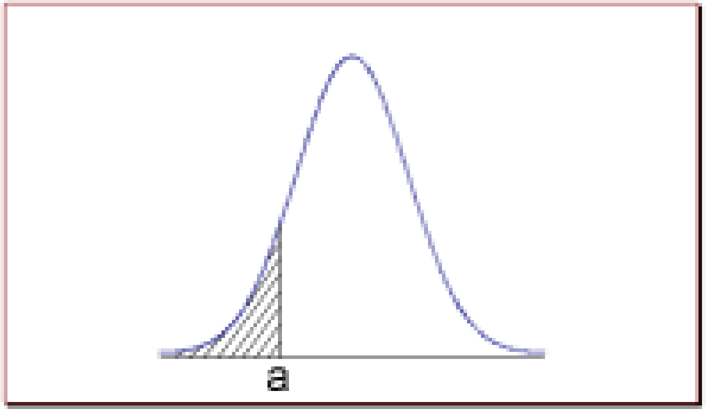
$$X = \{ 1/[\sigma * \text{sqrt}(2\pi)] \} * e^{(-(x-\mu)^{\wedge}2 /2 /\sigma^{\wedge}2)}$$



Normal Curve

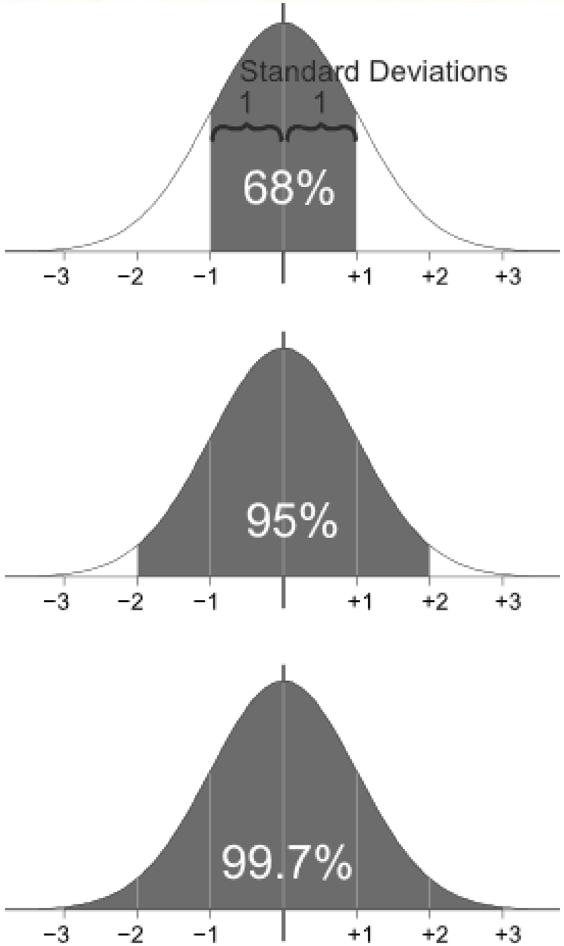


Probability and the Normal Curve

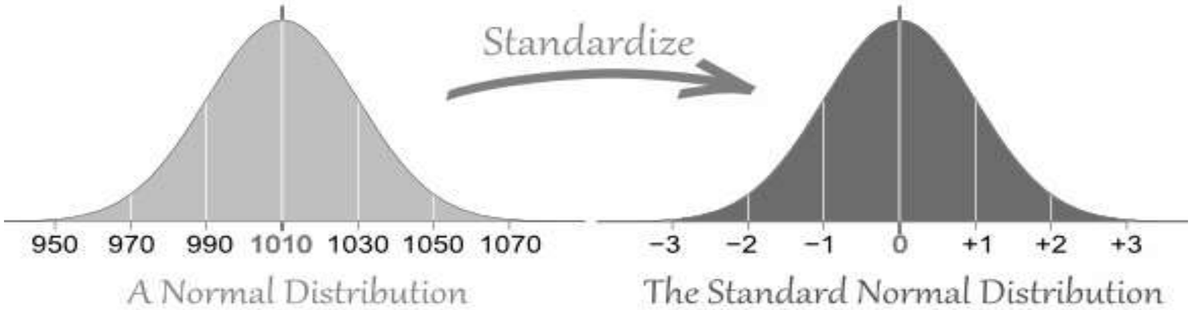


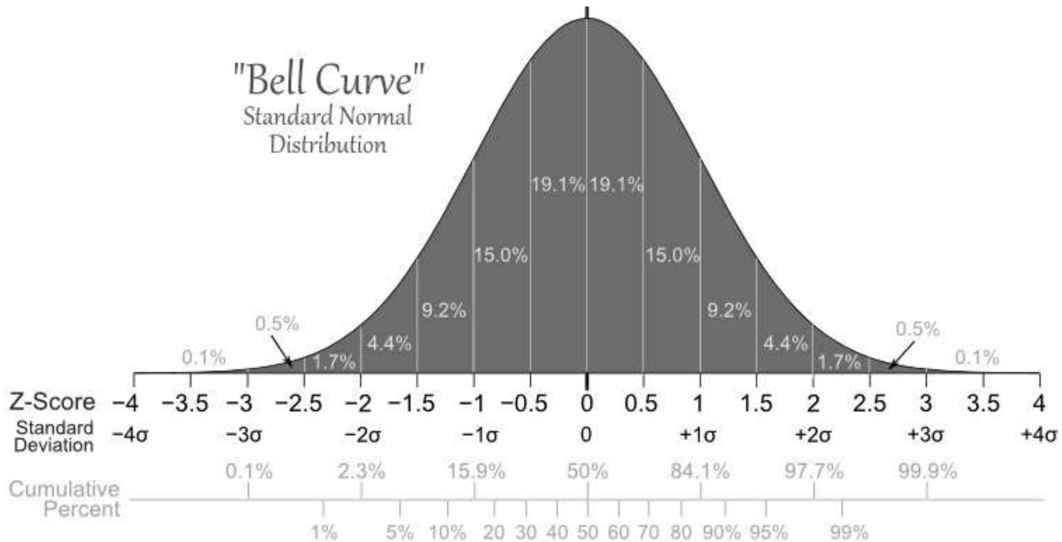
- The total area under the normal curve 1
- The probability that a normal random variable *X* equals any particular value is 0
- The probability that a random variable assumes a value between *a* and *b* is equal to the **area under the density function bounded by *a* and *b***.
- The probability that *X* is greater than *a* equals the area under the normal curve bounded by *a* and plus infinity
- The probability that *X* is less than *a* equals the area under the normal curve bounded by *a* and minus infinity

68-95-99.7 rule

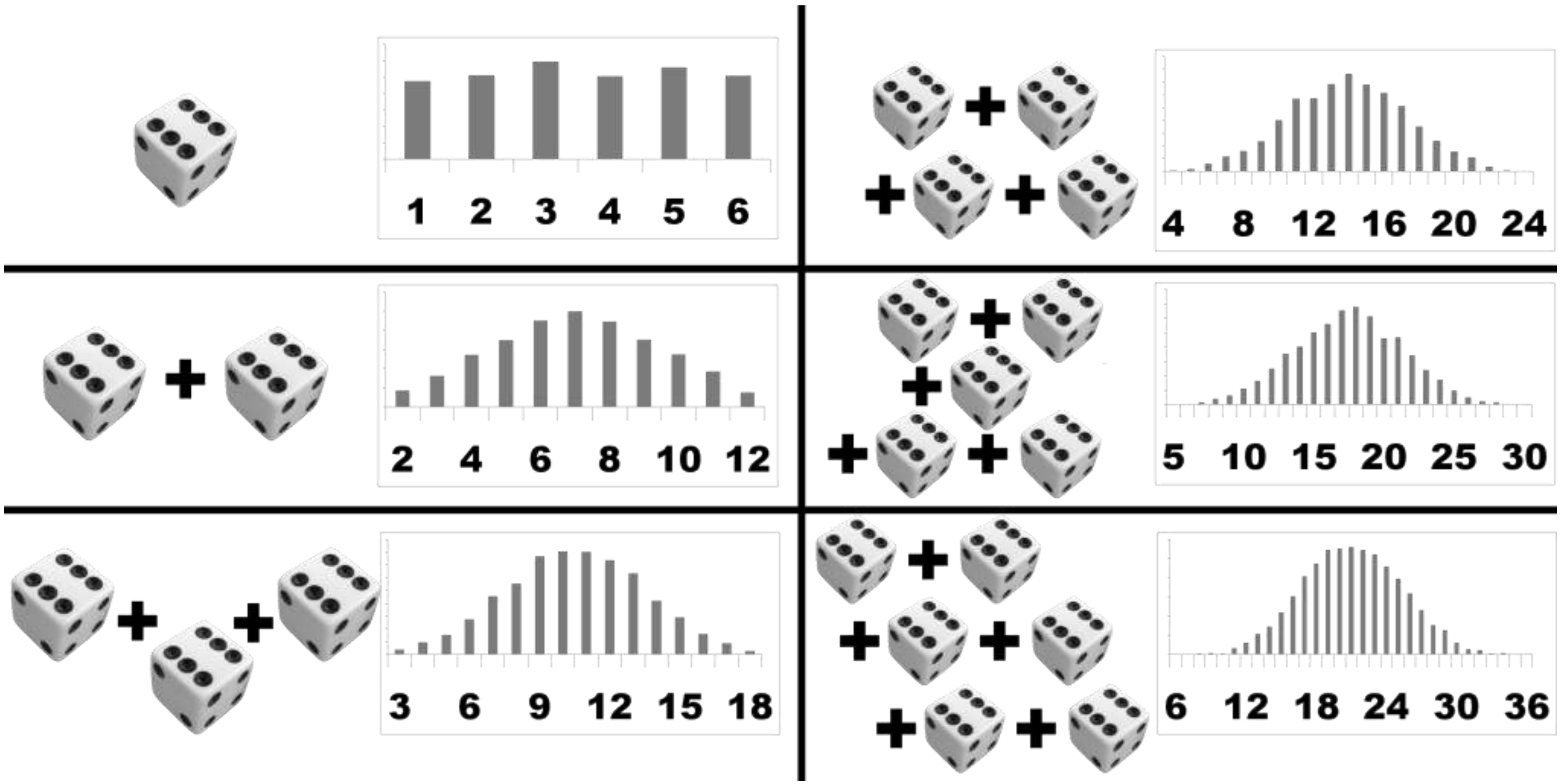


Standard Normal Distribution





Central Limit Theorem



Central Limit Theorem

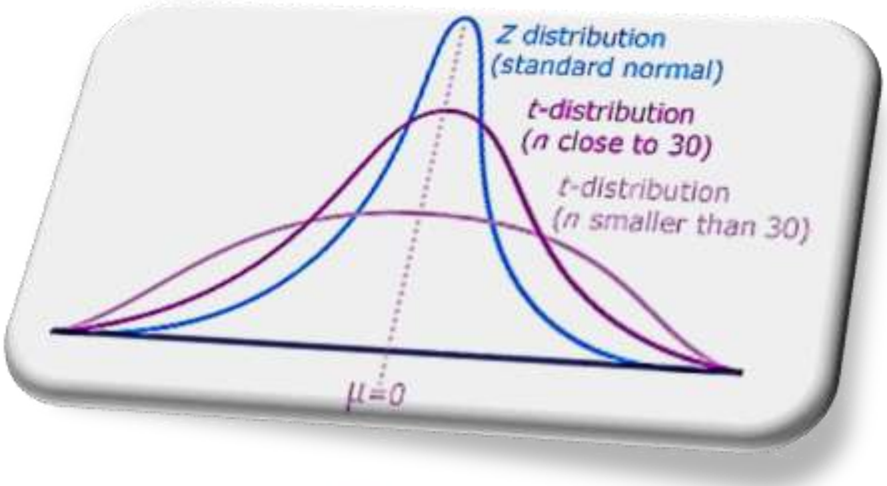
The sampling distribution will be normal for a large sample

Conditions:

- The population distribution is normal.
- The sample distribution is roughly symmetric, unimodal, without outliers, and the sample size is 15 or less.
- The sample distribution is moderately skewed, unimodal, without outliers, and sample size is between 16 and 40.
- The sample size is greater than 40, without outliers.

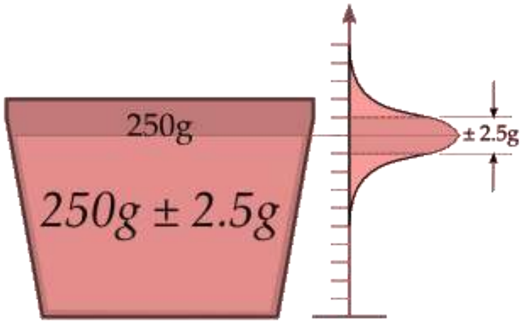
Student's t Distribution

- **t = [x - μ] / [s / sqrt(n)]**
- Small sample sizes



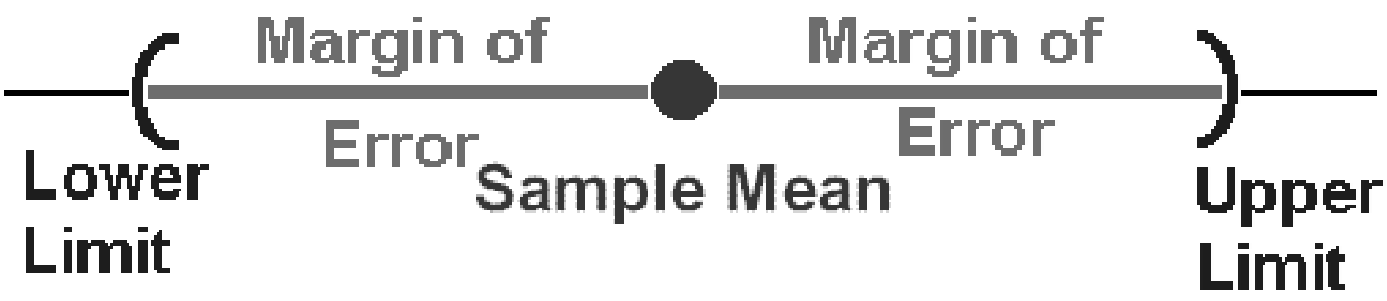
Estimation Theory

- Process to makes inferences about a population
  - based on information from a sample
  - Point Estimation
    - population mean μ, based on sample mean x
  - Interval Estimation

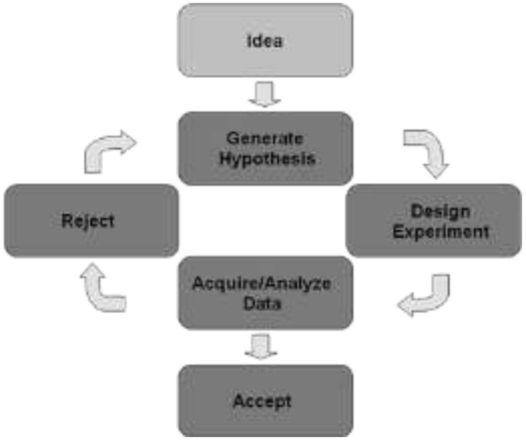


Confidence Interval

- Precision and uncertainty
  - Confidence level
  - Statistic
  - Margin of error



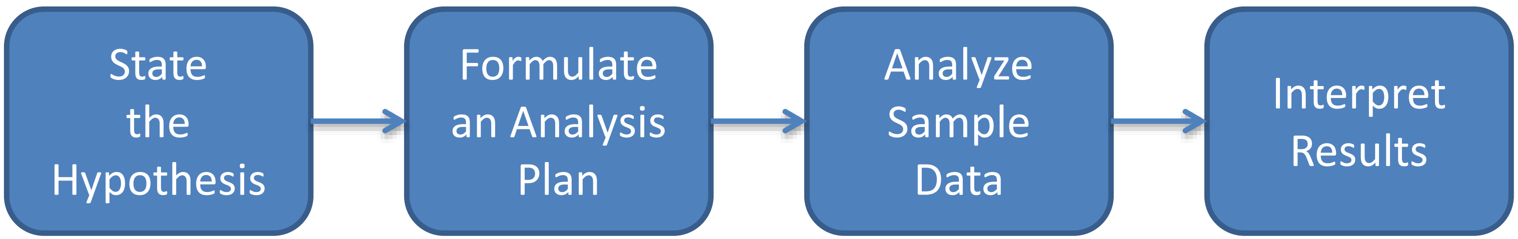
Hypothesis Testing



Statistical Hypotheses

- Examine random sample from a population
- **Null hypothesis**
  - H_0
  - observations result purely from chance.
- **Alternative hypothesis**
  - H_a
  - observations are influenced by some non-random cause.

Hypothesis Tests



Decision Error

Your Decision Based On a Random Sample	Given the Null Hypothesis Is: True	Given the Null Hypothesis Is: False
Reject	Type I Error	Correct Decision
Do Not Reject	Correct Decision	Type II Error

Two Types of Errors in Decision Making

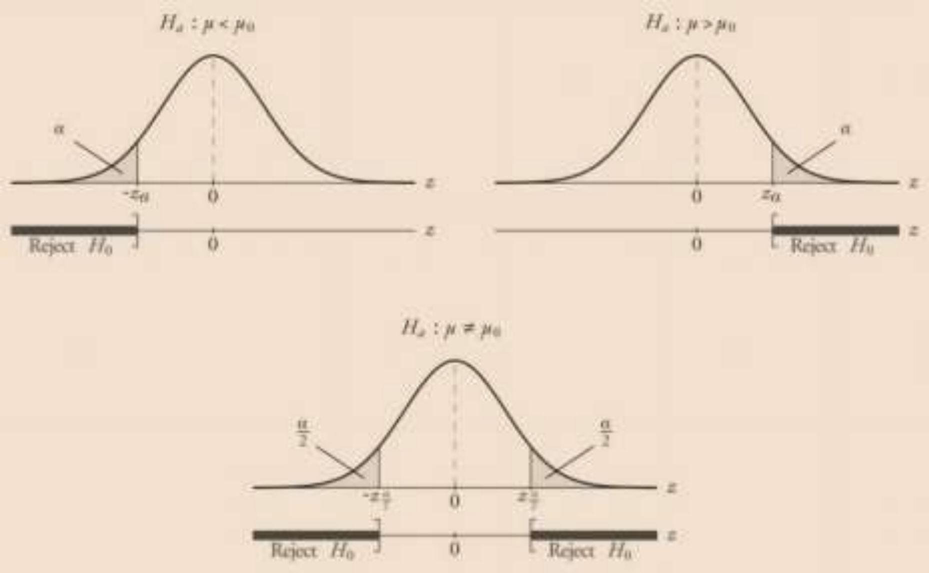
Decision Rules

- **P-value**.
  - Strength of evidence in support of a H_0
  - P-value is < significance level (0.05), reject H_0

- **Region of acceptance**.
  - range of values.
  - test statistic falls within the region of acceptance
  - H_0 is not rejected
  - defined so that the chance of making a Type I error is equal to the significance level

One-Tailed and Two-Tailed Tests

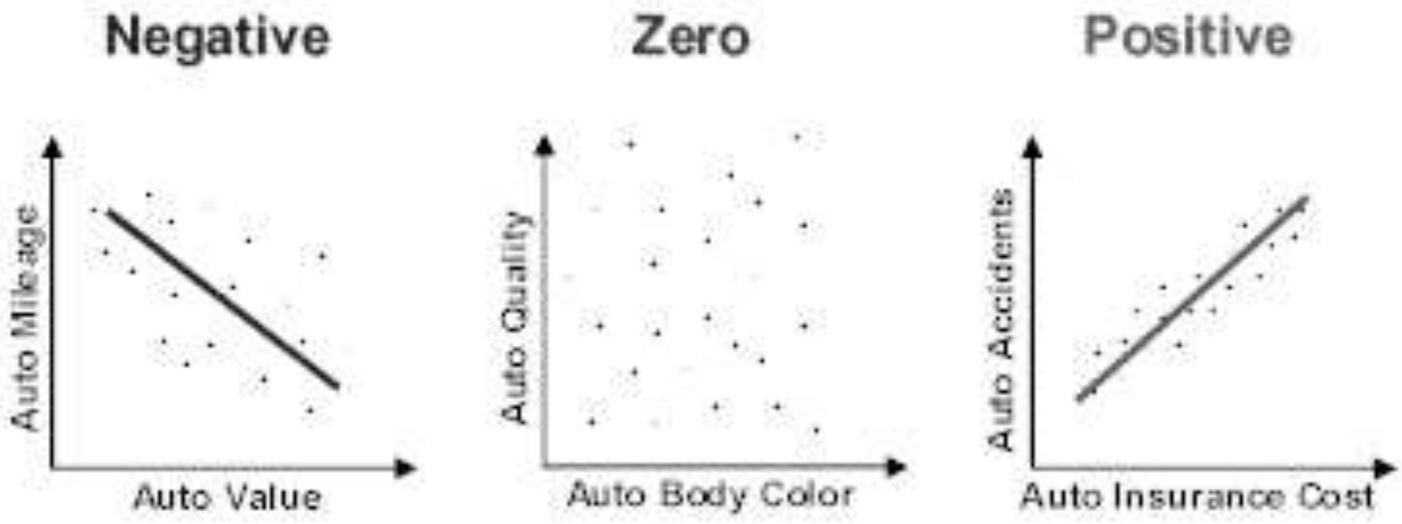
- One-tailed test
  - region of rejection is on only one side of the sampling distribution
  - Example: mean is less than or equal to 10
- Two-tailed test
  - region of rejection is on both sides
  - Example: mean is equal to 10



Linear Correlation

Correlation

Relationship Between Two Quantities Such That When One Changes, the Other Does



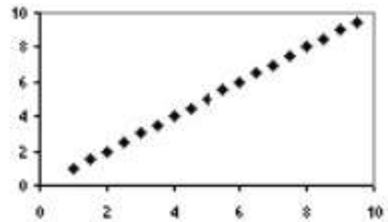
Correlation Coefficients

- Measure the strength of association between two variables
- Pearson product-moment correlation coefficient

$$r = \Sigma (xy) / \text{sqrt}[(\Sigma x^2) * (\Sigma y^2)]$$

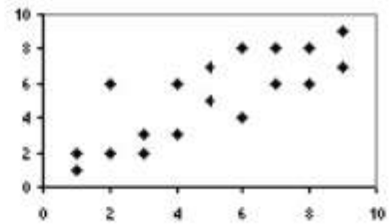
  - $x = x_i - x$,
  - x_i is the x value for observation i
  - x is the mean x value,
  - $y = y_i – y$
  - y_i is the y value for observation I
  - y is the mean y value.

Correlation Coefficients



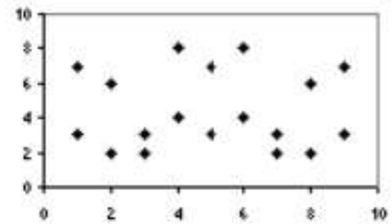
Maximum positive correlation

(r = 1.0)



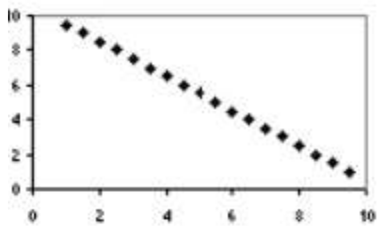
Strong positive correlation

(r = 0.80)



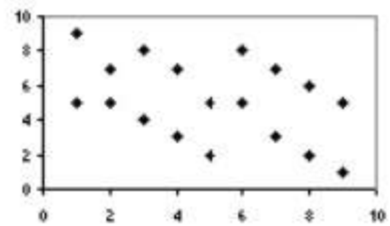
Zero correlation

(r = 0)



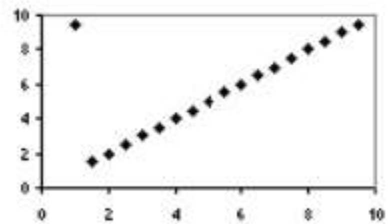
Maximum negative correlation

(r = -1.0)



Moderate negative correlation

(r = -0.43)



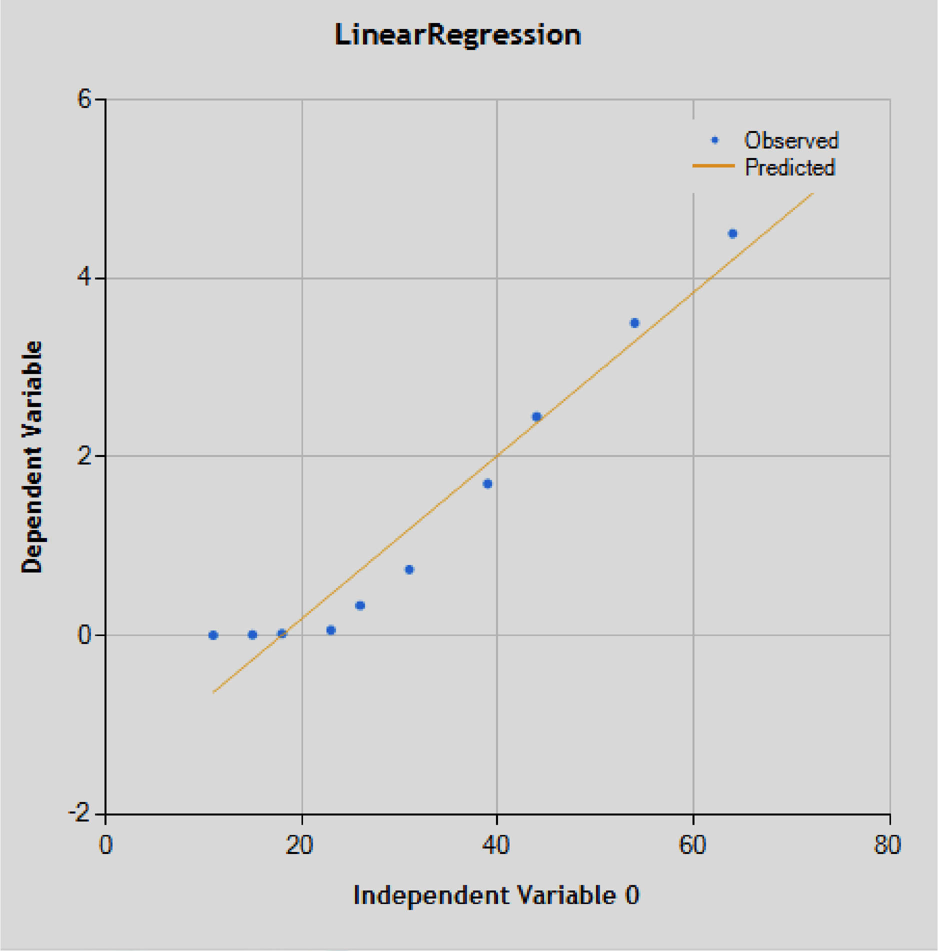
Strong correlation & outlier

(r = 0.71)

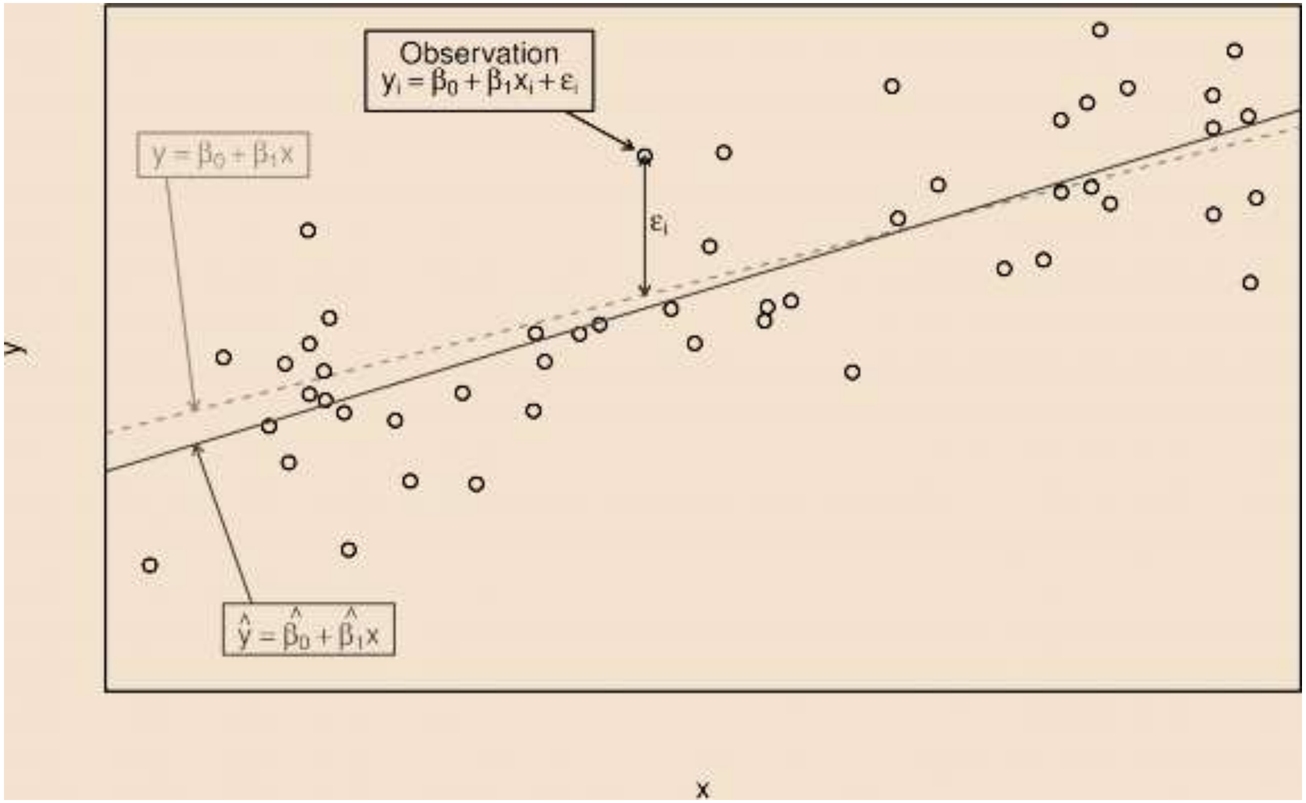
Note:

A correlation of 0 does not mean zero relationship between two variables; rather, it means zero linear relationship.

Linear Regression



Least Squares Regression Line

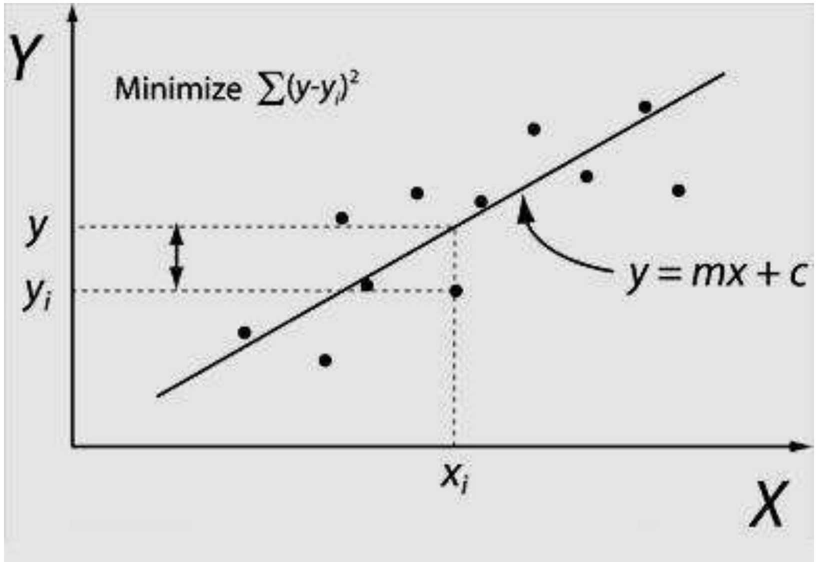


Regression line

- $\hat{y} = b_0 + b_1 x$
  - $b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$
  - $b_1 = r * (s_y / s_x)$
  - $b_0 = y - b_1 * x$
    - b_0 is the constant in the regression equation
    - b_1 is the regression coefficient
    - r is the correlation between x and y
    - x_i is the *X* value of observation *I*
    - y_i is the *Y* value of observation *I*
    - x is the mean of *X*
    - y is the mean of *Y*
    - s_x is the standard deviation of *X*
    - s_y is the standard deviation of *Y*

Properties of Regression Line

- Minimizes sum of squared differences
- Passes through mean of the *X* and *Y* values (x & y)
- (b_0) is equal to the y intercept
- (b_1) is the slope of the regression line

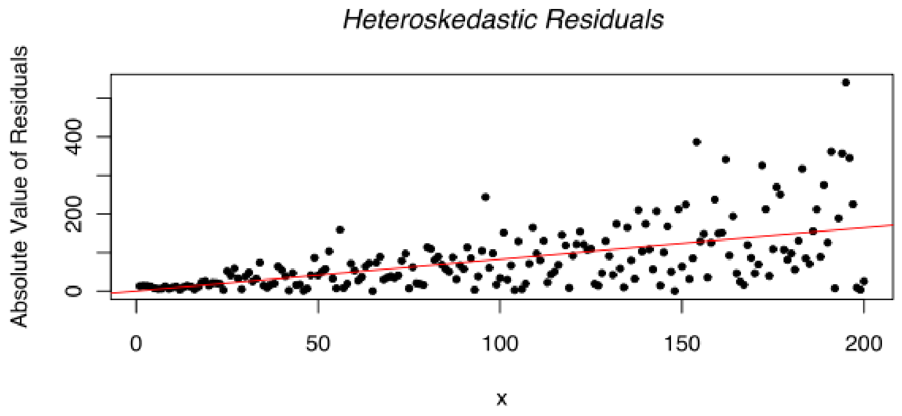


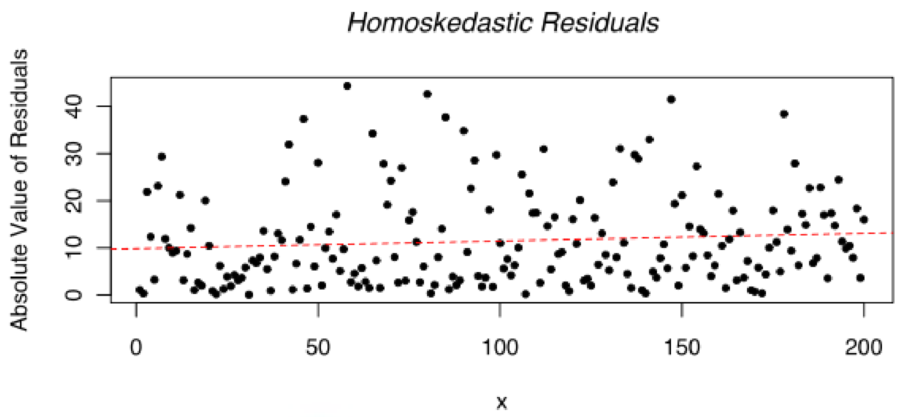
Coefficient of Determination

- R^2
  - Between 0 and 1
  - $R^2 = 0$, dependent variable cannot be predicted
  - $R^2 = 1$, dependent variable can be predicted without error
  - An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable.
    - $R^2 = 0.10$ means that 10% of the variance in *Y* is predictable from *X*
    - $R^2 = 0.20$ means that 20% is predictable

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

Homoscedasticity and Heteroscedasticity





Thank You!

To Learn Automated Trading

Connect With Us:

INDIA
A-309, Boomerang,
Chandivali Farm Road, Powai,
Mumbai - 400 072
Phone: +91-022-61691400

SINGAPORE
11 Collyer Quay,
#10-10, The Arcade,
Singapore - 049317
Phone: +65-6221-3654

Email: contact@quantinsti.com