Inferential Statistic

Probability and Statistics

Probability is the chance of an outcome in an experiment (also called event).

Event: Tossing a fair coin
Outcome: Head, Tail

Probability deals with **predicting** the likelihood of **future** events.	Statistics involves the **analysis** of the **frequency** of **past** events

Example: Consider there is a drawer containing 100 socks: 30 red, 20 blue and 50 black socks.

We can use probability to answer questions about the selection of a random sample of these socks.

- **PQ1.** What is the probability that we draw two blue socks or two red socks from the drawer?
- **PQ2.** What is the probability that we pull out three socks or have matching pair?
- **PQ3.** What is the probability that we draw five socks and they are all black?

Statistics

Instead, if we have no knowledge about the type of socks in the drawers, then we enter into the realm of statistics. Statistics helps us to infer properties about the population on the basis of the random sample.

Questions that would be statistical in nature are:

- **SQ1**: A random sample of 10 socks from the drawer produced one blue, four red, five black socks. What is the total population of black, blue or red socks in the drawer?
- **SQ2**: We randomly sample 10 socks, and write down the number of black socks and then return the socks to the drawer. The process is done for five times. The mean number of socks for each of these trial is 7. What is the true number of black socks in the drawer?
- etc.

Probability vs. Statistics

In other words:

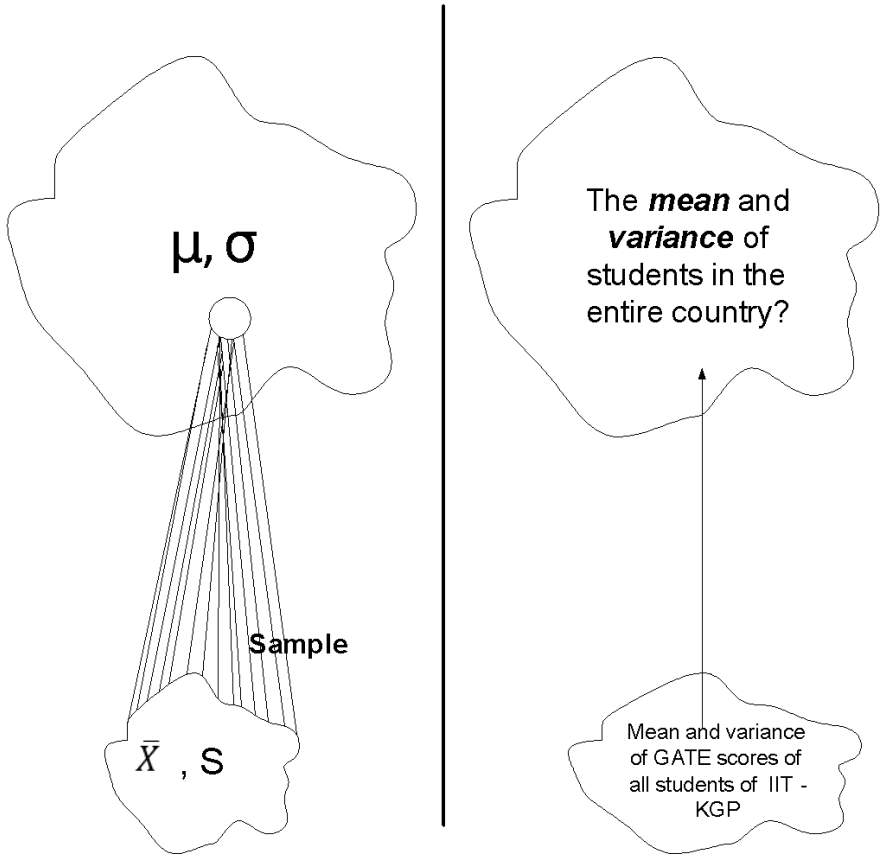
- In probability, we are given a model and asked what kind of data we are likely to see.
- In statistics, we are given data and asked what kind of model is likely to have generated it.

Measles Study

- A study on health is concerned with the incidence of childhood measles in parents of childbearing age in a city. For each couple, we would like to know how likely, it is that either the mother or father or both have had childhood measles.
- The current census data indicates that 20% adults between the ages 17 and 35 (regardless of sex) have had childhood measles.
  - This give us the probability that an individual in the city has had childhood measles.

Introduction

The primary objective of statistical analysis is to use data from a sample to make inferences about the population from which the sample was drawn.



Introduction

As a task of statistical inference, we usually follow the following steps:

- **Data collection**
  - Collect a sample from the population.
- **Statistics**
  - Compute a statistics from the sample.
- **Statistical inference**
  - From the statistics we made various statements concerning the values of population parameters.
    - For example, population mean from the sample mean, etc.

Basic terminologies

Some basic terminology which are closely associated to the above-mentioned tasks are reproduced below.

- **Population:** A population consists of the totality of the observation, with which we are concerned.
- **Sample:** A sample is a subset of a population.
- **Random variable:** A random variable is a function that associates a real number with each element in the sample.
- **Statistics:** Any function of the random variable constituting random sample is called a statistics.
- **Statistical inference:** It is an analysis basically concerned with generalization and prediction.

Statistical Inference

There are two facts, which are key to statistical inference.

1. Population parameters are fixed number whose values are usually unknown.
2. Sample statistics are known values for any given sample, but vary from sample to sample, even taken from the same population.

- In fact, it is unlikely for any two samples drawn independently, producing identical values of sample statistics.
- In other words, the variability of sample statistics is always present and must be accounted for in any inferential procedure.
- This variability is called sampling variation.

Note:

A sample statistics is random variable and like any other random variable, a sample statistics has a probability distribution.

Statistical Inference: Need for Sampling

- Working on entire population is not feasible
- Almost always experiments are conducted on samples taken from population or generated based on some model
- Focuses on drawing conclusions about population from samples
- Typical questions that are answered:
  - I have a sample of Indians and I want to **estimate** the mean salary of all Indians
    - **Estimation**
  - I **think** mean Indian salary is X. Does this sample of Indians which I have collected agree with my claim or disagrees with my claim?
    - **Hypothesis Testing**
- Applications
  - Quality control
  - Acceptance sampling

Basic Approaches

Approach 1: Hypothesis testing

- We conduct **test on hypothesis**.
  - We hypothesize that one (or more) parameter(s) has (have) some specific value(s) or relationship.
- Make our decision about the parameter(s) based on one (or more) sample statistic(s)
- Accuracy of the decision is expressed as the probability that the decision is incorrect.

Approach 2: Confidence interval measurement

- We estimate one (or more) parameter(s) using sample statistics.
  - This estimation usually done in the form of an interval.
- Accuracy of the decision is expressed as the **level of confidence** we have in the interval.

(Non) Sampling Error

- While dealing with sample and not the entire population, some error is inherent
- Such an error is called **sampling** or **statistical error**
- **Non sampling error** on the other hand occurs when the sample is not good and does not represent the entire population well
- Analysts want to eliminate non-sampling error and understand the nature of sampling error
- Sampling error can be minimized by increasing the size of the sample tradeoff with cost

Quick Revision: Coins

- Probability of getting a head?
  - 0.5
- Probability of getting HH?
  - 0.25
- Probability of getting exactly one H in two tosses?
  - 0.5
- Probability of getting exactly one head in three tosses?
  - 0.375

Random Variables

- Numerical description of the outcome of an experiment
- Discrete – number of possible outcomes can be counted
  - Need not necessarily be finite
  - Number of hits on a website link
- Continuous
  - Daily temperature
  - Time to complete a task
  - Time between failures of a machine

Defining Random Variable

Definition : **Random Variable**

A random variable is a rule that assigns a numerical value to an outcome of interest.

Example 4.2: In "measles Study", we define a random variable X as the number of parents in a married couple who have had childhood measles.

This random variable can take values of $0, 1$ *and* 2.

Note:

- Random variable is not exactly the same as the variable defining a data.
- The probability that the random variable takes a given value can be computed using the rules governing probability.
  - For example, the probability that $X = 1$ means either mother or father but not both has had measles is 0.32. Symbolically, it is denoted as **P(X=1) = 0.32**

Probability Distributions

- Characterization of the possible values that a RV may assume along with the respective probability
- Working knowledge of common families of probability distributions is important
  - Can help you to understand the underlying process that generates sample data
  - Many phenomena in business and nature follow some theoretical distribution and are thus useful for building decision models
  - Essential in computing probabilities of occurrence of outcomes to assess risk and make decisions

Probability Distribution

Definition : Probability distribution

A probability distribution is a definition of probabilities of the values of random variable.

Example 4.3: Given that 0.2 is the probability that a person (in the ages between 17 and 35) has had childhood measles. Then the probability distribution is given by

X	Probability
0	0.64
1	0.32
2	0.04

?

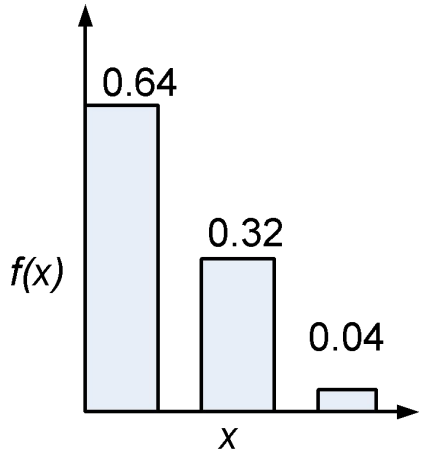
Probability Distribution

- In data analytics, the probability distribution is important with which many statistics making inferences about population can be derived .
  - In general, a probability distribution function takes the following form

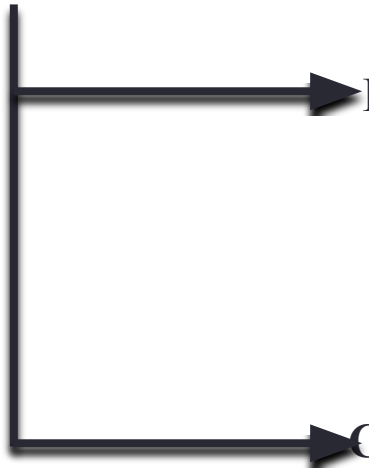
x	x_1	$x_2 \dots\dots\dots\dots\dots x_n$
$f(x) = P(X = x)$	$f(x_1)$	$f(x_2) \dots\dots\dots f(x_n)$

Example: Measles Study

x	0	1	2
$f(x)$	0.64	0.32	0.04



Taxonomy of Probability Distributions



Discrete probability distributions

- Binomial distribution
- Multinomial distribution
- Poisson distribution
- Hypergeometric distribution

Continuous probability distributions

- Normal distribution
- Standard normal distribution
- Gamma distribution
- Exponential distribution
- Chi square distribution
- Lognormal distribution
- Weibull distribution

Usage of Probability Distribution

- Distribution (discrete/continuous) function is widely used in simulation studies.
  - A simulation study uses a computer to simulate a real phenomenon or process as closely as possible.
  - The use of simulation studies can often eliminate the need of costly experiments and is also often used to study problems where actual experimentation is impossible.

Examples 4.4:

1) A study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use such a drug approximately follows a binomial distribution.

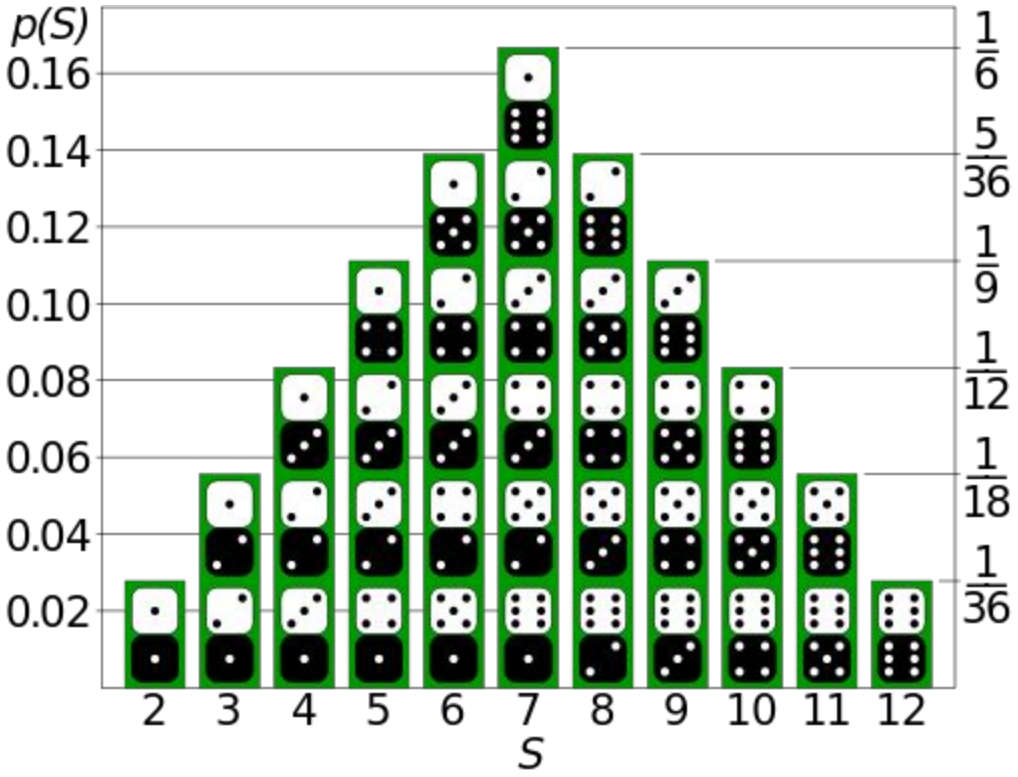
2) Operation of ticketing system in a busy public establishment (e.g., airport), the arrival of passengers can be simulated using Poisson distribution.

Discrete Probability Distributions

- Probability distribution for discrete RVs is called Probability Mass Function (PMF)
  - f(X=x) = P(X=x)
- Cumulative Distribution Function CDF
  - P(X<=x)

Example

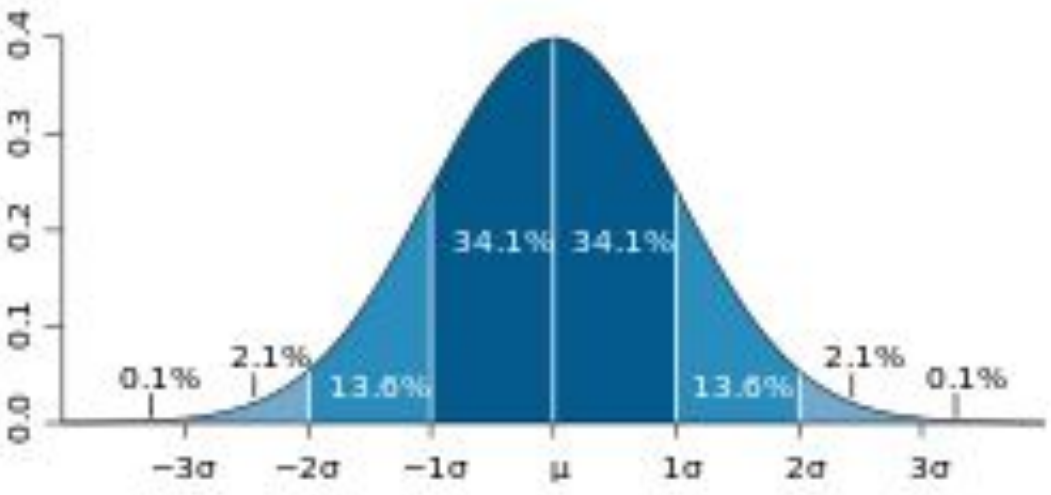
- Probability distribution for the sum S of counts from two dice
- The pmf allows the computation of probabilities of events such as P(S > 9) = 1/12 + 1/18 + 1/36 = 1/6



Continuous Probability Distributions

- PMF is now called PDF
- Histogram now becomes curve
- Area under curve = 1
- P(a<X<b) = area under curve between a and b
- Probabilities of X are defined only over intervals
  - F(x) = P(X<=x) = area under curve to the left of x
- P(a<=X<=b) = F(b) – F(a)

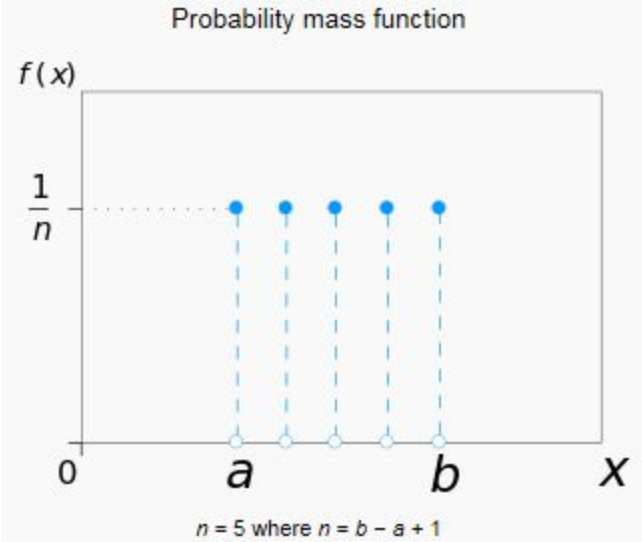
Example

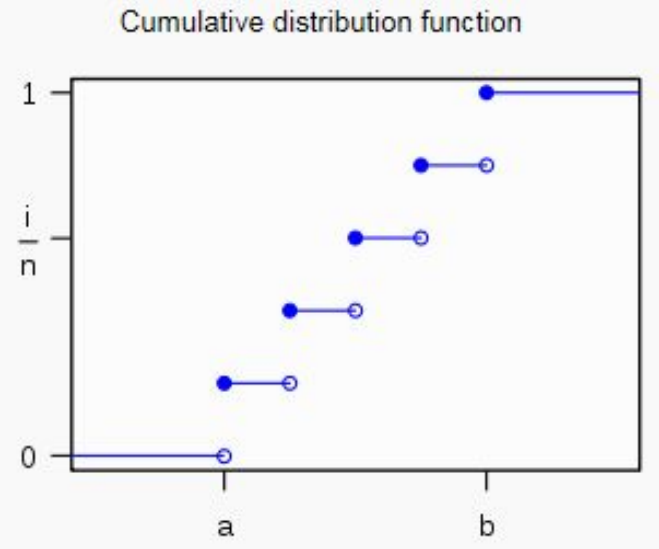


Discrete Probability Distributions

Discrete Uniform Distribution

- Rolling of a die
  - All outcomes (discrete) have same probability (1/6)





- A known, finite number of outcomes equally likely to happen

Bernoulli Distribution

- Characterizes a RV with two possible outcomes
- Coin flip
- Whether an individual responds negatively or positively to a telemarketing promotion

Binomial Distribution

- Models n independent replications of a Bernoulli experiment
- If *n* is large enough, you can use the normal distribution to get an approximate answer that's very close to what you would get with the binomial distribution

Binomial Distribution

- In many situations, an outcome has only two outcomes: success and failure.
  - Such outcome is called dichotomous outcome.
- An experiment when consists of repeated trials, each with dichotomous outcome is called Bernoulli process. Each trial in it is called a Bernoulli trial.

Example 4.5: Firing bullets to hit a target.

- Suppose, in a Bernoulli process, we define a random variable $X \equiv$ the number of successes in trials.
- Such a random variable obeys the binomial probability distribution, if the experiment satisfies the following conditions:

1) The experiment consists of n trials.
2) Each trial results in one of two mutually exclusive outcomes, one labelled a *"success"* and the other a *"failure"*.
3) The probability of a success on a single trial is equal to $\boldsymbol{p}$. The value of p remains constant throughout the experiment.
4) The trials are independent.

Defining Binomial Distribution

Definition : **Binomial distribution**

The function for computing the probability for the binomial probability distribution is given by

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

$$for\ x = 0, 1, 2, \ldots., n$$

Here, $f(x) = P(X = x)$, where X denotes "the number of success" and $X = x$ denotes the number of success in x trials.

Binomial Distribution

Example 4.6: Measles study

X = having had childhood measles a success

$p = 0.2$, the probability that a parent had childhood measles

$n = 2$, here a couple is an experiment and an individual a trial, and the number of trials is two.

Thus,

$$P(x = 0) = \frac{2!}{0!(2-0)!}(0.2)^0(0.8)^{2-0} = \mathbf{0.64}$$

$$P(x = 1) = \frac{2!}{1!\,(2-1)!}(0.2)^1(0.8)^{2-1} = \mathbf{0.32}$$

$$P(x = 2) = \frac{2!}{2!\,(2-2)!}(0.2)^2(0.8)^{2-2} = \mathbf{0.04}$$

Multinomial Distribution

- Can be used to compute the probabilities in situations in which there are more than two possible outcomes
- “If these two chess players played 12 games, what is the probability that Player A would win 7 games, Player B would win 2 games, and the remaining 3 games would be drawn?”

The Multinomial Distribution

The binomial experiment becomes a multinomial experiment, if we let each trial has more than two possible outcome.

Definition : **Multinomial distribution**

If a given trial can result in the ***k*** outcomes $E_1, E_2, \dots\dots, E_k$ with probabilities $p_1, p_2, \dots\dots, p_k$, then the probability distribution of the random variables $X_1, X_2, \dots\dots, X_k$ representing the number of occurrences for $E_1, E_2, \dots\dots, E_k$ in n independent trials is

$$f(x_1, x_2, \dots\dots, x_k) = \binom{n}{x_1, x_2, \dots\dots, x_k} {p_1}^{x_1} {p_2}^{x_2} \dots\dots {p_k}^{x_k}$$

$$\text{where } \binom{n}{x_1, x_2, \dots\dots, x_k} = \frac{n!}{x_1! x_2! \dots\dots x_k!}$$

$$\sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} p_i = 1$$

The Hypergeometric Distribution

- Collection of samples with two strategies
  - With replacement
  - Without replacement
- A necessary condition of the binomial distribution is that all trials are **independent** to each other.
  - When sample is collected "with replacement", then each trial in sample collection is independent.

Example 4.8:
Probability of observing three red cards in 5 draws from an ordinary deck of 52 playing cards.
  – You draw one card, note the result and then returned to the deck of cards
  – Reshuffled the deck well before the next drawing is made

- The hypergeometric distribution *does not require* independence and is based on the sampling done **without** replacement.

The Hypergeometric Distribution

- In general, the hypergeometric probability distribution enables us to find the probability of selecting x successes in n trials from N items.

Properties of Hypergeometric Distribution

- A random sample of size n is selected without replacement from N items.
- k of the N items may be classified as success and $N-k$ items are classified as failure.

Let X denotes a hypergeometric random variable defining the number of successes.

Definition : Hypergeometric Probability Distribution

The probability distribution of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labelled success and $N-k$ labelled as failure is given by

$$f(x) = P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

$$\max(0, n-(N-k)) \leq x \leq \min(n, k)$$

Multivariate Hypergeometric Distribution

The hypergeometric distribution can be extended to treat the case where the N items can be divided into k classes $A_1, A_2, \ldots\ldots, A_k$ with a_1 elements in the first class A_1, … and a_k elements in the k^{th} class. We are now interested in the probability that a random sample of size n yields x_1 elements from A_1, x_2 elements from $A_2, \ldots\ldots, x_k$ elements from A_k.

Definition: Multivariate Hypergeometric Distribution

If N items are partitioned into k classes $a_1, a_2, \ldots\ldots, a_k$ respectively, then the probability distribution of the random variables $X_1, X_2, \ldots\ldots, X_k$, representing the number of elements selected from $A_1, A_2, \ldots\ldots, A_k$ in a random sample of size n, is

$$f(x_1, x_2, \ldots\ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots\ldots, X_k = x_k) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\ldots\ldots\binom{a_k}{x_k}}{\binom{N}{n}}$$

$$\text{with } \sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} a_i = N$$

Poisson Distribution

- Used to model the number of occurrences in some unit of measure
- Number of customers arriving in 2 hours
- No. of visits to web page in a minute
- Number of errors per line of software code
- Telephone calls arriving in a system
- Customers arriving at a counter or call centre
- Cars arriving at a traffic light
- Number of Losses/Claims occurring in a given period of Time

The Poisson Distribution

There are some experiments, which involve the occurring of the number of outcomes during a given time interval (or in a region of space).

Such a process is called Poisson process.

Example :

Number of clients visiting a ticket selling counter in a metro station.

The Poisson Distribution

Properties of Poisson process

- The number of outcomes in one time interval is independent of the number that occurs in any other disjoint interval [Poisson process has no memory]
- The probability that a single outcome will occur during a very short interval is proportional to the length of the time interval and does not depend on the number of outcomes occurring outside this time interval.
- The probability that more than one outcome will occur in such a short time interval is negligible.

Definition : **Poisson distribution**

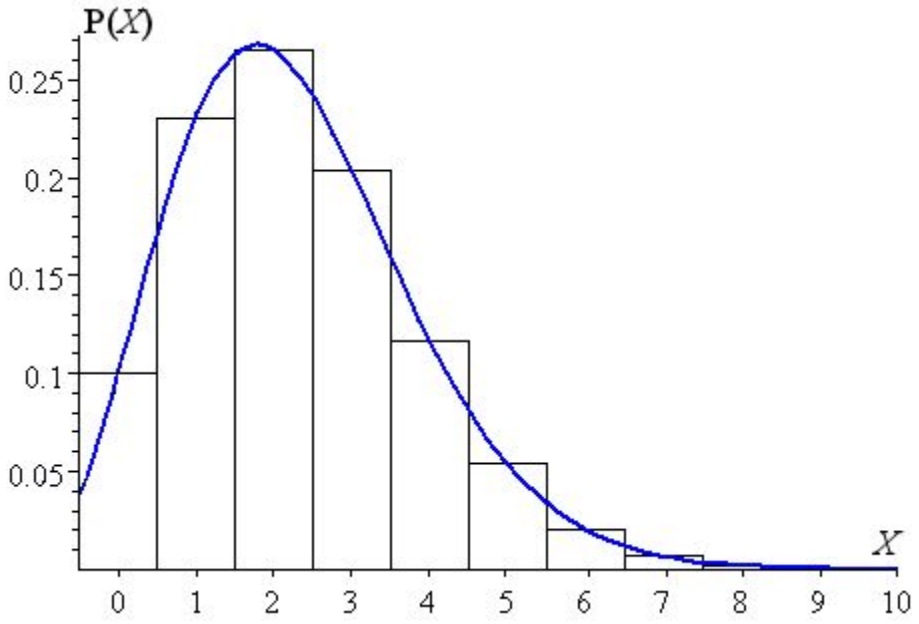
The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval t, is

$$f(x, \lambda t) = P(X = x) = \frac{e^{-\lambda t}.(\lambda t)^x}{x!}, x = 0, 1, \dots \dots$$

where λ is the average number of outcomes per unit time and $e = 2.71828 \dots$

Poisson Distribution

- x.poi <- rpois(n=200, lambda=2.5)
- hist(x.poi)



Descriptive measures

Given a random variable X in an experiment, we have denoted $f(x) = P(X = x)$, the probability that $X = x$. For discrete events $f(x) = 0$ for all values of x except $x = 0, 1, 2, \ldots\ldots$

Properties of discrete probability distribution

1. $0 \leq f(x) \leq 1$
2. $\sum f(x) = 1$
3. $\mu = \sum x.f(x)$ [is the mean]
4. $\sigma^2 = \sum(x - \mu)^2 . f(x)$ [is the variance]

In 2, 3 *and* 4, summation is extended for all possible discrete values of x.

Note: For discrete uniform distribution, $f(x) = \frac{1}{n}$ with $x = 1, 2, \ldots\ldots, n$

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

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

Descriptive measures

1. **Binomial distribution**

The binomial probability distribution is characterized with p (the probability of success) and n (is the number of trials). Then

$$\mu = n.p$$

$$\sigma^2 = np(1-p)$$

2. **Hypergeometric distribution**

The hypergeometric distribution function is characterized with the size of a sample (n), the number of items (N) and k labelled success. Then

$$\mu = \frac{nk}{N}$$

$$\sigma^2 = \frac{N-n}{N-1}.\mathrm{n}.\frac{k}{N}\left(1-\frac{k}{N}\right)$$

Descriptive measures

3. **Poisson Distribution**

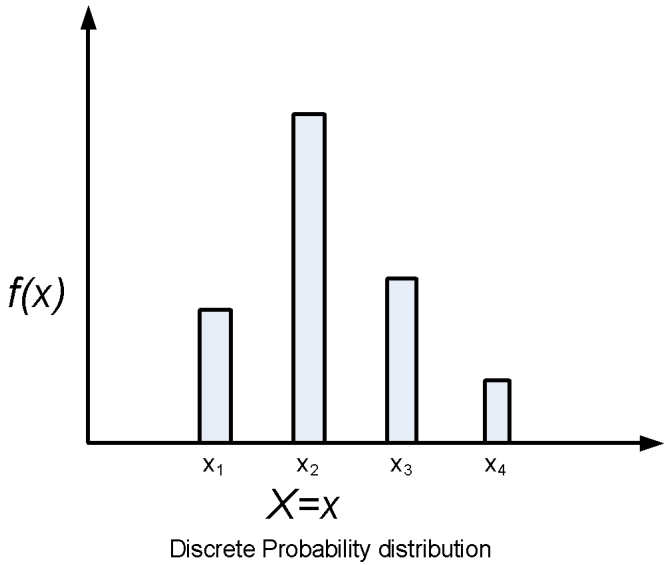
The Poisson distribution is characterized with λt where $\lambda = the\ mean\ of\ outcomes$ and $t = time\ interval$.

$$\mu = \lambda t$$

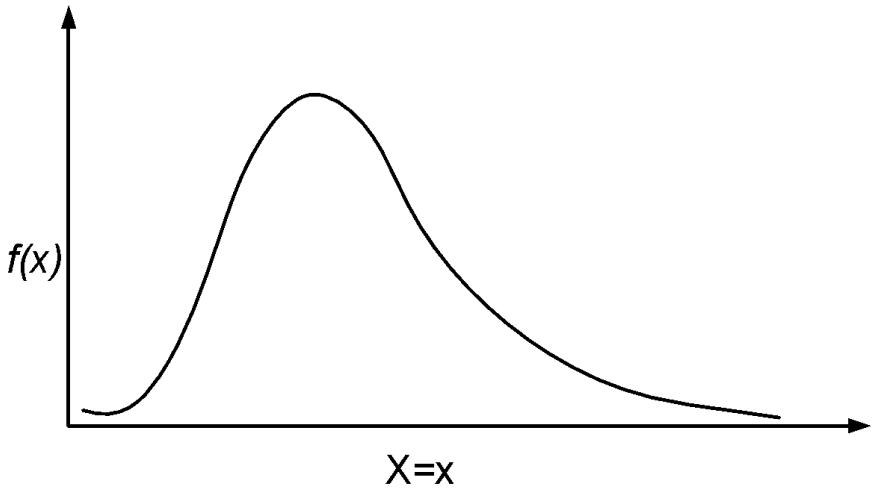
$$\sigma^2 = \lambda t$$

Continuous Probability Distributions

Continuous Probability Distributions



Discrete Probability distribution



Continuous Probability Distribution

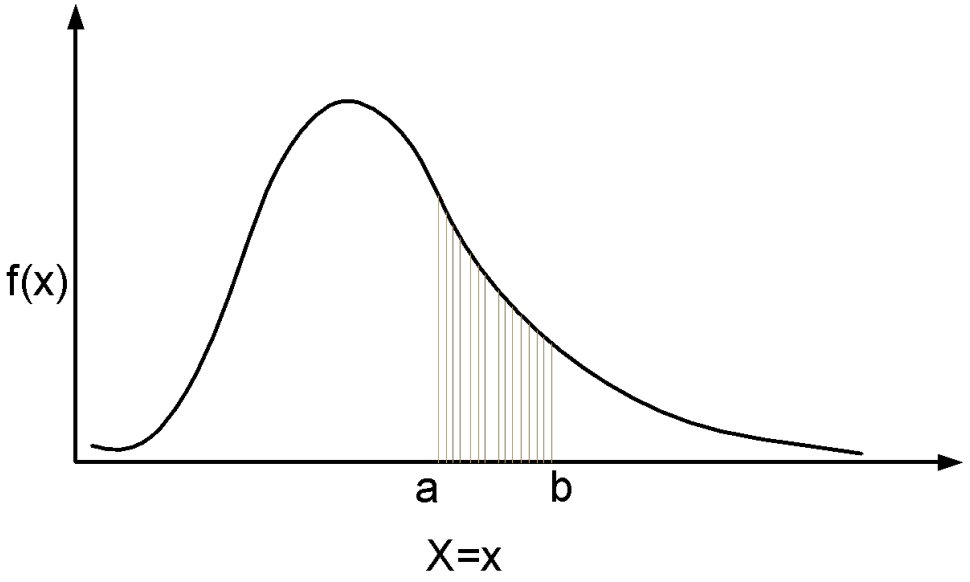
Continuous Probability Distributions

- When the random variable of interest can take any value in an interval, it is called continuous random variable.
  - Every continuous random variable has an infinite, uncountable number of possible values (i.e., any value in an interval)
- Consequently, continuous random variable differs from discrete random variable.

Properties of Probability Density Function

The function $f(x)$ is a probability density function for the continuous random variable X, defined over the set of real numbers R, if

1. $f(x) \geq 0$, for all $x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$
4. $\mu = \int_{-\infty}^{\infty} xf(x)\,dx$
5. $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$



Uniform Distribution

- All outcome between some minimum and maximum value are equally likely
- Often assumed as default in business scenarios when nothing other than min and max are known
- When a p-value is used as a test statistic for a simple null hypothesis, and the distribution of the test statistic is continuous, then the p-value is uniformly distributed between 0 and 1 if the null hypothesis is true
- Restricting and , the resulting distribution $U(0,1)$ is called a **standard uniform distribution**

Continuous Uniform Distribution

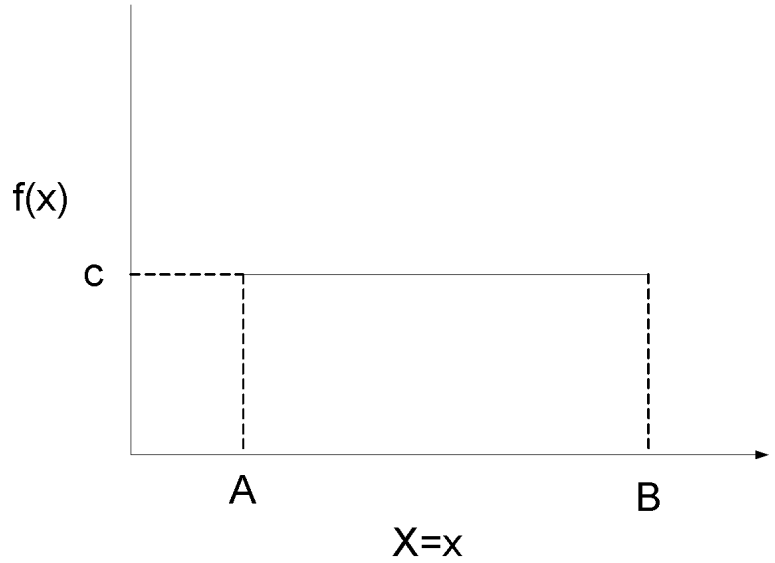
- One of the simplest continuous distribution in all of statistics is the continuous uniform distribution.

Definition 4.8: **Continuous Uniform Distribution**

The density function of the continuous uniform random variable X on the interval $[A, B]$ is:

$$f(x: A, B) = \begin{cases} \dfrac{1}{B - A} & A \leq x \leq B \\ 0 & Otherwise \end{cases}$$

Continuous Uniform Distribution



Note:

a) $\int_{\infty}^{-\infty} f(x)dx = \frac{1}{B-A} \times (B - A) = 1$

b) $P(c < x < d) = \frac{d-c}{B-A}$ where both c and d are in the interval (A,B)

c) $\mu = \frac{A+B}{2}$

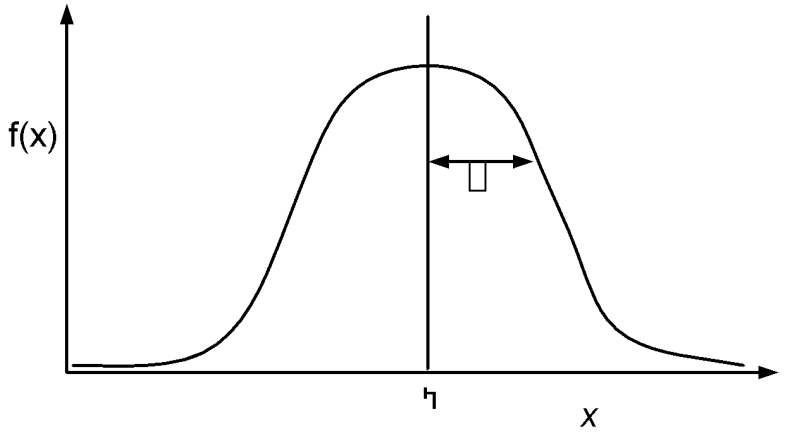
d) $\sigma^2 = \frac{(B-A)^2}{12}$

Normal Distribution

- The most often used continuous probability distribution is the normal distribution; it is also known as **Gaussian** distribution.
- Its graph called the normal curve is the bell-shaped curve.
- Such a curve approximately describes many phenomenon occur in nature, industry and research.
  - Physical measurement in areas such as meteorological experiments, rainfall studies and measurement of manufacturing parts are often more than adequately explained with normal distribution.
- A continuous random variable X having the bell-shaped distribution is called a normal random variable.

Normal Distribution

- The mathematical equation for the probability distribution of the normal variable depends upon the two parameters μ and σ, its mean and standard deviation.



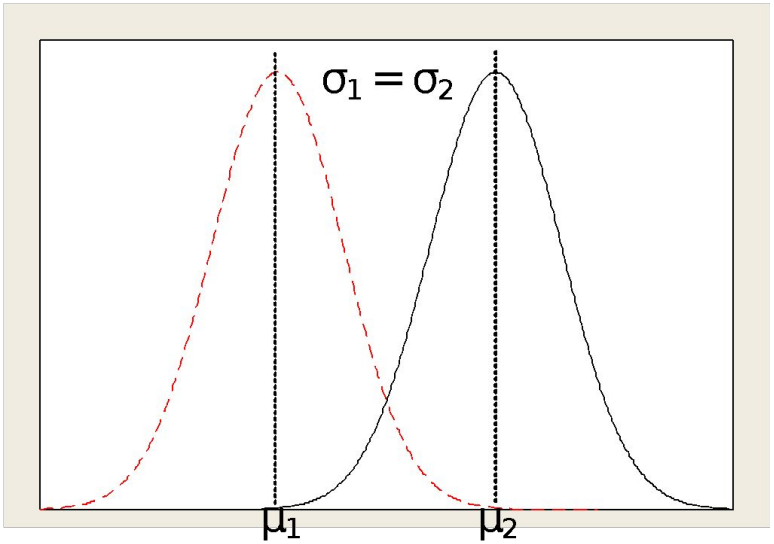
Definition : **Normal distribution**

The density of the normal variable x with mean μ and variance σ^2 is

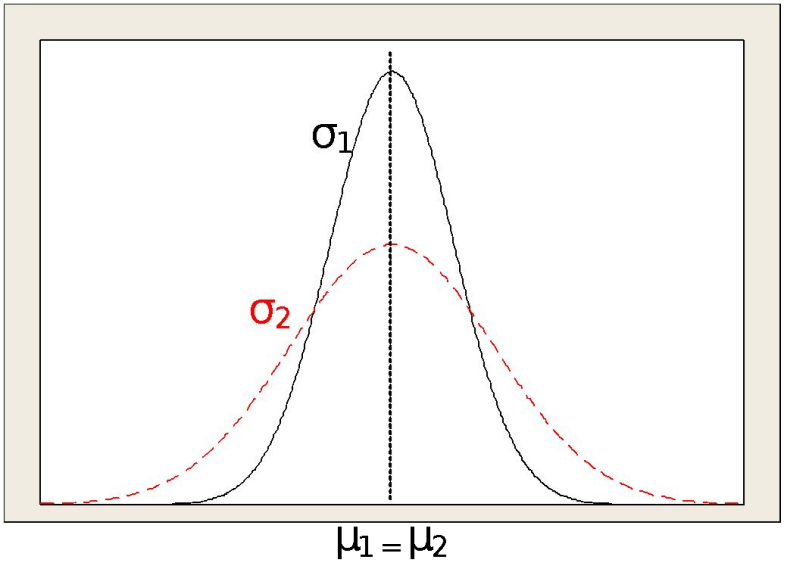
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

where $\pi = 3.14159\ldots$ and $e = 2.71828\ldots.$, the Naperian constant

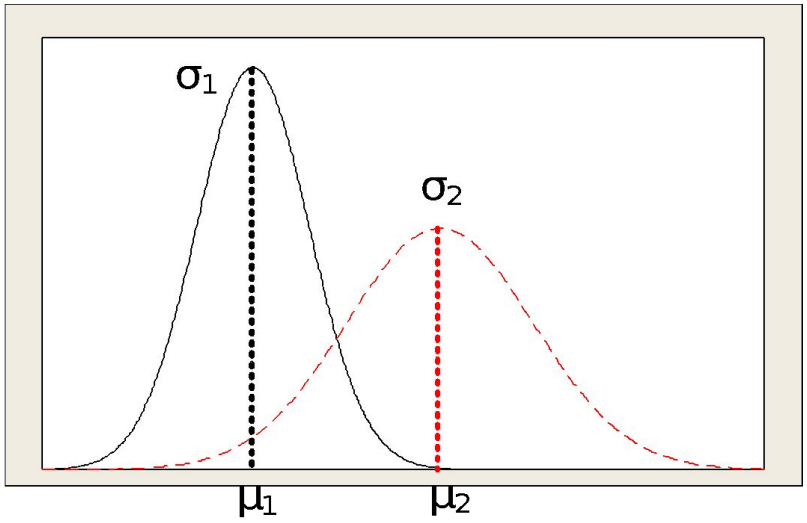
Normal Distribution



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$



Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$

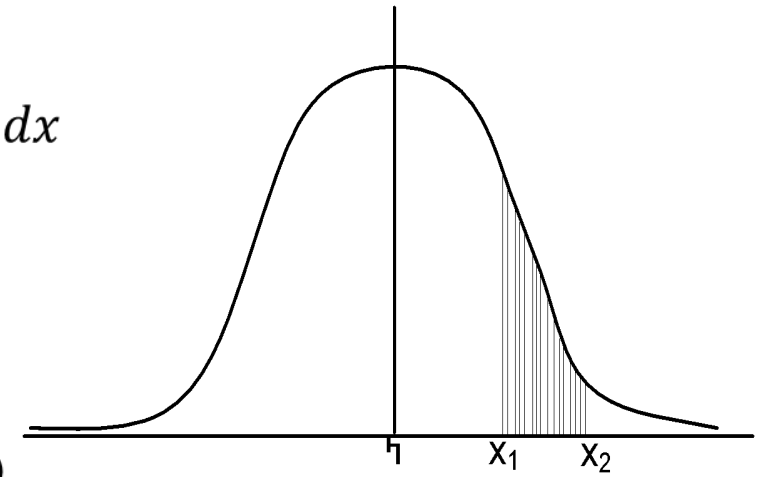
Properties of Normal Distribution

- The curve is symmetric about a vertical axis through the mean μ.
- The random variable x can take any value from $-\infty \; to \; \infty$.
- The most frequently used descriptive parameter s define the curve itself.
- The mode, which is the point on the horizontal axis where the curve is a maximum occurs at $x = \mu$.
- The total area under the curve and above the horizontal axis is equal to 1.

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1$$

- $\mu = \int_{-\infty}^{\infty} x.f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x.e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$
- $\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2.e^{-\frac{1}{2}[(x-\mu)/\sigma^2]} dx$
- $P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

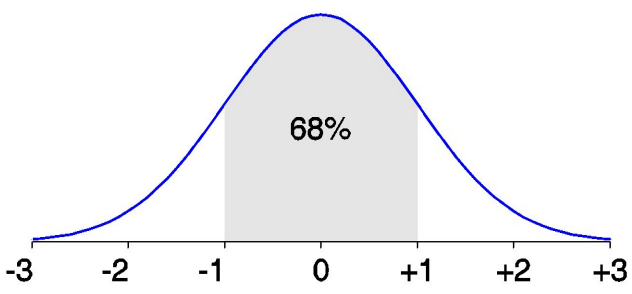
denotes the probability of x in the interval (x_1, x_2).

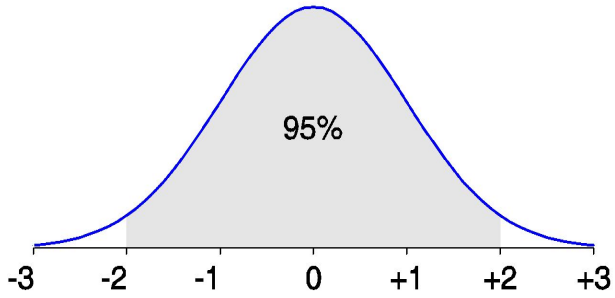


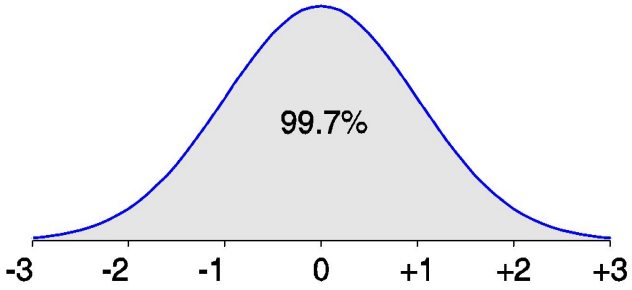
Standard Normal Distribution

- Special case of normal distribution
- Mean = 0, std dev = 1
- Pre calculated table
- Probability of any normal random variable X
  - Convert it to Z
  - Lookup the appropriate value in table
  - Allows to look up the cumulative probability for any value of z between -3 and 3

Standard Normal Distribution







Standard Normal Distribution

- The normal distribution has computational complexity to calculate $P(x_1 < x < x_2)$ for any two (x_1, x_2) and given μ and σ
- To avoid this difficulty, the concept of z-transformation is followed.

$$z = \frac{x-\mu}{\sigma} \qquad \text{[Z-transformation]}$$

- X: Normal distribution with mean μ and variance σ^2.
- Z: Standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.
- Therefore, if *f(x)* assumes a value, then the corresponding value of $f(z)$ is given by

$$f(x:\mu,\sigma): \quad P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

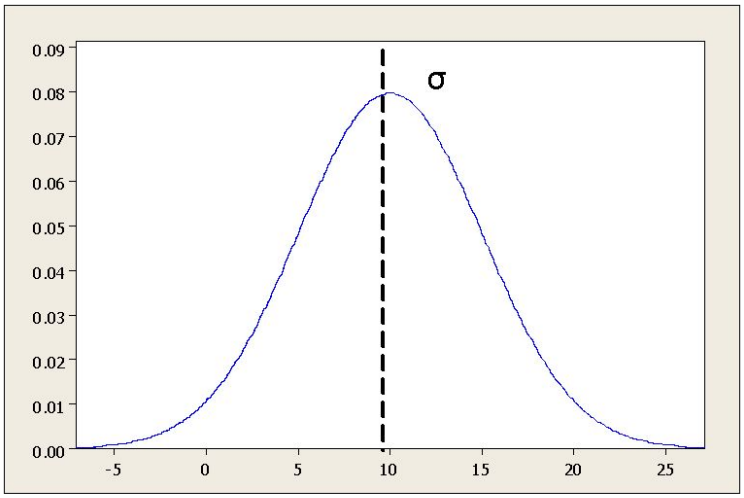
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$

$$= f(z:0,\sigma)$$

Standard Normal Distribution

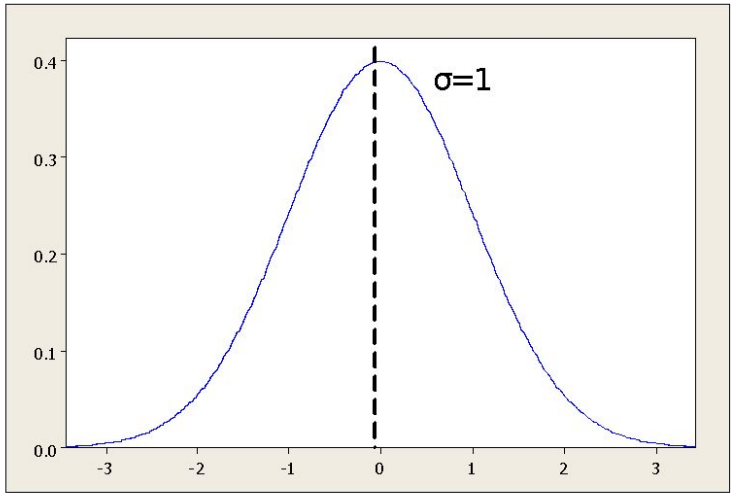
Definition : **Standard normal distribution**

The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution.



$x=\mu$

$f(x: \mu, \sigma)$



$\mu=0$

$f(z: 0, 1)$

Exponential Distribution

- Models time between randomly occurring events
- Time between failure
- Inter arrival times of customers
- If the number of events occurring **during** an interval of time is Poisson distributed, then the time **between** events is exponentially distributed
- Bounded below by 0
- Greatest density at 0, declines as x increases

Exponential Distribution

Definition : **Exponential Distribution**

The continuous random variable x has an exponential distribution with parameter β , where:

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & \text{where } \beta > 0 \\ 0 \end{cases}$$

Note:

1) The mean and variance of gamma distribution are

$$\mu = \alpha\beta$$
$$\sigma^2 = \alpha\beta^2$$

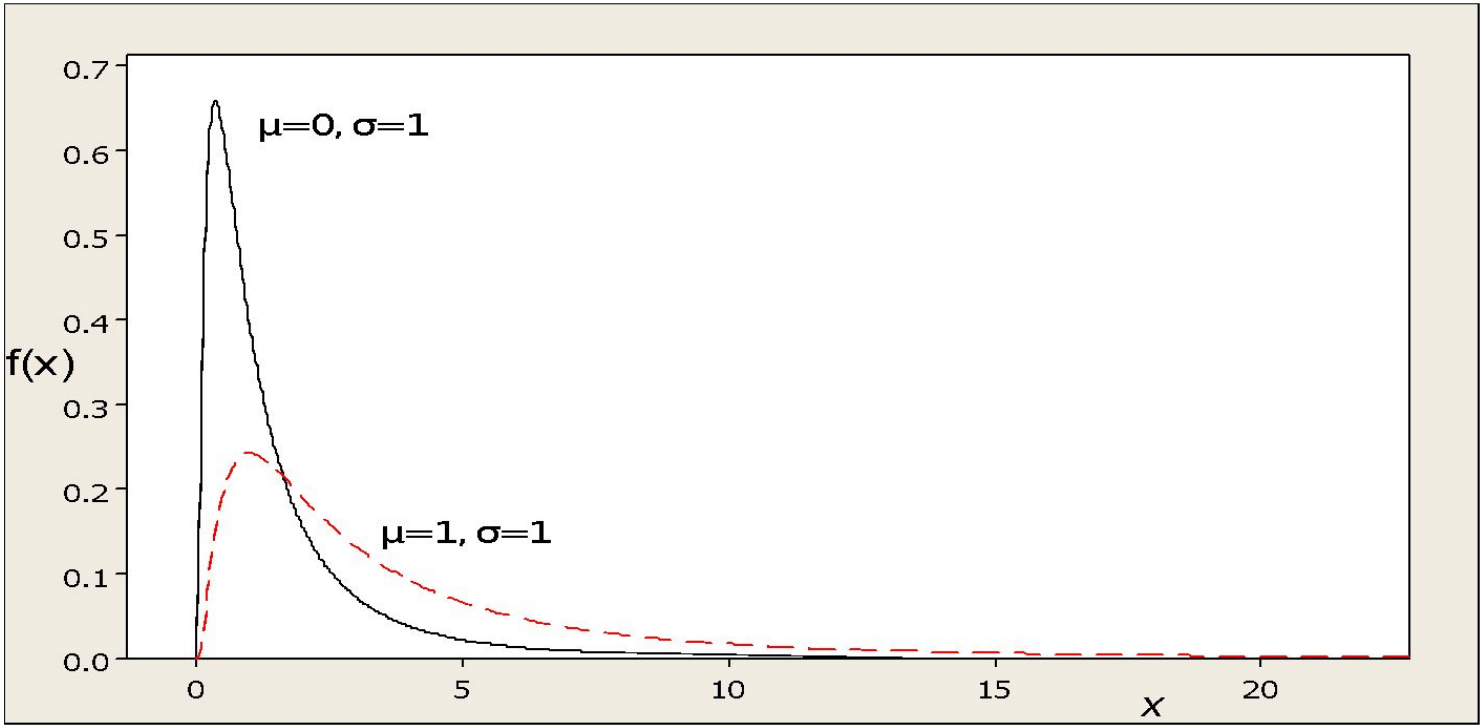
2) The mean and variance of exponential distribution are

$$\mu = \beta$$
$$\sigma^2 = \beta^2$$

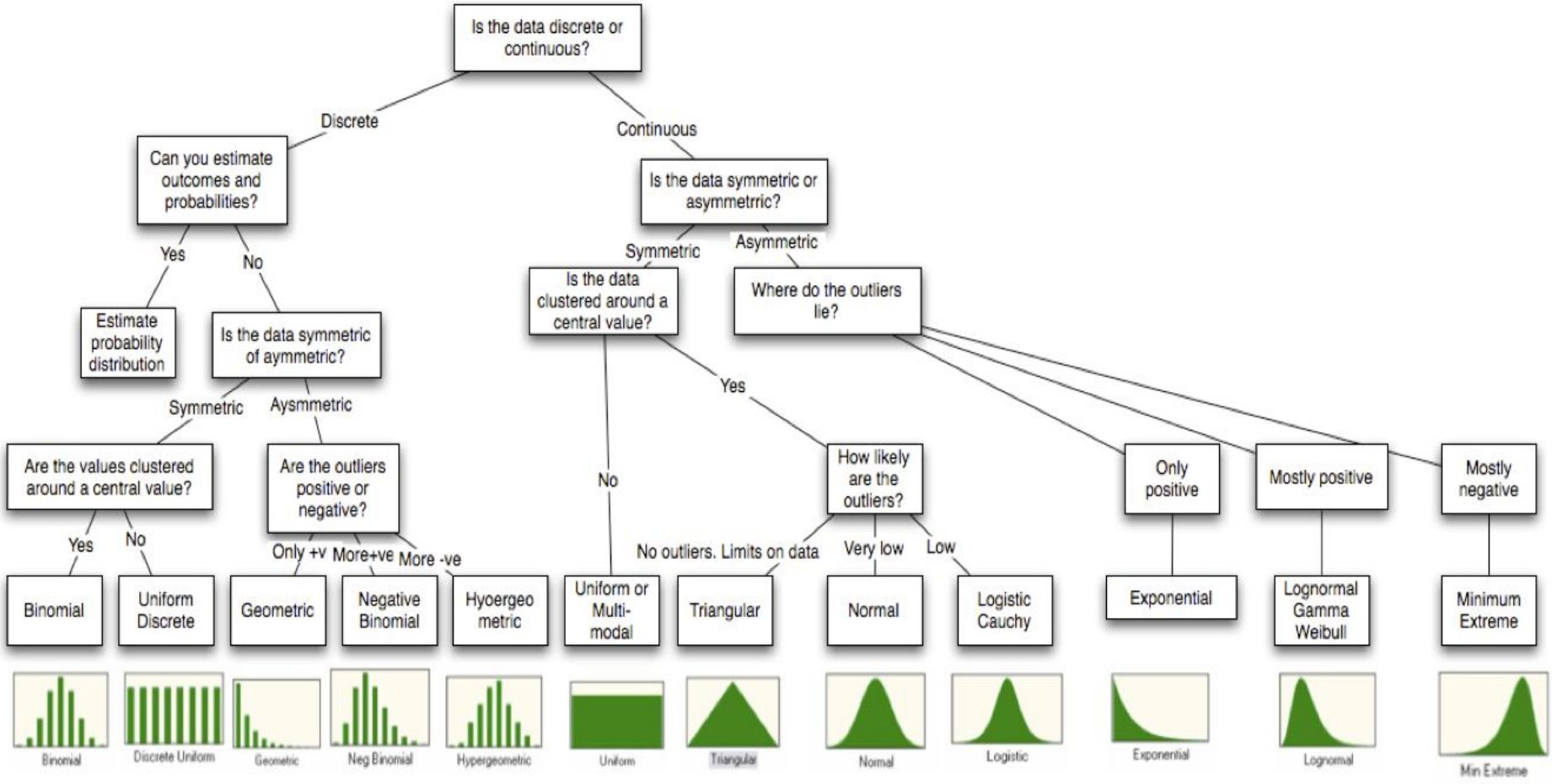
Other Practical Distributions

- Triangular distribution
  - Often used when no data are available to characterize an uncertain variable except min, max and most likely value
- Lognormal distribution
  - If natural logarithm of a RV X is normal, then X has a log normal distribution
  - Positively skewed, bounded below by zero
  - Used to model phenomena that have low probabilities of large values and cannot have negative values
  - E.g. time to complete a task, stock prices, real estate prices

Lognormal Distribution



Summary



Probabilistic approaches to risk by Aswath Damodaran

Sampling Distribution

- When you take a sample of data, it's important to realize the results will vary from sample to sample

- Sampling distribution is the distribution of means of all possible samples of a fixed size n from some population

Sampling Distribution

Example 5.1:

Consider five identical balls numbered and weighting as 1, 2, 3, 4 and 5. Consider an experiment consisting of drawing two balls, replacing the first before drawing the second, and then computing the mean of the values of the two balls.

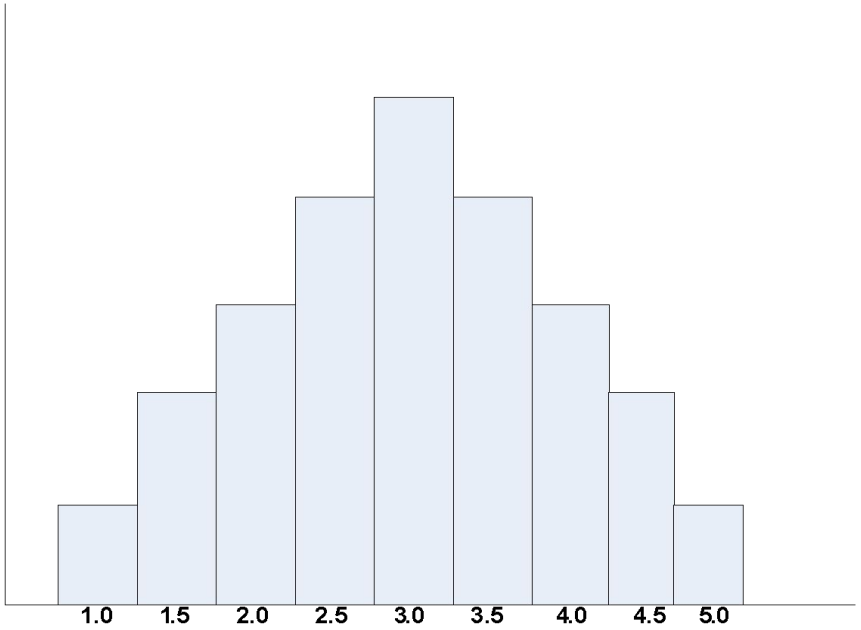
Following table lists all possible samples and their mean.

Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)
[1,1]	1.0	[2,4]	3.0	[4,2]	3.0
[1,2]	1.5	[2,5]	3.5	[4,3]	3.5
[1,3]	2.0	[3,1]	2.0	[4,4]	4.0
[1,4]	2.5	[3,2]	2.5	[4,5]	4.5
[1,5]	3.0	[3,3]	3.0	[5,1]	3.0
[2,1]	1.5	[3,4]	3.5	[5,2]	3.5
[2,2]	2.0	[3,5]	4.0	[5,3]	4.0
[2,3]	2.5	[4,1]	2.5	[5,4]	4.5
				[5,5]	5.0

Sampling Distribution

Sampling distribution of means

$\bar{X}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$



Sampling Distribution

More precisely, sampling distributions are probability distributions and used to describe the variability of sample statistics.

Definition : Sampling distribution

The sampling distribution of a statistics is the probability distribution of that statistics.

- The probability distribution of sample mean (hereafter, will be denoted as $\bar{X}$) is called the sampling distribution of the mean (also, referred to as the distribution of sample mean).
- Like $\bar{X}$, we call sampling distribution of variance (denoted as S^2).
- Using the values of $\bar{X}$ and S^2 for different random samples of a population, we are to make inference on the parameters μ and σ^2 (of the population).

Theorem on Sampling Distribution

- Famous theorem in Statistics

Theorem 5.1: Sampling distribution of mean and variance

The sampling distribution of a random sample of size n drawn from a population with mean μ and variance σ^2 will have mean $\bar{X} = \mu$ and variance $S^2 = \frac{\sigma^2}{n}$

Example 5.2: With reference to data in Example 5.1

For the population, $\mu = \frac{1+2+3+4+5}{5} = 3$

$$\sigma^2 = \frac{(25-1)}{12} = 2$$

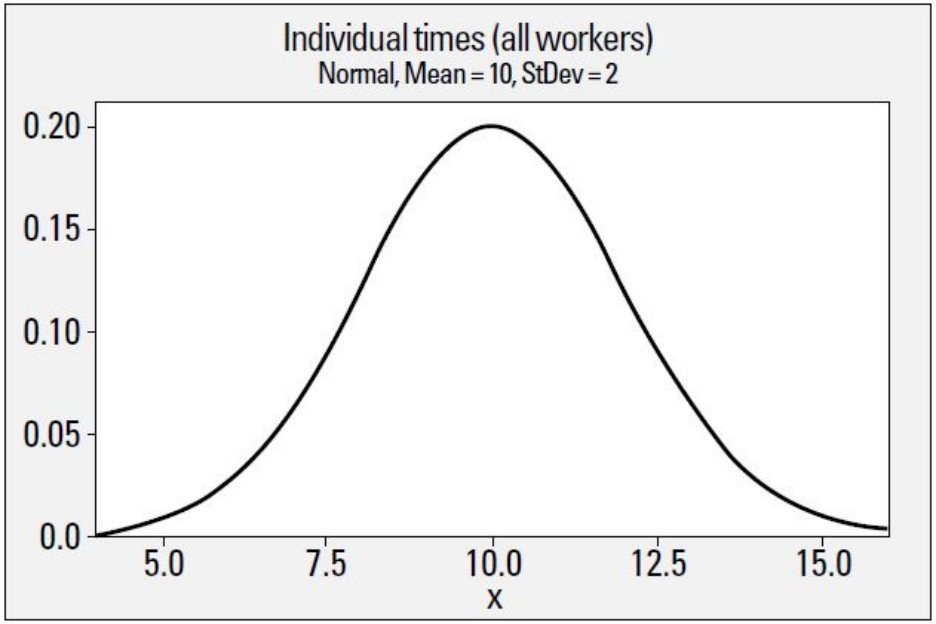
Applying the theorem, we have $\bar{X} = 3 \; and \; S^2 = 1$

Hence, the theorem is verified!

Example

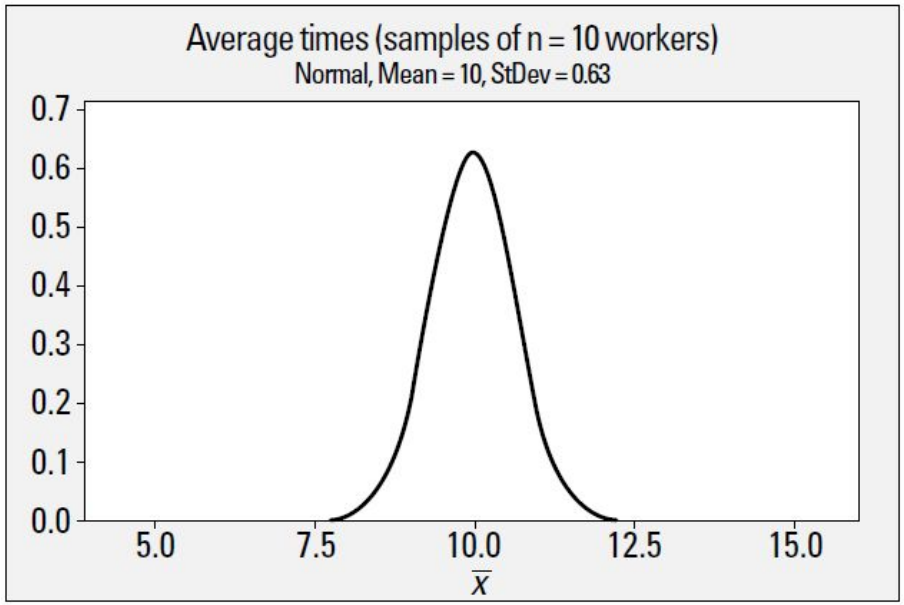
- Suppose *X* is the time it takes for a worker to type and send 10 letters of recommendation.

- Suppose *X* has a normal distribution with mean 10 minutes and standard deviation 2 minutes.



Example

- Now take a random sample of 10 workers, measure their times, and find the average, each time.
- Repeat this process over and over, and graph all of the possible results for all possible samples



Two Key Properties

- Std dev of such a distribution is called **standard error of the mean** and is computed as $\sigma/\sqrt{n}$
  - Thus large sample sizes have less sampling error

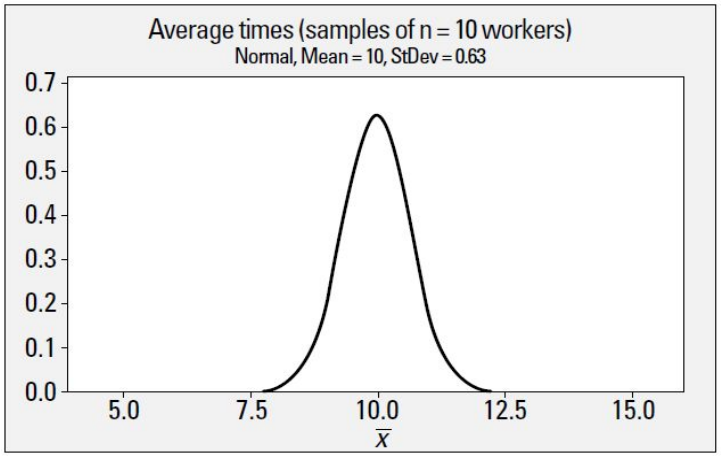
- **Central Limit Theorem**

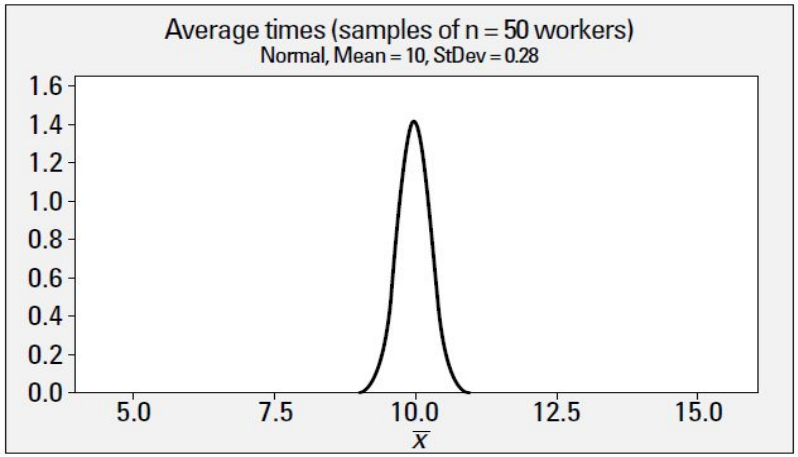
Example

- The average times are closer to 10 than the individual times
- Larger sample sizes mean more precision and less change from sample to sample
- Standard error =

$$\frac{\sigma_x}{\sqrt{n}} = \frac{2}{\sqrt{10}} = 0.63$$

$$\frac{2}{\sqrt{50}} = 0.28$$





Shape of Sampling Distribution

- If *X* has a normal distribution, then sampling distribution is also normal
- If the *X* distribution is *any* distribution that is not normal, or if its distribution is unknown, you can't automatically say the sample means have a normal distribution.
- But you can approximate its distribution with a normal distribution — if the sample size is large enough. This result is due to the *Central Limit Theorem*

Central Limit Theorem

- The Theorem 5.1 is an amazing result and in fact, also verified that if we sampling from a population with unknown distribution, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ provided that the sample size is large.

  This further, can be established with the famous "central limit theorem", which is stated below.

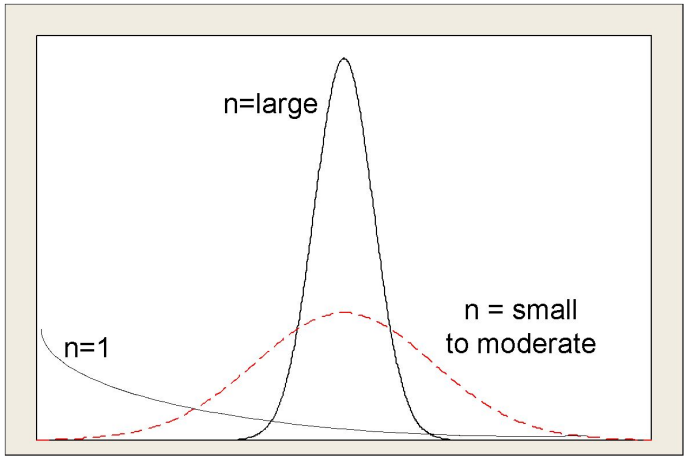
Theorem 5.3: Central Limit Theorem

If random samples each of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{X}$ will have a distribution approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.

The approximation becomes better as n increases.

Applicability of Central Limit Theorem

- The normal approximation of $\overline{X}$ will generally be good if $n \geq 30$
- The sample size $n = 30$ is, hence, a guideline for the central limit theorem.
- The normality on the distribution of $\overline{X}$ becomes more accurate as n grows larger.



One very important application of the **Central Limit Theorem** is the determination of reasonable values of the population mean μ and variance σ^2.

For standard normal distribution, we have the z-transformation

$$Z = \frac{\overline{X} - \mu}{S} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Central Limit Theorem

- One of the most important practical results in statistics
- If the sample size is large enough,
  - the sampling distribution of the mean is approximately normally distributed regardless of the distribution of the population
  - the mean of sampling distribution will be the same as that of the population
  - if the population is normally distributed, then the sampling distribution of the mean will also be normal for **any** sample size
- Most statisticians agree that if *n* is at least 30, it will do a reasonable job in most cases

Standard Sampling Distributions

- Apart from the normal distribution to describe sampling distribution, there are some other quite different sampling, which are extensively referred in the study of statistical inference.
  - χ^2: Describes the distribution of variance.
  - t: Describes the distribution of normally distributed random variable standardized by an estimate of the standard deviation.
  - F: Describes the distribution of the ratio of two variables.

Application 1

- Suppose *X* is the time it takes a worker to type and send 5 letters of recommendation.
- Suppose *X* (the times for all the workers) has a normal distribution and the reported mean is 10 minutes and the standard deviation 2 minutes
- You take a random sample of 50 workers and measure their times. What is the chance that their average time is less than 9.5 minutes?
- Essentially,

$$P(\bar{X} < 9.5) \qquad Z = \frac{\bar{X} - \mu_x}{\sigma_x / \sqrt{n}} \qquad P(Z < -1.77)$$

Application 2

- Size of individual customer orders, X, from a major discount book publisher web site is normally distributed with mean of $36 and std dev of $8

- Probability that next individual who places an order will make a purchase of > $40?
  - 1 – P(X < $40) = 1 – 0.6915 = 0.3085
  -

- Sample of 16 customers is chosen

- What is the probability that the **mean purchase for these 16 customers** will exceed $40?
  - Use sampling distribution of mean
  - Mean = $36, Std Dev = $2 (std error of mean)
  - 1-P(M<$40) = 1 – 0.9772 = 0.0228

Note

- If *n* is not large enough for the CLT, you use the *t*-distribution in many cases

Estimation

- Involves assessing the value of an unknown population parameter (mean, std dev etc.) using sample data
- An estimator of a population parameter an approximation depending solely on sample information
- Measures used for estimation are called **estimators**
- A specific value is called an estimate.
- For example, sample variance is an estimator of variance of population

Estimation

- **Point estimate**
  - Single number
  - Example: sample mean, sample variance

- **Interval estimate**
  - Range

- Point estimate is located exactly in the middle of the confidence interval.
- Confidence intervals provide much more information and are preferred when making inferences

- Two characteristics
  - Bias
  - Precision / Efficiency

efficient unbiased estimators

Bias

- Sample 1, point estimate 1; sample 2, point estimate 2; sample 3, point estimate 3; and so on
- Each individual point estimate will vary from the actual population parameter
- However, we would hope that expected value of all such point estimates would equal the actual population parameter
  - Then it is called **unbiased estimator**
  - Otherwise **biased**

Example

- You might estimate the average household income (parameter) based on the average household income from a random sample of 1,000 homes (statistic)

- Because sample results will vary, you need to add a measure of that variability to your estimate
  - This measure of variability is called the margin of error
  - The heart of a confidence interval

- The margin of error is not the chance a mistake was made; it measures variation in the random samples due to chance.

Efficient estimators are the ones with the least variability of outcomes

Point Estimates

- Sample mean

- Sample variance
  - It is not the variance of the sample
  - BUT an estimation of variance of population based on this sample

Interval Estimates

- Provides a range for a population characteristic based on a sample
- Provide more information than a point estimate
- Typically point estimate +- margin of error
- How to calculate margin of error?

Confidence Intervals

- Provides a way of assessing the accuracy of point estimates
- Range of values between which the value of the population parameter is believed to be along with a probability that the interval correctly estimates the true (unknown) population parameter
  - Called **level of confidence**
- Margin of error depends on level of confidence and sample size
- The confidence level describes the uncertainty associated with a sampling method

What it really means?

- Sample 1
  - Margin of error at 95% level of confidence and for this sample size is calculated to be 2
  - Point estimate = 10
  - Confidence interval = [8, 12]

- Sample 2
  - [8.4, 12.4]
- Neither of them may actually contain the real population parameter
- But out of 100 such samples, 95 would
- When not stated, 95% is the assumed level of confidence

Misinterpretation

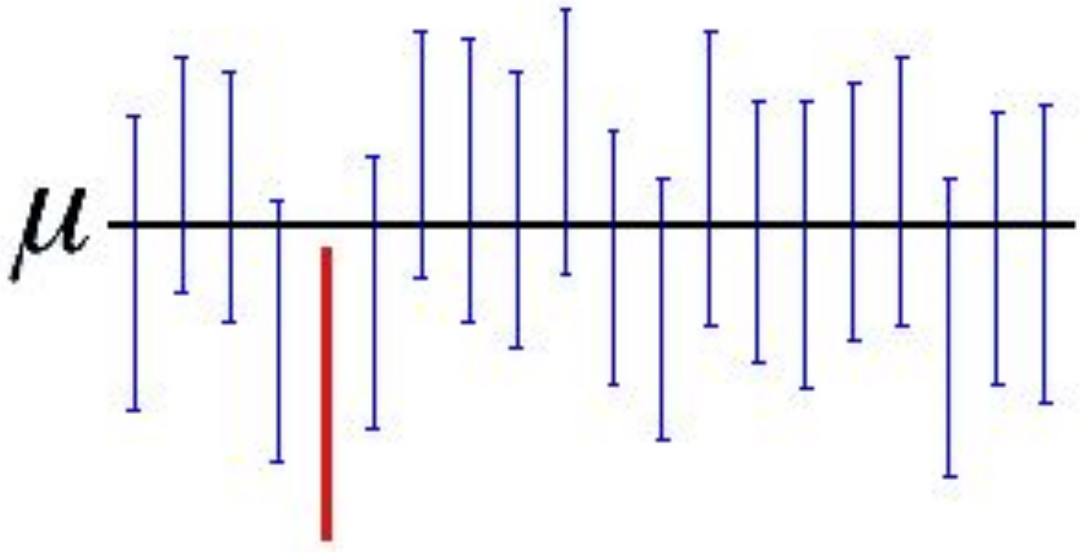
- I am 95% confident that the population mean is between A and B
  - Wrong!
  - **Confidence** intervals are not **probability** intervals

- Once your sample has been selected and your confidence interval is calculated, it either contains the population parameter or it doesn't; there is no probability involved

- Confidence level (in this case 95%) does not apply to a single confidence interval.

- Percentage of all possible samples of size *n* whose confidence intervals contain the population parameter

Correct Understanding

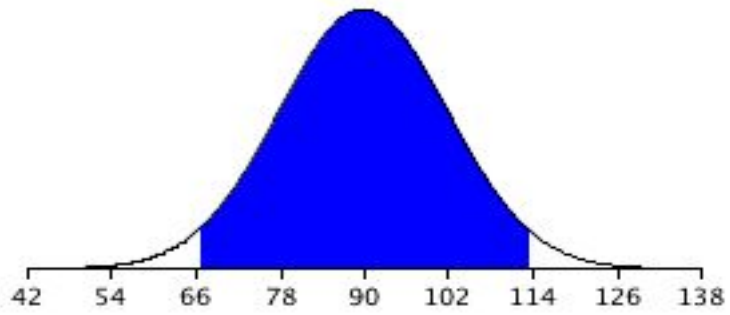


A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

Insight behind confidence intervals

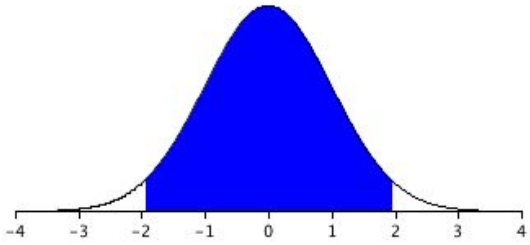
- Assume that the weights of 10-year-old children are normally distributed with a mean of 90 and a standard deviation of 36. What is the sampling distribution of the mean for a sample size of 9?

- Interval which would contain middle 95% of the distribution?



Writing in terms of z

- We want the z-value corresponding to, say 95% of area
  - **95% level of confidence**
- i.e. Find x such that P(-x < z < x) = 0.95
- i.e. Find x such that P(z<x) = 1 - (1-0.95)/2
- i.e. Find x such that F(z=x) = 1 - α/2
- Let us call this $z_{\alpha/2}$
- **CI = sample mean ± $z_{\alpha/2}$ ($\sigma/\sqrt{n}$)**



Example

- Automated process of filling bottles with liquid detergent
- Historical standard deviation is 15ml
- In filling 800ml bottles, a sample of 25 bottles found an average volume of 796ml
- What is the confidence interval of estimated population mean at 95%level of confidence?
- 796 +- 1.96 (15/√25) = [790.12, 801.88]
- Does this indicate a serious problem?
  - Not necessarily. Sample does not provide sufficient evidence that population mean is less than 800

Example

- If, however, the sample average was found to be 792
- Confidence interval turns out to be [786.12, 797.88]
- It is thus highly unlikely that population mean will be 800
- Manufacturer should check the equipment to maintain quality!

What to do when σ is unknown?

- Use t-distribution
  - Cousin of z-distribution
- Distribution of a mean divided by its estimate of the standard error
- Also called Student's t-distribution
  - W.S.Gosste worked for a brewery and because of a contractual agreement, published the paper under the pseudonym "Student"

T-Distribution

- Has larger variance than std normal, thus making confidence intervals wider, in essence correcting for the uncertainty about true (unknown) std dev
- As the df increases, t-distribution converges to z
- For sample sizes > 100, it is as good as z
- Even for sample sizes about 30-35, not much of a difference
- For any sample size, true sampling distribution of mean is t-distribution, so when in doubt, use t

T-Distribution

- However, with smaller sample sizes, the t distribution is leptokurtic, which means it has relatively more scores in its tails than does the normal distribution

- As a result, you have to extend farther from the mean to contain a given proportion of the area

- Recall that with a normal distribution, 95% of the distribution is within 1.96 standard deviations of the mean. Using the t distribution, if you have a sample size of only 5, 95% of the area is within 2.78 standard deviations of the mean. Therefore, the standard error of the mean would be multiplied by 2.78 rather than 1.96

T-table

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Example

- Following five numbers are sampled from a normal distribution: 2, 3, 5, 6, and 9; standard deviation of population is not known
- Sample mean? Sample variance?
  - 5, 7.5
- Standard error of mean?
- Instead, let us estimate the standard error of mean

$$s_M = \frac{s}{\sqrt{N}} = 1.225$$

- T-value?
  - 95% and 4 degrees of freedom

Degrees of Freedom

- The degrees of freedom of an estimate is the number of independent pieces of information that go into the estimate
- In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question
- For example, to estimate the population variance, one must first estimate the population mean
- Therefore, if the estimate of variance is based on N observations, there are N-1 degrees of freedom

Example

- CI = sample mean ± $t_{\alpha/2,\, n-1}\, s/\sqrt{n}$
- $t_{\alpha/2,\, n-1}$ = value of t distribution with n-1 degrees of freedom that has a cumulative probability of 1 – (α/2)

$$\text{Lower limit} = 5 - (2.776)(1.225) = 1.60$$
$$\text{Upper limit} = 5 + (2.776)(1.225) = 8.40$$

- Degrees of freedom = sample size - 1

Steps to find confidence interval

- Choose your confidence level and your sample size
- Select a random sample of individuals from the population
- Collect reliable and relevant data from the individuals in the sample
- Summarize the data into a statistic (for example, a sample mean)
- Calculate the margin of error
- Take the statistic plus or minus the margin of error to get your final estimate of the parameter

Goal: Accuracy vs Precision

- Have confidence interval as narrow as possible
- You want to think about this issue before collecting your data; after the data are collected, the width of the confidence interval is set
- Factors affecting margin of error
  - Confidence level
  - Sample size
  - Amount of variability in population
- Sample statistic itself only determines the midpoint of the confidence interval, not its width

Goal: Accuracy vs Precision

- As the level of confidence decreases, confidence interval becomes narrower
  - 99% confidence interval will be wider than 95% confidence interval

- Smaller risk (that actual value doesn't fall in the interval) □ wider confidence interval □ lesser accuracy

- Thus, instead, consider increasing the sample size to reduce the risk

Hypothesis Testing

What is Hypothesis?

- "A hypothesis is an educated prediction that can be tested" (study.com).
- "A hypothesis is a proposed explanation for a phenomenon" (Wikipedia).
- "A hypothesis is used to define the relationship between two variables" (Oxford dictionary).
- "A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation" (Walpole).

- **Example: Avogadro's Hypothesis(1811)**

"The volume of a gas is directly proportional to the number of molecules of the gas."

$$V = aN$$

Statistical Hypothesis

- If the hypothesis is stated in terms of population parameters (such as mean and variance), the hypothesis is called statistical hypothesis.
- Data from a sample (which may be an experiment) are used to test the validity of the hypothesis.
- A procedure that enables us to agree (or disagree) with the statistical hypothesis is called a **test of the hypothesis.**

Example :

1. To determine whether the wages of men and women are equal.
2. A product in the market is of standard quality.
3. Whether a particular medicine is effective to cure a disease.

The Hypotheses

- The main purpose of statistical hypothesis testing is to choose between two competing hypotheses.

Example 6.3:

One hypothesis might claim that wages of men and women are equal, while the **alternative** might claim that men make more than women.

- Hypothesis testing start by making a set of two statements about the parameter(s) in question.
- The hypothesis actually to be tested is usually given the symbol H_0 and is commonly referred as the **null hypothesis.**
- The other hypothesis, which is assumed to be true when null hypothesis is false, is referred as the **alternate hypothesis** and is often symbolized by H_1
- The two hypotheses are exclusive and exhaustive.

The Hypotheses

Example 6.4:

Ministry of Human Resource Development (MHRD), Government of India takes an initiative to improve the country's human resources and hence set up **23 IIT's** in the country.

To measure the engineering aptitudes of graduates, MHRD conducts GATE examination for a mark of 1000 in every year. A sample of 300 students who gave GATE examination in 2018 were collected and the mean is observed as 220.

In this context, statistical hypothesis testing is to determine the mean mark of the all GATE-2018 examinee.

The two hypotheses in this context are:

$$H_0: \mu = 220$$

$$H_1: \mu < 220$$

The Hypotheses

Note:

1. As null hypothesis, we could choose $H_0: \mu \leq 220$ or $H_0: \mu \geq 220$
2. It is customary to always have the null hypothesis with an equal sign.
3. As an alternative hypothesis there are many options available with us.

Examples 6.5:

I. $H_1: \mu > 220$

II. $H_1: \mu < 220$

III. $H_1: \mu \neq 220$

4. The two hypothesis should be chosen in such a way that they are **exclusive** and **exhaustive**.
   - One or other must be true, but they cannot both be true.

The Hypotheses

One-tailed test

- A statistical test in which the alternative hypothesis specifies that the population parameter lies entirely above or below the value specified in H_0 is called a one-sided (or one-tailed) test.

Example.

$$H_0: \mu = 100 \qquad H_1: \mu > 100$$

Two-tailed test

- An alternative hypothesis that specifies that the parameter can lie on their sides of the value specified by H_0 is called a two-sided (or two-tailed) test.

Example.

$$H_0: \mu = 100 \qquad H_1: \mu <> 100$$

The Hypotheses

Note:

In fact, a 1-tailed test such as:

$$H_0: \mu = 100$$

$$H_1: \mu > 100$$

is same as

$$H_0: \mu \leq 100$$

$$H_1: \mu > 100$$

In essence, $\mu > 100$, it does not imply that $\mu > 80, \mu > 90$, etc.

Hypothesis Testing Procedures

The following five steps are followed when testing hypothesis

1. Specify H_0 and H_1, the null and alternate hypothesis, and an **acceptable level of α.**
2. Determine an appropriate sample-based test statistics and the **rejection region** for the specified H_0.
3. Collect the sample data and calculate the test statistics.
4. Make a decision to either reject or fail to reject H_0.
5. Interpret the result in common language suitable for practitioners.

Hypothesis Testing Procedure

- In summary, we have to choose between H_0 and H_1
- The standard procedure is to assume H_0 is true.
  (Just we presume innocent until proven guilty)
- Using statistical test, we try to determine whether there is sufficient evidence to declare H_0 false.
- We reject H_0 only when the **chance is small** that H_0 is true.
- The procedure is based on probability theory, that is, there is a chance that we can make errors.

Errors in Hypothesis Testing

In hypothesis testing, there are two types of errors.

Type I error: A type I error occurs when we incorrectly reject H_0 (i.e., we reject the null hypothesis, when H_0 is true).

Type II error: A type II error occurs when we incorrectly fail to reject H_0 (i.e., we accept H_0 when it is not true).

	Observation	
Decision	H_0 is true	H_0 is false
H_0 is accepted	Decision is correct	Type II error
H_0 is rejected	Type I error	Decision is correct

Probabilities of Making Errors

Type I error calculation

α: denotes the probability of making a Type I error

$$\alpha = \mathbf{P}(\text{Rejecting } H_0 | H_0 \text{ is true})$$

Type II error calculation

β: denotes the probability of making a Type II error

$$\beta = \mathbf{P}(\text{Accepting } H_0 | H_0 \text{ is false})$$

Note:

- α and β are not independent of each other as one increases, the other decreases
- When the sample size increases, both to decrease since sampling error is reduced.
- In general, we focus on Type I error, but Type II error is also important, particularly when sample size is small.

Calculating α

Assuming that we have the results of random sample. Hence, we use the characteristics of sampling distribution to calculate the probabilities of making either Type I or Type II error.

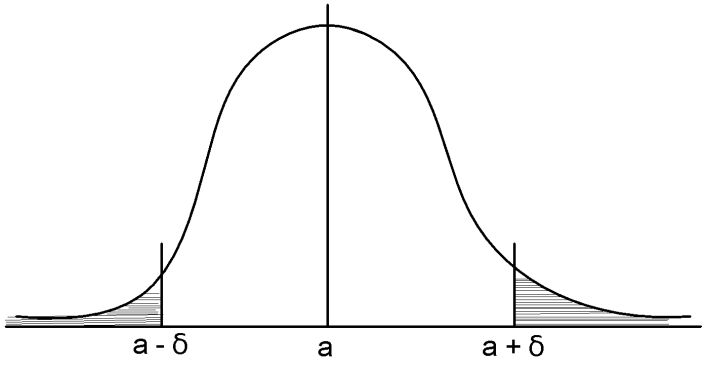
Example 6.6:

Suppose, two hypotheses in a statistical testing are:

$$H_0: \mu = a$$
$$H_1: \mu \neq a$$

Also, assume that for a given sample, population obeys normal distribution. A threshold limit say $a \pm \delta$ is used to say that they are significantly different from a.

Calculating α



Here, shaded region implies the probability that, $\overline{X} < a - \delta$ *or* $\overline{X} > a + \delta$

Thus the null hypothesis is to be rejected if the mean value is less than $a - \delta$ or greater than $a + \delta$.

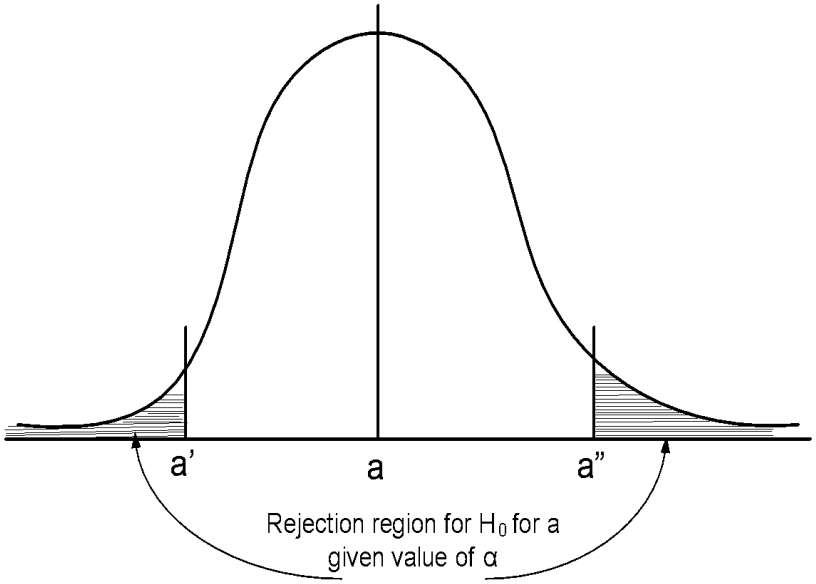
If $\overline{X}$ denotes the sample mean, then the Type I error is

$$\alpha = P(\overline{X} < a - \delta \text{ or } \overline{X} > a + \delta, \quad \text{when } \mu = a, \quad \text{i.e., } H_0 \text{ is true})$$

The Rejection Region

The rejection region comprises of value of the test statistics for which

1. The probability when the null hypothesis is true is less than or equal to the specified α.
2. Probability when H_1 is true are greater than they are under H_0.



Two-Tailed Test

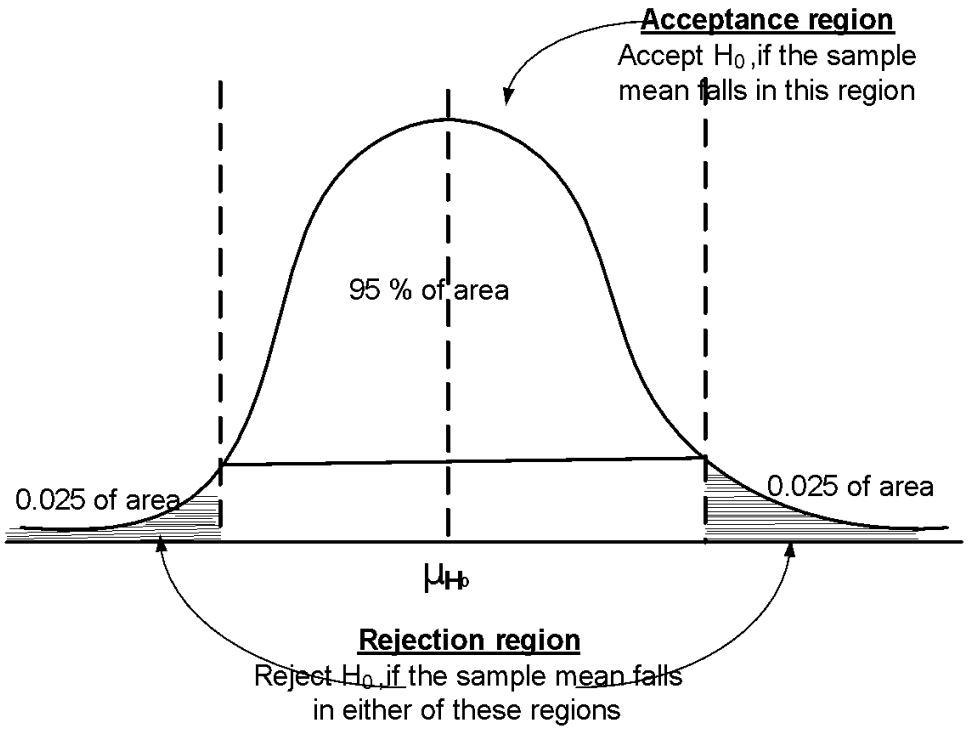
- For two-tailed hypothesis test, hypotheses take the form

$$H_0: \mu = \mu_{H_0}$$
$$H_1: \mu \neq \mu_{H_0}$$

In other words, to reject a null hypothesis, sample mean $\mu > \mu_{H_0}$ or $\mu < \mu_{H_0}$ under a given α.

Thus, in a two-tailed test, there are two rejection regions (also known as critical region), one on each tail of the sampling distribution curve.

Two-Tailed Test



Acceptance and rejection regions in case of a two-tailed test with 5% significance level.

One-Tailed Test

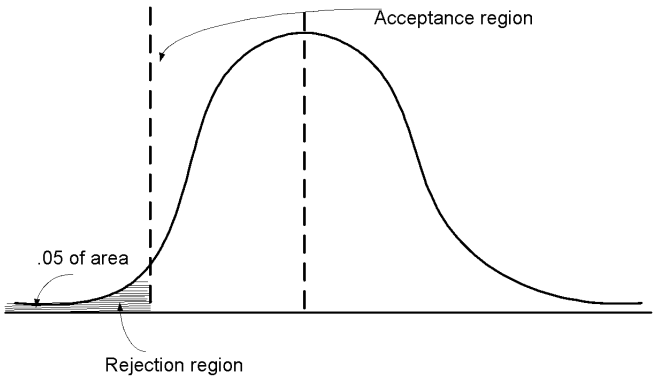
A one-tailed test would be used when we are to test, say, whether the population mean is either lower or higher than the hypothesis test value.

Symbolically,

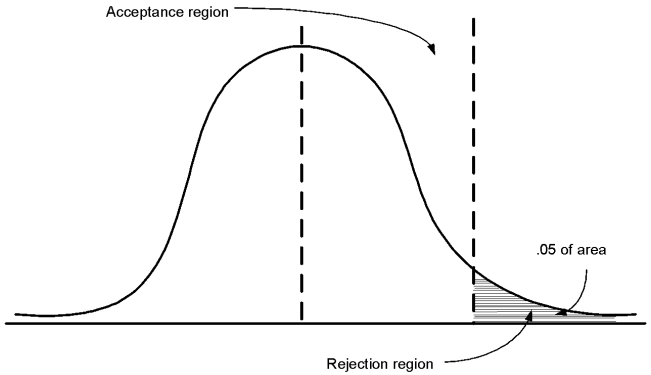
$$H_0: \mu = \mu_{H_0}$$

$$H_1: \mu < \mu_{H_0} \qquad [or\ \mu > \mu_{H_0}]$$

Wherein there is one rejection region only on the left-tail (or right-tail).



Left – tailed test



Right – tailed test

Example 6.7: Calculating α

Consider the two hypotheses are

The null hypothesis is

$$H_0: \mu = 8$$

The alternative hypothesis is

$$H_1: \mu \neq 8$$

Assume that given a sample of size 16 and standard deviation is 0.2 and sample follows normal distribution.

Example 6.7: Calculating α

We can decide the rejection region as follows.

Suppose, the null hypothesis is to be rejected if the mean value is less than 7.9 or greater than 8.1. If $\overline{X}$ is the sample mean, then the probability of Type I error is

$$\alpha = P(\overline{X} < 7.9 \; or \; \overline{X} > 8.1, \text{when } \mu = 8)$$

Given σ, the standard deviation of the sample is 0.2 and that the distribution follows normal distribution.

Thus,

$$P(\overline{X} < 7.9) = P\left[Z = \frac{7.9 - 8}{0.2/\sqrt{16}}\right] = P[Z < -2.0] = 0.0228$$

and

$$P(\overline{X} > 8.1) = P\left[Z = \frac{8.1 - 8}{0.2/\sqrt{16}}\right] = P[Z > 2.0] = 0.0228$$

Hence, $\alpha = 0.0228 + 0.0228 = 0.0456$

Example 6.8: Calculating α and β

There are two identically appearing boxes of chocolates. Box A contains 60 red and 40 black chocolates whereas box B contains 40 red and 60 black chocolates. There is no label on the either box. One box is placed on the table. We are to test the hypothesis that "Box B is on the table".

To test the hypothesis an experiment is planned, which is as follows:

- Draw at random five chocolates from the box.
- We replace each chocolates before selecting a new one.
- The number of red chocolates in an experiment is considered as the sample statistics.

Note: Since each draw is independent to each other, we can assume the sample distribution follows binomial probability distribution.

Example 6.8: Calculating α

Let us express the population parameter as p = the number of red chocolates in Box B.

The hypotheses of the problem can be stated as:

$H_0: p = 0.4$ // Box B is on the table

$H_1: p = 0.6$ // Box A is on the table

Calculating α:

In this example, the null hypothesis (H_0) specifies that the probability of drawing a red chocolate is 0.4. This means that, lower proportion of red chocolates in observations ($i.e., sample$) favors the null hypothesis. In other words, drawing all red chocolates provides sufficient evidence to reject the null hypothesis. Then, the probability of making a $Type\ I$ error is the probability of getting five red chocolates in a sample of five from Box B. That is,

$$\alpha = P(X = 5 \quad when\ p = 0.4)$$

Using the binomial distribution

$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}\ where\ n = 5, x = 5$$

$$= (0.4)^5 = 0.01024$$

Thus, the probability of rejecting a true null hypothesis is ≈ 0.01. That is, there is approximately 1 in 100 chance that the box B will be mislabeled as box A.

Example 6.8: Calculating β

The *Type II* error occurs if we fail to reject the null hypothesis when it is not true. For the current illustration, such a situation occurs, if Box A is on the table but we did not get the five red chocolates required to reject the hypothesis that Box B is on the table.

The probability of *Type II* error is then the probability of getting four or fewer red chocolates in a sample of five from Box A.

That is,

$$\beta = P(X \leq 4 \qquad \text{when } p = 0.6)$$

Using the probability rule:

$$P(X \leq 4) + P(X = 5) = 1$$

$$\text{That is, } P(X \leq 4) = 1 - P(X = 5)$$

$$\text{Now, } P(X = 5) = (0.6)^5$$

$$\text{Hence, } \beta = 1 - (0.6)^5$$

$$= 1 - 0.07776 = 0.92224$$

That is, the probability of making *Type II* error is over 92%. This means that, if Box A is on the table, the probability that we will be unable to detect it is 0.92.

Case Study 1: **Coffee Sale**

A coffee vendor nearby city railway station has been having average sales of 500 cups per day. Because of the development of a bus stand nearby, it expects to increase its sales. During the first 12 days, after the inauguration of the bus stand, the daily sales were as under:

550 570 490 615 505 580 570 460 600 580 530 526

On the basis of this sample information, can we conclude that the sales of coffee have increased?

Consider 5% level of significance.

Hypothesis Testing : 5 Steps

The following five steps are followed when testing hypothesis

1. Specify H_0 and H_1, the null and alternate hypothesis, and an **acceptable level of α.**
2. Determine an appropriate sample-based test statistics and the **rejection region** for the specified H_0.
3. Collect the sample data and calculate the test statistics.
4. Make a decision to either reject or fail to reject H_0.
5. Interpret the result in common language suitable for practitioner.

Case Study 1: **Step 1**

Step 1: Specification of hypothesis and acceptable level of α

Let us consider the hypotheses for the given problem as follows.

$H_0: \mu = 500$ cups per day

The null hypothesis that sales average 500 cups per day and they have not increased.

$H_0: \mu > 500$

The alternative hypothesis is that the sales have increased.

Given the acceptance level of $\alpha = 0.05$ $(i.e., 5\%\ level\ of\ significance)$

Case Study 1: **Step 2**

Step 2: Sample-based test statistics and the rejection region for specified H_0

Given the sample as

550 570 490 615 505 580 570 460 580 530 526

Since the sample size is small and the population standard deviation is not known, we shall use $t - test$ assuming normal population. The test statistics t is

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

To find $\bar{X}$ and S, we make the following computations.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{6576}{12} = 548$$

Case Study 1: **Step 2**

Sample #	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	550	2	4
2	570	22	484
3	490	−58	3364
4	615	67	4489
5	505	−43	1849
6	580	32	1024
7	570	22	484
8	460	−88	7744
9	600	52	2704
10	580	32	1024
11	530	−18	324
12	526	−22	484
$n = 12$	$\sum X_i = 6576$		$\sum (X_i - \bar{X})^2 = 23978$

Case Study 1: **Step 2**

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{23978}{12-1}} = 46.68$$

Hence, $t = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{48}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$

Note:

Statistical table for t-distributions gives a *t*-value given *n*, the degrees of freedom and α, the level of significance and vice-versa.

Case Study 1: **Step 3**

Step 3: Collect the sample data and calculate the test statistics

$$Degree\ of\ freedom = n - 1 = 12 - 1 = 11$$

As H_1 is one-tailed, we shall determine the rejection region applying one-tailed in the right tail because H_1 is more than type) at 5% level of significance.

Case Study 1: **Step 3**

Step 3: Collect the sample data and calculate the test statistics

$$Degree\ of\ freedom = n - 1 = 12 - 1 = 11$$

As H_1 is one-tailed, we shall determine the rejection region applying one-tailed in the right tail because H_1 is more than type) at 5% level of significance.

Using table of $t - distribution$ for 11 degrees of freedom and with 5% level of significance,

$$R: t > 1.796$$

Case Study 1: **Step 4**

Step 4: Make a decision to either reject or fail to reject H_0

The observed value of $t = 3.558$ which is in the rejection region and thus H_0 is rejected at 5% level of significance.

Case Study 1: **Step 5**

Step 5: Final comment and interpret the result

We can conclude that the sample data indicate that coffee sales have increased.

Case Study 2: **Machine Testing**

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the amount of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean amount of medicine in these 16 tubes will be used to test the hypothesis that the machine is indeed working properly.

Case Study 2: **Step 1**

Step 1: Specification of hypothesis and acceptable level of α

The hypotheses are given in terms of the population mean of medicine per tube.

The null hypothesis is

$$H_0: \mu = 8$$

The alternative hypothesis is

$$H_1: \mu \neq 8$$

We assume α, the significance level in our hypothesis testing ≈ 0.05.
(This signifies the probability that the machine needs to be adjusted less than 5%).

Case Study 2: **Step 2**

Step 2: Sample-based test statistics and the rejection region for specified H_0

Rejection region: Given $\alpha = 0.05$, which gives $|Z| > 1.96$ (obtained from standard normal calculation for $n(Z: 0,1) = 0.025$ for a rejection region with two-tailed test).

Case Study 2: **Step 3**

Step 3: Collect the sample data and calculate the test statistics

Sample results: $n = 16,\ \bar{x} = 7.89, \sigma = 0.2$

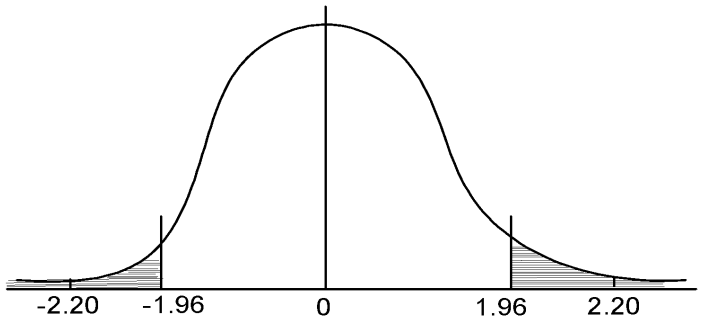
With the sample, the test statistics is

$$Z = \frac{7.89-8}{0.2/\sqrt{16}} = -2.20$$

Hence, $|Z| = 2.20$

Case Study 2: **Step 4**

Step 4: Make a decision to either reject or fail to reject H_0



Since $Z > 1.96$, we reject H_0

Case Study 2: **Step 5**

Step 5: Final comment and interpret the result

We conclude $\mu \neq 8$ and recommend that the machine be adjusted.

Case Study 2: **Alternative Test**

Suppose that in our initial setup of hypothesis test, if we choose $\alpha = 0.01$ instead of 0.05, then the test can be summarized as:

1. $H_0: \mu = 8$, $H_1: \mu \neq 8$ $\alpha = 0.01$
2. Reject H_0 if $Z > 2.576$
3. Sample result n =16, $\sigma = 0.2$, $\overline{X}$=7.89, $Z = \frac{7.89-8}{0.2/\sqrt{16}} = -2.20$, $|Z| = 2.20$
4. $|Z| < 2.20$, we fail to reject H_0= 8
5. We do not recommend that the machine be readjusted.

Hypothesis Testing Strategies

- The hypothesis testing determines the validity of an assumption (technically described as null hypothesis), with a view to choose between two conflicting hypothesis about the value of a population parameter.

- There are two types of tests of hypotheses
  - ✔ Non-parametric tests (also called distribution-free test of hypotheses)
  - ✔Parametric tests (also called standard test of hypotheses).

Parametric Tests : Applications

- Usually assume certain properties of the population from which we draw samples.
  - Observation come from a normal population
  - Sample size is small
  - Population parameters like mean, variance, etc. are hold good.
  - Requires measurement equivalent to interval scaled data.

Parametric Tests

Important Parametric Tests

The widely used sampling distribution for parametric tests are

- $Z - test$
- $t - test$
- $\chi^2 - test$
- $F - test$

Note:

All these tests are based on the assumption of normality (i.e., the source of data is considered to be normally distributed).

Parametric Tests : Z-test

***Z – test*: This is most frequently test in statistical analysis.**

- It is based on the normal probability distribution.
- Used for judging the significance of several statistical measures particularly the mean.
- It is used even when *binomial distribution* or $t - distribution$ is applicable with a condition that such a distribution tends to normal distribution when *n* becomes large.
- Typically it is used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when **population variance** is known.

Parametric Tests : t-test

***t − test*: It is based on the t-distribution.**

- It is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of
  - small sample(s)
  - **population variance is not known** (in this case, we use the variance of the sample as an estimate of the population variance)

Parametric Tests : χ^2 -test

$\chi^2 - test$: It is based on Chi-squared distribution.

- It is used for comparing a sample variance to a theoretical population variance.

Parametric Tests : F -test

$F-test$: It is based on F-distribution.

- It is used to compare the variance of two independent samples.
- This test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means.

Hypothesis Testing : Assumptions

Case 1: Normal population, population infinite, sample size may be large or small, variance of the population is known.

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}}$$

Case 2: Population normal, population finite, sample size may large or small……variance is known.

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}[\sqrt{(N-n)/(N-1)}]}$$

Case 3: Population normal, population infinite, sample size is small and variance of the population is unknown.

$$t = \frac{\bar{X} - \mu_{H_0}}{S/\sqrt{n}} \qquad with\ degree\ of\ freedom = (n-1)$$

and

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{(n-1)}}$$

Hypothesis Testing

Case 4: Population finite

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}\left[\sqrt{(N-n)/(N-1)}\right]} \; with\ degree\ of\ freedom = (n-1)$$

Note: If variance of population (σ) is known, replace S by σ. Population normal, population infinite, sample size is small and variance of the population is unknown.

Hypothesis Testing : Non-Parametric Test

- ***Non-Parametric tests***

✔Does not under any assumption

✔Assumes only nominal or ordinal data

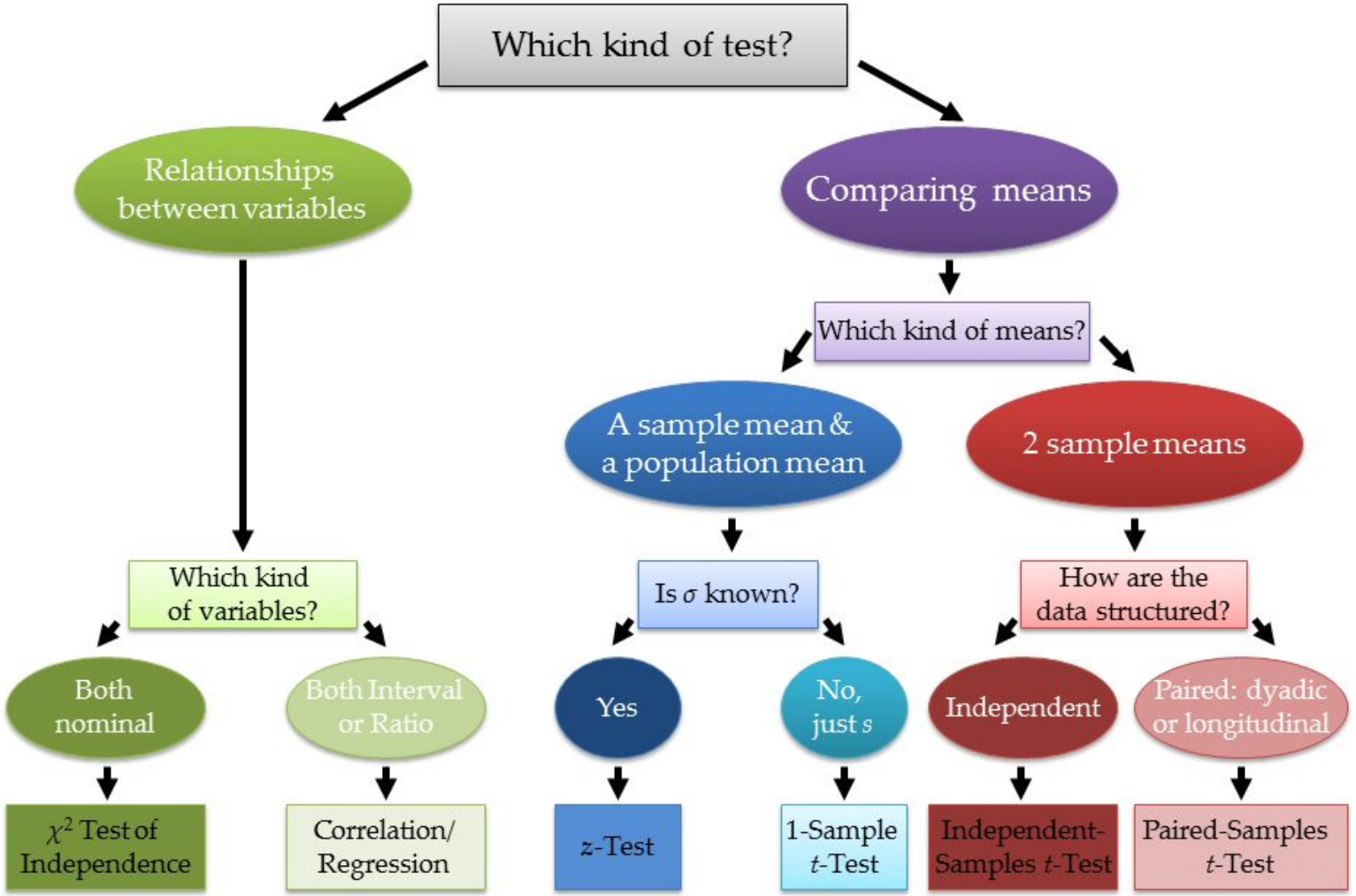
Note: Non-parametric tests need entire population (or very large sample size)

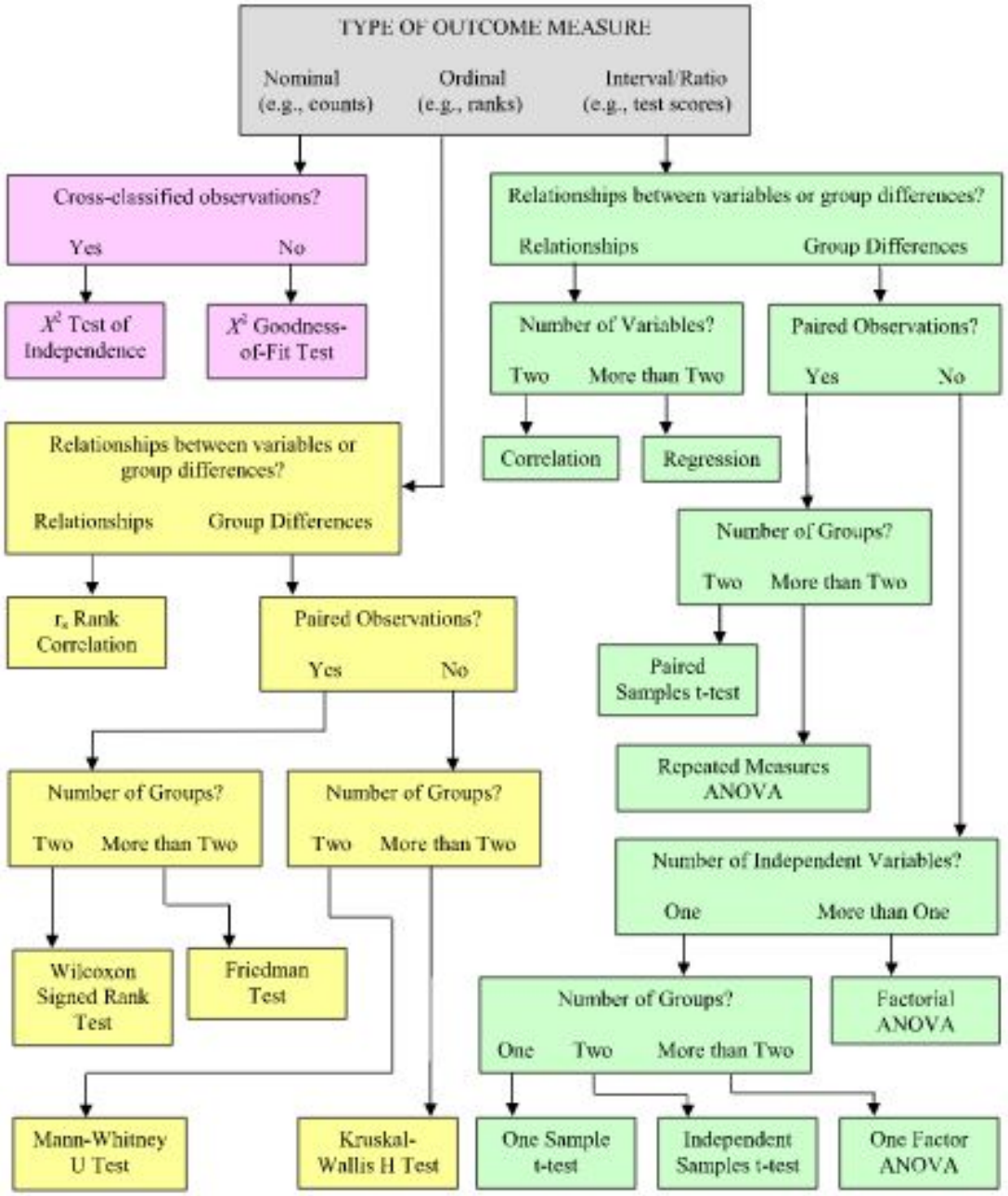
Reference

The detail material related to this can be found in

Probability and Statistics for Engineers and Scientists (8th Ed.) by Ronald E. Walpole, Sharon L. Myers, Keying Ye (Pearson), 2013.

Which test in which scenario?





http://case.truman.edu/files/2015/06/ProperTest-1.pdf