Topics:

- Code: t-test & z-test (Youtube ads)
- Z-proportions test
- χ^2-test ✓
- ANOVA → lengthy

KStest_Ttest.ipynb - Colaborat | Mann-Whitney U test - Wikipe

colab.research.google.com/drive/1lWuS8AxULBBold2GlfBKUeT1RSj3fFAJ#scrollTo=Y569EYQBiMeN&printMode=true

KStest_Tt

File Edit View

+ Code + Text

KS-Test

[ ] ↳ 11 cells h

T-test & Z-t

```
[ ] # Group
    # Group
    # Let us
    # H0: mu
    # H1: mu
```

```
[ ] import p
    import n
    import matplotlib.pyplot as plt
    import seaborn as sns
```

Print

Destination: Save as PDF

Pages: All

Layout: Portrait

More settings

Loading preview...

Cancel Save

2 / 93

Share Editing

✓ 0s completed at 21:18

[A/B Testing:]

Youtube

PM

1 ad

2 ads

break → 1 ad

↘ 2ads

Finance

$x_{1,1}, x_{1,2}, x_{1,3} \ldots x_{1,5000}$

10,000 randomly → A_{5000} (50%) → 1 ad

→ B_{5000} (50%) → 2 ads

[0 - 1]

< 0.5

$x_{2,1}, x_{2,2} \ldots x_{2,5000}$

watch-time

Code

H_0: mean watch time of gp A = MWT of gp B

H_a: " $\neq$ "

which test → Z-Test ✓

T-test ✓

n_1 & n_2 are small

Q

Biz-ques:

Compare WT of A & B → DA

KS-test disb of WT_A & WT_B

→ 95% C.I on MWT → CLT

medianWT → bootstrapping or CLT

→ QQplot $disb_A$ & $disb_B$ are same or not

T-test & Z-test

```
[ ] # Group A --> Treatment Group shown 2 ads per ad-break
    # Group B --> Control Group shown only 1 ad per ad break
    # Let us comapre mean watch-times per group
    # H0: mu1= mu2
    # H1: mu1 != mu2
```

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    from scipy import stats
    import scipy
```

```
[ ] # Download data
    # https://drive.google.com/file/d/1Hl96n6BWdl3ruJgCo_gaAWEb0kEYg__H/view?usp=sharing
    id = "1Hl96n6BWdl3ruJgCo_gaAWEb0kEYg__H"
    path = "https://docs.google.com/uc?export=download&id=" + id
    print(path)
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help

KStest_Ttest.ipynb - Colaborat

colab.research.google.com/drive/1lWuS8AxULBBold2GlfBKUeT1RSj3fFAJ#scrollTo=Y569EYQBiMeN

KStest_Ttest.ipynb

File Edit View Insert Runtime Tools Help Last saved at 17:39

Comment Share

+ Code + Text Connect Editing

```python
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import scipy
```

```python
# Download data
# https://drive.google.com/file/d/1Hl96n6BWdl3ruJgCo_gaAWEb0kEYg__H/view?usp=sharing
id = "1Hl96n6BWdl3ruJgCo_gaAWEb0kEYg__H"
path = "https://docs.google.com/uc?export=download&id=" + id
print(path)
```

https://docs.google.com/uc?export=download&id=1Hl96n6BWdl3ruJgCo_gaAWEb0kEYg__H

```python
!wget "https://docs.google.com/uc?export=download&id=1Hl96n6BWdl3ruJgCo_gaAWEb0kEYg__H" -O ab_test_data.csv
```

```
--2022-05-19 12:07:38--  https://docs.google.com/uc?export=download&id=1Hl96n6BWdl3ruJgCo_gaAWEb0kEYg__H
Resolving docs.google.com (docs.google.com)... 173.194.79.102, 173.194.79.139, 173.194.79.100, ...
Connecting to docs.google.com (docs.google.com)|173.194.79.102|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://doc-00-ag-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/ee6hdontug8dme
Warning: wildcards not supported in HTTP
```

7/7

```
!ls -lrt
```

```
total 872
drwxr-xr-x 1 root root   4096 May 17 13:39 sample_data
-rw-r--r-- 1 root root 887610 May 19 12:07 ab_test_data.csv
```

```
!cat ab_test_data.csv
```

```
date,customer_id,premium,watch_time_hrs,customer_segmnt
2018-09-11,402,0,7.173618291737698,control
2018-02-28,227,0,0.8361704091625795,control
2018-10-18,812,1,4.40207828983548,treatment
2018-05-22,43,0,3.982454213516833 7,control
2018-07-18,307,0,7.51330204299158 1,control
2018-09-10,238,0,1.45696098276263 7,control
2018-02-21,691,1,3.80037470551441 2,treatment
2018-04-27,199,0,4.5744457928113444,control
2018-05-28,105,0,3.4259417790693356,control
2018-09-24,604,0,3.959895726467226,treatment
2018-07-18,512,0,1.4222729738202025,treatment
2018-04-12,9,1,2.72394365641492 4,control
2018-08-20,216,0,2.1513605982069364,control
2018-10-24,591,0,3.051881058517045,treatment
2018-12-28,152,1,4.0154599567449765,control
2018-01-10,116,0,0.7329431665935803,control
2018-12-09,697,0,2.927355165799921 3,treatment
```

1ad-s

Control → A

Treatment → B

2-ads

```
2018-11-14,507,0,4.3198214955554874,treatment
2018-07-11,807,1,1.1820394825584641,treatment
2018-08-07,771,0,1.695615402882516,treatment
2018-07-21,704,0,1.1660962947526128,treatment
2018-05-10,620,1,1.1703272186648386,treatment
2018-04-30,681,1,2.2900538409710434,treatment
2018-01-17,761,0,1.7702776878978959,treatment
2018-12-19,951,0,2.66771754179069,treatment
```

```
ab_test_data = pd.read_csv("ab_test_data.csv")
ab_test_data.sample(10)
```

	date	customer_id	premium	watch_time_hrs	customer_segmnt
12648	2018-04-05	981	1	1.181891	treatment
299	2018-06-19	519	0	2.235622	treatment
16309	2018-10-30	842	0	1.811846	treatment
11240	2018-09-12	193	0	4.221867	control
3654	2018-08-10	109	0	2.936236	control
17109	2018-06-19	962	0	2.717192	treatment
18074	2018-03-12	833	1	1.949134	treatment
7985	2018-08-24	239	0	2.240057	control

```
ab_test_data = pd.read_csv("ab_test_data.csv")
ab_test_data.sample(100)
```

	date	customer_id	premium	watch_time_hrs	customer_segmnt
3269	2018-08-03	438	1	5.759168	control
6562	2018-03-30	275	0	2.534567	control
9975	2018-01-06	381	0	2.232723	control
1042	2018-12-02	621	1	3.279372	treatment
1860	2018-09-07	505	0	7.124559	treatment
...	...	...	...	...	...
4403	2018-05-18	644	0	3.742603	treatment
355	2018-01-09	915	0	2.659615	treatment
16787	2018-11-04	695	0	0.943913	treatment
16841	2018-02-02	475	0	2.246929	control
7323	2018-04-25	976	1	4.504133	treatment

100 rows × 5 columns

```
ab_test_data.shape
```

```
(20000, 5)
```

```
ab_test_data['customer_segmnt'].value_counts()
# n1=n2=10000 => we can do t-test or z-test to compare means.
```

```
control      10000
treatment    10000
Name: customer_segmnt, dtype: int64
```

```
ab_test_data.describe()
```

	customer_id	premium	watch_time_hrs
count	20000.000000	20000.000000	20000.000000
mean	499.001650	0.176750	9.362542
std	288.223444	0.381467	244.884839
min	0.000000	0.000000	0.160268
25%	249.000000	0.000000	1.678066
50%	500.000000	0.000000	2.670953

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:21

KStest_Ttest.ipynb - Colaborat

colab.research.google.com/drive/1lWuS8AxULBBold2GlfBKUeT1RSj3fFAJ#scrollTo=bi-Mwf5ugQ8j

+ Code + Text RAM Disk

```
[ ] ab_test_data.describe()
```

	customer_id	premium	watch_time_hrs
count	20000.000000	20000.000000	20000.000000
mean	499.001650	0.176750	9.362542
std	288.223444	0.381467	244.884839
min	0.000000	0.000000	0.160268
25%	249.000000	0.000000	1.678066
50%	500.000000	0.000000	2.670953
75%	747.000000	0.000000	4.204673
max	999.000000	1.000000	10007.648185

in a day

outliers ...

→ X error/noise

```
[ ] # remove extreme values as we dont want them to impact means
    ab_test_data["watch_time_hrs"].quantile(0.999)

    # NOTE: only 24 hrs in a day
```

26.036198684124518

12 / 12

50%	500.000000	0.000000	2.670953
75%	747.000000	0.000000	4.204673
max	999.000000	1.000000	10007.648185

```
# remove extreme values as we dont want them to impact means
ab_test_data["watch_time_hrs"].quantile(0.999)

# NOTE: only 24 hrs in a day
```

99.9[th]

```
26.036198684124518
```

```
ab_test_data["watch_time_hrs"].quantile(0.998)
```

```
21.35607722117484
```

```
q998 = ab_test_data["watch_time_hrs"].quantile(0.998)
ab_test_data_no_out = ab_test_data[~(ab_test_data["watch_time_hrs"] > q998)]
```

```
# disb of watch-time
sns.histplot(ab_test_data_no_out['watch_time_hrs'], bins=100)
plt.show()
```

+ Code + Text

RAM
Disk

50%	500.000000	0.000000	2.670953
75%	747.000000	0.000000	4.204673
max	999.000000	1.000000	10007.648185

box plot or IQR

drop
0.2 percentiles

```
# remove extreme values as we dont want them to impact means
ab_test_data["watch_time_hrs"].quantile(0.999)

# NOTE: only 24 hrs in a day
```

```
26.036198684124518
```

```
ab_test_data["watch_time_hrs"].quantile(0.998)
```

```
21.356607722117484
```

```
q998 = ab_test_data["watch_time_hrs"].quantile(0.998)
ab_test_data_no_out = ab_test_data[~(ab_test_data["watch_time_hrs"] > q998)]
```

```
# disb of watch-time
sns.histplot(ab_test_data_no_out['watch_time_hrs'], bins=100)
plt.show()
```

14 / 16

50%	500.000000	0.000000	2.670953
75%	747.000000	0.000000	4.204673
max	999.000000	1.000000	10007.648185

```
# remove extreme values as we dont want them to impact means
ab_test_data["watch_time_hrs"].quantile(0.999)

# NOTE: only 24 hrs in a day
```

```
26.036198684124518
```

99.8

0.2 percentiles

```
ab_test_data["watch_time_hrs"].quantile(0.998)
```

```
21.356607722117484
```

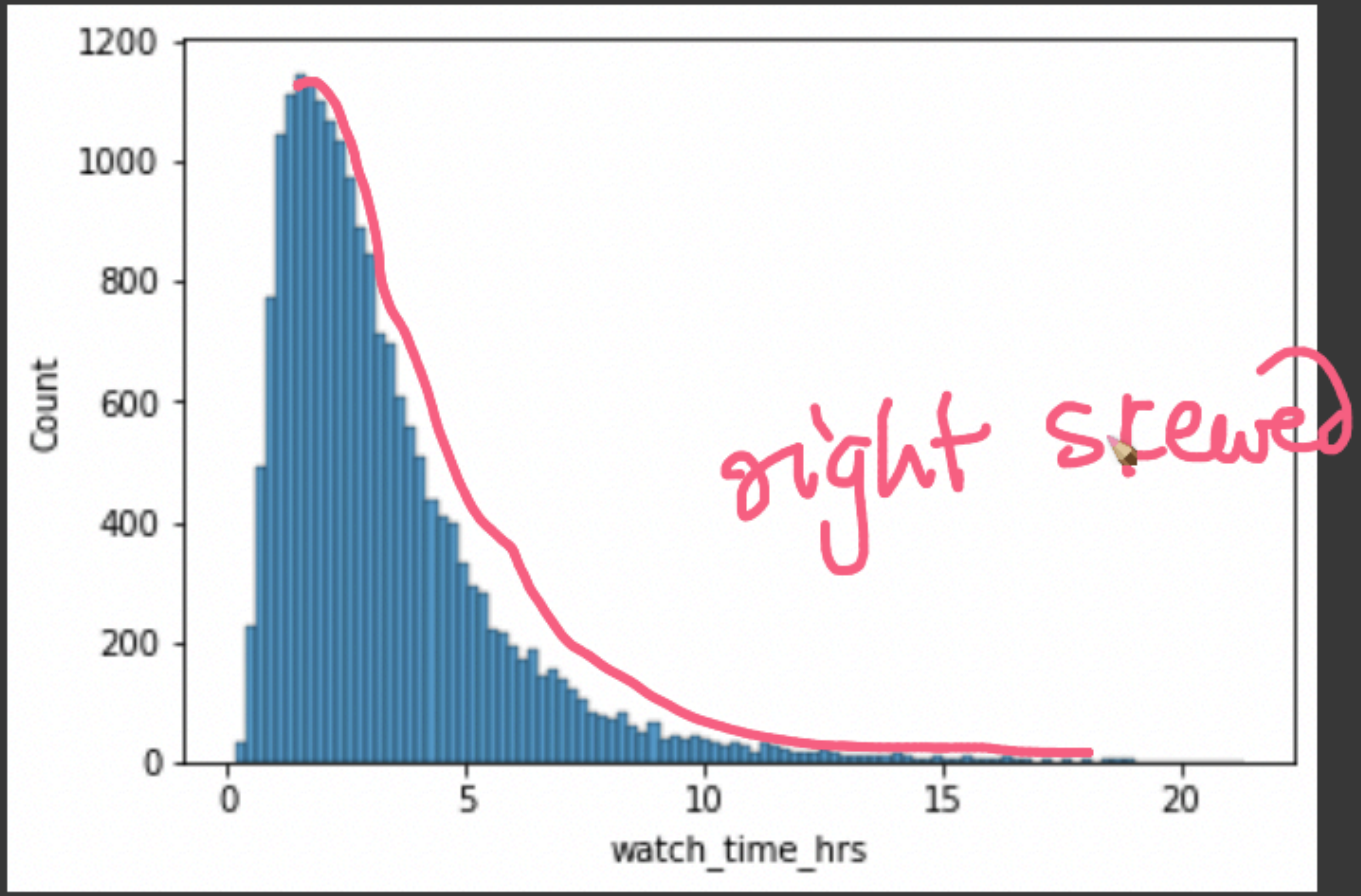
```
q998 = ab_test_data["watch_time_hrs"].quantile(0.998)
ab_test_data_no_out = ab_test_data[~(ab_test_data["watch_time_hrs"] > q998)]
```

21.35 hrs

```
# disb of watch-time
sns.histplot(ab_test_data_no_out['watch_time_hrs'], bins=100)
plt.show()
```

```python
q998 = ab_test_data["watch_time_hrs"].quantile(0.998)
ab_test_data_no_out = ab_test_data[~(ab_test_data["watch_time_hrs"] > q998)]
```

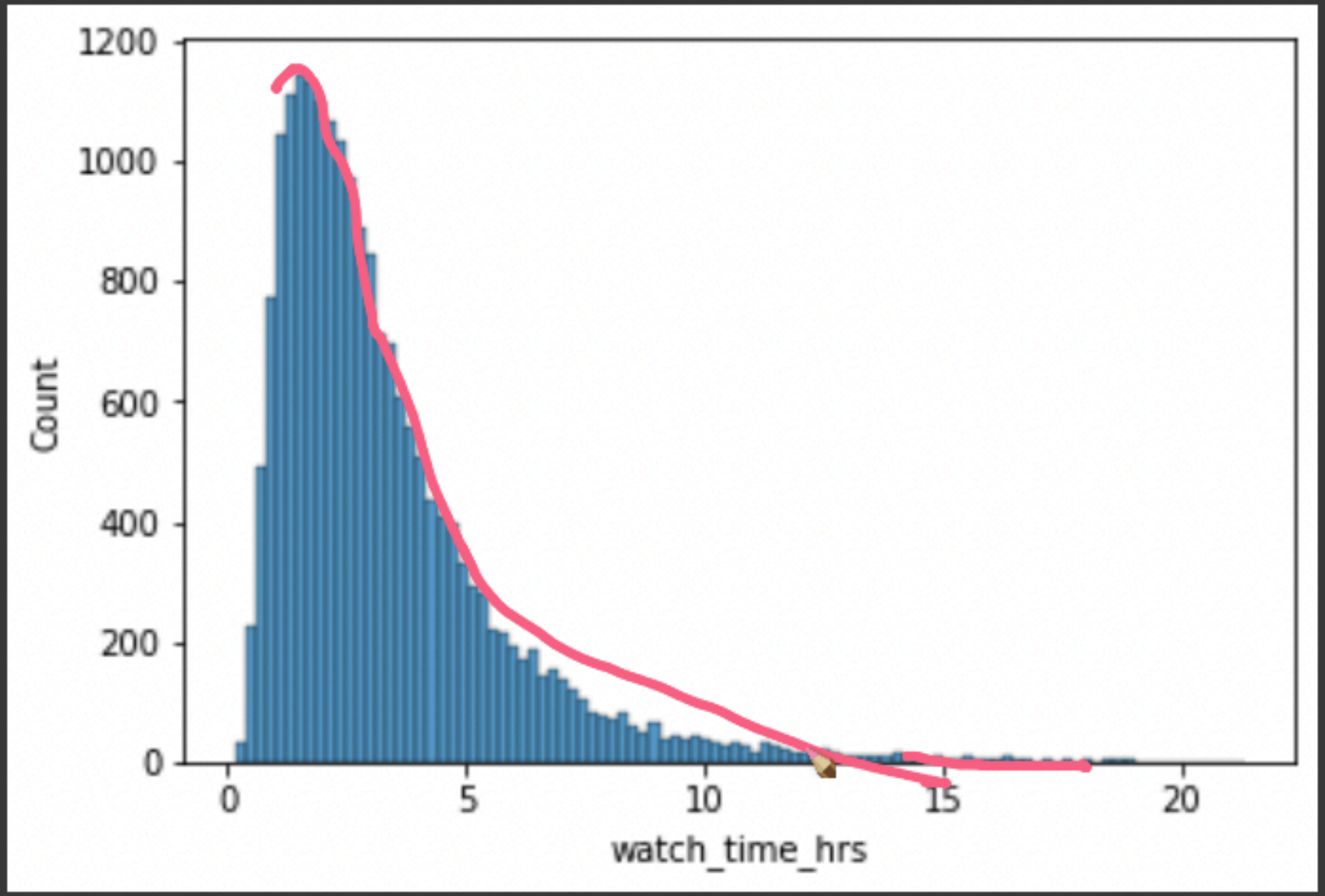
```python
# disb of watch-time
sns.histplot(ab_test_data_no_out['watch_time_hrs'], bins=100)
plt.show()
```




```python
#split the data
```

```
q998 = ab_test_data["watch_time_hrs"].quantile(0.998)
ab_test_data_no_out = ab_test_data[~(ab_test_data["watch_time_hrs"] > q998)]
```

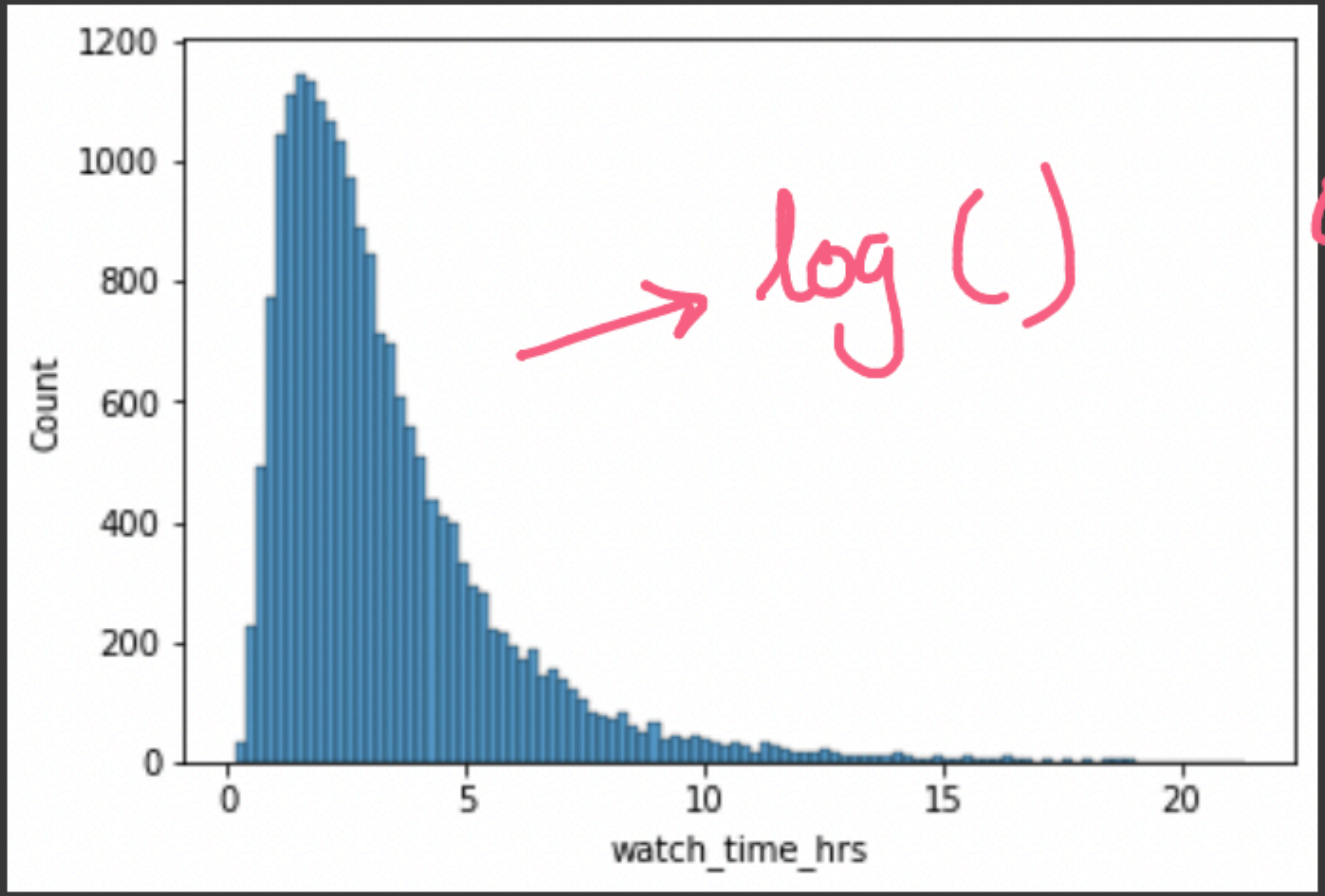
```
# disb of watch-time
sns.histplot(ab_test_data_no_out['watch_time_hrs'], bins=100)
plt.show()
```




```
#split the data
```

```
q998 = ab_test_data["watch_time_hrs"].quantile(0.998)
ab_test_data_no_out = ab_test_data[~(ab_test_data["watch_time_hrs"] > q998)]
```

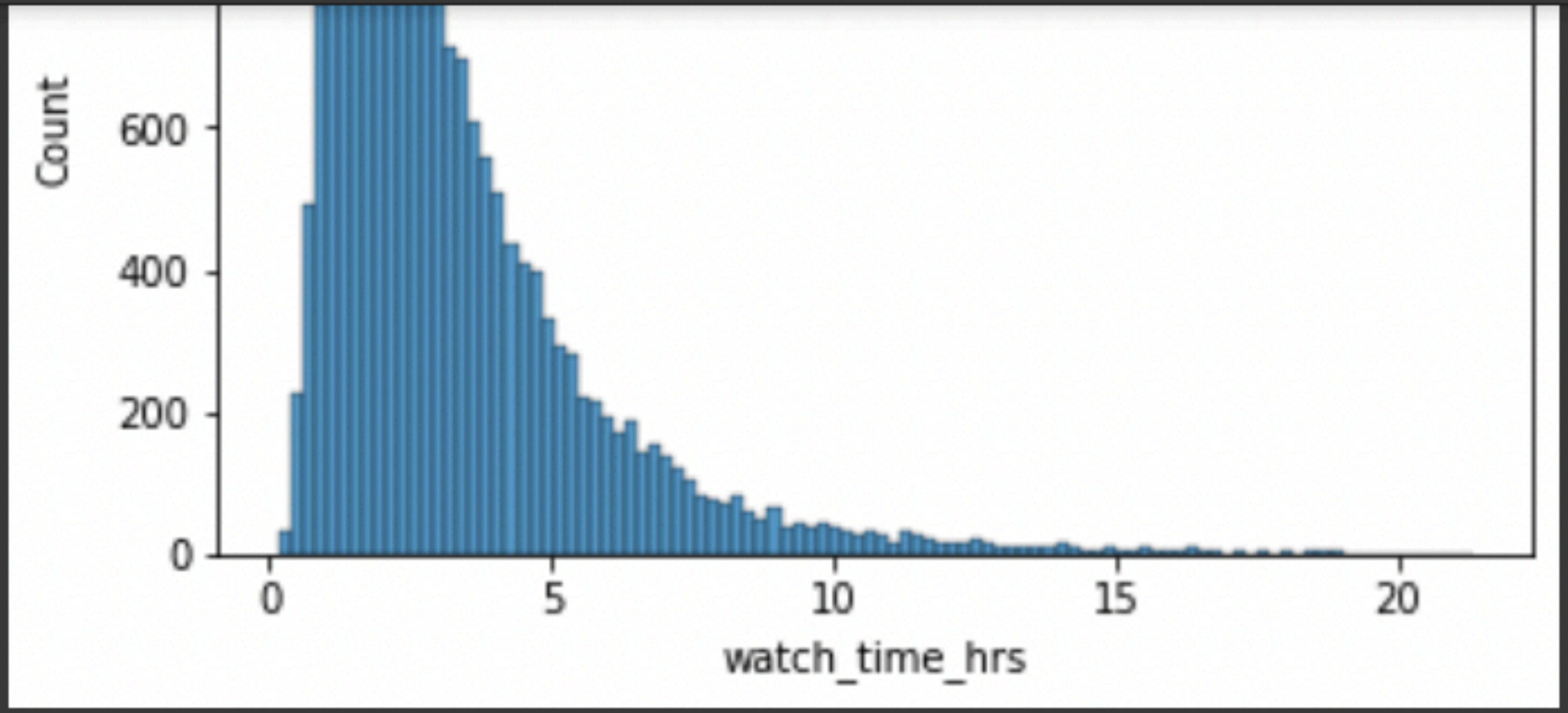
```
# disb of watch-time
sns.histplot(ab_test_data_no_out['watch_time_hrs'], bins=100)
plt.show()
```



→ log ( )

qqplot with Gaussian ✓

```
#split the data
ab_test_control_data = ab ... r_segmnt"] == "control"]
```



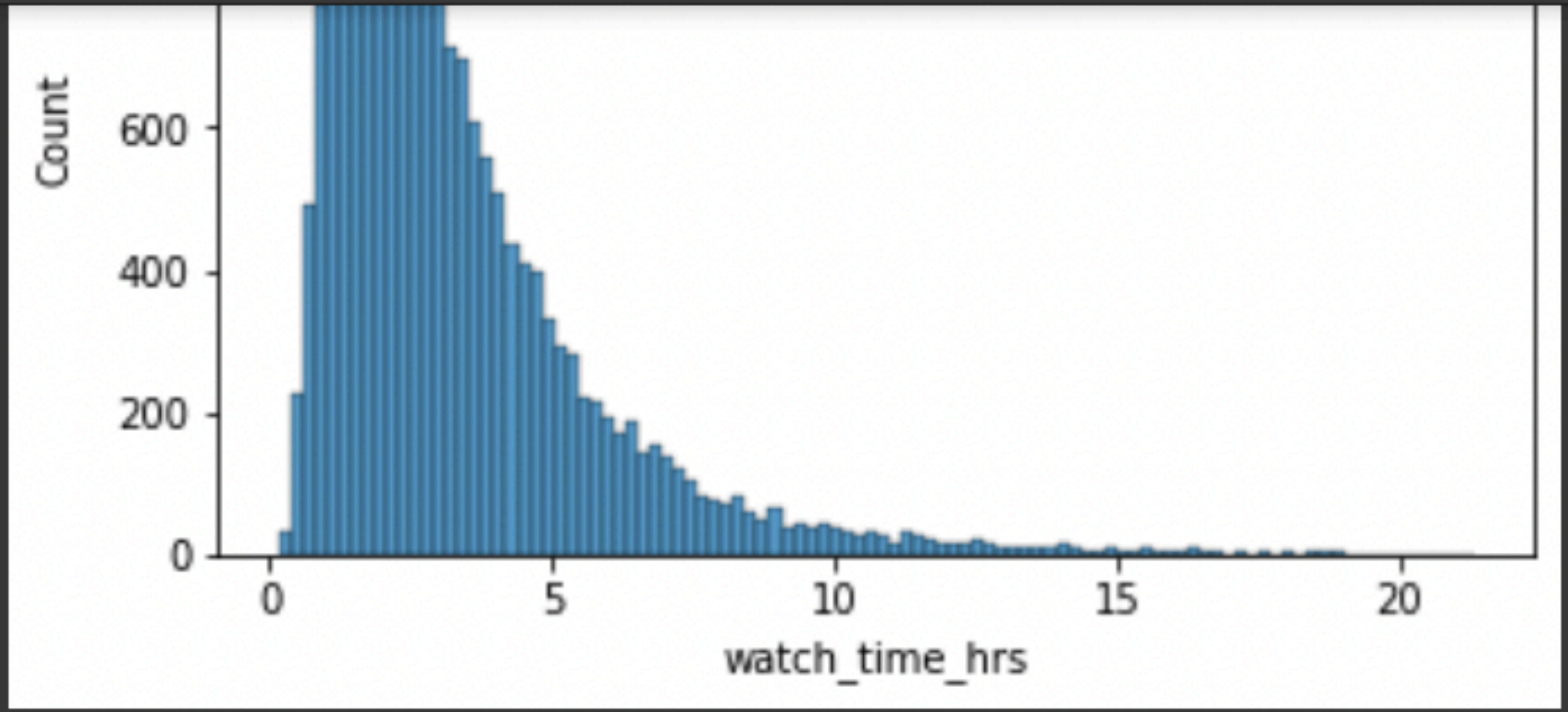

```
#split the data
ab_test_control_data = ab_test_data_no_out[ab_test_data_no_out["customer_segmnt"] == "control"]
ab_test_treatment_data = ab_test_data_no_out[ab_test_data_no_out["customer_segmnt"] == "treatment"]
```

```
ab_test_control_data.shape
```

```
(9973, 5)
```

```
ab_test_treatment_data.shape
```

```
(9987, 5)
```




```
#split the data
ab_test_control_data = ab_test_data_no_out[ab_test_data_no_out["customer_segmnt"] == "control"]
ab_test_treatment_data = ab_test_data_no_out[ab_test_data_no_out["customer_segmnt"] == "treatment"]
```

```
ab_test_control_data.shape
```

n_1 (9973, 5)

```
ab_test_treatment_data.shape
```

n_2 (9987, 5)

```
ab_test_control_data.shape
```

(9973, 5)

```
ab_test_treatment_data.shape
```

(9987, 5)

t-test (v)

```
dof = ab_test_control_data.shape[0] + ab_test_treatment_data.shape[0] - 2
dof
```

9987 9973

19958

```
diff_means = ab_test_control_data["watch_time_hrs"].mean() - ab_test_treatment_data["watch_time_hrs"].mean()
diff_means
```

0.5556665488445294

```
#2 sample t-test
stats.ttest_ind(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

```
ab_test_control_data.shape
```

(9973, 5)

t-disb(large v) → Gaussian

```
ab_test_treatment_data.shape
```

(9987, 5)

```
dof = ab_test_control_data.shape[0] + ab_test_treatment_data.shape[0] - 2
dof
```

19958

```
diff_means = ab_test_control_data["watch_time_hrs"].mean() - ab_test_treatment_data["watch_time_hrs"].mean()
diff_means
```

0.555665488445294

```
#2 sample t-test
stats.ttest_ind(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

KStest_Ttest.ipynb - Colaboratory

colab.research.google.com/drive/1lWuS8AxULBBold2GlfBKUeT1RSj3fFAJ#scrollTo=g-YGC1NQh78B

+ Code + Text

RAM
Disk

n1

```
(9973, 5)
```

```
[ ] ab_test_treatment_data.shape
```

n2

```
(9987, 5)
```

```
[ ] dof = ab_test_control_data.shape[0] + ab_test_treatment_data.shape[0] - 2
    dof
```

```
19958
```

1ad

2ads

```
diff_means = ab_test_control_data["watch_time_hrs"].mean() - ab_test_treatment_data["watch_time_hrs"].mean()
diff_means
```

```
0.5556665488445294
```

```
[ ] #2 sample t-test
    stats.ttest_ind(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

```
Ttest_indResult(statistic=15.96034913022092, pvalue=5.438408586231319e-57)
```

23 / 23

```
dof = ab_test_control_data.shape[0] + ab_test_treatment_data.shape[0] - 2
dof
```

19958

```
diff_means = ab_test_control_data["watch_time_hrs"].mean() - ab_test_treatment_data["watch_time_hrs"].mean()
diff_means
```

0.5556654884452 94

t-test

$$T_t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

```
#2 sample t-test
stats.ttest_ind(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

Ttest_indResult(statistic=15.96034913022092, pvalue=5.438408586231319e-57)

```
# 2-sample z-test as n1 nad n2 are large.
# Refer: https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html
from statsmodels.stats.weightstats import ztest as ztest
ztest(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is depre
  import pandas.util.testing as tm
(15.96034913022092, 2.4137738128170024e-57)

```
dof = ab_test_control_data.shape[0] + ab_test_treatment_data.shape[0] - 2
dof
```

```
19958
```

```
diff_means = ab_test_control_data["watch_time_hrs"].mean() - ab_test_treatment_data["watch_time_hrs"].mean()
diff_means
```

```
0.5556654884452940
```

```
#2 sample t-test
stats.ttest_ind(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

```
Ttest_indResult(statistic=15.96034913022092, pvalue=5.438408586231319e-57)
```

$5 \times 10^{-57} \ll \alpha = 1\%$

reject H_0 ✓

```
# 2-sample z-test as n1 nad n2 are large.
# Refer: https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html
from statsmodels.stats.weightstats import ztest as ztest
ztest(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is depre
  import pandas.util.testing as tm
(15.96034913022092, 2.4137738128170024e-57)
```

```
diff_means = ab_test_control_data["watch_time_hrs"].mean() - ab_test_treatment_data["watch_time_hrs"].mean()
diff_means
```

```
0.5556654884452942
```

n_1+n_2-2

T_t

```
#2 sample t-test
stats.ttest_ind(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

```
Ttest_indResult(statistic=15.96034913022092, pvalue=5.438408586231319e-57)
```

p-val_t

```
# 2-sample z-test as n1 nad n2 are large.
# Refer: https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html
from statsmodels.stats.weightstats import ztest as ztest
ztest(ab_test_control_data["watch_time_hrs"], ab_test_treatment_data["watch_time_hrs"])
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is depre
  import pandas.util.testing as tm
(15.96034913022092, 2.413773812817002e-57)
```

ztest T_z

p-val_z

T dist = z-dist as $\nu \to \infty$

[Z-proportions test]

Retail website

① A: old webpage → $\text{conv. rate}_1 = \dfrac{\text{conversations}_1}{\text{visits}_1} = \hat{p}_1$

1000

② B: new webpage → $\text{conv-rate}_2 = \dfrac{\text{conv}_2}{\text{visits}_2} = \hat{p}_2$

500

pop. prop

$H_0: p_1 = p_2$

$H_a: p_1 \neq p_2$

Conditions

→ sample size ≥ 30 ✓

→

28 / 31

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Sensible

$$T_{zp} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$0 \le \hat{p}_1 = \frac{\text{conv}_1}{\text{visits}_1} \le 1$$

$$0 \le \hat{p}_2 = \frac{\text{conv}_2}{\text{visits}_2} \le 1$$

under $H_0 \rightarrow (T_{zp} \rightarrow 0)$

Under $H_a \rightarrow$ larger +/-ve value

$$\hat{p} = \frac{\text{conv}_1 + \text{conv}_2}{\text{visits}_1 + \text{visits}_2}$$

$$0 \le \hat{p} \le 1$$

29 / 31

proof

$T_{zp} \sim Z(0,1)$ under H_0

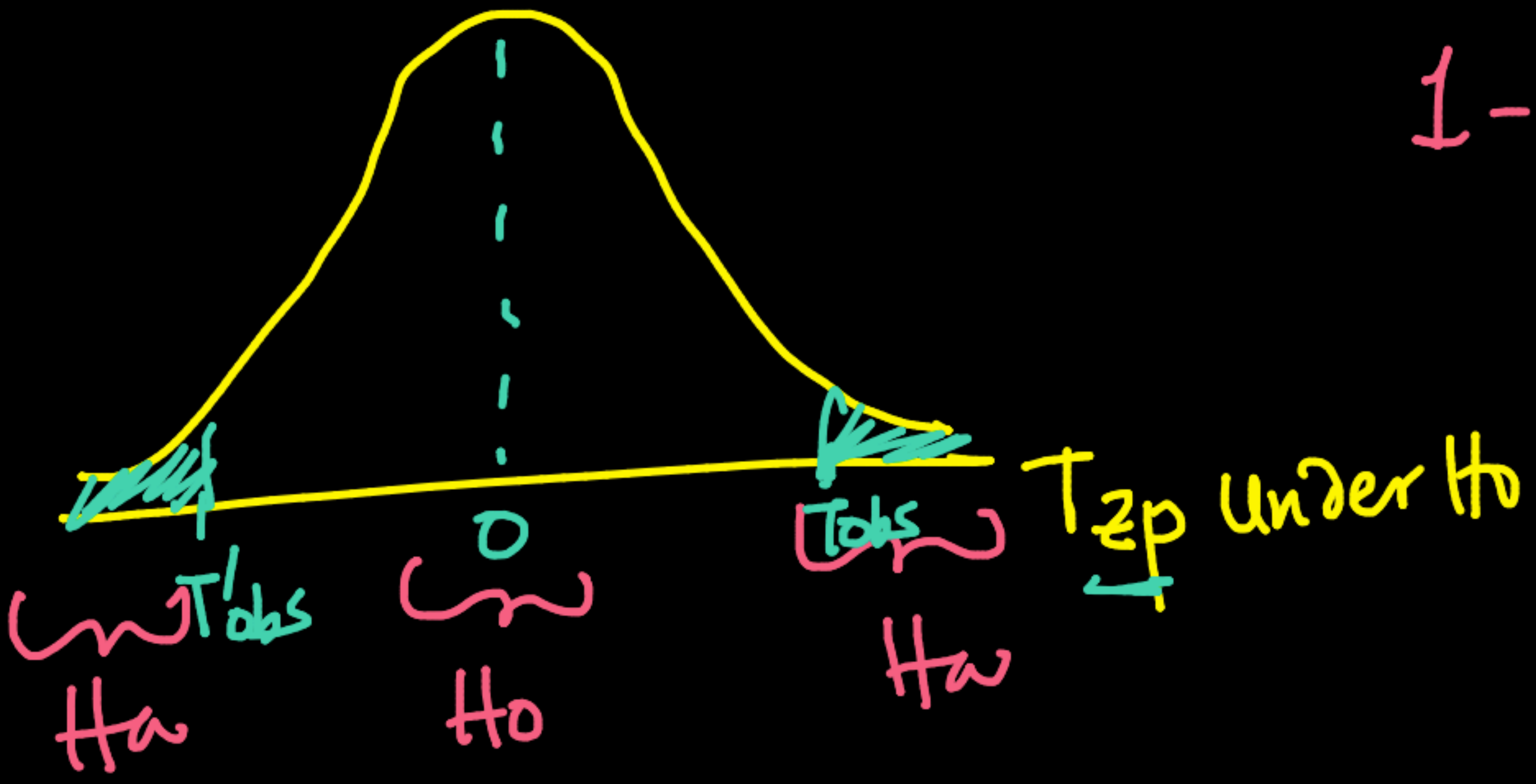
$= N(0,1)$

2-sided / 1-sided

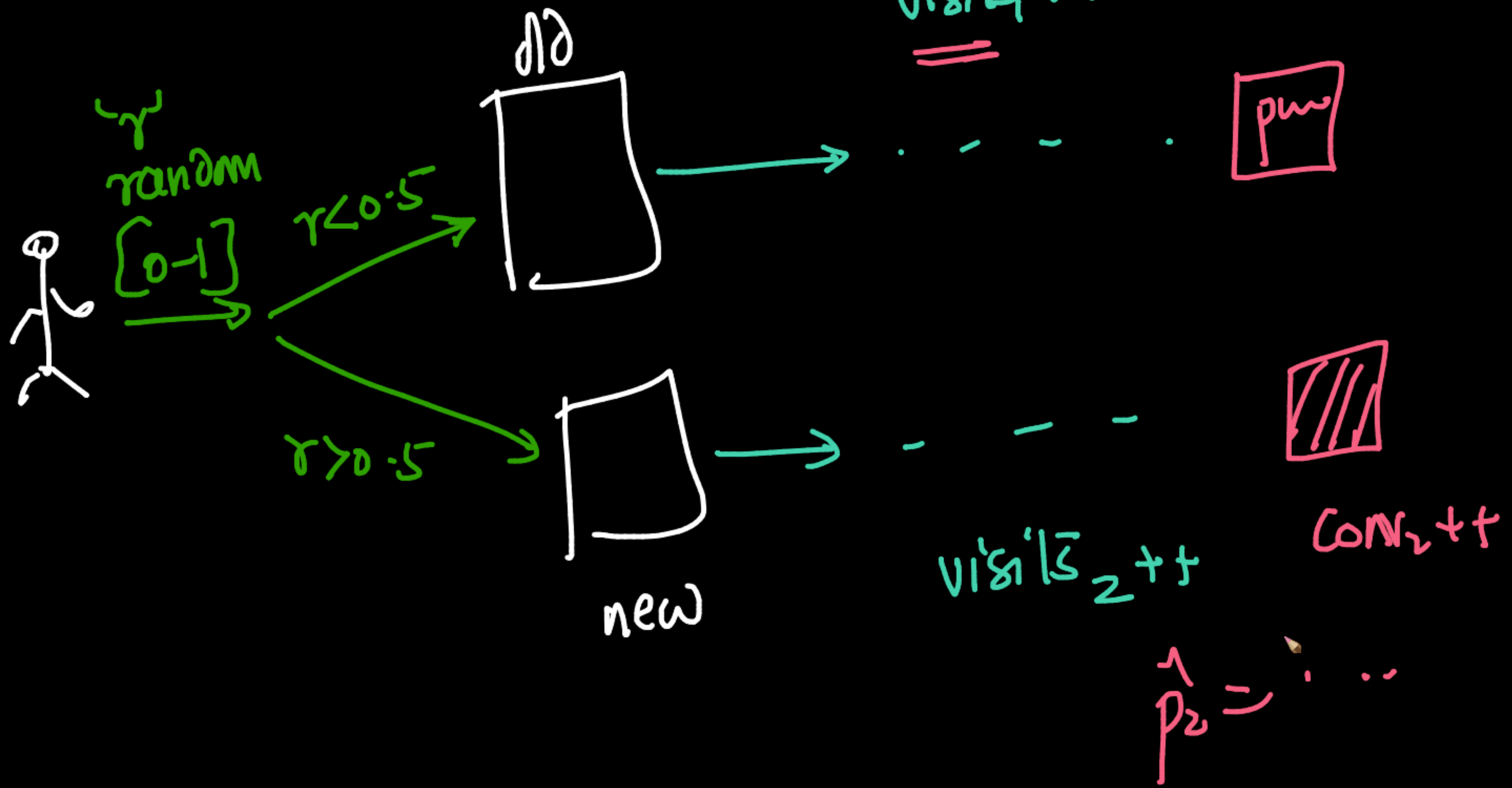
T_{obs}

$\boxed{} = \text{p-val}$

vs α



Framework



$\hat{p}_1 =$
random
r
[0-1]
r<0.5
r>0.5
old
new
visits₁++
conv₁++
pw
visits₂++
conv₂++
$\hat{p}_2 =$

Q

X

100

$x_1, x_2 \ldots x_{100}$

30 kms/ltr

< 30 kms/hr

$H_0: \bar{X}_{mean} = 30$ kms/hr

$H_a: \bar{X}_{mean} \neq 30$ km/hr

95 C.I

m

29

Z-test

T-test

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:53

H_0 : $\bar{X}$ mean $= 30$

Z-test

H_a : $\bar{X}$ mean < 30

Left tailed test

$T_Z \sim Z$ under H_0

33 / 33

Compare Mean n_1 & n_2 are large $\longrightarrow$ Z-test

" n_1 & n_2 are small $\longrightarrow$ T-test

Compare Prop. n_1 & n_2 are large $\longrightarrow$ Z-P-test

⋮

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:03

[Mutually exclusive] → H_0 :
→ H_a :

but not necessarily exhaustive

$\bar{X}_{mean} = 30$

$\bar{X}_{mean} < 30$

↳ left tailed test

~~$\bar{X}_{mean} < 30$~~

35 / 37

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:02

Z-prop : param / non-param

V_1 & V_2 > 30 ✓

prop:

$T_{zp} \sim Z(0,1)$

bernouli r.V

conv

non conv ✓

PROOF

H T

Conv nonconv

36 / 37

Z-test

1

$x_{11} \quad x_{12} \quad \cdots \quad x_{1n_1}$

pop μ & σ
should be finite

2

$x_{21} \quad x_{22} \quad \cdots \quad x_{2n_2}$

χ^2-test

[Test of independence]

Biz:

regular → premium

Q: does gender impact premium or not?

[isPremium indep of Gender]

isPremium: 1, 0

Gender: M, F

L- IQ
M-IQ
H-IQ

$\longrightarrow$

L-INC
M-INC
H-ING

Setup:

H_0: Gender has no effect on premium sub

H_a: Gender has some effect " "

observed: Data

Contigency Table:

Gender \ isPr	0	1	
F	346	149	→ 495
M	339	165	→ 504
	↓ 685	↓ 314	$\boxed{999}$

Observed:

G \ P	0	1	
F	O_1: 346	O_2: 149	495
M	O_3: 339	O_4: 165	504
	685	314	999

Under H_0; Expected

G \ P	0	1
F	E_1: $\frac{495 \times 685}{999}$ = 339	E_2: $\frac{495 \times 314}{999}$ = 155
M	E_3: $\frac{504 \times 685}{999}$ = 345	E_4: $\frac{504 \times 314}{999}$ = 158

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:24

$$T_{\chi^2} = \sum_{i=1}^{c \text{ cells}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(K)$$

$K \rightarrow (n_{rows}-1) \times (n_{cols}-1)$

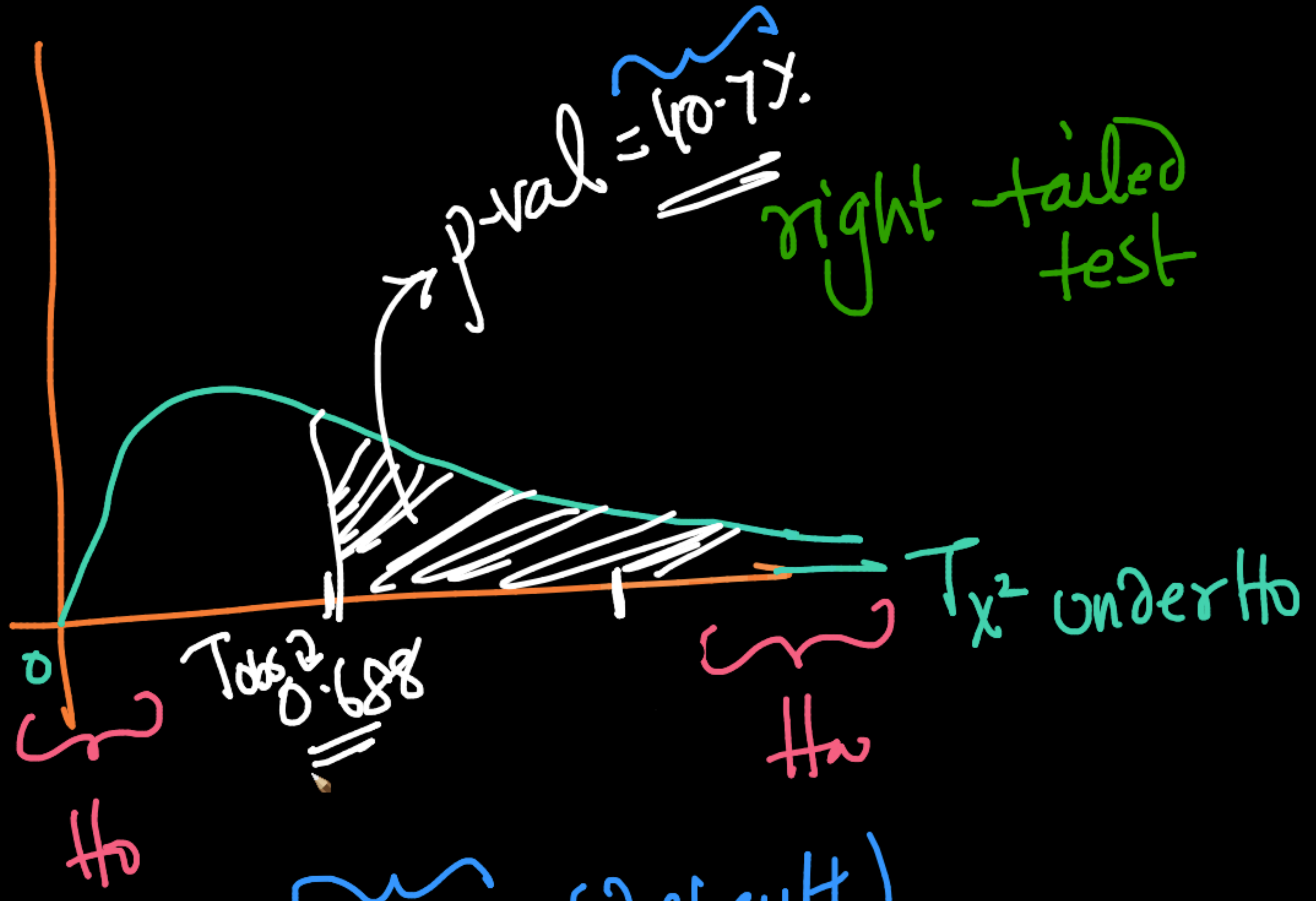
$K = 1 \times 1 = 1$

Under H_0 : $T_{\chi^2} \to 0$

Under H_a : T_{χ^2} large +ve value

43 / 45

Under H_0: T



right tailed test

accept H_0 ✓

$\underline{\alpha = 5\%}$ (default)

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:30

Ho | Ha

✓ 40.7%

p-val > α

40.7% 5%

as extreme as

$$P(\text{observing } T = 0.688 \mid H_0) = 40.7\%$$

✓

$\Rightarrow$ cannot reject Ho

$\Downarrow$

accept Ho

45 / 48

Continous → PCC & SRCC (next class)

Continous → bucketize → categorical → χ^2-test

L-IQ	L-INC
M-IQ	M-INC
H-IQ	H-INC

10:31

ANOVA

X 2

$K \geq 2$

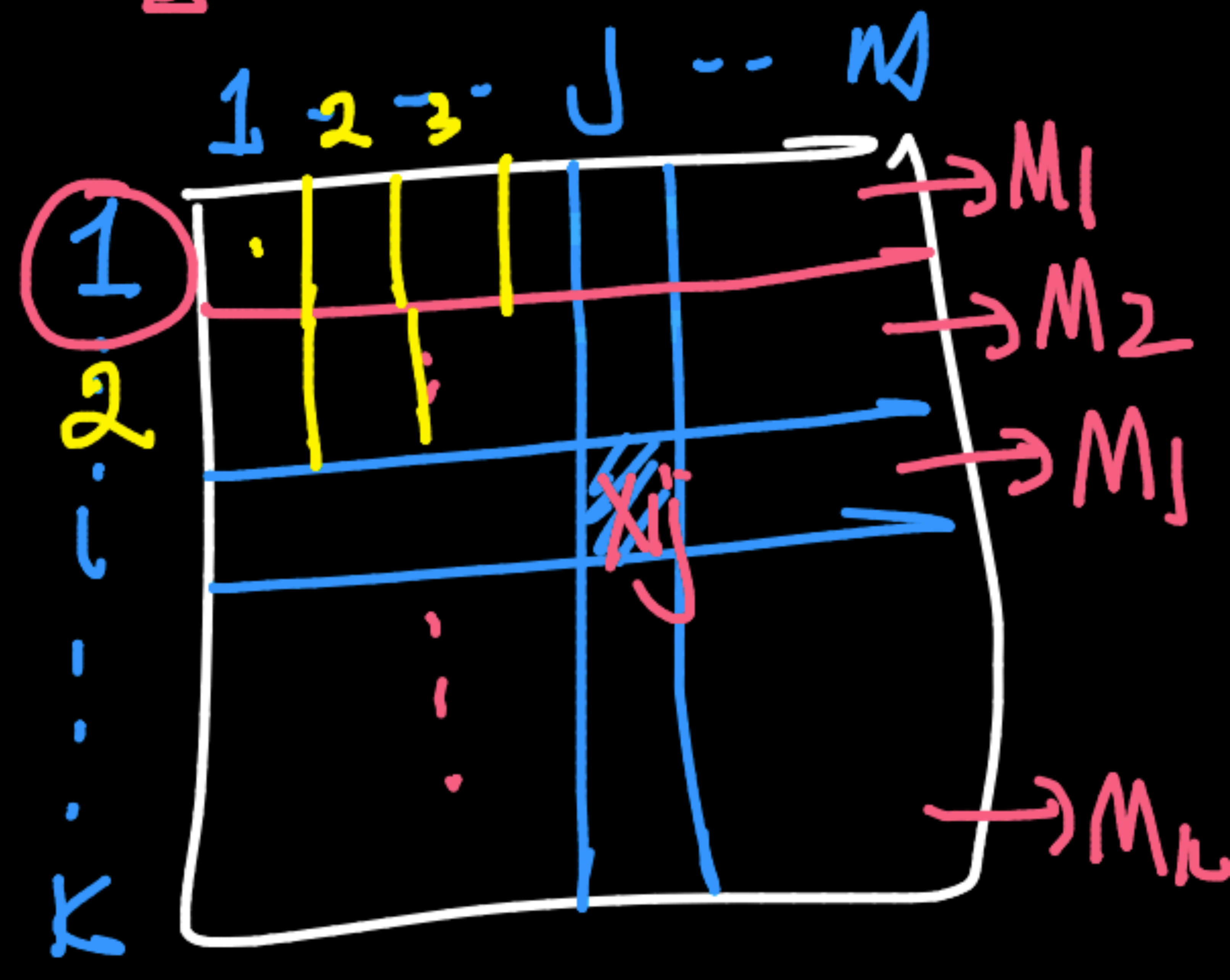
Observed data

→ K - drugs → K - groups

✓ → m patients per group

→ $n = m \times K$

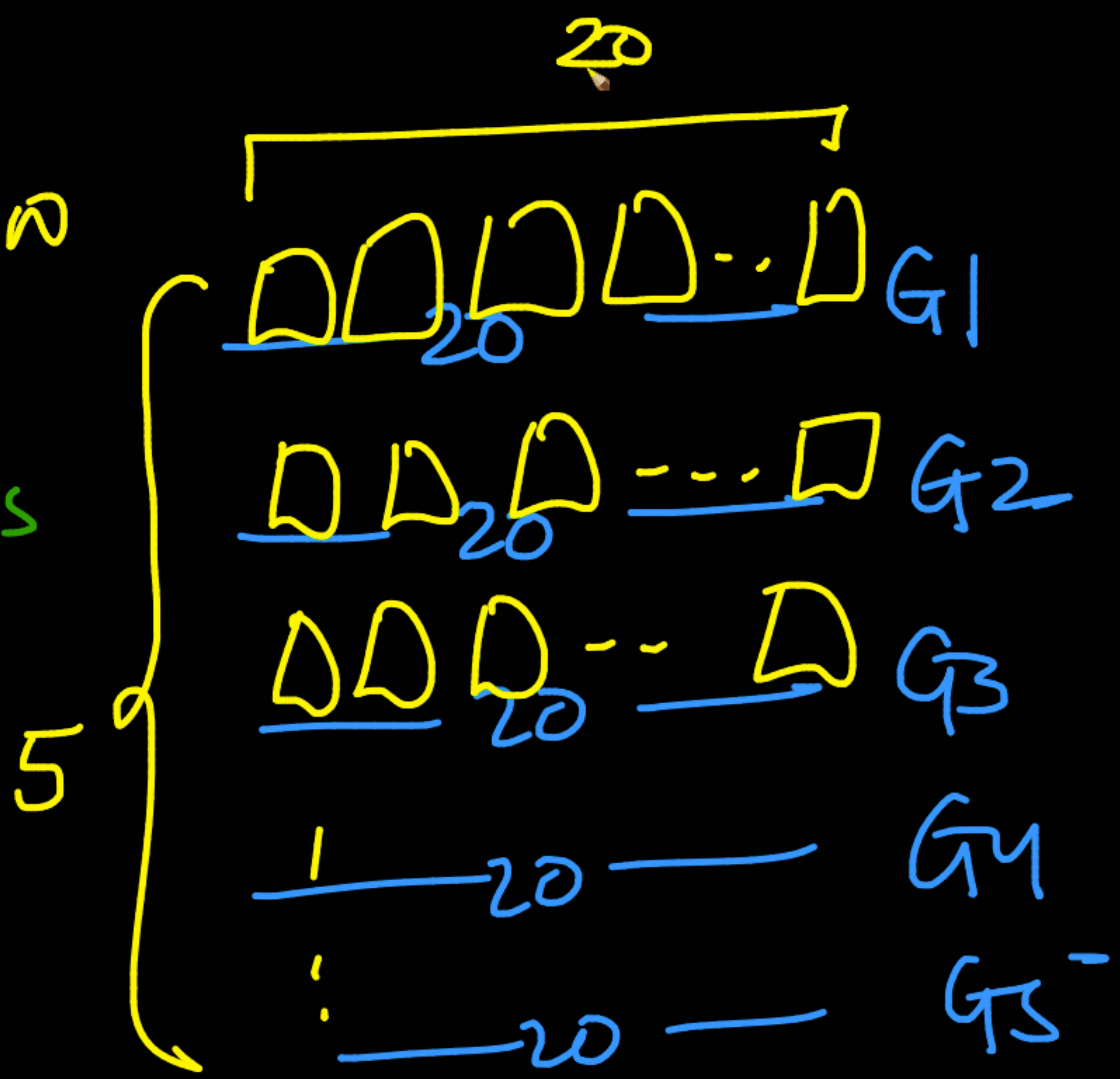
patients



X_{ij} = rec. time of j^{th} patient in i^{th} group

P1 P2 P3 P4 --- P100

5 grps



Task: all these medicines have same avg rec. time

H_0: there is no difference between group means

H_a: there is some diff b/w group means

$\boxed{Q}$ ✓

K - groups

? $\boxed{}$ t-tests

${}^kC_2$

$k = 10$

${}^kC_2 = \frac{10 \times 9}{2} = \boxed{45}$

$k = 5$ ✗

1, 2, 3, 4, 5

1 - 2 → ✓

1 - 3 → ✓

10

1 - 4

1 - 5

2 - 3 → ✗

2 - 4

2 - 5

3 - 4

4 - 5

3 - 5

$k = 10$ groups

$\boxed{{}^{k}C_2 = 45}$ t-tests

compute
& not an
issue

{ HINT: $\alpha = 5\%$

can this mess up something?

$t_1 \rightarrow$ Hb

$t_2 \rightarrow$ Hb

t_3

$\vdots \qquad \vdots$

$t_{45} \rightarrow$ Hb

(let) $\alpha = 5\%$

there is a 5% accept the incorrectly (prev class)

mistake

✓ 45 tests

1 or more tests

Q

Prob of making a mistake (accepting Ha incorrectly) = 5% = p(H)

losses

45 tests

$$P\left(\text{1 or more tests are making a mistake}\right) = ?$$

$$= 1 - P(\text{no mistake})$$

$$P(\geq 1 \text{ heads in } 45 \text{ } \cancel{\text{heads}} \text{ tosses})$$

$$= 1 - P(45 \text{ tails in } 45 \text{ tosses})$$

$$= \boxed{1 - 0.95^{45}} = 90\%$$

$P(\text{heads}) = 5\%$

45 tosses

$$1 - P(0 \text{ heads}) = 1 - P(45 \text{ tails})$$

$$= 1 - (0.95)(0.95)(0.95) \cdots (0.95)$$

$$= \underline{1 - 0.95^{45}}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:34

ANOVA

f-test

Test-statistic

$$f = \frac{MSB}{MSW} = \frac{\text{Mean-ss-dist between}}{\text{Mean-ss-dist-within}}$$

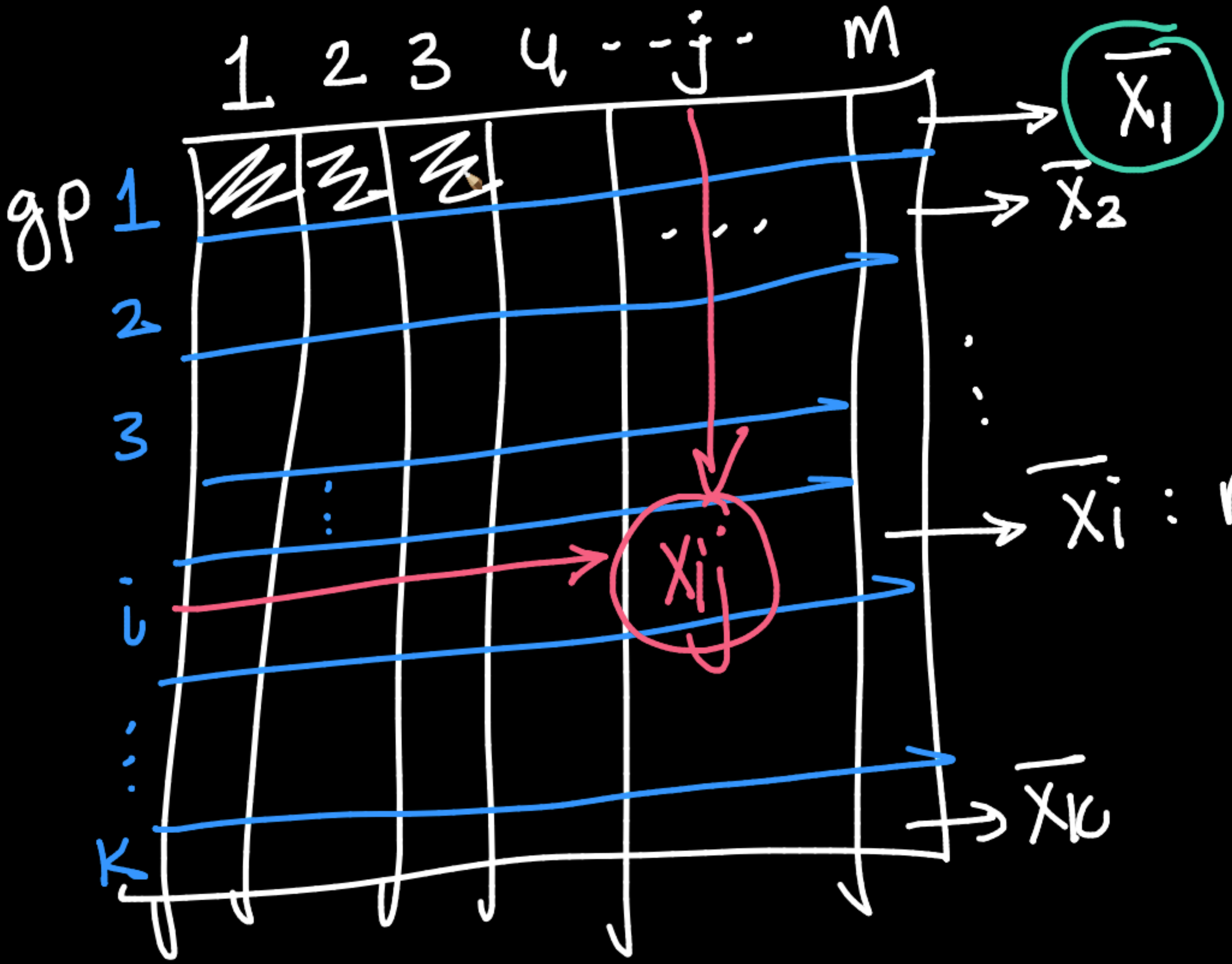
DOF/params

F-distb(K-1, n-K)

KxM

58 / 73

X :



$n = m \times k$

$\bar{x}_i$: mean. rec. time of group i

$\bar{\bar{X}}$ = overall mean rec. time

Chrome File Edit View History Bookmarks Profiles Tab Window Help

$$f = \frac{MSB}{MSW}$$

$$\sum_{i=1}^{K} (\bar{X}_i - \bar{\bar{X}})^2 / (K-1)$$

$$\sum_{i=1}^{K} \underbrace{\sum_{j=1}^{m} (X_{ij} - \bar{X}_i)^2}_{\text{within gp ss diff}} \Big/ \underbrace{(n-K)}_{\text{Unbiased estimate}}$$

$n = m \times K$

60 / 63

$(k-1) \longrightarrow$ unbiased estimate

$k \longrightarrow$ biased estimate

[later class]

61 / 62

Chrome File Edit View History Bookmarks Profiles Tab Window Help

ANOVA assumptions

→ each groups data are gaussian disb → close to gaussian

k-gps

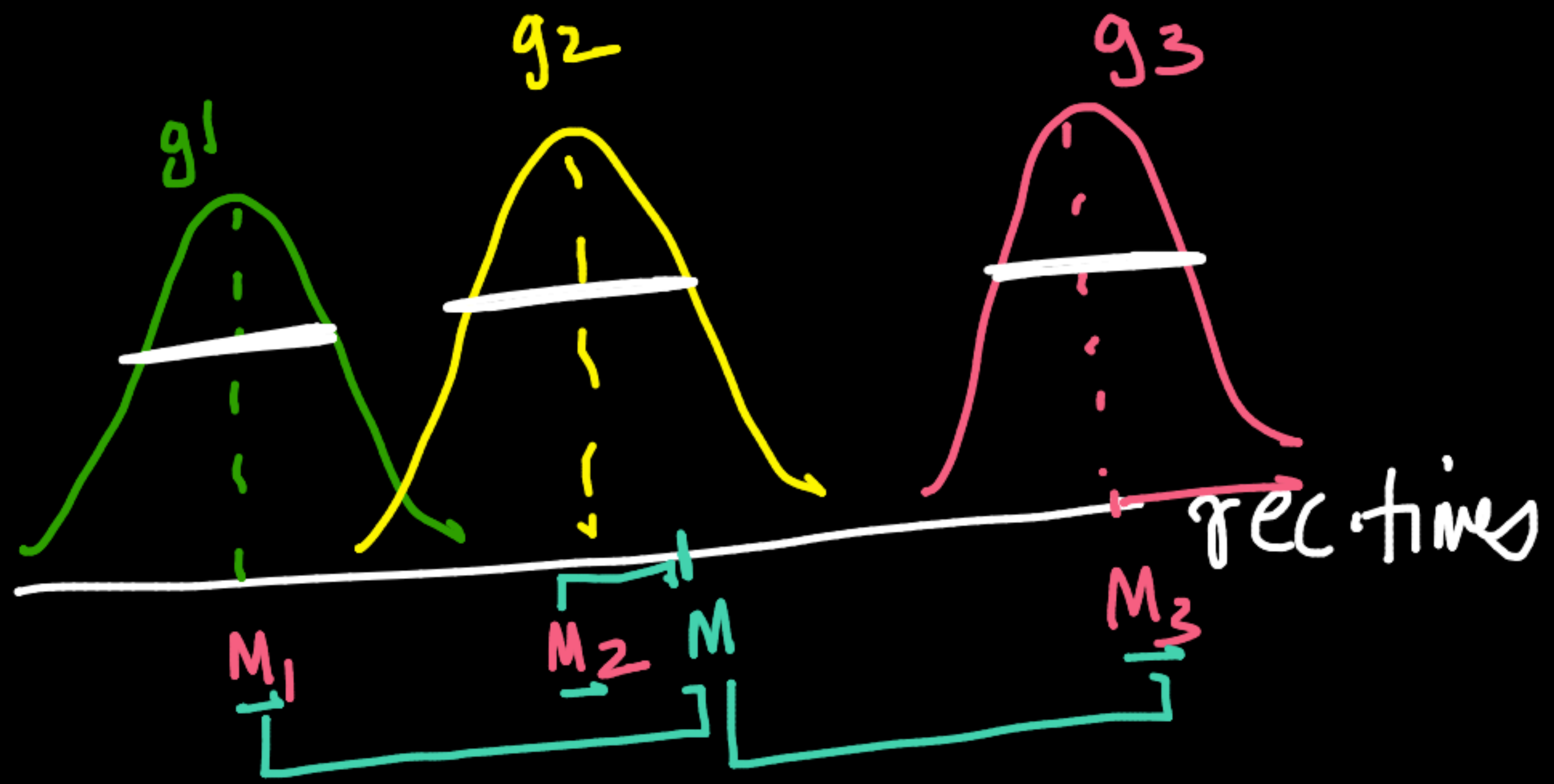
→ [Var of each-gp is same / similar] → close (practice)

→ n obs are indep of one another

parametric test

62 / 73

$$f = \frac{MSB}{MSW}$$



Case1:

$$f_1 = \frac{MSB\uparrow}{MSW\uparrow}$$

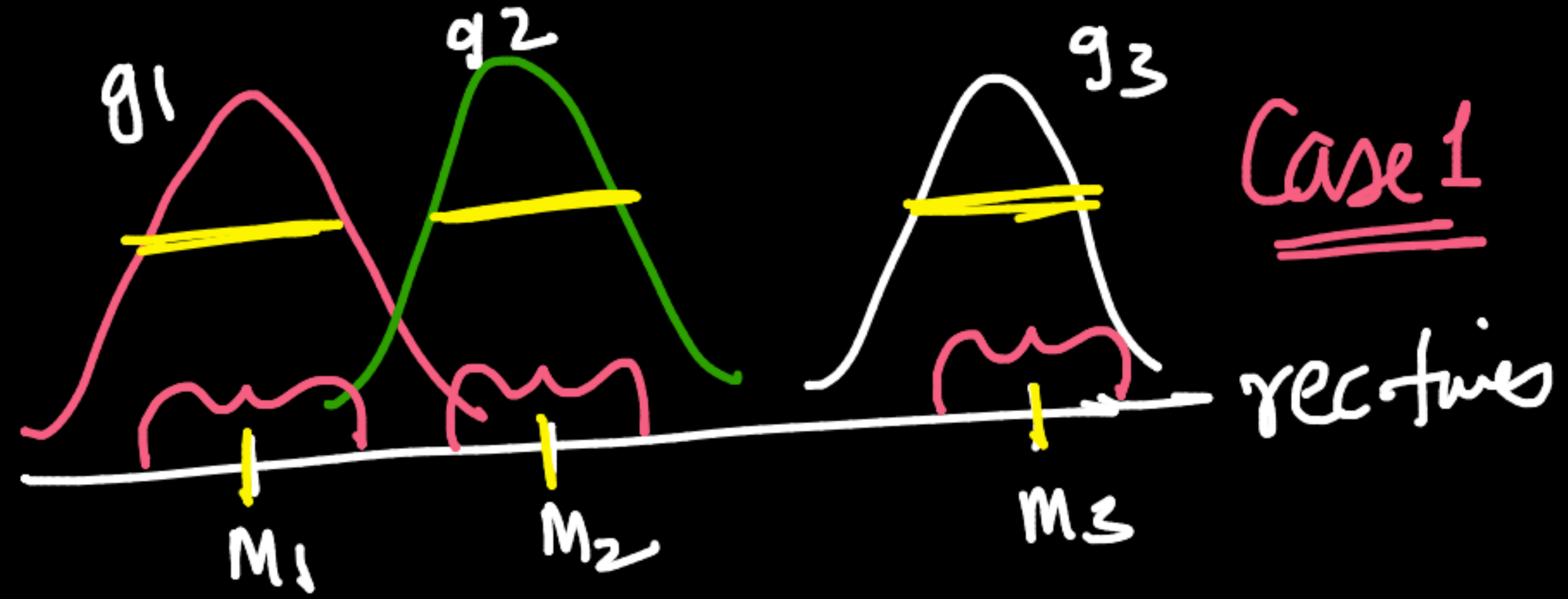
$$f = \frac{MSB}{MSW}$$

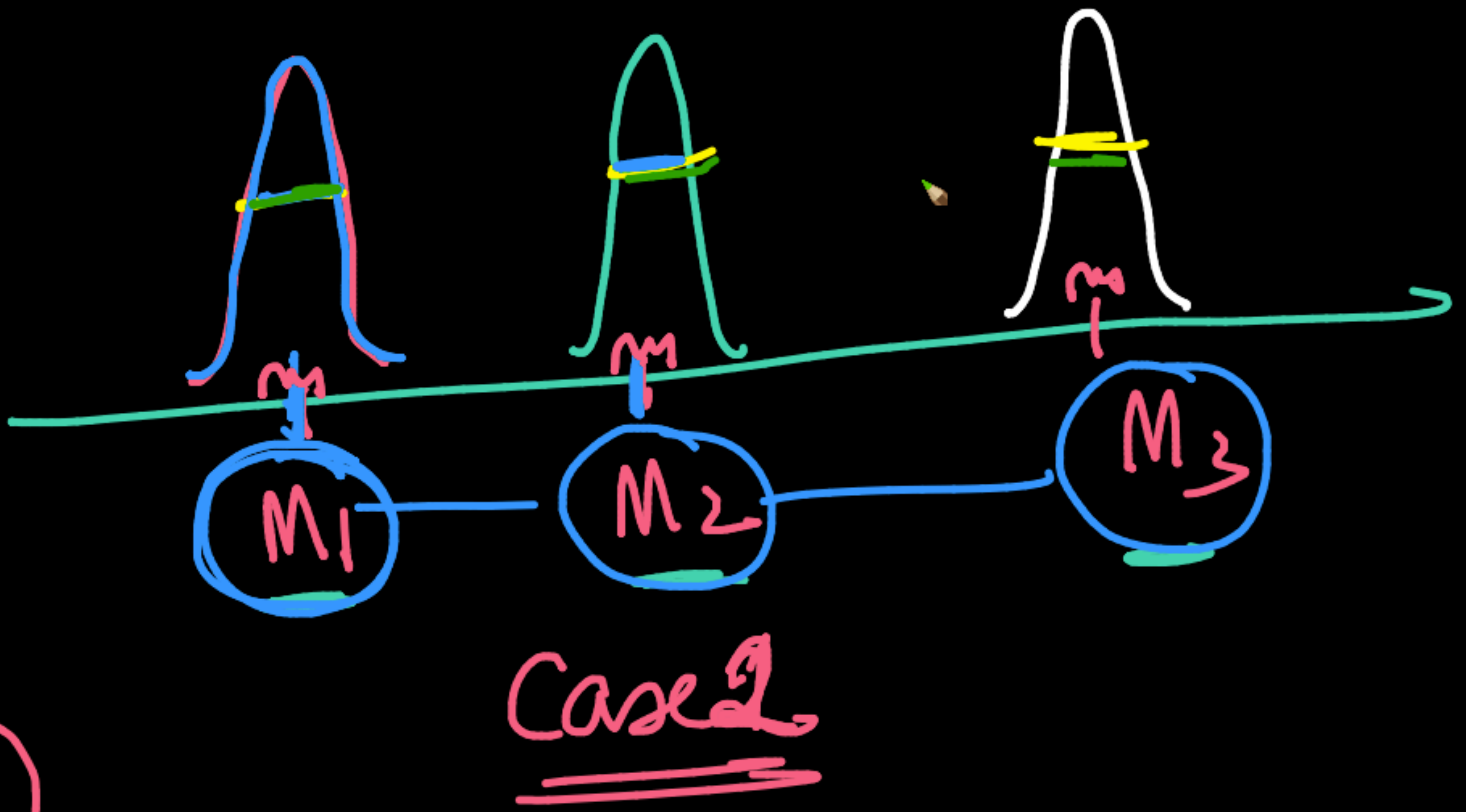
$$f_1 : \frac{MSB_1}{MSW_1}$$

$$f_2 : \frac{MSB_2}{MSW_2}$$

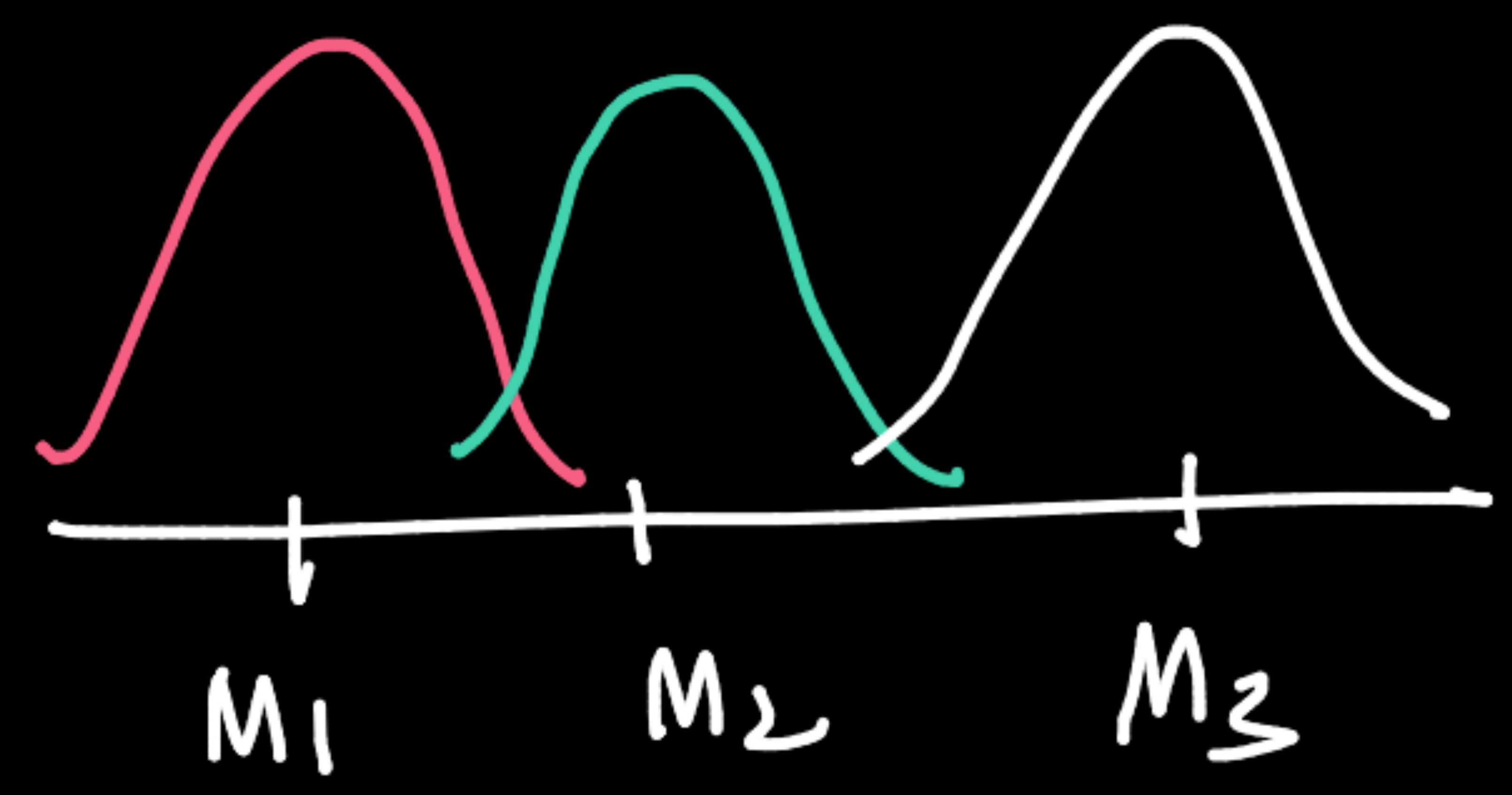
$$f_2 > f_1$$

Q



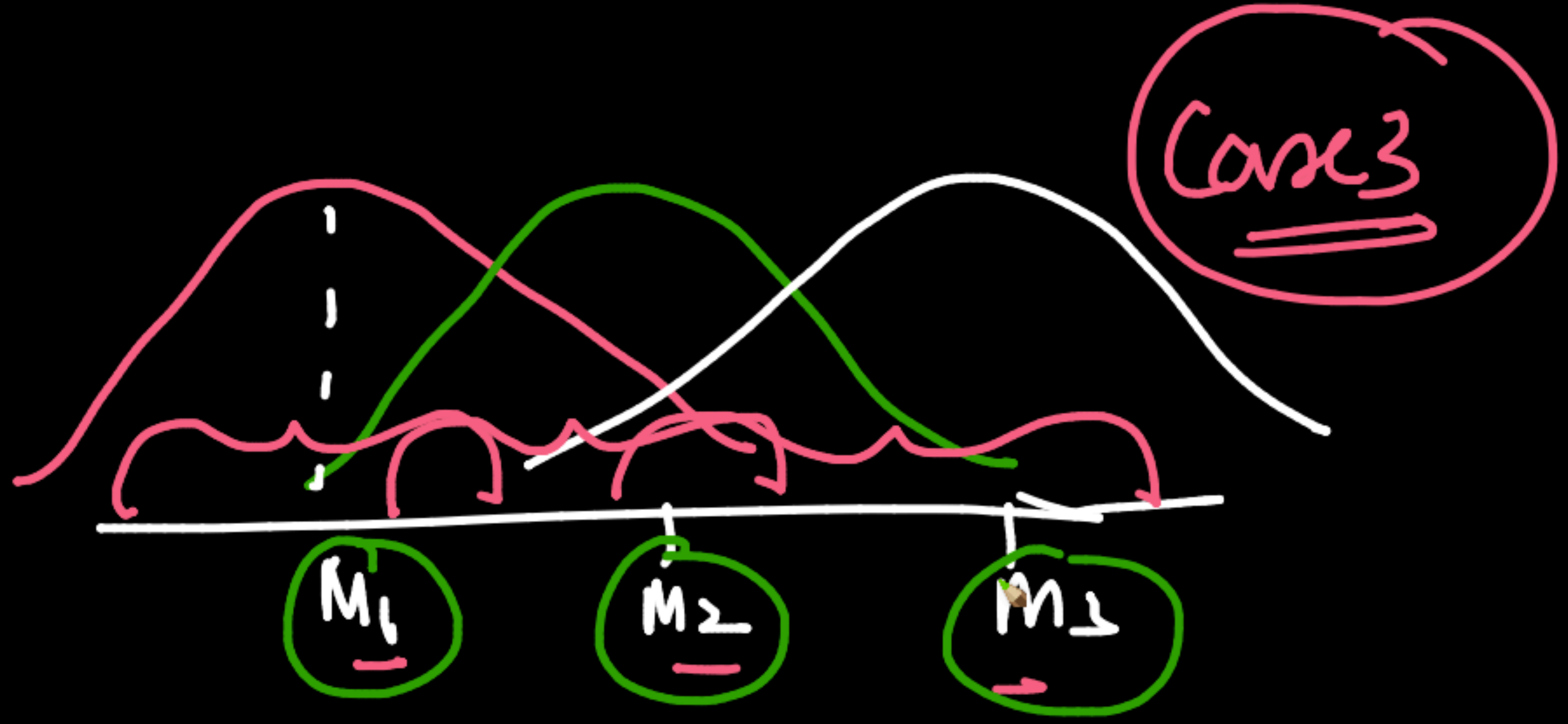


$$f_1 : \frac{MSB}{MSW}$$

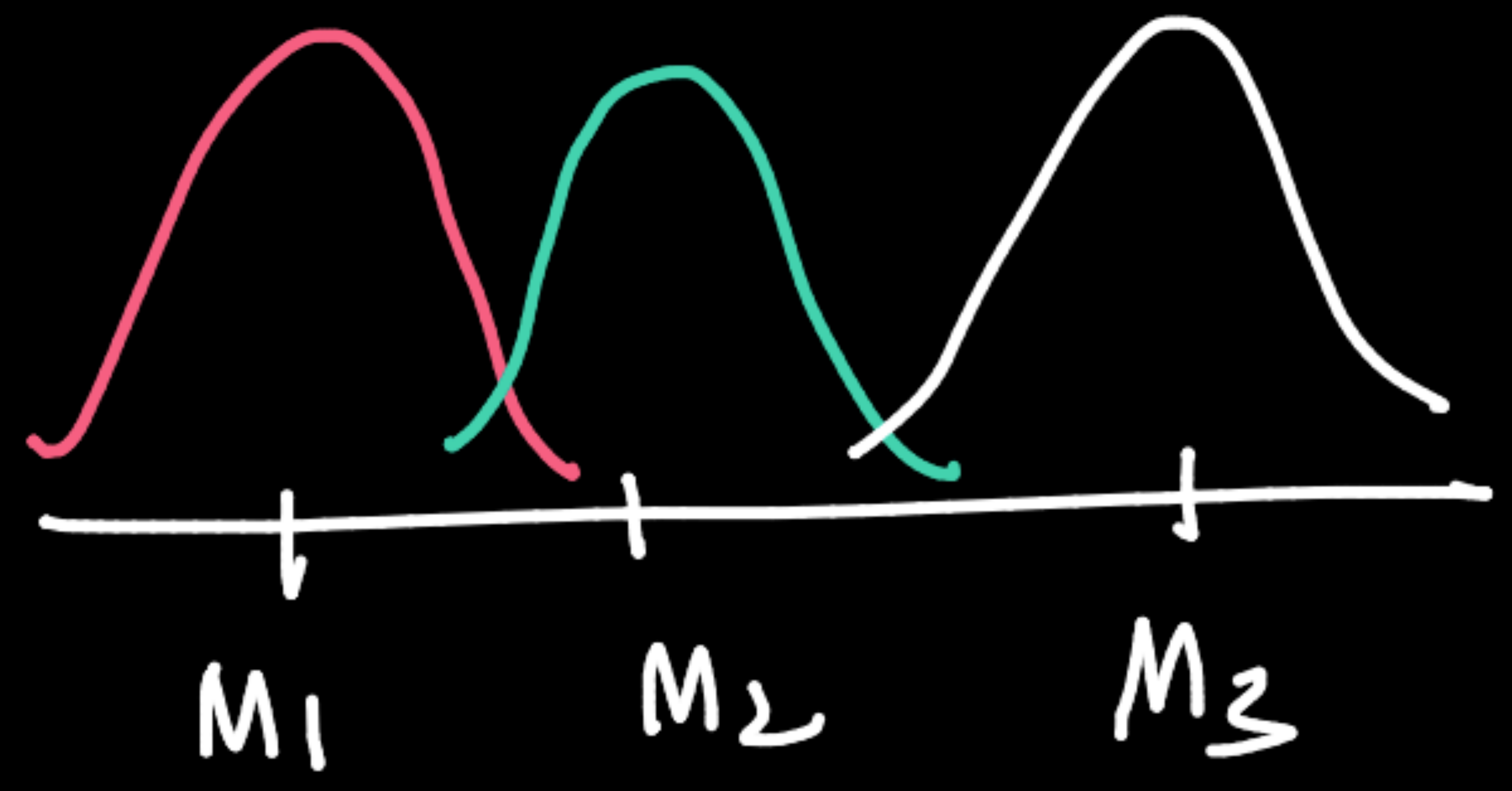


$$\underline{\underline{f_2 < f_1}}$$

$$\downarrow f_2 : \frac{MSB}{MSW \uparrow}$$



$$f_1 : \frac{MSB}{MSW}$$

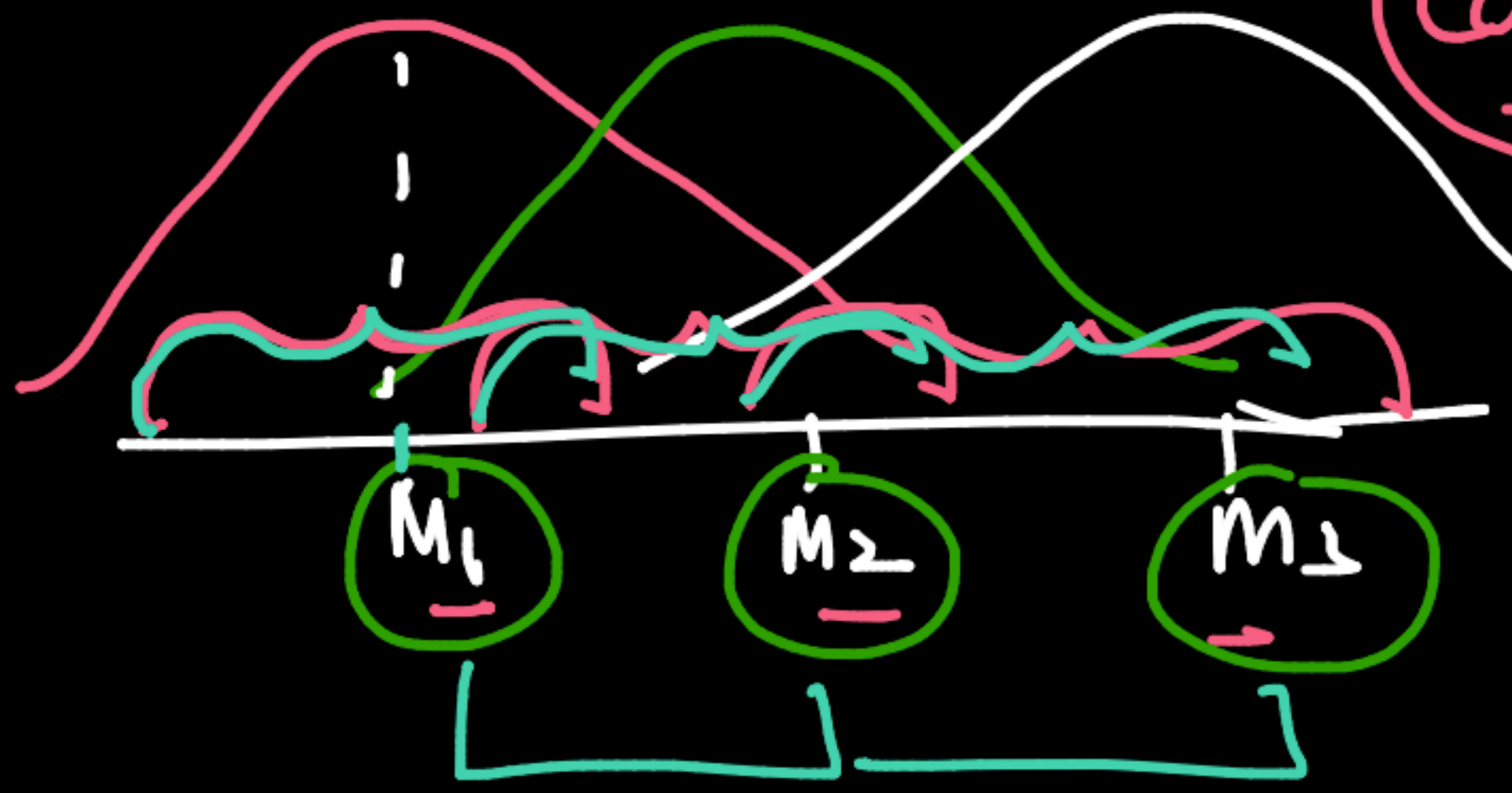


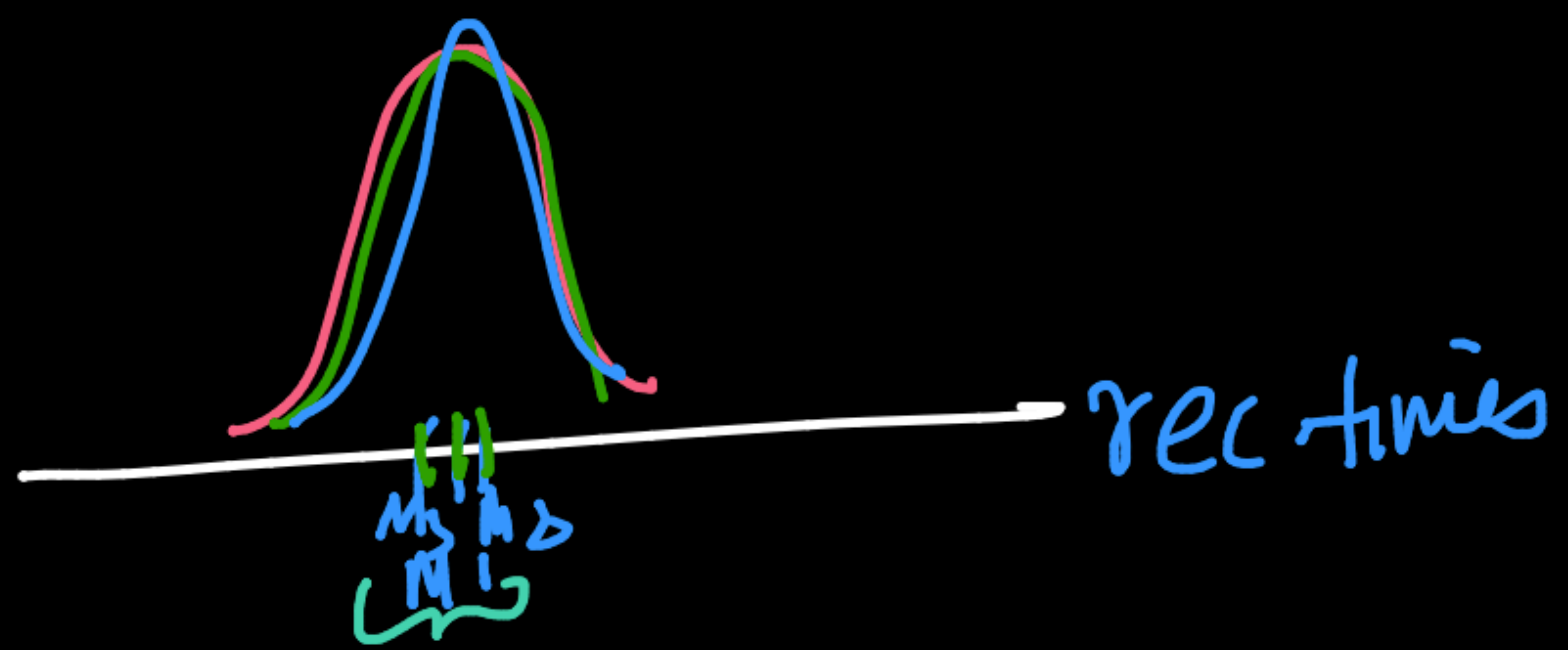
$$f_2 < f_1$$

F-dist

$$f_2 : \frac{MSB}{MSW}$$

Case3




rec times

$$f = \frac{MSB}{MSW} \sim F(k-1, n-k)$$

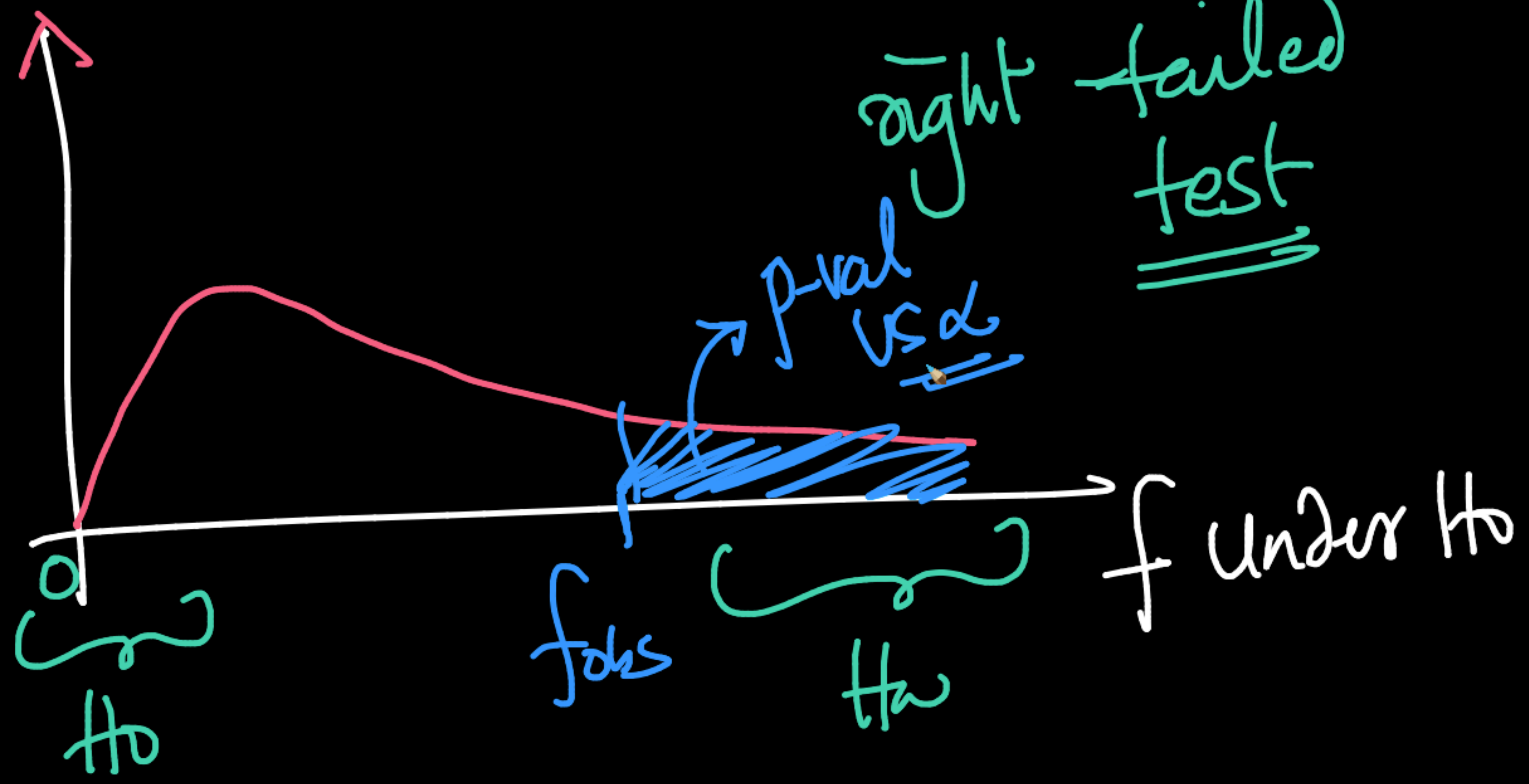
H_0: no difference in gp-means

H_a: $\exists$ a difference in " "

k-1 n-k

right tailed test

f



ANOVA: K-gps ($K \geq 2$)

$k = 2$

$\downarrow$

t test / z test

$\downarrow$

comparison of means

$$f = \frac{MSB}{MSW}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:25

- Variations of ANOVA
- T1 error / Type 2 error

⋮

71 / 71

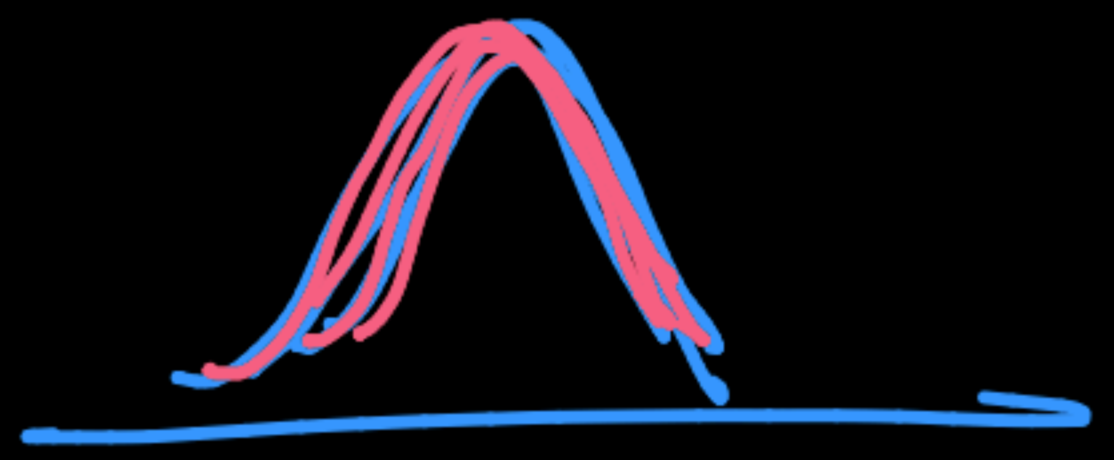
Q&A

ANOVA — continous

PCC / SRCC --->

(~~'dependent'~~)

f↓



k-drugs — ANOVA

Stats

Complex variations ⟷ Special case

ANOVA

Linear reg

ML

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:34

✓ QQ

✓ t/z-test

✓ ANOVA

74 / 74

χ^2 :-

H_0: gender is indep of is Premium

H_1: gender is ~~dep~~ on isPremium

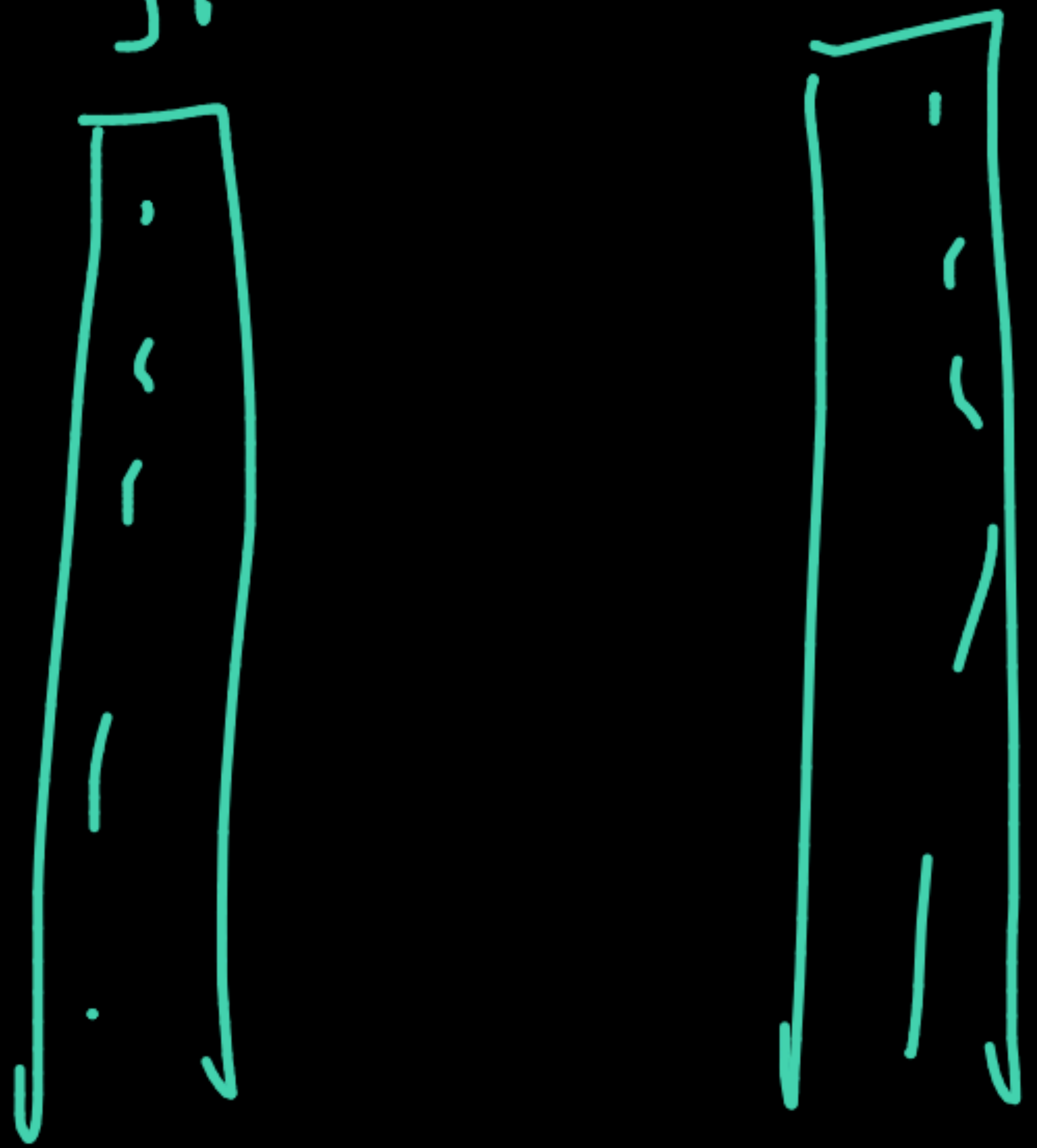
T_x under H_0 → disb not known

ML:

χ^2-test ✓

f_i

y_i : labels



Cat

Cate.

AB - test

- outliers
- A vs B

C T

hyp-test

C.I

hyp-test

Ho: $\text{mean } WT_A = MWT_B$

Ha: " $\neq$ "

p-val = v.v small

< 5%

reject Ho

MWT_A

MWT_B

Con'd → median C.I

→ hyp-test

ks-test

X

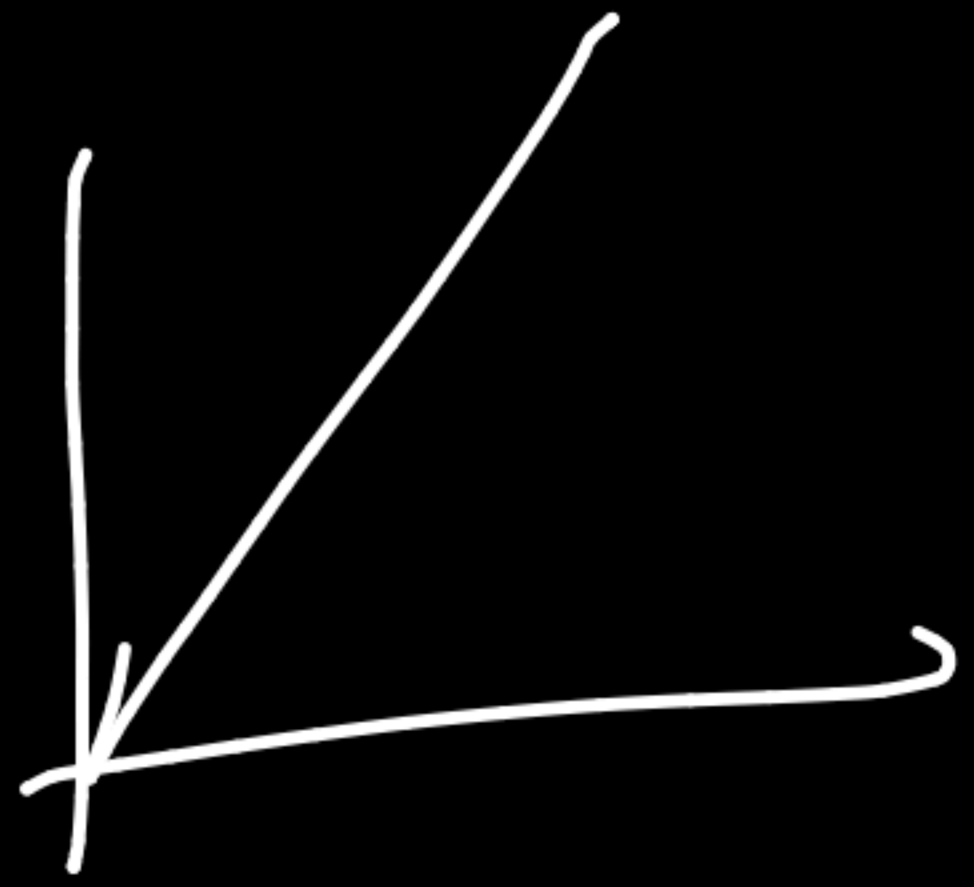
Y

$X \sim Y$

$X \sim \text{log-normal}(\mu, \sigma)$

1000 pts

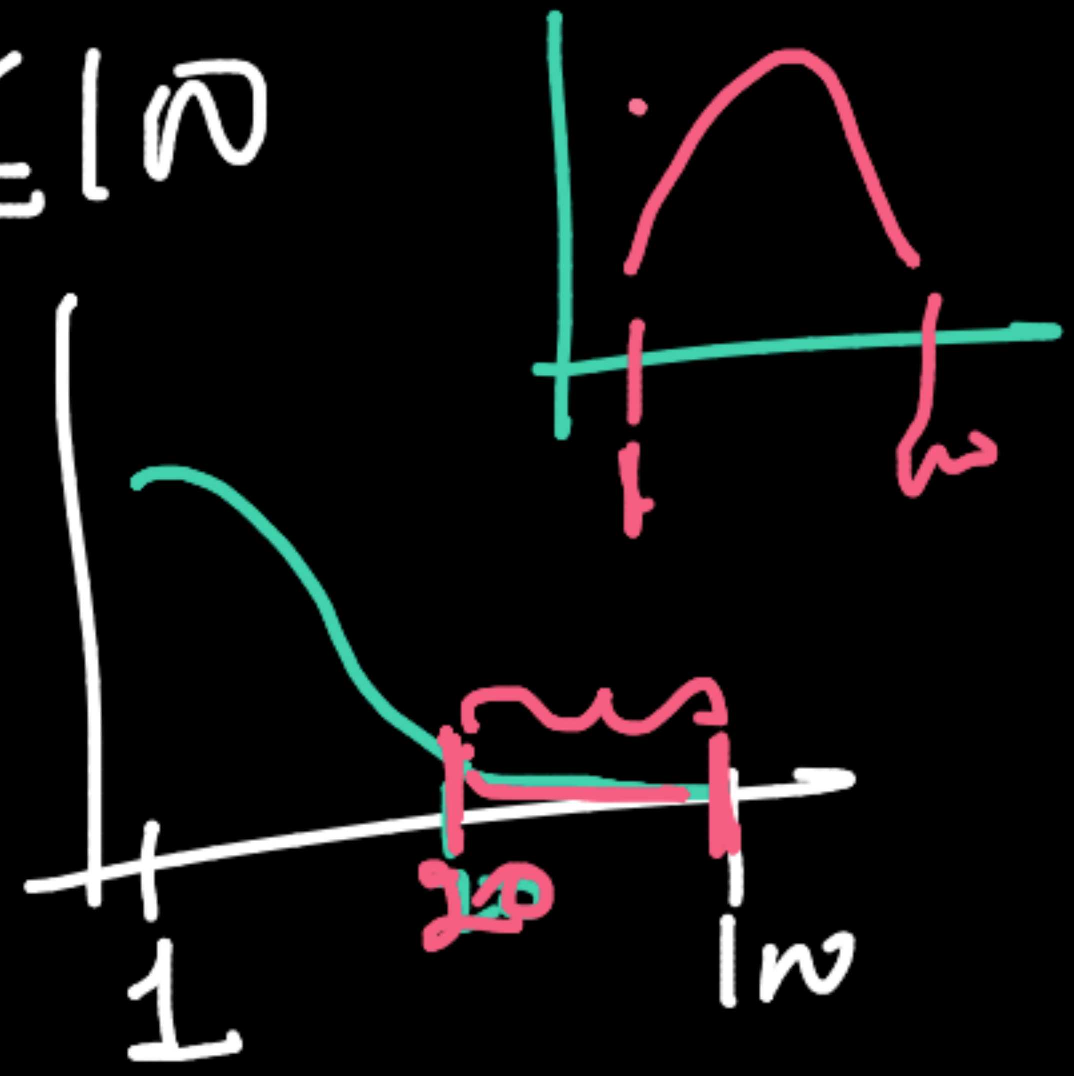
$\boxed{X \sim Y}$



X pop: 1 2 - - - - 100

$1 \le X \le 100$

PDF



Sample: 1 20

$n \geq 30$

random

C.I of pop mean.

#samples

CLT

1-100

pop!

Sample

1-20

bootstrapping + CLT

Uniform

1

10 1's

10 2's

10 3's

⋮

10 100's

100

random sample

V.V V small

1-20

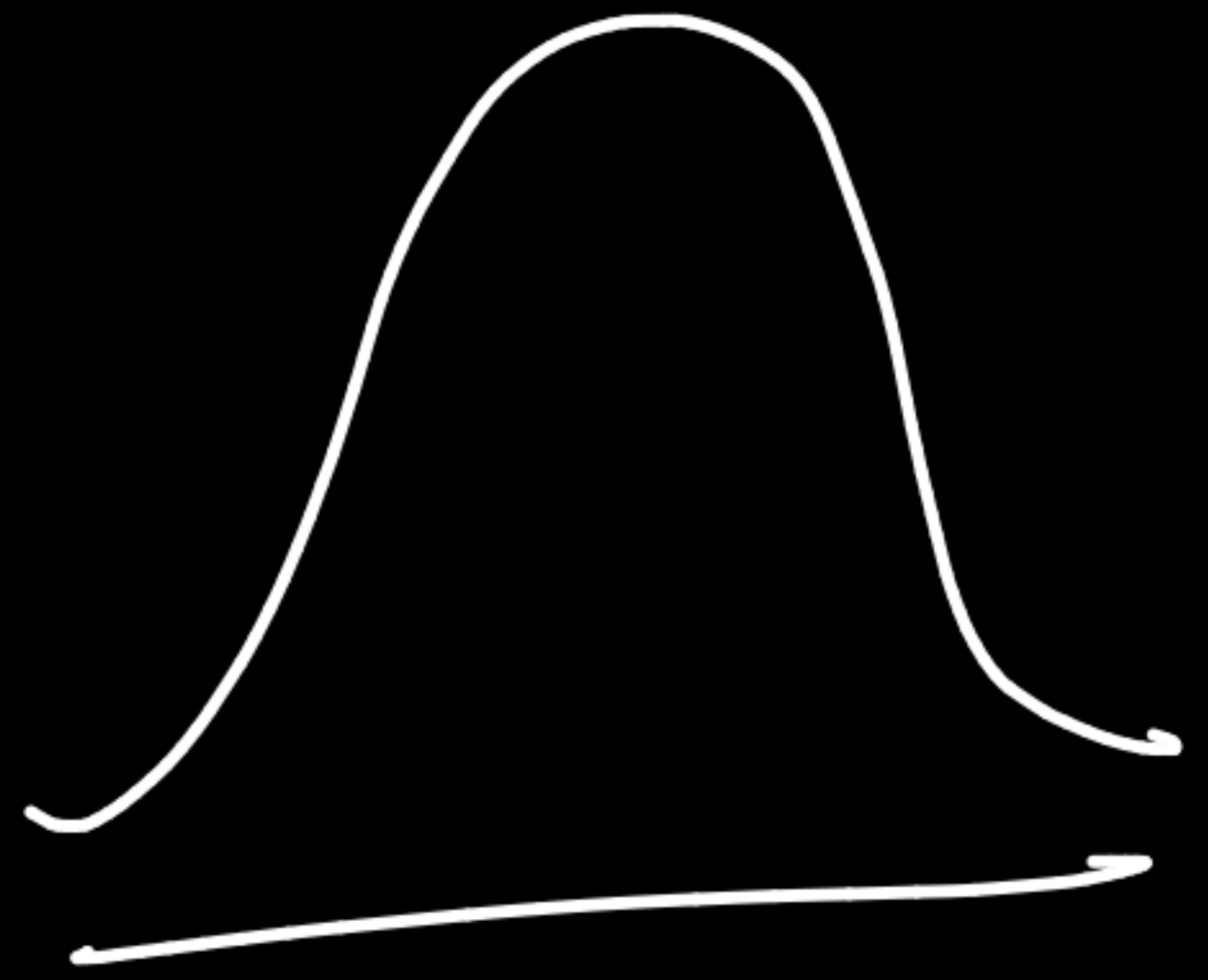
1-20

$S1^{n} \rightarrow M_1$

$S2^{n} \rightarrow M_2$

$S_3^{n} \rightarrow M_3$

$\vdots$



data collected

INC - - -

1 - - - - . 500

random sample

Sample ~ pop

ML

10 indep-models

bin-classfn:

$x_q \to M1 \to$ prob-score

$x_q \to M2 \to$ "

⋮

$x_q \to M10 \to$ "

Test

$i : 1 \to n$

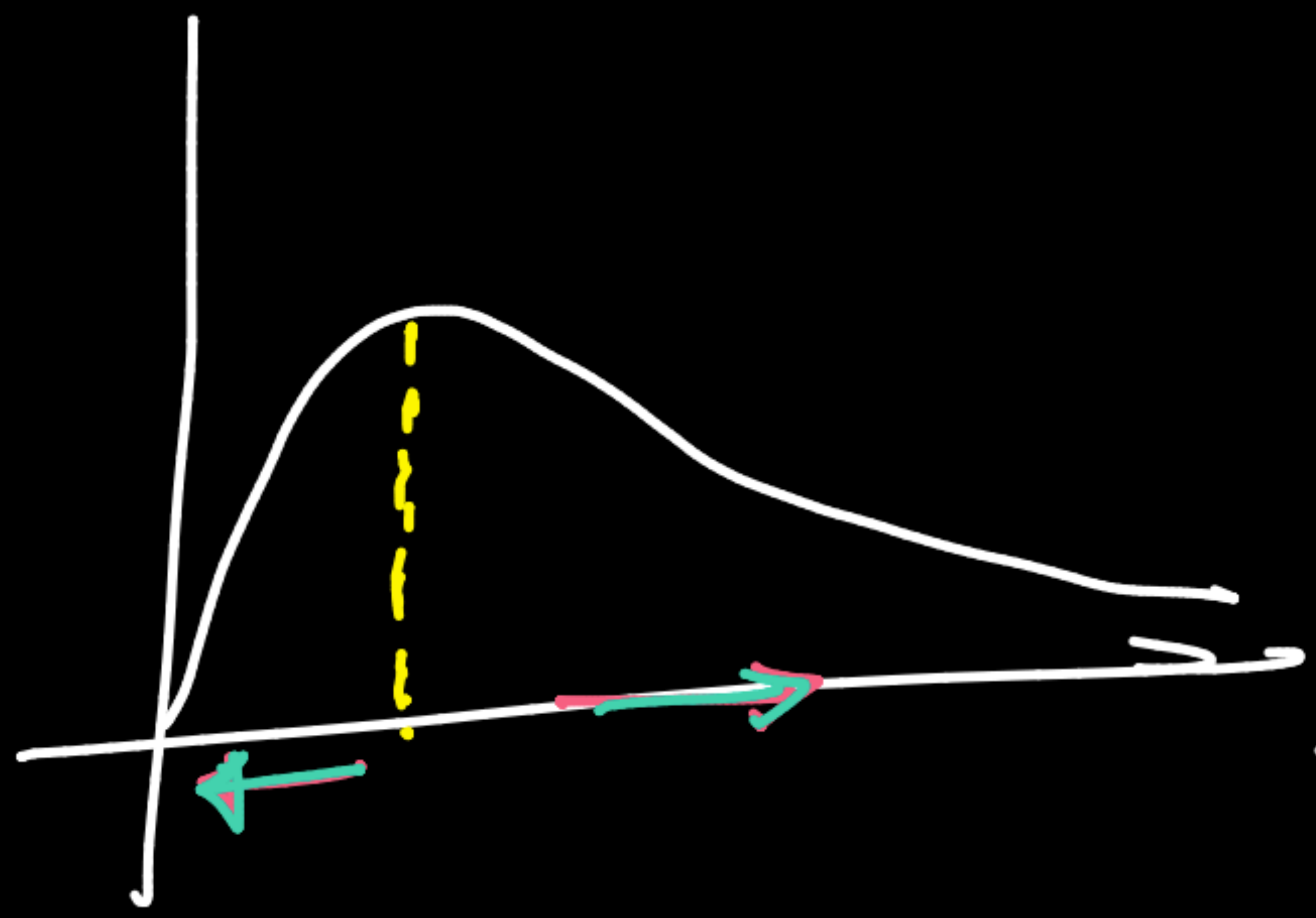
$$
\begin{array}{ccc}
 & & \qquad\qquad\qquad\qquad\quad i,k \\
M_1 & \longrightarrow & \underline{P_{11}\ P_{12}}\quad P_{13}\quad \boxed{}\quad P_{1n} \\
 & & \text{Corr.coeff} \\
M_2 & \longrightarrow & \underline{P_{21}\ P_{22}}\quad P_{23}\quad \boxed{}\quad P_{2n}
\end{array}
$$

ANOVA

f under H_0.

$\boxed{X}$ f_{obs}

$$f = \frac{MSB}{MSW}$$



f under H_0

means are same

error :- α = incorrectly accept the

T1 & T2 → next class

→ Variations of ANOVA ; T1, T2 errors (β, α)

→ Correlation coeff...

:

:

https://colab.research.google.com/drive/1lWuS8AxULBBold2GlfBKUeT1RSj3fFAJ#scrollTo=Y569EYQBiMeN

KS-Test

[ ] ↳ 11 cells hidden

T-test & Z-test

```
# Group A --> Treatment Group shown 2 ads per ad-break
# Group B --> Control Group shown only 1 ad per ad break
# Let us comapre mean watch-times per group
# H0: mu1= mu2
# H1: mu1 != mu2
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import scipy
```

93 / 93

✓ 0s completed at 21:18