

# Statistics n Probability for ML

---

Personal Notes – Bhadri Vaidhyanathan

## Contents

Data Basics .....	4
Types: .....	4
Basic Premise: .....	5
Statistics .....	5
Notation and Random Variables .....	5
What is Statistics? .....	8
Central Tendencies: Mean, median, and mode .....	8
Inter-quartile range IQR: .....	9
Finding IQR: .....	9
Measures of Spread: Range, Variance and SD .....	9
Linear transformation and central tendencies: .....	11
Quartiles/ Percentile: .....	12
Shape: .....	12
Robust Statistics: .....	15
Visualizations: .....	15
Histogram: .....	15
Normal Distribution .....	18
Standard Normal Distribution: .....	18
Z scores: .....	18
Median Absolute Deviation (MAD) : .....	20
Inferential Statistics: .....	20
Important Theorems: .....	21
Law of large numbers: .....	21
Central Limit Theorem .....	21
Bootstrapping .....	22
Law of Total Probability .....	24

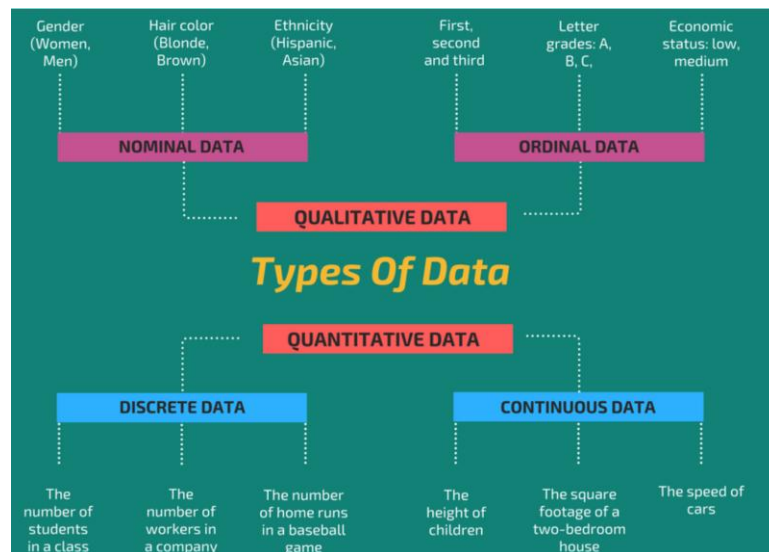
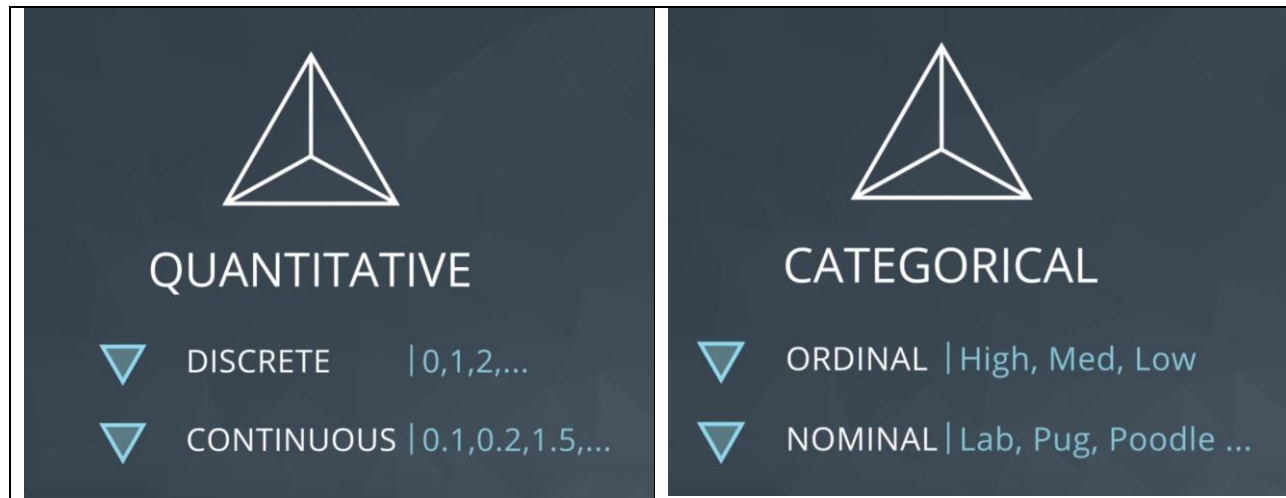
Confidence Intervals: .....	25
Degrees of Freedom: .....	26
Hypothesis Testing:.....	26
T test and Z test .....	29
Disadvantages of Hypothesis testing approach (against Machine Learning): .....	31
Mitigating errors in Hypothesis testing: .....	31
Other Techniques.....	31
P – value.....	31
What are these pvalues? .....	32
Acceptable conclusions:.....	32
Which test? Z test or T test or Chi Square test: .....	33
Statistical Power: .....	33
Probability .....	34
Probability Distribution Functions: .....	34
Conditions for PMF: .....	35
Sampling Distribution.....	39
Normal Distribution .....	39
Normal Distribution Probability Density function: .....	39
Binomial Distribution .....	41
Formula break up:.....	42
Comparison .....	45
Difference Binomial and Bernoulli:.....	45
Conditional Probability .....	46
Baye’s Theorem: .....	47
Baye’s in Python.....	49
Law of total probability.....	50
Regression .....	50
Simple Linear Regression .....	51
Finding the “Line of Best Fit” .....	53
How to Interpret a Least Squares Regression Line .....	53
How to Use the Least Squares Regression Line .....	54
The Coefficient of Determination .....	54

Assumptions of Linear Regression .....	55
What to do if this assumption is violated .....	58
Assumption 2: Independence .....	58
Explanation .....	58
Assumption 3: Homoscedasticity.....	59
What to do if this assumption is violated .....	60
Assumption 4: Normality .....	60
Explanation .....	60
How to determine if this assumption is met .....	60
What to do if this assumption is violated: .....	61
Doing Stats/Prob in Python.....	62
Simulating Coin flips and Dice rolls in Python.....	62
Simulating Coin flips.....	62
Np.random.choice.....	64
<PDdataframe>.sample .....	64
Simulating Binomials: random.binomial .....	64
Simulating sampling distribution .....	65
Confidence intervals .....	66
Simulating Null Hypothesis testing .....	66
ANOVA: .....	66
GLM eqn for ANOVA: .....	70
Which Test to use?.....	71

## Data Basics

### Types:

Numerical and Categorical



(Qualitative aka Categorical, Quantitative aka Numerical)

### Quantitative:

**Discrete:**

are countable in a finite amount of time. a variable whose value is obtained by counting.

**Continuous:**

would (literally) take forever to count. is a variable whose value is obtained by measuring. **Is usually rounded off but has infinite level of sub-unit measurements like time: hours, mins, secs, picoseconds...** while discrete above is fixed like number of people (there is no half person measurements)

**Categorical:**

**Ordinal:** There is an order to the choices available eg: Low/ High, First/Second/Third

**Nominal:** No order eg: Gender (M/F)

## Basic Premise:

A random unbiased sample with sufficient sample size from the population is more likely to contain number of successes that are equal to or near the actual number of successes in a population.

<https://towardsdatascience.com/demystifying-the-binomial-distribution-580475b2bb2a>

# Statistics

## Notation and Random Variables

As a quick recap, **capital letters** signify **random variables**. When we look at **individual instances** of a particular random variable, we identify these as **lowercase letters** with subscripts attach themselves to each specific observation.

For example, we might have **X** be the amount of time an individual spends on our website. Our first visitor arrives and spends 10 minutes on our website, and we would say **X<sub>1</sub>** is 10 minutes.

We might imagine the random variables as columns in our dataset, while a particular value would be notated with the lower case letters.

Notation	English	Example
$X$	A random variable	Time spent on website
$x_1$	First observed value of the random variable $X$	15 mins
$\sum_{i=1}^n x_i$	Sum values beginning at the first observation and ending at the last	$5 + 2 + \dots + 3$
$\frac{1}{n} \sum_{i=1}^n x_i$	Sum values beginning at the first observation and ending at the last and divide by the number of observations (the mean)	$(5 + 2 + 3)/3$
$\bar{x}$	Exactly the same as the above - the mean of our data.	$(5 + 2 + 3)/3$

### **Notation for the Mean**

We took our notation even farther by introducing the notation for summation  $\sum$  (sigma). Using this we were able to calculate the mean as:

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

### Population and Sample Notations in Statistics:

	PARAMETER	STATISTIC
Mean	$\mu$	$\bar{x} \quad \hat{\mu}$
Standard Deviation	$\sigma$	$S \quad \hat{\sigma}$
Variance	$\sigma^2$	$s^2 \quad \sigma^2$
Proportion	$\pi$	$p \quad \hat{\pi}$
Regression Coefficient	$\beta$	$b \quad \hat{\beta}$

Parameter	Statistic	Description
$\mu$	$\bar{x}$	"The mean of a dataset"
$\pi$	$p$	A percent of a population with a certain attribute. Can be found by "The mean of a dataset with only 0 and 1 values - a proportion" also gives the probability of picking 1
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	"The difference in means"
$\pi_1 - \pi_2$	$p_1 - p_2$	"The difference in proportions"
$\beta$	$b$	"A regression coefficient - frequently used with subscripts"
$\sigma$	$s$	"The standard deviation"
$\sigma^2$	$s^2$	"The variance"
$\rho$	$r$	"The correlation coefficient"

## What is Statistics?

Statistics is a set of mathematical methods and tools that enable us to answer important questions about data. It is divided into two categories:

**Descriptive Statistics** - this offers methods to summarise data by transforming raw observations into meaningful information that is easy to interpret and share.

**Inferential Statistics** - this offers methods to study experiments done on small samples of data and chalk out the inferences to the entire population (entire domain). Setting hypothesis and inferring on p value...

## Central Tendencies: Mean, median, and mode

Mean, median, and mode are different measures of center in a numerical data set. They each try to summarize a dataset with a single number to represent a "typical" data point from the dataset.

**Mean:** The "average" number; found by adding all data points and dividing by the number of data points.

**Median:** The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

**Mode:** The most frequent number—that is, the number that occurs the highest number of times.

• Population mean

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

• Sample mean

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



## Inter-quartile range IQR:

Difference between P25 and P75

or difference between median of first half and median of second half

or it is the distance between the first quartile and the third quartile

- Find the first quartile. The first quartile is the median of the data points to the left of the median in the ordered list.
- Find the third quartile. The thirs quartile is the median of the data points to the right of the median in the ordered list.

## Finding IQR:

- A. Even number of values: (Average and separate)
  1. Find the median: average the middle two terms
  2. Find medians of lower and upper halves(odd number of values): pick the middle terms
  3. Find the range between the two medians
- B. Odd number of values: (Just pick and skip)
  - a. Find median: pick the middle value
  - b. Median of two halves: Skip median and take values besides it as boundary value to form the two halves(even number of values). Average the middle two values to find the two medians
  - c. Difference between the two gives the IQR

## Measures of Spread: Range, Variance and SD

Range:  $\text{Max} - \text{Min}$

**Variance** (mean of the squared distance from mean):

## Variance

- Population

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

- Sample

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

Sample is divided by n-1 because when n is used the answers results seem to be skewed while when n-1 is the results are closer to Population mean when validated.

Checkout Khan Academy lessons on this:

<https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/more-standard-deviation/v/review-and-intuition-why-we-divide-by-n-1-for-the-unbiased-sample-variance>

Unbiased sample variance: dividing by n-1 makes the result less biased hence called unbiased sample variance.

### Standard Deviation:

Variance is squared distance from mean while SD is just average distance of each data point from mean.

Standard deviation measures the spread of a data distribution. The more spread out a data distribution is, the greater its standard deviation.

## Standard Deviation

- The standard deviation is the square root of the variance
- The variance is in "square units" so the standard deviation is in the same units as  $x$

- Population

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

- Sample

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

Sample SD even though is divided by  $n-1$  does not make it unbiased since the  $\text{sq.root}$  is non-linear and makes the result biased.

### Linear transformation and central tendencies:

if a variable  $X$  has a mean of  $\mu$ , a standard deviation of  $\sigma$ , and a variance of  $\sigma^2$ , then a new variable  $Y$  created using the linear transformation

$$Y = bX + A$$

will have a mean of  $b\mu + A$ , a standard deviation of  $b\sigma$ , and a variance of  $b^2\sigma^2$ .

The form  $y = wx + b$  is a universal representation of linear transformations. Whether in regression, neural networks, feature scaling, or PCA, it's the fundamental way to map inputs to outputs in a structured way. You will see this in many places:

#### 1. Linear Transformations in General: $y=ax+b$

- **a (or w):** Scaling factor (weights in ML).
- **b:** Translation (bias in ML).
- **Purpose:** Converts one space (input) into another space (output).

#### 2. Linear Regression: $y=mx+b$

- Here, **m (or w)** represents the slope (weight), and  $b$  represents the intercept (bias).
- It models **how an input variable  $x$  affects the output  $y$  in a linear way**.
- **Connection:** Linear regression finds the best  $w$  and  $b$  to minimize the difference between predictions and real values.

#### 3. Neural Networks (Perceptrons & Deep Learning) : $y=wx+b$

- Each **neuron** in a neural network applies this transformation.
- **w** (weight) determines how important an input  $x$  is.
- **b** (bias) allows flexibility in learning.
- Then, an **activation function** (like ReLU or sigmoid) is applied to introduce non-linearity.

→ **Connection:** Each neuron learns a weighted sum of inputs, just like linear regression, but stacked in layers for more complex modeling.

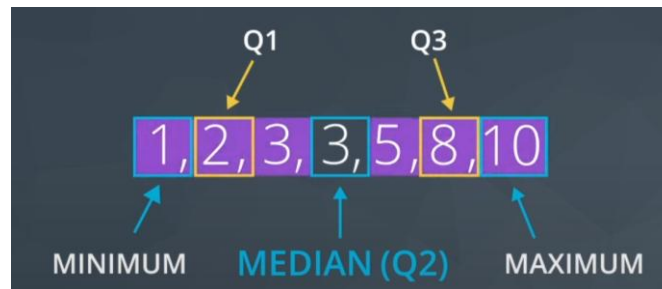
#### Why This Form Is Everywhere

- **Simplicity:** The simplest transformation that keeps the structure of data.
- **Interpretability:** Easy to understand and visualize.
- **Building Block:** Even in complex models (like deep learning), they start with linear operations before adding non-linearity.

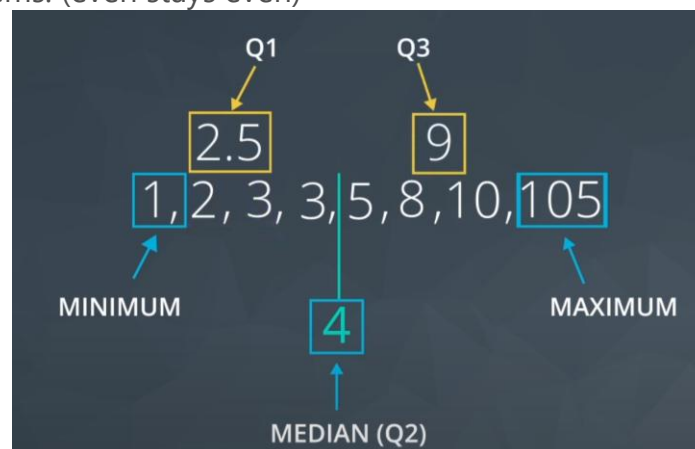
## Quartiles/ Percentile:

Odd number of items: (keeping it odd)

Middle item is skipped for finding the Q1 and Q3



Even number of items: (even stays even)



## Shape:

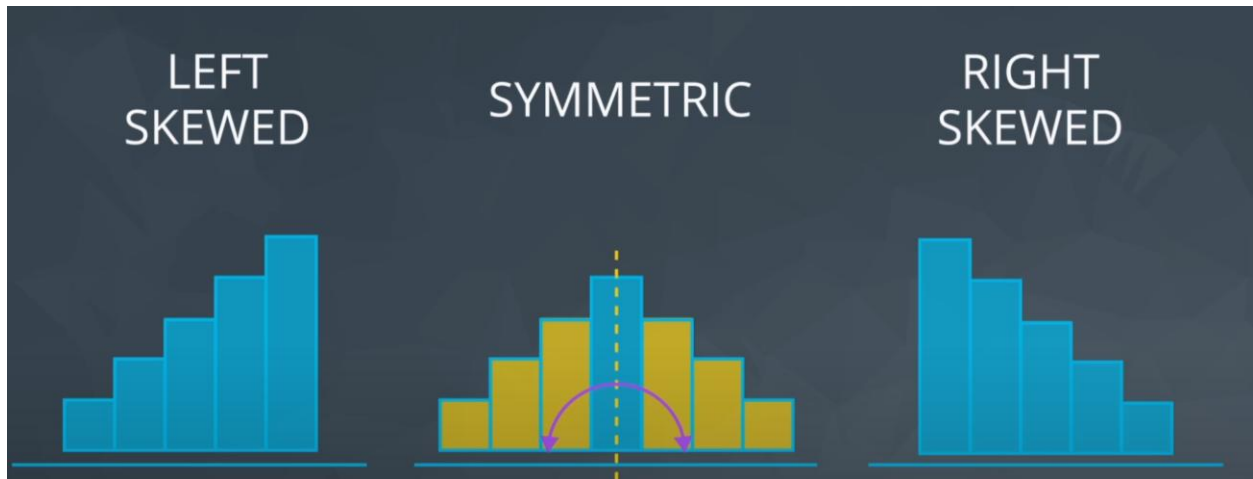
data is frequently associated with one of the three shapes:

1. Right-skewed (positively skewed – from perspective of bell curve)
2. Left-skewed (negatively skewed)
3. Symmetric (frequently normally distributed)

Depending on the shape associated with our dataset, certain measures of center or spread may be better for summarizing our dataset.

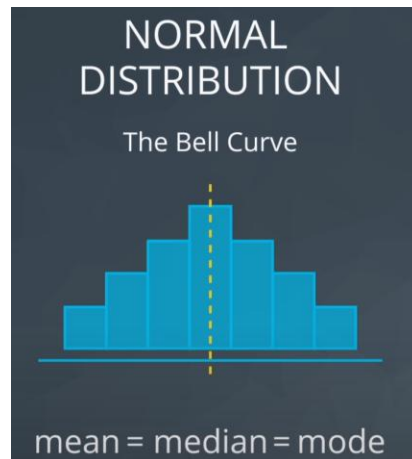
When we have data that follows a normal distribution, we can completely understand our dataset using the mean and standard deviation.

However, if our dataset is skewed, the 5 number summary (and measures of center associated with it) might be better to summarize our dataset.



***Symmetric/ Normal Distribution:***

Mean = Median = Mode

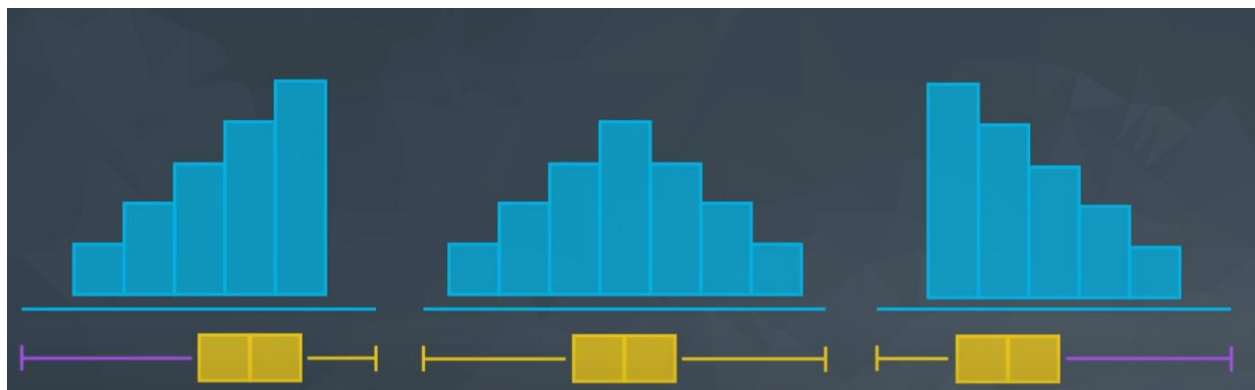


***Skewed:***

The skew pulls the mean away while the median remains close to the mode.



Boxplots for shapes:



The left whisker is longer for left skewed  
Right whisker is longer for right skewed

Is my data normally distributed:

If you aren't sure if your data are normally distributed, there are plots called [normal quantile plots](#) and statistical methods like the [Kolmogorov-Smirnov test](#) that are aimed to help you understand whether or not your data are normally distributed.

Implementing this test is beyond the scope of this class, but can be used as a fun fact.

## Robust Statistics:

	robust	non-robust
center	median	mean
spread	IQR	SD, range

*skewed,  
with extreme  
observations*

*symmetric*

**Median and IQR Spread** are robust stats for data that are skewed or high extreme observations / Outliers.

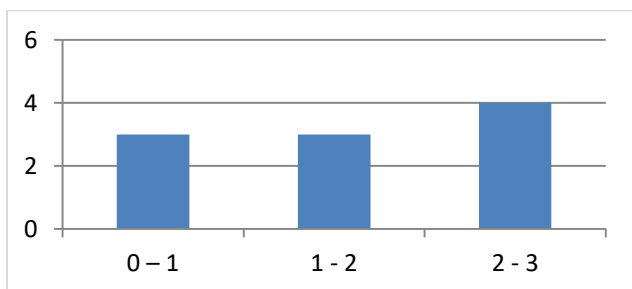
## Visualizations:

### Histogram:

#### Frequency Histogram:

Data: 0.3, 0.4, 0.8, 1, 1.3, 1.35, 2, 2.5, 2.6, 2.7

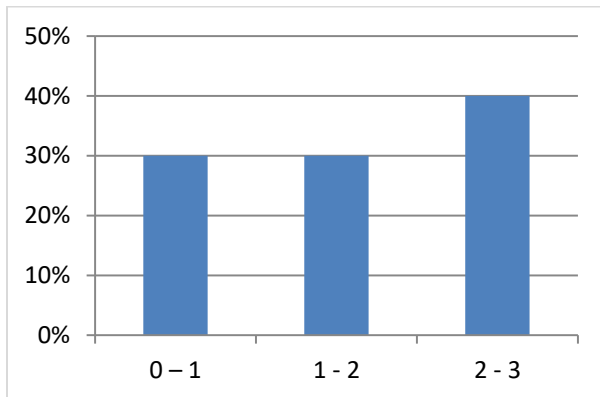
0 - 1	3
1 - 2	3
2 - 3	4



## Relative Frequency Histogram

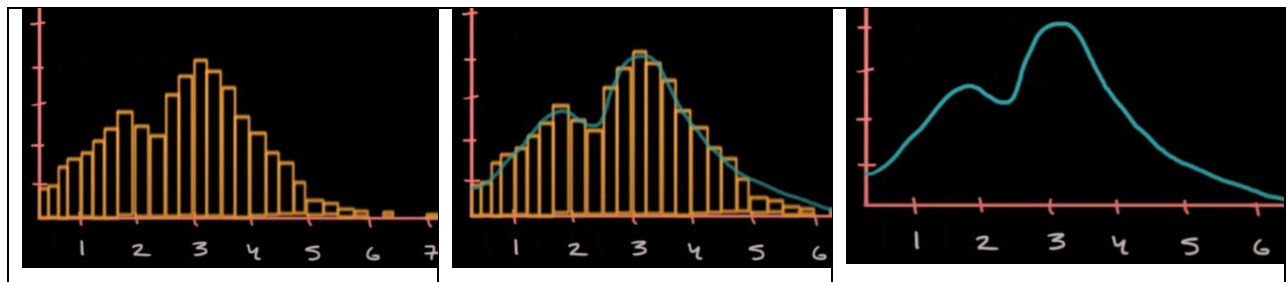
More useful for very large data set compared to normal freq. histogram

Data Ranges	Freq.	Relative Freq.
0 - 1	3	$3 / 10 = 30\%$
1 - 2	3	$3 / 10 = 30\%$
2 - 3	4	$4 / 10 = 40\%$



## Density Curve:

Line joining the top of the bars in relative freq. distribution



Characteristics:

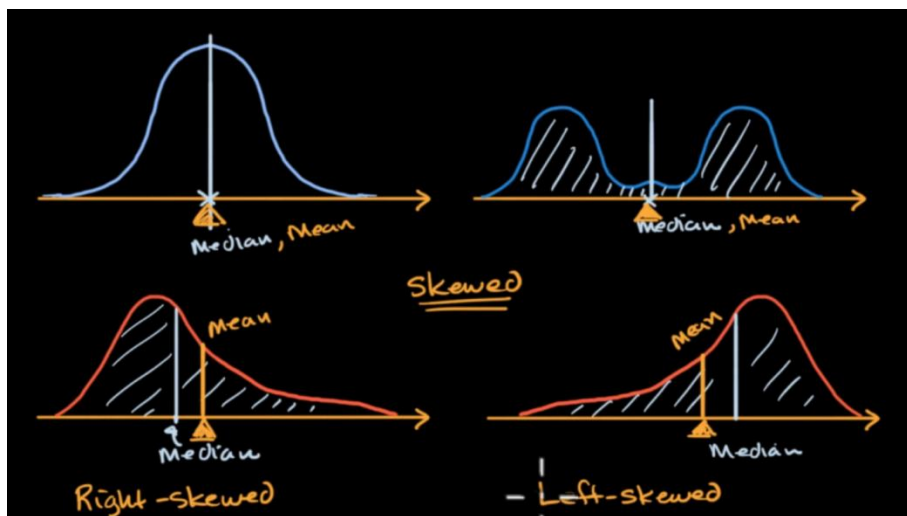
- Area under the curve is 100%
- Area under curve gives the % of values that falls in that range
- Eg: % of values between 2 and 4





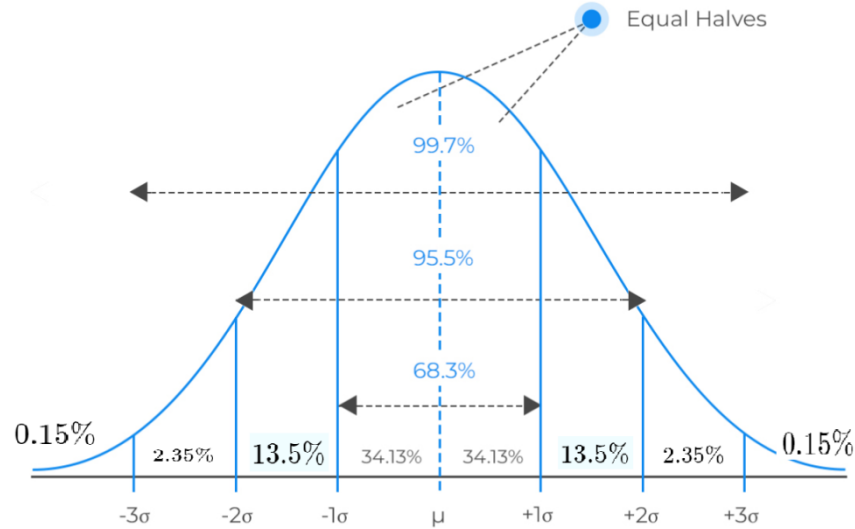
- You must not use above concept to look for % of values that meet one value eg: 3 but only for range like 3 to 4.
- Values that exactly meet one value is a line with no width which has no area hence this results in zero area
- Alternative approach is to consider a range like 2.9 to 3.1 and area of this give approximate no. of values at 3

Skewed:



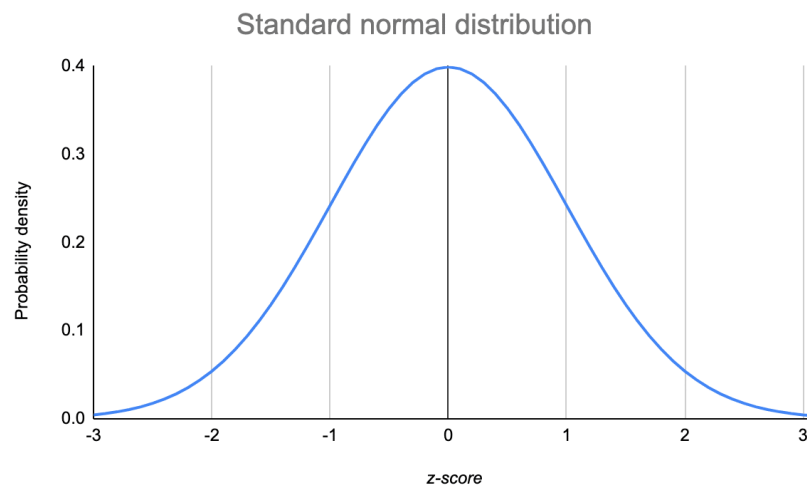
- Skewed when median is not equal to mean
- Median is where the area below the graph are same on both sides
- Mean (visually) is the point where we can physically put a fulcrum and both sides will be balanced.
  - How fulcrum? Why cant median point be the fulcrum.  
Because physically, a small value that is far from the fulcrum can balance a heavier weight closer to the fulcrum. Mean is that point that can balance weights on both sides even when they are not the same.

## Normal Distribution



### Standard Normal Distribution:

Zscore (how many SDs from mean)



### Z scores:

How many SDs away is a point from mean?

A z-score measures exactly how many standard deviations above or below the mean a data point is.

$$z = \frac{\text{data point} - \text{mean}}{\text{standard deviation}}$$

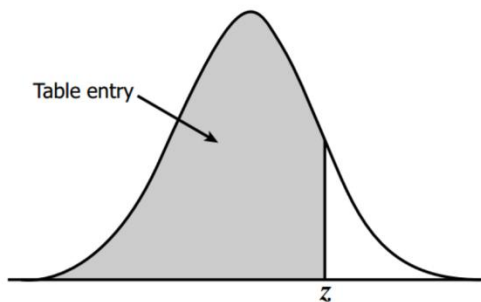
$$z = \frac{x - \mu}{\sigma}$$

A Z score can be calculated for any distribution and its formula is not only for normal distribution. Z score with above formula can be found for any distribution with mean and SD.

### **Z Table: (Area of curve for a z score)**

Z Tables give the area under the curve for a given z score value.

Positive Z tables give the areas above mean and it includes the 50% area below the mean.



Find values on the right of the mean in this

z-table. Table entries for z represent the

area under the bell curve to the left of z.

Positive scores in the Z-table correspond to

the values which are greater than the

mean.

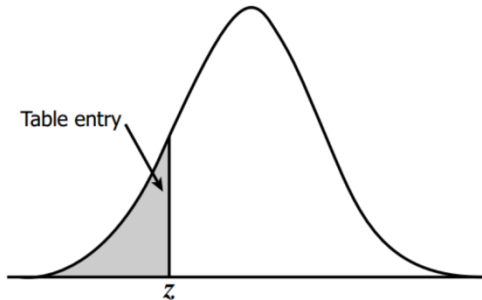
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Eg:

We find that Z score is 0.57 SD above the mean. Then using Z table we can tell what % of values are below 0.57 and what % above 0.57.

Using above table, we see that 0.57 has .7157 value in table thus 71.57% of values are below 0.57 SD.

### Negative Z table:



Find values on the left of the mean in this negative Z score table. Table entries for z represent the area under the bell curve to the left of z. Negative scores in the z-table correspond to the values which are less than the mean.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048

### Median Absolute Deviation (MAD) :

Modified Z score:

Z score gets affected by outliers since mean gets moved towards outlier skewing our numbers.

Median is a good measure of central tendency when we have serious outliers.

Using median instead of mean in calculation of z stat gives modified Z score.

$$\text{Modified } z = (x - \text{median}(x)) / \text{MAD}$$

$$\text{MAD} = \text{median} | (x - \text{median}(x)) |$$

### Inferential Statistics:

Descriptive Statistics

Descriptive statistics is about describing our collected data.

Measures of center, measures of spread, shape of our distribution, and outliers.

---

## Inferential Statistics

**Inferential Statistics** is about using our collected data to draw conclusions to a larger population.

We looked at specific examples that allowed us to identify the

1. **Population** - our entire group of interest.
2. **Parameter** - numeric summary about a population - **Parameter is abt Population - PP**
3. **Sample** - subset of the population
4. **Statistic** numeric summary about a sample - **Statistic is abt Sample - SS**

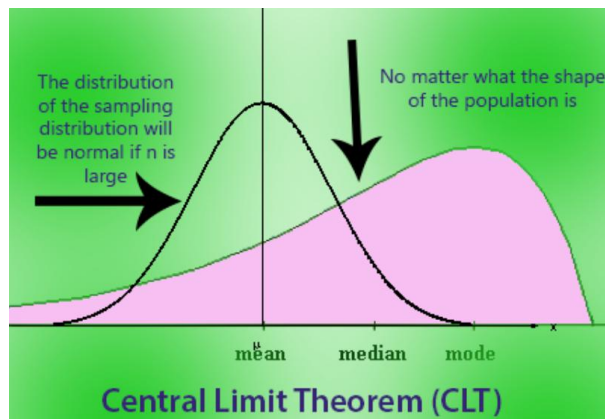
## Important Theorems:

### Law of large numbers:

The Law of Large Numbers says that as our sample size increases, the **sample mean** gets closer to the **population mean**.

As the number of trials increase, the actual value of a probability will converge to the expected means like say head-toss is a 50:50 probability when in reality it might not be exactly 50:50 and in only few trials it could be worse like 75:25

## Central Limit Theorem



If we take large random samples ( $n \geq 30$ ) from the population **with replacement**, then the distribution of the means of sample will be normally distributed.

- If you take many random samples from any population (regardless of its shape or distribution like Poisson or Binomial or Exponential) **WITH REPLACEMENT**, and calculate the mean of each sample...

- What if there is no replacement: In the **CLT**, the assumption is that the sampling is done **with replacement**, which ensures independence between samples. If sampling is done **without replacement**, the independence assumption is violated because the population changes as values are removed.
- The distribution of those sample means will look like a **normal distribution** (bell-shaped curve)...
- **As long as the sample size is large enough** (usually  $n \geq 30$ ).

(Basic idea behind bootstrapping with replacement)

No matter how weird or skewed the original population is, when you take large enough samples and look at their means, those means will follow a normal distribution. The CLT is why we can use normal distribution-based methods so often in statistics!

## Whats with the name?

It's called the **Central** Limit Theorem (CLT) because it plays a **central** role in probability and statistics. The term "limit" refers to the fact that the theorem describes the **limiting behavior** of sample means as the sample size increases.

The CLT is fundamental because it shows that, regardless of the original population distribution, the sampling distribution of the mean approaches a normal distribution as the sample size grows. This is why it's "central"—it underpins many statistical methods and justifies the use of the normal distribution in various real-world applications.

## Why Is the CLT Important?

1. **Simplifies analysis:** Even if the population is not normal, we can still make inferences using normal distribution tools if we use sample means.
2. **Foundation for hypothesis testing:** Many statistical tests assume normality of the sample mean, which the CLT ensures.

The **Central Limit Theorem** actually applies for these well known statistics:

1. Sample means ( $\bar{X}$ )
2. Sample proportions ( $p$ )
3. Difference in sample means ( $\bar{X}_1 - \bar{X}_2$ )
4. Difference in sample proportions ( $p_1 - p_2$ )

## Bootstrapping

Sampling from a sample a million times (as per CLT, the sampling distribution will turn out to be normal)

<https://statisticsbyjim.com/hypothesis-testing/bootstrapping/>

**Bootstrapping** is sampling with replacement in such way that resampling (resample is same size as original sample) can pick the same item again and again (since we replace it) but if done million times the variability in each resample is as if you have been taking samples from original population.

Original	Bootstrap1	Bootstrap2	Bootstrap3	Bootstrap4
1	1	2	1	1
2	1	3	2	1
3	3	3	3	1
4	3	3	5	4
5	5	4	5	5

There is a small possibility that the original sample is so unrepresentative and in that case it was bad luck. But other methods like traditional would have been equally bad which makes bootstrapping superior. Various studies over the intervening decades have determined that bootstrap sampling distributions approximate the correct sampling distributions. As the sample size increases, bootstrapping converges on the correct sampling distribution under most conditions.

Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples. This process allows you to calculate standard errors, construct confidence intervals, and perform hypothesis testing for numerous types of sample statistics. It mitigates some of the pitfalls encountered within the traditional approach, or take a large sample approach.

Bootstrapping resamples the original dataset with replacement many thousands of times to create simulated datasets. This process involves drawing random samples from the original dataset. Here's how it works:

1. The bootstrap method has an equal probability of randomly drawing each original data point for inclusion in the resampled datasets.
2. The procedure can select a data point more than once for a resampled dataset. This property is the "with replacement" aspect of the process.
3. The procedure creates resampled datasets that are the same size as the original dataset.

The process ends with your simulated datasets having many different combinations of the values that exist in the original dataset. Each simulated dataset has its own set of sample statistics, such as the mean, median, and standard deviation.

<https://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307>

Traditional Approach	Bootstrapping
The theory states that, under certain conditions such as large sample sizes, the sampling distribution will be approximately normal, and the standard deviation of the distribution will be equal to the standard error.	
Sample is used to find different statistics (mean, median, std), sample is used to calculate population estimates to then make inferences on.	Sample is used as population and resampling is done with replacement
What if the	
Assumption: data are normally distributed	No assumption
	bootstrapping is consistent and more accurate
Standard intervals/confidence levels obtained from sample variance	Standard intervals/confidence levels obtained from sampling distribution
A primary difference is how they estimate sampling distributions.	
Traditional hypothesis testing procedures require <b>equations</b> that estimate sampling distributions using the properties of the sample data, the <b>experimental design</b> , and a <b>test statistic</b> . To obtain valid results, you'll need to use the proper test statistic and satisfy the assumptions.	This method takes the sample data that a study obtains, and then resamples it over and over to create many simulated samples. Each of these simulated samples has its own properties, such as the mean. When you graph the distribution of these means on a histogram, you can observe the sampling distribution of the mean. You <b>don't need to worry about test statistics, formulas, and assumptions</b> .

## Law of Total Probability

The **Law of Total Probability** is a fundamental rule in probability that helps us calculate the probability of an event by breaking it into smaller, simpler parts based on a partition of the sample space.

B1, B2, B3 are mutually exclusive events. A is a common event that happens separately in B1, B2 and B3 sample space: A is a subset of all 3. then if P(A) in B1 space, P(A) in B2 space and P(A) in B3 space is

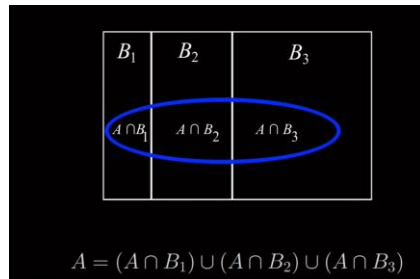


known separately, to get overall  $P(A)$  you can just add them together after multiplying with the appropriate  $P(B_x)$  (conditional probability formula).

$$P(A) = P(B_1 | A) + P(B_2 | A) + P(B_3 | A)$$

$P(A)_{B_1}$ ,  $P(A)_{B_2}$  and  $P(A)_{B_3}$  are actually arrived at using conditional probability formula like

$$P(A)_{B_1} = P(B_1) * P(A | B_1).$$



<https://www.youtube.com/watch?v=7t9jyikrG7w>

### Simple Example: Scenario

Suppose there are three factories  $B_1, B_2, B_3$  that produce the same type of lightbulbs. The probabilities that a randomly chosen bulb comes from each factory are:

- $P(B_1)=0.4$  = 0.4 (40% from Factory 1),
- $P(B_2)=0.35$  = 0.35 (35% from Factory 2),
- $P(B_3)=0.25$  = 0.25 (25% from Factory 3).

The probability that a bulb is defective ( $A$ ) depends on the factory:

- $P(A|B_1)=0.02$  (2% defective from Factory 1),
- $P(A|B_2)=0.03$  (3% defective from Factory 2),
- $P(A|B_3)=0.05$  (5% defective from Factory 3).

Using the Law of Total Probability:

$$P(A) = P(A|B_1)*P(B_1) + P(A|B_2)*P(B_2) + P(A|B_3)*P(B_3)$$

Substitute the values:

$$P(A) = (0.02)(0.4) + (0.03)(0.35) + (0.05)(0.25)$$

$$P(A) = 0.008 + 0.0105 + 0.0125 = 0.031$$

So, the total probability that a bulb is defective is **3.1%**.

## Confidence Intervals:

### Degrees of Freedom:

The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based. The degrees of freedom of an estimate of variance is equal to  $N - 1$ , where  $N$  is the number of observations.

Bias:

Bias refers to whether an estimator tends to either over or underestimate the parameter.

Eg: Weighing scale 1 is highly accurate but incorrectly calibrated such that it overstates weight by 1 pound. This is bias.

A statistic is biased if the mean of the sampling distribution of the statistic is not equal to the parameter. The mean of the sampling distribution of a statistic is sometimes referred to as the expected value of the statistic.

Sampling variability

Sampling variability refers to how much the estimate varies from sample to sample.

Eg: Weighing scale 2 is not that accurate but it is equally likely to give a positive or negative result such that it is unbiased.

sampling variability of a statistic refers to how much the statistic varies from sample to sample and is usually measured by its standard error ; the smaller the standard error, the less the sampling variability.

standard error of the mean is:

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

Statistics differ in their sampling variability even with the same sample size. For example, for normal distributions, the standard error of the median is larger than the standard error of the mean. The smaller the standard error of a statistic, the more efficient the statistic.

The **relative efficiency** of two statistics is typically defined as the ratio of their standard errors. However, it is sometimes defined as the ratio of their squared standard errors.

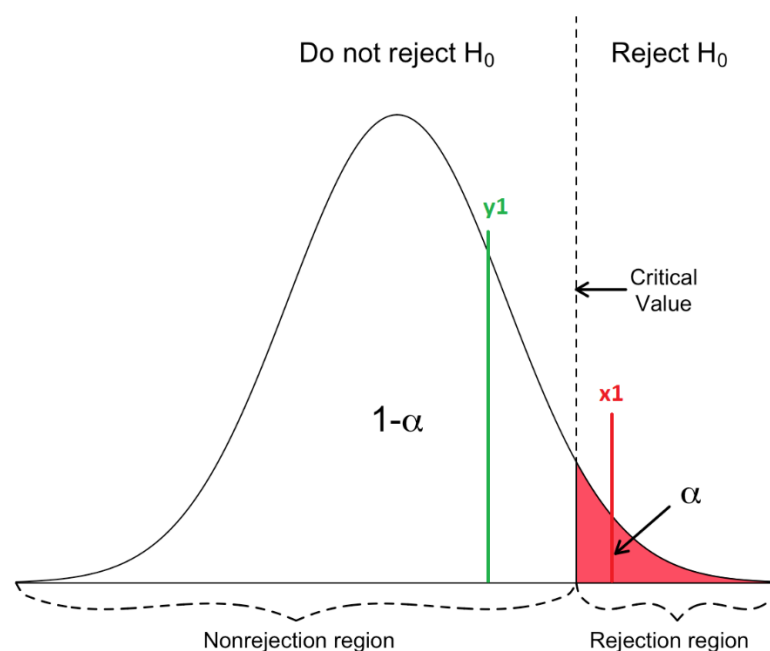
## Hypothesis Testing:

Refer Stats Key Concepts File

In layman terms:

Hypothesis testing is a framework to check if a value  $x$  is close enough to population mean or too faraway. This is done by assuming the population is in a certain distribution (like normal distribution) with some mean  $\mu$  and sd  $\sigma$ . Knowing these values,  $x$ ,  $\mu$  and  $\sigma$  and other facts, one can find the p value/probability (pink area below) of the value  $x$ . If that p value is less than 5% then it implies it is far (many  $\sigma$ s away from mean) away from population mean. If probability (p value

Note: pvalue looked up in charts gives the area under curve and not a value for one point.



Based on p value, you can tell if value its falls in the critical region(too faraway) or acceptance region(close to pop. mean). The threshold is the alpha  $\alpha$  value usually 0.05. Hence p value is less then 0.05 then it means accept alternate Hyp and reject null. If p value is above 0.05 then null hyp. is not rejected.

One sided or two sided?

Alt. Hypothesis dictates if we sld select two-tail or one-tail test.

If close enough, one can say it is from same population (null not rejected) and not "different" from rest of population

but if too faraway, it is not from same population and likely from a different population (null rejected).

Prereqs: 1) distribution of sampling is a normal distribution (CLT)

It is difficult to prove many inferences but it is easy to disprove an inference by identifying existing exception to the inference. This is the primary basis of Inferential Statistics.

1. Hypotheses
2. Significance
3. Sample
4. P Value
5. Decision

#### 1. Hypotheses

H0: Null Hypotheses

Statement of status quo/no change or no effect

The hypothesis that an apparent effect is due to chance.

When the null hypothesis is rejected, the effect is said to be statistically significant.

H1: Alternative Hypotheses

Usually the statement that is being proven.

There is an effect of...

H1 is first setup and the opposite is H0.

Few rules:

- A. The  $H_0$  is true before you collect any data.
- B.  $H_0$  usually states there is no effect or that two groups are equal.
- C.  $H_0$  and  $H_1$  are competing, non-overlapping hypotheses.
- D. Hypotheses always are inferences on Population Parameters and not Sample Statistic.
- E. Null Hypotheses should have equality statement ( $=$  or  $\geq$  or  $\leq$ ) while alternative hypotheses should not have equality ( $\neq$  or  $<$  or  $>$ )

#### 2. Significance Level

Significance level is the probability that Null Hypotheses is wrong.  
Usually 0.05 is selected as significance level aka alpha ( $\alpha$ ) value

3. Sample

Take a sample from population and derive the statistic in question

4. P Value

Calculate/find the p value

5. Decide

If the p value is less than significance, then null hypotheses is rejected

When the null hypothesis is rejected, the effect is said to be **statistically significant**.

Null hypothesis of different tests:

**Pearson's correlation test :**

the null hypothesis for the Pearson's correlation test is that there is no relationship between two variables.

**Student's t test:**

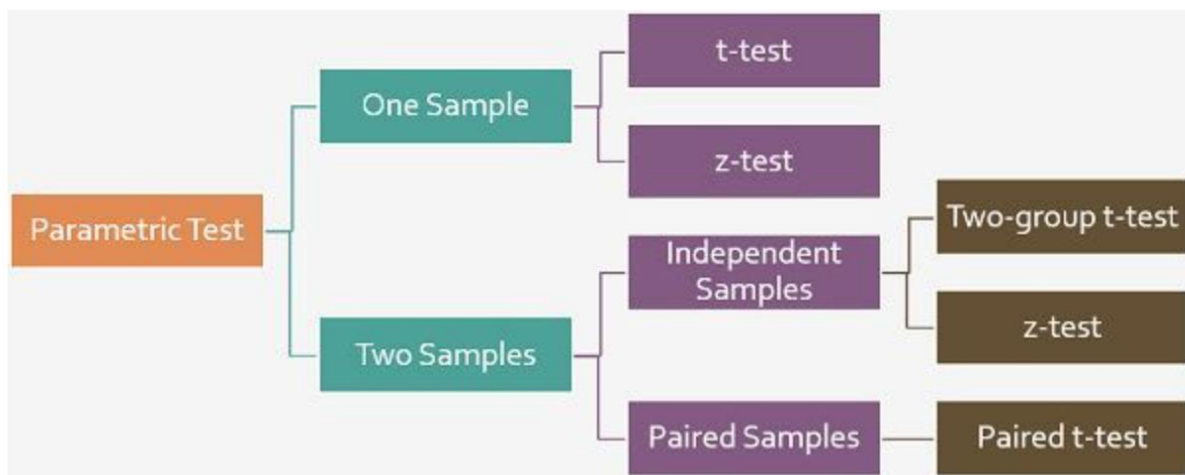
The null hypothesis for the Student's t test is that there is no difference between the means of two populations.

**T test and Z test**

[Refer Stats Key Concepts File](#)

BASIS FOR COMPARISON	T-TEST	Z-TEST
Meaning	T-test refers to a type of parametric test that is applied to identify, how the means of two sets of data differ from one another when variance is not given.	Z-test implies a hypothesis test which ascertains if the means of two datasets are different from each other when variance is given.
Based on	Student-t distribution	Normal distribution
Population variance	Unknown	Known

BASIS FOR COMPARISON	T-TEST	Z-TEST
	(approx. from sample)	
Sample Size	Small	Large



T test:

### Assumptions of T-test:

- All data points are independent.
- The sample size is small. Generally, a sample size exceeding 30 sample units is regarded as large, otherwise small but that should not be less than 5, to apply t-test.
- Sample values are to be taken and recorded accurately.

The test statistic is:

$$\text{T-test} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$\bar{x}$  is the sample mean  
 $s$  is sample standard deviation  
 $n$  is sample size  
 $\mu$  is the population mean

### Setting up $H_0$ and $H_1$ :

One tailed	Two tailed
Testing X is greater than Y	Testing if X is same or different from Y (don't care if X is greater or lesser, just different)
$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

### Disadvantages of Hypothesis testing approach (against Machine Learning):

Eg: We want to open a mikshake shop and want to know if chocolate or vanilla shake is more popular. We did a survey with 1 million people and hypothesis testing approach showed that chocolate was more popular so we went with it.

What was apparent is that chocolate was more popular but 51% more popular and vanilla was 49% popular which practically means that the shop should sell both flavors since they difference is small. This is not apparent in Hypothesis testing approach hence Machine learning approach (like linear and logistic regression) is better.

### Mitigating errors in Hypothesis testing:

If 0.05 p value or 5% p value is used, then 1 in 20 is wrong. Our test could have been that 1 in 20 and thus giving us wrong conclusions. How do we mitigate this? One approach is called **Bonferroni** correction.

When performing more than one hypothesis test, your type I error compounds. In order to correct for this, a common technique is called the **Bonferroni** correction. This correction is **very conservative**, but says that your new type I error rate should be the error rate you actually want divided by the number of tests you are performing. Therefore, if you would like to hold a type I error rate of 1% for each of 20 hypothesis tests, the **Bonferroni** corrected rate would be  $0.01/20 = 0.0005$ . This would be the new rate you should use as your comparison to the p-value for each of the 20 tests to make your decision.

## Other Techniques

Additional techniques to protect against compounding type I errors include:

1. [Tukey correction](#)
2. [Q-values](#)

## P – value

REFER Key Stats docx.

pvalue of 1 corresponds to the value at the mean this pvalue of 0.05 is far away and thus indicating that it is not part of the population and part of another.

## What are these pvalues?

Refer Stats Key Concepts File

1. pvalues against each beta value in regression results like below.

This pvalue tells if that predictor or that beta is = 0 or not.

high pvalue means it is probably =0 and thus might have to removed.

pvalue below 5% or 0.05 means it is far from null hypo that says that it is 0.

### Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-4.3576	1.0353	-4.209	2.57e-05	***
gpa	1.0511	0.2989	3.517	0.000437	***
---					

By comparing our p-value to our type I error threshold ( $\alpha$ ), we can make our decision about which hypothesis we will choose.

$P \text{ value} \leq \alpha(0.05) \Rightarrow \text{Reject } H_0$

$pval > \alpha \Rightarrow \text{Fail to reject } H_0$



Eg:

P-VALUE/ALPHA	CONCLUSION
p-value = 0.03, alpha = 0.05	Reject the null
p-value = 0.20, alpha = 0.01	Fail to reject the null
p-value = 0.10, alpha = 0.05	Fail to reject the null

#### Acceptable conclusions:

1. We have evidence to reject the null hypothesis
2. We fail to reject the null hypothesis

#### Which test? Z test or T test or Chi Square test:

- If the test is about a value and mean and its normal distribution then its Z test
- If the test is about comparing variances of two sets of data then it is F test
- If the test is to check if shape of a distribution is uniform or not then its Chi Sq test

**Simpson's paradox**, which also goes by several other names, is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

Example:

DATA:

	MALE			FEMALE		
	APPLIED	ADMITTED	RATE	APPLIED	ADMITTED	RATE
MAJOR A	900	450	50%	100	80	80%
MAJOR B	100	10	10%	900	180	20%
BOTH	1,000	460	46%	1,000	260	26%

WHO IS BEING FAVORED ?

- MALE
- FEMALE

- 1) When Female data alone is studied and %s calculated; once can tell Female are being favored since the acceptance rate for each major is higher than males
- 2) When the data for majors is combined, the trend reverses and male seem favored

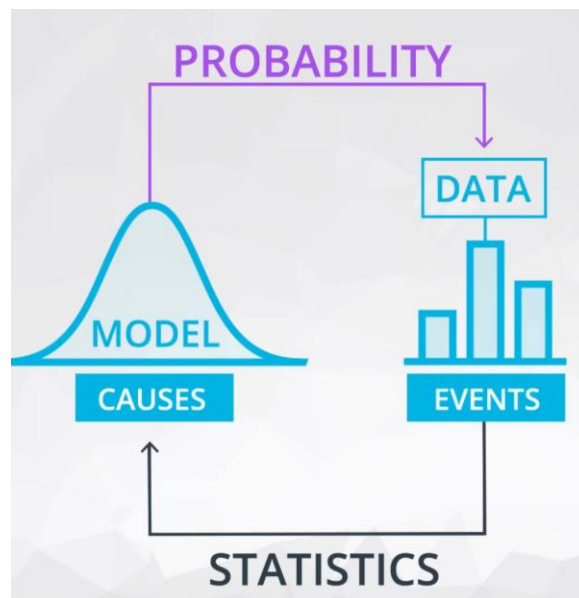
### Statistical Power:

Refer Stats Key Concepts File

The statistical power of a hypothesis test is the probability of detecting an effect, if there is a true effect present to detect.

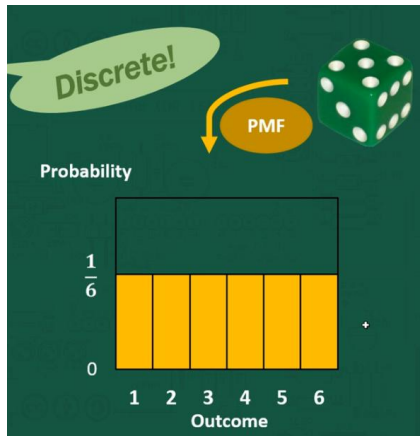
- A power analysis can be used to estimate the minimum sample size required for an experiment, given a desired significance level, effect size, and statistical power.

### Probability

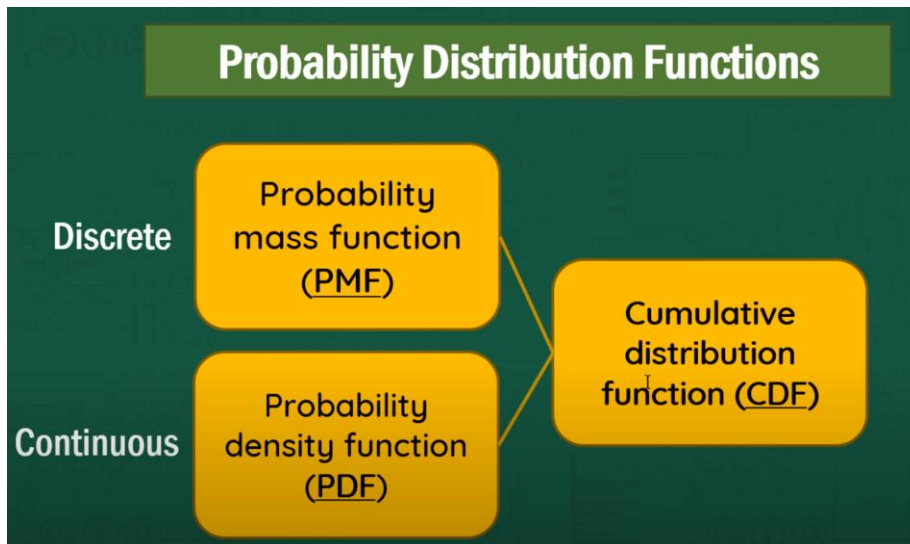


### Probability Distribution Functions:

Probability distribution function tells the probability of getting different values possible for a measurement around the mean of a population. Eg: the prob. of each number on a dice



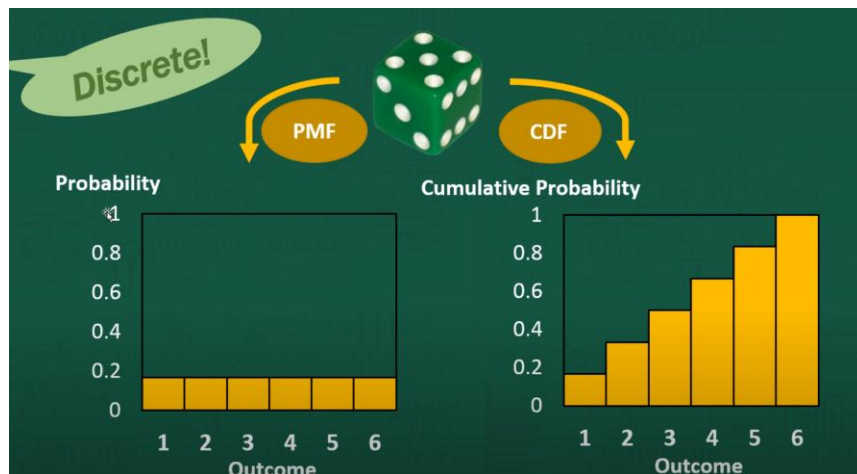
<https://www.youtube.com/watch?v=YXLVjCKVP7U> – zedstatistics



Discrete:

Eg: roll of dice

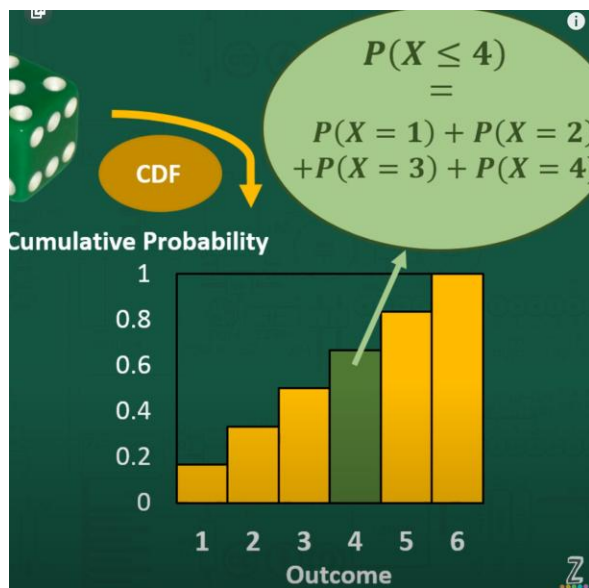
Probability of each number in dice =  $1/6$



### Conditions for PMF:

1. Sum of probabilities of all the discrete var = 1

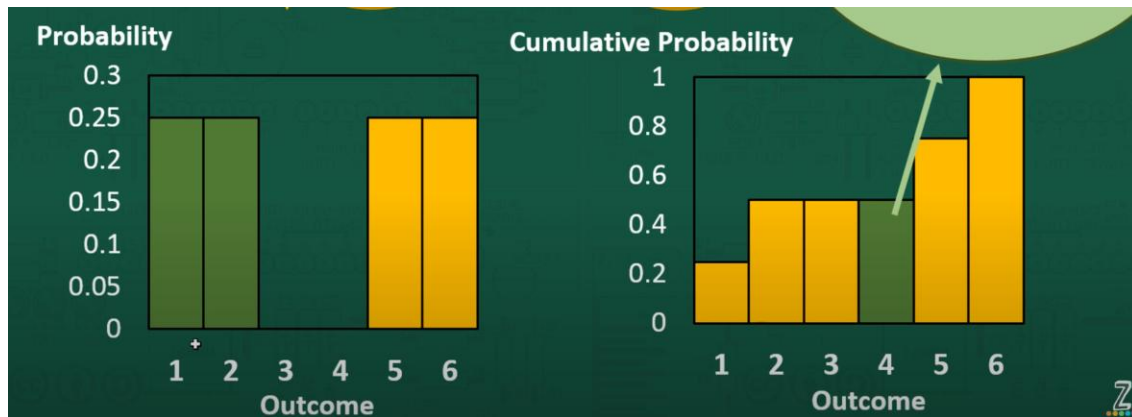
Cumulative density function is for questions such as  $P(X \leq 4)$  which is  $P(X=1)+P(X=2)+P(X=3)+P(X=4)$  or pick observe in the graph



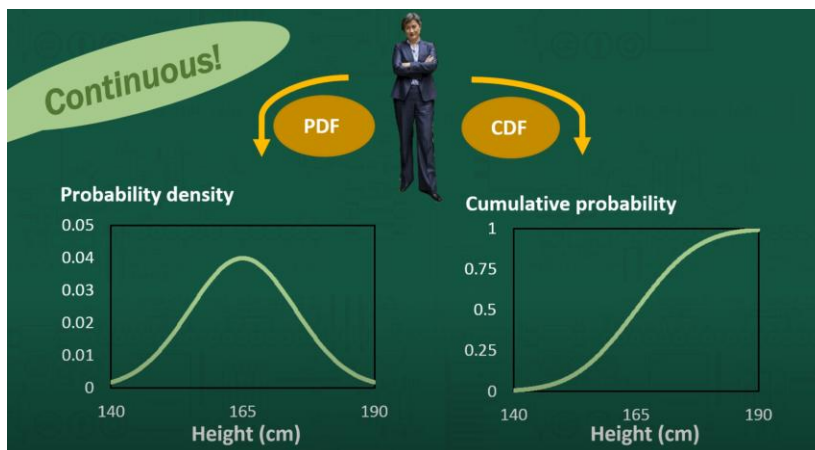
Conditions of CDF:

1. The final bar should be 1

Note: if CMF is flat that means  $P(X)$  for those are zero.



Probability Density Function:



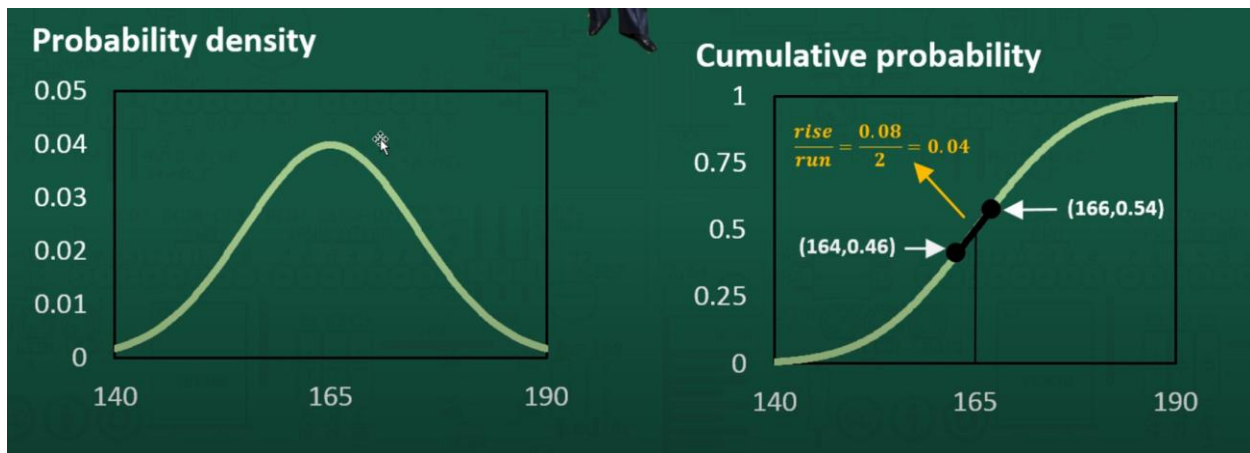
Overall:

PMF or PDF can be used for the  $P(X)$  for a specific value like  $X=1 / 13$  (Discrete) or a minute range like  $X=13.1$  to  $13.2$  (Continuous) while CMF is useful for  $P(X \leq )$  like  $P(X \leq 4)$  or  $P(X < = 143.2)$

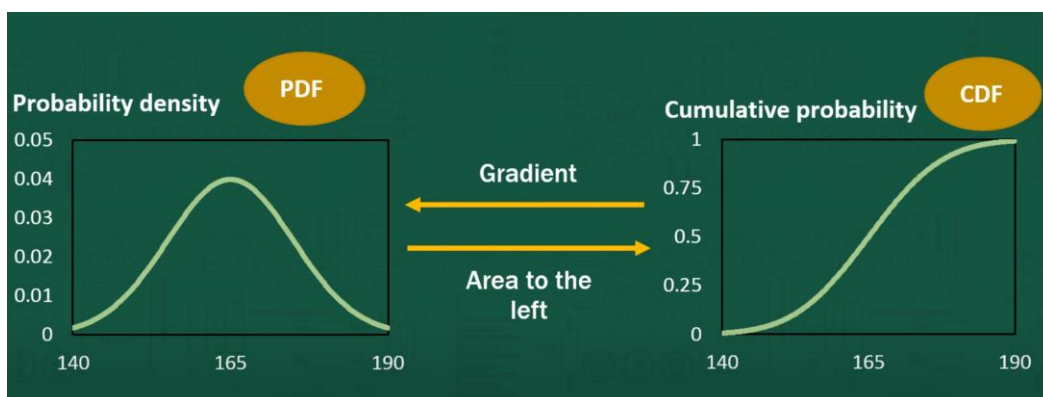
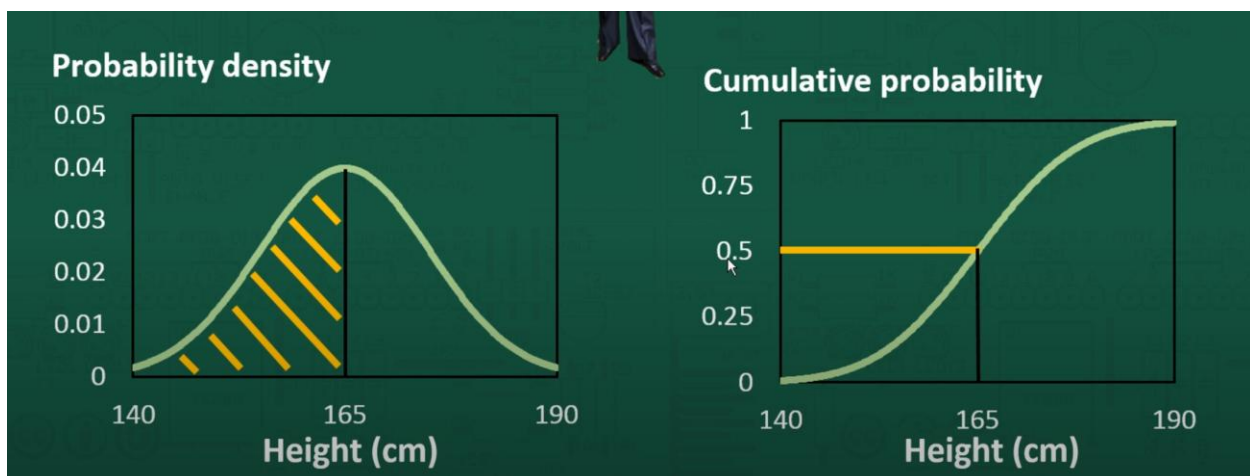
Relation between PDF and CDF:

1) slope of CDF can be used to find the PDF at a certain value

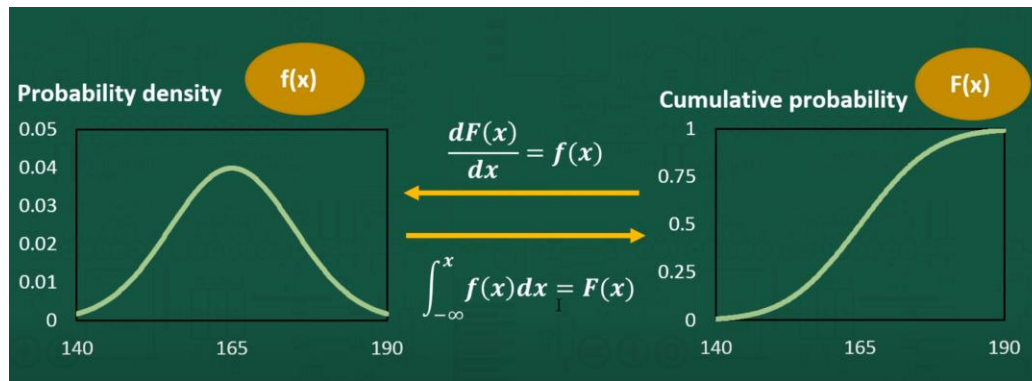
the slope of CDF at  $P(X=165) = 0.04$  is  $P(X=165)$



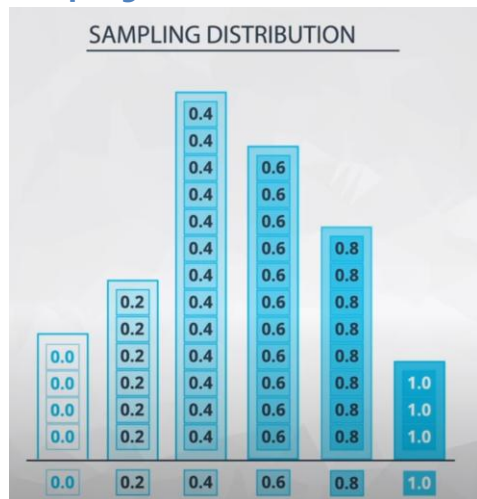
2) Area to the left of a point in PDF gives the value of CDF



With Calculus and Integral formulae:



## Sampling Distribution



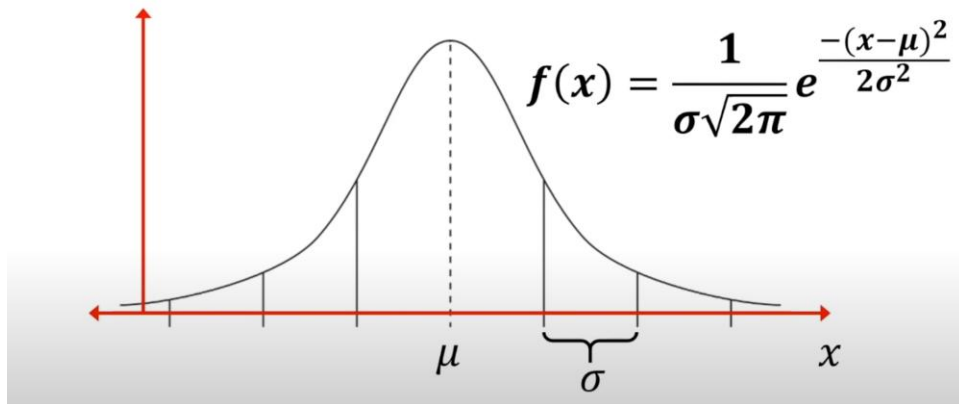
## Normal Distribution

.... yret to fill

Normal Distribution Probability Density function:

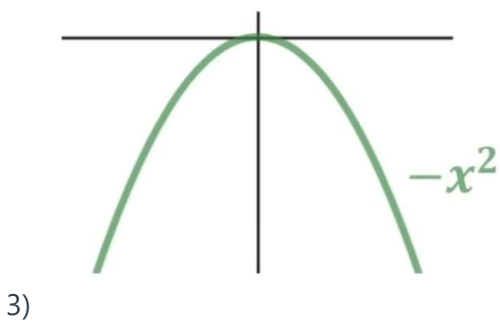
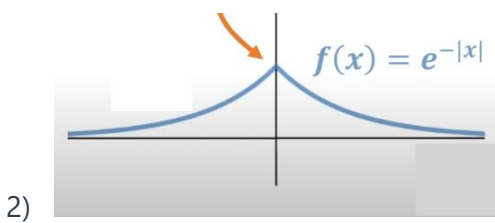
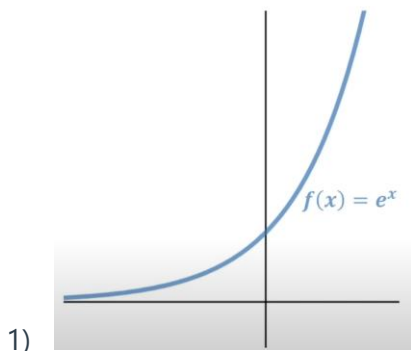
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$	<p>where</p> $z = \frac{x - \mu}{\sigma}$
--	--	---

## Probability Density Function Of The Normal Distribution

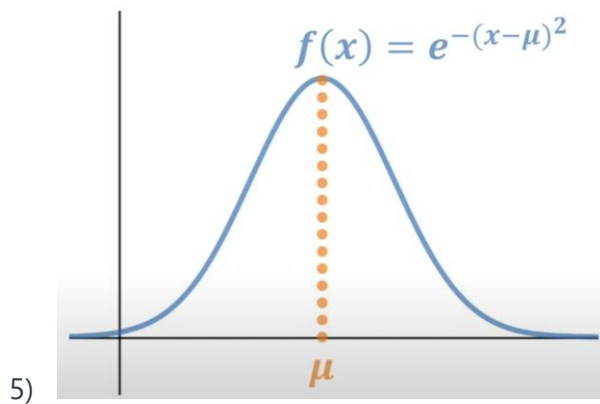
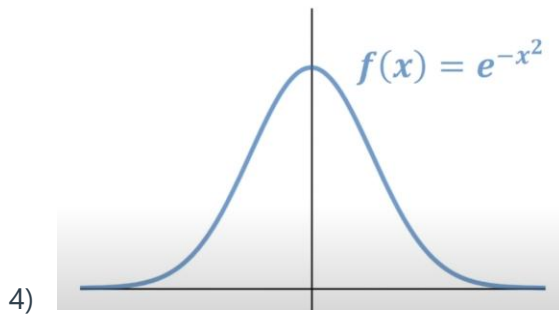


The formula gives the point on the graph line which will a probability.

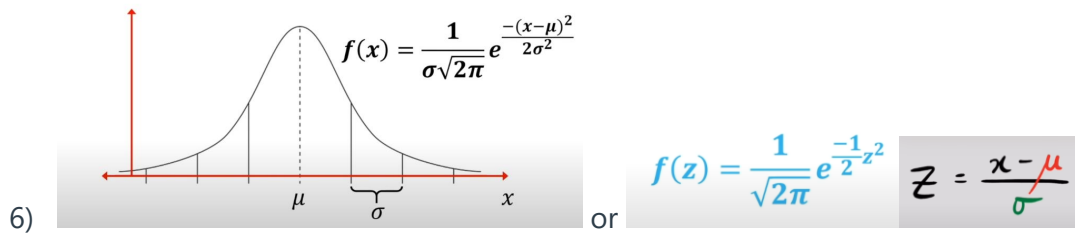
Deriving Normal PDF graph:







Probability Density Function Of The Normal Distribution



## Binomial Distribution

**Binomial Distribution** is used for the probability of another probability.

In the theory of probability and statistics, a Bernoulli trial (Binomial Trial) is a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment (independent events) is conducted.

Eg: Heads (0.5) or Tails (0.5) for coin flip

Eg: probability of heads or tails is 0.5.

Now if I flip a coin 10 times, I will get different combinations of heads and tails. Like 4H&6T or 3H&7T etc. If I keep repeating 10 flips million times then I could probably be able to see what proportions of the 10flips are 4H&6T or 3H&7T etc. Knowing these proportions leads us to calculation of the probability of each of these combinations.

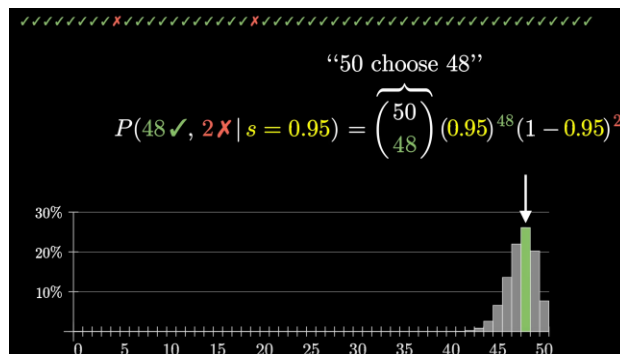
Eg: what is the probability of getting 4H&6T in 10 flips?

= (count no. times you got 4H&6T) / million

The Binomial mass function provides us just this without having to do the 10flips million times.

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Say probability of an event is 0.8. If the test is repeated, what is the probability of 0.8 will happen x out of n times.



The **Binomial Distribution** helps us determine the probability of a string of independent 'coin flip like events'.

### Formula break up:

Binomial distribution function= Combinations X probability of Binomial trials

In Binomial trials if P (one event) = x then P (other event) = n-x

Probability of independent event (A,A,B...) = P(A) \* P(A) \* P(B)...

Probability of heads(/tails) in one flip is 0.5 => P(H)=0.5

Eg: What is the prob. of getting Heads 2 times (or x times) in 3 flips of coin = aka 2 heads and 1 tail = combinations \* prob. of first \* prob of second \* prob of third...

### **Why Combination?**

2H can occur as HHT or HTH or THH and so each of these combination must be included and this can be done with combination formula  $nCr$

Eg of Combination:

Pick 2 from group of 4 (A,B,C,D)

AB(or BA), AC (or CA),AD,BC,BD,CD = 6 total

Instead of manually writing down all combinations, we can easily do with combination formula.

6 can also be obtained via formula  $4C2 = \frac{4!}{(2! * 2!)} = \frac{4 * 3 * 2}{2} = 6$

$$= nCx * P(H) * P(H) * P(T)...$$

$$= nCx * 0.5 * 0.5 * 0.5 * ...$$

Say in a biased coin  $P(H) = 0.8$  and so  $P(T) = 1 - 0.8 = 0.2$  above will look like

$$= nCx * 0.8 * 0.8 * 0.2$$

If you notice above 0.8 appears r times and 0.2 appears (n-r) times.

$$= nCx * 0.8^r * 0.2^{n-r}$$

Which gives the Binomial Distribution mass function:

$$= nCx * P(A)^x * P(B)^{n-x}$$

$$= \text{combinations} * P(\text{one event})^x * P(\text{other event})^{n-x}$$

Since it is binomial trials,  $P(\text{other event}) = 1 - P(\text{one event})$

$$= \text{combinations} * P(\text{one event})^x * P(1 - \text{one event})^{n-x}$$

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}$$

Probability of another probable event happening x out of n times = combinations \*  $P(\text{one event})^x * P(1 - \text{one event})^{n-x}$

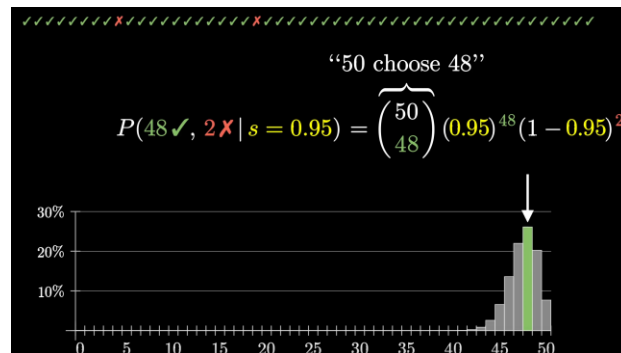
This can also be obtained by repeating the test million times and noting the results. The results plotted will give idea practically how many times we got the desired combination.

Eg:

Event: probability of 48 success out of 50 trials where probability of success is 0.95 for each trial.

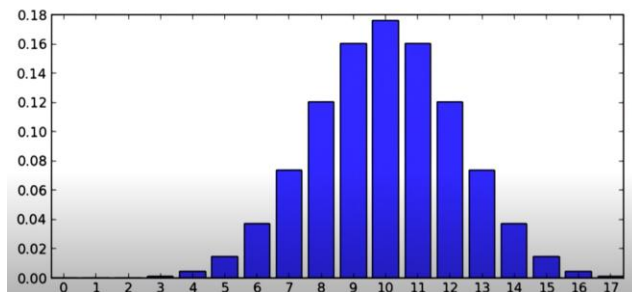
Below is a cumulative % distribution of say 10000 times, and we see that 48S and 2F occurs around 25%.

The binomial formula give this number (25%) without having to simulate the 10K experiments

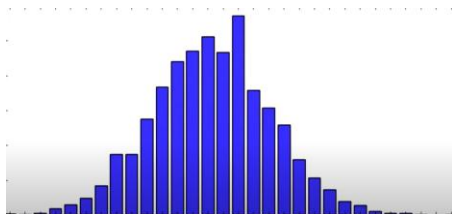


Example 2 of application of Binomial m. function.:

Coin is flipped 20 times what is the probability of 5H+15T or 7H+13T... If this is simulated the typical graph will look like a bell curve:



It can also be random like below but the idea is mean will eventually be like above



From this graph you can tell 5H+15T is 0.01 something or 1.x%.

7H+13T is 0.07 something or 7.x%.

This 1.x% or 7.x% can be found with the binomial m. function without having to simulate the trials.

Binomial to Normal Approximation:

Even the distribution of the random variable of Binomial type can be approximated to normal if  $n * p$  or  $n(1 - p) > 5$

### **Variance of Binomial Distribution:**

The mean of a binomial distribution with parameters

$N$  (the number of trials) and  $p$  (the probability of success for each trial) is  $m = Np$ .

The variance of the binomial distribution is  $s^2 = Np(1 - p)$ , where  $s^2$  is the variance of the binomial distribution.

$$\text{Mean}, \mu = np$$

$$\text{Variance}, \text{Var}(X) = np(1 - p)$$

### **Comparison**

	$P(A) = a/n$	Binomial	Normal(Gaussian)
Best for	Single	A few	Many (infinity)
	probability of the flip being heads?	probability of a given number of flips being heads?	probability of a given proportion of flips being heads?

### **Difference Binomial and Bernoulli:**

Bernoulli trial	Binomial
$n = 1$	$n = 500$ where $N = 100$ (100 coin throws done 500 times)
$P(X = k) = \binom{n}{k} p^k * (1 - p)^{n-k}$ <p>The probability mass function of binomial distribution . Bernoulli is when <math>n</math> is simply = 1</p>	

**Bernoulli distribution is simply a special case of the binomial distribution with  $N = 1$  such that  $Y$  (predicted value) is only 1 or 0 while in Binomial it can be anything between 0 and 1.**

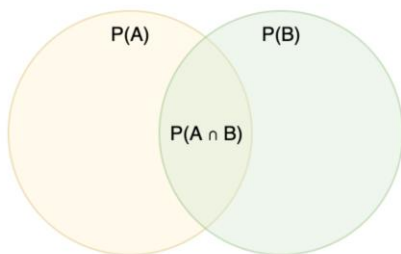
**Bernoulli deals with the outcome of the single trial of the event**, whereas Binomial deals with the outcome of the multiple trials of the single event. Bernoulli is used when the outcome of an event is required for only one time, whereas the Binomial is used when the outcome of an event is required multiple times.

## Conditional Probability

Probability of an event given another has occurred.

For example, the probability of obtaining a positive test result is actually dependent on whether or not you have a particular condition. If you have a condition, it is more likely that a test result is positive. We can formulate conditional probabilities for any two events in the following way:

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



In this case, we could have this as:

$$P(\text{positive} | \text{disease}) = P(\text{positive} \cap \text{disease}) / P(\text{disease})$$

This formula comes from  $P(A \cap B) = P(A) * P(A|B)$  (where A and B are dependent events).

Another way to understand is that, the basic probability formula is:

$$P() = \text{count or probability of an event} / \text{count or probability of all possible events}$$

$P(B)$  is all possible events for  $P(A | B)$  and  $P(A \cap B)$  are the events in which A and B are happening.

$$\text{Hence } P(A|B) = P(A \cap B) / P(B)$$

## Baye's Theorem:

App in ML: In Naive Bayes Classifier – refer ML Modelling: Naive Bayes Classifier

Moral of Baye's thm/ conditional prob formula: When given a conditional case, the probability formula is actually more complex than just  $n(A)/\text{total}$ .

$P(A B) = \frac{P(B A)P(A)}{P(B)}$	$= \frac{P(A \cap B)}{P(B)}$ or
Baye's thm	simple conditional prob. formula

<https://www.youtube.com/watch?v=XQoLVI31ZfQ>

Why used? Probability of a conditional event in which the P( reverse condition) is known

Popular Terms used in Bayes Thm Topic:

$$\boxed{P(A|B)}_{\text{posterior}} = \boxed{P(A)}_{\text{prior}} \times \frac{\boxed{P(B|A)}_{\text{likelihood}}}{\boxed{P(B)}_{\text{marginal}}}$$

Proof:

$$P(A|B) = P(A \cap B) / P(B) \quad \text{--(1)} \quad \text{and}$$

$$P(B|A) = P(B \cap A) / P(A) \quad \text{--(2)}$$

Now:  $P(A \cap B) = P(B \cap A)$  so switch it in (2) thus (2) rearranged becomes  $\Rightarrow$   
 $P(A \cap B) = P(B|A) * P(A)$  and now apply this to (1)

$$P(A|B) = P(A \cap B) / P(B) = P(B|A) * P(A) / P(B)$$

P(B) can be further expanded and is needed because it is more difficult to find.

$$P(B) = P(B \cap A) + P(B \cap \sim A)$$

Eg: B = +ve test result and A is having disease

$$P(+ve \text{ test}) = P(+ve \text{ test and have disease}) + P(+ve \text{ test and no disease})$$

$$\text{Lastly, } P(B \cap A) = P(B|A).P(A) \quad \& \quad P(B \cap \sim A) = P(B|\sim A).P(\sim A)$$

$$P(A|B) = P(B|A) * P(A) / P(B \cap A) + P(B \cap \sim A) \quad \text{or}$$

$$P(A|B) = P(B|A) * P(A) / [P(B|A).P(A) + P(B|\sim A).P(\sim A)]$$

### **Alternate explanation:**

<https://www.youtube.com/watch?v=HZGCoVF3YvM>

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

Basic prob:  $P(A) = (\text{total A}) / (\text{total possible events}) \rightarrow$

(total A) in above case is total number of A given B has happened

$$(\text{total A}) \rightarrow [\text{total count of A+B} * P(A) * P(B|A)]$$

(total possible events) is total number of A given B and total number of B who is not A.



(total possible events) → Above events + Events when case isn't true

→ [total count of A+B \* P(A) \* P(B|A)] + [total count of A+B \* P(~A) \* P(B | ~A) ]

simplified by cancelling total count of A+B from num and denom →

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

Although Baye's is expressed in shorter format, person using above formula will have to find all the values in above expression.

**Benefit of Baye's:** In real world, it might be easy to calculate the probability of one of the conditional prob and so using Baye's, the reverse conditional prob is found.

## Baye's in Python

$P(A|B) \Rightarrow (df.query('colB == condition1')[colA]==condition2).mean()$

$df.query('colB == condition1')$  : reduces to rows that meet B condition1

$[colA]==condition2$  : creates an index array based on fulfilling condition2

$( )$  : ensures that index array can be operated on by mean()

Eg:

A = have cancer

B = Test positive in cancer test

$P(\text{cancer}|\text{positive}) =$

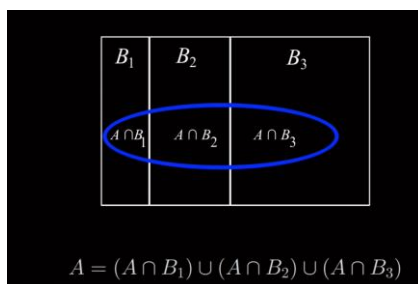
```
( df.query('test_result == "Positive"') ['has_cancer'] == True ).mean()
```

### Law of total probability

$B_1, B_2, B_3$  are mutually exclusive events.  $A$  is an event that is a subset of all 3. then:

$$P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + P(B_3) \cdot P(A|B_3)$$

$$= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$



<https://www.youtube.com/watch?v=7t9jyikrG7w>

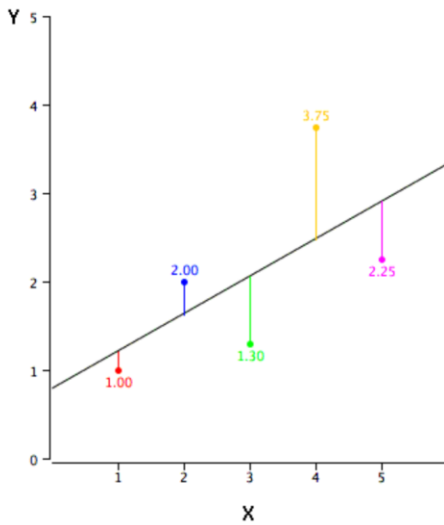
## Regression

Regression is all about predicting the value of one variable based on other variables.

**Correlation** quantifies the direction and strength of the relationship between two numeric variables,  $X$  and  $Y$ , and always lies between  $-1.0$  and  $1.0$ . **Simple linear regression** relates  $X$  to  $Y$  through an equation of the form  $Y = a + bX$ .

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as  $Y$ . The variable we are basing our predictions on is called the predictor variable and is referred to as  $X$ .

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is the line that minimizes the sum of the squared errors of prediction.



The error of prediction / residual for a point is the distance between the point and the best-fitting line.

$$y = mx + b$$

$$\text{slope} = r \cdot s_y / s_x$$

$$\text{intercept} = M_y - \text{slope} \cdot M_x \quad (M_y - \text{mean of } Y \text{ and } M_x - \text{mean of } x)$$

## Simple Linear Regression

**Simple linear regression** is a statistical method you can use to understand the relationship between two variables,  $x$  and  $y$ .

One variable,  $x$ , is known as the **predictor variable**.

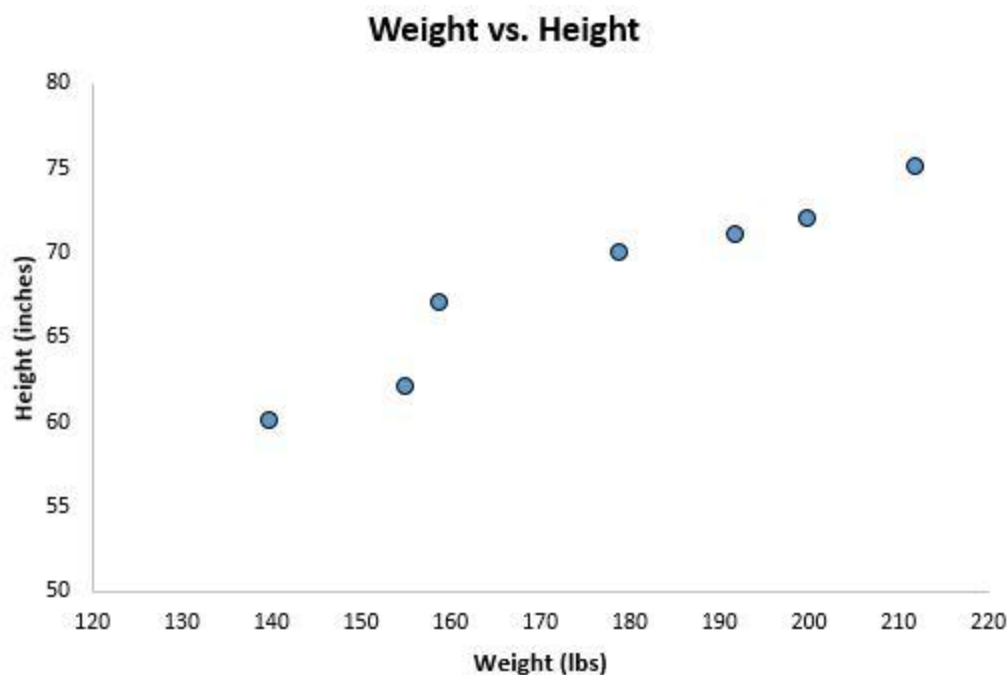
The other variable,  $y$ , is known as the **response variable**.

For example, suppose we have the following dataset with the weight and height of seven individuals:

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

Let *weight* be the predictor variable and let *height* be the response variable.

If we graph these two variables using a [scatterplot](#), with weight on the x-axis and height on the y-axis, here's what it would look like:



Suppose we're interested in understanding the relationship between weight and height. From the scatterplot we can clearly see that as weight increases, height tends to increase as well, but to actually *quantify* this relationship between weight and height, we need to use linear regression.

Using linear regression, we can find the line that best "fits" our data. This line is known as the **least squares regression line** and it can be used to help us

understand the relationships between weight and height. Usually you would use software like Microsoft Excel, SPSS, or a graphing calculator to actually find the equation for this line.

The formula for the line of best fit is written as:

$$\hat{y} = b_0 + b_1x$$

$$y = \text{intercept} + \text{slope} \cdot x$$

where  $\hat{y}$  is the predicted value of the response variable,  $b_0$  is the y-intercept,  $b_1$  is the regression coefficient, and  $x$  is the value of the predictor variable.

## Finding the “Line of Best Fit”

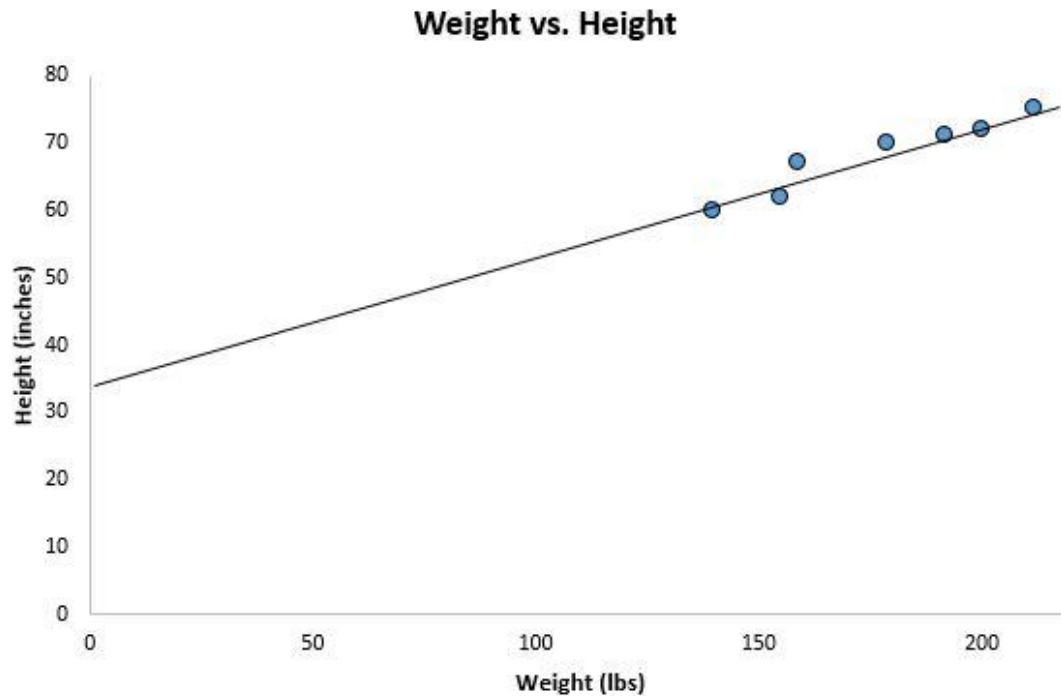
slope =  $r \cdot s_y / s_x$  ( $r$  correlation coeff between  $x$  and  $y$ ,  $s_y$  – std. dev for response vars, and  $s_x$  – std. dev for predictor vars)

intercept =  $M_y - \text{slope} \cdot M_x$  ( $M_y$  – mean of  $Y$  and  $M_x$  – mean of  $x$ )

The calculator automatically finds the **least squares regression line**:

$$\hat{y} = 32.7830 + 0.2001x$$

If we zoom out on our scatterplot from earlier and added this line to the chart, here’s what it would look like:



Notice how our data points are scattered closely around this line. That's because this least squares regression line is the best fitting line for our data out of all the possible lines we could draw.

## How to Interpret a Least Squares Regression Line

Here is how to interpret this least squares regression line:  $\hat{y} = 32.7830 + 0.2001x$

**$b_0 = 32.7830$** . This means when the predictor variable *weight* is zero pounds, the predicted height is 32.7830 inches. Sometimes the value for  $b_0$  can be useful to know, but in this specific example it doesn't actually make sense to interpret  $b_0$  since a person can't weight zero pounds.

**$b_1 = 0.2001$** . This means that a one unit increase in  $x$  is associated with a 0.2001 unit increase in  $y$ . In this case, a one pound increase in weight is associated with a 0.2001 inch increase in height.

## How to Use the Least Squares Regression Line

Using this least squares regression line, we can answer questions like:

*For a person who weighs 170 pounds, how tall would we expect them to be?*

To answer this, we can simply plug in 170 into our regression line for  $x$  and solve for  $y$ :

$$\hat{y} = 32.7830 + 0.2001(170) = \mathbf{66.8 \text{ inches}}$$

*For a person who weighs 150 pounds, how tall would we expect them to be?*

To answer this, we can plug in 150 into our regression line for  $x$  and solve for  $y$ :

$$\hat{y} = 32.7830 + 0.2001(150) = \mathbf{62.798 \text{ inches}}$$

**Caution:** *When using a regression equation to answer questions like these, make sure you only use values for the predictor variable that are within the range of the predictor variable in the original dataset we used to generate the least squares regression line. For example, the weights in our dataset ranged from 140 lbs to 212 lbs, so it only makes sense to answer questions about predicted height when the weight is between 140 lbs and 212 lbs.*

## **The Coefficient of Determination**

One way to measure how well the least squares regression line “fits” the data is using the **coefficient of determination**, denoted as  $R^2$ .

The coefficient of determination is the proportion of the variance in the response variable that can be explained by the predictor variable.

The coefficient of determination can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

An  $R^2$  between 0 and 1 indicates just how well the response variable can be explained by the predictor variable. For example, an  $R^2$  of 0.2 indicates that 20% of the variance in the response variable can be explained by the predictor variable; an  $R^2$  of 0.77 indicates that 77% of the variance in the response variable can be explained by the predictor variable.

Notice in our output from earlier we got an  $R^2$  of 0.9311, which indicates that 93.11% of the variability in height can be explained by the predictor variable of weight:

CALCULATE

Linear Regression Equation:

$$\hat{y} = 32.7830 + (0.2001) \cdot x$$

Goodness of Fit:

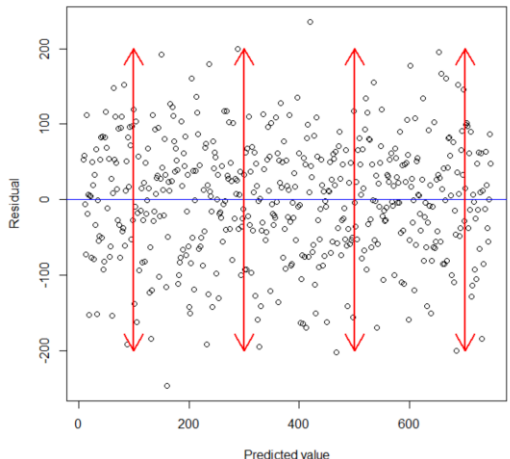
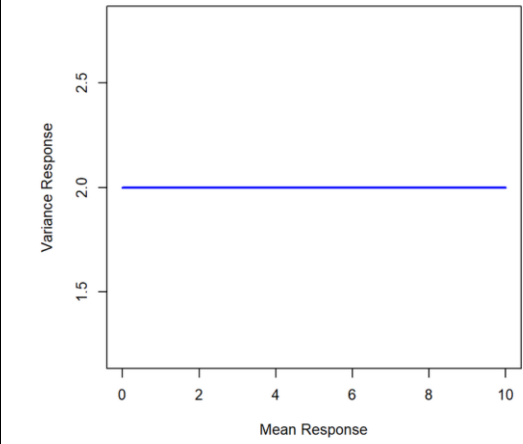
R Square: 0.9311

This tells us that weight is a very good predictor of height.

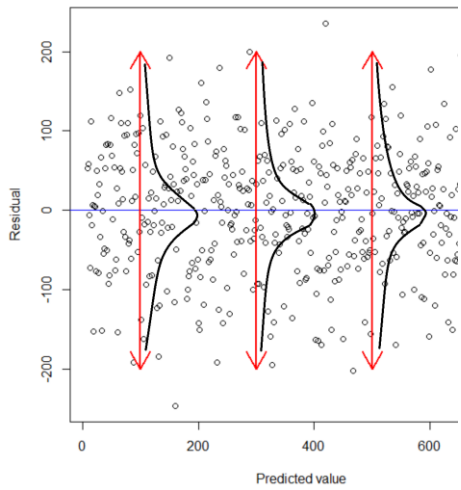
## Assumptions of Linear Regression

Few assumptions used in Linear Model:

- The distribution of predictors is normal
- Residuals are normally distributed and variance of residual is same

Constant variance (homoscedasticity)		
 <p>here predicted value is the linear regression point on line</p>	or	<p>Mean alone</p> 



Normally distributed residuals		
 <p>denser close to the mean and lesser away from mean</p>		

For the results of a linear regression model to be valid and reliable, we need to Linear regression is a useful statistical method we can use to understand the relationship between two variables,  $x$  and  $y$ . However, before we conduct linear regression, we must first make sure that four assumptions are met:

1. **Linear relationship:** There exists a linear relationship between the independent variable,  $x$ , and the dependent variable,  $y$ .
2. **Independence:** The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
3. **Homoscedasticity:** The residuals have constant variance at every level of  $x$ .
4. **Normality:** The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

### Assumption 1: Linear Relationship

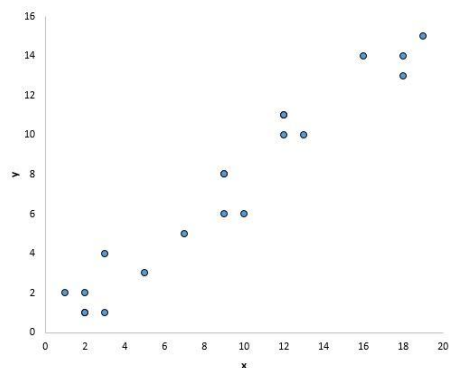
## Explanation

The first assumption of linear regression is that there is a linear relationship between the independent variable,  $x$ , and the dependent variable,  $y$ .

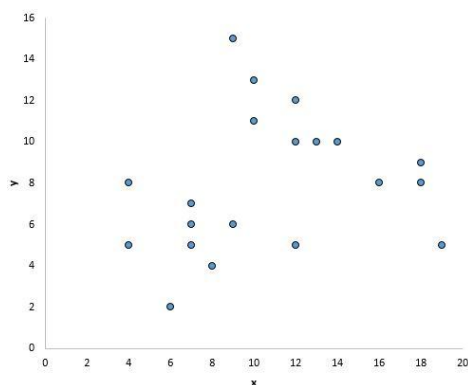
How to determine if this assumption is met

The easiest way to detect if this assumption is met is to create a scatter plot of  $x$  vs.  $y$ . This allows you to visually see if there is a linear relationship between the two variables. If it looks like the points in the plot could fall along a straight line, then there exists some type of linear relationship between the two variables and this assumption is met.

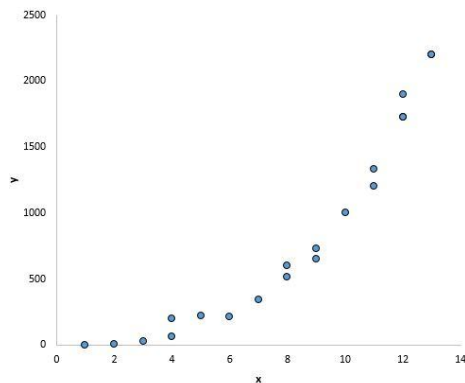
For example, the points in the plot below look like they fall on roughly a straight line, which indicates that there is a linear relationship between  $x$  and  $y$ :



However, there doesn't appear to be a linear relationship between  $x$  and  $y$  in the plot below:



And in this plot there appears to be a clear relationship between  $x$  and  $y$ , *but not a linear relationship*:



### What to do if this assumption is violated

If you create a scatter plot of values for  $x$  and  $y$  and see that there is *not* a linear relationship between the two variables, then you have a couple options:

1. Apply a nonlinear transformation to the independent and/or dependent variable. Common examples include taking the log, the square root, or the reciprocal of the independent and/or dependent variable.
2. Add another independent variable to the model. For example, if the plot of  $x$  vs.  $y$  has a parabolic shape then it might make sense to add  $X^2$  as an additional independent variable in the model.

### Assumption 2: Independence

#### Explanation

The next assumption of linear regression is that the residuals are independent. This is mostly relevant when working with time series data. Ideally, we don't want there to be a pattern among consecutive residuals. For example, residuals shouldn't steadily grow larger as time goes on.

How to determine if this assumption is met

The simplest way to test if this assumption is met is to look at a residual time series plot, which is a plot of residuals vs. time. Ideally, most of the residual autocorrelations should fall within the 95% confidence bands around zero, which are located at about  $\pm 2/\sqrt{n}$ , where  $n$  is the sample size. You can also formally test if this assumption is met using the [Durbin-Watson test](#).

What to do if this assumption is violated

Depending on the nature of the way this assumption is violated, you have a few options:

For positive serial correlation, consider adding lags of the dependent and/or independent variable to the model.

For negative serial correlation, check to make sure that none of your variables are *overdifferenced*.

- For seasonal correlation, consider adding seasonal dummy variables to the model.

### **Assumption 3: Homoscedasticity**

#### **Explanation**

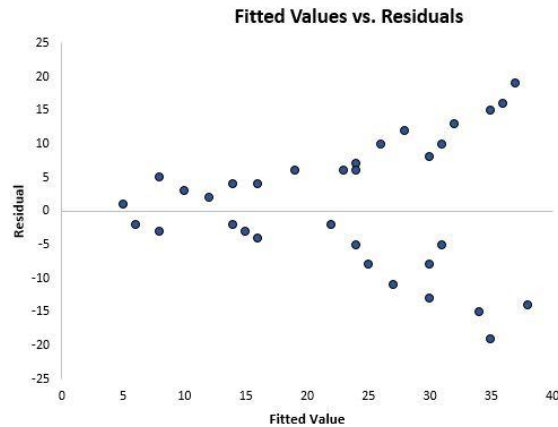
The next assumption of linear regression is that the residuals have constant variance at every level of  $x$ . This is known as *homoscedasticity*. When this is not the case, the residuals are said to suffer from *heteroscedasticity*.

When heteroscedasticity is present in a regression analysis, the results of the analysis become hard to trust. Specifically, heteroscedasticity increases the variance of the regression coefficient estimates, but the regression model doesn't pick up on this. This makes it much more likely for a regression model to declare that a term in the model is statistically significant, when in fact it is not.

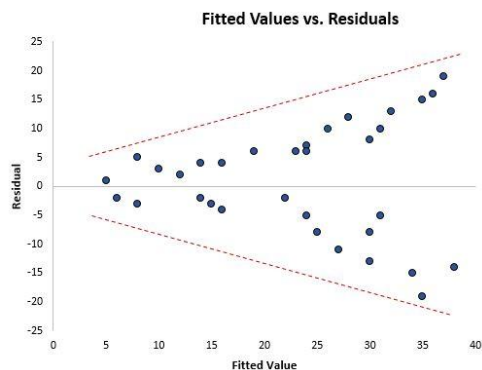
How to determine if this assumption is met

The simplest way to detect heteroscedasticity is by creating a *fitted value vs. residual plot*.

Once you fit a regression line to a set of data, you can then create a scatterplot that shows the fitted values of the model vs. the residuals of those fitted values. The scatterplot below shows a typical *fitted value vs. residual plot* in which heteroscedasticity is present.



Notice how the residuals become much more spread out as the fitted values get larger. This “cone” shape is a classic sign of heteroscedasticity:



## What to do if this assumption is violated

There are three common ways to fix heteroscedasticity:

1. Transform the dependent variable. One common transformation is to simply take the log of the dependent variable. For example, if we are using population size (independent variable) to predict the number of flower shops in a city (dependent variable), we may instead try to use population size to predict the log of the number of flower shops in a city. Using the log of the dependent variable, rather than the original dependent variable, often causes heteroskedasticity to go away.

2. Redefine the dependent variable. One common way to redefine the dependent variable is to use a *rate*, rather than the raw value. For example, instead of using the population size to predict the number of flower shops in a city, we may instead use population size to predict the number of flower shops per capita. In most cases, this reduces the variability that naturally occurs among larger populations since we're

measuring the number of flower shops per person, rather than the sheer amount of flower shops.

3. Use weighted regression. Another way to fix heteroscedasticity is to use weighted regression. This type of regression assigns a weight to each data point based on the variance of its fitted value. Essentially, this gives small weights to data points that have higher variances, which shrinks their squared residuals. When the proper weights are used, this can eliminate the problem of heteroscedasticity.

### **Assumption 4: Normality**

#### **Explanation**

The next assumption of linear regression is that the residuals are normally distributed.

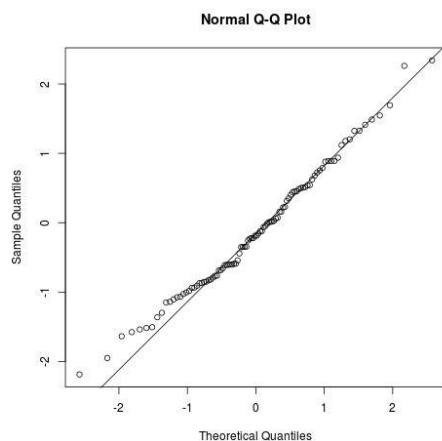
#### **How to determine if this assumption is met**

There are two common ways to check if this assumption is met:

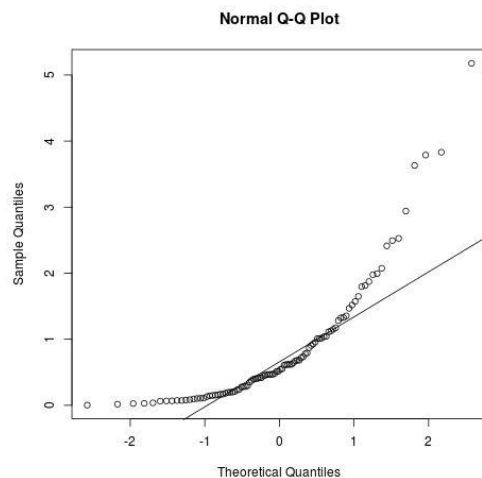
1. Check the assumption visually using [Q-Q plots](#).

A Q-Q plot, short for quantile-quantile plot, is a type of plot that we can use to determine whether or not the residuals of a model follow a normal distribution. If the points on the plot roughly form a straight diagonal line, then the normality assumption is met.

The following Q-Q plot shows an example of residuals that roughly follow a normal distribution:



However, the Q-Q plot below shows an example of when the residuals clearly depart from a straight diagonal line, which indicates that they do not follow normal distribution:



2. You can also check the normality assumption using formal statistical tests like Shapiro-Wilk, Kolmogorov-Smirnov, Jarque-Barre, or D’Agostino-Pearson. However, keep in mind that these tests are sensitive to large sample sizes – that is, they often conclude that the residuals are not normal when your sample size is large. This is why it’s often easier to just use graphical methods like a Q-Q plot to check this assumption.

### What to do if this assumption is violated:

If the normality assumption is violated, you have a few options:

- First, verify that any outliers aren’t having a huge impact on the distribution. If there are outliers present, make sure that they are real values and that they aren’t data entry errors.
- Next, you can apply a nonlinear transformation to the independent and/or dependent variable. Common examples include taking the log, the square root, or the reciprocal of the independent and/or dependent variable.

### Difference between Correlation and Regression:

Correlation	Regression
<b>Key similarities</b>	
Both quantify the direction and strength of the relationship between two numeric variables.	
When the correlation ( $r$ ) is negative, the regression slope ( $b$ ) will be negative.	
When the correlation is positive, the regression slope will be positive.	

<b>Key Differences:</b>	
X and Y variables are interchangeable.	Regression attempts to establish how X causes Y to change and the results of the analysis will change if X and Y are swapped.
X and Y are typically both random variables*, such as height and weight or blood pressure and heart rate.	assumes X is fixed with no error, such as a dose amount or temperature setting.
single statistic	produces an entire equation
Correlation is a more concise (single value) summary of the relationship between two variables than regression. In result, many pairwise correlations can be viewed together at the same time in one table.	Regression provides a more detailed analysis which includes an equation which can be used for prediction and/or optimization.

## Doing Stats/Prob in Python

### Simulating Coin flips and Dice rolls in Python

#### *Numpy.Random*

Numpy.random.randint

Used to simulate coin flips and dice rolls:

#### Simulating Coin flips

```
# simulate 1 million tests of a fair coin flips
# 0 is heads and 1 is tails
tests = np.random.randint(2, size=(int(1e6)))

#2 – indicates end of integer range, starting point 0 is skipped because 0 is default else it will be like randint(1,11,...)
```

Out:array([0, 0, 1, ..., 1, 0, 1])

```
# proportion of tests that produced heads
# statement below produces a index array with 1s when true and 0s when false
# mean is count of all 1s and divided by count of all constituents which gives the proportion of 1s or probability of an event coded 1
# sum gives the count of event(1)
(tests == 0).mean()
```



```
# simulate 1 million tests of a fair coin flips  
tests = np.random.randint(2, size=(int(1e6)))
```

```
(tests == 0).mean()
```

<pre>np.random.randint(2,size=(int(1e6),3))</pre>	<pre>array([[0, 1, 1],        [1, 1, 1],        [0, 1, 1],        ...,        [0, 1, 0],        [1, 1, 0],        [0, 1, 1]])</pre>
---	---

```
# simulate 1 million tests of three fair coin flips  
tests = np.random.randint(2,size=(int(1e6),3))
```

```
# sums of all tests  
test_sums = tests.sum(axis=1)
```

```
array([1, 3, 2, ..., 2, 0, 0])
```

```
# proportion of tests that produced exactly one head  
(test_sums == 1).mean()
```

```
# A die rolls an even number: simulate 1 million tests of one die roll  
tests = np.random.randint(1,7,size=(int(1e6)))
```

```
# proportion of tests that produced an even number  
(tests%2 == 0).mean()
```

```
# Two dice roll a double: simulate the first million die rolls
```

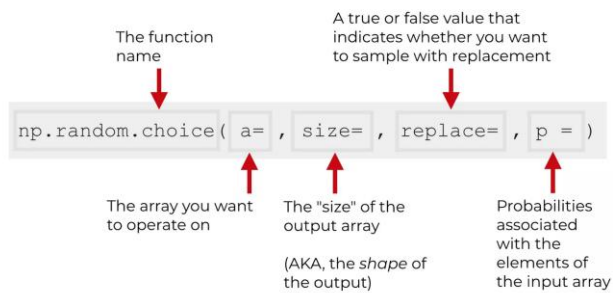
```
first = np.random.randint(1,7,size=(int(1e6)))
```

```
# simulate the second million die rolls
```

```
second = np.random.randint(1,7,size=(int(1e6)))
```

```
# proportion of tests where the 1st and 2nd die rolled the same number  
(first == second).mean()
```

## Np.random.choice



## <PDDataframe>.sample

```
<dataframe>.sample(size=)
```

## Simulating Binomials: random.binomial

Binomial:

2 events; heads or tails; with a probability like 0.5 or other

Flip coin 10 times and how many times is heads (can be tails also; binomial so results reversible)?

`np.random.binomial(10, 0.5)` tells how many heads in 10 flips

```
# number of heads from 10 fair coin flips
np.random.binomial(10, 0.5)
5
```

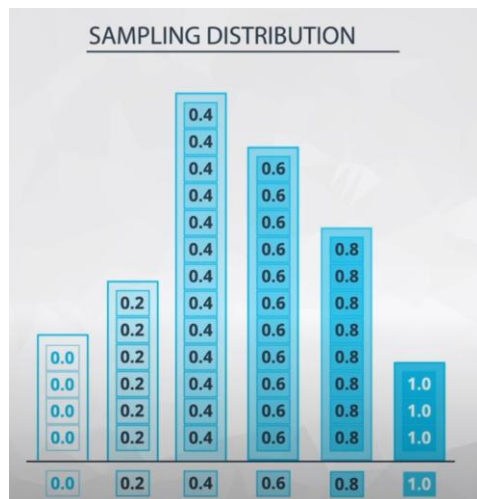
to repeat the test 20 times:

```
# results from 20 tests with 10 coin flips
np.random.binomial(10, 0.5, 20)
array([7, 7, 5, 6, 5, 6, 4, 8, 5, 4, 7, 5, 5, 9, 8, 7, 7, 5, 5, 4])
```

get mean to see the avg of no. of heads:

```
# mean number of heads from the 20 tests
np.random.binomial(10, 0.5, 20).mean()
4.8499999999999996
```

## Sampling Distribution



## Simulating sampling distribution

Numpy:

```
np.random.choice(<array>, <sample size>).mean() for i in range(10000)
```

or

```
sampling_dist = []
for i in range (10000):
    sample = np.random.choice(<array>, size=)
    sampling_dist.append(<array>.mean())
```

Pandas:

```
sampling_dist = []
for i in range (10000):
    sample = dataframe.sample(size=)
    sampling_dist.append(dataframe[dataframe index array]== condition][<column>.mean())
```

## Confidence intervals

```
np.percentile(<array>, 2.5), np.percentile(<array>, 97.5)
```

OUT: (start, end)

## Simulating Null Hypothesis testing

1. Bootstrap the values being studied
2. Find standard dev.
3. Use `x = np.random.normal (<mean>, <std dev>, #no. of draws)`
4. Plot x to see if it is normally distributed and the limits

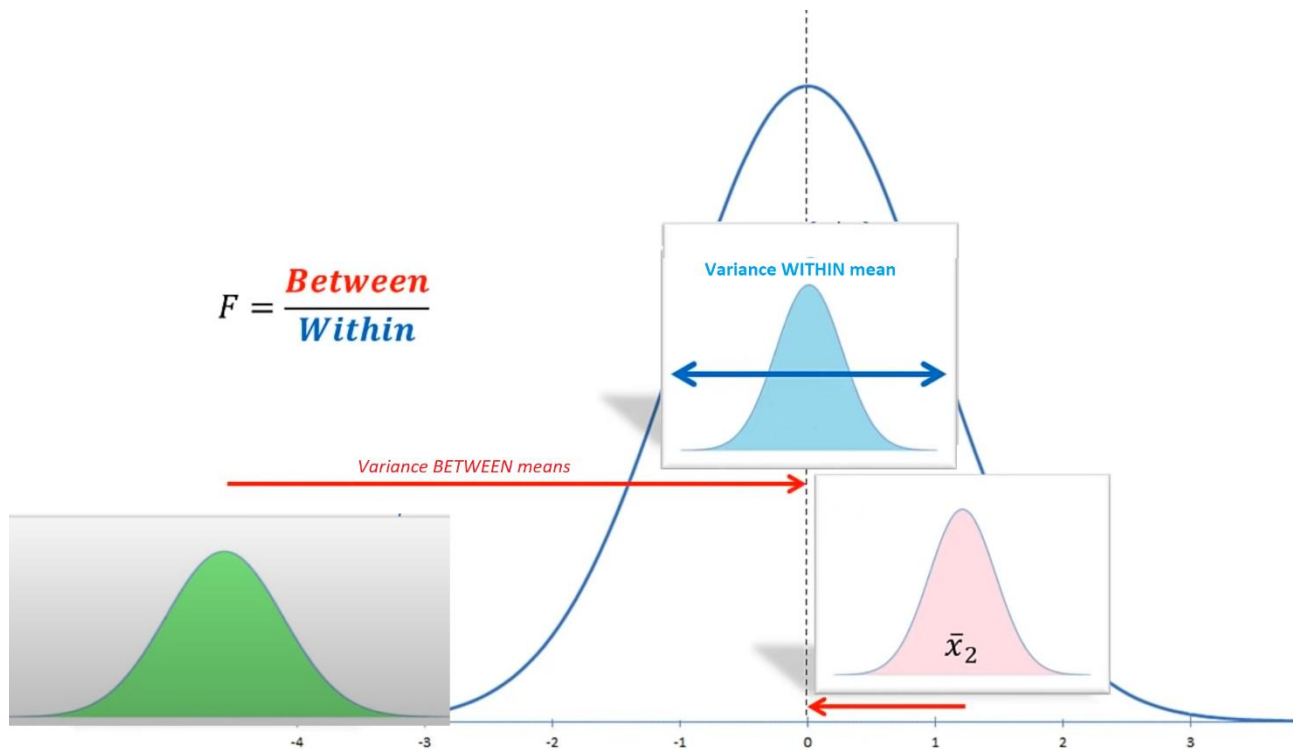
## ANOVA:

So far we have been comparing only two population means, Independent samples (t test) or (correlated) paired t test.

<https://www.youtube.com/watch?v=0Vj2V2qRU10>

Why ANOVA?

- Compare multiple populations and wven groups within groups which is not possible with earlier methods.
- Besides, if we have multiple sample distributions, ANOVA allows to check if the means are from the same population or it is so wide apart that it is not.



$$\text{Variance Between} + \text{Variance Within} = \text{Total Variance}$$

Variance within AKA Error variance

## ANOVA: Analysis of Variance is a *variability ratio*

$$\frac{\text{Variance Between}}{\text{Variance Within}} \left. \vphantom{\frac{\text{Variance Between}}{\text{Variance Within}}} \right\} \text{Total Variance Components}$$

$$\text{Variance Between} + \text{Variance Within} = \text{Total Variance}$$

“Partitioning” – separating total variance into its component parts

If the variability **BETWEEN** the means (distance from overall mean) in the numerator is relatively large compared to the variance **WITHIN** the samples (internal spread) in the denominator, the ratio will be much larger than 1. The samples then most likely do NOT come from a common population; **REJECT NULL HYPOTHESIS** that means are equal.

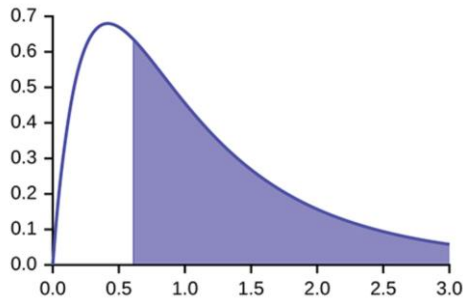
	$\frac{\text{LARGE}}{\text{small}} = \text{Reject } H_0$	At least one mean is an outlier and each distribution is narrow; distinct from each other.
$\frac{\text{Variance Between}}{\text{Variance Within}}$	$\frac{\text{similar}}{\text{similar}} = \text{Fail to Reject } H_0$	Means are fairly close to overall mean and/or distributions overlap a bit; hard to distinguish.
	$\frac{\text{small}}{\text{LARGE}} = \text{Fail to Reject } H_0$	The means are very close to overall mean and/or distributions “melt” together.

F test is performed

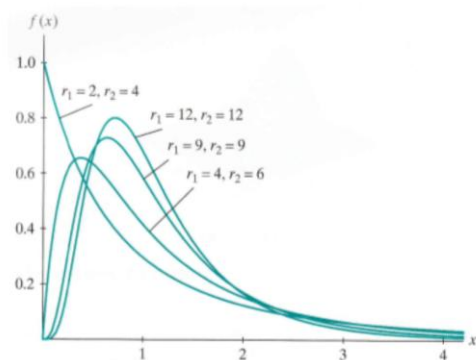
$$F = \frac{\text{Between}}{\text{Within}} \quad \text{or AKA} \quad F = \frac{\text{Among}}{\text{Around}}$$

Facts about the F distribution.

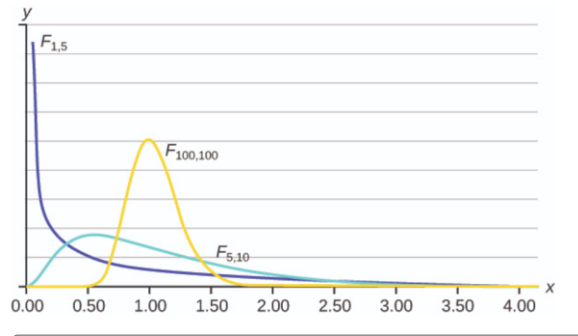
1. The curve is not symmetrical but skewed to the right.



2. There is a different curve for each set of degree of freedom for numerator,  $r_1$  and degree of freedom for denominator  $r_2$



3. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.



4. The F statistic is greater than or equal to zero.
5. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.

Why not paired t test?

paired t test is not performed since with each mean being compared, the error rate compounds as below reducing the confidence level.

**error COMPOUNDS with each t-test:**

Eg: three means  $(.95)(.95)(.95) = .857$

$$\alpha = 1 - 0.857 = 0.143 \text{ or only } 14.3\%$$

Tukey HSD?

ANOVA tests the non-specific null hypothesis that all four population means are equal. That is,

$$\mu_{\text{false}} = \mu_{\text{felt}} = \mu_{\text{miserable}} = \mu_{\text{neutral}}.$$

This non-specific null hypothesis is sometimes called the *omnibus null hypothesis*. When the omnibus null hypothesis is rejected, the conclusion is that at least one population mean is different from at least one other mean. However, since the ANOVA does not reveal which means are different from which, it offers less specific information than the [Tukey HSD test](#). The Tukey HSD is therefore preferable to ANOVA in this situation. Some textbooks introduce the Tukey test only as a follow-up to an ANOVA. However, there is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA. But all said, complex

types of analyses that can be done with ANOVA and not with the Tukey test. Besides, ANOVA is more popularly used in general.

### GLM eqn for ANOVA:

$$Y = B_0 + B_1 \text{GROUP}_1 + B_2 \text{GROUP}_2 + E$$

(ANOVA)

### One Way and Two way ANOVA

An ANOVA conducted on a design in which there is only one variable/factor is called a one-way ANOVA. If an experiment has two variables/factors, then the ANOVA is called a two-way ANOVA.

## Which Test to use?

