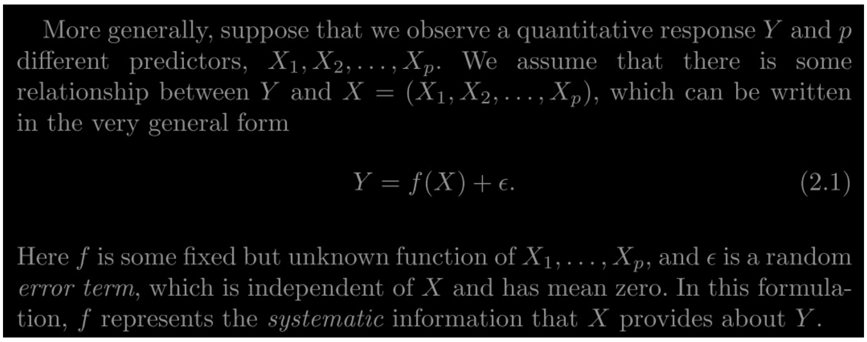
Statistical Learning : The extract form of prediction function is unknown but the predicted values are calculated.

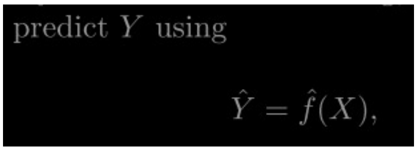


More generally, suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

$$Y = f(X) + \epsilon. \qquad (2.1)$$

Here f is some fixed but unknown function of $X_1, \ldots, X_p$, and ϵ is a random *error term*, which is independent of X and has mean zero. In this formulation, f represents the *systematic* information that X provides about Y.

Prediction :

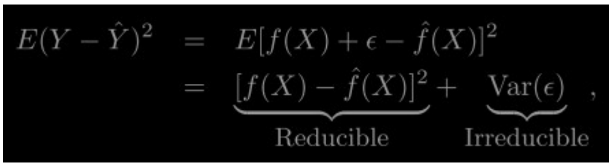


predict Y using

$$\hat{Y} = \hat{f}(X),$$

# Accuracy of the predictions Y depends upon two factors :

a. Reducible Error: The error is Reducible if we can potentially improve the accuracy of function f.
b. Irreducible Error: Irreducible errors cannot be improved by any means and variability associated with epsilon affects the accuracy of the predictions. It provides an upper bound on the accuracy of our prediction and this bound is almost unknown.



$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}, \end{aligned}$$

Inference : The prediction function is known properly and the prediction values are calculated but the effect of independent variables on the dependent (target) variables is estimated.

*Descriptive Statistics: Collecting, Presenting and describing data
*Inferential Statistics: Drawing conclusions and/or making decisions concerning a population-based only on the sample data.

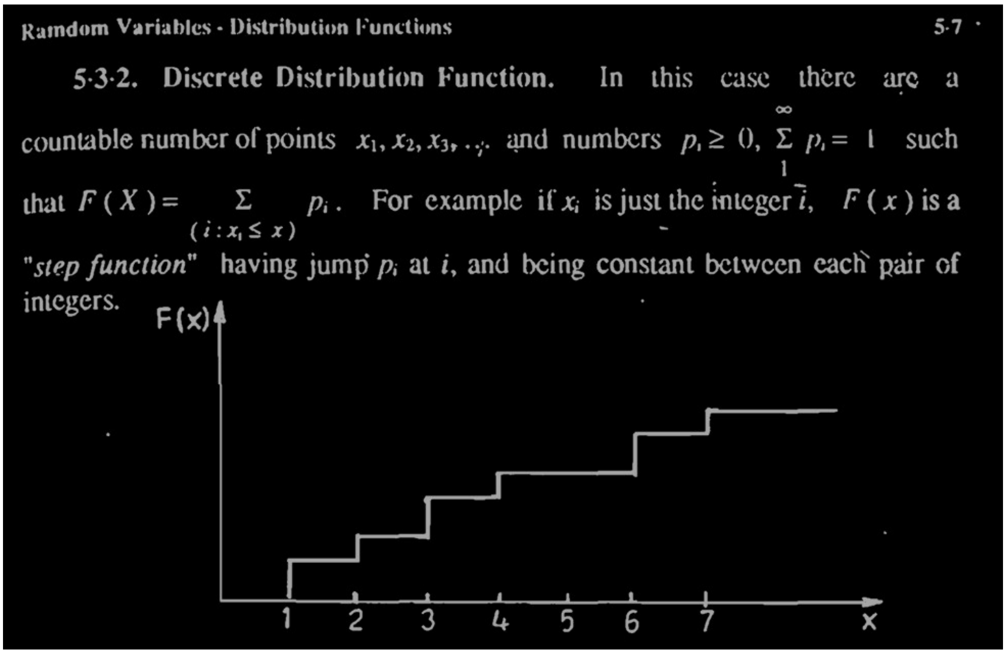
Why Sample ?

1. Less time consuming and less costly to administer than a census.
2. Possible to obtain high precision based on samples.
3. Sampling is only option when access to the population is not possible.

Random Variables: By a Random Variable (r,v) we mean a real number X connected with the outcome of a random experiment E.

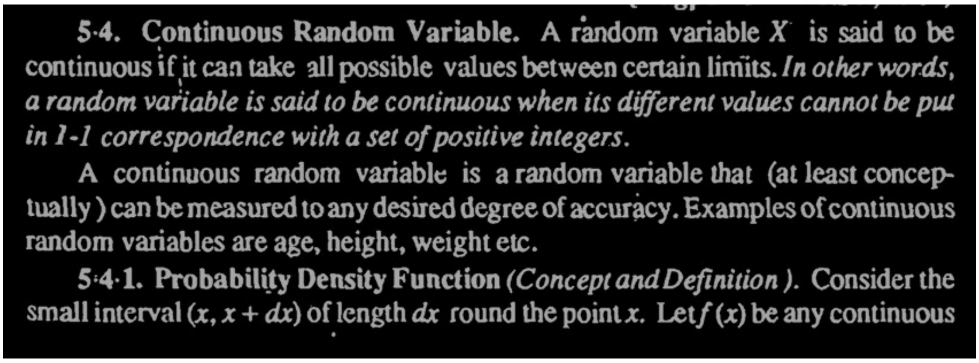
A random variable is a rule (real-valued function) that assigns a numerical value to each outcome in a sample space.

# The domain of the Random Variable is a Sample Space.
# Discrete Random Variable: A discrete R.V can take only finite distinct values of the provided range.



5·3·2. Discrete Distribution Function. In this case there are a countable number of points $x_1, x_2, x_3, \ldots$ and numbers $p_i \geq 0$, $\sum_{1}^{\infty} p_i = 1$ such that $F(X) = \sum_{(i : x_i \leq x)} p_i$. For example if x_i is just the integer i, $F(x)$ is a "*step function*" having jump p_i at i, and being constant between each pair of integers.

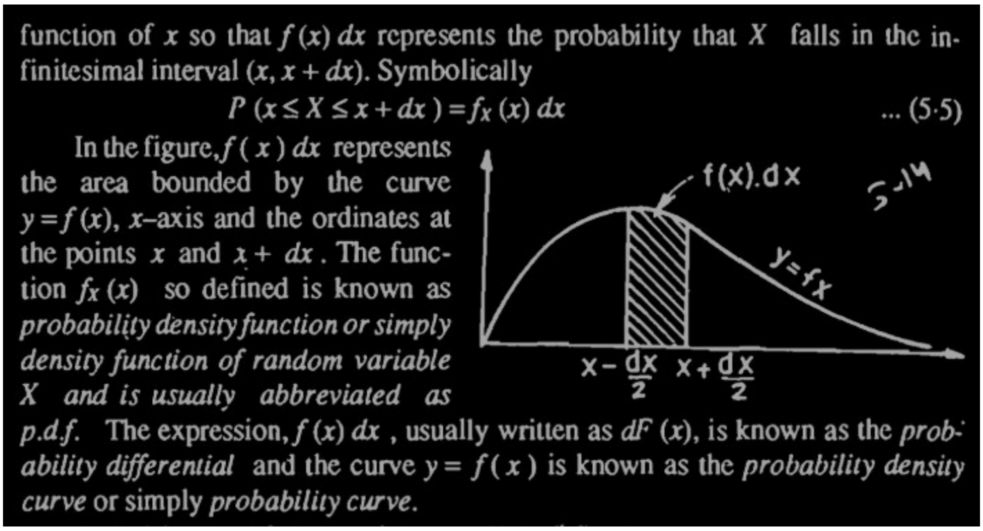
F(x)
1
2
3
4
5
6
7
X



5·4. Continuous Random Variable. A random variable X is said to be continuous if it can take all possible values between certain limits. *In other words, a random variable is said to be continuous when its different values cannot be put in 1-1 correspondence with a set of positive integers.*

A continuous random variable is a random variable that (at least conceptually) can be measured to any desired degree of accuracy. Examples of continuous random variables are age, height, weight etc.

5·4·1. Probability Density Function *(Concept and Definition).* Consider the small interval $(x, x + dx)$ of length dx round the point x. Let $f(x)$ be any continuous

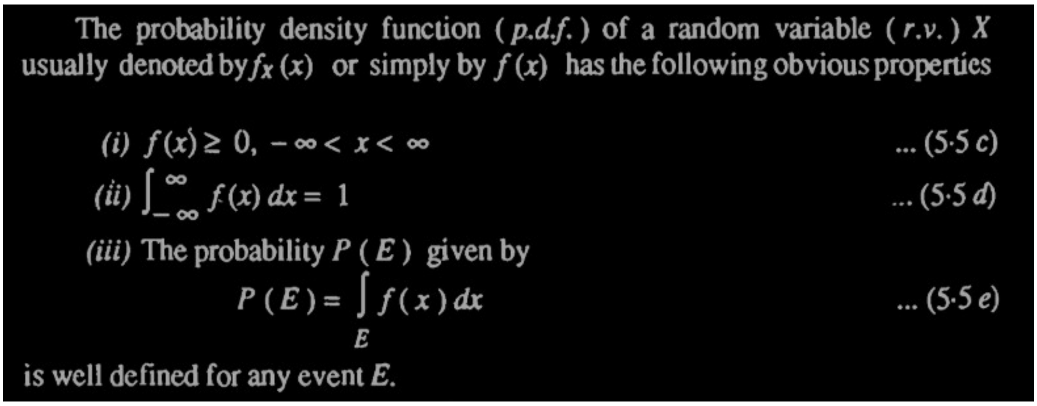


function of x so that $f(x)\,dx$ represents the probability that X falls in the infinitesimal interval $(x, x + dx)$. Symbolically

$$P(x \leq X \leq x + dx) = f_X(x)\,dx \qquad \ldots (5{\cdot}5)$$

In the figure, $f(x)\,dx$ represents the area bounded by the curve $y = f(x)$, x–axis and the ordinates at the points x and $x + dx$. The function $f_X(x)$ so defined is known as *probability density function or simply density function of random variable X and is usually abbreviated as p.d.f.* The expression, $f(x)\,dx$, usually written as $dF(x)$, is known as the *probability differential* and the curve $y = f(x)$ is known as the *probability density curve* or simply *probability curve.*

f(x).dx
y=fx
x−dx/2
x+dx/2



The probability density function (*p.d.f.*) of a random variable (*r.v.*) X usually denoted by $f_X(x)$ or simply by $f(x)$ has the following obvious properties

(i) $f(x) \geq 0, \; -\infty < x < \infty$ ... (5·5 *c*)

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ ... (5·5 *d*)

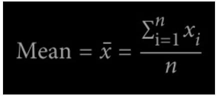
(iii) The probability $P(E)$ given by

$$P(E) = \int_E f(x)\,dx \qquad \ldots (5{\cdot}5\,e)$$

is well defined for any event E.

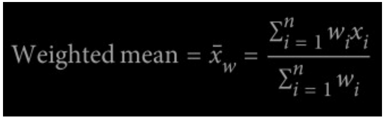
Key Terms for Estimates of Location

1) Mean: The sum of all values divided by the number of values.
   Synonym: average



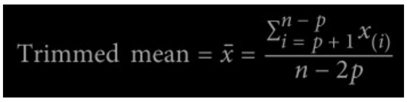
$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2) Weighted mean: The sum of all values times a weight divided by the sum of the weights.
   Synonym: weighted average



$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

3) Median: The value such that one-half of the data lies above and below.
   Synonym: 50th percentile
   # The median is referred to as a robust estimate of location since it is not influenced by outliers (extreme cases) that could skew the results.
4) Percentile: The value such that P percent of the data lies below.
   Synonym: Quantile
5) Weighted median: The value such that one-half of the sum of the weights lies above and below the sorted data.
6) Trimmed Mean: The average of all values after dropping a fixed number of extreme values.
   Synonym: truncated mean
   # A trimmed mean is widely used to avoid the influence of outliers.
7) Truncated Mean: Robust Not sensitive to extreme values.
   Synonym: Resistant



$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$$

8) Outlier: A data value that is very different from most of the data.
   Synonym: extreme value

Key Terms for Variability Metrics (Measure of Dispersion)

Deviations: The difference between the observed values and the estimate of location.
Synonyms: Errors, Residuals

Variance: The sum of squared deviations from the mean divided by n-1 where n is the number of data values.
Synonyms: Mean-Squared-Error

Standard Deviation: The square root of the variance.

# Variance and standard deviation are the most widespread and routinely reported statistics of variability.
# Both are sensitive to outliers.

Mean Absolute Deviation: The mean of the absolute values of the deviations from the mean.
Synonyms: L1-Norm, Manhattan Norm

Range: It is the difference between the largest and smallest value in a data set.

Order Statistics: Metrics based on the data values sorted from smallest to biggest.
Synonym: Ranks

Percentile: The value such that P percent of the values take on this value or less and (100-P) percent take on this value or more.
Synonym: Quantile

Interquartile Range: The difference between the 75th percentile and the 25th percentile.
Synonym: IQR

Degrees of Freedom, and n or n – 1?: Having n – 1 in the denominator in the variance formula, instead of n, leads to the concept of degrees of freedom.

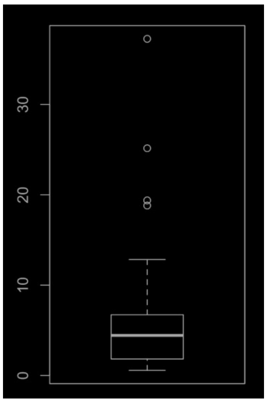
If you use the intuitive denominator of n in the variance formula, you will underestimate the true value of the variance and the standard deviation in the population. This is referred to as a biased estimate. However, if you divide by n – 1 instead of n, the variance becomes an unbiased estimate.

## In statistical theory, location and variability are referred to as the first and second moments of a distribution. The third and fourth moments are called skewness and kurtosis. Skewness refers to whether the data is skewed to larger or smaller values, and kurtosis indicates the propensity of the data to have extreme values. Generally, metrics are not used to measure skewness and kurtosis; instead, these are discovered through visual displays.

Key Terms for Exploring the Distribution

Boxplot: A plot introduced by Tukey as a quick way to visualize the distribution of data.

Synonym: box and whiskers plot



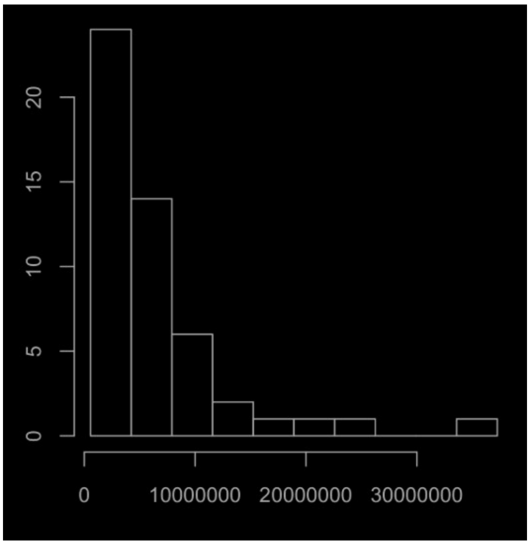
# The median is shown by the horizontal line in the box.
# The dashed lines, referred to as whiskers.
# The top and bottom of the box are the 75th and 25th percentiles, respectively.

Frequency table: A tally of the count of numeric data values that fall into a set of intervals (bins).

## Both frequency tables and percentiles summarize the data by creating bins. In general, quartiles and deciles will have the same count in each bin (equal-count bins), but the bin sizes will be different. The frequency table, by contrast, will have different counts in the bins (equal-size bins), and the bin sizes will be the same.

Histogram: A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis. While visually similar, bar charts should not be confused with histograms.

- Empty bins are included in the graph.
- Bins are of equal width.
- The number of bins (or, equivalently, bin size) is up to the user.
- Bars are contiguous—no empty space shows between bars, unless there is an empty bin.



Density plot: A smoothed version of the histogram, often based on a kernel density estimate. It requires a function to estimate a plot based on the data.

Key Terms for Exploring Categorical Data

Mode: The most commonly occurring category or value in a data set.
Expected value: When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.

The expected value is calculated as follows:
1. Multiply each outcome by its probability of occurrence.
2. Sum these values.

Bar charts: The frequency or proportion for each category is plotted as bars.
Pie charts: The frequency or proportion for each category is plotted as wedges in a pie.

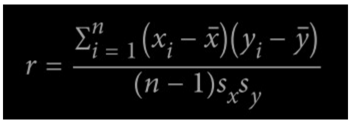
# Difference between Bar Chart and Histogram: In a bar chart the x-axis represents different categories of a factor variable, while in a histogram the x-axis represents values of a single variable on a numeric scale.

In a histogram, the bars are typically shown touching each other, with gaps indicating values that did not occur in the data. In a bar chart, the bars are shown separately from one another.

## Correlation:

Correlation coefficient: A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1).

# Pearson's Correlation Coefficient:



$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{(n-1)s_x s_y}$$

# Correlation Coefficient is sensitive to the outlier in the data.
# A correlation coefficient of zero indicates no correlation, but be aware that random arrangements of data will produce both positive and negative values for the correlation coefficient just by chance.

Correlation matrix: A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.

# Spearman's rho or Kendall's tau is rank-based correlation coefficients that work with rank rather than values.

- Robust to outliers.
- Can handle certain type of non-linearities.
- Rank-Based estimates are mostly for smaller datasets and specific hypothesis tests.

Scatterplot: A plot in which the x-axis is the value of one variable, and the y-axis the value of another.

Covariance and Coorelation:

https://towardsdatascience.com/covariance-and-correlation-321fdacab168

Pearson Correlation:

https://towardsdatascience.com/pearson-coefficient-of-correlation-explained-369991d93404

Spearman's Rank-Order Correlation:
https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf
https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Random Sampling and Sample Bias

# A sample is a subset of data from a larger data set
# Population ------> larger dataset -----> from where our sample is drawn.

Simple Random Sampling: Random Sampling is the process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw. The sample that results is called a simple random sample.

> # Sampling With Replacement ----> In this sampling, observations are put back in the population after each draw for possible future reselection.
> # Sampling Without Replacement ----> In this sampling, observations once selected are not available for future draws.

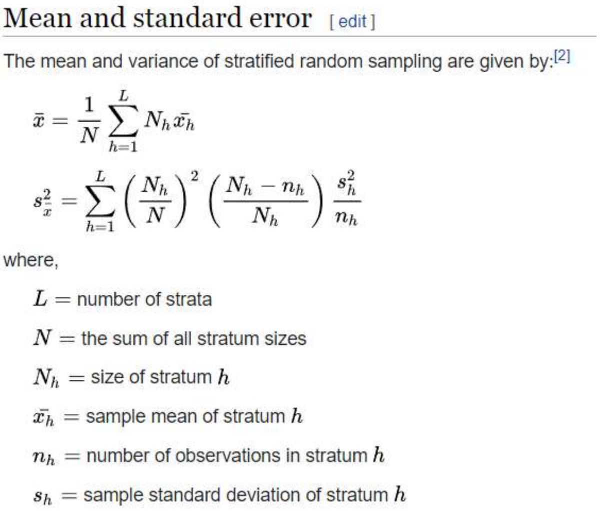
# In general, the sample mean and variance differ from the true mean and variance because they depend upon specific n values of the sample., they are only estimates of the true mean and variance and are called estimators.

# An estimator is unbiased if its expected value is equal to the true value, otherwise, it is biased.

Stratified Sampling: When the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata" and dividing the population into strata and randomly sampling from each strata.
# Advantages:

1. If measurements within strata have lower standard deviation (as compared to the overall standard deviation in the population), stratification gives smaller error in estimation.
2. For many applications, measurements become more manageable and/or cheaper when the population is grouped into strata.

# Stratum (pl., strata): A homogeneous subgroup of a population with common characteristics.
# Sampling Fraction: The ratio of the size of the sample to the size of the population is called a sampling fraction.



Mean and standard error [edit]

The mean and variance of stratified random sampling are given by:[2]

$$\bar{x} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{x_h}$$

$$s_{\bar{x}}^2 = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

where,

L = number of strata

N = the sum of all stratum sizes

N_h = size of stratum h

$\bar{x_h}$ = sample mean of stratum h

n_h = number of observations in stratum h

s_h = sample standard deviation of stratum h

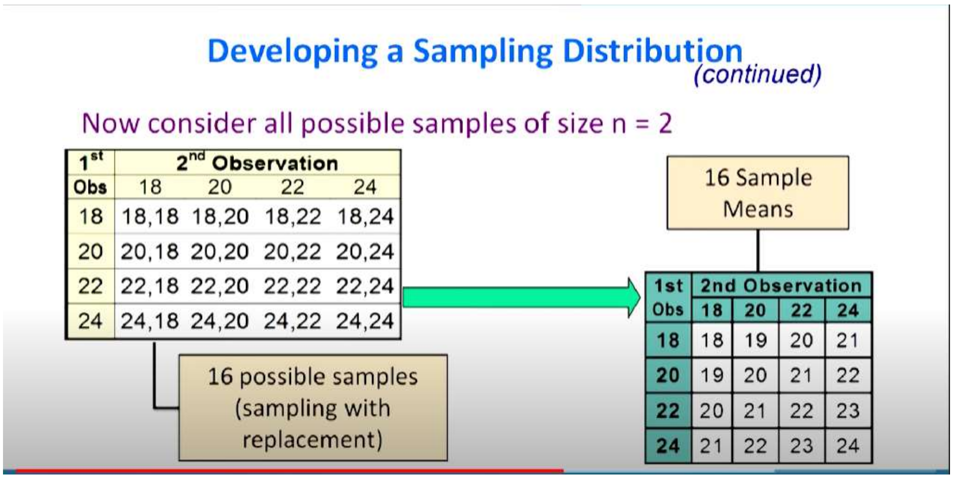
Bias: Statistical Bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process.

> # Bias may be observable or invisible.
> # It can arise when any important variable is left-out or any statistical/ML model is not specified.

Sampling Distribution of a Statistic:

Sampling Distribution of a statistic refers to the distribution of some sample statistic over many samples drawn from the same population.

# The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of the data itself.
# The larger the sample the statistic is based on, the more this is true. Also, the larger the sample, the narrower the distribution of the sample statistic.



Developing a Sampling Distribution

(continued)

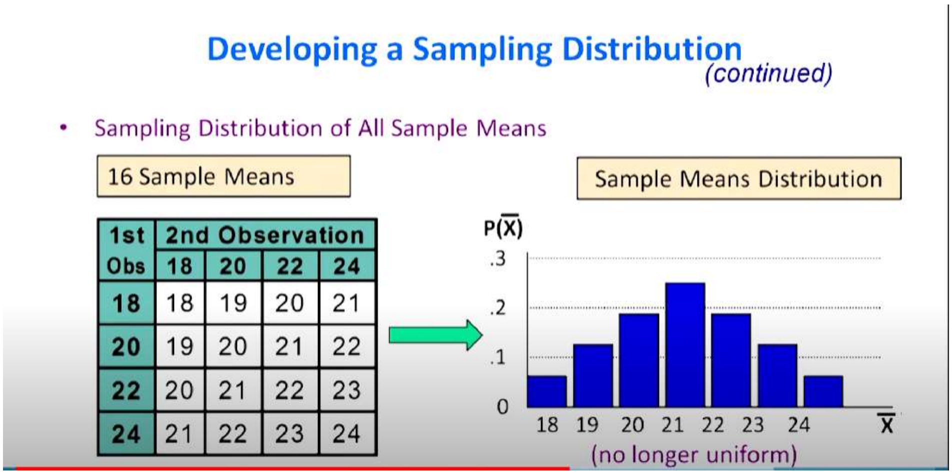
Now consider all possible samples of size n = 2

1st Obs	2nd Observation 18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

16 Sample Means

1st Obs	2nd Observation 18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24



Developing a Sampling Distribution

(continued)

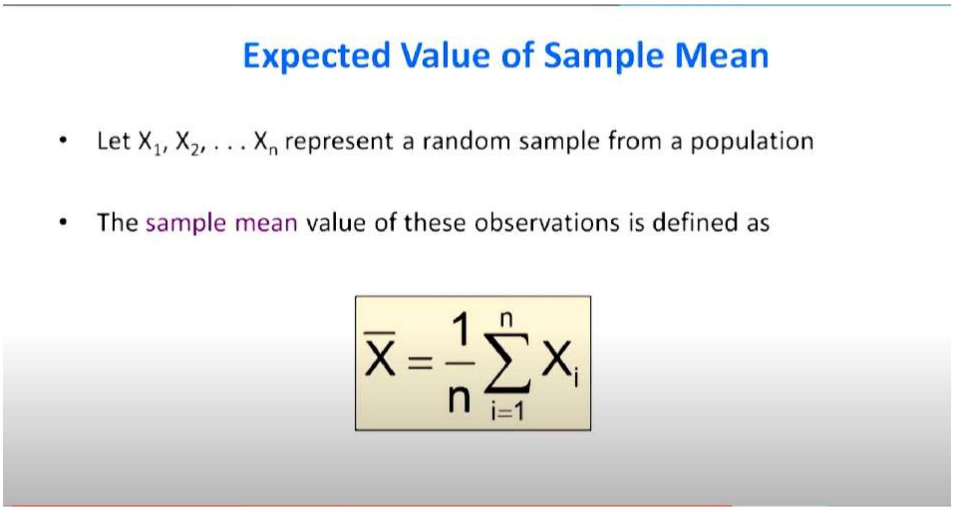
- Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation 18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution

(no longer uniform)

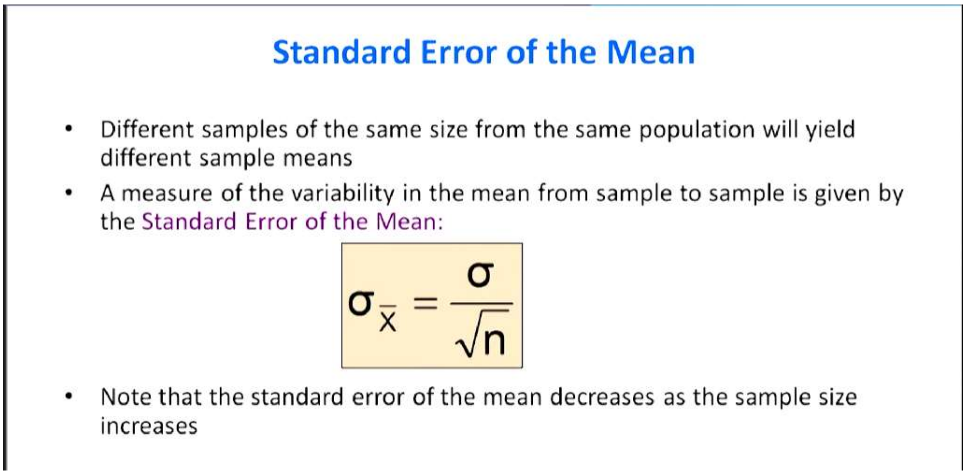


Expected Value of Sample Mean

- Let $X_1, X_2, \ldots X_n$ represent a random sample from a population

- The sample mean value of these observations is defined as

$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

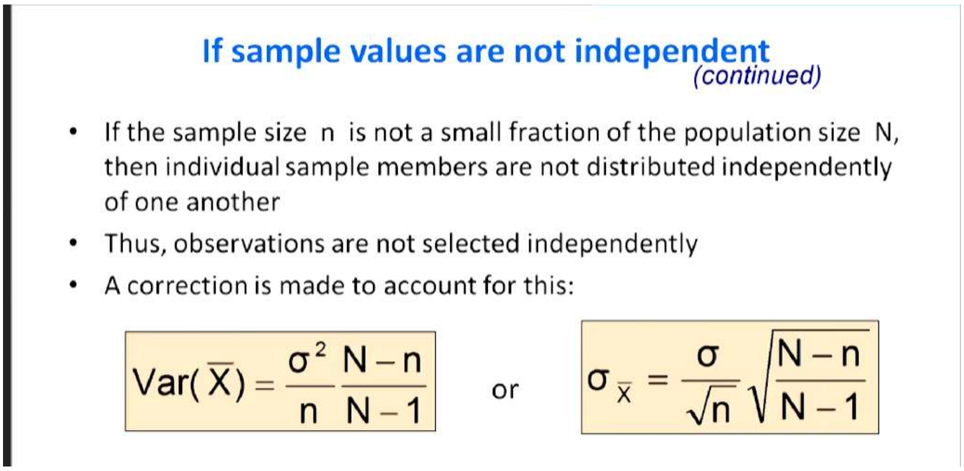


Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the Standard Error of the Mean:

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$$

- Note that the standard error of the mean decreases as the sample size increases

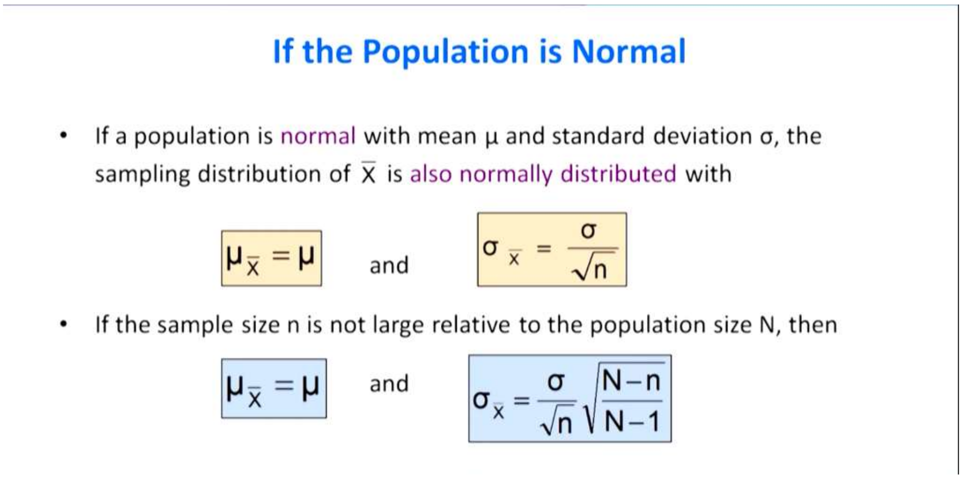


If sample values are not independent

(continued)

- If the sample size n is not a small fraction of the population size N, then individual sample members are not distributed independently of one another
- Thus, observations are not selected independently
- A correction is made to account for this:

$$\boxed{\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}} \quad \text{or} \quad \boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$



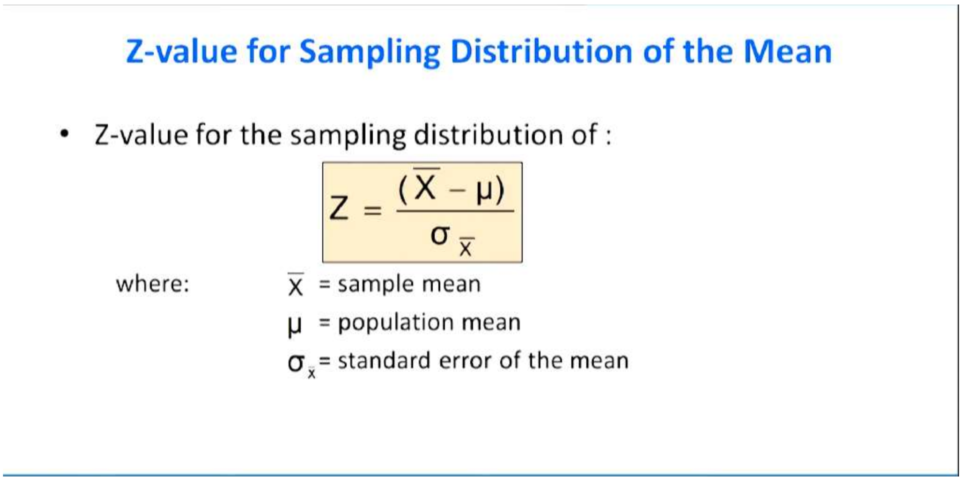
If the Population is Normal

- If a population is normal with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is also normally distributed with

$$\boxed{\mu_{\bar{X}} = \mu} \quad \text{and} \quad \boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$$

- If the sample size n is not large relative to the population size N, then

$$\boxed{\mu_{\bar{X}} = \mu} \quad \text{and} \quad \boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$



Z-value for Sampling Distribution of the Mean

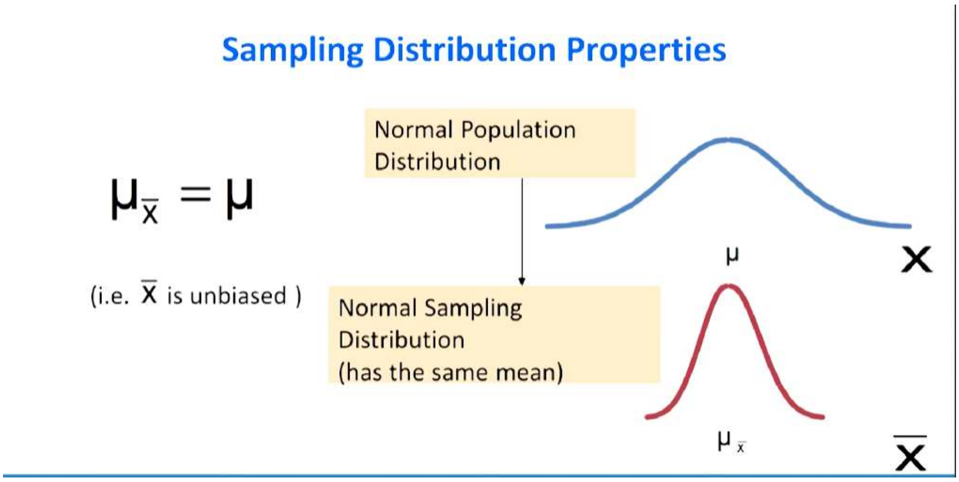
- Z-value for the sampling distribution of :

$$\boxed{Z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}}}$$

where: $\bar{X}$ = sample mean

μ = population mean

σ_{x} = standard error of the mean



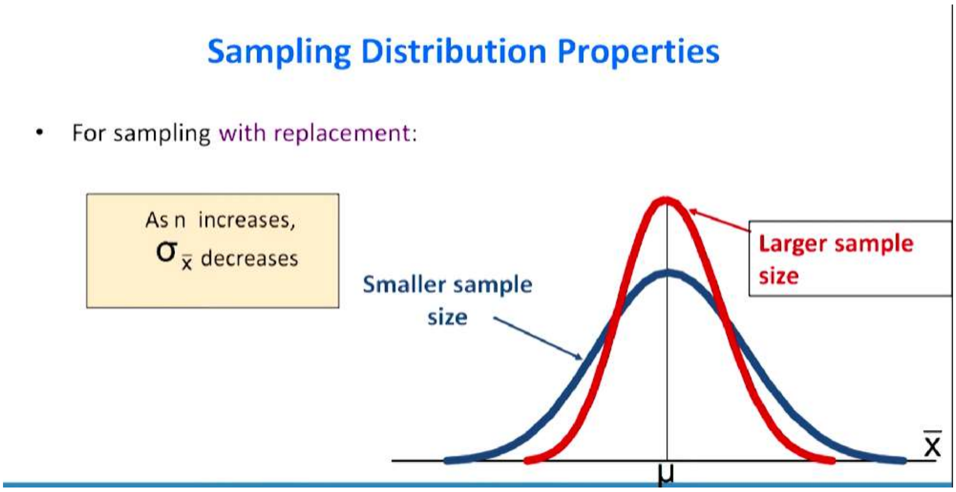
Sampling Distribution Properties

$\mu_{\bar{x}} = \mu$

(i.e. $\bar{X}$ is unbiased)

Normal Population Distribution

Normal Sampling Distribution (has the same mean)



Sampling Distribution Properties

- For sampling with replacement:

As n increases, $\sigma_{\bar{x}}$ decreases

Smaller sample size

Larger sample size

Central Limit Theorem: It says that the means drawn from multiple samples will resemble the familiar bell-shaped normal curve (see "Normal Distribution" on page 69), even if the source population is not normally distributed, provided that the sample size is large enough and the departure of the data from normality is not too great.

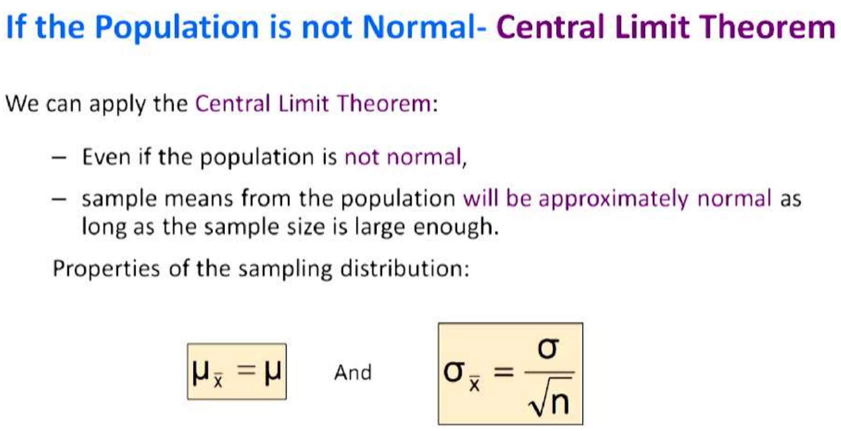
Standard Error: The standard error is a single metric that sums up the variability of the sampling distribution for a statistic.

The standard error can be estimated using a statistic based on the standard deviation s of the sample values, and the sample size n.

Standard Error = SE = s/sqrt(n)

# As sample size (n) increases the SE decreases.
# Standard Deviation Versus Standard Error: Standard deviation (which measures the variability of individual data points) with standard error (which measures the variability of a sample metric).

# The Sampling Distribution can be estimated via the bootstrap, or via formulas that rely on Central Limit Theorem.



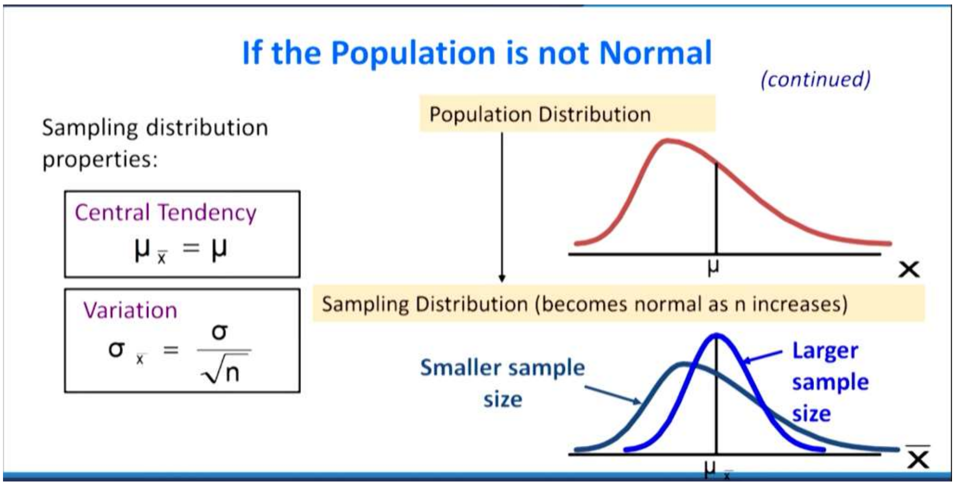
If the Population is not Normal- Central Limit Theorem

We can apply the Central Limit Theorem:

- Even if the population is not normal,
- sample means from the population will be approximately normal as long as the sample size is large enough.

Properties of the sampling distribution:

$\boxed{\mu_{\bar{x}} = \mu}$ And $\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$



If the Population is not Normal

(continued)

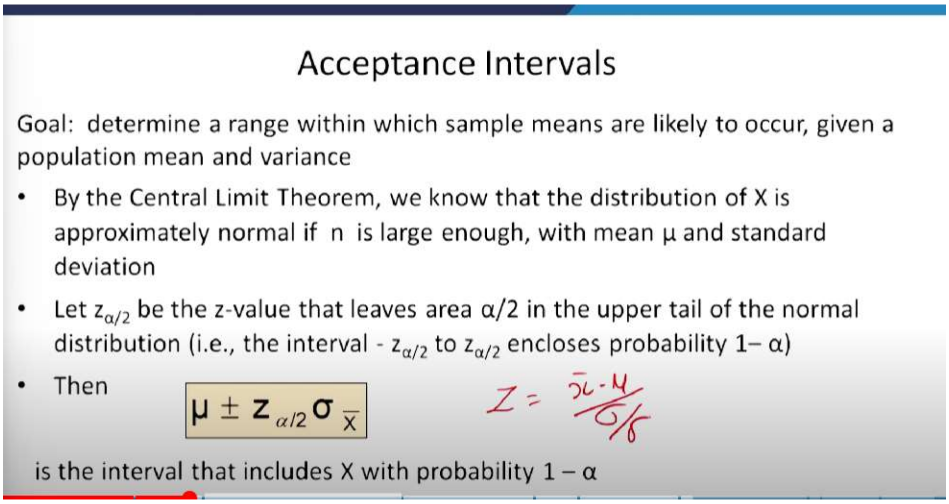
Sampling distribution properties:

Central Tendency
$\mu_{\bar{x}} = \mu$

Variation
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Population Distribution

Sampling Distribution (becomes normal as n increases)

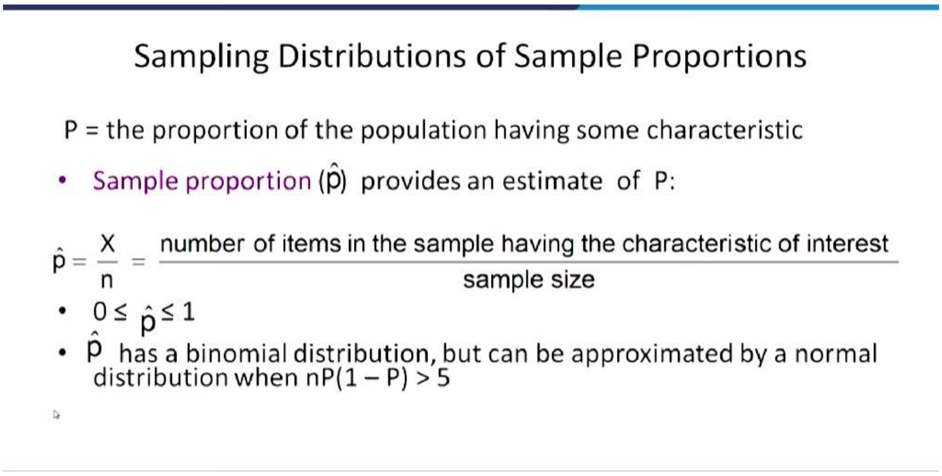


Acceptance Intervals

Goal: determine a range within which sample means are likely to occur, given a population mean and variance

- By the Central Limit Theorem, we know that the distribution of X is approximately normal if n is large enough, with mean μ and standard deviation
- Let $z_{\alpha/2}$ be the z-value that leaves area α/2 in the upper tail of the normal distribution (i.e., the interval - $z_{\alpha/2}$ to $z_{\alpha/2}$ encloses probability 1– α)
- Then $\mu \pm z_{\alpha/2}\sigma_{\bar{X}}$ $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

is the interval that includes X with probability 1 – α



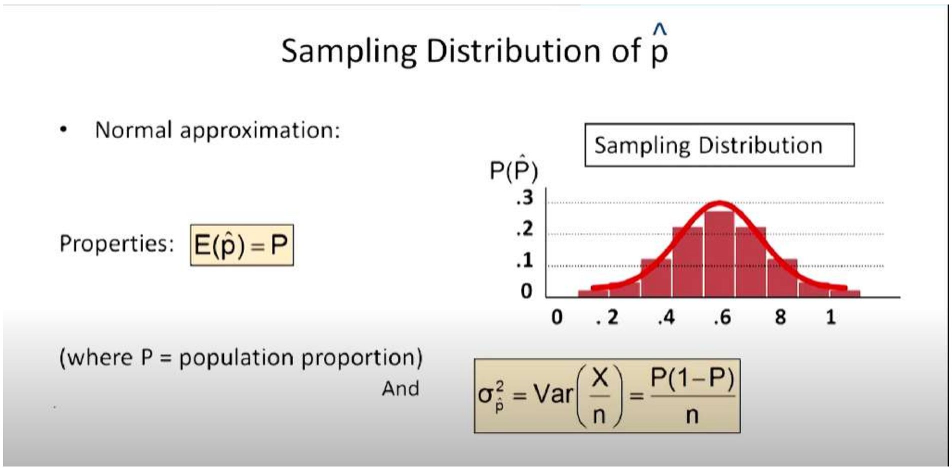
Sampling Distributions of Sample Proportions

P = the proportion of the population having some characteristic

- Sample proportion ($\hat{p}$) provides an estimate of P:

$$\hat{p} = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

- $0 \leq \hat{p} \leq 1$
- $\hat{p}$ has a binomial distribution, but can be approximated by a normal distribution when nP(1 – P) > 5



Sampling Distribution of $\hat{p}$

- Normal approximation:

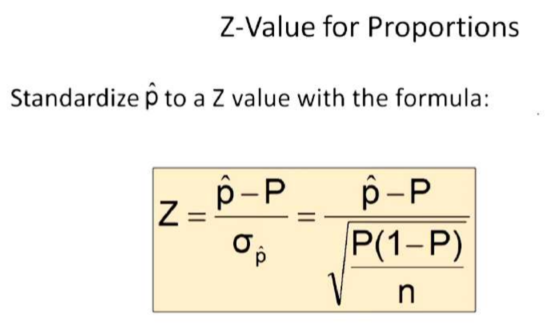
Properties: $E(\hat{p}) = P$

(where P = population proportion)

And

$$\sigma_{\hat{p}}^2 = \text{Var}\left(\frac{X}{n}\right) = \frac{P(1-P)}{n}$$

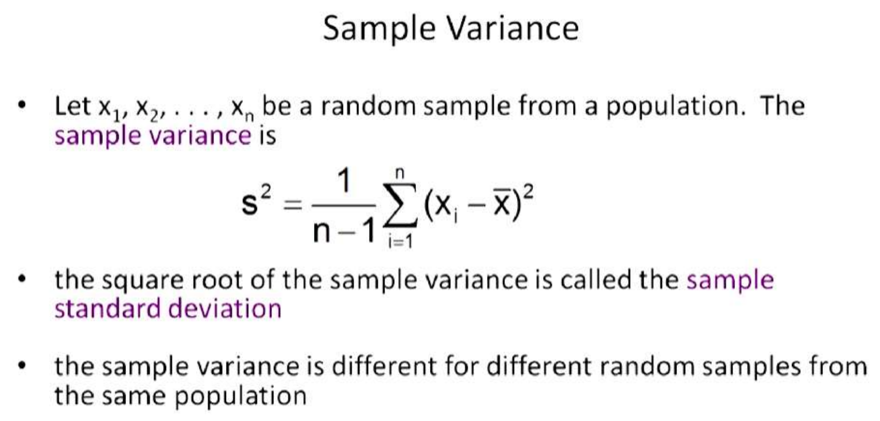
Sampling Distribution



Z-Value for Proportions

Standardize $\hat{p}$ to a Z value with the formula:

$$\boxed{Z = \frac{\hat{p} - P}{\sigma_{\hat{p}}} = \frac{\hat{p} - P}{\sqrt{\dfrac{P(1-P)}{n}}}}$$

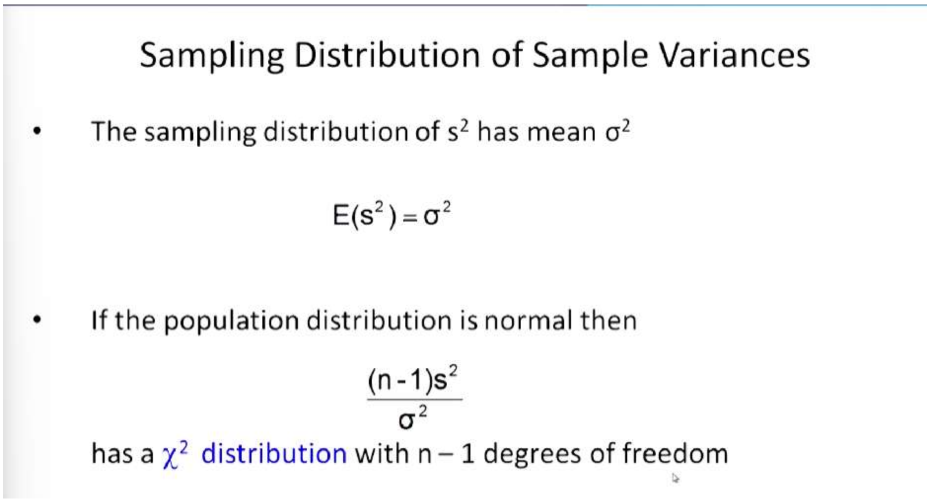


Sample Variance

- Let $x_1, x_2, \ldots, x_n$ be a random sample from a population. The sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- the square root of the sample variance is called the sample standard deviation
- the sample variance is different for different random samples from the same population



Sampling Distribution of Sample Variances

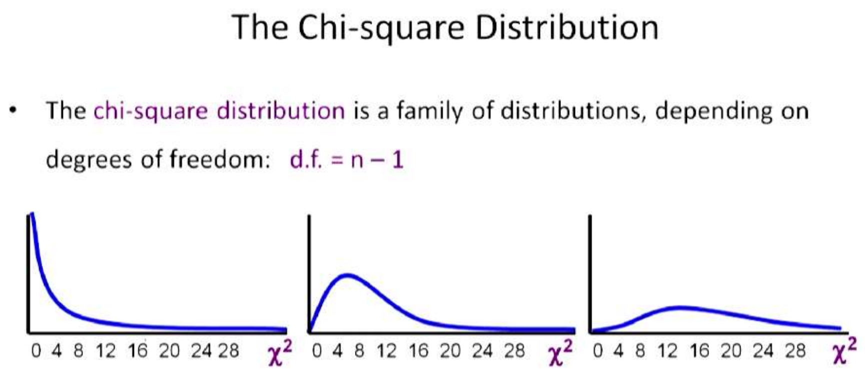
- The sampling distribution of s^2 has mean σ^2

$$E(s^2) = \sigma^2$$

- If the population distribution is normal then

$$\frac{(n-1)s^2}{\sigma^2}$$

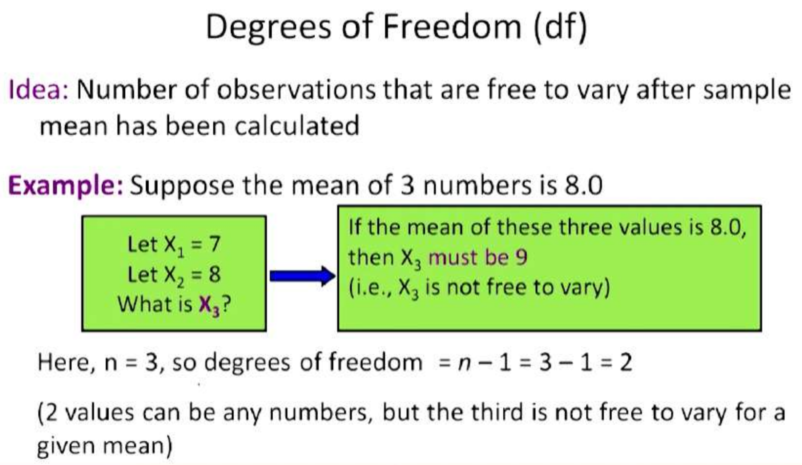
has a χ^2 distribution with n – 1 degrees of freedom



The Chi-square Distribution

- The chi-square distribution is a family of distributions, depending on degrees of freedom: d.f. = n – 1

As sample size(n) increases, chi-square distribution become normal distribution.



Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$
Let $X_2 = 8$
What is X_3?

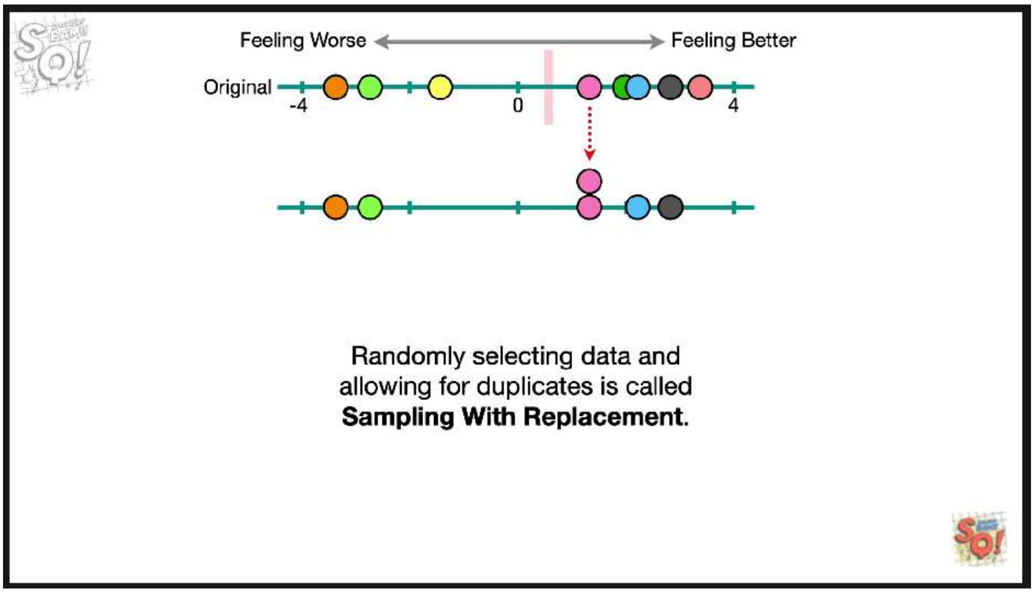
If the mean of these three values is 8.0, then X_3 must be 9
(i.e., X_3 is not free to vary)

Here, n = 3, so degrees of freedom $= n - 1 = 3 - 1 = 2$

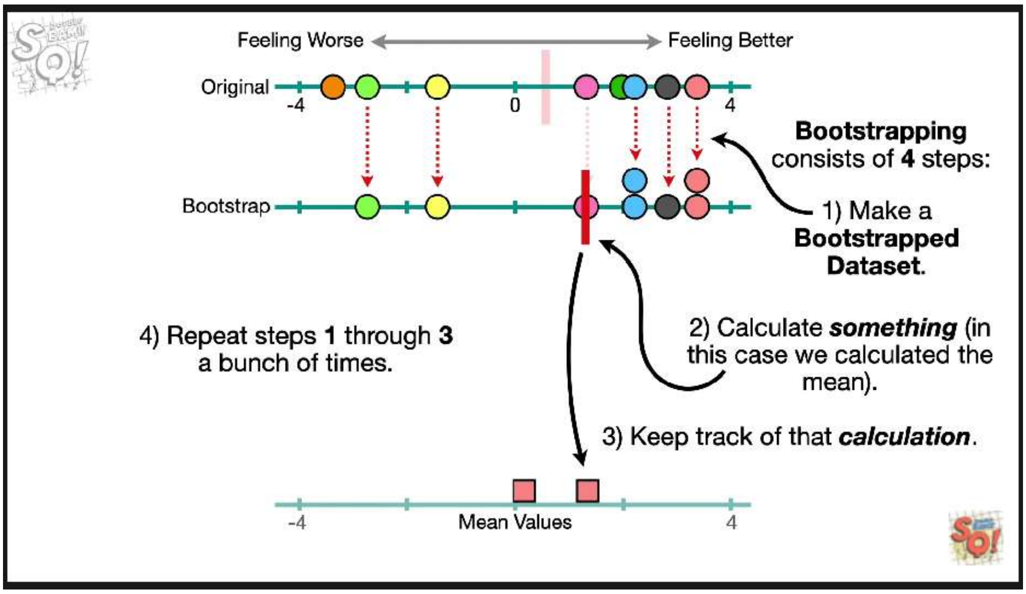
(2 values can be any numbers, but the third is not free to vary for a given mean)

Bootstrap Theory:

# The bootstrap (sampling with replacement from the data set) is powerful tool that provide us the variability of a sample statistic.
# When applied to predictive models, aggregating multiple bootstrap sample predictions (bagging) outperforms the use of a single model.



Randomly selecting data and allowing for duplicates is called **Sampling With Replacement.**



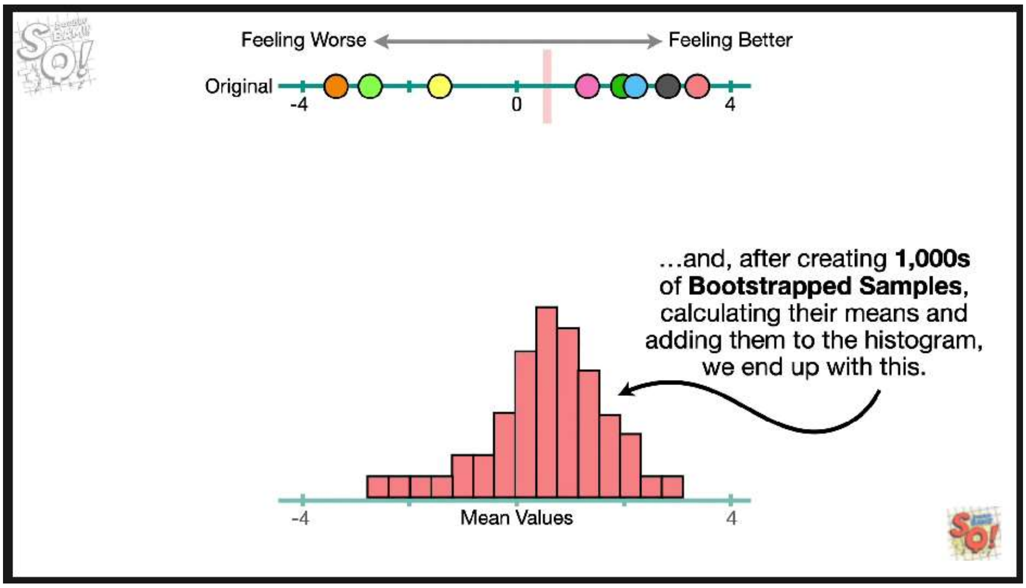
Bootstrapping consists of **4** steps:

1) Make a **Bootstrapped Dataset.**

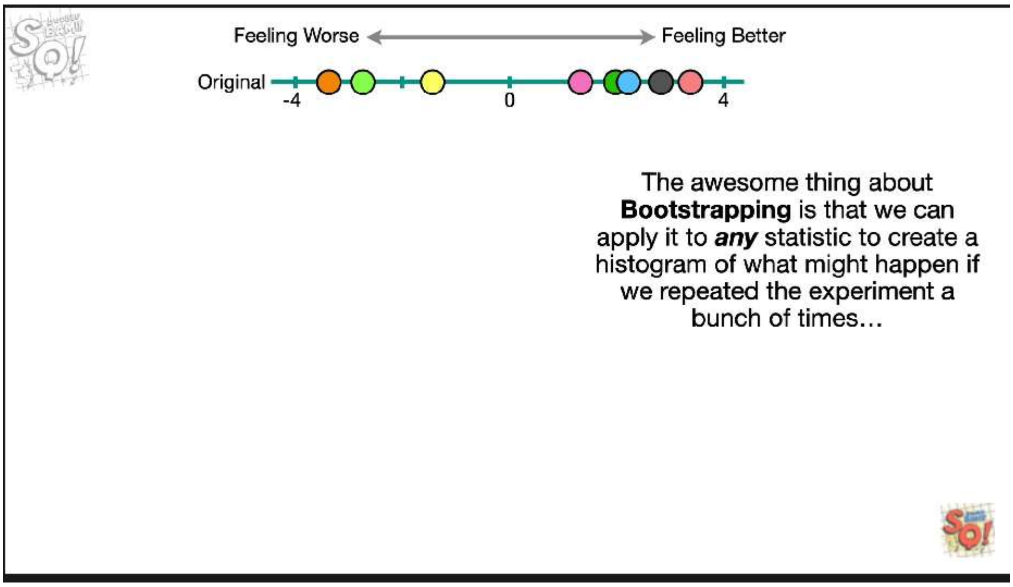
2) Calculate ***something*** (in this case we calculated the mean).

3) Keep track of that ***calculation***.

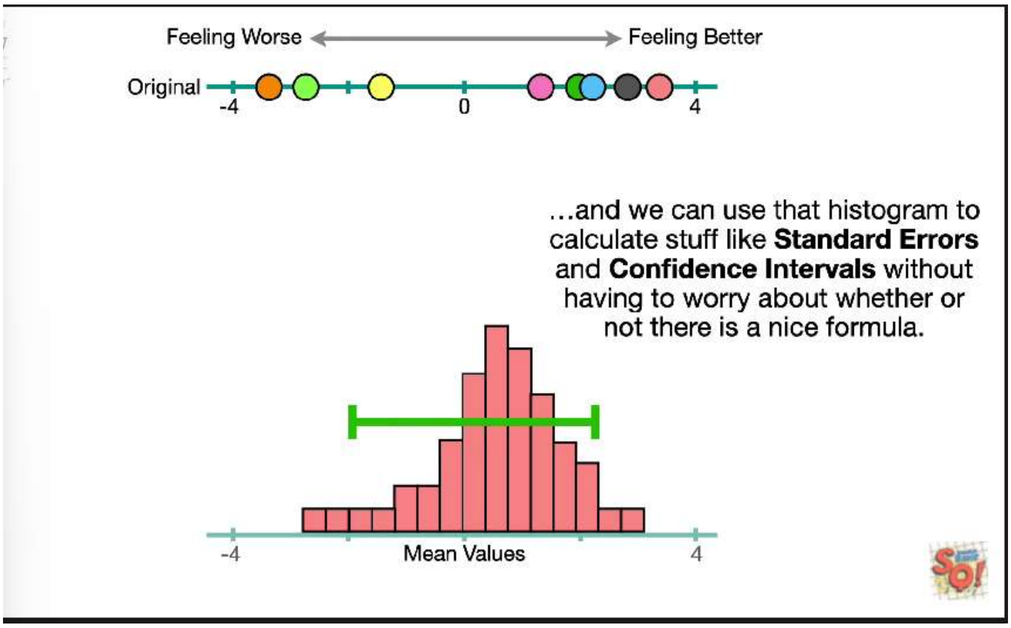
4) Repeat steps **1** through **3** a bunch of times.



…and, after creating **1,000s** of **Bootstrapped Samples**, calculating their means and adding them to the histogram, we end up with this.



The awesome thing about **Bootstrapping** is that we can apply it to ***any*** statistic to create a histogram of what might happen if we repeated the experiment a bunch of times…



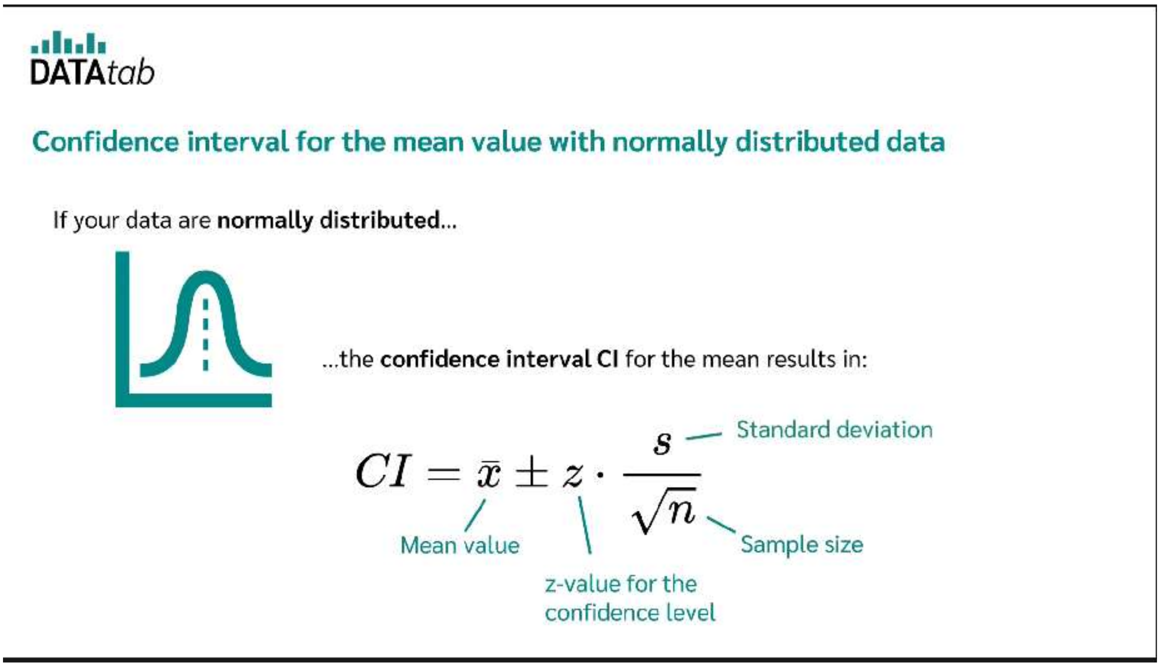
…and we can use that histogram to calculate stuff like **Standard Errors** and **Confidence Intervals** without having to worry about whether or not there is a nice formula.

# Bootstrap can work on multivariate data
# Model might run on bootstrapped data, to estimate the stability(variability) of model parameters, or to improve predictive power.

# In Decision Tree algorithm, multiple trees run on bootstrapped data and then averaging their predictions generally performs better than using single tree.	This process is called Bagging (bootstrap aggregating).

Confidence Interval: The confidence interval is the range in which the parameter will lie with a high probability.

If different samples are taken from a population, a parameter such as the mean value will be different in each sample. The confidence interval indicates the range of variation.



DATAtab

Confidence interval for the mean value with normally distributed data

If your data are **normally distributed**...

...the **confidence interval CI** for the mean results in:

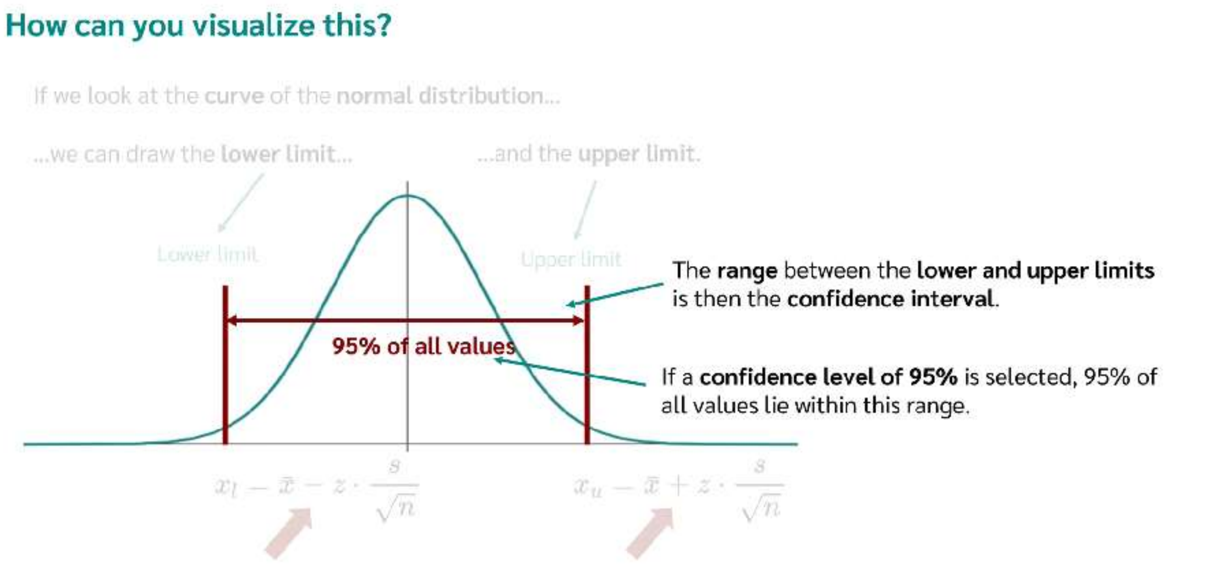
$$CI = \bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$$

Standard deviation

Mean value

z-value for the confidence level

Sample size



How can you visualize this?

If we look at the **curve** of the **normal distribution**...

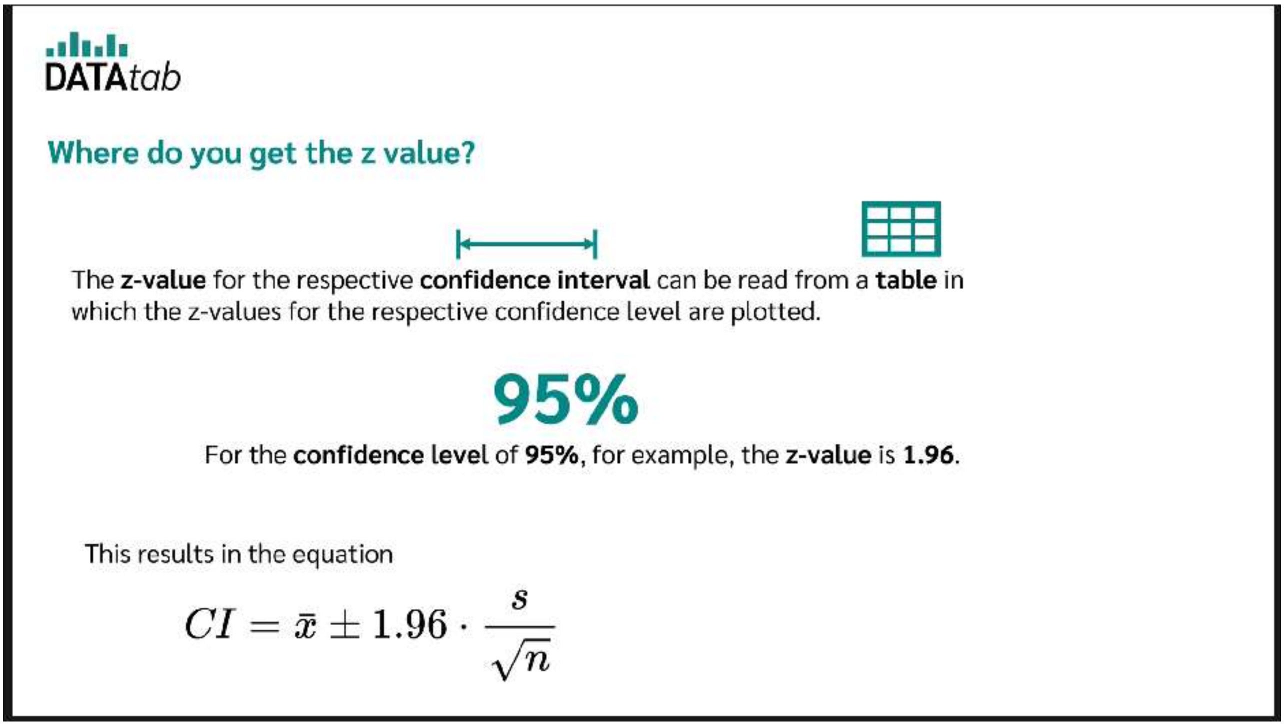
...we can draw the **lower limit**...

...and the **upper limit**.

The **range** between the **lower and upper limits** is then the **confidence interval**.

If a **confidence level of 95%** is selected, 95% of all values lie within this range.

$x_l = \bar{x} - z \cdot \frac{s}{\sqrt{n}}$ $\quad$ $x_u = \bar{x} + z \cdot \frac{s}{\sqrt{n}}$



DATAtab

Where do you get the z value?

The **z-value** for the respective **confidence interval** can be read from a **table** in which the z-values for the respective confidence level are plotted.

95%

For the **confidence level** of **95%**, for example, the **z-value** is **1.96**.

This results in the equation

$$CI = \bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

# The percentage associated with the confidence interval is termed the level of confidence.
# The higher the level of confidence, the wider the interval.
# Smaller the sample, the wider the interval (i.e., the greater the uncertainty).

Normal Distribution:

# Error: The difference between a data point and a predicted or average value.
# Standardize: Subtract the mean and divide by the standard deviation.
# Z-score: The result of standardizing an individual data point.

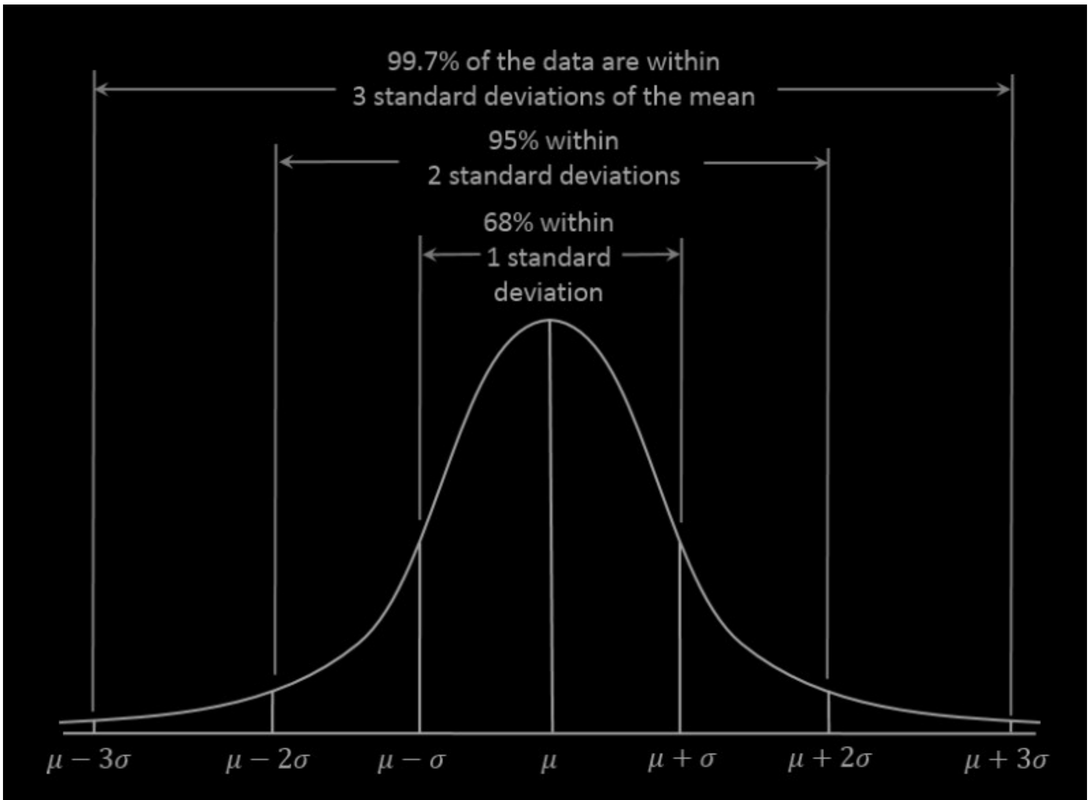


Figure 2-10. Normal curve

# Standard normal: A normal distribution with mean = 0 and standard deviation = 1.

QQ-Plot (Quantile-Quantile Plot): A plot to visualize how close a sample distribution is to a specified distribution, e.g., the normal distribution.

The QQ-Plot orders the z-scores from low to high and plots each value's z-score on the y-axis, the x-axis is the corresponding quantile of a normal distribution for that value's rank. Since the data is normalized, the units correspond to the number of standard deviations away from the mean.

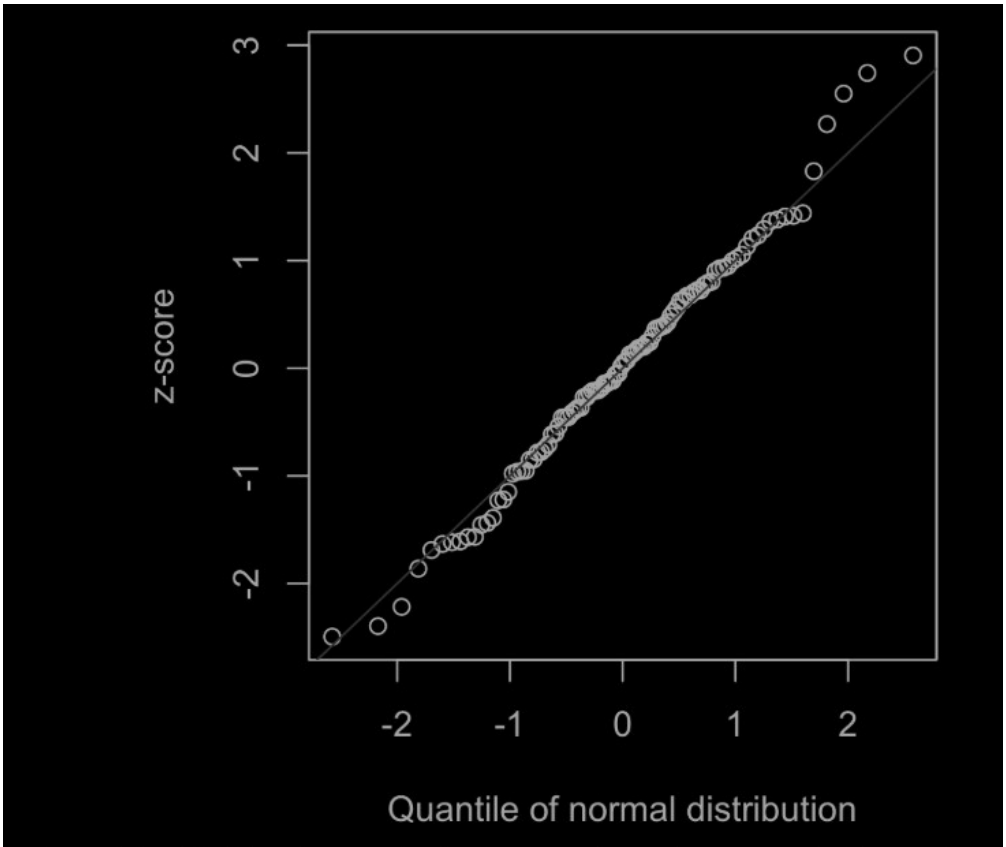


Figure 2-11. QQ-Plot of a sample of 100 values drawn from a standard normal distribution

# To convert data to z-scores, you subtract the mean of the data and divide by the standard deviation; you can then compare the data to a normal distribution.

Long Tailed Distribution:

# Tail: The long narrow portion of the distribution where relatively extreme values occur at low frequency.
# Skew: It measures the difference between a long tail and a shorter tail.

Student’s t-Distribution:

https://www.youtube.com/watch?v=T0xRanwAIiI

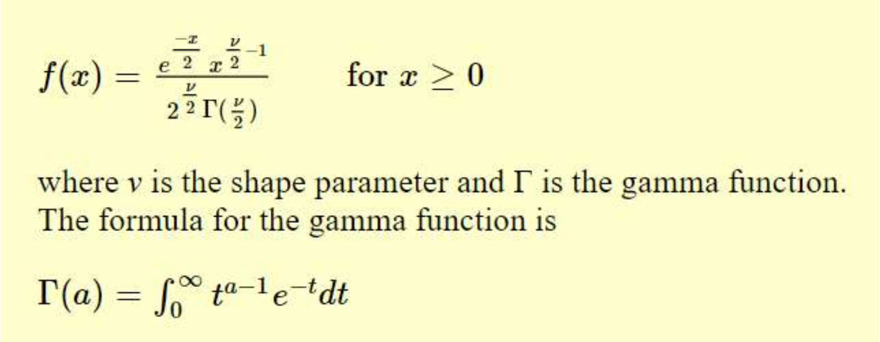
- The t-distribution is a family of distributions resembling the normal distribution but with thicker tails.
- The t-distribution is widely used as a reference basis for the distribution of sample means, differences between two sample mean, regression parameters, and more.

Binomial Distribution:

- A binomial trial is an experiment with two possible outcomes: one with probability p and the other with probability 1 – p.
- With large n, and provided p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.
- The mean of a binomial distribution is np and variance is np(1 — p)

Chi-Squared Distribution: The chi-square distribution results when ν independent variables with standard normal distributions are squared and summed.

https://www.youtube.com/watch?v=hcDb12fsbBU

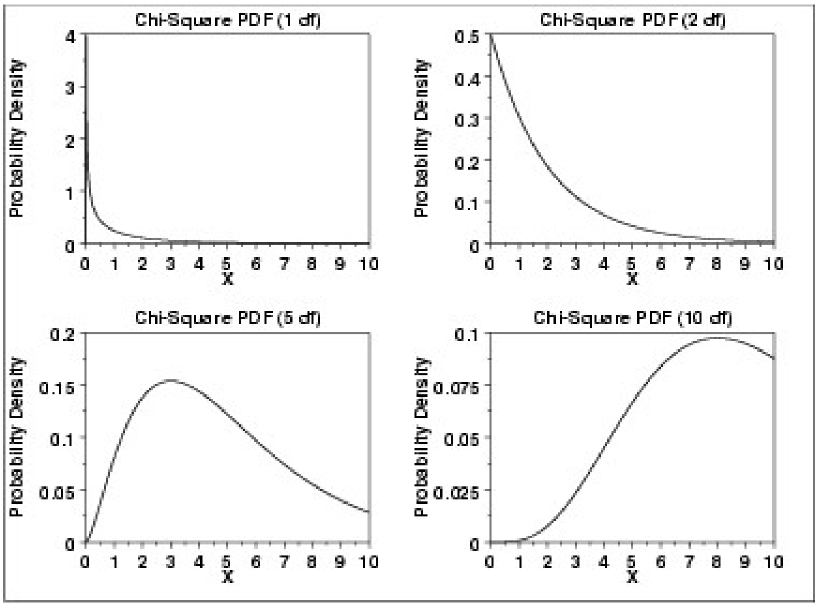


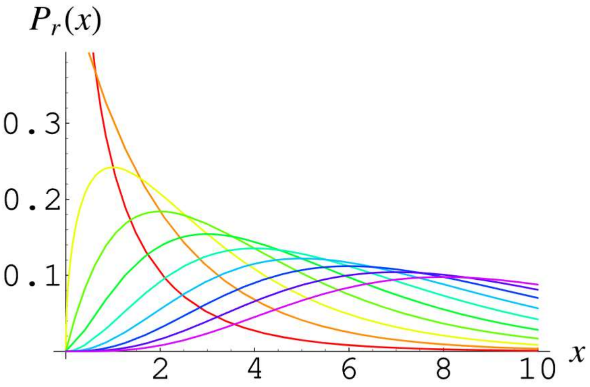
$$f(x) = \frac{e^{\frac{-x}{2}}\, x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\,\Gamma(\frac{\nu}{2})} \qquad \text{for } x \geq 0$$

where v is the shape parameter and Γ is the gamma function. The formula for the gamma function is

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

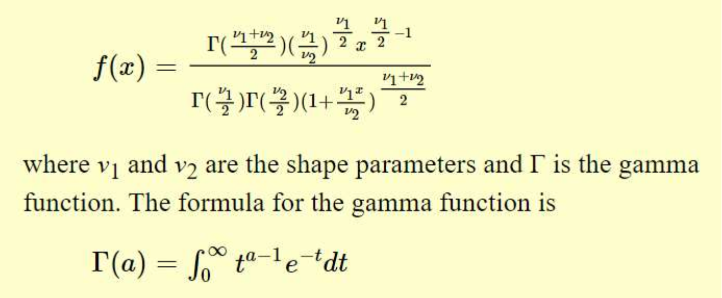
# The chi-square statistic measures the extent of departure from what you would expect in a null model.





F Distribution: The F distribution is the ratio of two chi-distribution with degree of freedom v1 and v2 respectively, where each chi-distribution is divided by its degrees of freedom.

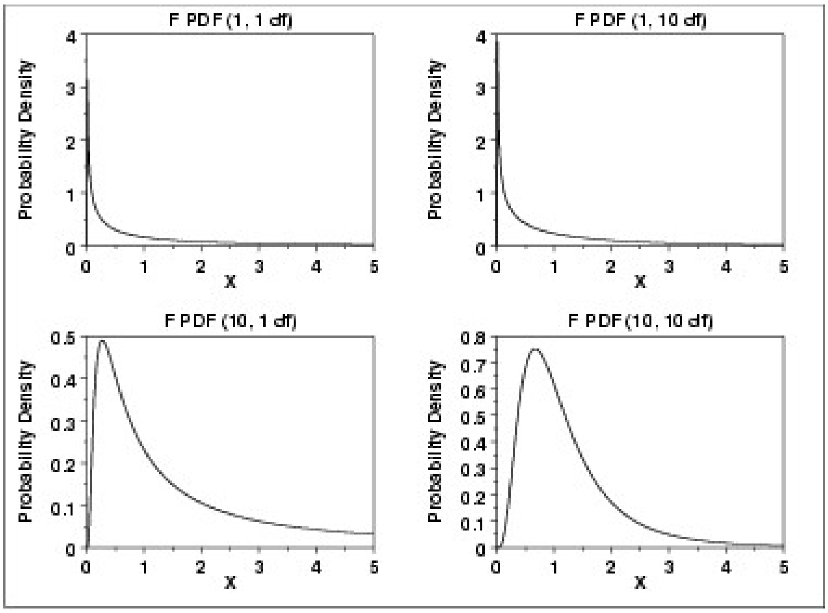
https://www.youtube.com/watch?v=G_RDxAZJ-ug



$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})(1+\frac{\nu_1 x}{\nu_2})^{\frac{\nu_1+\nu_2}{2}}}$$

where ν_1 and ν_2 are the shape parameters and Γ is the gamma function. The formula for the gamma function is

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$$



Poisson Distribution: For events that occur at a constant rate, the number of events per time or space can be defined in terms of poisson
distribution.

Exponential Distribution: This distribution gives the time or distance between two consecutive events.

	Exponential	Poisson
Question	How much time between given no. of events?	How many no. of events in a given time?
Random Variable	Time - Continuous	No. of events - Discrete
Parameter	Lambda = rate of occurrence (unit - 1/time)	Lambda = mean number of events that occurs in a specified interval of time or space.
Notes	It is a special case of gamma function Exp = gamma(shape = 1 , scale = 1/lambda)	Both Mean and Variance of Poisson distribution is equal to lambda

Weibull Distribution: The Weibull distribution is an exponential distribution in which the event rate is allowed to change, as specified by a shape parameter β.
If β < 1, the probability of an event decreases over the time.
If β > 1, the probability of an event increases over the time.

Statistical Experiments and Significance Testing

Hypothesis: Hypothesis testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between the two statistical variables.

Examples:

- A teacher assumes that 60% of his college students come from lower-middle-class families.

- A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.

# Depending on the population distribution, you can classify the statistical hypothesis into two types.

1. Simple Hypothesis: A simple hypothesis specifies an exact value for the parameter.

   Example: A company is claiming that their average sales for this quarter are 1000 units. This is an example of a simple hypothesis.

2. Composite Hypothesis: A composite hypothesis specifies a range of values.

   Example: Suppose the company claims that the sales are in the range of 900 to 1000 units. Then this is a case of a composite hypothesis.

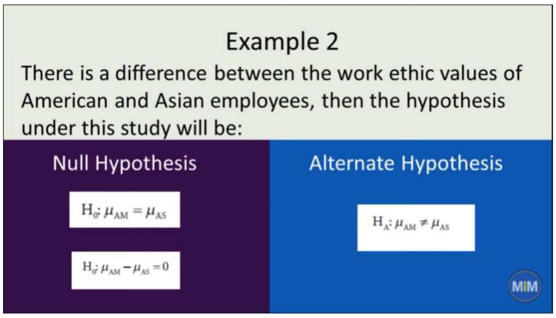
Test of a Statistical Hypothesis: A test of a statistical hypothesis is a two-action decision problem after the experimental sample has been obtained, the two-action being the acceptance or rejection of the hypothesis under consideration.

Important:

Hypothesis Testing: https://www.cuemath.com/data/hypothesis-testing/
Type 1 and Type 2 Error: https://towardsdatascience.com/clarifying-type-i-and-type-ii-errors-in-hypothesis-testing-1a4e6a5ba616

@ https://www.investopedia.com/articles/active-trading/092214/hypothesis-testing-finance-concept-examples.asp

# In mathematical expressions null hypothesis is denoted by 'equal to =' sign whereas the alternate hypothesis is represented by 'not equal to' sign, < or >.
# If the null hypothesis is accepted, the result of the study becomes insignificant and if the alternative hypothesis is accepted then the results are significant.



Example 2

There is a difference between the work ethic values of American and Asian employees, then the hypothesis under this study will be:

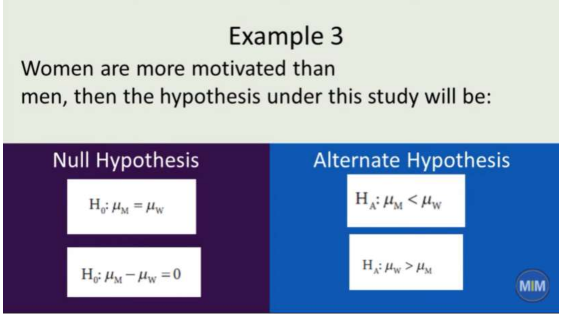
Null Hypothesis

$H_0: \mu_{AM} = \mu_{AS}$

$H_0: \mu_{AM} - \mu_{AS} = 0$

Alternate Hypothesis

$H_A: \mu_{AM} \neq \mu_{AS}$



Example 3

Women are more motivated than men, then the hypothesis under this study will be:

Null Hypothesis

$H_0: \mu_M = \mu_W$

$H_0: \mu_M - \mu_W = 0$

Alternate Hypothesis

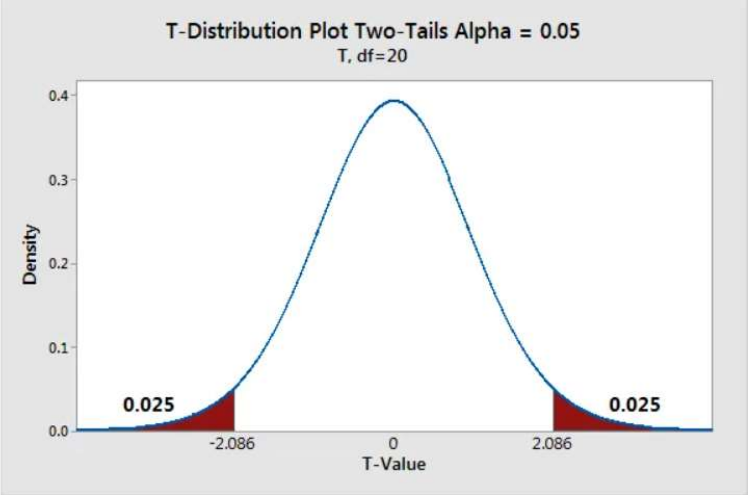
$H_A: \mu_M < \mu_W$

$H_A: \mu_W > \mu_M$

# One should try to structure the problem so that the null hypothesis is simple rather than composite.

Two-Tailed Hypothesis: Two-Tailed hypothesis tests are also known as nondirectional and two-sided tests because you can test for effects in both directions.

When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution.



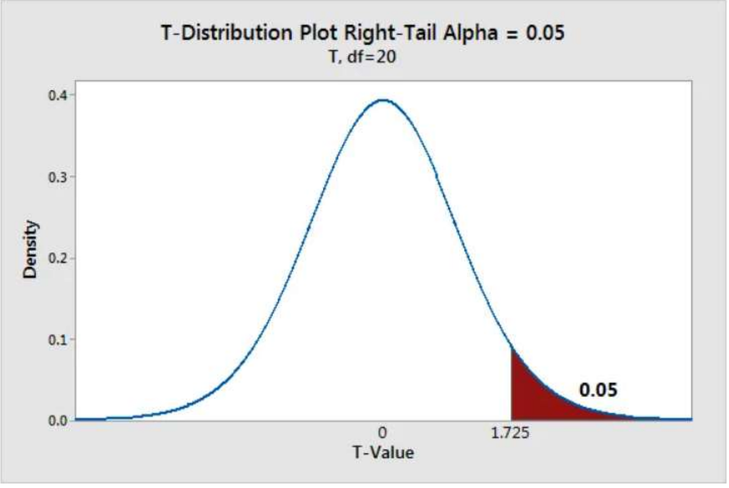
# Using Two-Tailed Hypothesis test, one can detect both positive or negative effects.

One-Tailed Hypothesis: One-Tailed Hypothesis tests are known as a directional and one-sided test because you test effects only in one direction. When the One-Tailed Hypothesis is performed, the entire significance level percentage goes to extreme end of one tail of the distribution.

Sets of generic hypothesis:

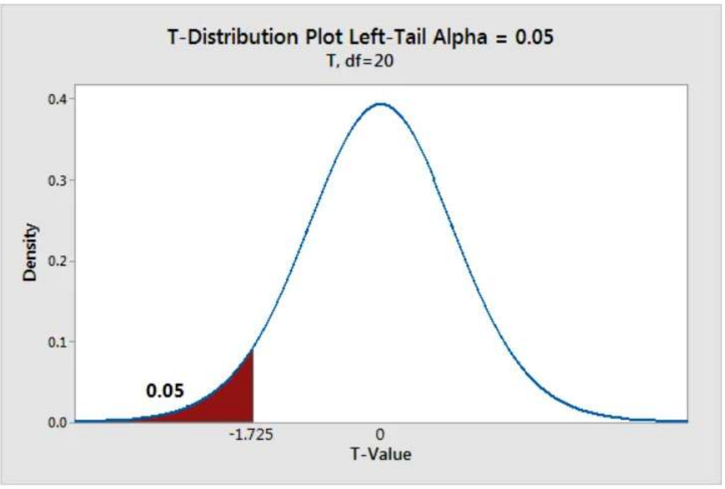
1)

- **Null**: The effect is less than or equal to zero.
- **Alternative**: The effect is greater than zero.



2)

- **Null**: The effect is greater than or equal to zero.
- **Alternative**: The effect is less than zero.



Z-Score:
https://www.youtube.com/watch?v=2tuBREK_mgE&t=207s

Z-test:

https://towardsdatascience.com/z-test-simply-explained-80b346e0e239
https://www.cuemath.com/data/z-test/

t-test:

https://towardsdatascience.com/the-statistical-analysis-t-test-explained-for-beginners-and-experts-fd0e358bbb62
https://www.youtube.com/watch?v=_7IW2PUqe64
https://www.analyticsvidhya.com/blog/2019/05/statistics-t-test-introduction-r-implementation/

f-test:

https://www.cuemath.com/data/f-test/

Measure of Dispersion, Skewness and Kurtosis:

https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/
https://gacbe.ac.in/pdf/ematerial/18BPS34A-U3.pdf

Five Number Summary:
https://www.analyticsvidhya.com/blog/2021/05/five-number-summary-for-analysis/

Chebyshev's Inequalty:

https://www.analyticsvidhya.com/blog/2021/06/complete-guide-to-chebyshevs-inequality-and-wlln-in-statistics-for-data-science/