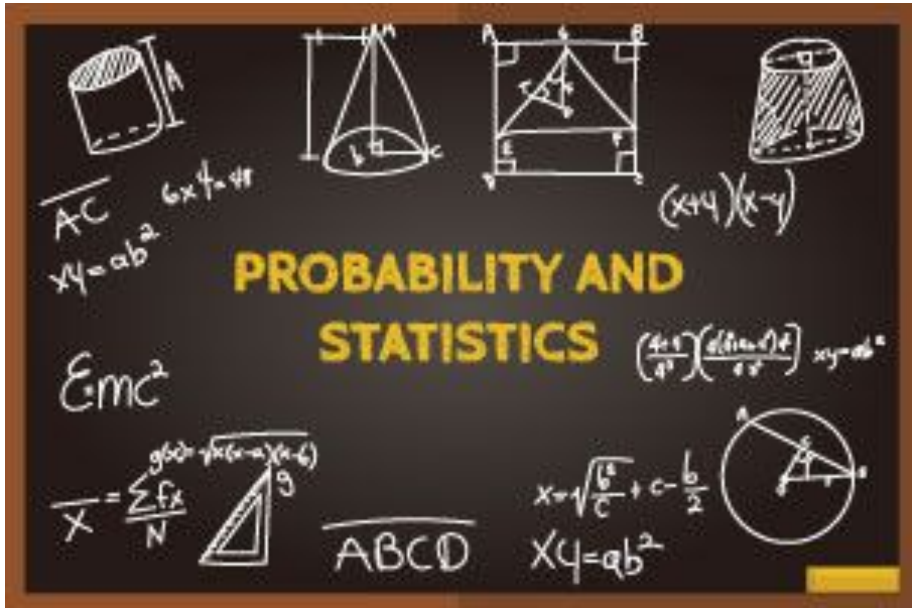


Probability, Stats for ML & A.I

Index

- DESCRIPTIVE STATS
- PROBABILITY
- INFERENTIAL STATS
- HYPOTHESIS TESTING

Descriptive Statistics - Central Tendency

- **Mean** : The average
  - Add all number and divide by their count
- **Median**: The middle value
  - Order the numbers and find the middle value
  - If the count is even, find average of the two middle values
- **Mode**: The most occurring value
  - The value that occurs most
  - Usage depends on situation

mean(), min(), max(), median

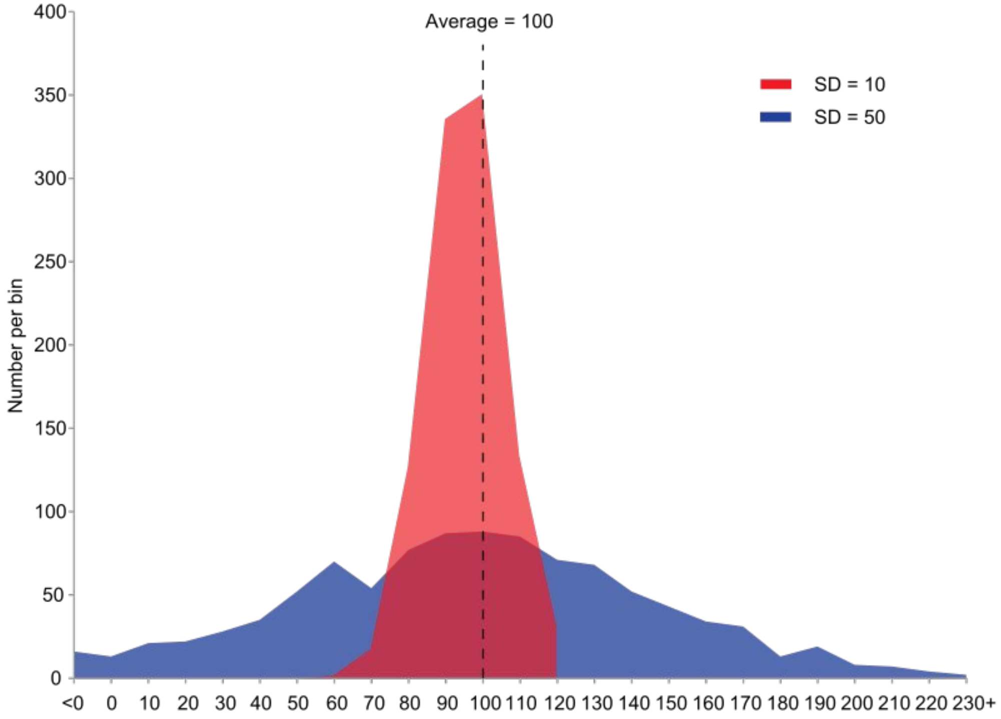
Example-

- Observations : 1, 3, 4, 5, 5, 7, 8, 9, 9, 9
- Count: 10
- Sum: 60
- Mean : Sum / Count = 60/10 = 6
- Median : Middle Value = (5 + 7) / 2 = 6
- Mode: 9

$$\bar{x} = (\Sigma x_i) / n$$

Dispersion/variation

- In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed)- Diff. is high between max and mean
- Example of samples from two populations with the same mean but different dispersion. The blue population is much more dispersed than the red population.



Mean Deviation

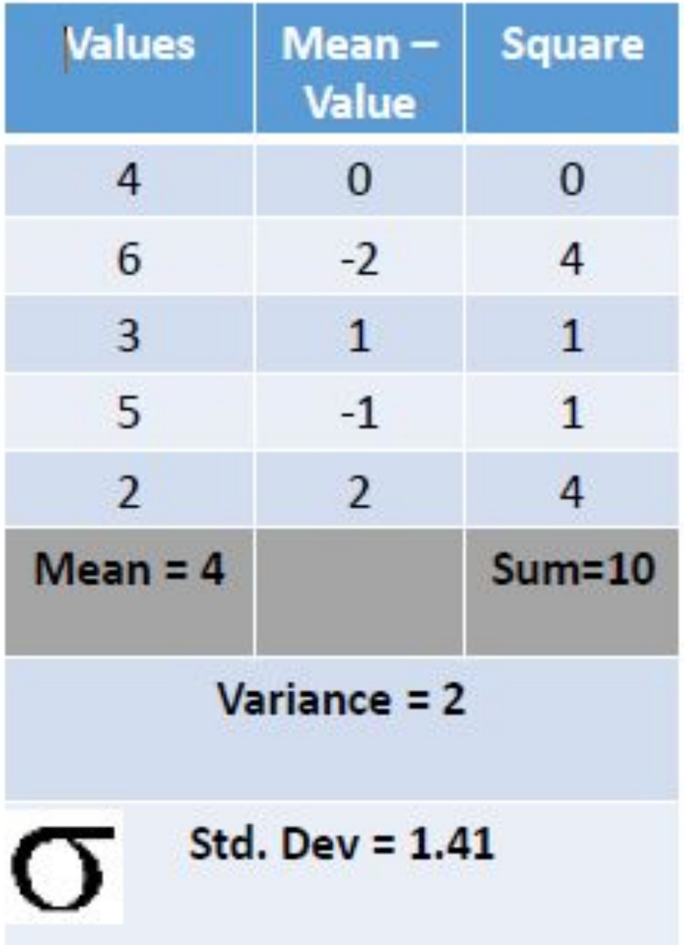
- Mean deviation is defined as average of the sum of the absolute values of deviation from any arbitrary value viz. mean, median, mode, etc. It is often suggested to calculate it from the median because it gives least value when measured from the median.

The deviation of an observation xi from the assumed mean A is defined as

- (xi – A).
- Therefore, the mean deviation can be defined as
- Summation(|xi - A|)/n

Variance

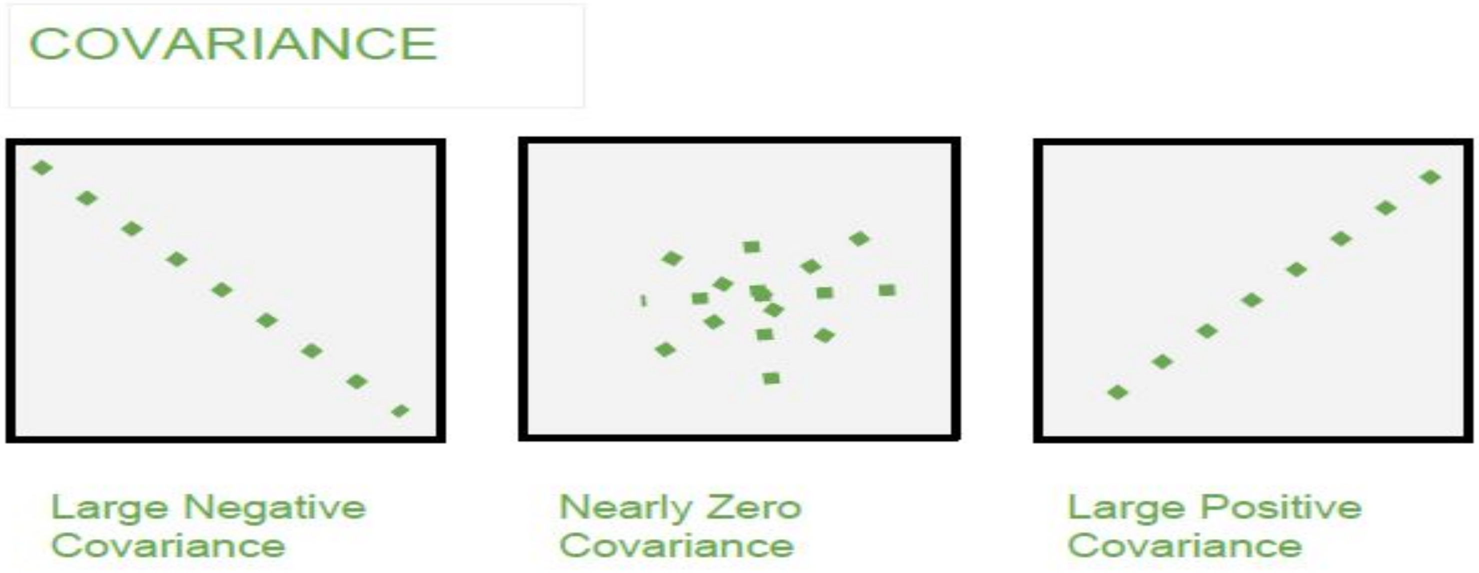
- Describes how values are distributed around the mean
  - If most values are closer to mean, low variance
  - If significant differences in values, then high variance
- To compute
  - Square the differences from the mean
  - Sum of Squares
  - Divide by count
  - Find the mean
- Standard Deviation is Square Root of variance
- variance is .. $\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$ s.d is .. $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$



Values	Mean – Value	Square
4	0	0
6	-2	4
3	1	1
5	-1	1
2	2	4
Mean = 4		Sum=10
Variance = 2		
σ	Std. Dev = 1.41	

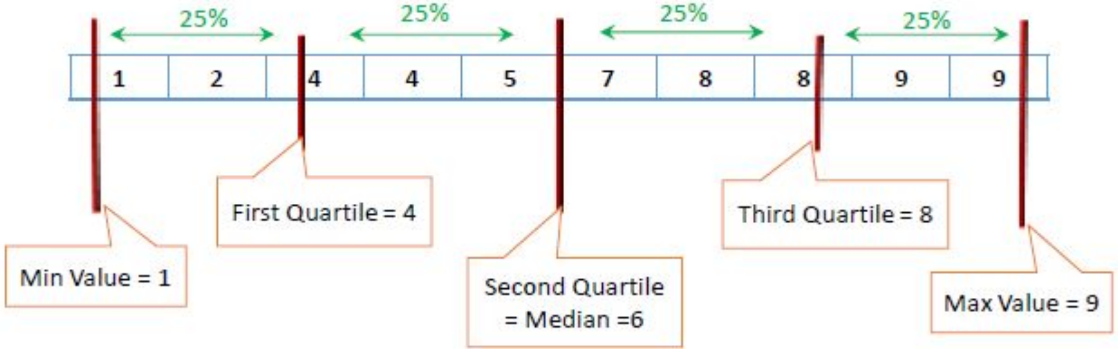
Covariance

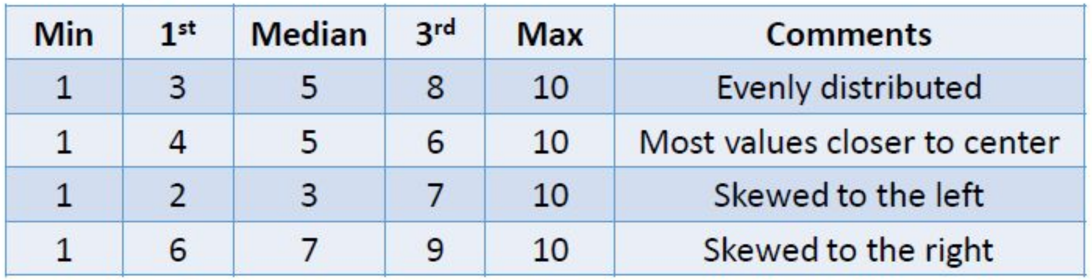
Covariance measures the directional relationship between the two variables. A positive covariance means that they move together while a negative covariance means they move inversely.



Range & Quartiles

- Range of a set of data is the difference between the largest and smallest values
- Describes the central tendency, distribution, range and skew in one set of measures
- Interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentile.
- Given a set of observations, we divide them into 4 equal sets.
- The boundaries form the quartiles-

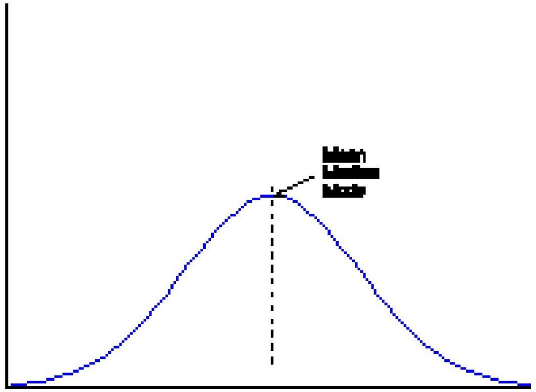




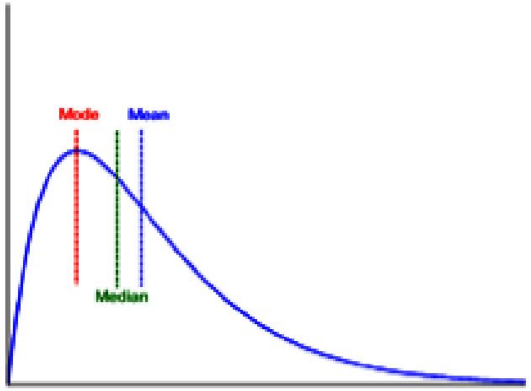
Min	1st	Median	3rd	Max	Comments
1	3	5	8	10	Evenly distributed
1	4	5	6	10	Most values closer to center
1	2	3	7	10	Skewed to the left
1	6	7	9	10	Skewed to the right

Symmetric and skewed data

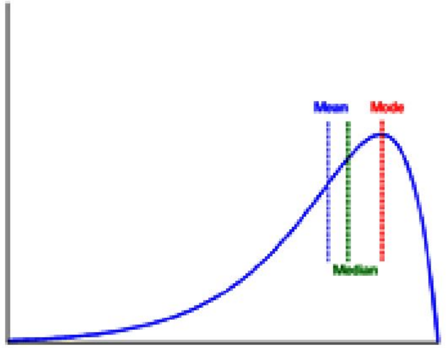
- Median, mean and mode of symmetric, positively and negatively skewed data can be represented as follows -**positively skewed distribution for which mean > median > mode** while if the **left tail is longer, we get a negatively skewed distribution for which mean < median < mode.**



Symmetric



Positively Skewed



Negatively Skewed

Pearson’s Coefficient of Skewness #1 uses the mode. The formula is:

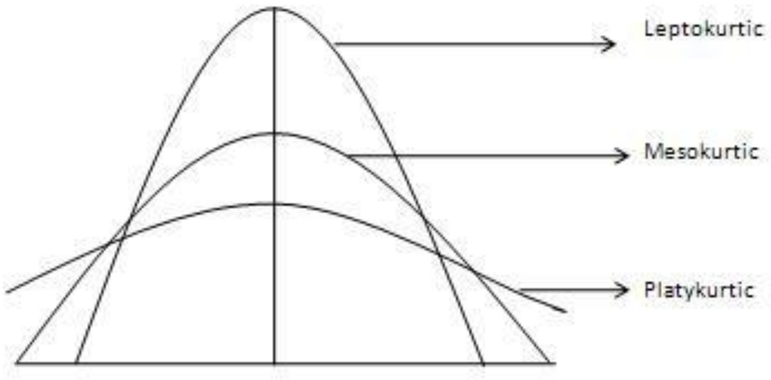
$$Sk_1 = \frac{\bar{X} - Mo}{s}$$

Where $\bar{X}$ = the mean, Mo = the mode and s = the standard deviation for the sample.

See: Pearson Mode Skewness.

Kurtosis

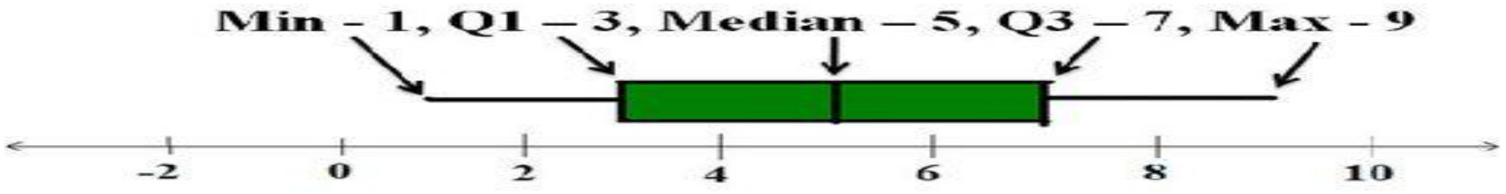
The degree of tailedness of a distribution is measured by kurtosis. It tells us the extent to which the distribution is **more or less outlier-prone** (heavier or light-tailed) than the normal distribution. Three different types of curves, are shown as follows –



The kurtosis of any univariate normal distribution is 3. It is common to compare the kurtosis of a distribution to this value. Distributions with kurtosis less than 3 are said to be *platykurtic*, although this does not imply the distribution is "flat-topped" as sometimes reported. Rather, it means the distribution produces fewer and less extreme outliers than does the normal distribution.

Five Number Summary

The **five-number summary** of a data set consists of the **five numbers** determined by computing the minimum, Q_1 , median, Q_3 , and maximum of the data set.



Probability

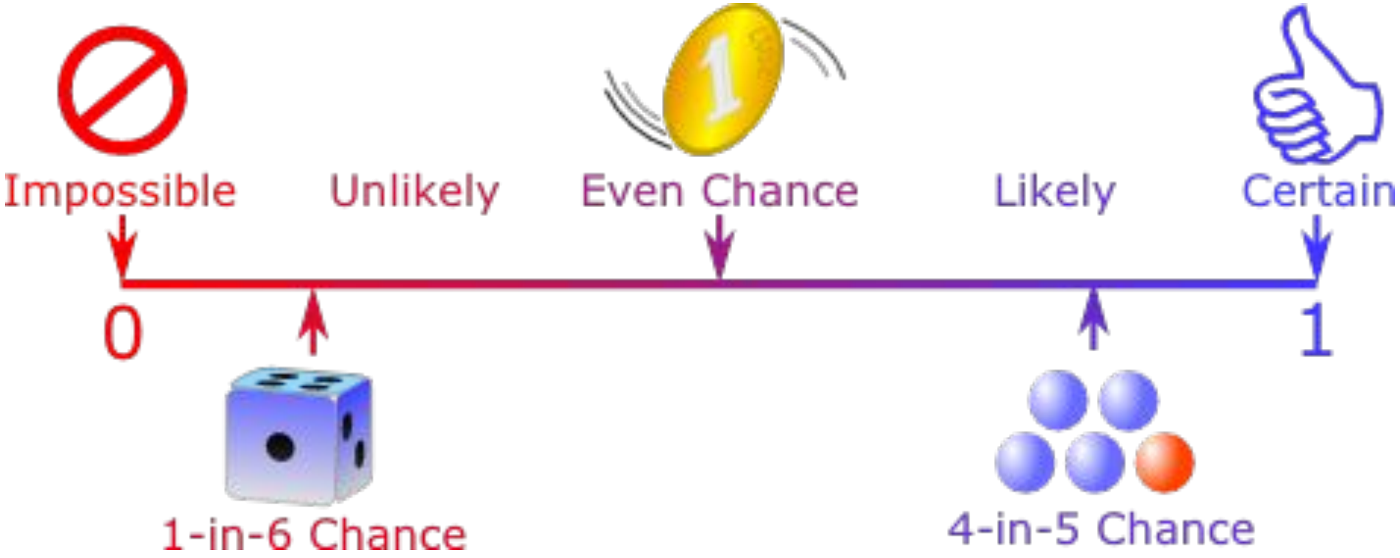
*How **likely** something is to happen.*

Many events can't be predicted with total certainty. The best we can say is how **likely** they are to happen, using the idea of probability.

Throwing Dice

When a single die is thrown, there are six possible outcomes: **1, 2, 3, 4, 5, 6**. The probability of any one of them is ⅙.

Probability of an event happening = Number of ways it can happen/ Total number of outcomes



Probability

- Random Experiment/ sample space/ mutually likely outcome/ equally likely outcome/ exhaustive outcome
- How to calculate the probability in simple cases?(What are the prerequisite to learn this? Basic set theory and concept of permutations and combinations.)
- Example (two coins/ three coins/one dice/ two dice)
- Conditional Probability(Concept)
- Conditional Probability(Formula)
- Or P(A/B) = P(A∩B)/P(B), concept independence and how to check ?

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

Probability……….

Example 1:

Toss three coins what is the probability of getting more than one heads?

Sol’n: S = (HHH, HHT, HTH, THH, HTT,THT,TTH,TTT)

event A=(HHH, HHT, HTH, THH)

P(A)= 4/8=1/2

Example 2:

Throw 2 dice,(a)what is the probability of getting even number on the first die?

(b) what is the probability of getting odd number on the second die?

(c) Are these two events independent?

(d) What is the probability that sum of the numbers on both the dice is 11?

Sol’n: S = (1,1) (1,2)………………..(1,6)
(2,1)…………………………(2,6)
………………………………………..
(6,1)……………………………(6,6) total 36 points in the sample space

Probability……….

(a)what is the probability of getting even number on the first die?= 18/36=1/2

(b) what is the probability of getting odd number on the second die? = 18/36=1/2

(c) Are these two events independent?

P(A)= ½, P(B)= ½ and P(A⋂B)= 9/36=1/4 (how), P(A)xP(B)=P(A⋂B)=1/4 independent

(d) What is the probability that sum of the numbers on both the dice is 11?

Sum can be any number from 2 to 12 but 11 can occur in following ways:

(5,6) and (6,5) so probability of getting 11 is = 2/36

Let me know what is the probability of getting sum 9?

Addition Rules for Probability

Addition Rule 1: When two events, A and B, are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event.

P(A or B) = P(A) + P(B)

Question 1: A single 6-sided die is rolled. What is the probability of rolling a 2 or a 5?
Answer : P(2 or 5) = P(2)+ P(5) = ⅙ + ⅙ = ⅓

Additional Rule 2: When two events, A and B, are non-mutually exclusive, the probability that A or B will occur is:

P(A or B) = P(A) + P(B) - P(A and B)

Question : In a math class of 30 students, 17 are boys and 13 are girls. On a unit test, 4 boys and 5 girls made an A grade. If a student is chosen at random from the class, what is the probability of choosing a girl or an A student?

Probabilities: P(girl or A) = P(girl) + P(A) - P(girl and A) = 13/30 + 9/30 - 5/30 = 17/30

Independent Events

Two events, A and B, are **independent** if the fact that A occurs does not affect the probability of B occurring.

Multiplication Rule 1: When two events, A and B, are independent, the probability of both occurring is:

P(A and B) = P(A) ▪ P(B)

Question: A dresser drawer contains one pair of socks with each of the following colors: blue, brown, red, white and black. Each pair is folded together in a matching set. You reach into the sock drawer and choose a pair of socks without looking. You replace this pair and then choose another pair of socks. What is the probability that you will choose the red pair of socks both times?

Probabilities: P(red and red) = P(red)▪ P(red) = ⅕ * ⅕ = 1/25

Independent Events

Factorial

There are n! ways of arranging n distinct objects into an ordered sequence, permutations where n = r.

Combination

The number of ways to choose a sample of r elements from a set of n distinct objects where order does not matter and replacements are not allowed.

Permutation

The number of ways to choose a sample of r elements from a set of n distinct objects where order does matter and replacements are not allowed. When n = r this reduces to n!, a simple factorial of n.

Combination Replacement

The number of ways to choose a sample of r elements from a set of n distinct objects where order does not matter and replacements are allowed.

Permutation Replacement

The number of ways to choose a sample of r elements from a set of n distinct objects where order does matter and replacements are allowed.

n - the set or population

r - subset of n or sample set

$C(n,r) = n! \,/\, (r!(n-r)!)$ For n ≥ r ≥ 0.

Example: You have won first place in a contest and are allowed to choose 2 prizes from a table that has 6 prizes numbered 1 through 6. How many different combinations of 2 prizes could you possibly choose?

In this example, we are taking a subset of 2 prizes (r) from a larger set of 6 prizes (n). Looking at the formula, we must calculate “6 choose 2.”

C (6,2)= 6!/(2! * (6-2)!) = 6!/(2! * 4!) = **15 Possible Prize Combinations**

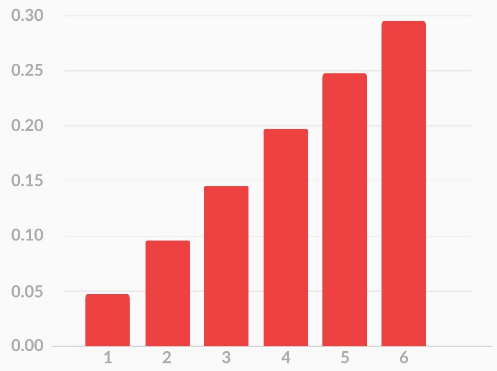
Random Variables & Probability Distribution

Random variable X basically converts outcomes of experiments to something measurable.
Now that the data is entirely in quantitative terms, it becomes possible to perform a number of different kinds of statistical analyses on it.

Probability distribution is ANY form of representation that tells us the probability for all possible values of X. It could be any of the following:

- A table
- A chart
- An equation- $P(x) = x/21$ (for $x = 1, 2, 3, 4, 5$ and 6)

x	P(x)
1	1/21
2	2/21
3	3/21
4	4/21
5	5/21
6	6/21



Random Variable…………

EXAMPLE : In the experiment of tossing a fair coin three times , ***the sample space S, consists of*** eight equally likely sample points ***S, = (HHH, . . . , TTT). If X is the r.v. giving the number of heads obtained, find***

(a) ***P(X = 2); (b) P(X < 2).***

(a) Let A c S, be the event defined by ***X = 2. Then, from Prob. example, we have***

- ***A = (X = 2) = {C: X(C) = 2) = {HHT, HTH, THH)***

Since the sample points are equally likely, we have

- ***P(X = 2) = P(A) = 3/8***
- Let ***B c S , be the event defined by X < 2. Then***
- ***B = (X < 2) = {c: X(c) < 2) = (HTT, THT, TTH, TTT)***
- and ***P(X < 2) = P(B) = 4/8***

Expected Value

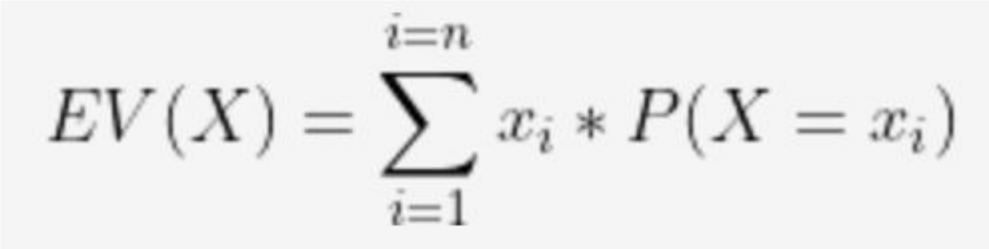
Expected value for a variable X is the value of X we would “expect” to get after performing the experiment once.

It is also called the expectation, average, and mean value. Mathematically speaking, for a random variable X that

can take values

x1,x2,x3,……….,xn, the **expected value (EV)** is given by:

EV(X)=x1*P(X=x1)+x2*P(X=x2)+x3*P(X=x3)+……….+xn*P(X=xn)



$$EV(X) = \sum_{i=1}^{i=n} x_i * P(X = x_i)$$

Cumulative Probability

.

Cumulative probability of X, denoted by F(x), is defined as the probability of the variable being less than or equal to x.

Consider a coin flip experiment. If we flip a coin two times, we might ask: What is the probability that the coin flips would result in one or fewer heads? The answer would be a cumulative probability. It would be the probability that the coin flip results in zero heads plus the probability that the coin flip results in one head. Thus, the cumulative probability would equal:

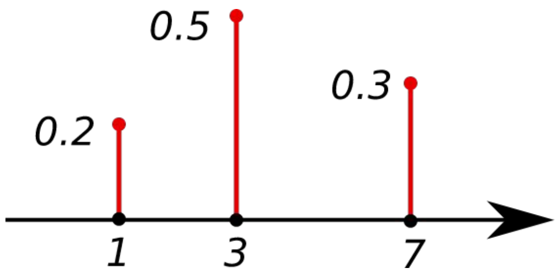
$$P(X < 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75$$

The table below shows both the probabilities and the cumulative probabilities associated with this experiment.

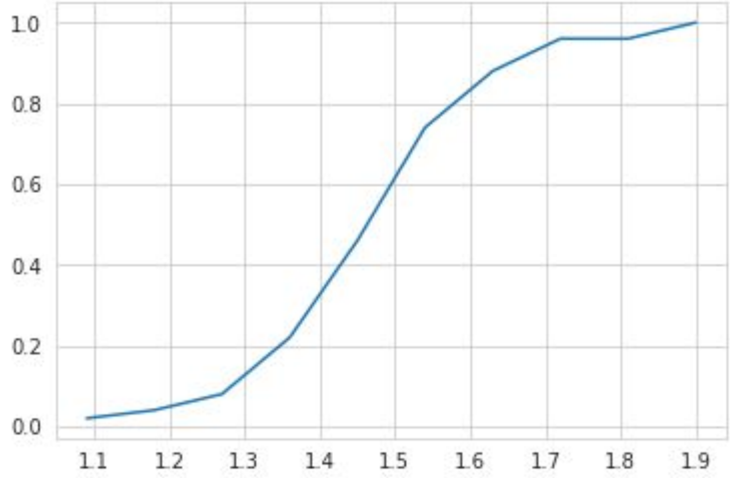
Number of heads	Probability	Cumulative Probability
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00

Probability Functions

PMF, is a statistical term that describes the probability distribution of the Discrete random variable.



CDF, or a cumulative distribution function, is a distribution which plots the cumulative probability of X against X.



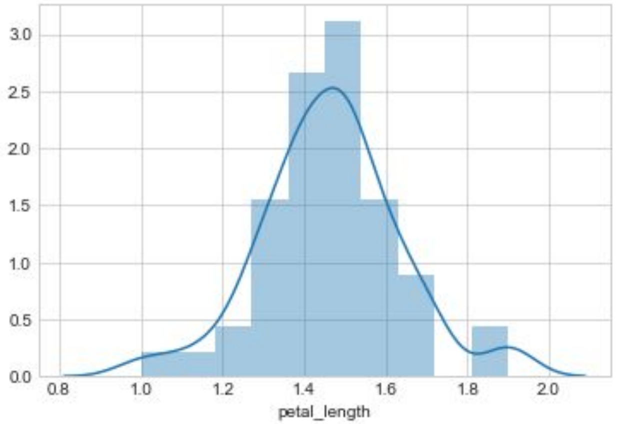
A PDF, or Probability Density Function, however, is a function in which the area under the curve, gives you the cumulative probability.

The main difference between the cumulative probability distribution of a continuous random variable and a discrete one, is the way you plot them. While the continuous variables' cumulative distribution is a curve, the distribution for discrete variables looks more like a bar chart.

PDFs are more commonly used in real life.

The reason is that it is much easier to see patterns in PDFs as compared to CDFs.

For example, here are the PDF and the CDF of a uniformly distributed continuous random variable:



PMF Example..

Example1 : if we toss three coins write down the sample space, define on rv on this and write its pmf. Also calculate P(X≤2)

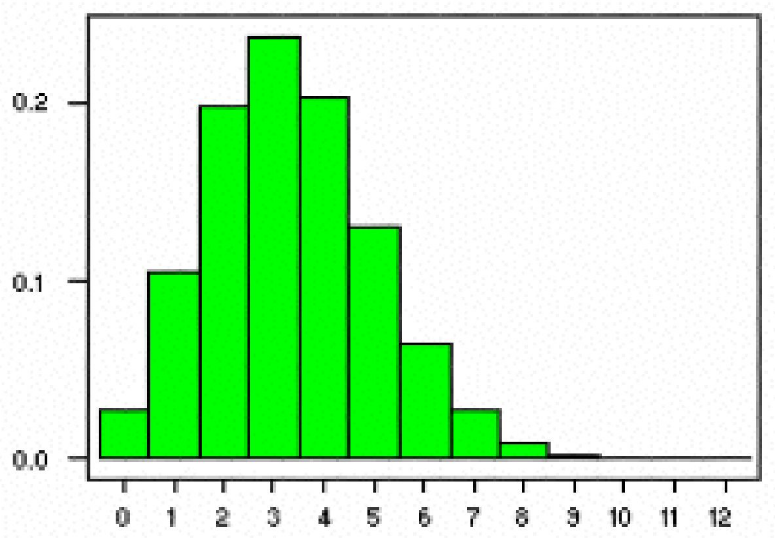
Sol’n: S = (HHH, HHT, HTH, THH, HTT,THT,TTH,TTT), let us define rv Y as number of tails then y= 0,1,2,3 and corresponding pmf is given by:

x	0	1	2	3	Total
f(x)	1/8	3/8	3/8	1/8	1

Binomial Distribution

- Describes the probability of a Boolean outcome (Yes/ No)
- If n is the number of trials , x is the particular trial
- p is the probability of success
- Plots the probabilities of all values of x->

$$P[X=x]=\begin{cases} {}^{n}C_{x}p^{x}q^{n-x} \; ; & x=0,1,2,\ldots,n \\ 0; & \text{elsewhere} \end{cases}$$



PMF

There are some conditions that need to be followed in order for us to be able to apply the formula.

1. Total number of trials is fixed at n
2. Each trial is binary, i.e., has only two possible outcomes - success or failure
3. Probability of success is same in all trials, denoted by p

Example

A coin is tossed 10 times. What is the probability of getting exactly 6 heads?

I’m going to use this formula: b(x; n, P) – nCx * Px * (1 – P)n – x

The number of trials (n) is 10

The odds of success (“tossing a heads”) is 0.5 (So 1-p = 0.5)

x = 6

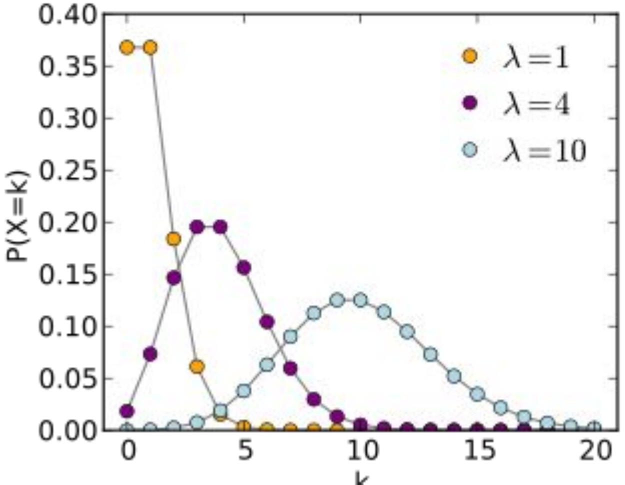
P(x=6) = 10C6 * 0.5^6 * 0.5^4 = 210 * 0.015625 * 0.0625 = 0.205078125

Poisson Distribution

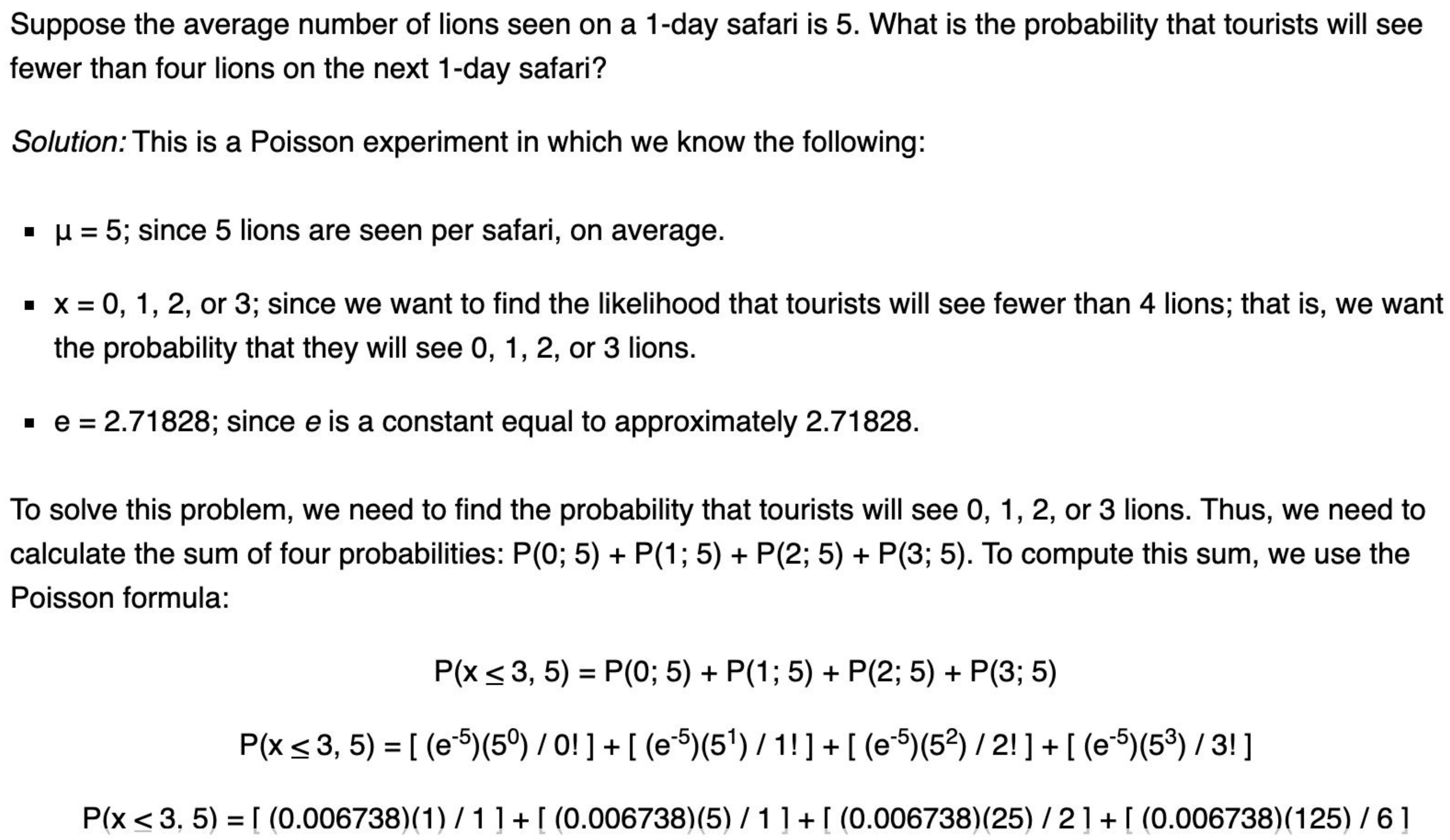
- We can derive or get Poisson from binomial with some conditions.
- The probability distribution of the number of occurrences, X, of some random event, in an interval of time or space.

  P(X= x) = e^(-λ). λ^ x/ xị   x = 0,1,2……………..

- The mean and variances of the distribution are both λ . The skewness of the distribution is 1/√ λ and its kurtosis is 3 +1/λ
- Its range, its parameter
- In which situations this can be used a probability models?



Example



Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 5$; since 5 lions are seen per safari, on average.
- $x = 0, 1, 2,$ or 3; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
- $e = 2.71828$; since *e* is a constant equal to approximately 2.71828.

To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: $P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)$. To compute this sum, we use the Poisson formula:

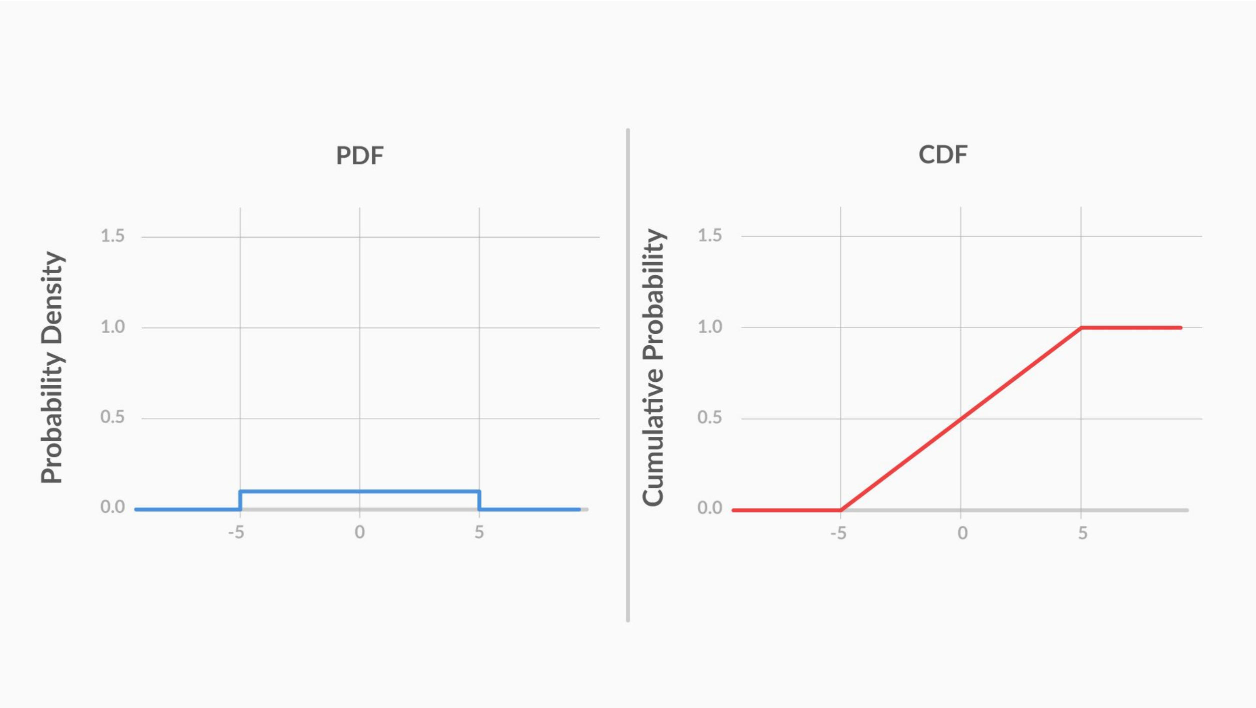
$$P(x \leq 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)$$

$$P(x \leq 3, 5) = [\ (e^{-5})(5^0) / 0!\ ] + [\ (e^{-5})(5^1) / 1!\ ] + [\ (e^{-5})(5^2) / 2!\ ] + [\ (e^{-5})(5^3) / 3!\ ]$$

$$P(x \leq 3, 5) = [\ (0.006738)(1) / 1\ ] + [\ (0.006738)(5) / 1\ ] + [\ (0.006738)(25) / 2\ ] + [\ (0.006738)(125) / 6\ ]$$

Uniform Distribution

The PDF clearly shows uniformity, as the probability density's value remains constant for all possible values. However, the CDF does not show any trends that help you identify quickly that the variable is uniformly distributed.



PDF and CDF for a Uniformly Distributed Variable

Example

A fountain erupts every 91 minutes. You arrive there at random and wait for 20 minutes ... what is the probability you will see it erupt?

This is actually easy to calculate, 20 minutes out of 91 minutes is:

p = 20/91 = 0.22 (to 2 decimals)

But let's use the Uniform Distribution for practice.

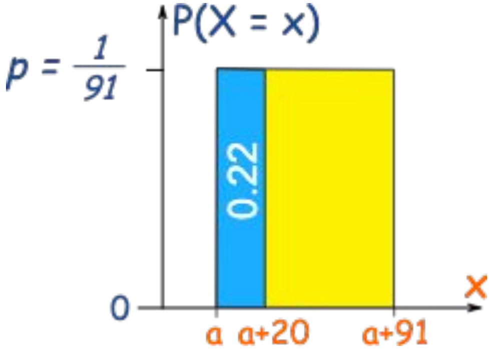
To find the probability between **a** and **a+20**, find the blue area:

Area = (1/91) x (a+20 − a)

= (1/91) x 20

= 20/91

= **0.22** (to 2 decimals)



So there is a 0.22 probability you will see Old Faithful erupt.

If you waited the full 91 minutes you would be sure (**p=1**) to have seen it erupt.

But remember this is a random thing! It might erupt the moment you arrive, or any time in the 91 minutes.

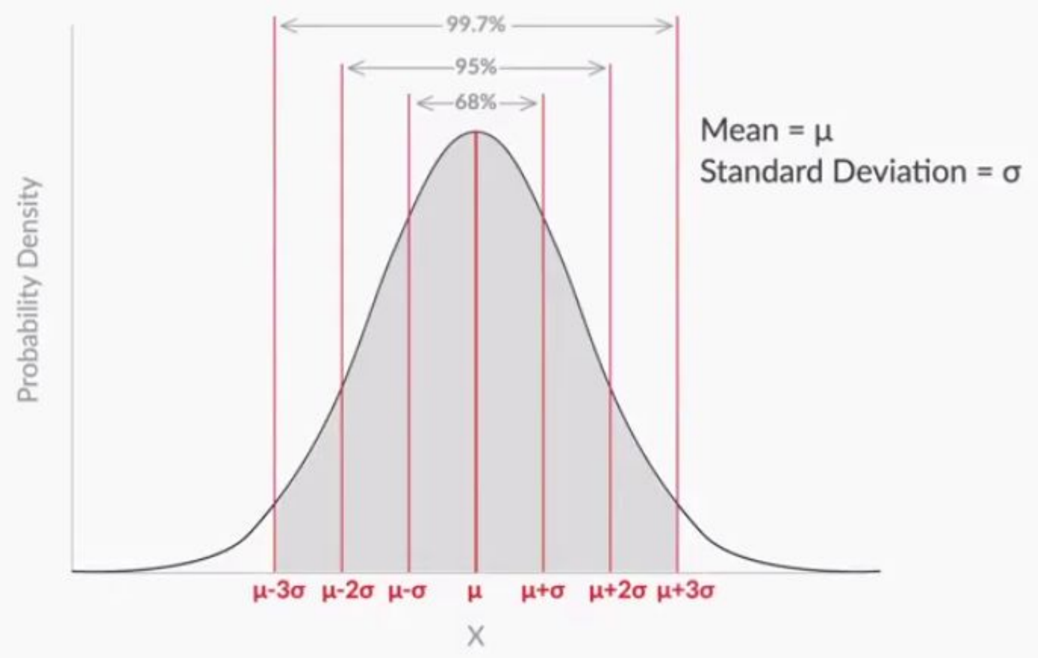
Normal Distribution

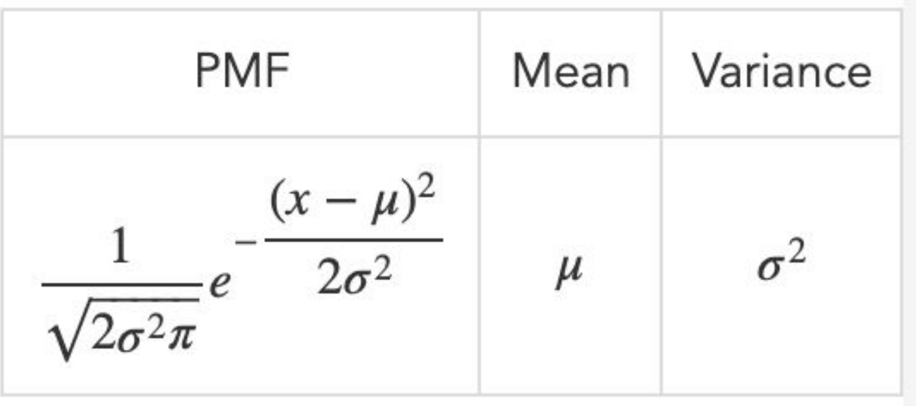
It is symmetric and its mean, median and mode lie at the centre.All data that is normally distributed follows the 1-2-3 rule. This rule states that there is a -

1. 68% probability of the variable lying within 1 standard deviation of the mean
2. 95% probability of the variable lying within 2 standard deviations of the mean
3. 99.7% probability of the variable lying within 3 standard deviations of the mean

The normal (or Gaussian) distribution is used in the natural and social sciences to represent real-valued random variables whose distributions are not known.

For example the normal distribution might be used to model people's height, where it is assumed most people are around the same height.





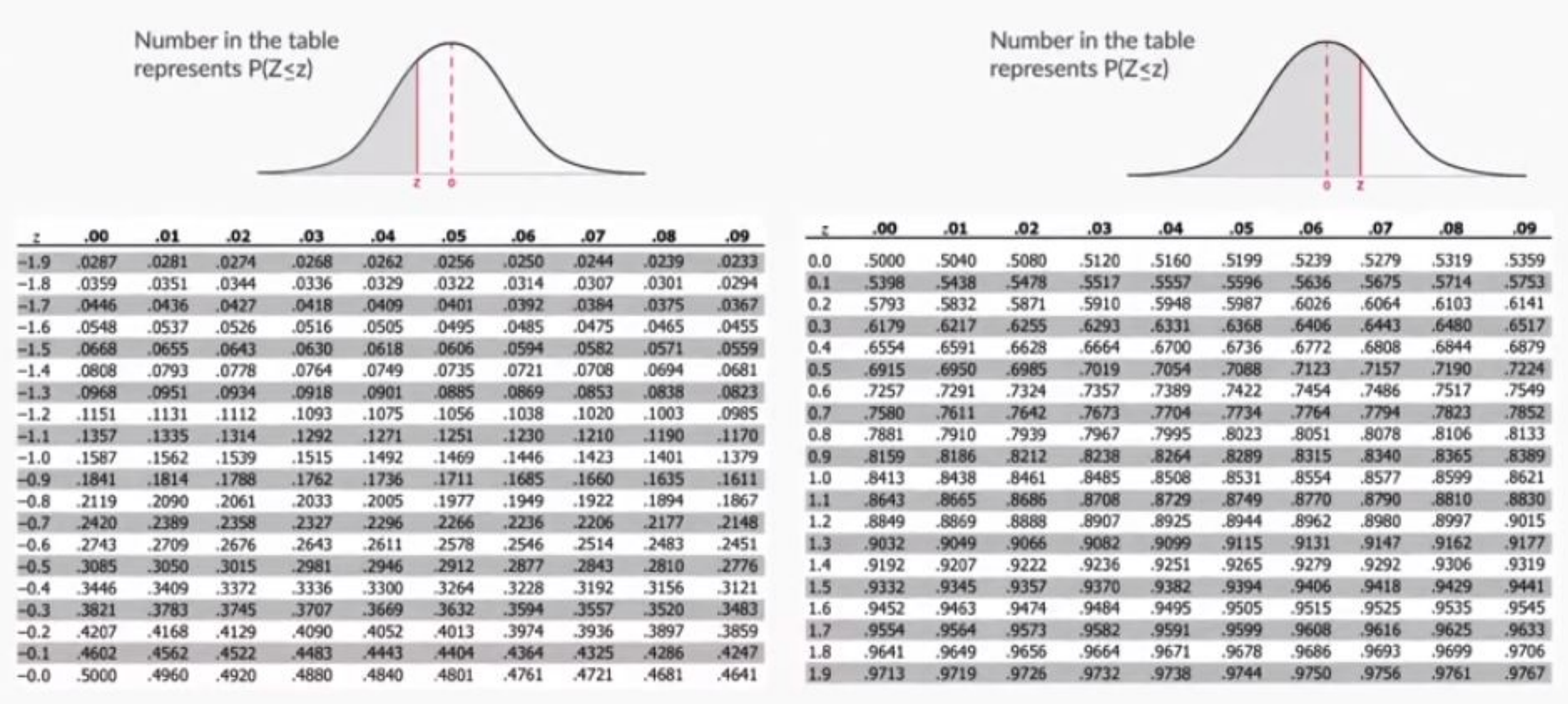
PMF	Mean	Variance
$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

Normal Distribution

Standardised random variable or Z Score is an important parameter. It is given by

$$Z = \frac{X-\mu}{\sigma}$$

Basically, it tells you how many standard deviations away from the mean your random variable is. As you just saw, you can find the cumulative probability corresponding to a given value of Z, using the Z table:



Number in the table represents P(Z≤z)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Number in the table represents P(Z≤z)

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
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Inferential Stats

Inferential statistics is used to find some population parameter (mostly population mean) when you have no initial number to start with. So, you start with the sampling activity and find out the sample mean. Then, you estimate the population mean from the sample mean using the confidence interval.

Population vs sample

- First, you need to identify the target population of your research.
- The **population** is the entire group that you want to draw conclusions about.
- The **sample** is the specific group of individuals that you will collect data from.
- The number of individuals in your sample depends on the size of the population, and on how precisely you want the results to represent the population as a whole.
- You can use a sample size calculator to determine how big your sample should be. In general, the larger the sample size, the more accurately and confidently you can make inferences about the whole population.

Types of Sampling Techniques

Probability Sampling

1. **Simple Random**
2. **Systematic**
3. **Stratified**
4. **Cluster**

Probability sampling means that every member of the population has a ***specified*** chance of being selected in the sample. If you want to produce results that are representative of the whole population, you need to use a probability sampling technique.

Non-Probability Sampling

1. **Convenience**
2. **Quota**
3. **Judgement**
4. **Snowball**

In Non Probability Sampling, elements do not have any specified chance of being selected in the sample. Consequently, there is a significant risk of ending up with a non-representative sample which doesn't produce results which we can generalize.

Types of Sampling Techniques

Simple Random Sampling

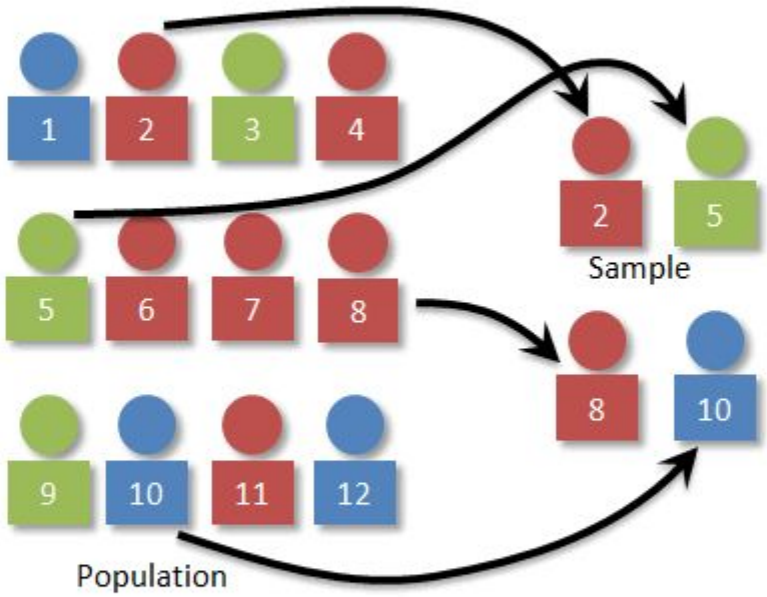
- In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.
- To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.(SRS/ SRSWOR/SRSWR, When/why)

Example

- You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

What are the major hurdles in this??

1. Frame 2. operational inconvenience 3. some times give non random looking results(when population is heterogeneous)



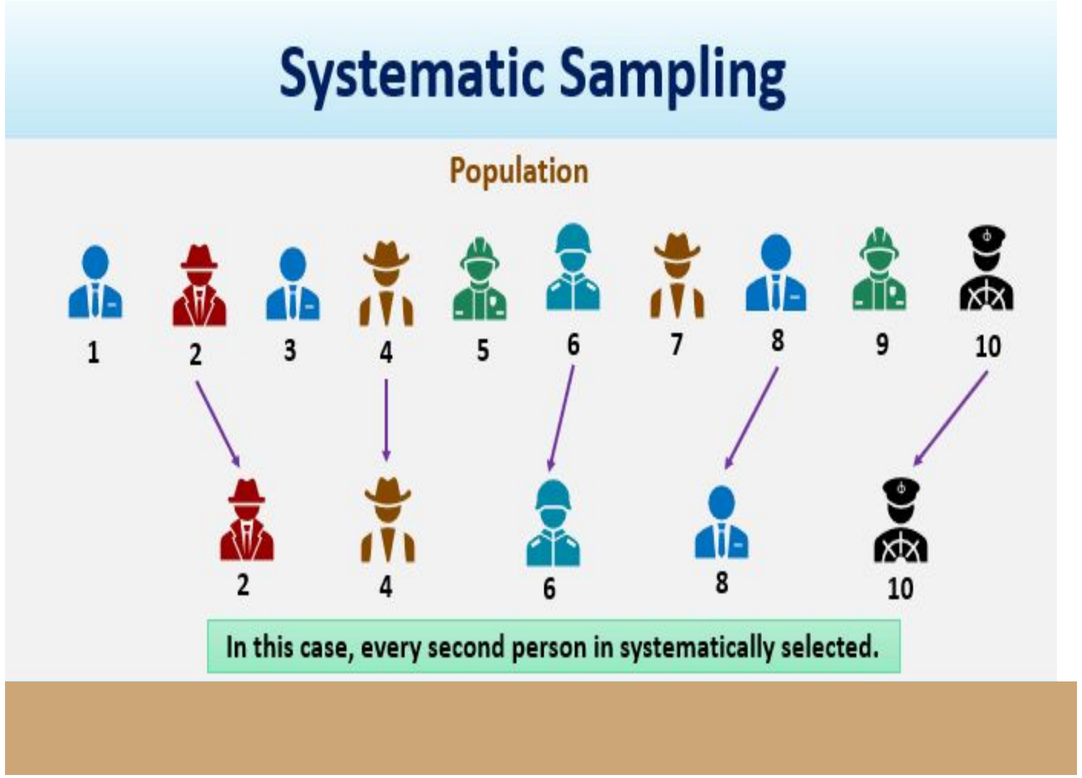
Types of Sampling Techniques

Systematic Sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.
All 1000 employees of the company are listed in alphabetical order and we want to have a sample of 100, as 1000/100=10. From the first 10 numbers, you randomly select a starting point: suppose that number is 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

(how we have improved on SRS)


Systematic Sampling

In this case, every second person in systematically selected.

Types of Sampling Techniques

Stratified Sampling

- This sampling method is appropriate when the population has mixed characteristics, and you want to ensure that every characteristic is proportionally represented in the sample.
- You divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role). The factor which is basis for their grouping is known as stratification factor.

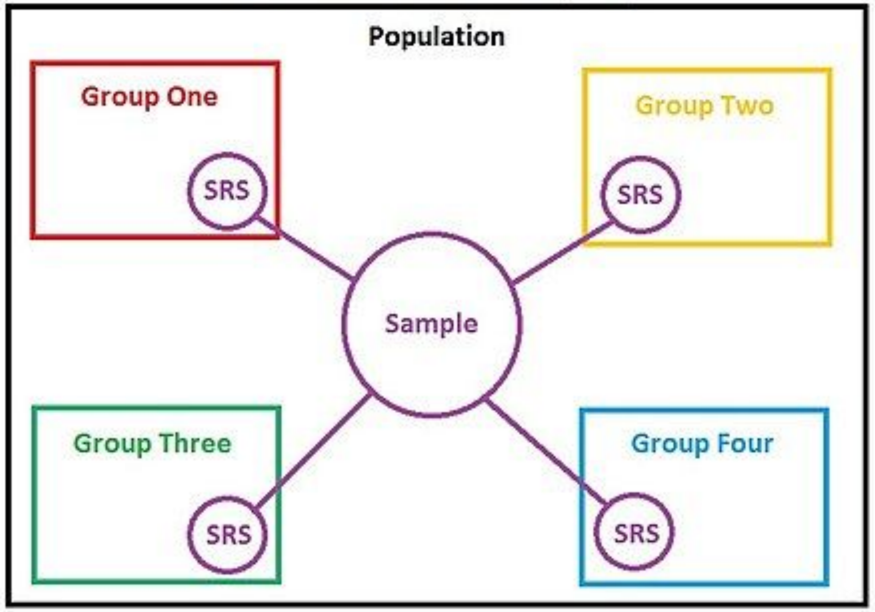
Problem of allocation:

1. Proportional Allocation 2. Neyman or optimum allocation

Example

- The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

Stratified Random Sampling



beingdatum
the data society

Non-probability sampling methods

- In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included. This type of sample is easier and cheaper to access, but you can't use it to make valid statistical inferences about the whole population.

- Non-probability sampling techniques are often appropriate for exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.

1. **Convenience sampling: A** convenience sample simply includes the individuals who happen to be most accessible to the researcher.

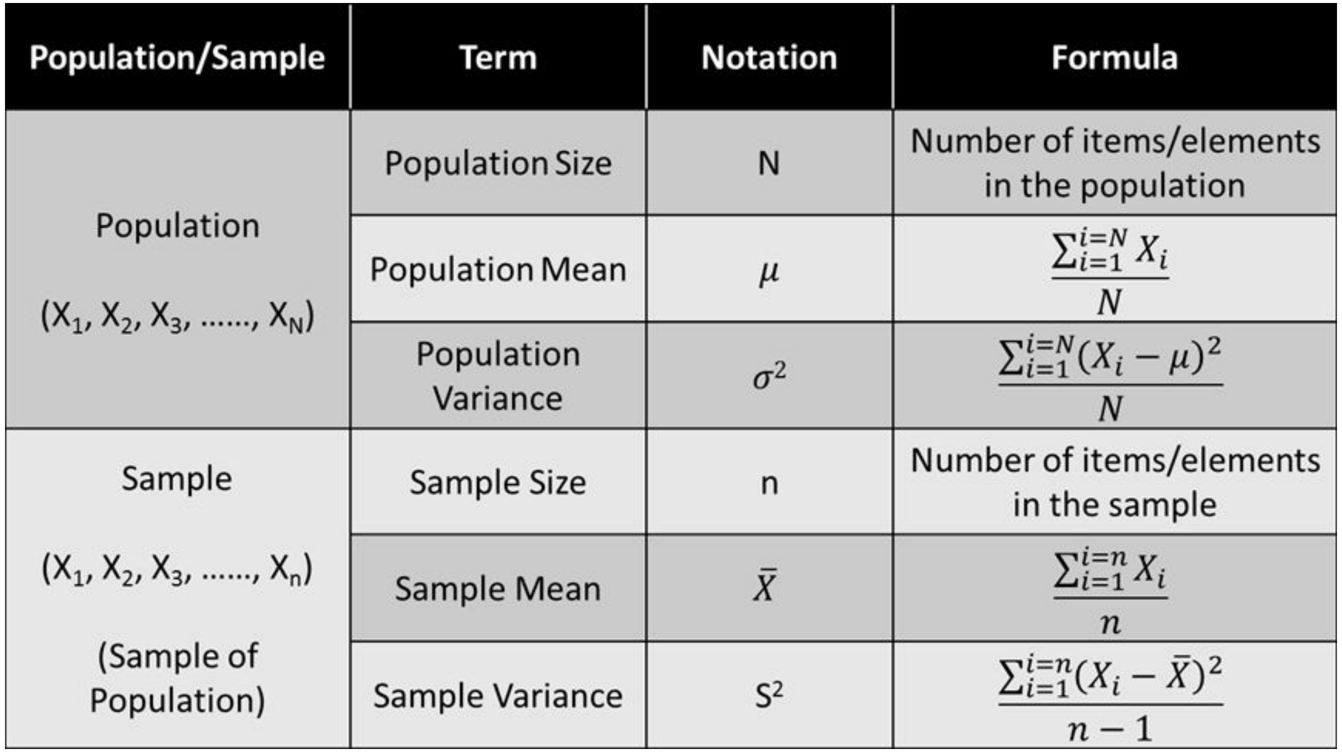
2. **Purposive sampling:** This type of sampling involves the researcher using their judgment to select a sample that is most useful to the purposes of the research.

3. **Snowball sampling:** If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people.

There are many other probability and non probability sampling methods but we have discussed the major ones

Population Sampling

Instead of finding the mean and standard deviation for the entire population, it is sometimes beneficial to find the mean and standard deviation for only a small representative sample. You may have to do this because of time and/or money constraints. It would not be fair to infer that the population mean is exactly equal to the sample mean. This is because the flaws of the sampling process must have led to some error. Hence, the sample mean's value has to be reported with some **margin of error.** The most useful functions are listed below:

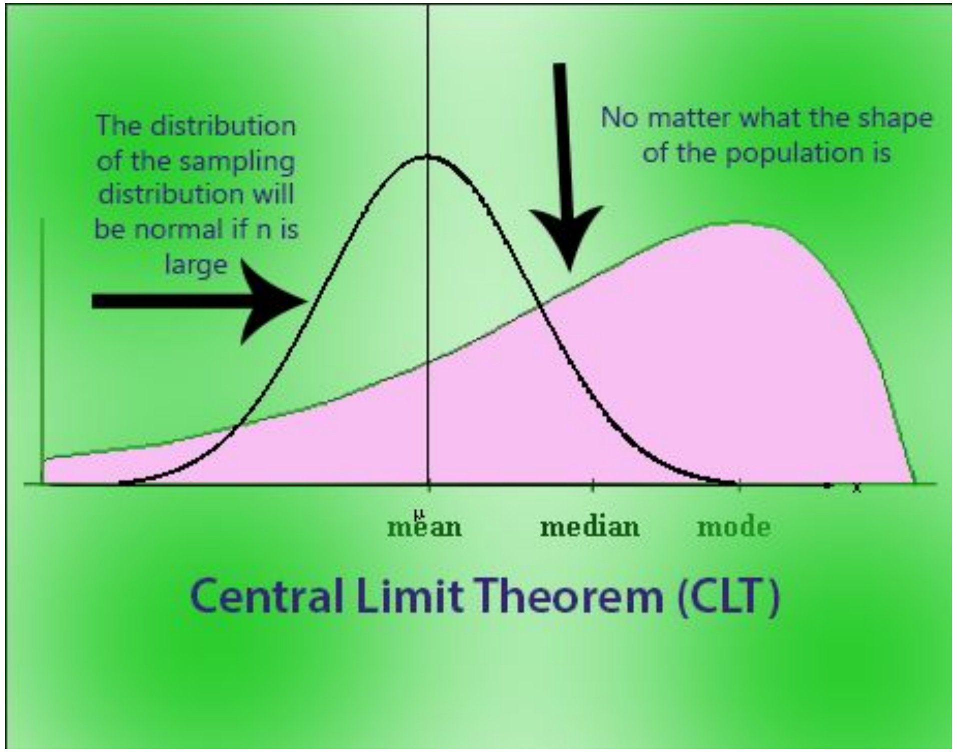


Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N} (X_i - \mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n} (X_i - \bar{X})^2}{n-1}$

Central limit Theorem

The central limit theorem says that, for any kind of data, provided a high number of samples has been taken, the following properties hold true:

- Sampling distribution mean (μ¯X) = Population mean (μ)
- Sampling distribution’s standard deviation (Standard error) = σ / √n
- For n > 30, the sampling distribution becomes a normal distribution


The distribution of the sampling distribution will be normal if n is large
No matter what the shape of the population is
mean
median
mode
Central Limit Theorem (CLT)

Confidence Interval

- A Confidence Interval is a **range of values** we are fairly sure our **true value** lies in.
- Example: Average Height
- We measure the heights of **40** randomly chosen men, and get a mean height of **175cm**,
- We also know the standard deviation of men's heights is **20cm**.
- The **95% Confidence Interval** (we show how to calculate it later) is:**175cm ± 6.2cm**
- https://www.mathsisfun.com/data/confidence-interval.html

Standard Error(SE), Margin of Error, Confidence Interval and Sample Size

$$SE = \frac{\sigma}{\sqrt{n}}$$

sigma is the standard deviation of the population.
n is the size (number of observations) of the sample.

$$Margin\ of\ Error = \mathrm{Z} * SE$$

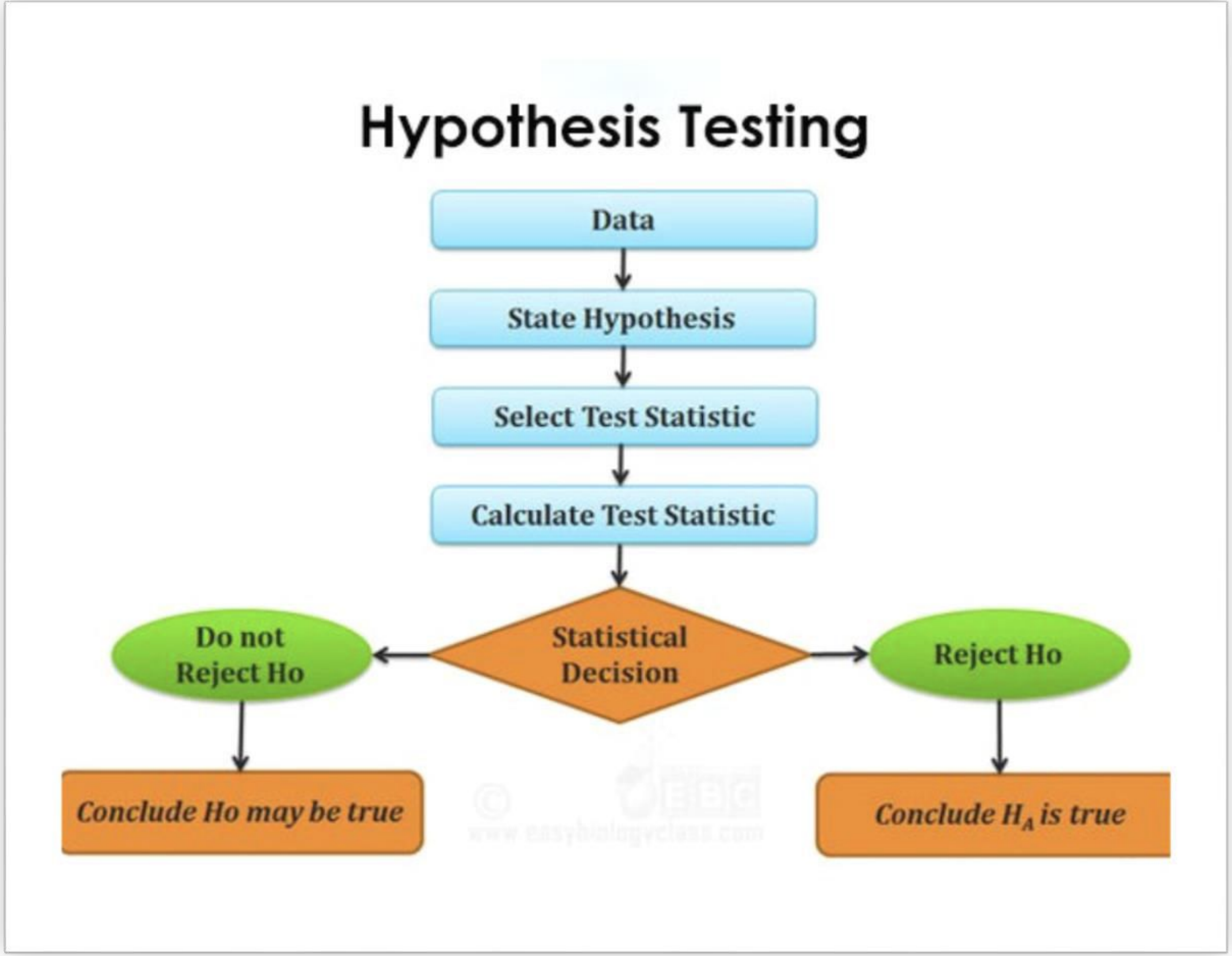
Margin of error is the radius or half-width of a confidence interval.

z-score is the number of standard deviations from the mean a data point is.

$z = (x – \mu) / \sigma$

Significance Level

- 5% is called **Significance Level** also known as alpha level (symbolized as α). It means that if random chance probability is less than 5% then we can conclude that there is difference in behavior of two different population. (1- Significance level) is also known as **Confidence Level** i.e. we can say that I am 95% confident that it is not driven by randomness.



Hypothesis Testing

Data

State Hypothesis

Select Test Statistic

Calculate Test Statistic

Statistical Decision

Do not Reject Ho

Reject Ho

Conclude Ho may be true

Conclude H_A is true

Hypothesis testing

Hypothesis testing is used to confirm your conclusion (or hypothesis) about the population parameter (which you know from EDA or your intuition). Through hypothesis testing, you can determine whether there is enough evidence to conclude if the hypothesis about the population parameter is true or not.

Hypothesis Testing starts with the formulation of these two hypotheses:

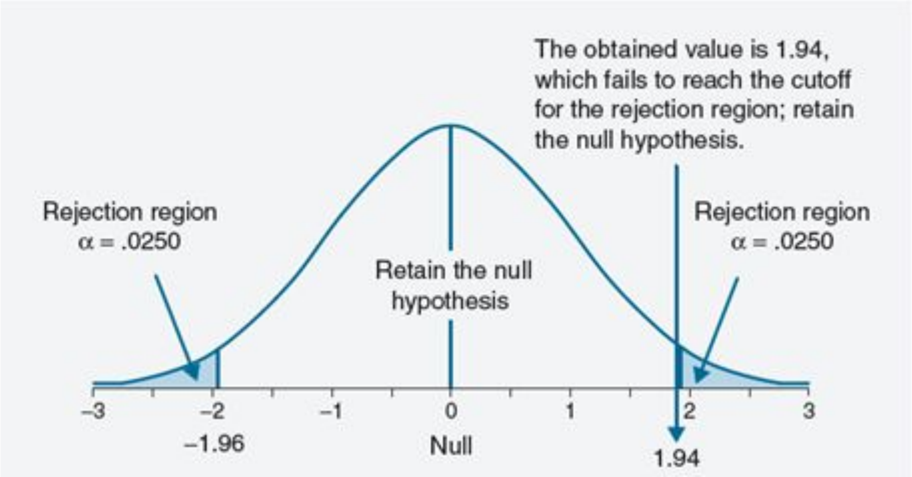
- **Null hypothesis** (H_0): The status quo
- **Alternate hypothesis** (Ha): The challenge to the status quo

Choosing a test statistic

No.	Application	Example	Test Used
1	To check whether Sample Mean =Population Mean	Avg salary of Company XYZ, Sales Dept employees is Rs. 20,000	One Sample T Test or Z Test, if population size <30 go for t-test
2	To compare Mean of One Population Vs Mean of another population	Apple Vs. Samsung mobile phone sales in a region	Independent T Test or Z Test for both companies separately & compare
3	To compare the effect before and after a particular event, on the same sample	Effect of a BP drug on patients, before and after consumption of the drug	Paired T Test
4	To find out Goodness of Fit (Observed =Expected or not)-attributes are usually categorical/qualitative	Estimated Vs actual sales	Chi-Square Test

Hypothesis testing

Making the decision to either reject or fail to reject the null hypothesis is based on the critical values and the position of the sample mean on the distribution.

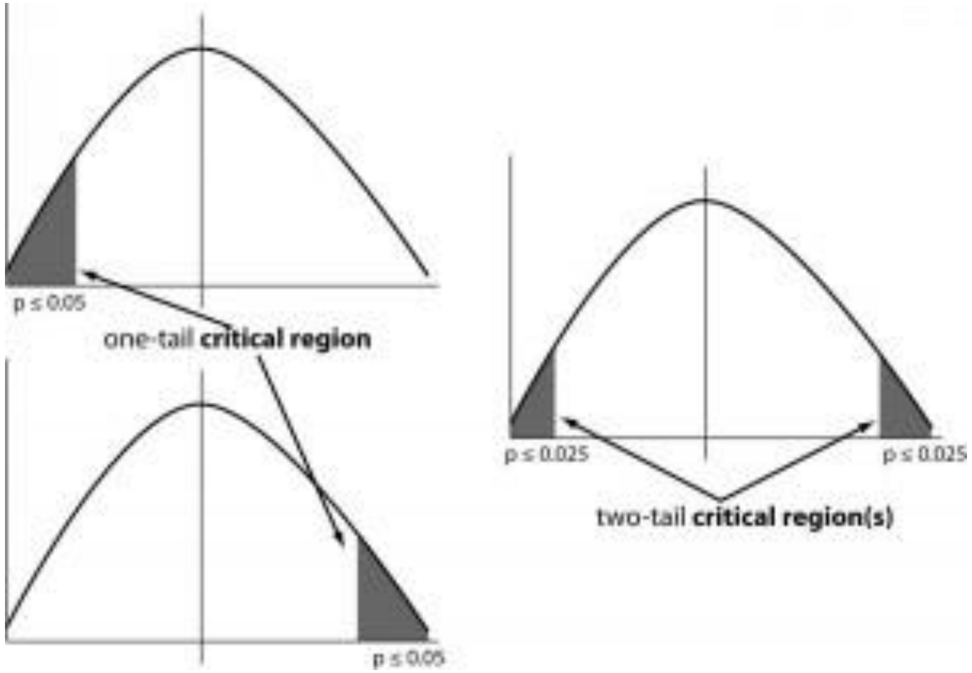


Making a decision - Critical value method:

- Calculate the value of Zc from the given value of α (significance level)
- Calculate the critical values (UCV and LCV) from the value of Zc
- Make the decision on the basis of the value of the sample mean ¯x with respect to the critical values (UCV AND LCV)

Directional/ Non Directional Hypothesis Testing

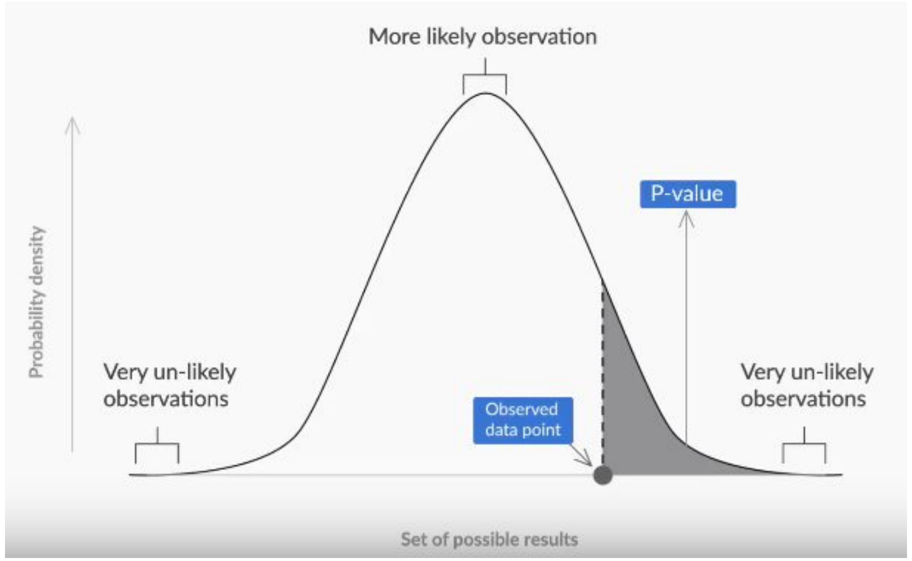
- In one tail test, our alternate hypothesis is greater or less than the observed mean so it is also known as Directional Hypothesis test. On the other hand, if you don’t know whether the impact of test is greater or lower then we go with Two tail test also known as Non Directional Hypothesis test.



p-value

p-value is the probability of the null hypothesis being accepted (or more aptly, not being rejected). This statement is not technically correct (or formal) definition of p-value, but it is used for better understanding of the p-value.

Higher the p-value, higher is the probability of failing to reject a null hypothesis. On the other hand, lower the p-value, higher is the probability of the null hypothesis being rejected.



p-value

To make a decision using the p-value method are as follows:

1. Calculate the value of z-score for the sample mean point on the distribution
2. Calculate the p-value from the cumulative probability for the given z-score using the z-table
3. Make a decision on the basis of the p-value (multiply it by 2 for a two-tailed test) with respect to the given value of α (significance value).

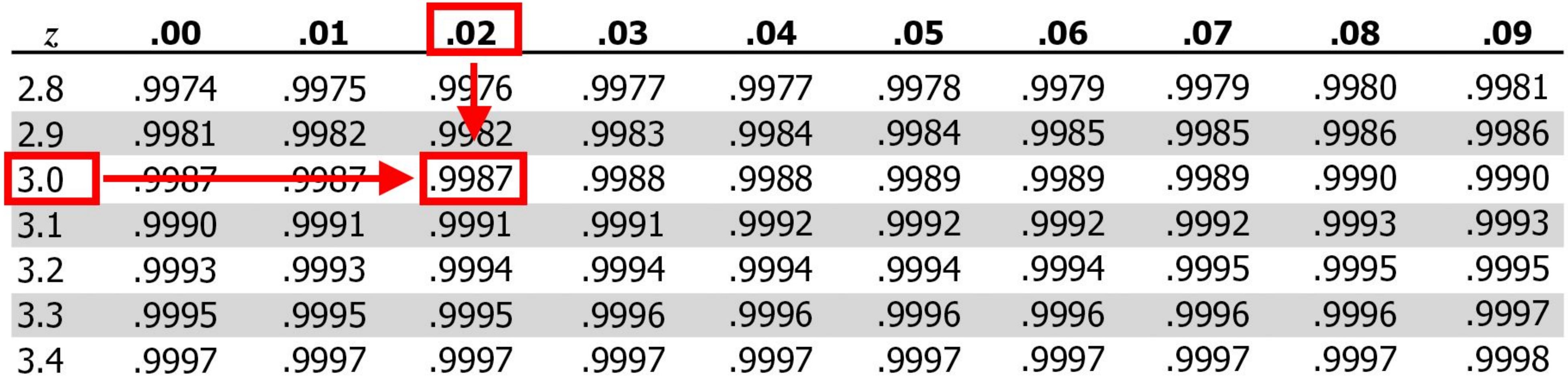
To find the correct p-value from the z-score, first find the cumulative probability by simply looking at the z-table, which gives you the area under the curve till that point.

z-score for sample point = + 3.02

Cumulative probability of sample point = 0.9987

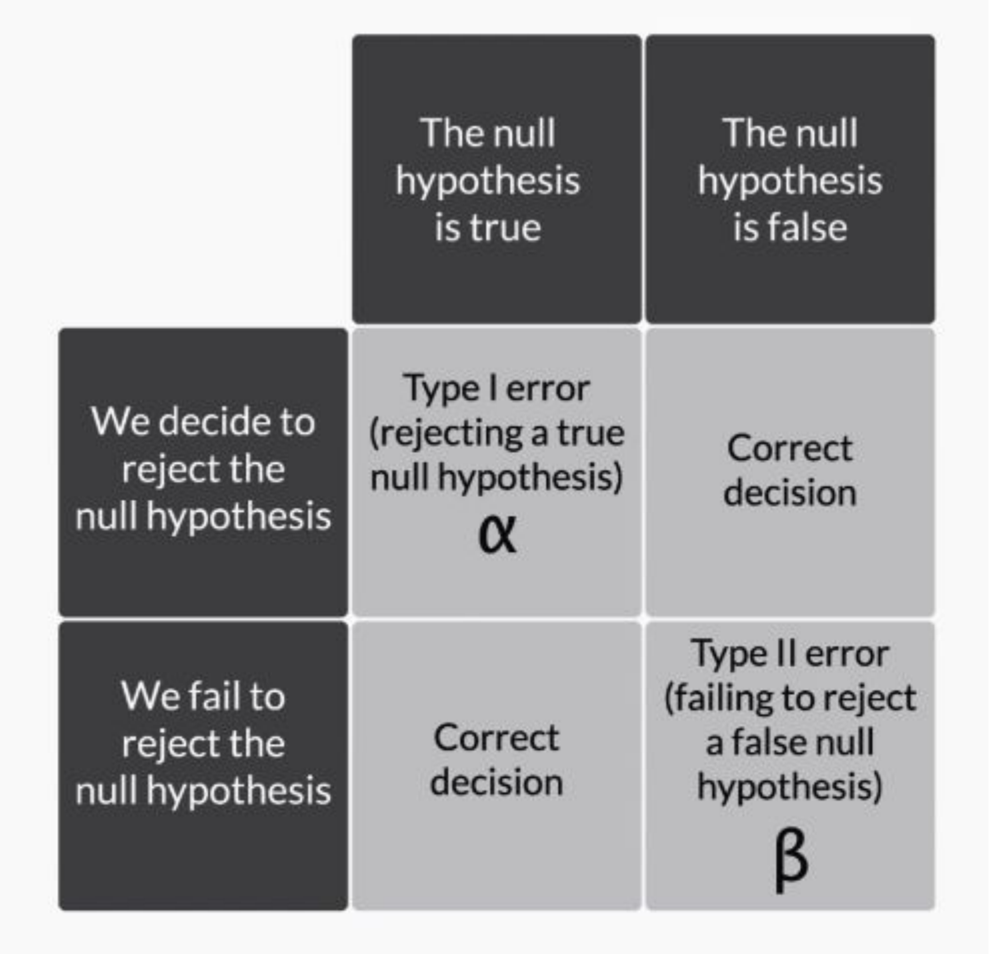
For one-tailed test → p = 1 - 0.9987 = 0.0013

For two-tailed test → p = 2 (1 - 0.9987) = 2 * 0.0013 = 0.0026

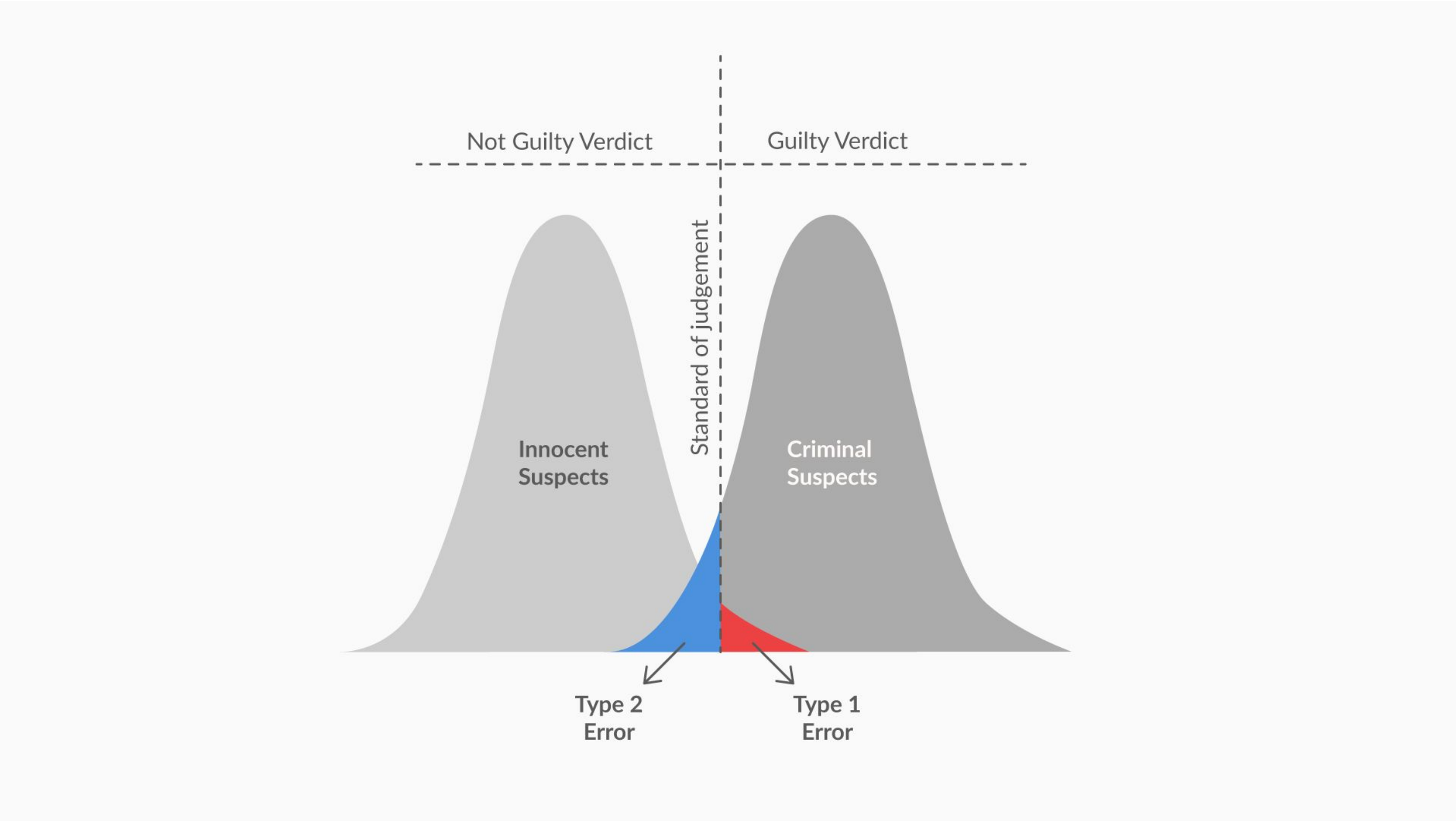


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Errors



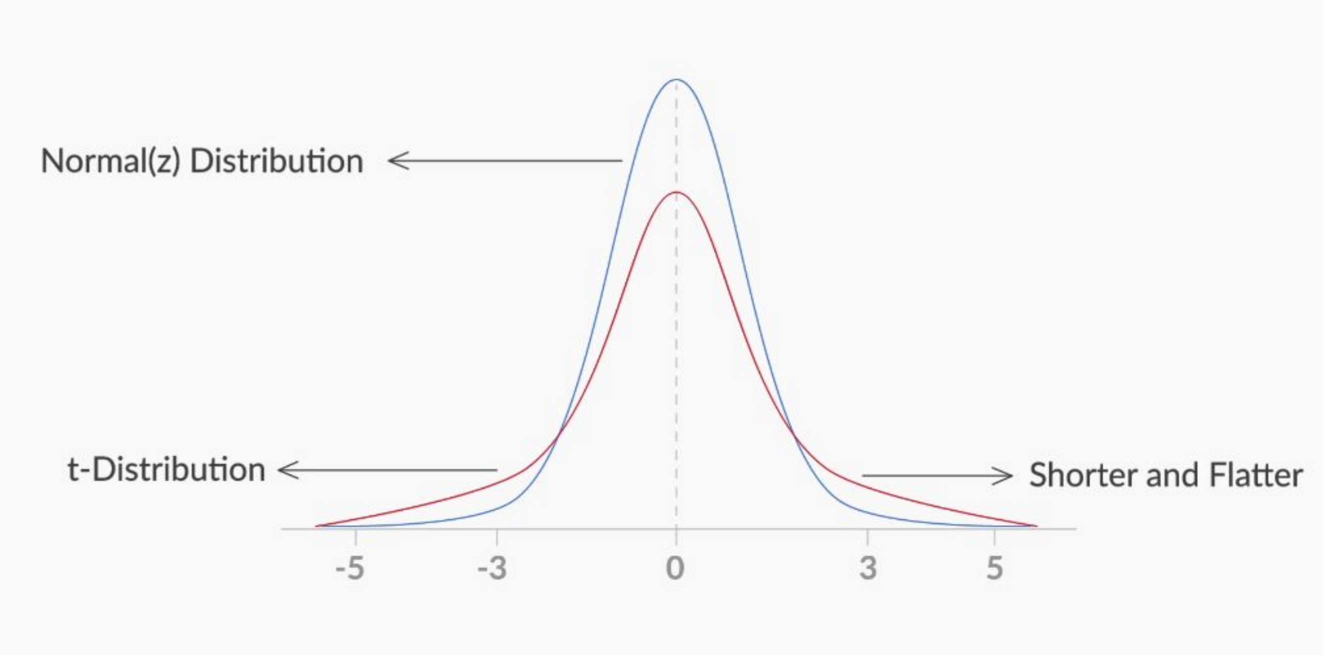
	The null hypothesis is true	The null hypothesis is false
We decide to reject the null hypothesis	Type I error (rejecting a true null hypothesis) α	Correct decision
We fail to reject the null hypothesis	Correct decision	Type II error (failing to reject a false null hypothesis) β

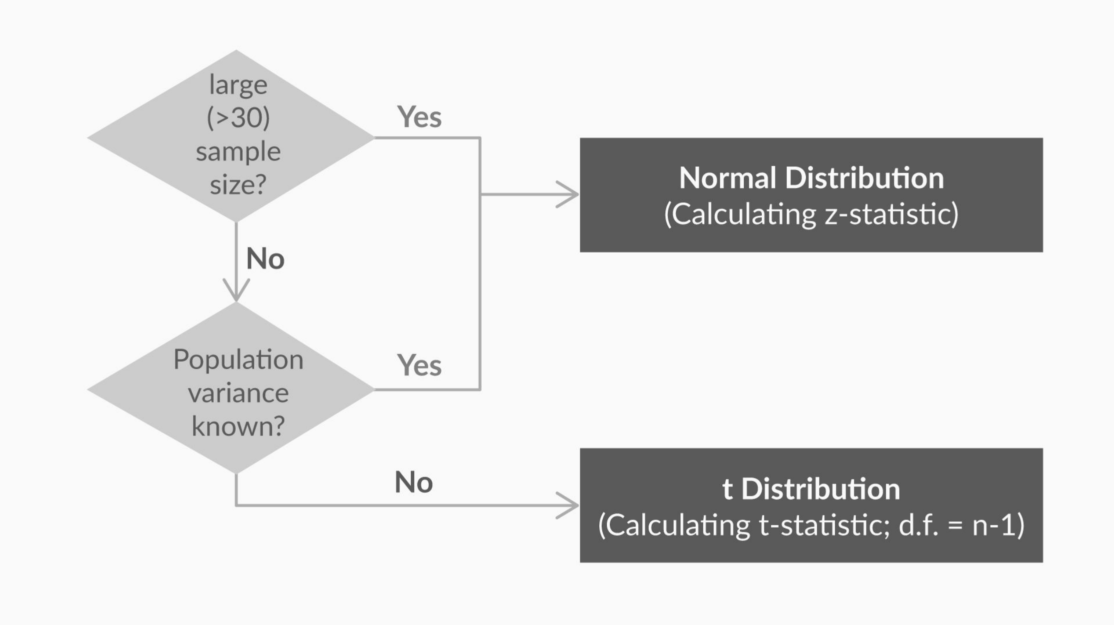


t-distribution

A t-distribution is also referred to as **Student's T distribution**. A t-distribution is similar to the normal distribution in many cases; for example, it is symmetrical about its central tendency. However, it is shorter than the normal distribution and has a flatter tail, which would eventually mean that it has a larger standard deviation.

Calculate the value of Zc from the given value of α (significance level). Take it as 5% if not specified in the problem.So, to find Zc, you would use the **t-table** instead of the z-table. The **t-table** contains values of Zc for a given degree of freedom and value of α (significance level). Zc, in this case, can also be called as t-statistic (critical).




large (>30) sample size?
Yes
No
Population variance known?
Yes
No
Normal Distribution
(Calculating z-statistic)
t Distribution
(Calculating t-statistic; d.f. = n-1)

Chi-square Goodness of Fit Test

- Chi-square test is used when we have one single categorical variable from the population.
- Let us understand this with help of an example. Suppose a company that manufactures chocolates, states that they manufacture **30% dairy milk, 60% temptation and 10% kit-kat**. Now suppose a random sample of 10000 chocolates has 50 dairy milk, 45 temptation and 5 kitkats. Does this support the claim made by the company?
- Let us state our Hypothesis first.
- Null Hypothesis: The claims are True
- Alternate Hypothesis: The claims are False.
- Chi-Square Test is given by:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where, O_i = sample or observed values

E_i = population values

- Let us now calculate the Expected values of all the levels.
- E (dairy milk)= 100 * 30% = 30
- E (temptation) = 100 * 60% =60
- E (kitkat) = 100 * 10% = 10
- Calculating chi-square = [(50-30)^2/30 + (45-60)^2/60 + (5-10)^2/10] =**19.58**
- Now, checking for p (chi-square < 19.58) using chi-square calculator, we get p=0.0001~0. This is significantly lower than the alpha(0.05).
- So we reject the Null Hypothesis.The claim made my the company is false.

https://www.analyticsvidhya.com/blog/2017/01/comprehensive-practical-guide-inferential-statistics-data-science/?utm_source=linkedinHA&utm_medium=linkedinHA

Examples of Hypothesis testing, z

$$z = \frac{\hat{p} - p}{\sqrt{p\,q/n}}$$

The CEO of a large electric utility claims that 80 percent of his 1,000,000 customers are very satisfied with the service they receive. To test this claim, the local newspaper surveyed 100 customers, using simple random sampling. Among the sampled customers, 73 percent say they are very satisfied. Based on these findings, can we reject the CEO's hypothesis that 80% of the customers are very satisfied? Use a 0.05 level of significance.

***Solution**:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

  Null hypothesis: P = 0.80

  Alternative hypothesis: P ≠ 0.80

  Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the sample proportion is too big or if it is too small.
- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. The test method, shown in the next section, is a one-sample z-test.
- **Analyze sample data**. Using sample data, we calculate the standard deviation (σ) and compute the z-score test statistic (z).

  σ = sqrt[P * (1 - P) / n]

  σ = sqrt [(0.8 * 0.2) / 100]

  σ = sqrt(0.0016) = 0.04

  z = (p - P) / σ = (.73 - .80)/0.04 = -1.75
- Since we have a two-tailed test, the P-value is the probability that the z-score is less than -1.75 or greater than 1.75.
- We use the z-table to find P(z < -1.75) = 0.04, and P(z > 1.75) = 0.04. Thus, the P-value = 0.04 + 0.04 = 0.08.
- **Interpret results**. Since the P-value (0.08) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Examples of Hypothesis testing

$$z = \frac{\hat{p} - p}{\sqrt{p\,q/n}}$$

Suppose the previous example is stated a little bit differently. Suppose the CEO claims that *at least* 80 percent of the company's 1,000,000 customers are very satisfied. Again, 100 customers are surveyed using simple random sampling. The result: 73 percent are very satisfied. Based on these results, should we accept or reject the CEO's hypothesis? Assume a significance level of 0.05.

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

  Null hypothesis: P >= 0.80

  Alternative hypothesis: P < 0.80

  Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected only if the sample proportion is too small.
- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. The test method, shown in the next section, is a one-sample z-test.
- **Analyze sample data**. Using sample data, we calculate the standard deviation (σ) and compute the z-score test statistic (z).

  σ = sqrt[P * (1 - P) / n] = sqrt [(0.8 * 0.2) / 100]

  σ = sqrt(0.0016) = 0.04

  z = (p - P) / σ = (.73 - .80)/0.04 = -1.75

  where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and n is the sample size.

  Since we have a one-tailed test, the P-value is the probability that the z-score is less than -1.75. We use the z-table to find P(z < -1.75) = 0.04. Thus, the P-value = 0.04.
- **Interpret results**. Since the P-value (0.04) is less than the significance level (0.05), we cannot accept the null hypothesis.

Examples of Hypothesis testing, t -value

An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline. From his stock of 2000 engines, the inventor selects a simple random sample of 50 engines for testing. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes. Test the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes. Use a 0.05 level of significance. (Assume that run times for the population of engines are normally distributed.)

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

  Null hypothesis: $\mu = 300$

  Alternative hypothesis: $\mu \neq 300$

  Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the sample mean is too big or if it is too small.
- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. The test method is a one-sample t-test.

- **Analyze sample data**. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

  SE = s / sqrt(n) = 20 / sqrt(50) = 20/7.07 = 2.83 (CLT)

  DF = n - 1 = 50 - 1 = 49

  t = (x - μ) / SE = (295 - 300)/2.83 = -1.77

  where s is the standard deviation of the sample, x is the sample mean, μ is the hypothesized population mean, and n is the sample size.

  Since we have a two-tailed test, the P-value is the probability that the t statistic having 49 degrees of freedom is less than -1.77 or greater than 1.77.

  We use the t Distribution Calculator to find P(t < -1.77) = 0.04, and P(t > 1.77) = 0.04. Thus, the P-value = 0.04 + 0.04 = 0.08.
- **Interpret results**. Since the P-value (0.08) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Examples of Hypothesis testing-chi sq.

Problem

Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans but not All-Stars, and 10% are veteran All-Stars.

Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  - Null hypothesis: The proportion of rookies, veterans, and All-Stars is 30%, 60% and 10%, respectively.
  - Alternative hypothesis: At least one of the proportions in the null hypothesis is false.
- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square goodness of fit test of the null hypothesis.

- **Analyze sample data**. Applying the chi-square goodness of fit test to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

  DF = k - 1 = 3 - 1 = 2
- $(E_i) = n * p_i$

  $(E_1) = 100 * 0.30 = 30$

  $(E_2) = 100 * 0.60 = 60$

  $(E_3) = 100 * 0.10 = 10$

  $X_2 = \Sigma [(O_i - E_i)_2 / E_i]$

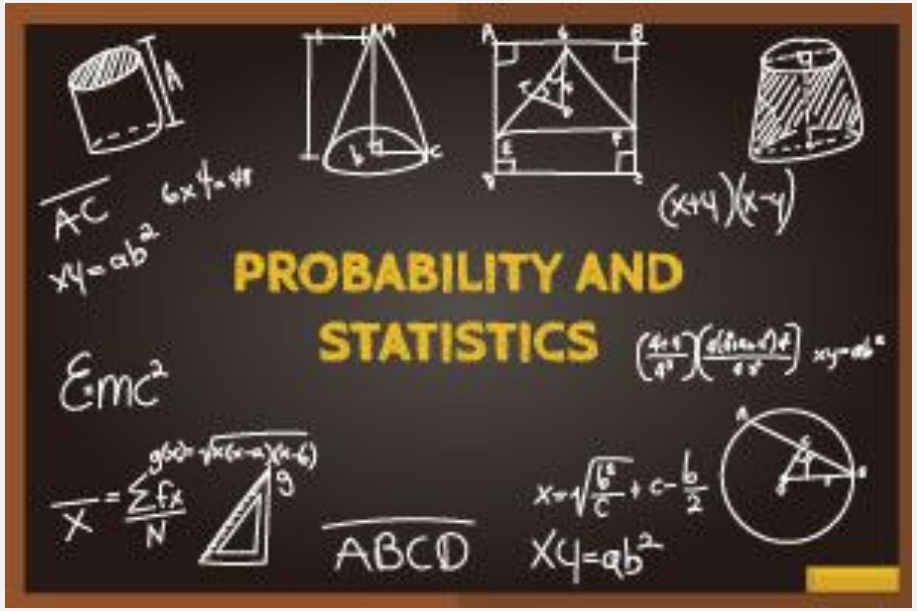
  $X_2 = [(50 - 30)_2 / 30] + [(45 - 60)_2 / 60] + [(5 - 10)_2 / 10]$

  $X_2 = (400 / 30) + (225 / 60) + (25 / 10) = 13.33 + 3.75 + 2.50 = 19.58$

  where DF is the degrees of freedom, k is the number of levels of the categorical variable, n is the number of observations in the sample, E_i is the expected frequency count for level i, O_i is the observed frequency count for level i, and X_2 is the chi-square test statistic.

  The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.58.

  We use the Chi-Square Distribution Calculator to find $P(X_2 > 19.58) = 0.0001$.
- **Interpret results**. Since the P-value (0.0001) is less than the significance level (0.05), we cannot accept the null hypothesis.



Thank You!

BeingDatum.com
contact@beingdatum.com