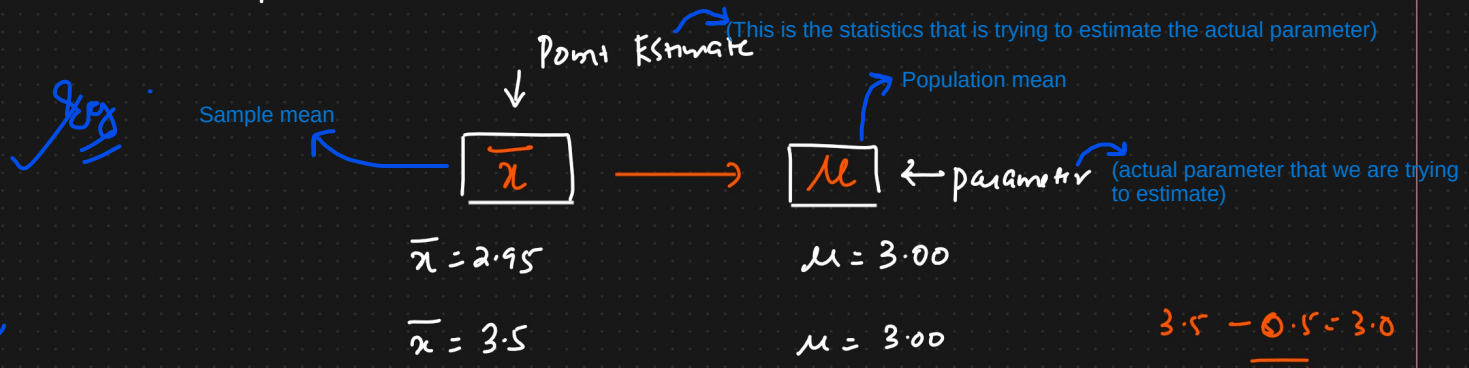
Agenda

1. Point Estimate ✓
2. Range of C.I ✓
3. Chi Square distribution ✓
4. F distribution. → F test ✓
5. [ANOVA] → Assignment ⇒ more than 2 groups ⇓ F Test

1. Point Estimate

Defn: The value of any statistic that estimates the value of a parameter (actual) is called point Estimate



Now the question is whether the estimate that we are making can be accurate? Answer is no since the estimate that we are trying to make (based several tests/hypothesis testing) can either be greater than or less than the actual value. This is the reason why we define the range of point estimate using Confidence Interval

We rarely know if our point Estimate is correct because it is an estimation of the actual value $\bar{x}$

We construct C.I to help estimate what the actual value of unknown population mean is.

Mathematically C.I range of point estimate is defined using margin of error

Point Estimate ± Margin of Error

Lower Range C.I = Point Estimate − Margin of Error

Higher Range C.I = Point Estimate + Margin of Error

How can we calculate this margin of error?
Ans: Using Ztest and Test. We can below numerical as an example for the same

① On the Verbal section of the CAT exam, a sample of 25 test takers has a mean of 520. with a standard deviation of 80. Construct a 95% C.I about the mean?

Ans) $\bar{x} = 520$ $n = 25$ $\boxed{s = 80}$ C.I = 0.95 $\alpha = 0.05$

95% ⇓ C.I

Note here since std dev for sample is given hence we will be using Ttest.

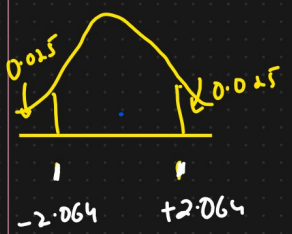
C.I = Point Estimate ± Margin of Error

[425 - 564]

$$= \bar{x} \pm \boxed{t_{\alpha/2}} \left(s/\sqrt{n} \right) \Rightarrow \text{Standard Error}$$

representing Tscore at significance level for 2 tail test

0.05



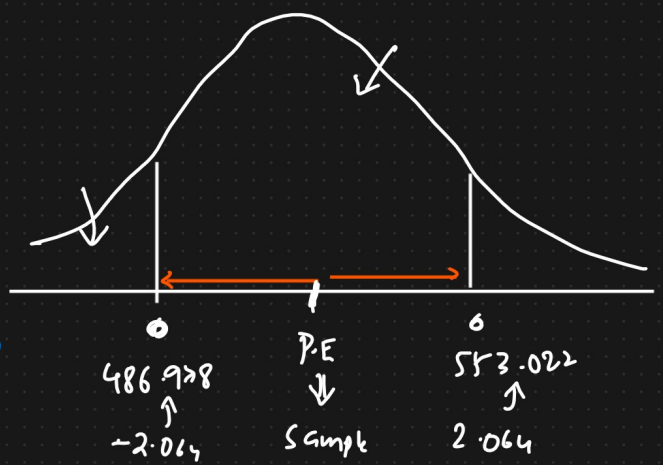
$$= 520 \pm 2.064 \left[80/\sqrt{25} \right]$$

degree of freedom : n-1 = 24

= 24 //.

Lower C.I : $520 - 2.064 \left[80/5 \right] = 486.978$ //.

Higher C.I : $520 + 2.064 \left[80/5 \right] = 553.022$ //



Simple fanda is that it is wrong to say that any statistics say mean for an estimate (eg sample mean) will be accurate. This is the reason why we are defining the range within which we can say mean is right estimate for population and outside which we can say that mean is not the right estimate of the population.
Eg: Event manager suggesting the number of plates to order. Compare with the numerical that we just did assuming manager has data of say 10 groups(sample) where at least 100 people attended and number of food plates used

Similarly we can determine the point estimate interval for a 2 tail test where Population standard deviation is given using Ztest. Simply in the margin of error formula we will replace Tscore for significance level with Zscore and sample std dev with population dev:

refer:
https://prwatech.in/blog/data-science/margin-of-error-tutorial/

mean

② CHI SQUARE TEST

The chi square Test for Goodness of fit tests claims about population proportion.

Categorical variables which can be either ordinal or nominal

[categorical variables]

It is a non parametric test that is performed on categorical data

[ordinal, nominal data].

I case of normal distribution we can make predictions about parameters like mean just by seeing the graph hence it is parametric. But in case non non gaussian distribution we are not able to make predictions easily that is why they are considered as the non parametric. Like we wont be able to predict what could be possible range of mean std dev etc....

To be more precise non parametric test is the test where we are not assuming anything about the population

Eg: There is a population of Male who likes different Color of Bike

	Theory	Sample
Yellow Bike	1/3	22
Orange Bike	1/3	17
Red Bike	1/3	59

=> Goodness of fit.

=> Observed categorical distribution

(This represents the sample. We would have gone to a group of individuals and have asked about their favorite color bike)

Now using Chi Sqaure test, using observer categorical dist(sample) we are going to decide whether the claim made by Theoretical categorical dist is correct or not(goodness of fit)

⇓

Theoretical categorical distribution

(This represents the population. Also this is in the proportion hence that's why in definition we used word population proportion)

①

In 2010 Census of the city, the weight of the individuals in a small city were found to be the following

<50kg	50 - 75	>75
20%	30%	50%

Population

group!

See population data is given in proportion whereas the sample data in whole numbers for categorical feature. This is how we are going to identify that we would need to use Chi Square Test

In 2020, weight of n=500 individuals were sampled. Below are the results

<50	50-75	>75
140	160	200

Sample

Using $\alpha = 0.05$, would you conclude the population difference of weights

has changed in last 10 years?

Ans)

In 2010

Expected

<50kg	50-75	>75
20%	30%	50%
⇓ 0.2	⇓ 0.3	⇓ 0.5

In 2020

n=500

Observed

<50	50-75	>75
140	160	200

In 2010

Expected

<50kg	50-75	>75
500 × 0.2 = 100	500 × 0.3 = 150	500 × 0.5 = 250

① Null Hypothesis H_0 : The data meets the expectation

Alternate hypo H_1 : The data does not meet the expectation.

② $K = 0.05$ C.I = 95%

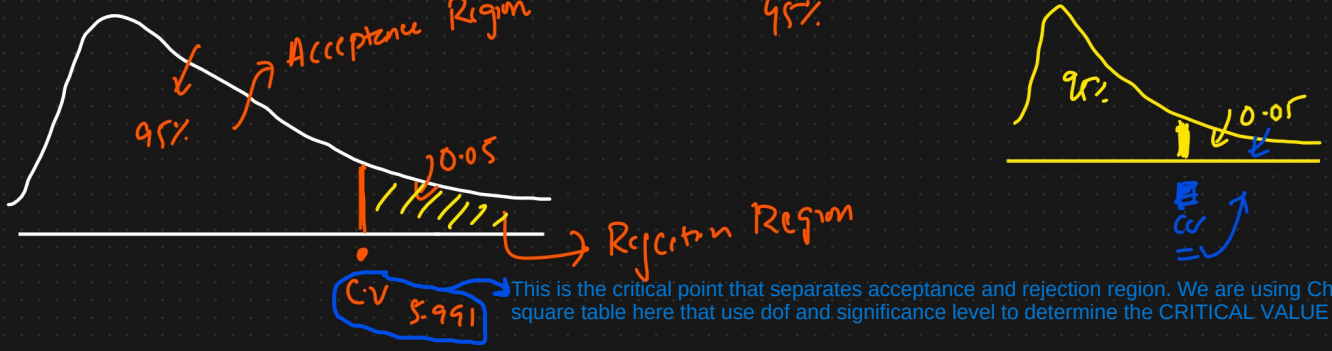
③ Degree of freedom

$dof = K - 1 = 3 - 1 = 2$

Here in dof for a Chi Square test, K represents the number of distinct categorical groups.

④ Decision Boundary

Just check in Wikipedia how the Chi Square pdf function looks like. You will observe that generally it is right skewed. Also at K=3, Chi Square distribution is right skewed. This is the reason why decision boundary is one tail and that too skewed positively



This is the critical point that separates acceptance and rejection region. We are using Chi square table here that use dof and significance level to determine the CRITICAL VALUE

Chi square Test $\chi^2 > 5.991$ { Reject the Null Hypothesis }.

Now we will be calculating the the Chi Square test statistics(represented by Xsqaured) and if it is greater than critical value calculated above then in that case we would reject the null hypothesis that we have stated.

⑤ Calculate Chi Square Test Statistics

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O represents the observed(sample) and E represents the expected(Population)

$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$\chi^2 = 26.67$$

Means data does not meet the expectation. If data has meet the expectation, then we would have said that according to question weights have not changed much in the last 10 years. But since here we have rejected the null hypothesis which mean that the data does not meet the expectation, which means data of weights have changed in the last 10 years. Since, Chi Test is applied on +ve skewed (generally) hence this weight change is in the +ve direction meaning that the weights have increased in 10 years which can be due to technological advancements and more people doing WFH

$\therefore \chi^2 > 5.991$ { Reject the Null Hypothesis }.

⑥ F-distribution

This distribution is generally used to compare the 2 groups(S1 and S2 defined below). If we are trying to compare more than 2 then it becomes the ANOVA

The F-distribution with d_1 and d_2 degree of freedom is the distribution of

$$\chi = \frac{S_1/d_1}{S_2/d_2}$$

Here S means sample

$S_1 \rightarrow$ Independant random variables

$S_2 \rightarrow$ Independ random Variable

These samples are following the chi square distribution } Chi square distribution

$d_1 \rightarrow$ Degree of freedom (S1)

$d_2 \rightarrow$ Degree of freedom (S2)

Here we are just trying to understand the F distribution and this mathematical intuition has been picked up from the Wikipedia

F test $\rightarrow$ Variance Ratio Test { Comparing the variance between 2 groups }.

How to Identify Ftest need to be applied?
F test is used to compare the variance of 2 groups(independent variables). If we want to compare the variance of more than 2 groups then we will be using the ANOVA test

F Test [Variance Ratio Test].

Understanding the Ftest with the numerical

① The following data shows the no. of bulbs produced daily for some days by 2 workers A and B

A	B
40	39
30	38
38	41
41	33
38	32
35	39
	40
	34

Can we consider based on the data or not Worker B is more stable and efficient

$\alpha = 0.05$ 95% C.I.

If spread of data for B(variance) is more as compared to A then that means that B is more efficient than A else not. Also, if spread is same for both the distribution then we would say that both are equally efficient. This is how we are making use of variance to basically compared that 2 independent groups/samples

Ans) Null Hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$

Alternate Hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

A

X_1	$\bar{X}$	$(x_i - \bar{x})^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
$\bar{X}_1 = 37$		$\Sigma(x_i - \bar{x})^2 = 80$

Here mean for both the distribution are same. This is the reason why we are using variance as the measure to compare both the groups

B

X_2	$\bar{X}_2$	$(X_2 - \bar{x})^2$
39	37	4
38	37	1
41	37	16
33	37	16
32	37	25
39	37	4
40	37	9
34	37	9
$\bar{X}_2 = 37$		84.

$$S_1^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{80}{5} = 16.$$

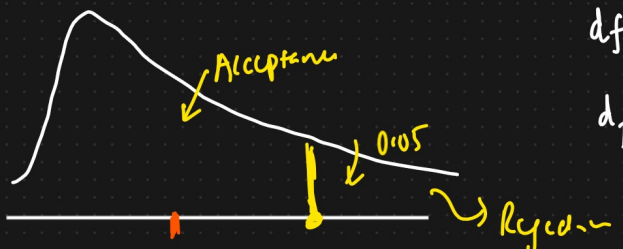
$$S_2^2 = \frac{84}{7} = 12$$

S1 and S2 represents the sample variance with baseian correlation

⊛ Variance Ratio. [F-test]

$$F = \frac{S_1^2}{S_2^2} = \frac{16}{12} = 1.33$$

⊛ Decision Rule [F distribution]



$df_1 = 6-1 = 5$

$df_2 = 8-1 = 7$

df1 and df2 is the degree of freedom for 1st and 2nd sample which is worker A and worker B

$\alpha = 0.05$

3.9715

Here to calculate the critical value we are using F table which use dof1, dof2 and significance level for the calculation of the same.

If, F-test > 3.9715 {Reject H_0}.

Is $1.33 > 3.9715$ ⇒ False

We fail to Reject the Null Hypothesis

Worker A ≈ Worker B

Here, we fail to reject the null hypothesis(means that we are accepting the null hypothesis) which signifies that both the worker A and worker B are equally efficient.