Day10

LetsUpgrade class notes

LIVE 262

Process of Hypothesis Testing

Formulate the Null Hypothesis and the alternative hypothesis.

↓

Select the appropriate test statistic.

↓

Choose the level of significance, α, and the Degree of Freedom

↓

Compute the calculated test value of the test statistic

↓

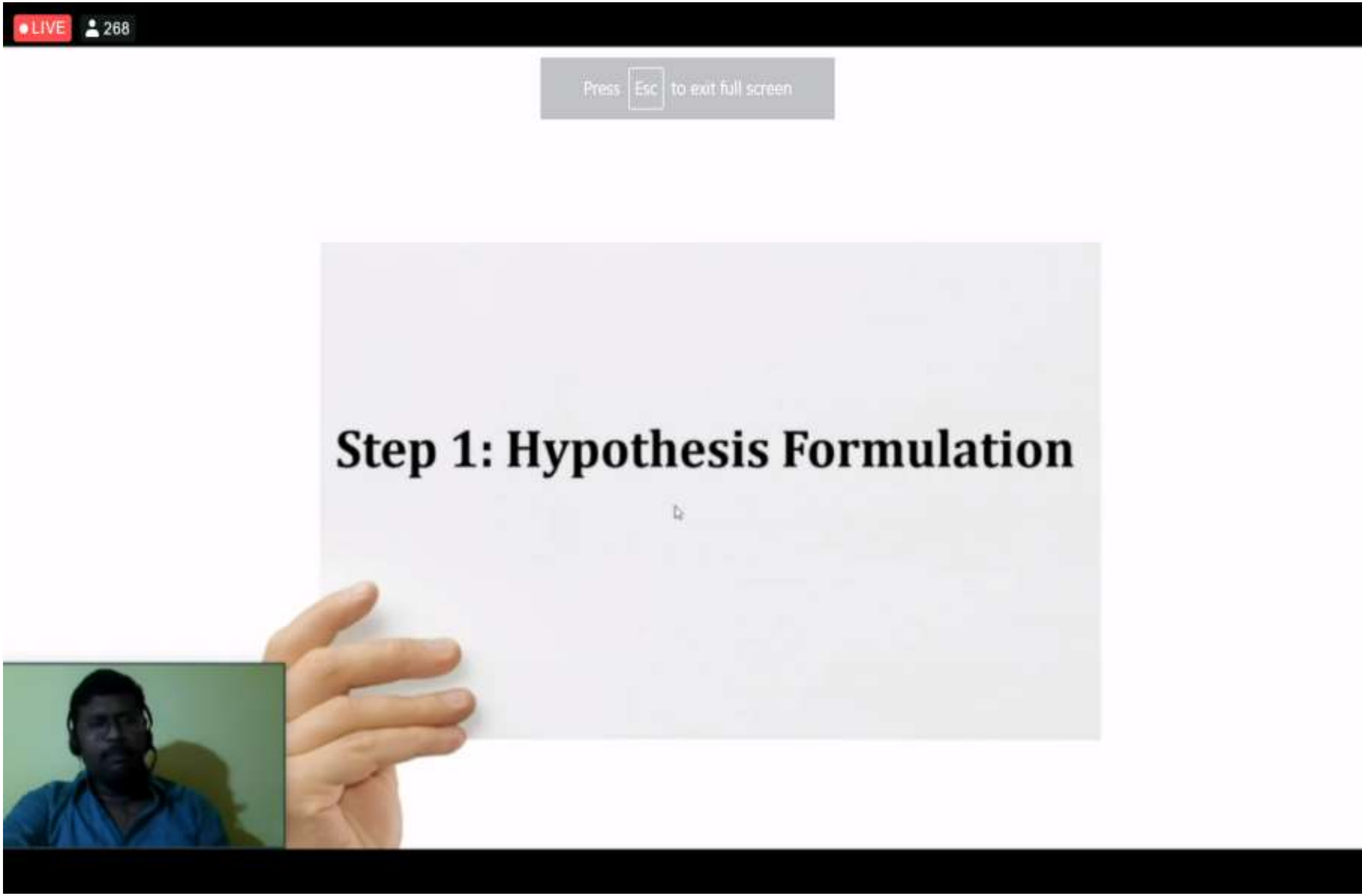
Compute the table test value of the test statisitc

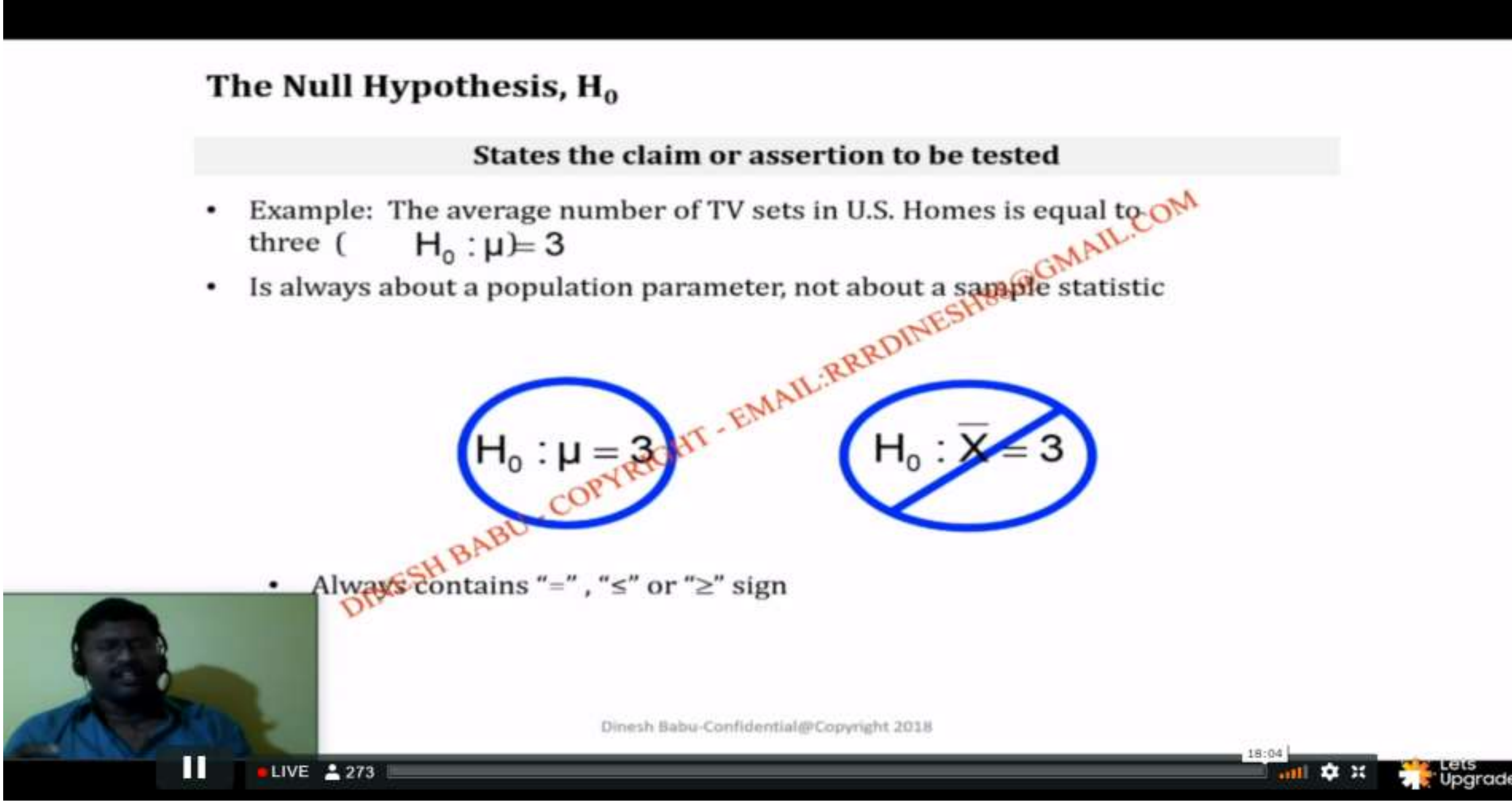
↓

Compare the calculated values and table values

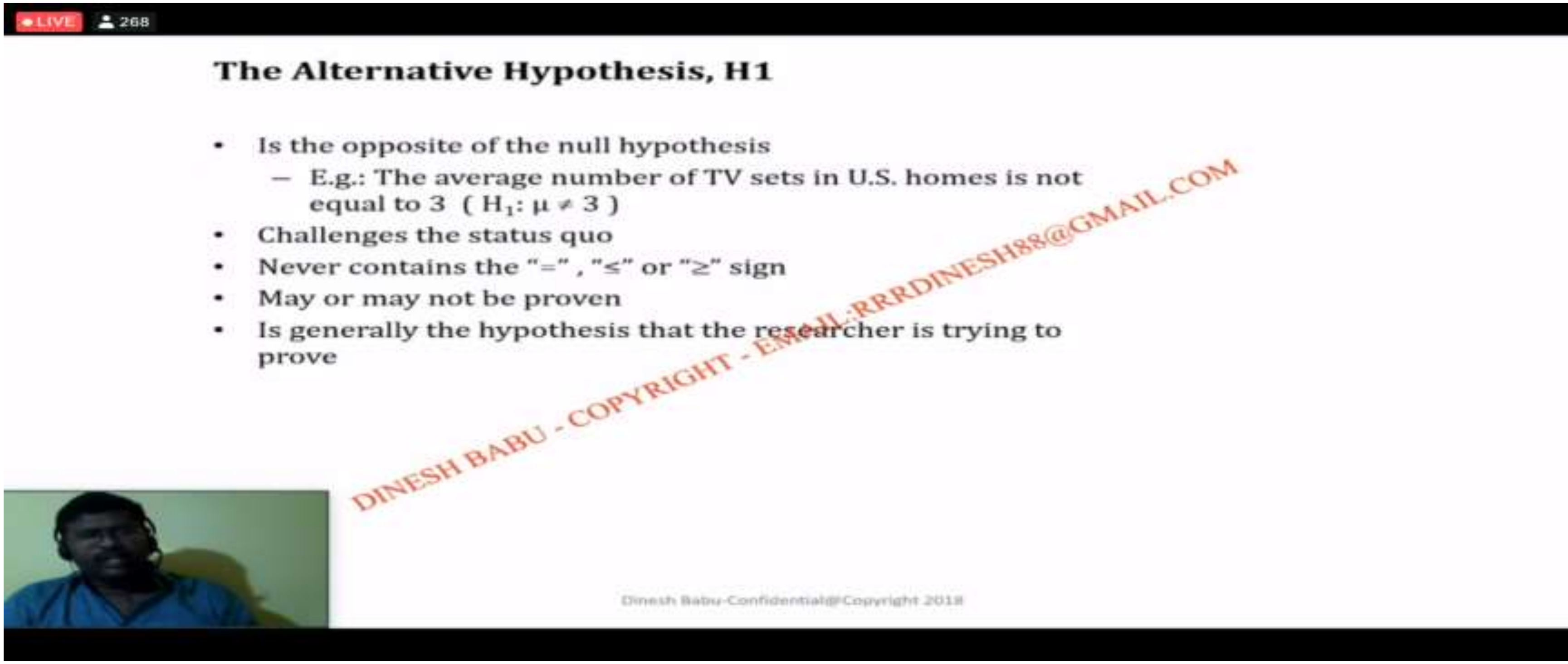
↓

Make the statistical decision and state the managerial conclusion.


LIVE
268
Press Esc to exit full screen
Step 1: Hypothesis Formulation

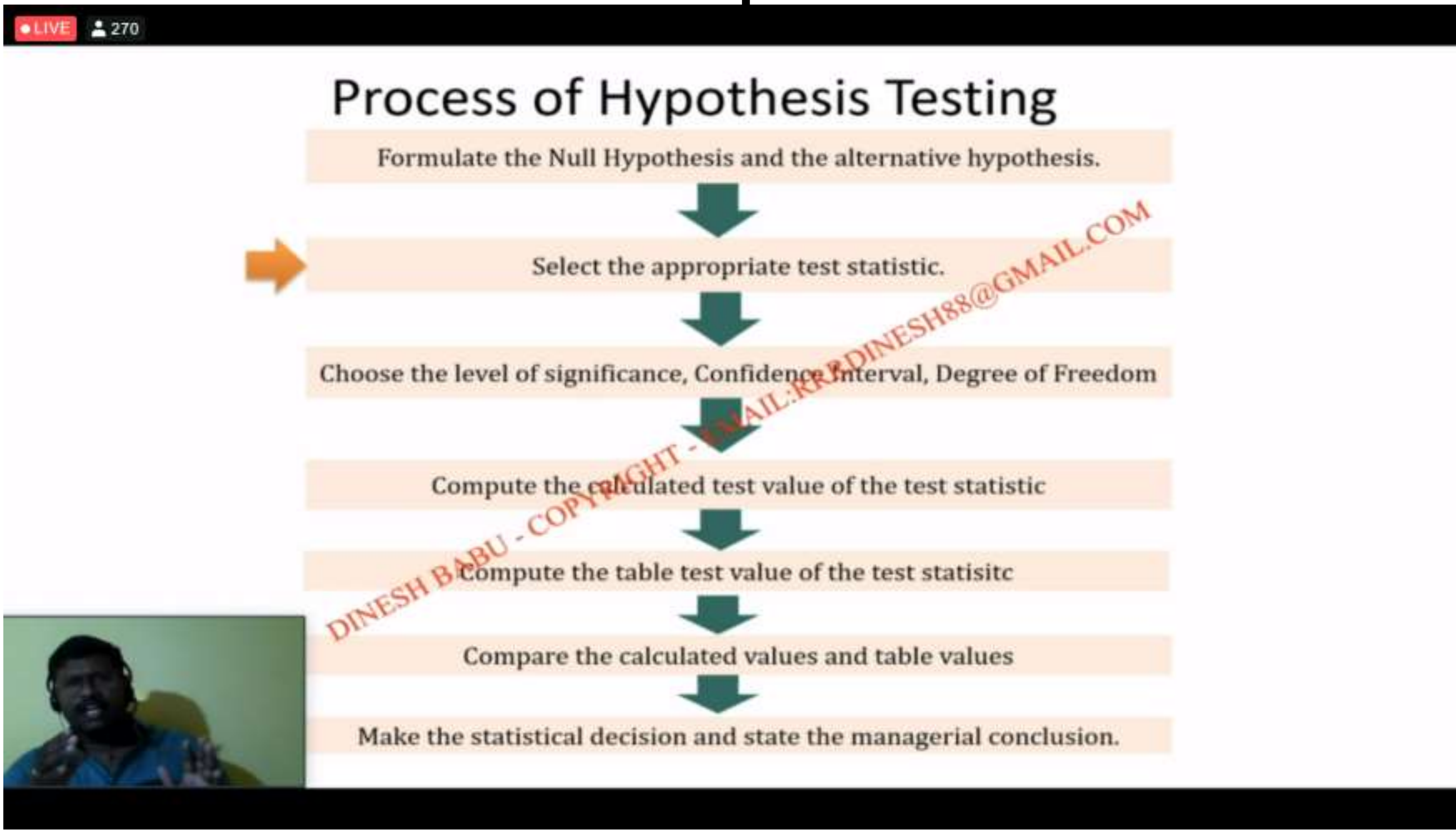


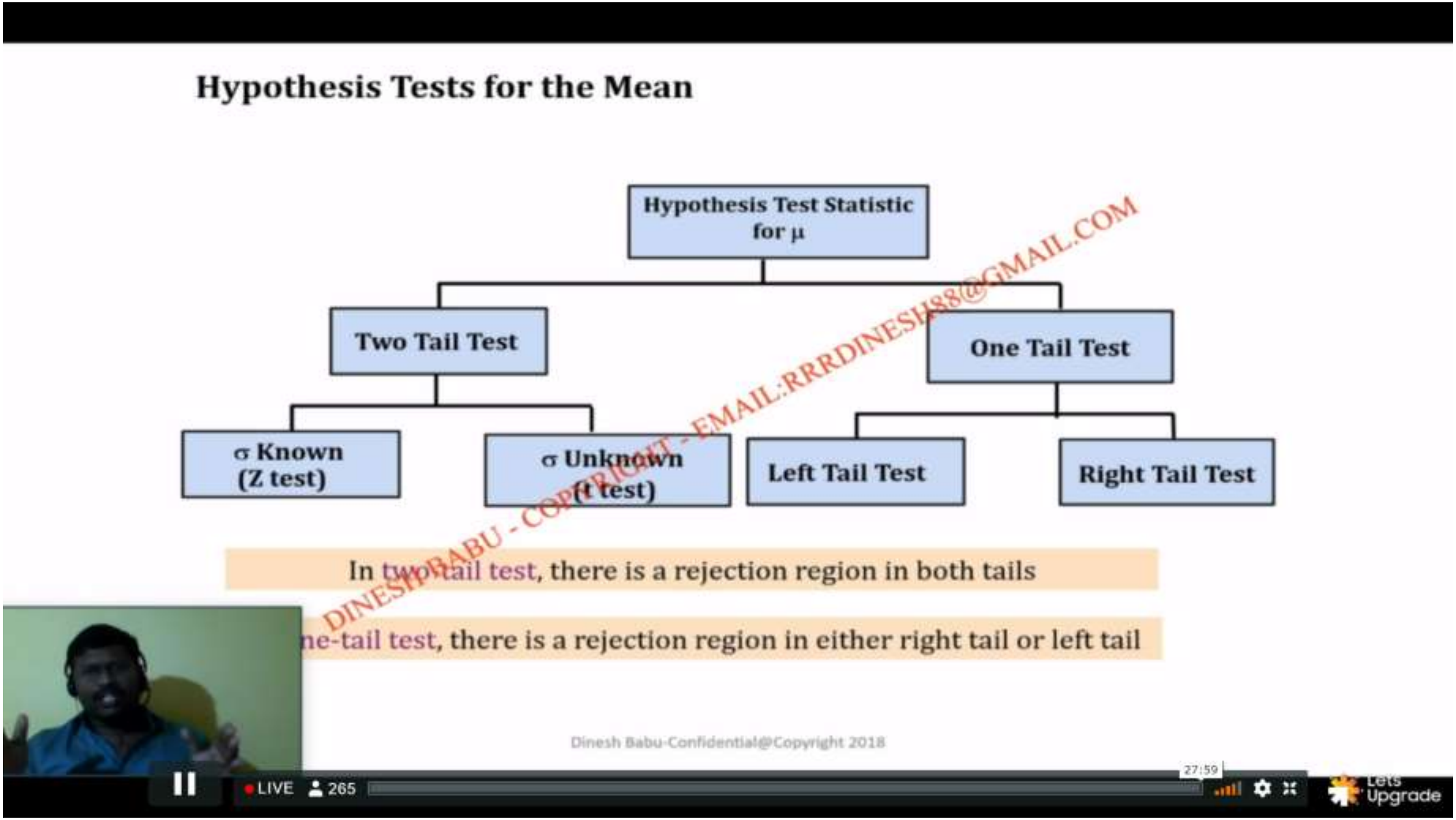
- Null hypothesis:
  1.set by us,2. alws must have three sings shown in fig.



- Alternate: alws opposite to null!.
  Alws not contained signs!
- May or may not b proven

Step2:


LIVE
270
Process of Hypothesis Testing
Formulate the Null Hypothesis and the alternative hypothesis.
Select the appropriate test statistic.
Choose the level of significance, Confidence Interval, Degree of Freedom
Compute the calculated test value of the test statistic
Compute the table test value of the test statisitc
Compare the calculated values and table values
Make the statistical decision and state the managerial conclusion.



- T test: std devt not kwn, Z tes: when std deviation is given for population.
- LHS / RHS tail : based on side of sampling distribution needed!

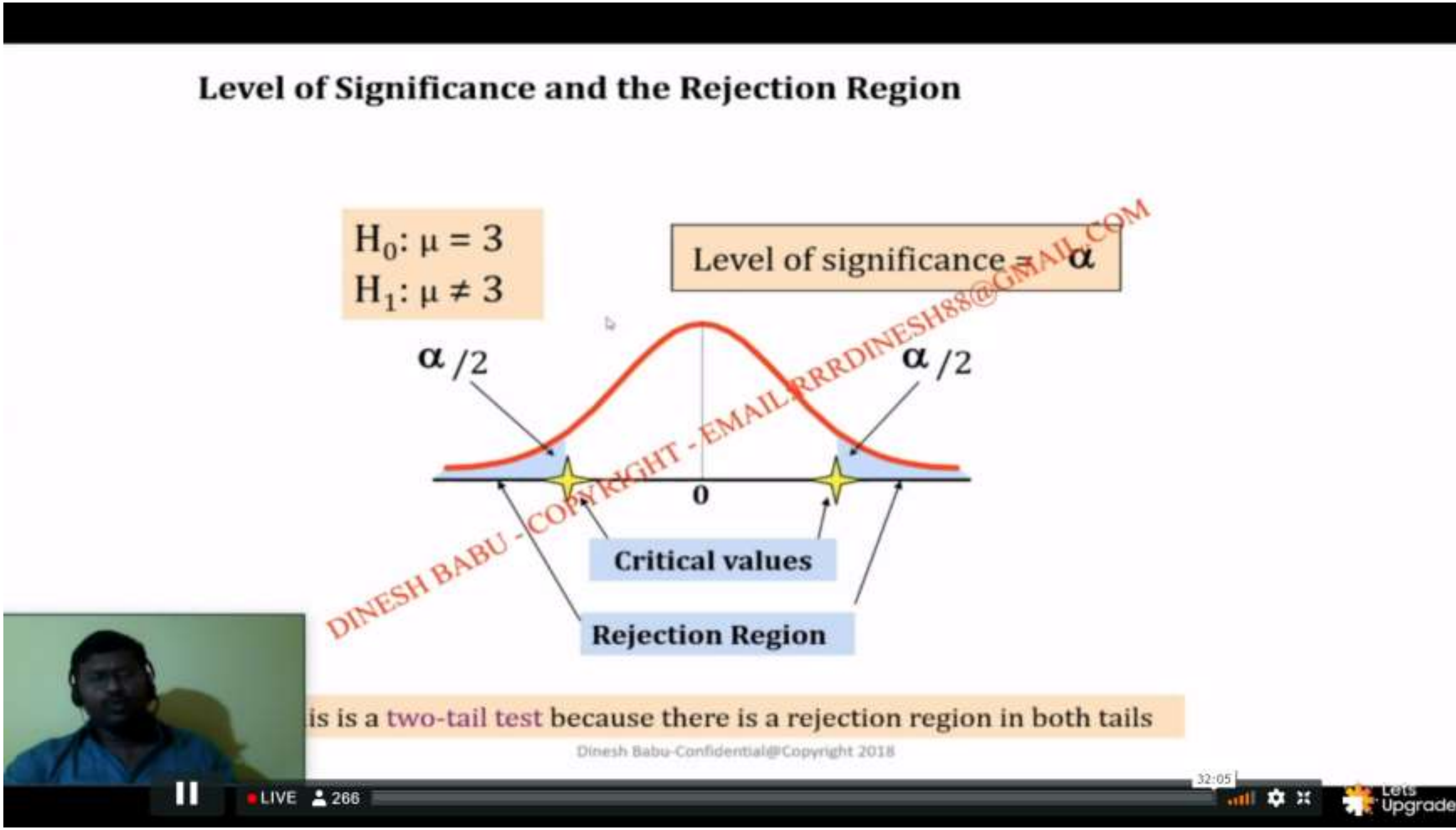
Two-Tailed Tests

Test where the region of rejection is on **both sides** of the sampling distribution.

Example

Speed limit in a freeway 60 – 80 mph (acceptable range of values).

gion of rejection would be numbers from both sides of the distribution,
t is, both <60 and >80 are defects.



- One tail : < += or >+=
- Two tail : = must there. (rejection at both sides)

Z Test of Hypothesis for the Mean (σ Known)

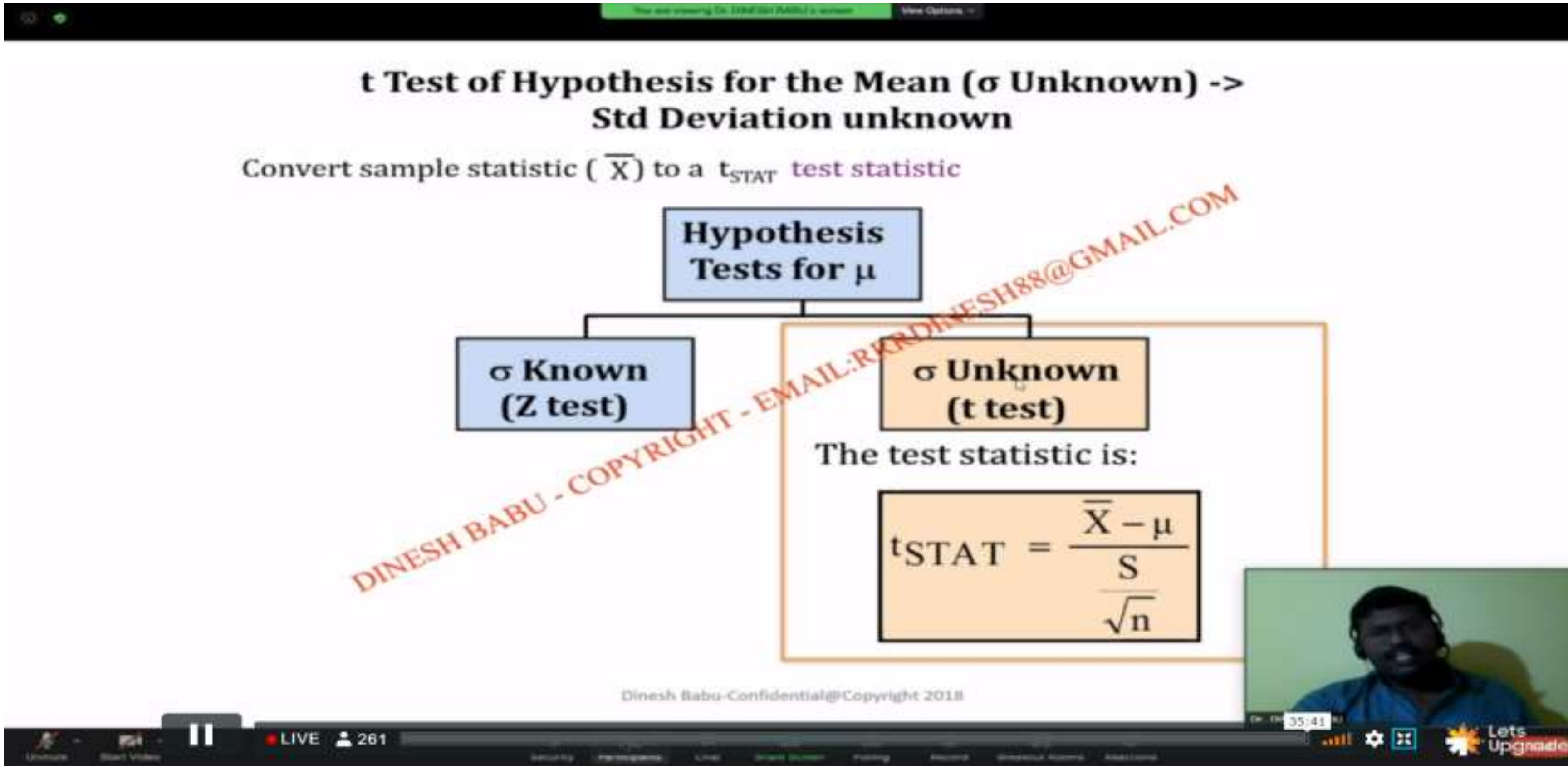
Convert sample statistic ($\overline{X}$) to a Z_{STAT} test statistic

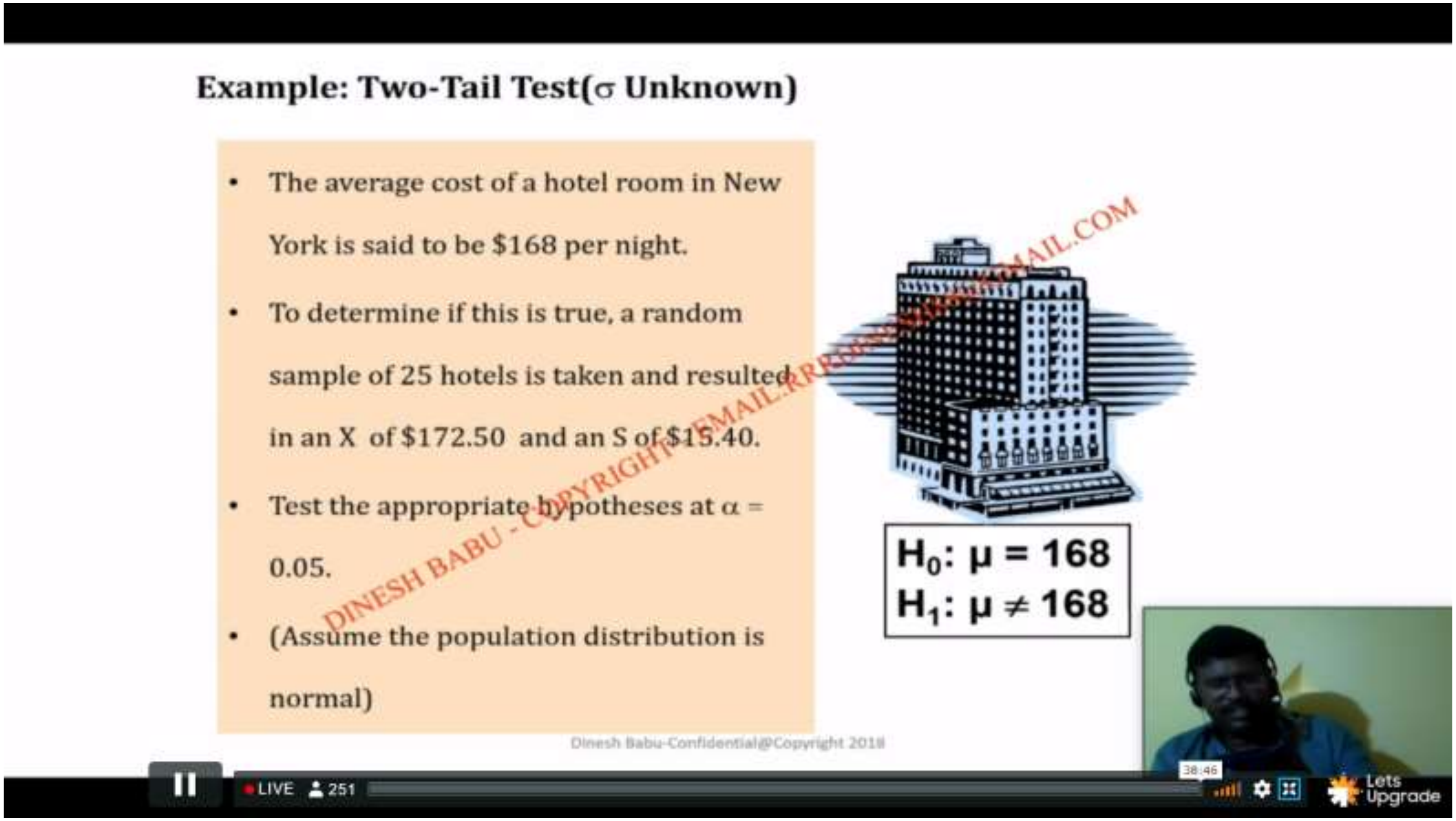
Hypothesis Tests for μ

- σ Known (Z test)
- σ Unknown (t test)

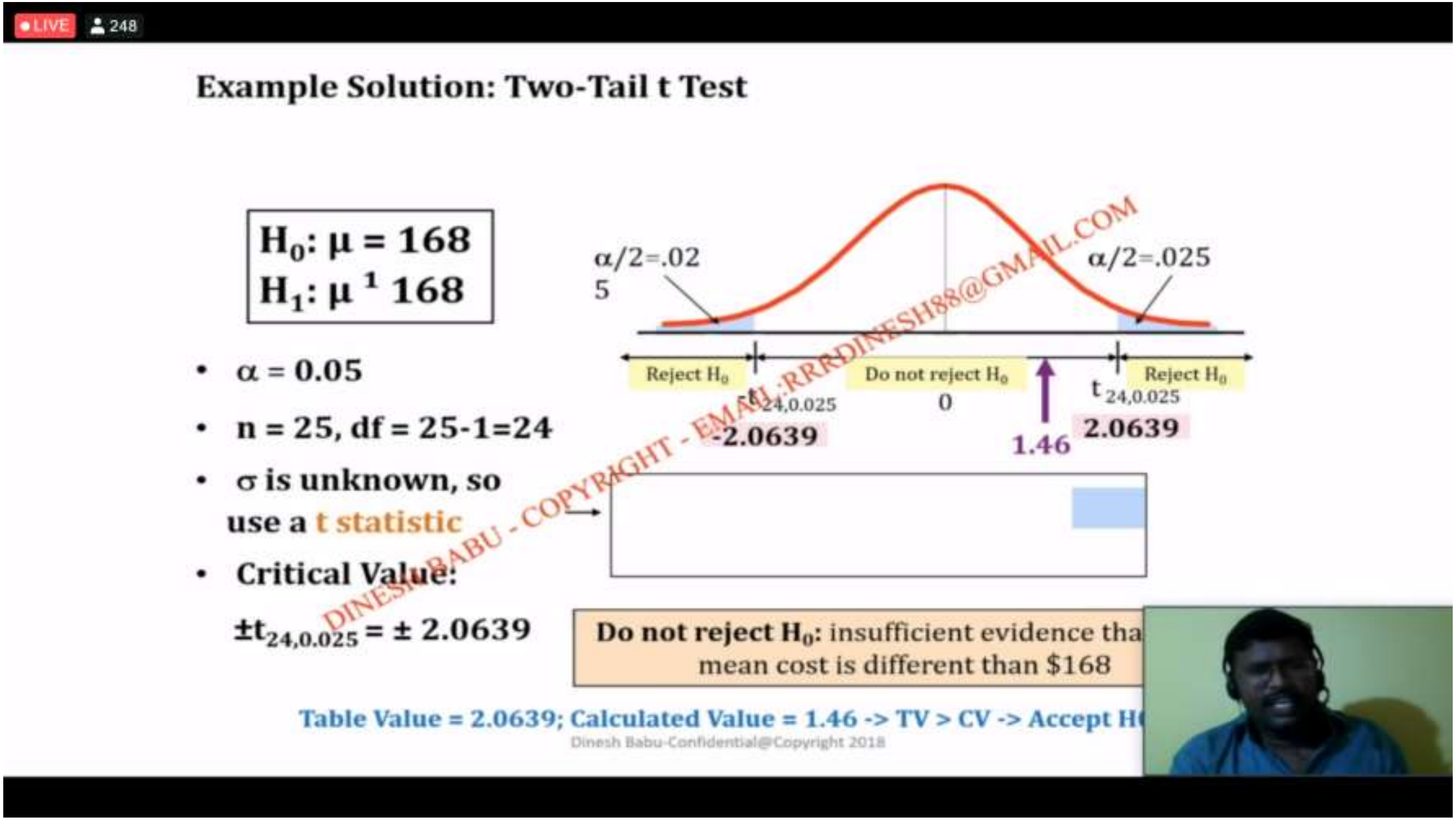
The test statistic is:

$$Z_{STAT} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

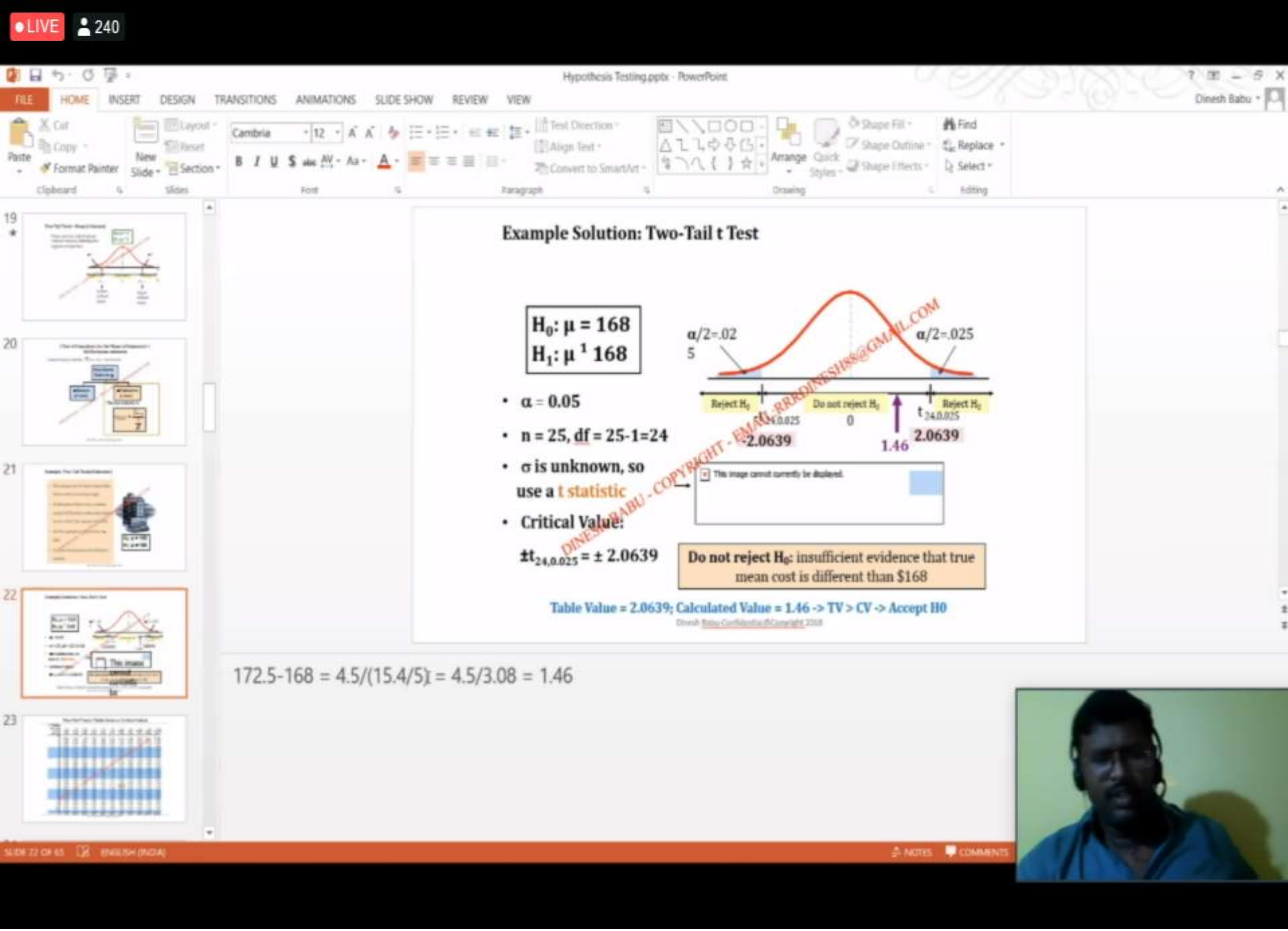

t Test of Hypothesis for the Mean (σ Unknown) -> Std Deviation unknown
Convert sample statistic ($\overline{X}$) to a t_{STAT} test statistic
Hypothesis Tests for μ
σ Known (Z test)
σ Unknown (t test)
The test statistic is:
$$t_{STAT} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$
DINESH BABU - COPYRIGHT - EMAIL:RKRDINESH88@GMAIL.COM
Dinesh Babu-Confidential@Copyright 2018
LIVE 261
35:41
Lets Upgrade



- Its two tail prob(= sing), ɥ is not avbl, only mean is given , hence T-test can be done.



- ά = nothing but error rate, it is given, 1.46>TV>CV: -> accept hypothesis.
- Ά = comes from quality of data, mostly given by customer.
- Rejection area: defined based on table values.


Example Solution: Two-Tail t Test
H0: μ = 168
H1: μ ¹ 168
α/2=.025
α/2=.025
Reject H0
Do not reject H0
Reject H0
0
-2.0639
1.46
2.0639
• α = 0.05
• n = 25, df = 25-1=24
• σ is unknown, so use a t statistic
• Critical Value:
±t24,0.025 = ± 2.0639
This image cannot currently be displayed.
Do not reject H0: insufficient evidence that true mean cost is different than $168
Table Value = 2.0639; Calculated Value = 1.46 -> TV > CV -> Accept H0
DINESH BABU - COPYRIGHT - EMAIL:RRRDINESH88@GMAIL.COM
172.5-168 = 4.5/(15.4/5) = 4.5/3.08 = 1.46

Two-Tail T test (Table Value or Critical Value)

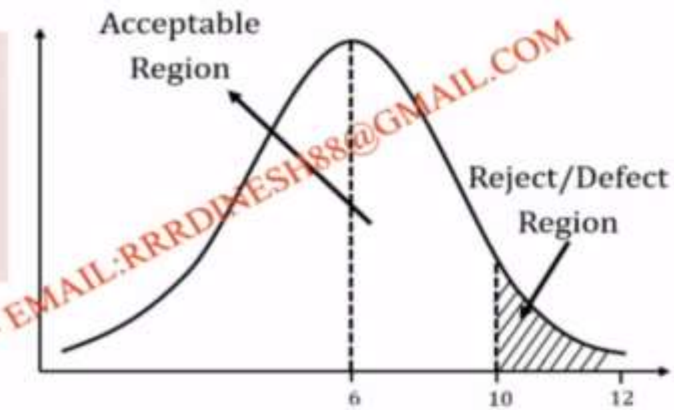
t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	[illegible]	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	[illegible]	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	[illegible]	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	[illegible]	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	[illegible]	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	[illegible]	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	[illegible]	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	[illegible]	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	[illegible]	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	[illegible]	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	[illegible]	[illegible]
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	[illegible]	[illegible]
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	[illegible]	[illegible]
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	[illegible]	[illegible]
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	[illegible]	[illegible]
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	[illegible]	[illegible]
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	[illegible]	[illegible]
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	[illegible]	[illegible]
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	[illegible]


LIVE
240
One Tail Test

One-Tailed Tests

Test where the region of rejection is on **only one side** of the sampling distribution.



Example

Null Hypothesis: Response time to customer query <=10 minutes

Alternative Hypothesis: Response time > 10 minutes
Region of rejection would be the numbers greater than 10 (there is no bo
the lesser time interval)

One-Tail Tests

In many cases, the alternative hypothesis focuses on a particular direction

Left Tail Test

$H_0: \mu \geq 3$

$H_1: \mu < 3$

This is a lower-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3

Right Tail Test

$H_0: \mu \leq 3$

$H_1: \mu > 3$

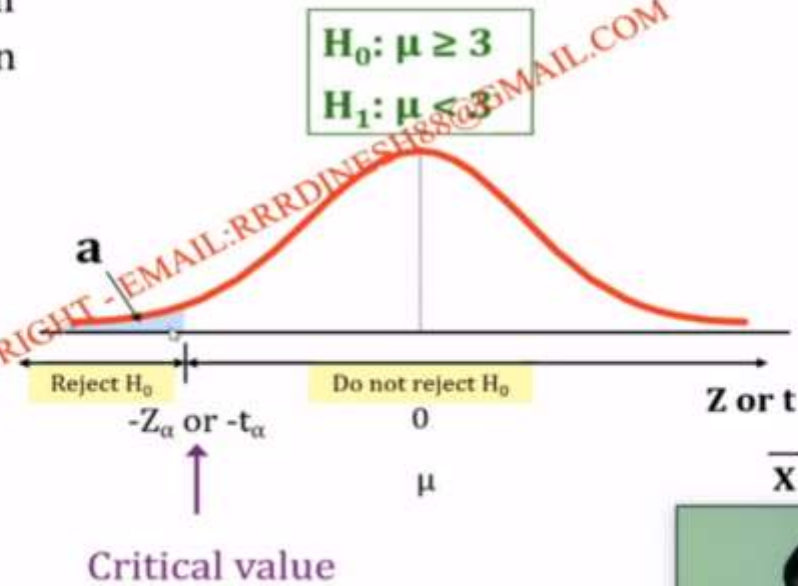
This is an upper-tail test since the alternative hypothesis is focused on the upper tail above the mean of 3

In one-tail test, there is a rejection region in either right tail or left

LIVE 242

Lower-Tail Tests -> One Tail Test -> Left Tail Test

There is only one critical value, since the rejection area is in only one tail.



Example: Upper-Tail t Test for Mean (σ unknown)

- A phone industry manager thinks that customer monthly cell phone bills have decreased, and now average less than \$52 per month.
- The company wishes to test this claim.
- Assume a normal population

Form hypothesis test:

$H_0: \mu \leq 52$ the average is not over \$52 per month

$H_1: \mu > 52$ the average **is** greater than \$52 per month (i.e., sufficient evidence exists to support the manager's claim)

LIVE 244

Process of Hypothesis Testing

Formulate the Null Hypothesis and the alternative hypothesis.

↓

Select the appropriate test statistic.

↓

Choose the level of significance, Confidence Interval, Degree of Freedom

↓

Compute the calculated test value of the test statistic

↓

Compute the table test value of the test statisitc

↓

Compare the calculated values and table values

↓

Make the statistical decision and state the managerial conclusion.

Step 3: Choose a Level of Significance α

- Level of Significance is also called as Error Rate

Scenarios:

- If $\alpha = 0.01$ then 1% error in the sample and remaining 99% accurate
- If $\alpha = 0.05$ then 5% error in the sample and remaining 95% accurate
- If $\alpha = 0.10$ then 10% error in the sample and remaining 90% accurate

Key Terms

Two key terms that you need to understand in Hypothesis Testing are:

Confidence Interval: Measure for reliability of an estimate; sample is used for estimating a population parameter so we need to know the reliability of that estimate

Degrees of Freedom: Number of values that are free to vary in a study

Confidence Interval

Confidence Interval

- Describes the reliability of an estimate
- Range of values (lower and upper boundary) within which the population parameter is included
- Width of the interval indicates the uncertainty associated with the estimate

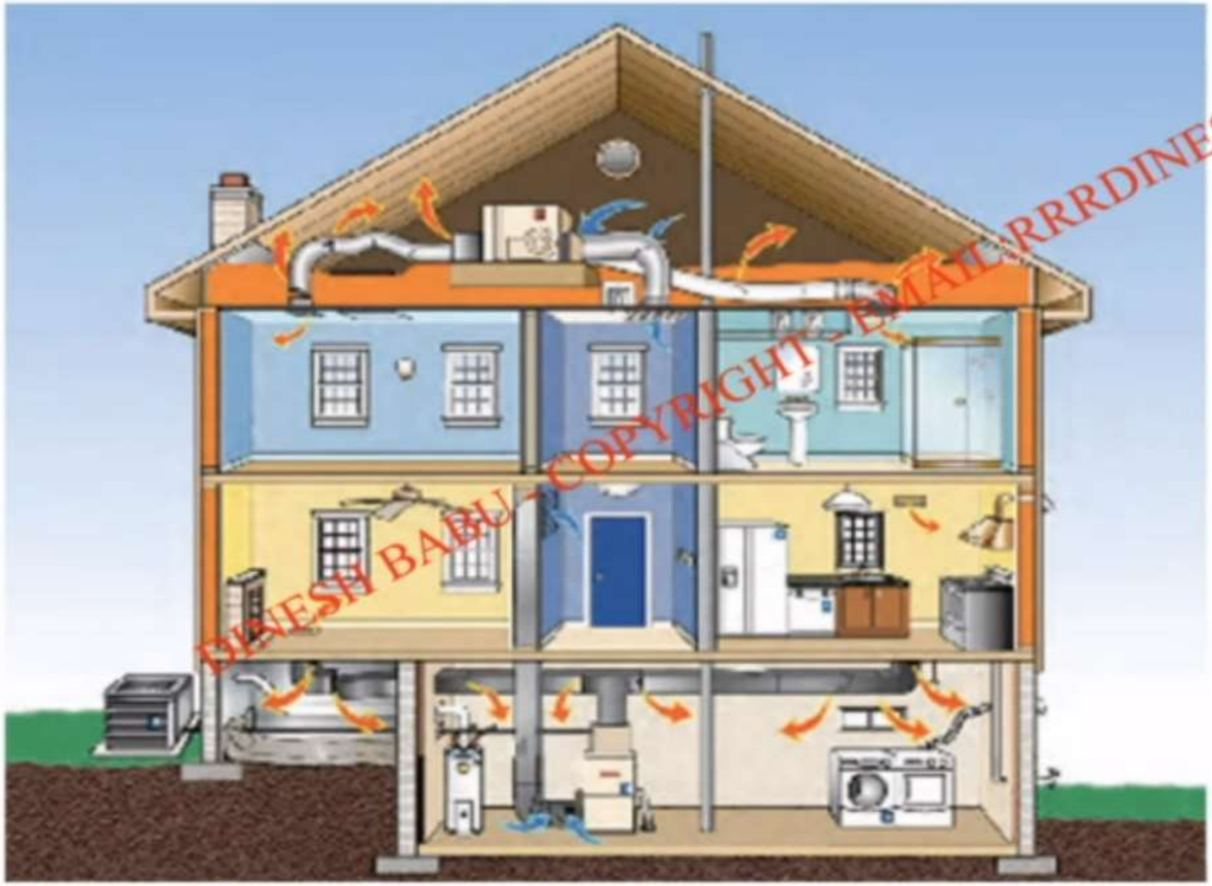
Confidence level

Probability associated with the confidence interval

Example 1: Confidence Inter

"Mean energy consumption of various houses in a colony is 200 units w

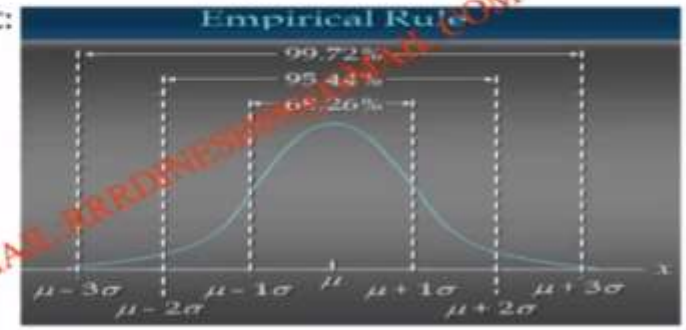
Standard Deviation of 20 units."



Discussion (Cont'd)

SOLUTION:

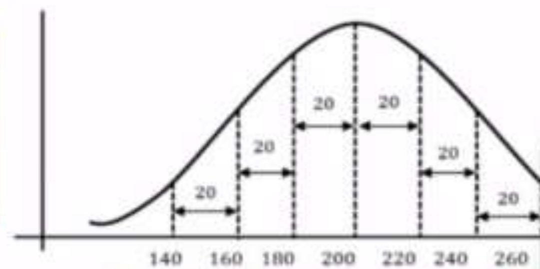
If the mean energy consumption of various houses in a colony is 200 units with a standard deviation of 20 units, it means that:



68.2% consume energy between 180 to 220 units

99% have their energy consumption between 140 to 260 units

Thus for any given household in the colony, there is a 99% confidence that the energy consumption of the household would be between 140 and 260 units.



-0.01

0.

Example 2: Confidence Interval

- Consider mean demand for computers during assembly lead time is 350 units. our operations manager wants to know ***whether the mean is different from 350 units***.

  Null Hypothesis - > H_0: = 350

  Thus, our research hypothesis becomes: H_1: $\neq$ 350

- Recall that the standard deviation [σ] was assumed to be 75, the sample size [n] was 25, and the sample mean was calculated to be 370.16

B. Common confidence levels and their critical values

You don't have to perform the above calculations every time. This list of critical values and their associated two-tailed test confidence levels were calculated using the above steps:

Confidence Level	Critical Value (Z-score)
0.90	1.645
0.91	1.70
0.92	1.75
0.93	1.81
0.94	1.88
0.95	1.96
0.96	2.05
0.97	2.17
0.98	2.33
0.99	2.575

https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-critical-values/

Critical Value Approach

- If we define the guts as the center 95% of the distribution [this 0.05], then the critical values that define the guts will be 1.96 st deviations of X-Bar on either side of the mean of the sampling distribution [350], or
- Upper Confidence Interval = Mean + (Table Value) * Std Dev.
- UCV = 350 + 1.96*15 = 350 + 29.4 = 379.4
- Lower Confidence Interval = Mean - (Table Value) * Std Dev.

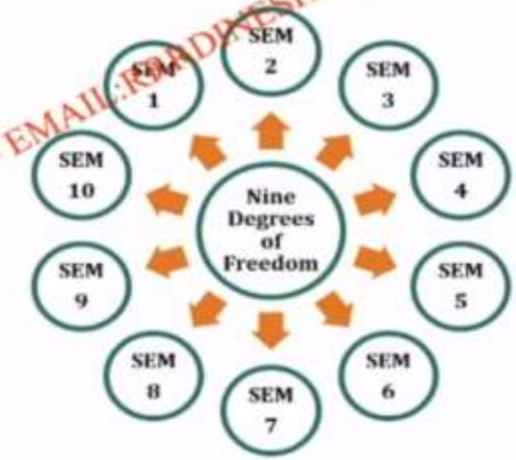
  LCV = 350 – 1.96*15 = 350 – 29.4 = 320.6
- Table Value (α = 0.05, Df = 24) = 1.96

Degrees of Freedom

Degrees of Freedom is the measure of number of values in a study that are free to vary.

For example, if you have to take ten different courses to graduate, and only ten different courses are offered, then you have nine degrees of freedom.

In nine semesters, you will be able to choose which class to take. In the tenth semester, there will only be one class left to take – there is no choice.



Degrees of freedom = No. of Rows – No. of Columns = 10 -1 = 9

Process of Hypothesis Testing

Formulate the Null Hypothesis and the alternative hypothesis.

↓

Select the appropriate test statistic.

↓

Choose the level of significance, Confidence Interval, Degree of Freedom

↓

Compute the calculated test value of the test statistic

↓

Compute the table test value of the test statistic

↓

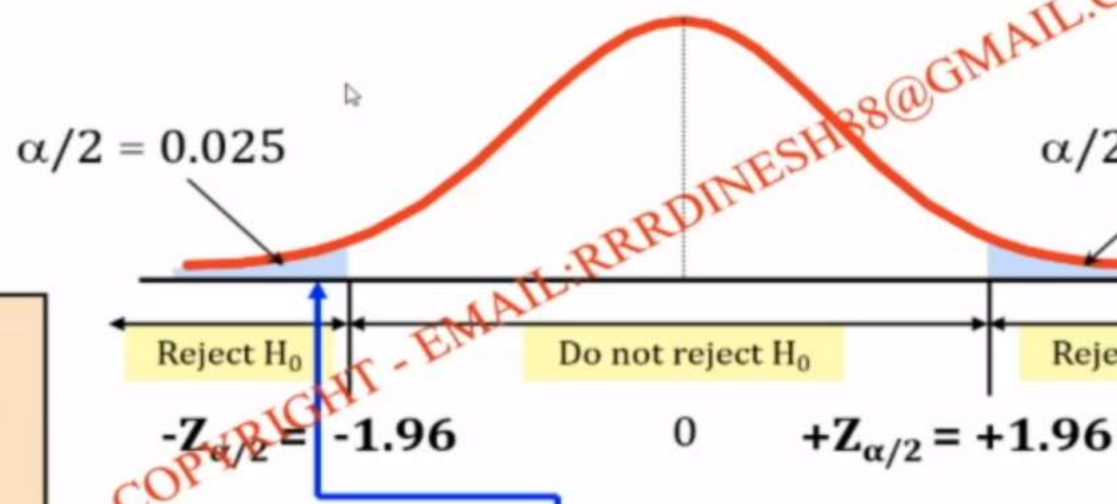
Compare the calculated values and table values

↓

Make the statistical decision and state the managerial conclusion.

Hypothesis Comparison

Is the test statistic in the rejection region?



Reject H_0 if $Z_{STAT} < -1.96$ or $Z_{STAT} > 1.96$; otherwise do not reject H_0

Here, $Z_{STAT} = -2.0 < -1.96$, so the test statistic is in the rejection region

Critical Value Approach to Testing

Test the claim that the true mean # of TV sets in US homes is equ

(Assume σ = 0.8)

1. State the appropriate null and alternative hypotheses
   - H_0: $\mu = 3$ H_1: $\mu \neq 3$ (This is a two-tail test)

2. Specify the desired level of significance and the sample size
   - Suppose that $\alpha = 0.05$ and $n = 100$ are chosen for this te
   - σ is assumed known so this is a Z test.

3. Collect the data and compute the test statistic
   - Suppose the sample results are

     $n = 100$, $X = 2.84$ ($\sigma = 0.8$ is assumed known)

   So the test statistic is:

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -$$

Process of Hypothesis Testing

Formulate the Null Hypothesis and the alternative hypothesis.

↓

Select the appropriate test statistic.

↓

Choose the level of significance, Confidence Interval, Degree of Freedom

↓

Compute the calculated test value of the test statistic

↓

Compute the table test value of the test statistic

↓

Compare the calculated values and table values

↓

Make the statistical decision and state the managerial conclusion.

P-Value Approach to Testing

- P-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the Null Hypothesis is true
- When P-value is less than a certain significance level (often 0.05), you "reject the null hypothesis". This result indicates that the observed result is not due to a random occurrence but a true difference.

P-value < 0.05 → Null Hypotheses → Result is due to a true difference

REJECT

- Compare the p-value with $\alpha = 0.05$
  - If p-value $<$ **0.05**, reject H_0
  - If p-value $\geq$ **0.05**, do not reject H_0

Possible Scenarios in Hypothesis

Four possible scenarios:

Truth about the population

Decision based on sample

	H_0 true	H_a true
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

Type I Error (α):
Reject the Null Hypothesis when it is true

Type II Error (β):
Accept the Null Hypothesis when it is false

Example

- A criminal trial is an example of hypothesis testing without the statistics.
- In a trial a jury must decide between two hypotheses. The null hypothesis is:

  H_0: The defendant is innocent

- The alternative hypothesis or research hypothesis is:

  H_1: The defendant is guilty

- The jury does not know which hypothesis is true. They must make a decision on the basis of evidence presented.

Justice

Null Hypothesis = "Person is innocent"

		Decision	
		Prison	Set free
True State	Innocent	Type I error	Correct decision
	Guilty	Correct decision	Type II error

Testing of hypotheses

Type I and Type II Errors. Example

Suppose there is a test for a particular disease.

If the disease really exists and is diagnosed early, it can be successfully treated

If it is not diagnosed and treated, the person will become severely disabled

If a person is erroneously diagnosed as having the disease and treated, no physical damage is done.

Testing of hypotheses

Type I and Type II Errors. Example.

Decision	No disease	Disease
Not diagnosed	OK	Type II error
Diagnosed	Type I error	OK

treated but not harmed
by the treatment

irreparable damage
would be done

Decision: to avoid Type error II, have high level of significance

How Do We Control Type I Errors?

- The Type I error rate is controlled by the researcher.
- It is called the **alpha rate**, and corresponds to the probability cut-off that one uses in a significance test.
- By convention, researchers use an alpha rate of .05. In other words, they will only reject the null hypothesis when a statistic is likely to occur 5% of the time or less when the null hypothesis is true.
- In principle, any probability value could be chosen for making the accept/reject decision. 5% is used by convention.

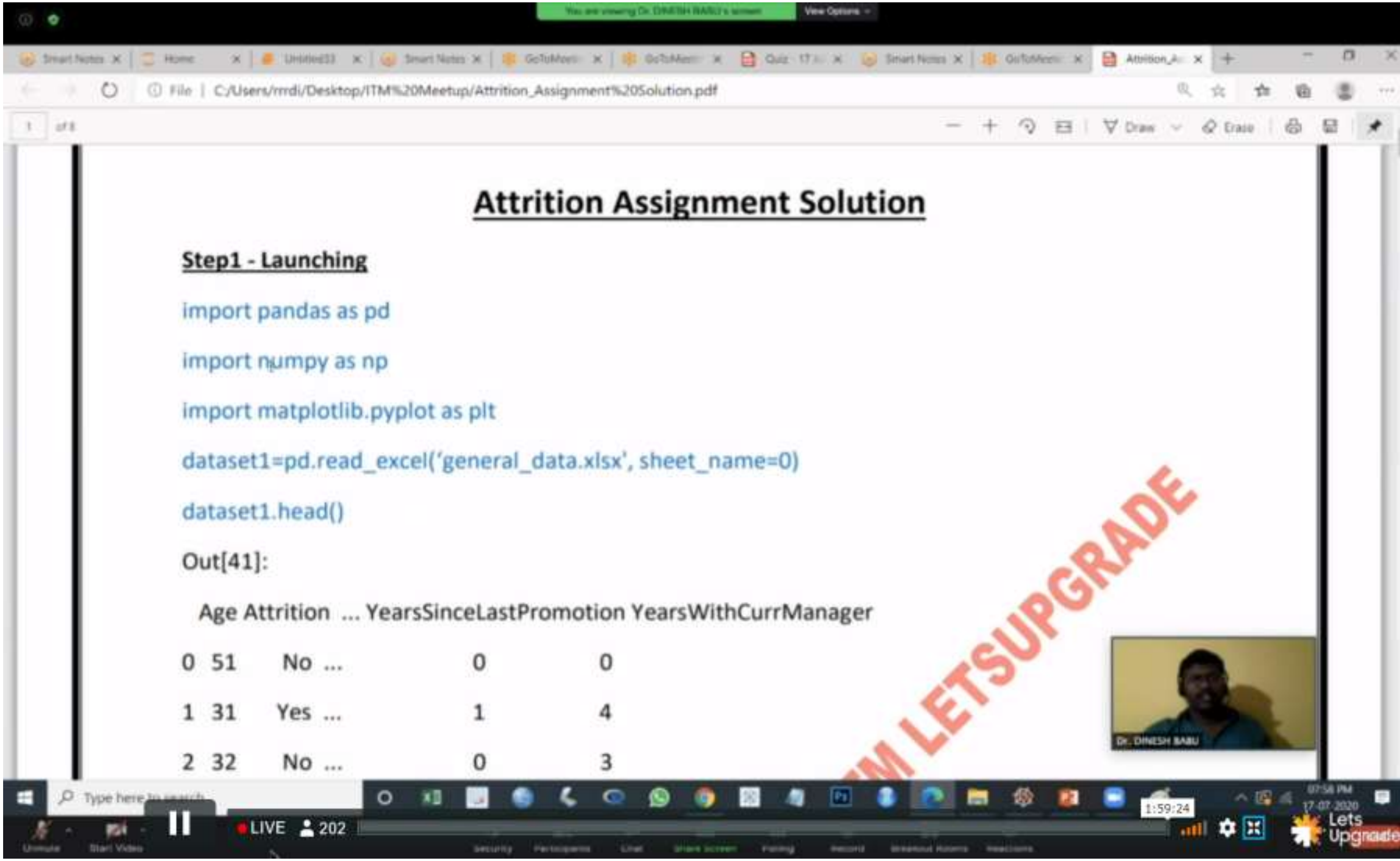
Type I & II Error Relationship

Type I and Type II errors cannot happen at the same time

- A Type I error can only occur if H_0 is true
- A Type II error can only occur if H_0 is false

If Type I error probability (α) $\uparrow$, then

Type II error probability (β) $\downarrow$



Attrition Assignment Solution

Step1 - Launching

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_excel('general_data.xlsx', sheet_name=0)
dataset1.head()
```

Out[41]:

	Age	Attrition	...	YearsSinceLastPromotion	YearsWithCurrManager
0	51	No	...	0	0
1	31	Yes	...	1	4
2	32	No	...	0	3

- Sheet_name= 0, first sheet of excel

Out[41]:

```
  Age Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
0  51        No  ...                       0                    0
1  31       Yes  ...                       1                    4
2  32        No  ...                       0                    3
3  38        No  ...                       7                    5
4  32        No  ...                       0                    4

[5 rows x 18 columns]
```

```
dataset1.columns
```

Out[42]:

```
Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',
       'Education', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus',
       'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike',
```

```
4  32     No  ...           0            4

[5 rows x 18 columns]

dataset1.columns

Out[42]:

Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',
       'Education', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus',
       'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

Step 2 - Data Treatment:

```
dataset1.isnull()
```

Out[47]:

	Age	Attrition	...	YearsSinceLastPromotion	YearsWithCurrManager
0	False	False	...	False	False
1	False	False	...	False	False
2	False	False	...	False	False
3	False	False	...	False	False
4	False	False	...	False	False

C:/Users/rrrdi/Desktop/ITM%20Meetup/Attrition_Assignment%20Solution.pdf

4405	False	False	...	False	False
4406	False	False	...	False	False
4407	False	False	...	False	False
4408	False	False	...	False	False
4409	False	False	...	False	False

[4410 rows x 18 columns]

```
dataset1.duplicated()
```

Out[50]:

0	False
1	False
2	False
3	False
4	False

```
dataset1.drop_duplicates()
```

Out[53]:

	Age	Attrition	...	YearsSinceLastPromotion	YearsWithCurrManager
0	51	No	...	0	0
1	31	Yes	...	1	4
2	32	No	...	0	3

Step 3 – Univariate Analysis:

```
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

dataset3
```

Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastP	
count	4410	4410	4410	4410	4391	4410	4401	4410	4410	4410	
mean	36...	9.19252	2.91293	65029.3	2.69483	15.2095	11.2799	2.79932	7.00816	2.18776	4.12313
std	9.1...	8.10503	1.02393	47068.9	2.49889	3.65911	7.78222	1.28898	6.12514	3.2217	3.56733

Step 3 – Univariate Analysis:

```
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

dataset3
```

Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastP	
count	4410	4410	4410	4410	4391	4410	4401	4410	4410	4410	
mean	36...	9.19252	2.91293	65029.3	2.69483	15.2095	11.2799	2.79932	7.00816	2.18776	4.12313
std	9.1...	8.10503	1.02393	47068.9	2.49889	3.65911	7.78222	1.28898	6.12514	3.2217	3.56733

```
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()

dataset3
```

Out[67]:

Age	36.0
DistanceFromHome	7.0
Education	3.0
MonthlyIncome	49190.0
NumCompaniesWorked	2.0
PercentSalaryHike	14.0
TotalWorkingYears	10.0
TrainingTimesLastYear	3.0

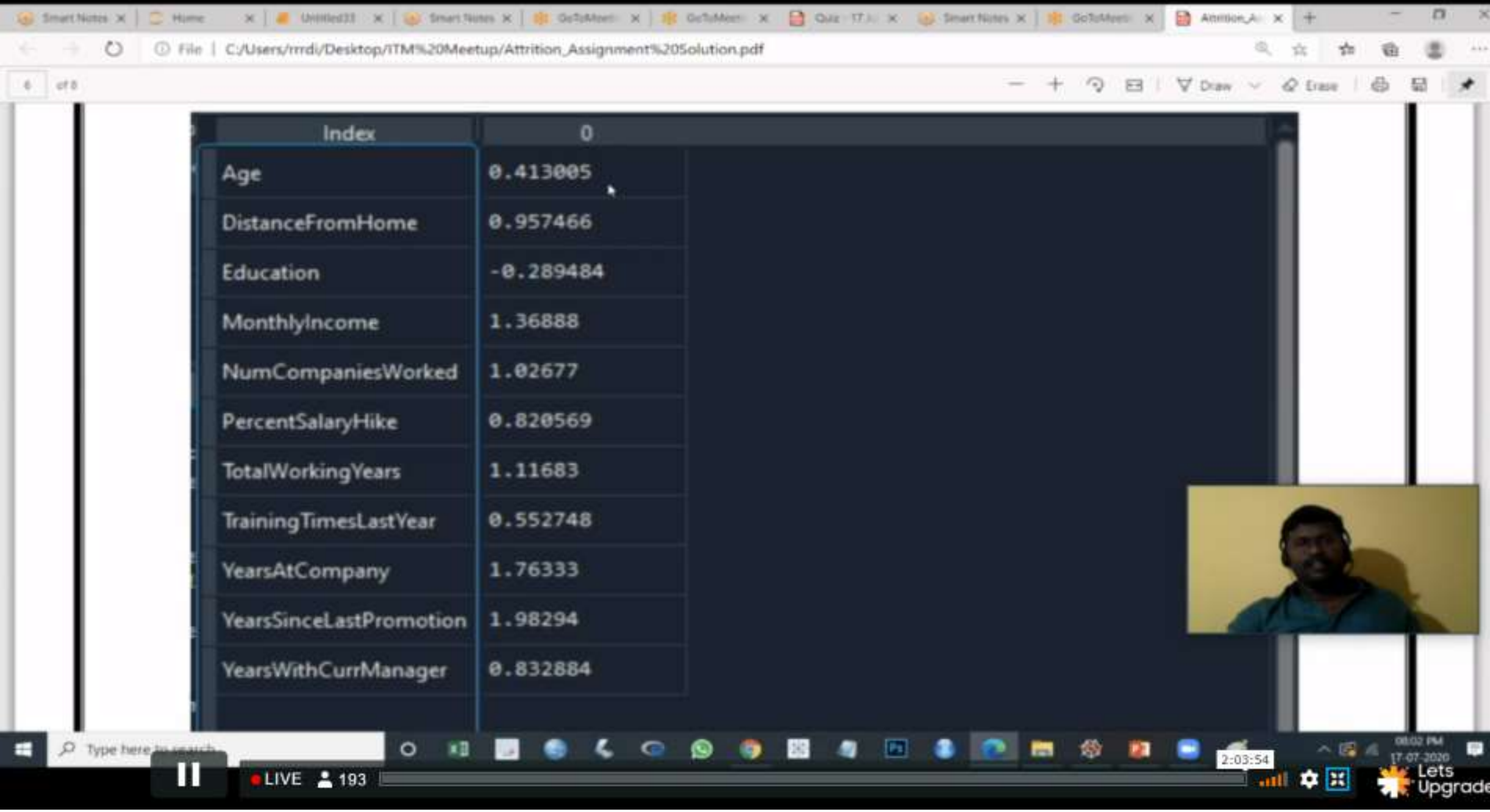
```
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLast
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()

dataset3
```

1

dataset3 - Series

Index	0
Age	83.4172
DistanceFromHome	65.6914
Education	1.04844
MonthlyIncome	2.21548e+09
NumCompaniesWorked	6.24444
PercentSalaryHike	13.3891
TotalWorkingYears	60.563
TrainingTimesLastYear	1.66146
YearsAtCompany	37.5173
YearsSinceLastPromotion	10.3793
YearsWithCurrManager	12.7258



Index	0
Age	0.413005
DistanceFromHome	0.957466
Education	-0.289484
MonthlyIncome	1.36888
NumCompaniesWorked	1.02677
PercentSalaryHike	0.820569
TotalWorkingYears	1.11683
TrainingTimesLastYear	0.552748
YearsAtCompany	1.76333
YearsSinceLastPromotion	1.98294
YearsWithCurrManager	0.832884

- Skewness

	Mean	Median	Mode	Variance	Std Deviation	IQR	Skewness	Kurtosis
Mean Age (Yrs)	36	36	35	83.14	9.1	13	0.418	-0.4
Mean Distance from Home (Kms)	9	7	2	65.69	8.1	2	0.957	-0.22
Mean Monthly Income (Rs)	65000	49190	23420	2215480000	47068	54000	1.36	1
Mean Work Experience (Yrs)	11.29	10	10	60	7.72	9	1.11	0.91
Mean Years at Company (Yrs)	7	5	5	37.51	6.12	6	1.76	3.92
Mean Years since last promotion (Yrs)	2	1	0	10.37	3.22	3	1.98	3.6
Mean Years with Current Manager (Yrs)	4	3	2	12.72	3.56	5	0.83	0.16

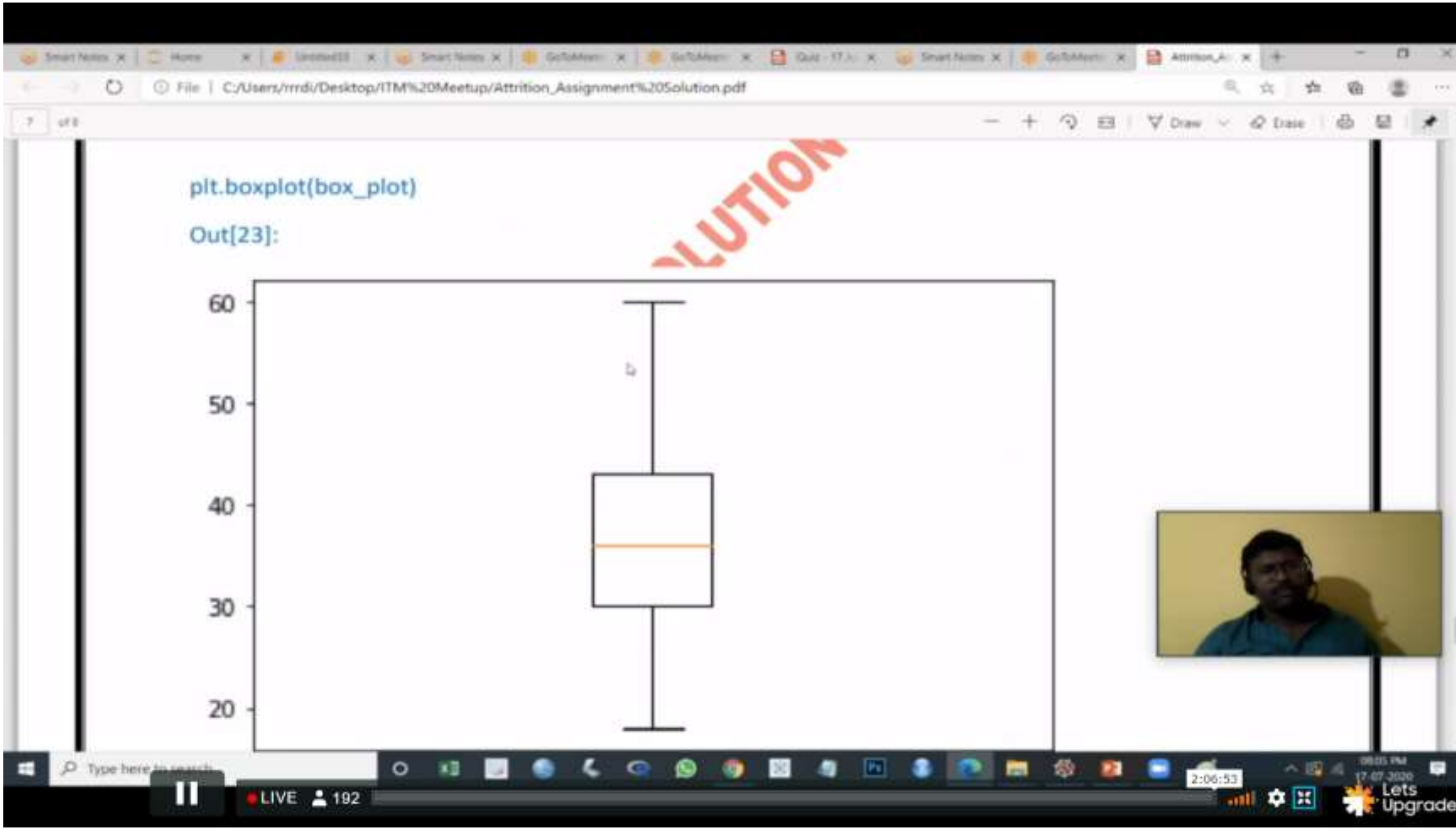
Inference from the analysis:

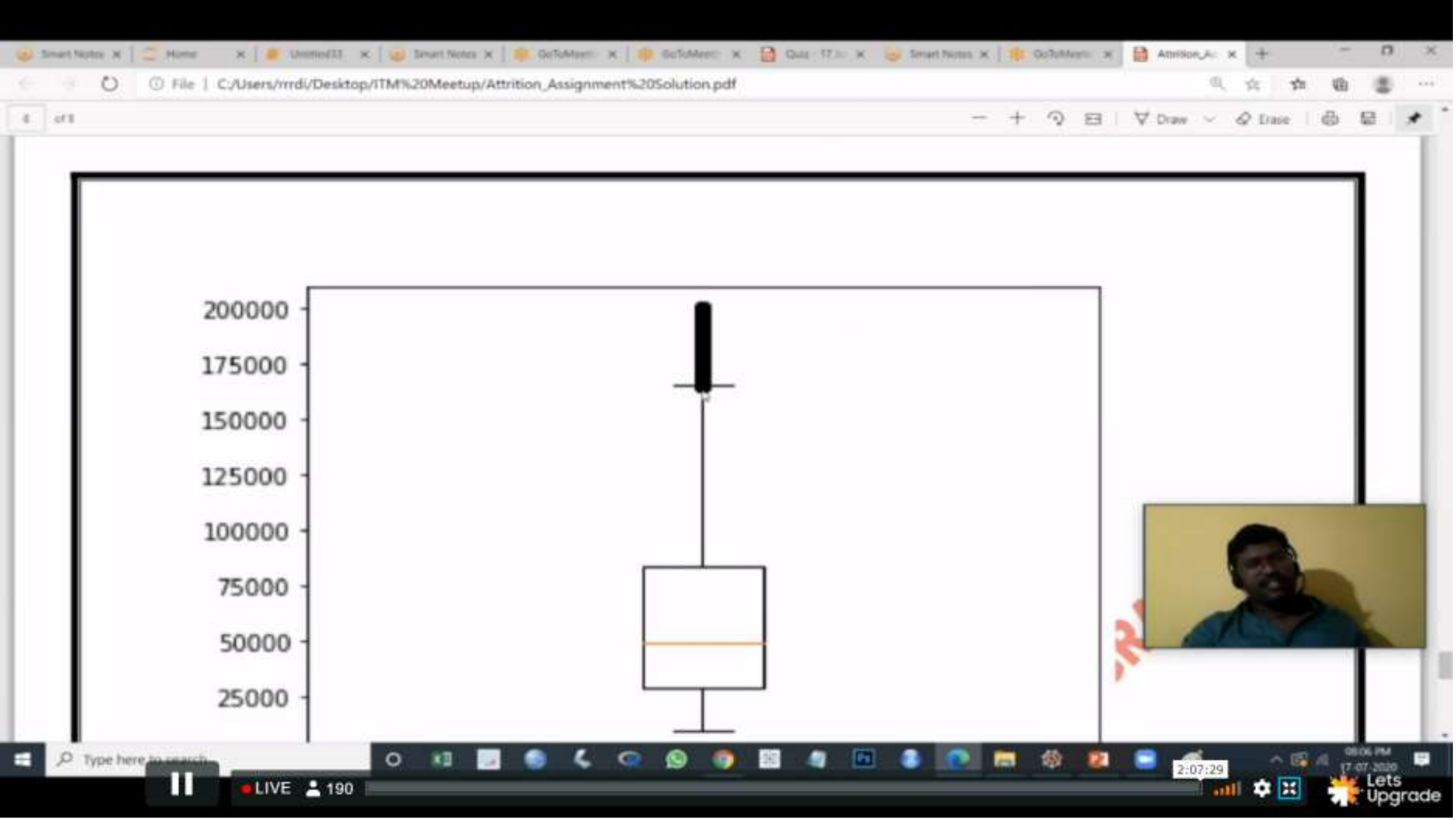
- All the above variables show positive skewness; while Age & Mean_distance_from_home are leptokurtic and all other variables are platykurtic.
- The Mean_Monthly_Income's IQR is at 54K suggesting company wide attrition across all income bands
- Mean age forms a near normal distribution with 13 years of IQR

Outliers:

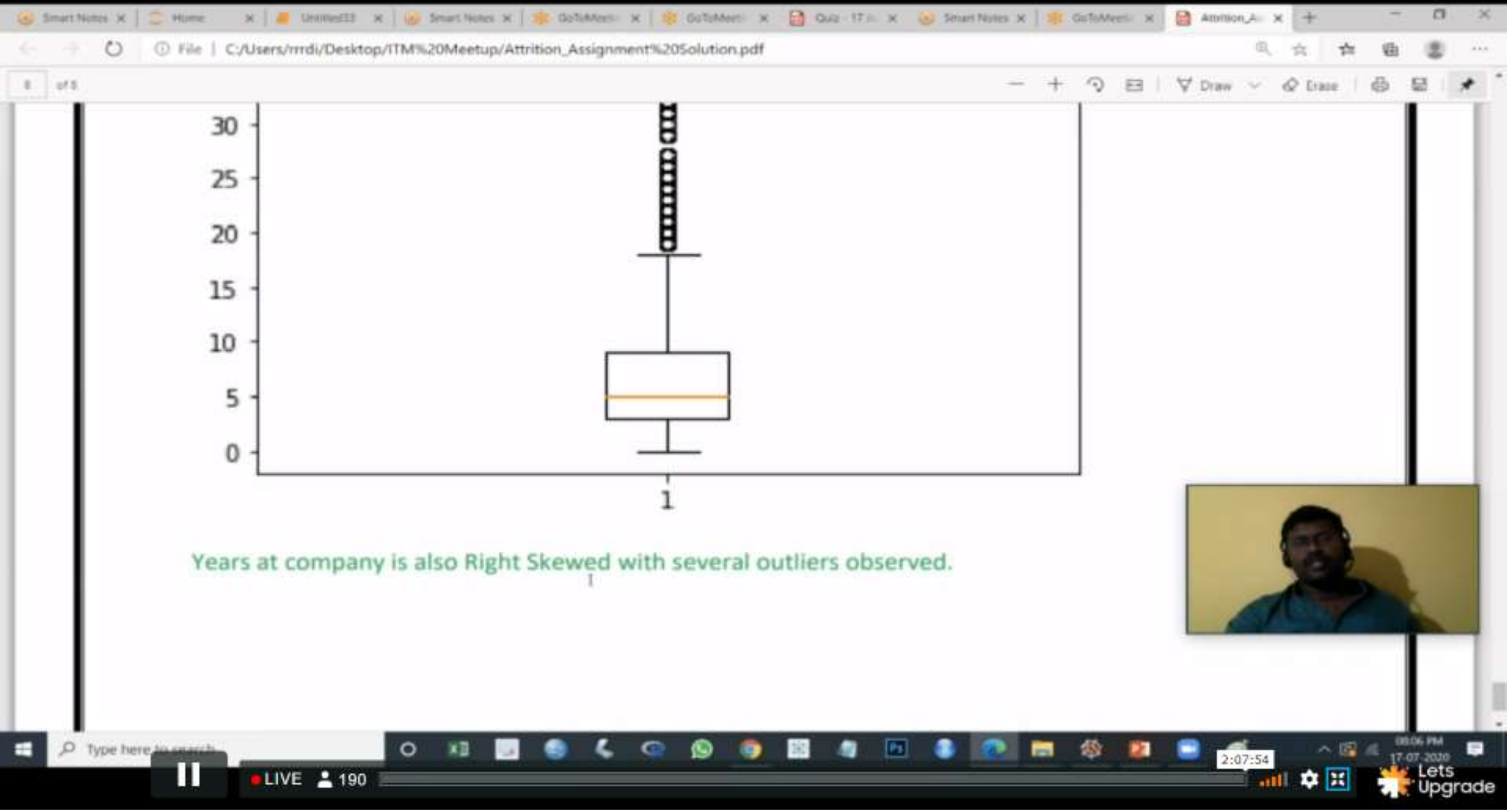
There's no regression found while plotting Age, MonthlyIncome, TotalWorkingYears, YearsAtCompany, etc., on a scatter plot

box_plot=dataset1.Age


C:/Users/rrrdi/Desktop/ITM%20Meetup/Attrition_Assignment%20Solution.pdf
plt.boxplot(box_plot)
Out[23]:
LUTION
60
50
40
30
20
LIVE
192
2:06:53
Lets Upgrade



- Monthly income


File | C:/Users/rrrdi/Desktop/ITM%20Meetup/Attrition_Assignment%20Solution.pdf
Draw
Erase
30
25
20
15
10
5
0
1
Years at company is also Right Skewed with several outliers observed.
Type here to search
2:07:54
LIVE
190
Lets Upgrade