Hypothesis testing & ANOVA

What, why & when

Lets outline solutions to few of these problems

How much stock of apple juice should I order for this week for powai location

By what amount should we increase credit card limits for this group of customers

Should BCCI select Ishant Sharma for Australia tour considering his poor performance in last 4 test matches in India

How many litres of water should we give to each household?

TATA salt plant manager has to investigate complaints of lower weights in packs

Should we provide additional tuitions for language courses to science students

Structure of this lecture

Before starting with Hypothesis Testing

- We'd discuss few building blocks and few building blocks for the building blocks.... Its not as confusing as it sound.
- Building blocks for building blocks:
  - Population & Sample
  - Histogram
  - Distributions
- Building blocks for hypothesis testing
  - Probabilities and cumulative probabilities
  - Standardization
  - Standard normal distribution and how it helps
  - Central Limit Theorem

What's to follow :

We'll talk about a couple of seemingly disconnected sections, but towards the end things will start to fall in place. So, don't worry about the disconnect, just make sure that you understand these sections/topics individually. Lets begin!

Population & Samples : Population

- Population is where data is coming from. Population is practically infinite . You never get to know all aspects of the population “exactly”.
  - Age of graduating engineers in 2012 across india from various colleges
  - Money spent by individuals on buying gifts on christmas across country in stores of shopper’s stop

Population & Samples : Sample

- Sample is subset of population, which might or might not be representative of entire population. [But its good if it is]
  - Data on age of 1000 graduating students from 100 colleges across india
  - Data on money spent for 1573 customers from 10 selected stores across country

& Estimates....

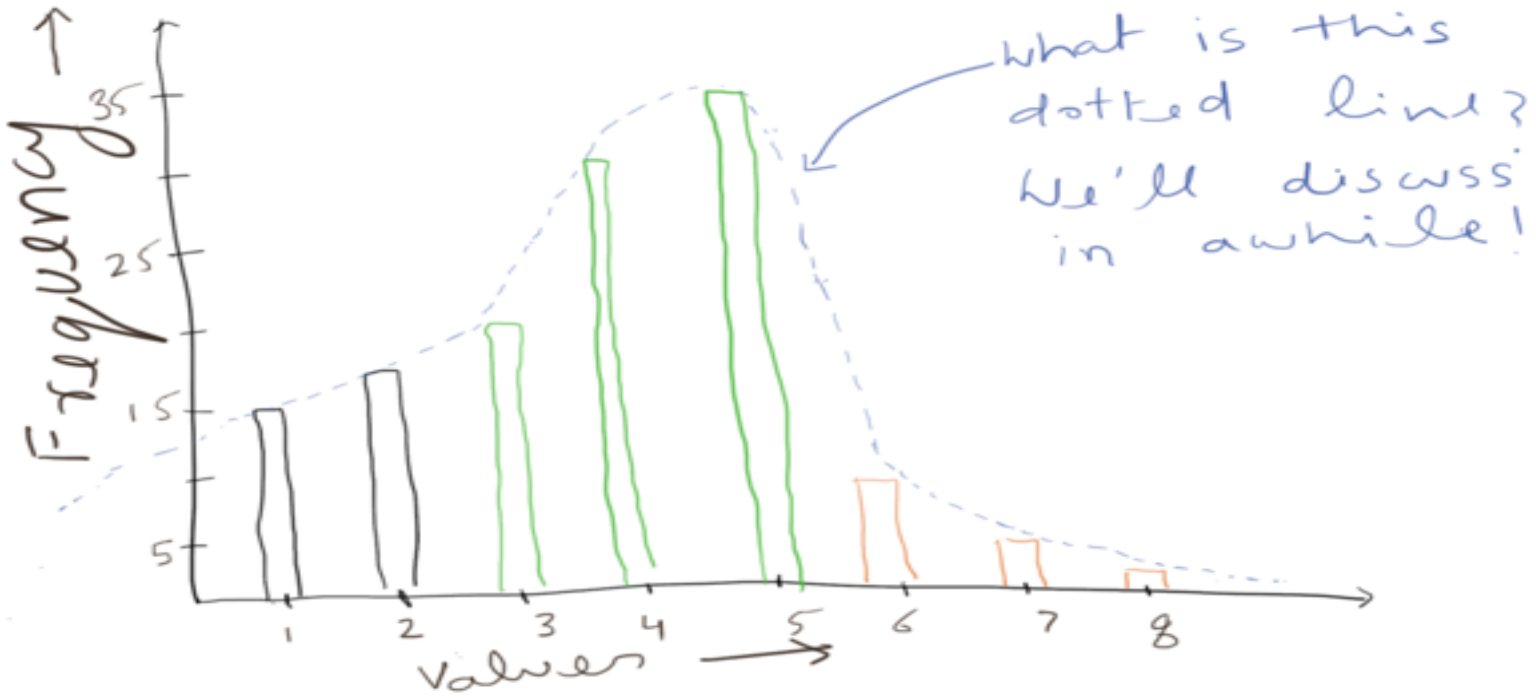
We’d never know “exact” average age or its standard deviation and other summary statistics for entire population of variable age ,coming from “all” graduating engineers. Same for the money spent data from stores.

Estimates contd..

- However if our “samples” are representative of the population they can be used to take a “guess” at the summary statistics of the “population”. This is called estimation.
  - Every such estimate would have errors [deviation from the real “exact” value]
  - If we draw another such sample , you’ll get different value of the estimates.

Histograms

- They are simple bar charts in the most simplistic sense. Like this

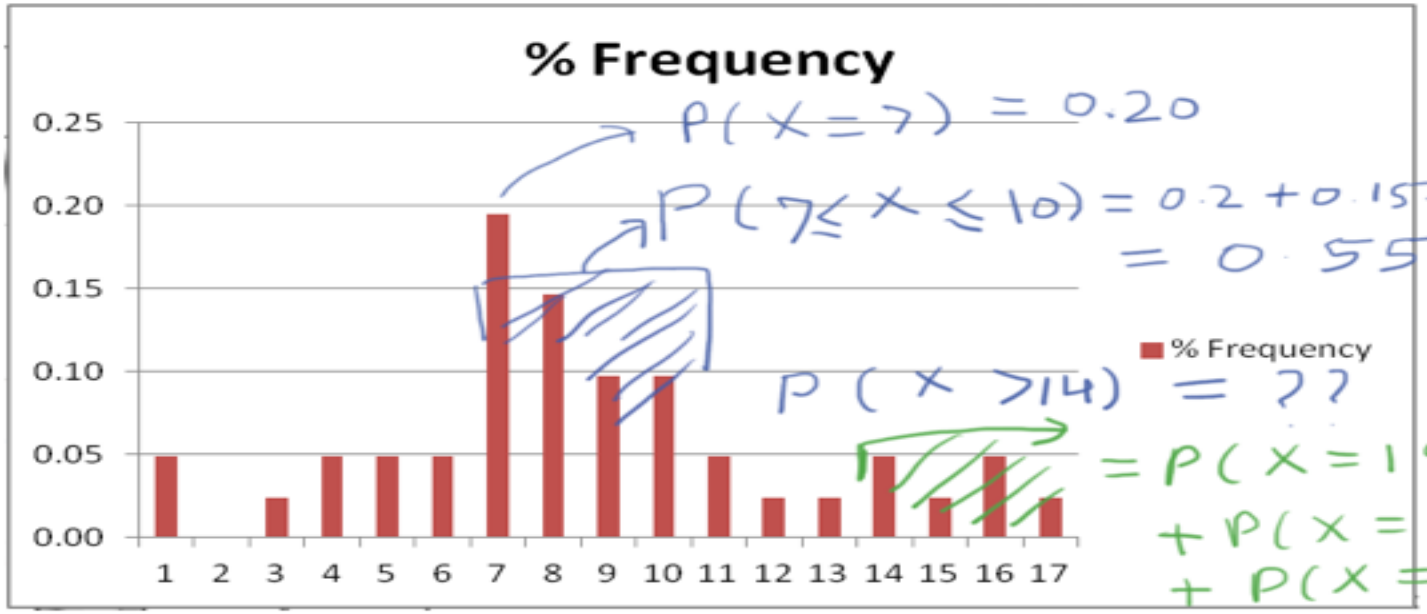


Percent frequency bar charts, Probabilities & cumulative Probabilities

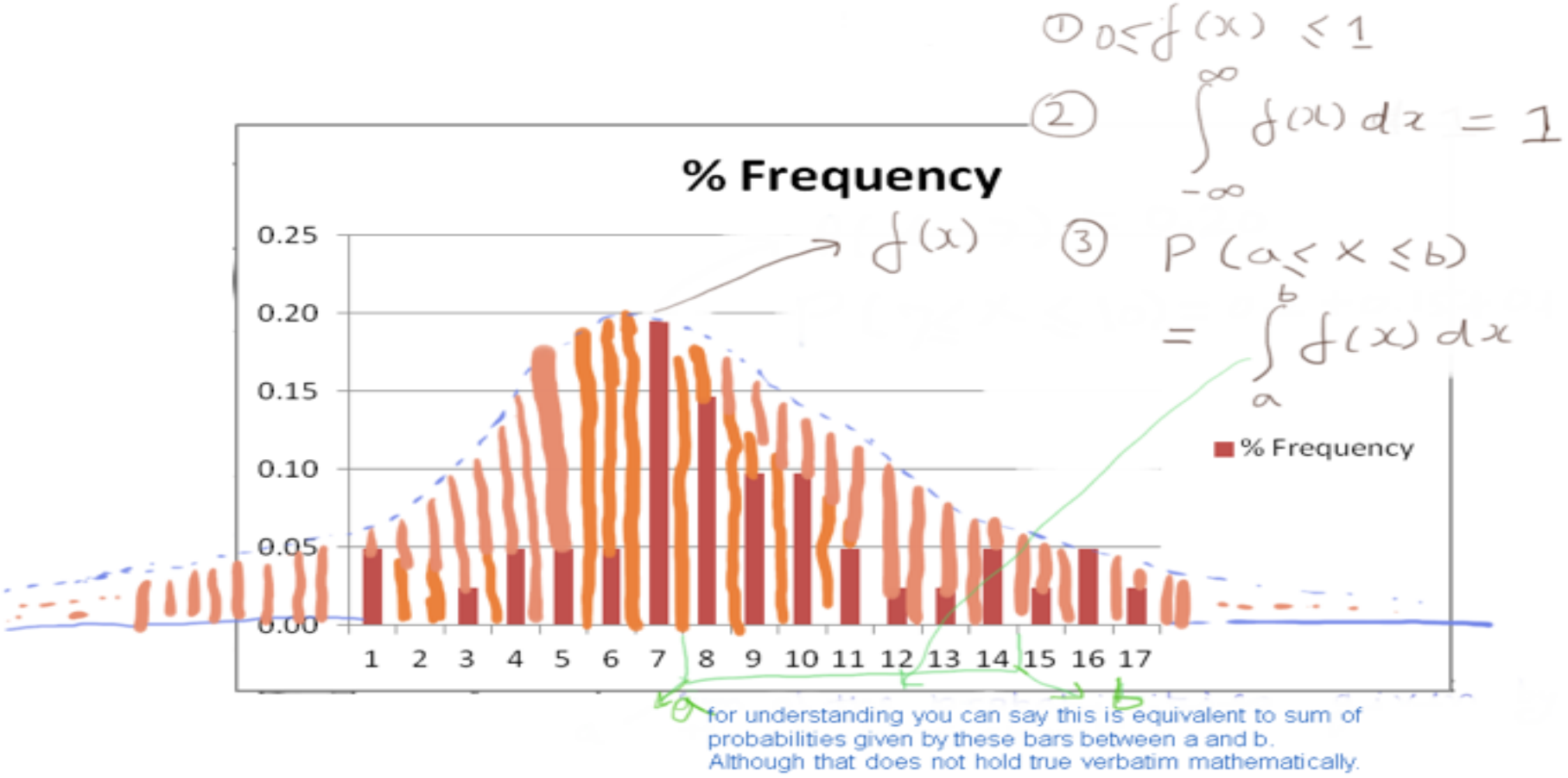
Value	Freq	% Frequency
2	2	0.05
3	0	0.00
4	1	0.02
5	2	0.05
6	2	0.05
7	2	0.05
8	8	0.20
9	6	0.15
10	4	0.10
11	4	0.10
12	2	0.05
13	1	0.02
14	1	0.02
15	2	0.05
16	1	0.02
17	2	0.05
18	1	0.02

Percent frequency bar charts, Probabilities & cumulative Probabilities

Quick Quiz => $P(1 \leq x \leq 17) = ???$

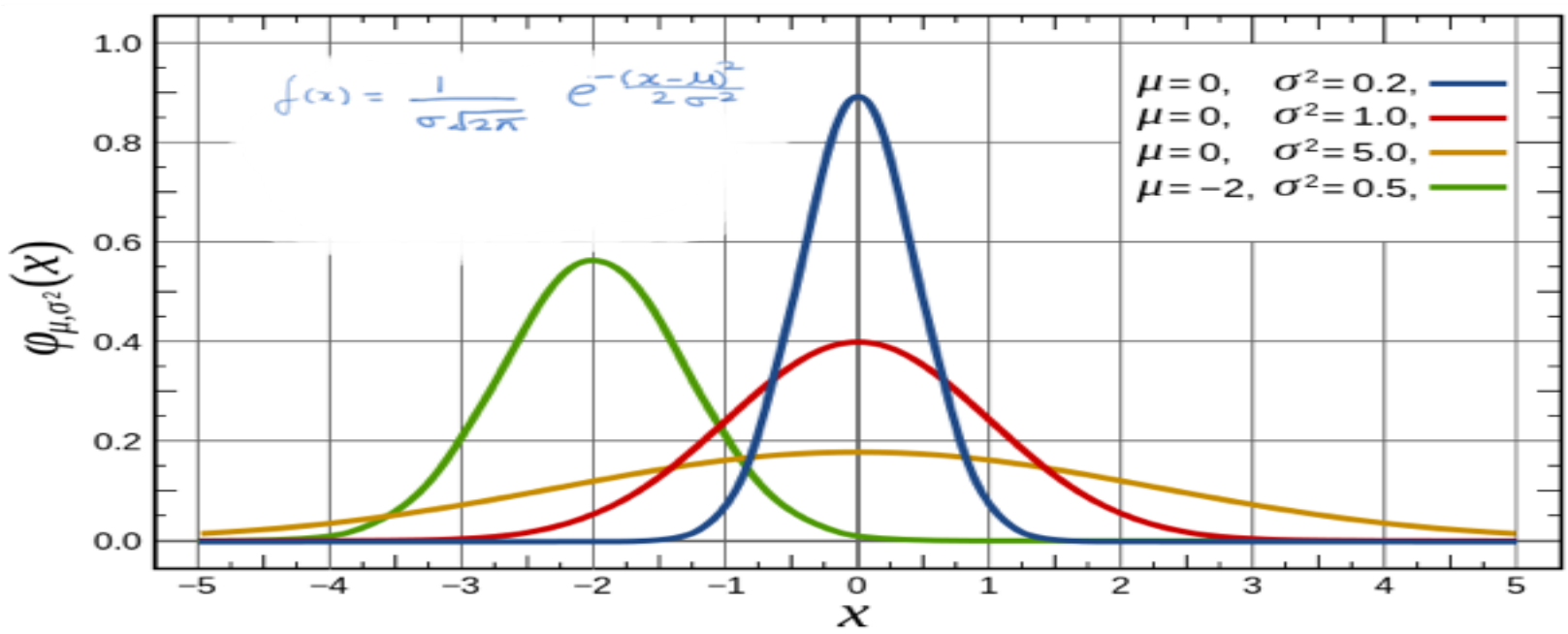


Distributions

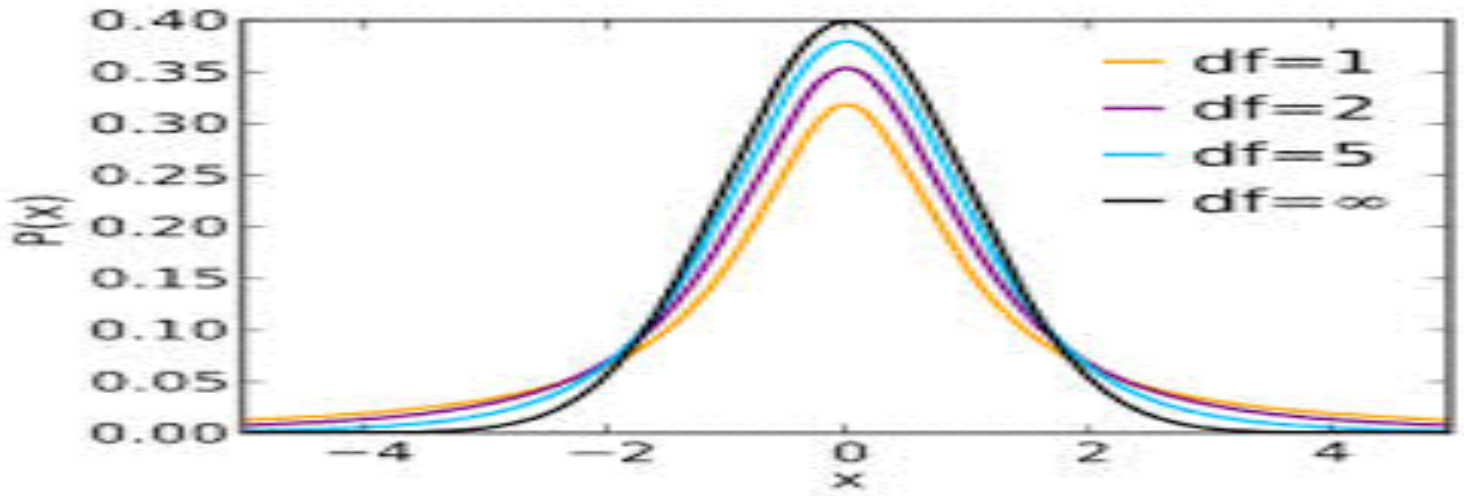


Normal distribution

- It is one of the very common f(x) which is a good fit/approximation for distribution of majority of the data that we come across . It has two parameters: mean, variance.



T-distribution



➢T distribution is symmetric bell shaped just like normal distribution.

➢t distribution with sample size = ∞ or large sample size corresponds to N(0, 1) distribution

Standardization

- Process of standardization makes mean of your variable 0 and standard deviation =1

- $X_{standard} = (X-\mu)/\sigma \mid X = X_{standard} * \sigma + \mu$

Standard Normal Distribution

- The curve in red in previous picture is standard normal distribution (Z)

- standard results from standard normal distribution can be extrapolated back to those other set of normal distributions as well.

- It has equation : $$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

Few standard results from standard normal distribution

- $P(Z \geq 0) = P(Z \leq 0) = 0.50$
- $P(-1 \leq Z \leq 1) = 0.682$
- $P(-2 \leq Z \leq 2) = 0.954$
- $P(-3 \leq Z \leq 3) = 0.996$

Some more..

- Since the distribution is symmetric, one sided probabilities as well as remainder probabilities can be calculated easily from these results

- $P (-1.645 <= Z <= 1.645) = 0.90$
- $P(-1.96 <= Z <= 1.96) = 0.95$
- $P (-2.576 <= Z <= 2.576) = 0.99$

Confidence Intervals

Deriving from standard results that we just saw we can say that 90% of the variable values coming from standard normal distribution lie in the range [-1.645 , 1.645]

Confidence Intervals contd..

- Now say a variable follows normal distribution with mean 2 and standard deviation 0.5 then the same 90% range can be derived from the above.

  - X_1 = -1.645 * 0.5 +2 = 1.177
  - X_2 = 1.645 * 0.5 +2 = 2.822
  - For this 90% of the values lie between [1.177, 2.822]

Answer to: why the hell people go gaga over normal distribution!

Central Limit Theorem

Averages of samples [of sufficient large size] follow normal distribution with mean μ and variance σ^2/n where (μ , σ^2) are mean and variance of the population and n is the sample size. If sample size is small, normal distribution is replaced by t-distribution. [This is irrespective of distribution being followed by the variable values].

Intro to Hypothesis Testing

Hypothesis testing a simple statistical framework to check your otherwise “informed guesses” or “hunches”. For example:

An Example

We want to check whether average marks in mathematics received by 12^{th} student in board exams belonging to CBSE board in Maharashtra is 85 or not. We drew a sample of 100 students which had average 80 and standard deviation 10. Does that tells us that average is less than 85?

How CLT comes into play : Sample Averages ~ N

we know that the marks themselves could be following any distribution but if we draw different samples their averages would follow normal distribution with mean 85 and standard deviation 10/sqrt(100) [=1].

How CLT comes into play : Constructing Hypothesis

Here pre established fact is that the average is 85, we have great faith in this pre established fact and call it null hypothesis. Now alternate hypothesis is that average is less than 85 which we are going to examine through a sample, this is called alternate hypothesis

How CLT comes into play : Formal Notation

- $H_0 \Rightarrow \mu = 85$
- $H_A \Rightarrow \mu < 85$

- Understanding this in terms of distributions, null hypothesis says that these sample averages follow a normal distribution with mean 85 whereas Alt-Hypo says they follow a normal distribution with mean less than 85.

One and Two tailed tests

$z = \frac{80-85}{1} = -5$

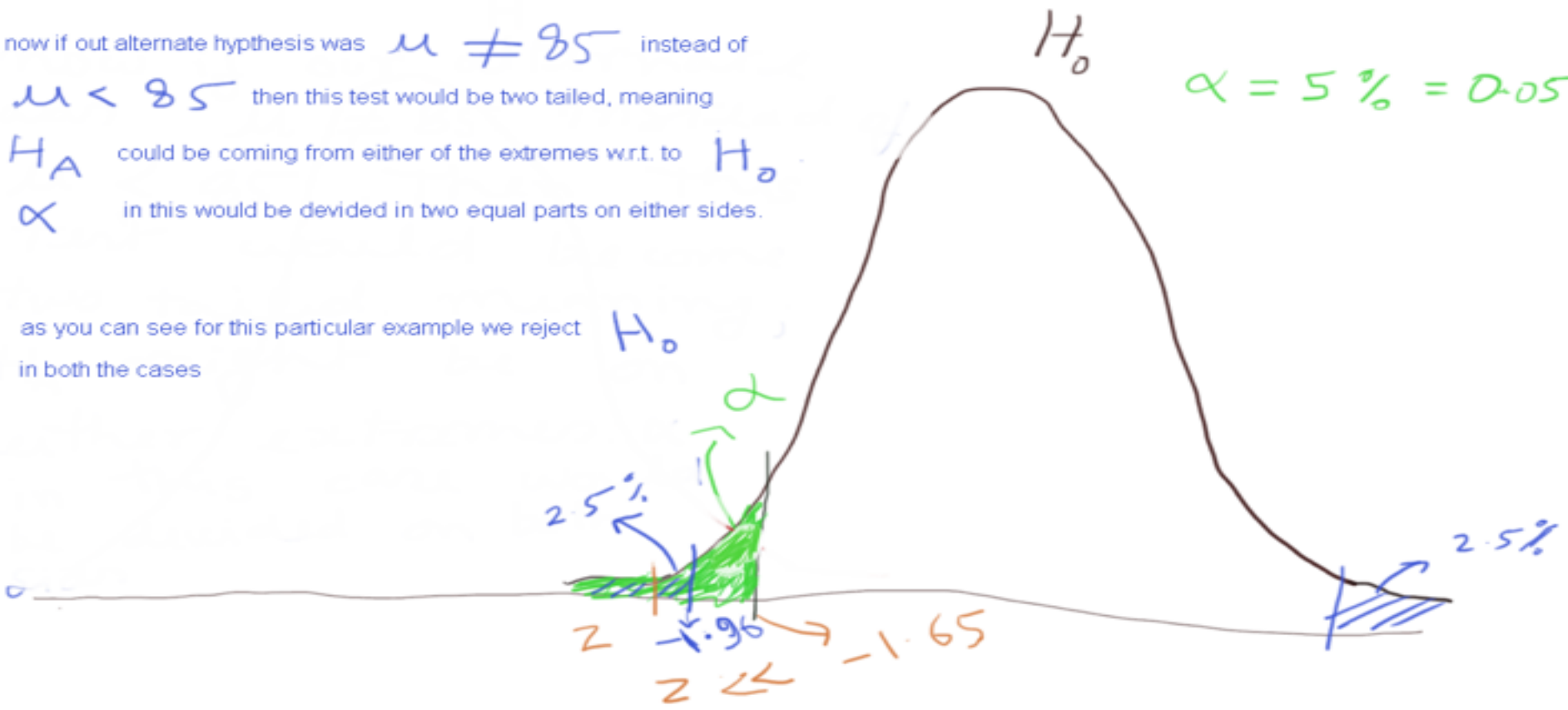
now if out alternate hypthesis was $\mu \neq 85$ instead of

$\mu < 85$ then this test would be two tailed, meaning

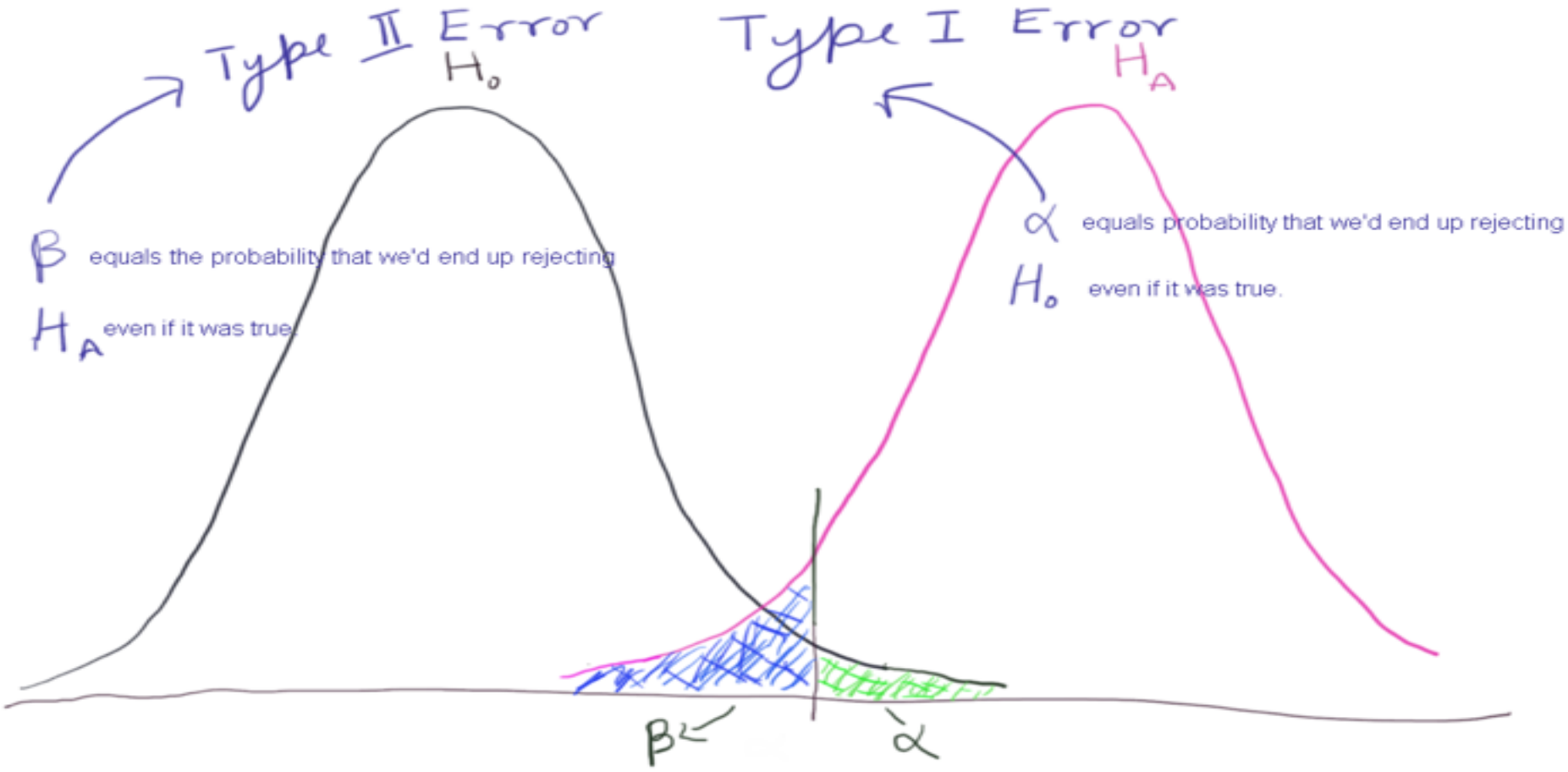
H_A could be coming from either of the extremes w.r.t. to H_0

α in this would be devided in two equal parts on either sides.

as you can see for this particular example we reject H_0 in both the cases



Type I and Type II errors



One Sample Ttest

Revisting One Sample Test

- To carry out the test we used a test statistic
  - (sample mean- population mean/sample std)

  - This was nothing but standardized sample mean and we compared this to values obtained from standard normal distribution

Two Sample Test

- Here we use similar test statistic for the difference of the sample mean.
- Null hypothesis being that samples are coming from same population, population mean of the differences is zero.
- Sample std is replaced by pooled sample std

Two Sample Ttest

ANOVA: Many Sample Problem

Ho: All Group Means are Equal, i.e. $\mu_1 = \mu_2 = \ldots = \mu_k$

H_A: At least one population mean is not equal to the rest

- **The test is developed under the assumption that variances across groups are equal.**

Sum of squares

- Within group sum of squares: [SSW]
  - Individual sum of squares of differences for each group separately

- Between Group sum of squares: [SSB]
  - Sum of squares of differences between mean of each sample from the **grand mean**

Sum of Square Contd..

- **SST = SSB + SSW**

- SSW will give us within group variances and SSB will give us between group variance.

- if groups are different from each other; SSW would be small but SSB would be large

ANOVA Test Statistic

$$F = MSB / MSE = (SSB / k-1) / (SSE / n1+n2+n3....nk)$$

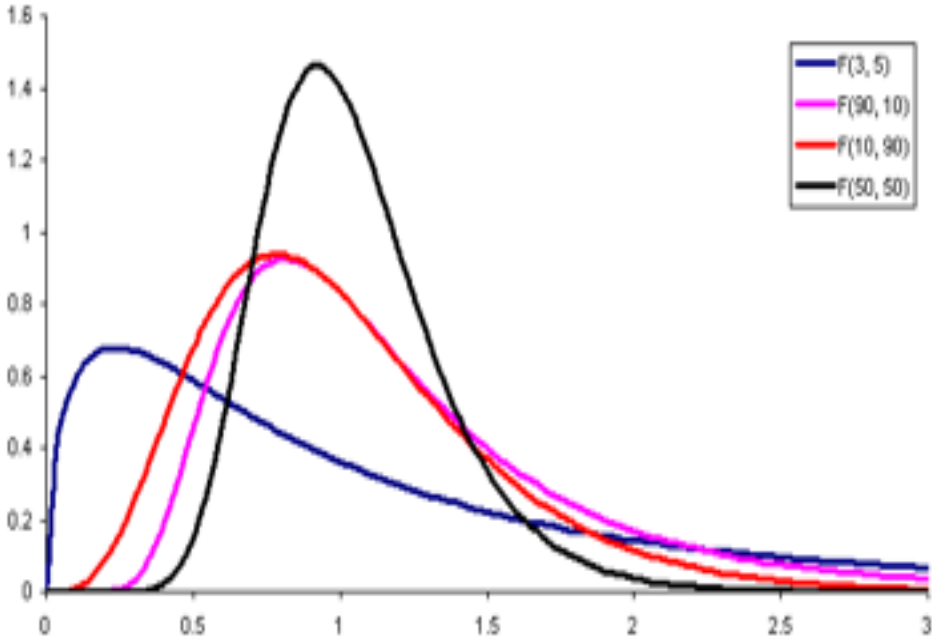
where

MSB : Mean sum of squares between groups

SSE : Mean square error

ANOVA Rejection Region

- Test statistic $F \sim F(DF_B, DF_E)$
- F distribution is characterized by two degrees of freedom
- F distribution is positively skewed and takes only positive values



Anova Test

- **Null hypothesis of equality of means is rejected if observed value of test statistic is too large**
- **Rejection region is right-sided only**
- **MSB is large compared to MSE only if all groups means are not identical**

Applications of One-way ANOVA

- A company has three manufacturing plans and wants to determine whether there is a difference in the average age of workers in the three locations.

- Salaries of freshers entering analytics in India varies according to locations. A multi-city survey on freshers from Gurgaon, Kolkata, Bengaluru, Mumbai and Hyderabad is undertaken to verify authenticity of this claim.

ANOVA

Categorical Variable

- Distribution of categorical variable's frequencies can be affected by another categorical variable's categories

- Looking at cross tables gives you an idea whether this is the case

Chisq test