Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:03

Topics:

easier

1. Z-test
2. KS-test
3. T-test
4. Z-proportions test (if possible)

- Hyp-testing Framework
- Applied/Practical
- Math + intuition
- Code + case-studies

1 / 3

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:04

Ops:

Questions → tab ✓

Questions → "END: ..."

Chat → Y/N & interactivity

✓ [Misc. topics

↳ dedicated session

'n-1' vs 'n'

2 / 3

Q1

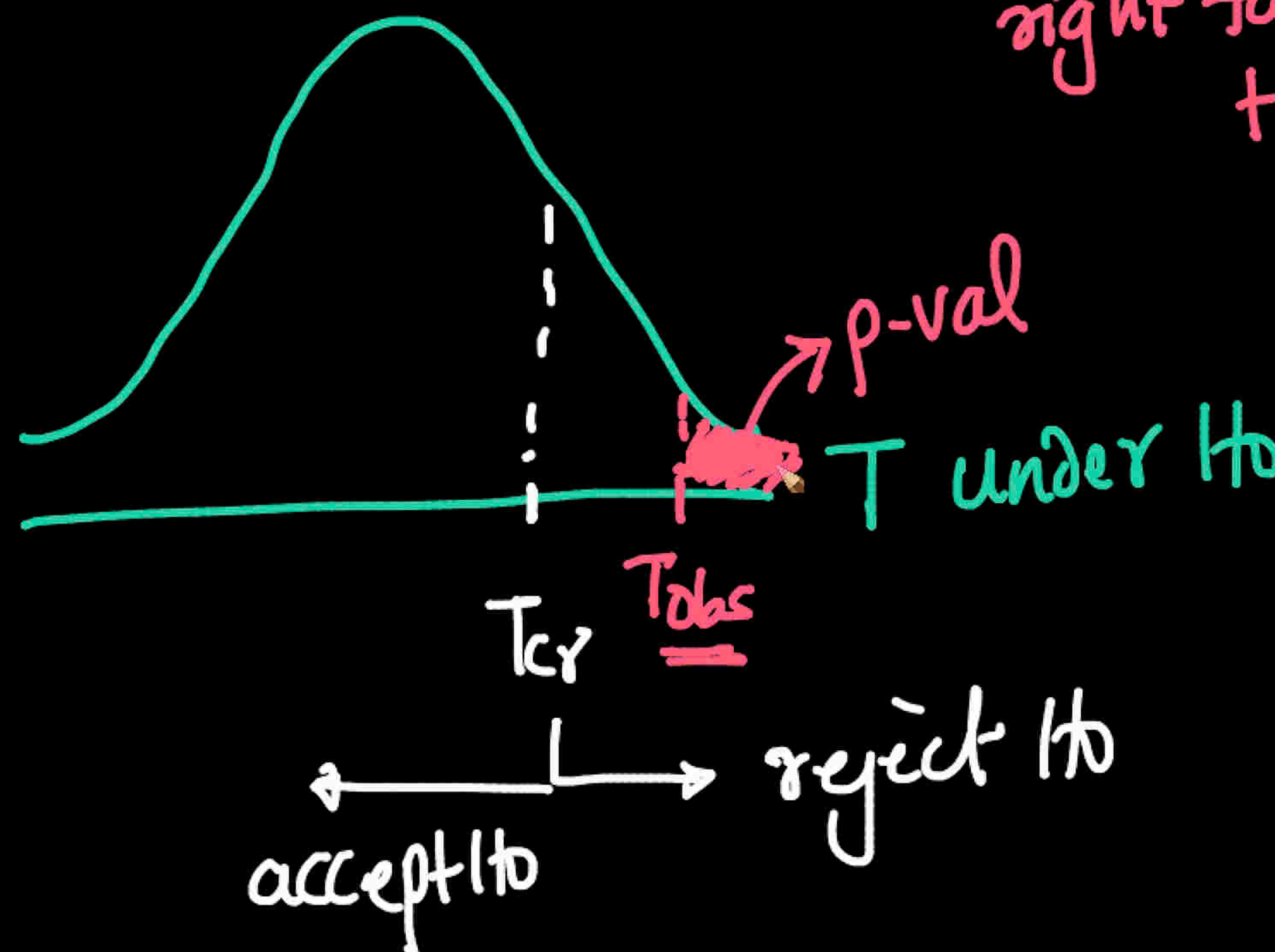
critical value vs p-value

T_{cr} : that value beyond which if T_{obs} lies $\Rightarrow$ reject H_0

$P(T \text{ is as extreme as } T_{obs} \mid H_0)$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:08

right tailed test



4 / 5

(Q2)

if p-value < α then reject H_0

area of the dist beyond T_{obs}

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:08

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:16

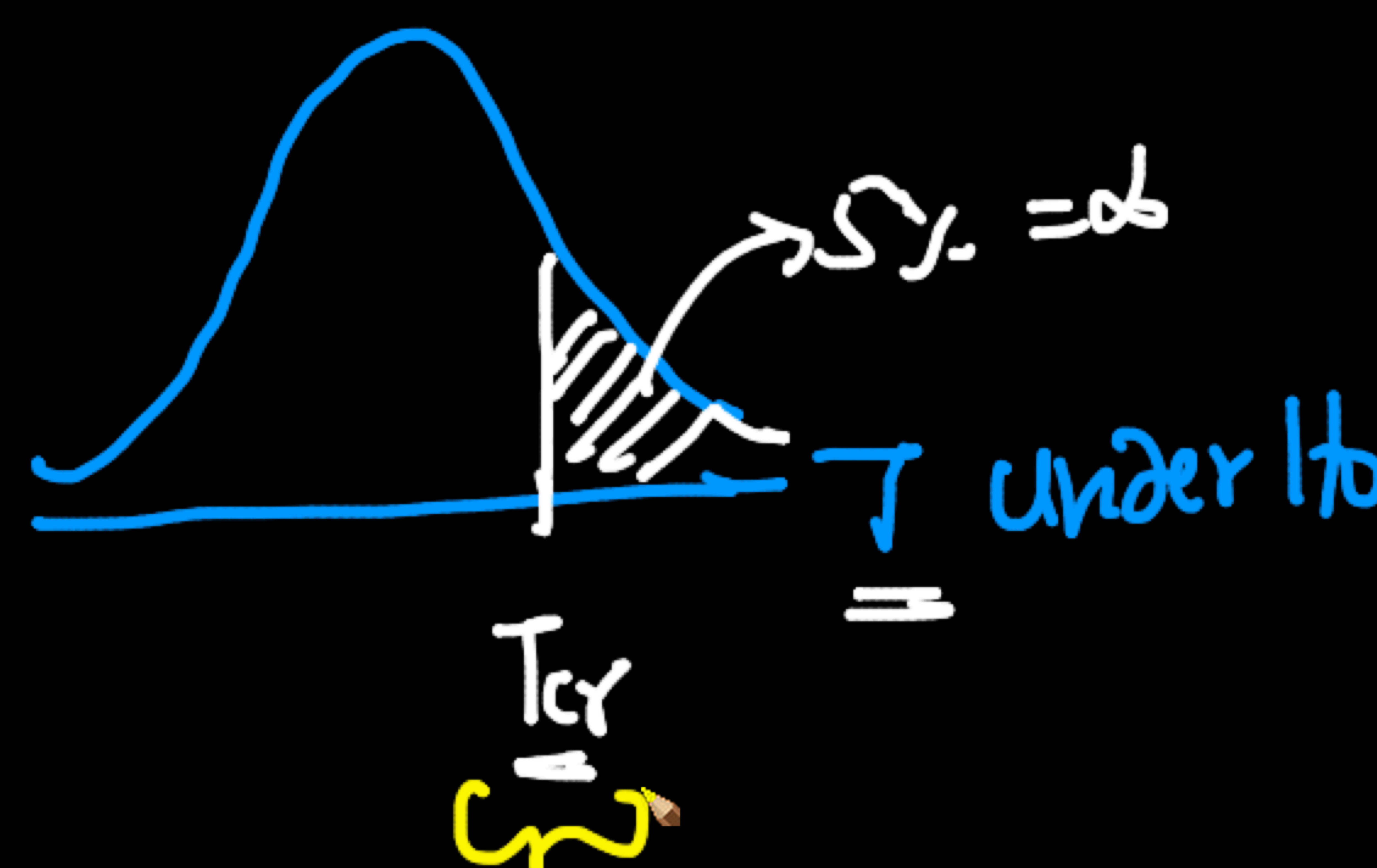
(Q3)

$\alpha = 5\%$

① right tailed

② T_{cr}

$T \sim Bin(\overset{\nearrow 100}{n}, p = 0.5)$



6 / 10

Under H_0

$P(\text{observing } \bar{T} \text{ beyond } T_{cr} \mid H_0)$

$= 5\%$

$\boxed{\text{right-tailed test}}$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:13

Q

H_0: coin is fair

H_a: coin is unfair

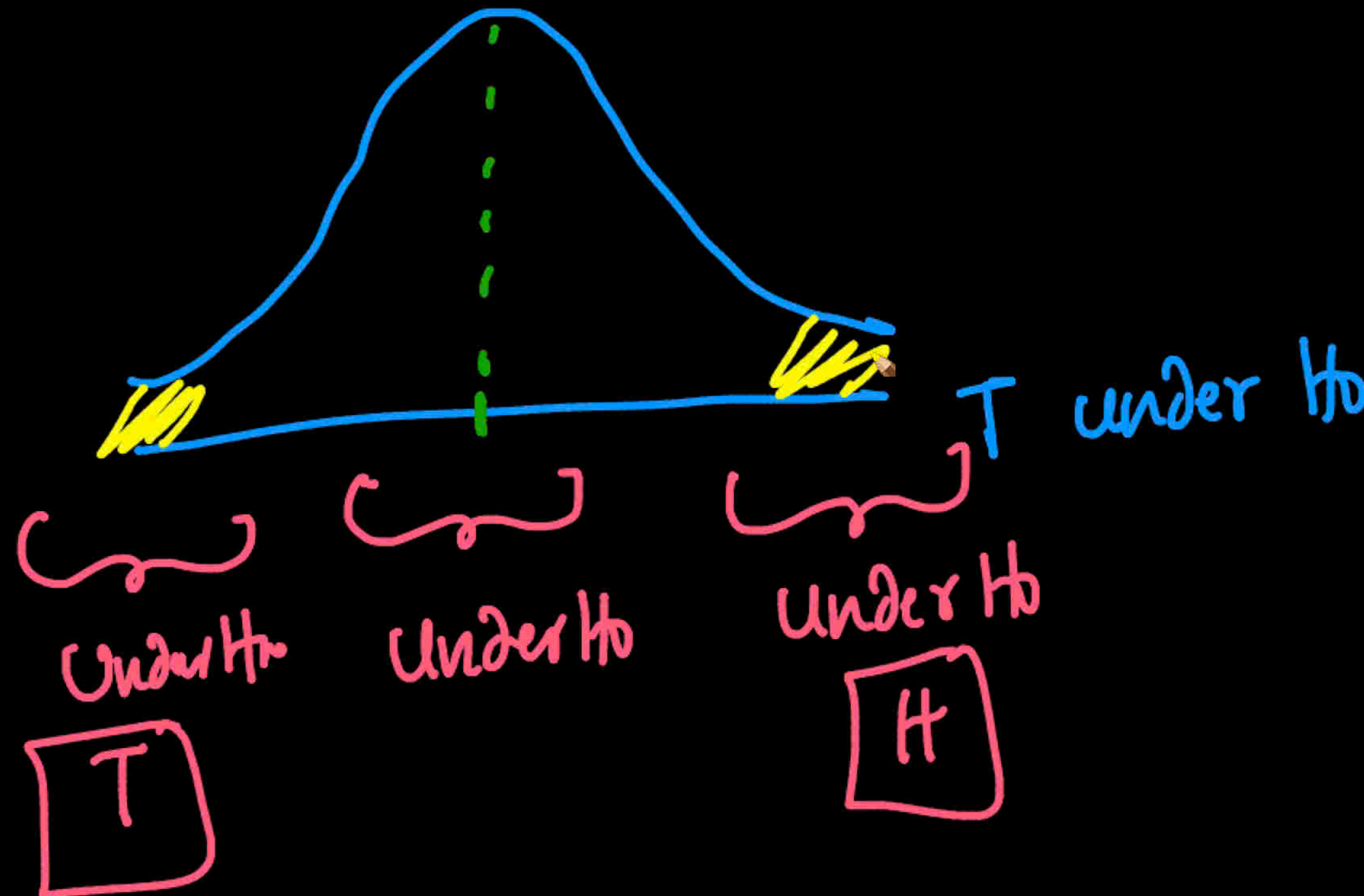
T: #heads in 100 tosses

coin-toss

2-tailed test

8 / 8

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:14



Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:15

10 / 10

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:12

Z-test

Expt

covid medicines

Remdesivir ✓

Med1 → $x_1, x_2, \ldots, x_{110}$ ← n_1

rec. time

Med2 → $x_1', x_2', \ldots, x_{90}'$ ← n_2

placebo

ICMR/FDA

PFIZER

$\boxed{\geq 30}$

11 / 35

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:33

$$\text{Sample}_{10} \rightarrow \text{Med } 1 \rightarrow M_1, S_1 \leftarrow \text{sample}$$

$$\text{Sample}'_{90} \rightarrow \text{Med}_2 \rightarrow M_2, S_2$$

12 / 16

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:29

$\boxed{Q}$ Are the pop. means of Med1 & Med2 same or not

$\boxed{Q} \rightarrow$ Is $\mu_1 = \mu_2$?

13 / 16

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:28

Q

1. μ_1 vs μ_2 } → Z-test (later)

2. 95 C.I on μ_1 & 95% C.I on μ_2

↑ samples

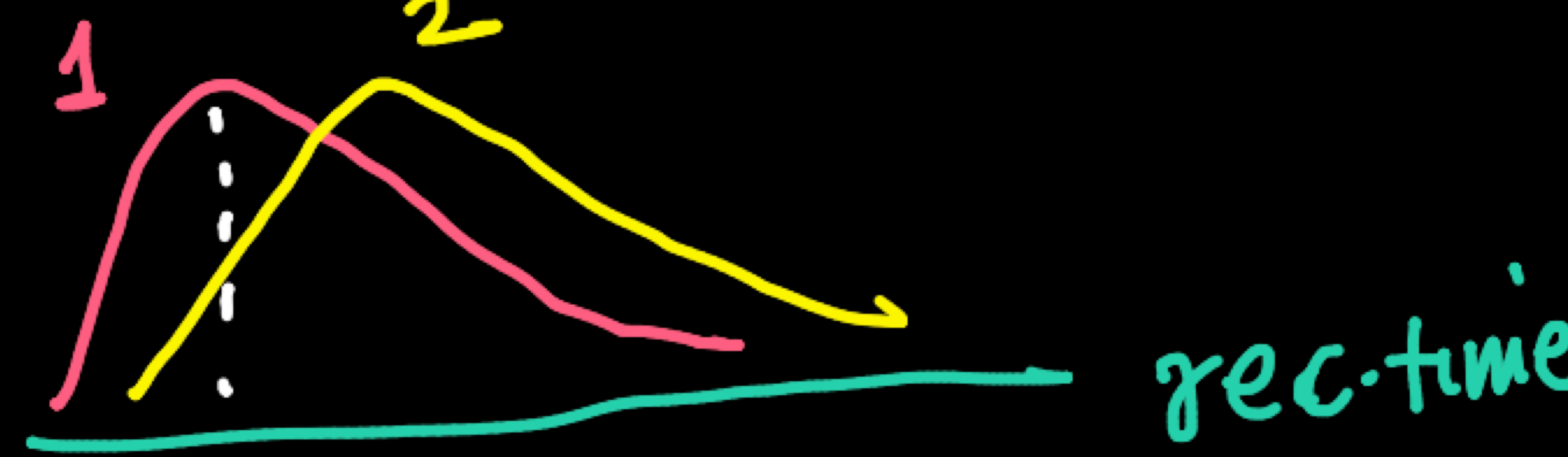
Tricky 3. pop median 1 vs pop median 2

14 / 16

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:27

✓ 4 QQ-plot

$disb_1 \neq disb_2$



5 KS-test → (later)

↳ uses gap b/w CDFs

15 / 15

Notation

$\mu \rightarrow$ pop mean

$\sigma \rightarrow$ pop. std-dev

$\bar{x}, m \rightarrow$ sample Mean

$s \rightarrow$ sample std-dev

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:47

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:56

Input → Two samples → n_1, → n_2

$$\boxed{\text{Z-Test}}$$

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

n_1 & n_2 are not small (≥ 30)

in practice

$S_1 \approx \sigma_1$ $S_2 \approx \sigma_2$

① assumes sample means: M_1 & M_2 follow CLT → Gaussian

② σ_1 & σ_2 : pop std. dev are known

17 / 24

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:40

$$T = \frac{\overbrace{m_1 - m_2}^{\checkmark}}{\left\{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\}}$$

Math proof $\rightarrow$

why not

$\underline{T = M_1 - M_2}$

$\downarrow$

do not know distb of T under Ho

T under Ho $\sim$ $\underline{Z(0,1)} = N(0,1)$

18 / 19

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:56

$$T_{obs} = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Math proof →

T under $H_0 \sim \underline{Z(0,1)} = N(0,1)$

why not $T = M_1 - M_2$

Can use it if we know its distb under H_0

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:55

Under H_0

$$T \sim Z(0,1)$$

$$\left\{ \frac{M_1 - M_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right.$$

$$\left\{ T = \tilde{M}_1 - M_2 \right.$$

Gaussi

Gau

$$\hookrightarrow N\left(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

✓

20 / 24

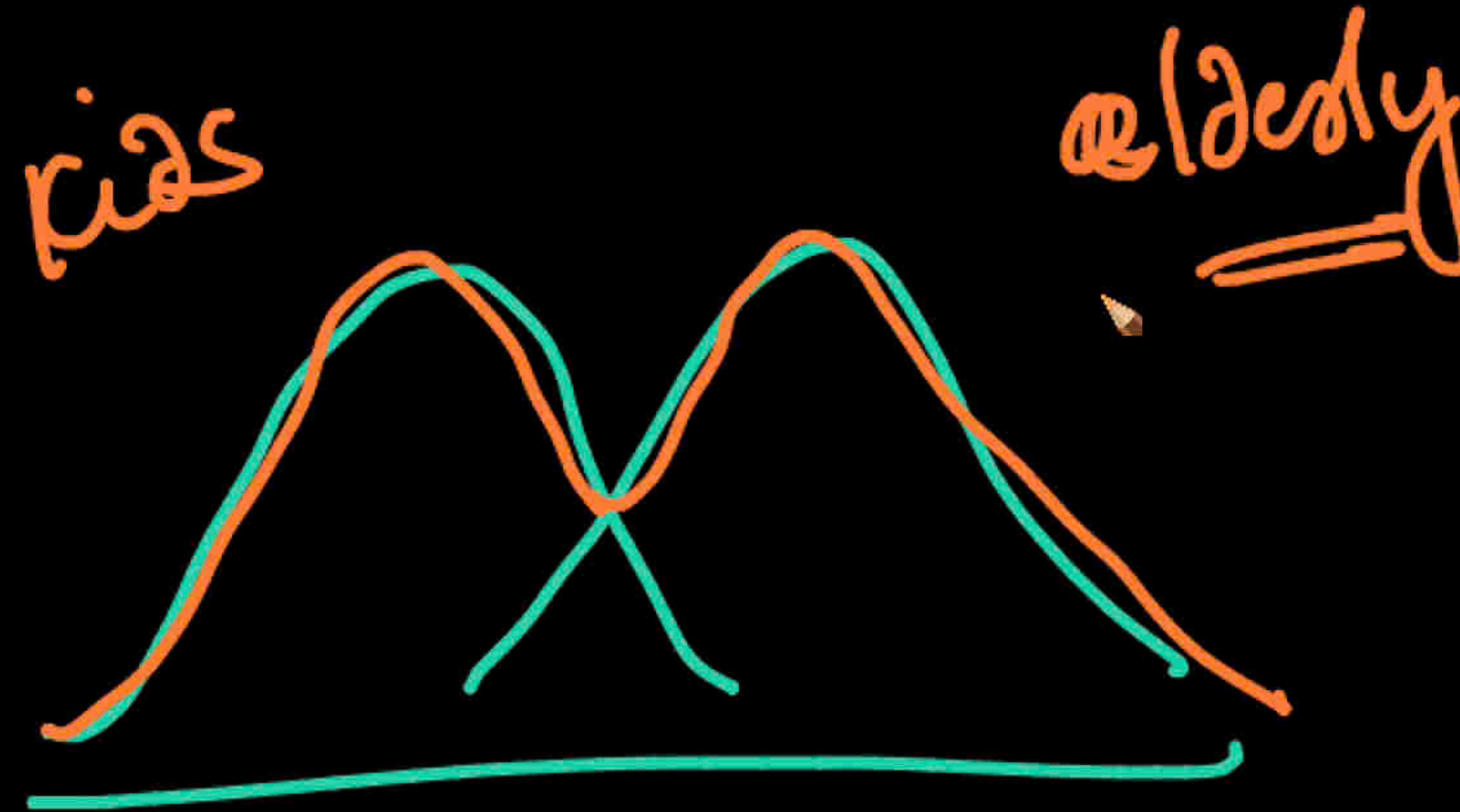
Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:44

CLT: If pop Mean & std-dev → finite

Then $\left\{ \text{Sample means follows Gaussian } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \right.$

21 / 21

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:59



Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:00

Φ

$\bar{X}$

$$\boxed{S = \frac{\sigma}{\sqrt{n}}}$$

X

CLT: $\underbrace{\text{sample-means}} \sim \underbrace{N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$ as $n \to \infty$

Std. error


Chrome File Edit View History Bookmarks Profiles Tab Window Help
9:56
25 / 25

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:58

CLT

Proof using CLT

if $\mu_1 = \mu_2$ (H_0)

samples → size n_1 size n_2

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

25 / 25

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:01

28 / 28

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:01

28 / 28

recap:

Z-test:

assumptions

→ Pop mean & std. dev are finite (CLT)

→ σ_1 & σ_2 are known

✓

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:21

1. $H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

2. $$T_{obs} = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

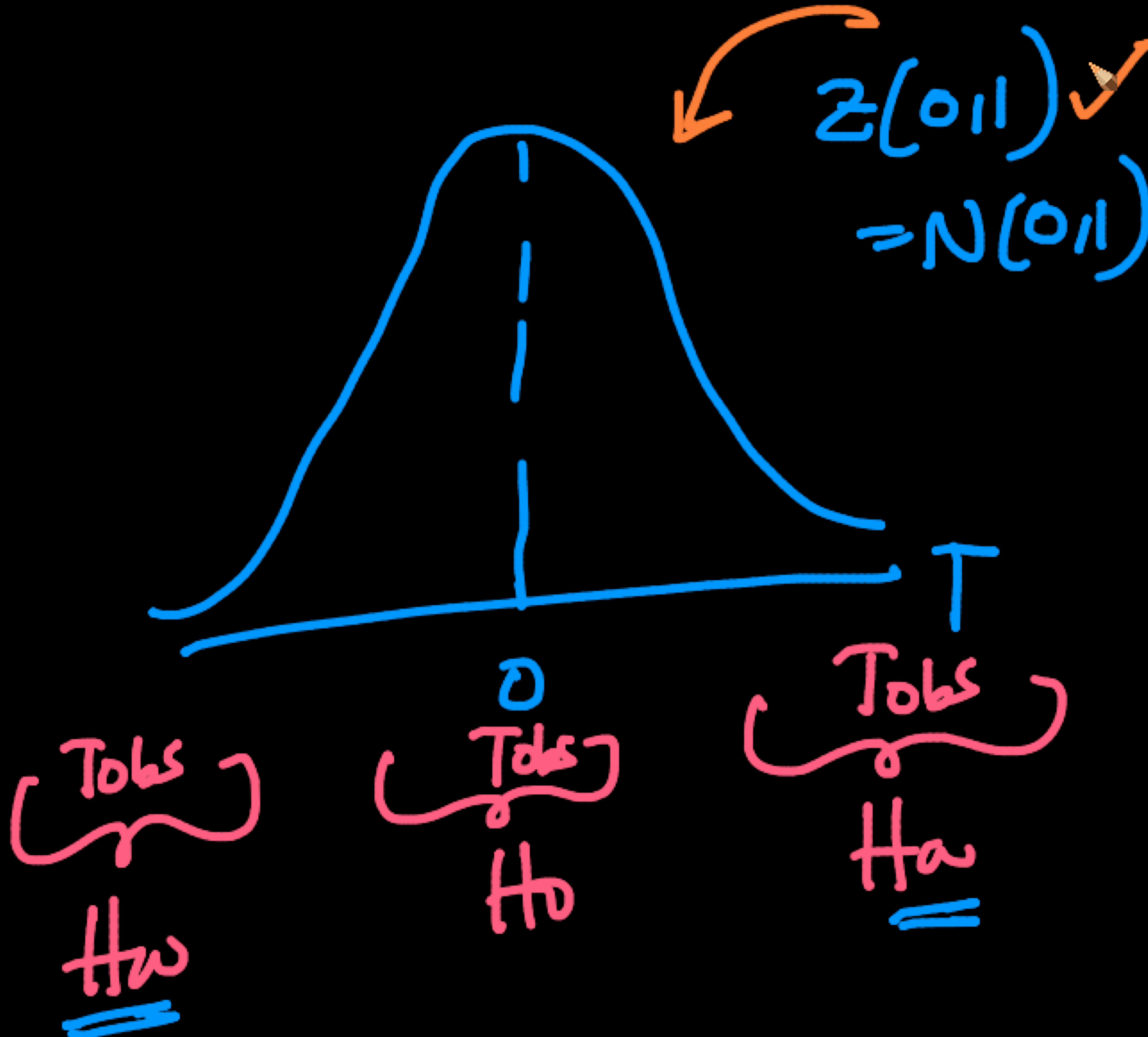
29 / 40

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:21

3

Under H_0

2-sided



Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:04

4

$x_1 \ldots x_{100}$

$x_1' \ldots x_{90}'$

$n_1 = 100$

$n_2 = 90$

$$\begin{cases} \mu_1 = \checkmark \\ \mu_2 = \checkmark \end{cases}$$

T_{obs} ✓

$\sigma_1 \approx s_1 \quad (n_1 > 30)$

$\sigma_2 \approx s_2 \quad (n_2 > 30)$

31 / 31

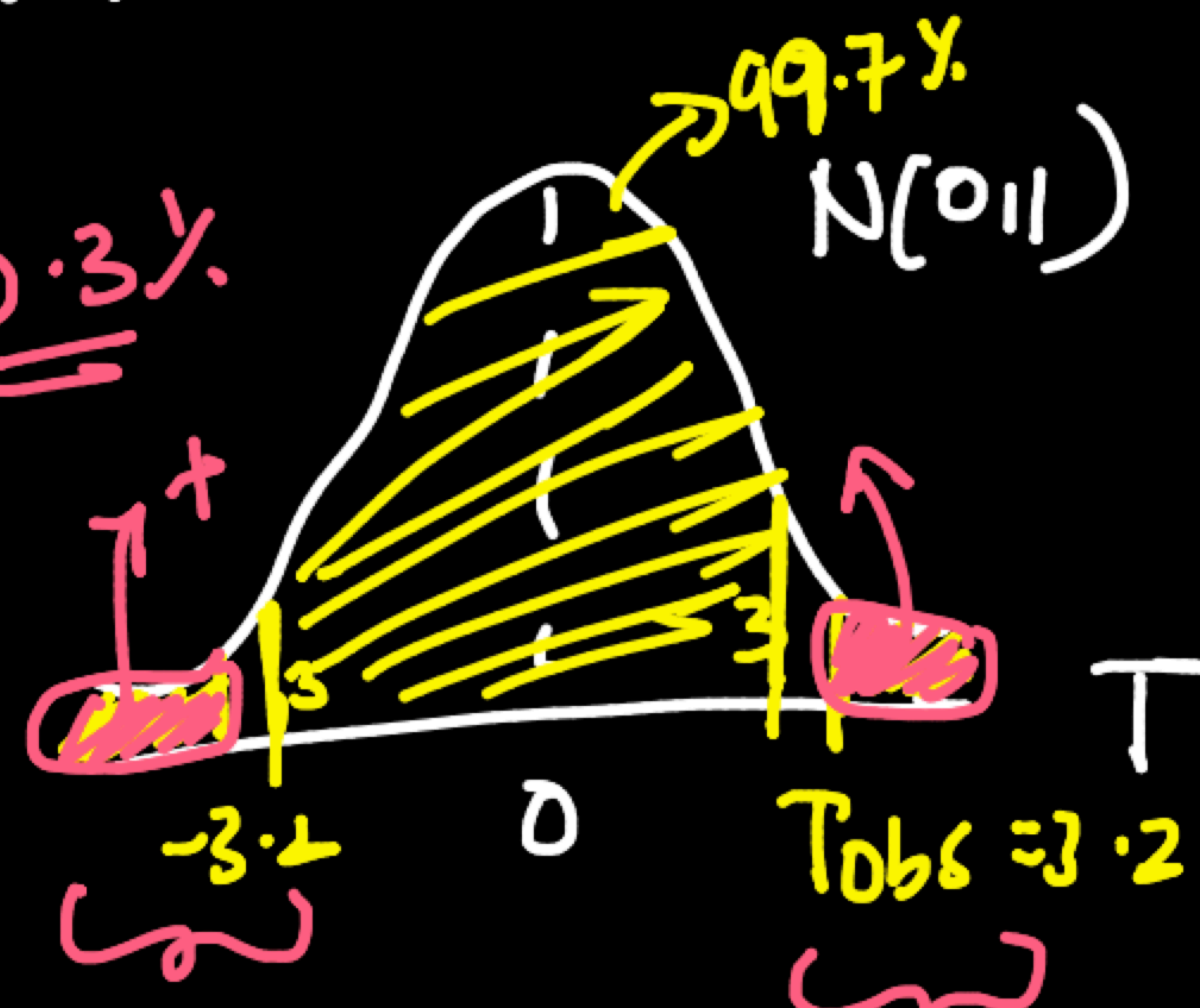
Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:08

Compute $T_{obs} = 3.2$ (let)

5 α : sig-level → 1%

6 p-val < 0.3%

p-val < 1%



7 reject H0

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

Q

2-Sample

Z-Test : → Compare pop means from 2-samples

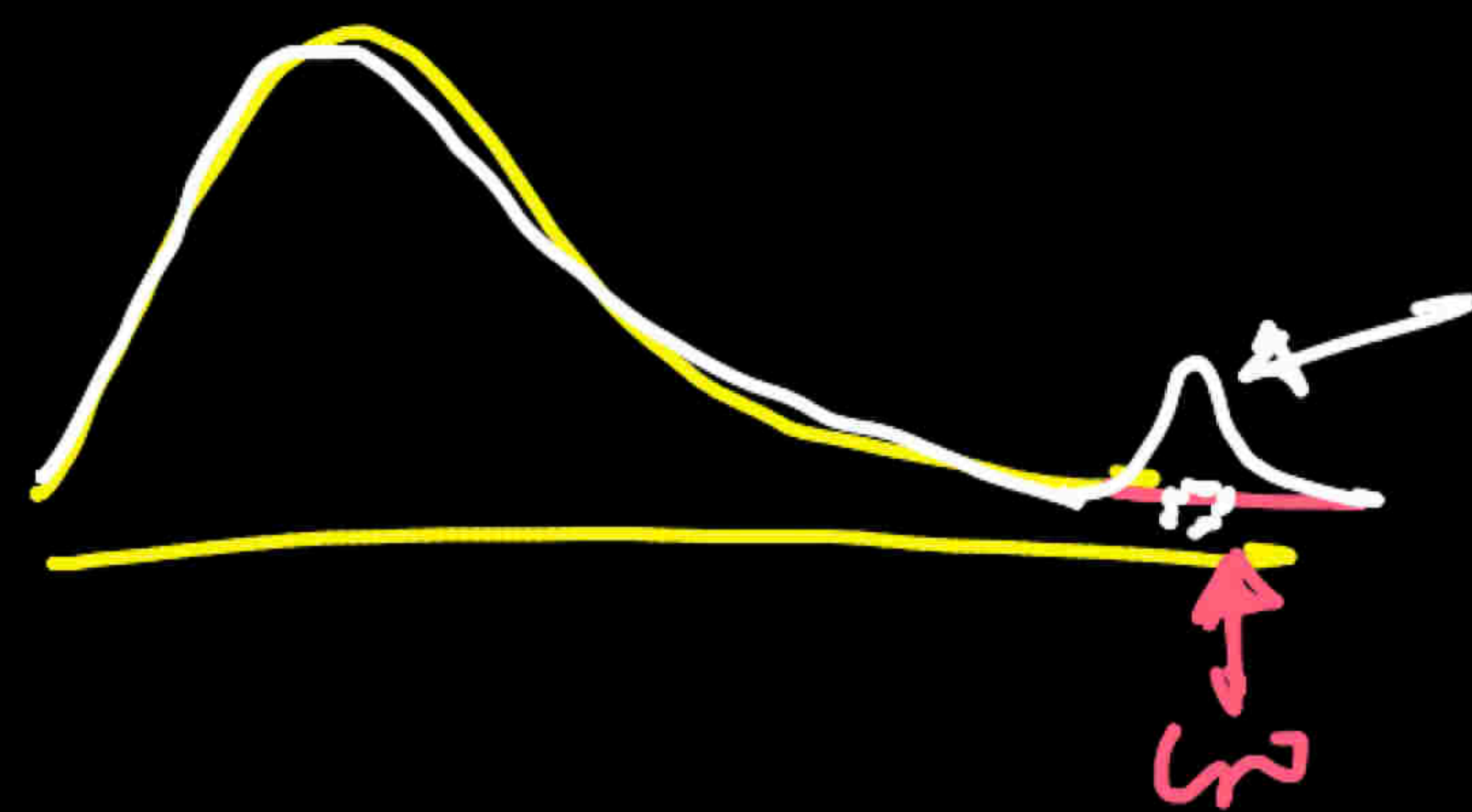
34 / 34

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:11

Q

log-normal

practical



actual outliers

35 / 35

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:14

Q

last class

① H_0: Coin is fair
H_a: coin is unfair

② $T = \#\text{heads}$

Under H_0:
$P(H) = 0.5$

$T \sim Bin(n, p = 0.5)$

Z-test → compare means

$H_0: \mu_1 = \mu_2$
$H_a: \mu_1 \neq \mu_2$

$$T = \frac{M_1 - M_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim Z(0,1)$$

36 / 36

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:31

Q

Task

z-test : $\underbrace{\text{Compare pop. means}}_{\text{Purpose}}$

&

✓ { n_1 & n_2 are not small

&

✓ { σ_1 & σ_2 are known or estimatable

{ pop mean & stdev are finite

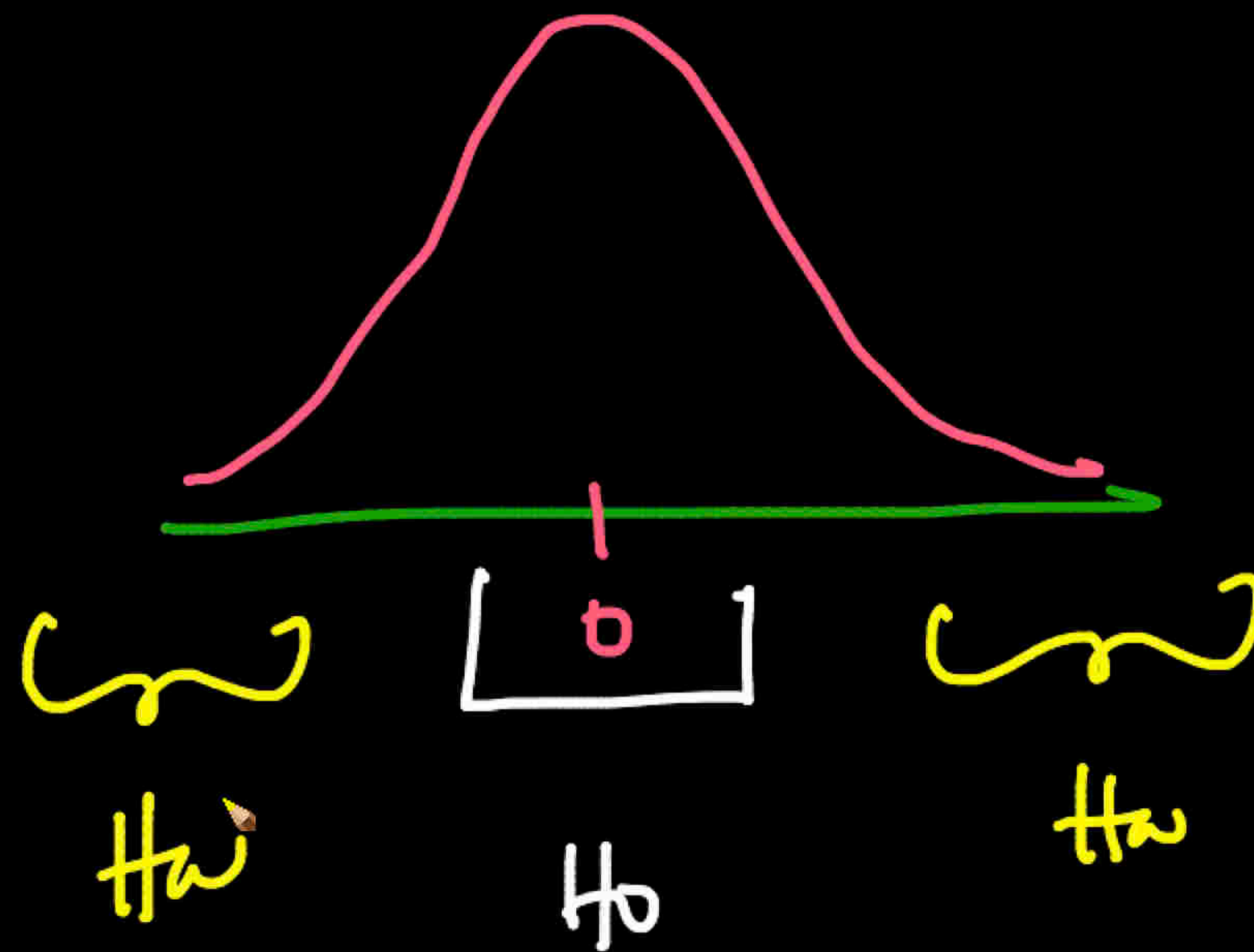
37 / 41

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:18

Q

$$T_{obs} = \frac{\overset{10}{M_1} - \overset{15}{M_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

+ve



38 / 38

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:20

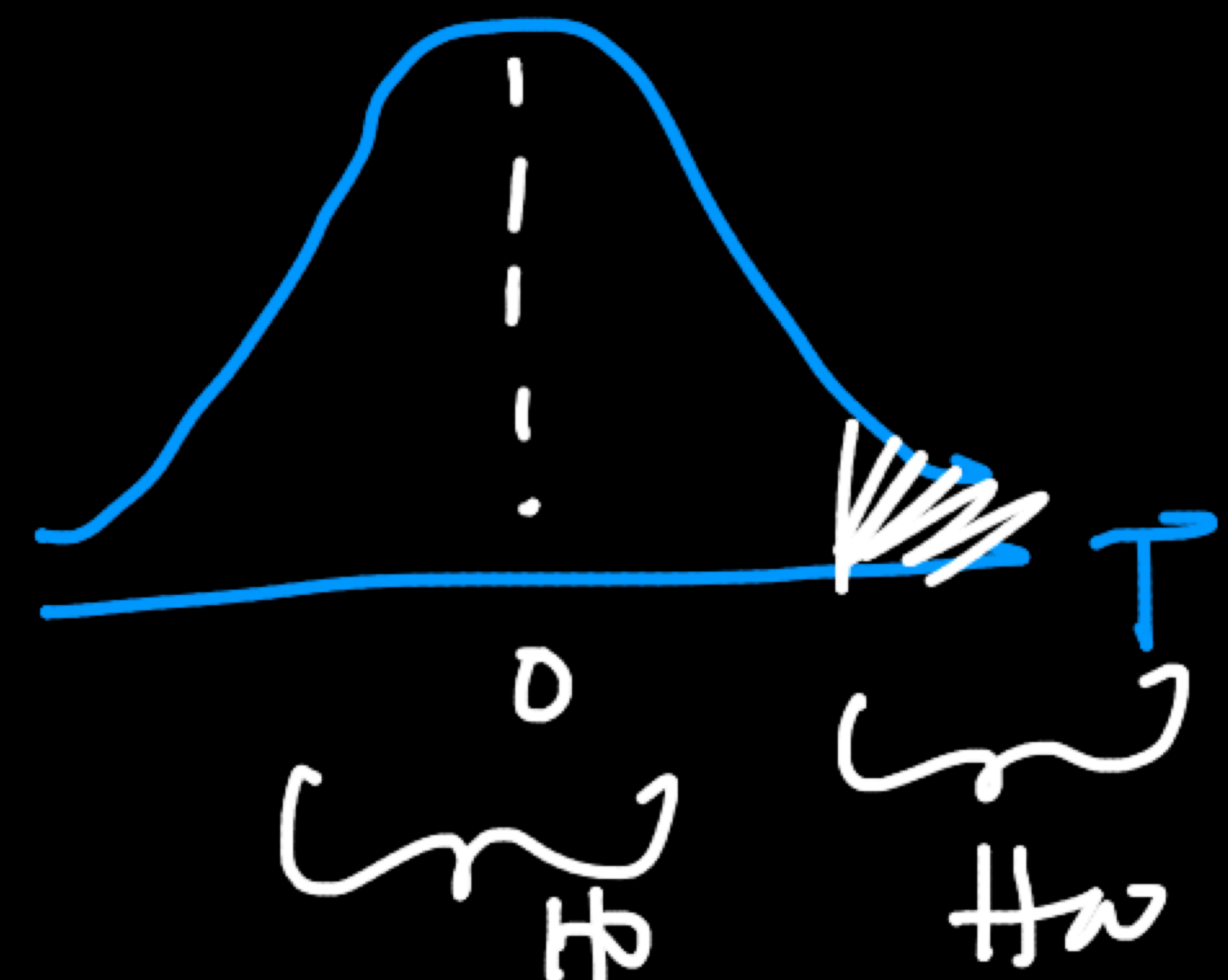
Q

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 > \mu_2$

$z(0,1)$

under H_0



$$T_{obs} = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

39 / 40

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:27

Q

(a) Hyp. test
z-test → $H_0: \mu_1 = \mu_2$
$H_a: \mu_1 \neq \mu_2$

(b) 95 C.I on pop. mean (CLT)

Med1 — 8 to 12

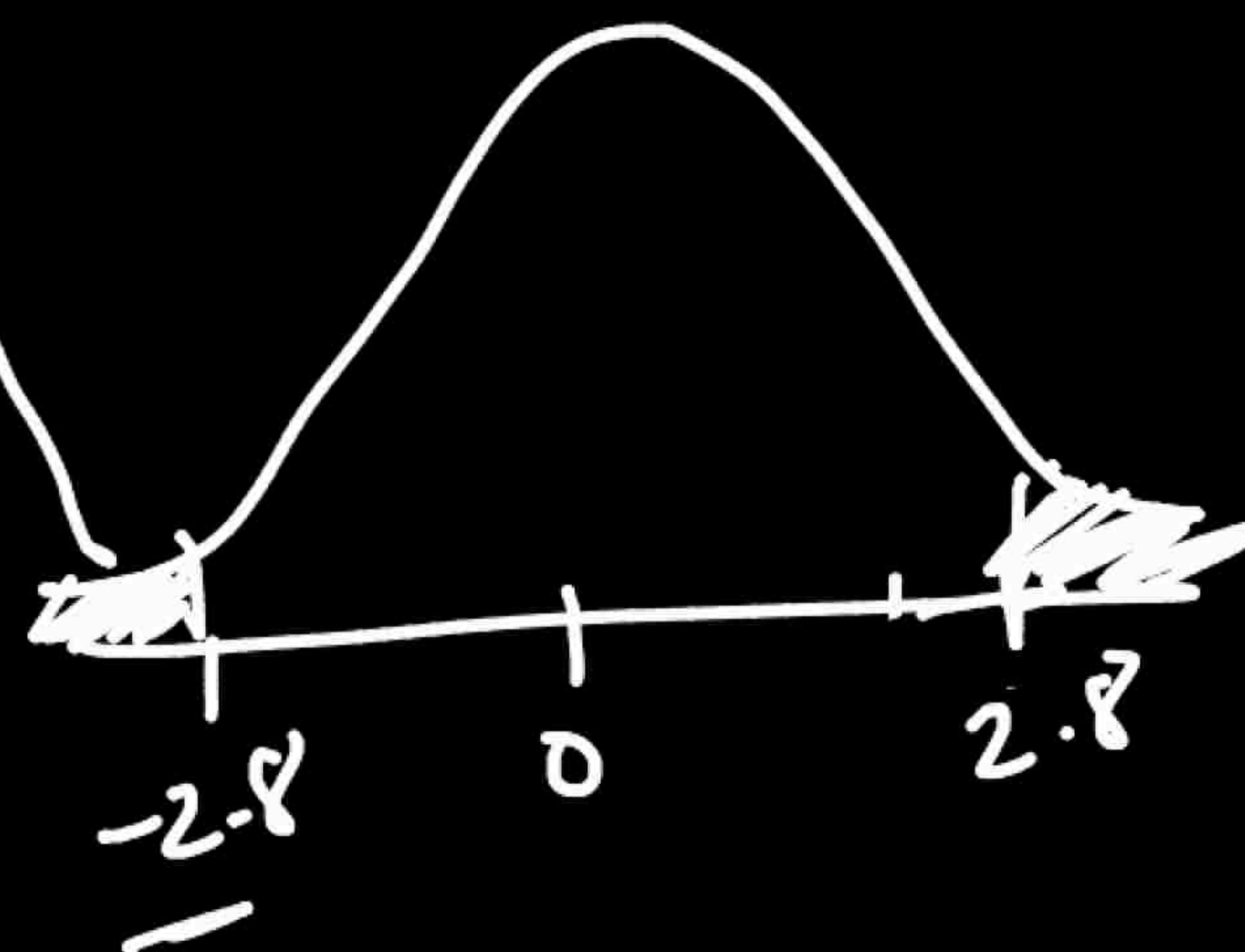
Med2 — 11 to 15

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:29

Q

$T_{obs} = 2.8$

$CDF(-2.8) \times 2$

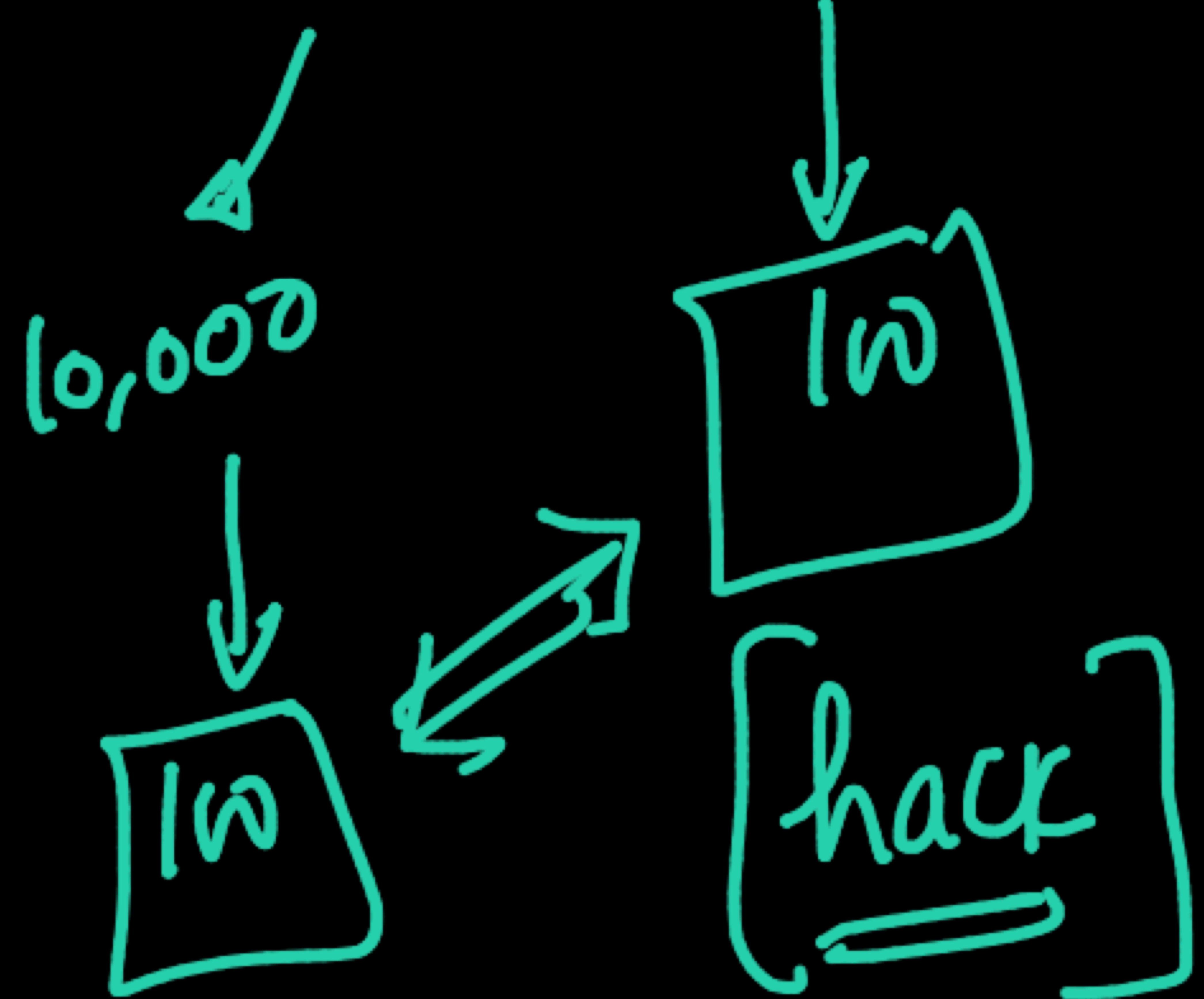


Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:34

Q

$n_1 >> n_2$

{ n_1 and n_2
case Compatible



42 / 42

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:35

Q

hack

$$\sigma_1 \approx s_1$$
$$\sigma_2 \approx s_2$$

{as n_1 & n_2 are large}

$$\mu_1 \approx M_1$$
$$\mu_2 \approx M_2$$

43 / 43

Q

$$\left.\begin{matrix} n_1 = 100 \\ n_2 = 90 \end{matrix}\right\} \rightarrow \text{T-test} \Rightarrow \text{(later)}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:36

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:39

Q

✓ Sample 1 → non Gaussian → μ_1 & σ_1 are not ∞

✓ Sample 2 → Gaussian

Z-test → σ_1 & σ_2 ✓

↳ n_1 & n_2 are large

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)
```

https://docs.

46 / 46

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:43

Case:

Youtube

More Money

{ 2 - ads per ad break

Compare the watch time

47 / 48

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:43

Z-test

$1000 = n_1$

2-ads $\longrightarrow$ $\overbrace{x_1, x_2 \;----\; x_{1000}}$ $\boxed{\mu_1}$

$\updownarrow$

A/B

1-ad $\longrightarrow$ $x_1' \; x_2' \;---\; x_{1000}'$ $\boxed{\mu_2}$

$1000 = n_2$

Task: Q does the (pop) mean w.r.t change due to 2-ads

48 / 48

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:43

KStest_Ttest.ipynb - Colabora

colab.research.google.com/drive/1lWuS8AxULBBold2GlfBKUeT1RSj3fFAJ#scrollTo=x1PceS2FecLE

+ Code + Text Reconnect

Control / Treatment

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)
```

https://docs.

49 / 49

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)
```

https://docs.

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)
```

https://docs.

KStest_Ttest.ipynb - Colaboratory

colab.research.google.com/drive/1lWuS8AxULBBold2GIfBKUeT1RSj3fFAJ#scrollTo=x1PceS2FecLE

+ Code + Text

Reconnect

```
!ls -lrt
```

```
total 872
drwxr-xr-x 1 root root   4096 Jul 13 13:43 sample_data
-rw-r--r-- 1 root root 887610 Jul 15 09:58 ab_test_data.csv
```

```
!cat ab_test_data.csv
```

```
2018-04-24,672,0,1.204038745515143,treatment
2018-10-27,735,0,1.720859204825507,treatment
2018-03-21,756,0,1.6596750020799753,treatment
2018-08-09,665,0,1.6993040851610293,treatment
2018-05-31,662,0,2.156321897846803,treatment
2018-05-28,267,0,1.2101014751056542,control
2018-12-05,157,1,3.262553148000151,control
2018-06-20,921,0,2.227408455175863,treatment
2018-08-14,143,0,3.190808386607947,control
2018-08-29,17,0,1.8798011760876796,control
2018-01-16,628,0,0.67496075296494,treatment
2018-08-30,621,0,10.644234566020007,treatment
2018-02-11,832,0,1.0737883660596945,treatment
2018-01-09,166,0,6.257587343737802,control
2018-03-21,570,0,2.692056020667736,treatment
2018-03-06,205,0,3.9752558540897067,control
2018-12-22,782,0,2.0935524794099583,treatment
2018-08-22,282,1,2.7272375275111767,control
2018-01-20,83
```

52 / 52

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:46

KStest_Ttest.ipynb - Colaborat

colab.research.google.com/drive/1lWuS8AxULBBold2GIfBKUeT1RSj3fFAJ#scrollTo=bi-Mwf5ugQ8j

+ Code + Text Reconnect

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

	date	customer_id	premium	watch_time_hrs	customer_segmnt
15560	2018-12-09	298	0	1.181086	control
11054	2018-12-17	562	1	2.537064	treatment
6255	2018-09-04	732	0	5.251925	treatment
9187	2018-08-24	521	0	9.019050	treatment
12888	2018-07-19	88	0	1.495338	control
...	...	...	...	...	...
6432	2018-06-15	612	0	5.108603	treatment
6727	2018-07-27	62	0	2.944847	control
17074	2018-01-11	244	0	7.780596	control
15845	2018-12-22	618	0	1.459799	treatment
10471	2018-12-08	948	0	2.855943	treatment

100 rows × 5 columns

53 / 53

10471	2018-12-08	948	0	2.855943	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	0.0[illegible]		

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

2

```
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

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

```
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
75%	747.000000	0.000000	4.204673
max	999.000000	1.000000	$\boxed{10007.648185}$

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

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:48
KStest_Ttest.ipynb - Colaborat
colab.research.google.com/drive/1lWuS8AxULBBold2GIfBKUeT1RSj3fFAJ#scrollTo=Kxc8aPWsgmOX
+ Code + Text Reconnect

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

99.9 percentile

```
# 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_
```

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

```python
# 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
```

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

```
21.356607722117484
```

```python
q998 = ab_test_data["watch_time_hrs"].quantile(0.998)
ab_test_data_
```

KStest_Ttest.ipynb - Colaboratory

```python
# 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 th percentile

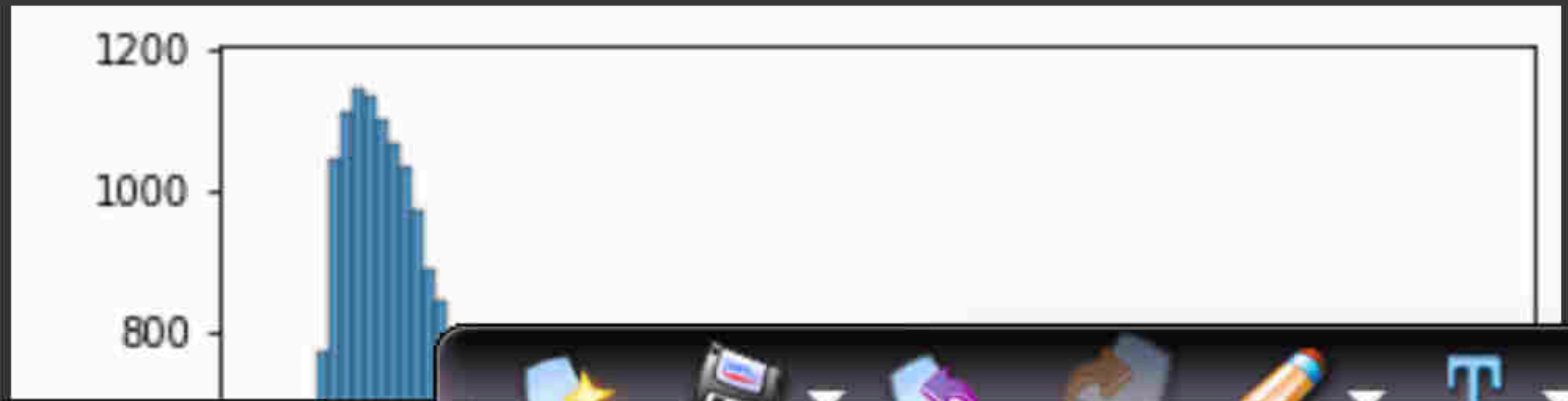
```python
ab_test_data["watch_time_hrs"].quantile(0.998)
```

```
21.356607722117484
```

remove 0.2 pe

```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()
```



KStest_Ttest.ipynb

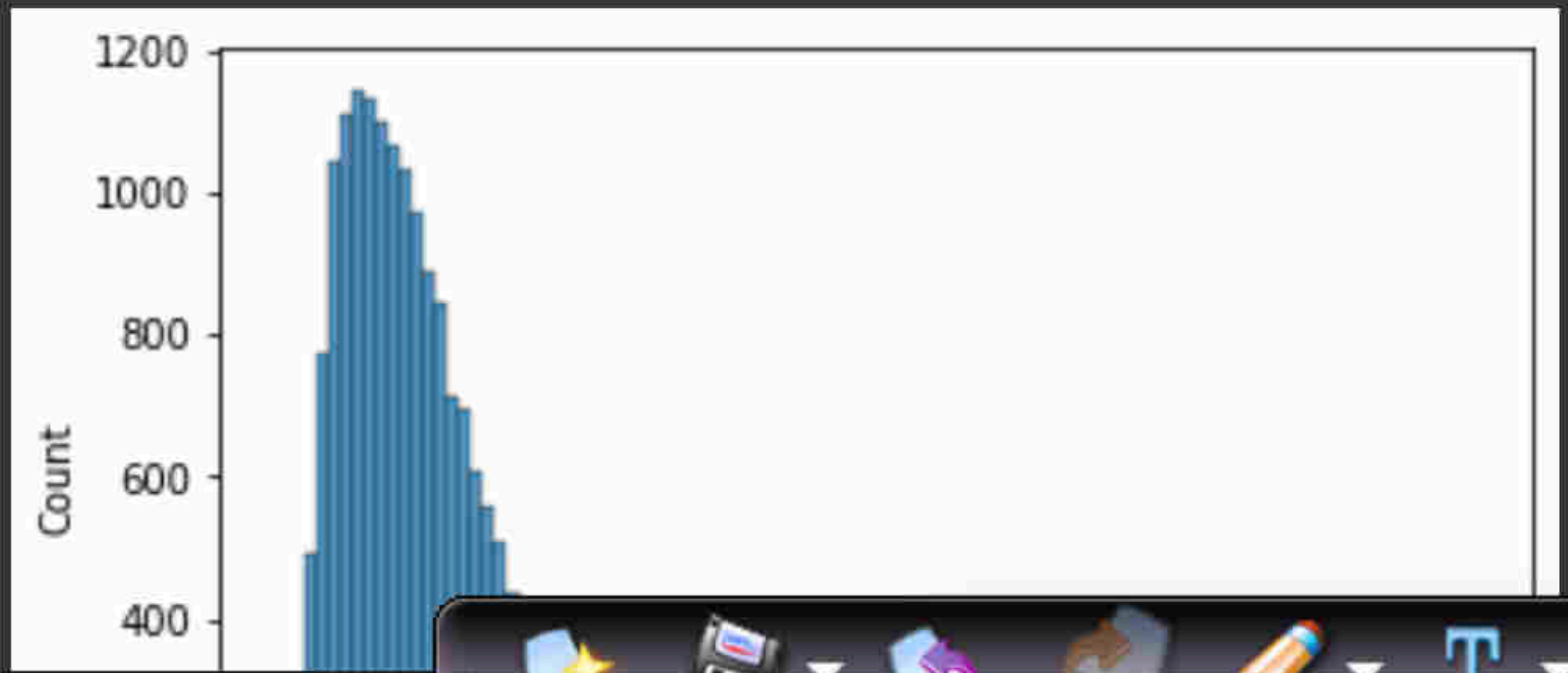
File Edit View Insert Runtime Tools Help All changes saved

```
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()
```



60 / 60

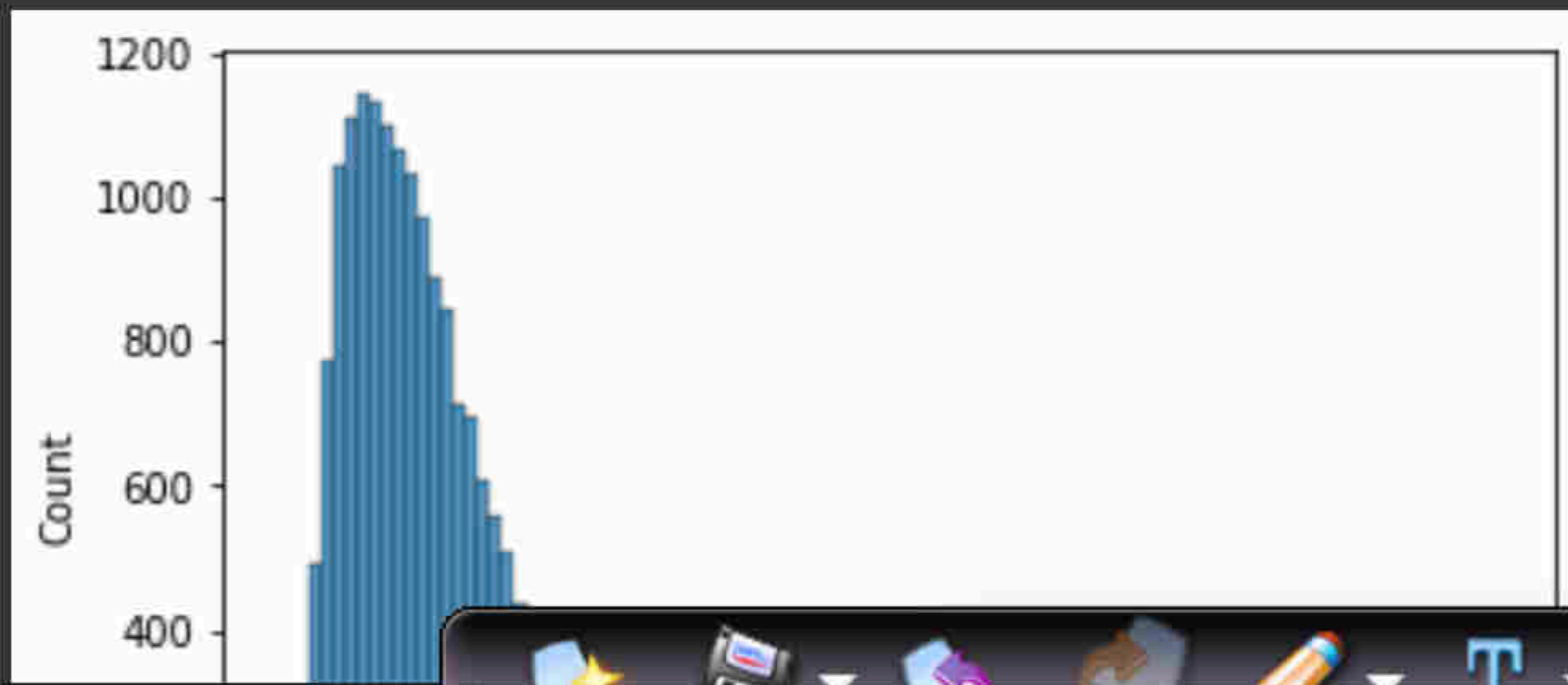
KStest_Ttest.ipynb

```
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)]
```

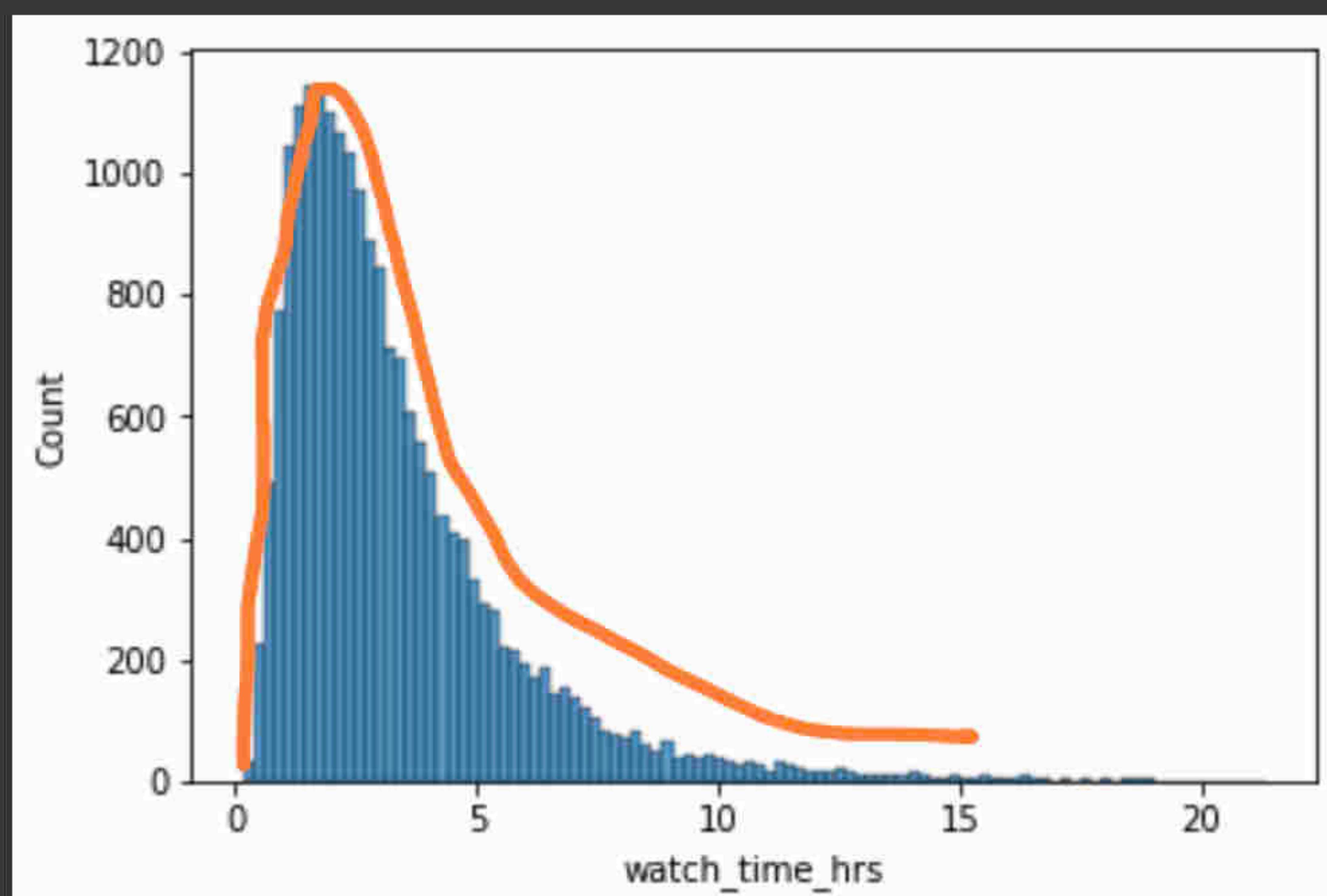
```
# disb of watch-time
sns.histplot(ab_test_data_no_out['watch_time_hrs'], bins=100)
plt.show()
```



61 / 61

KStest_Ttest.ipynb

```
sns.histplot(ab_test_data_no_out['watch_time_hrs'], bins=100)
plt.show()
```



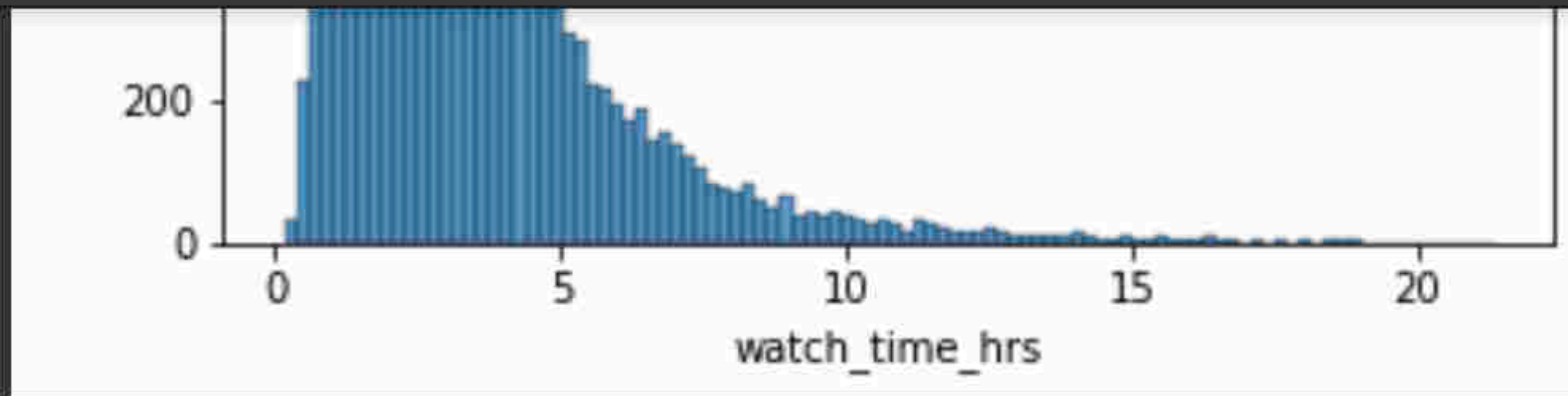
Small task
↓
log-normal - Check

```
#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"]
```

62 / 62

KStest_Ttest.ipynb - Colaboratory

colab.research.google.com/drive/1lWuS8AxULBBold2GIfBKUeT1RSj3fFAJ#scrollTo=Kxc8aPWsgmOX




```
#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
```

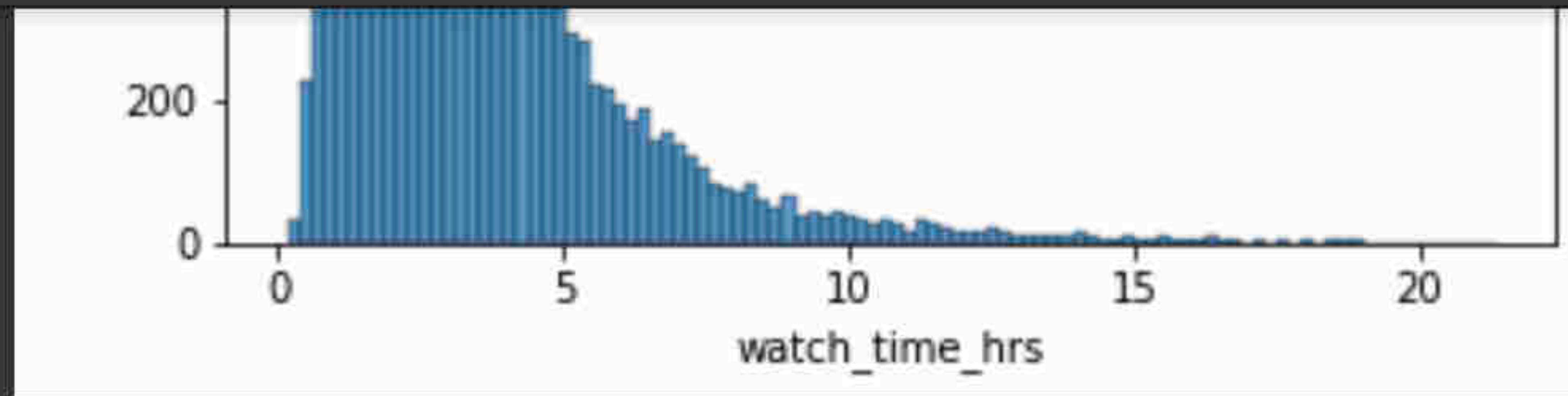
```
(9987, 5)
```

```
# 2-sample z-test as n1 nad n2 are large.
# Refer: http                                   test.html
```

63 / 63

KStest_Ttest.ipynb - Colaboratory

colab.research.google.com/drive/1lWuS8AxULBBold2GIfBKUeT1RSj3fFAJ#scrollTo=Kxc8aPWsgmOX




```
#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"]
```

100

```
ab_test_control_data.shape
```

n_1 (9973, 5)

```
ab_test_treatment_data.shape
```

(9987, 5)

n_2

```
# 2-sample z-test as n1 nad n2 are large.
# Refer: http ... test.html
```

64 / 64

```
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)

n_1 & $n_2 < 30$

↓

T-test

```
# 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"])
```

(15.96034913022092, 2.4137738128170024e-57)

T-Test

```
ab_test_control_data.shape
```

n1

(9973, 5)

```
ab_test_treatment_data.shape
```

(9987, 5)

n2

```
# 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"])
```

(15.96034913022092, 2.4137738128170024e-57)

T-Test

↳ 3 cells hidden

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:53

KStest_Ttest.ipynb - Colaboratory

colab.research.google.com/drive/1lWuS8AxULBBold2GIfBKUeT1RSj3fFAJ#scrollTo=Kxc8aPWsgmOX

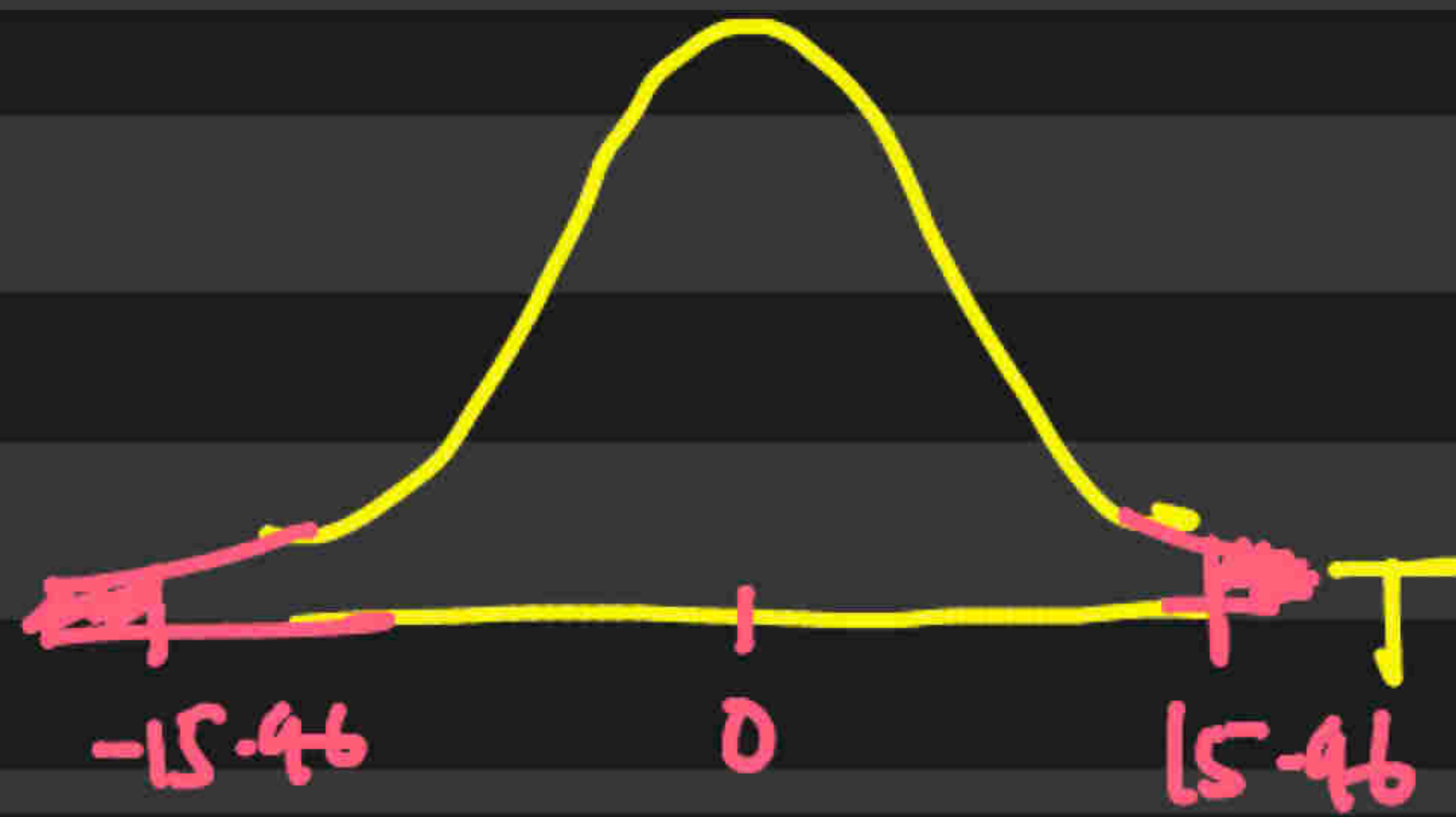
+ Code + Text Reconnect

```
ab_test_control_data.shape
```

(9973, 5)

```
ab_test_treatment_data.shape
```

(9987, 5)




```
# 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"])
```

(15.96034913022092, 2.4137738128170024e-57)

Tobs

p-val

$\alpha = 0.001\%$

2.4×10^{-57} pval

T-Test

↳ 3 cells hidden

67 / 67

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:54

✓

2-ads vs 1-ad

m_1 vs M_2

n_1 & n_2 lage

z-test fail

1. CLT assumptions are broken
2. n_1 & n_2 are < 30
3. σ_1 & σ_2 are unknown & unestimatable

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:03

Q:

✓

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 > \mu_2 \end{cases}$$

Right tailed test

$$T_{obs} = \frac{m_1 - m_2}{\sqrt{\cdots}}$$

$T \sim N(0,1)$

under H_0

70 / 70

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:07

T-test

Sample 1

med 1 $\longrightarrow n_1 \leq 30$

$x_1 \ldots x_{n_1}$

med 2 $\longrightarrow n_2 \leq 30$

$x_1' \; x_2' \ldots x_{n_2}'$

Sample 2

- very popular
- small sample sizes
- Barley @ Guiness

71 / 73

Purpose / Task

Compare pop. Means

when n_1 & n_2 are small

do NOT know σ_1 & σ_2

cannot estimate

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:16

① $\begin{cases} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 \neq \mu_2 \end{cases}$

sensible

② $T_{obs} = \dfrac{m_1 - m_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

replaced σ_1 with s_1

σ_2 with s_2

$s_1 \approx \sigma_1$ ✗

73 / 79

$Z(\underline{0},\underline{1})$

Under H_0

[dist of $T \sim t\text{-}\underline{d}ist\,(\nu)$]

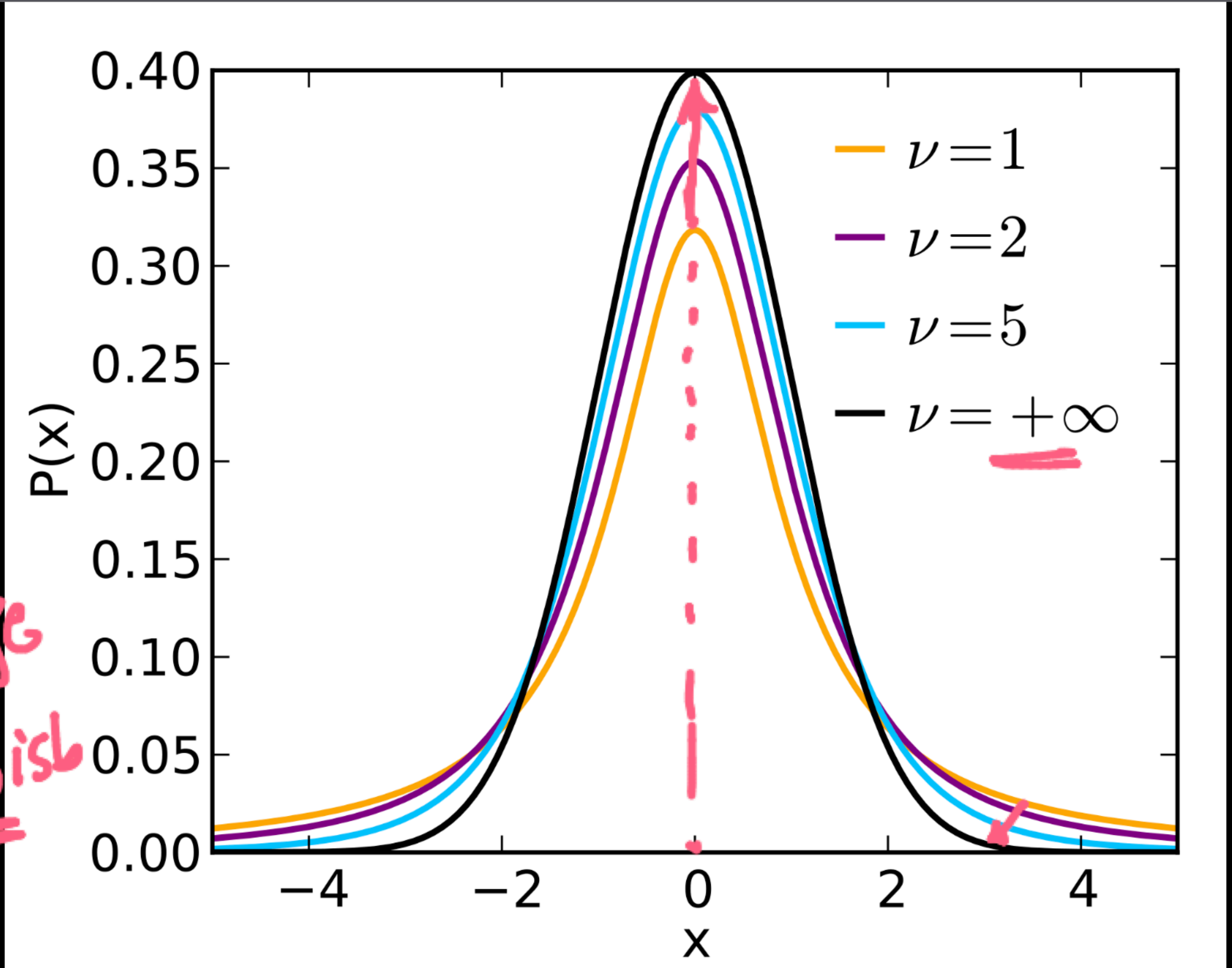
$\nu \rightarrow \underline{n_1 + n_2 - 2}$

[degrees of freedom]

PDF

- Symm
- Centered at zero
- ν is v. large

t-dist ~ Z-dist



$\nu \uparrow$

Plot of the density function for several members of the Student t family.

More details

Skbkekas - Own work

CC BY 3.0

Print/export
Download as PDF
Printable version

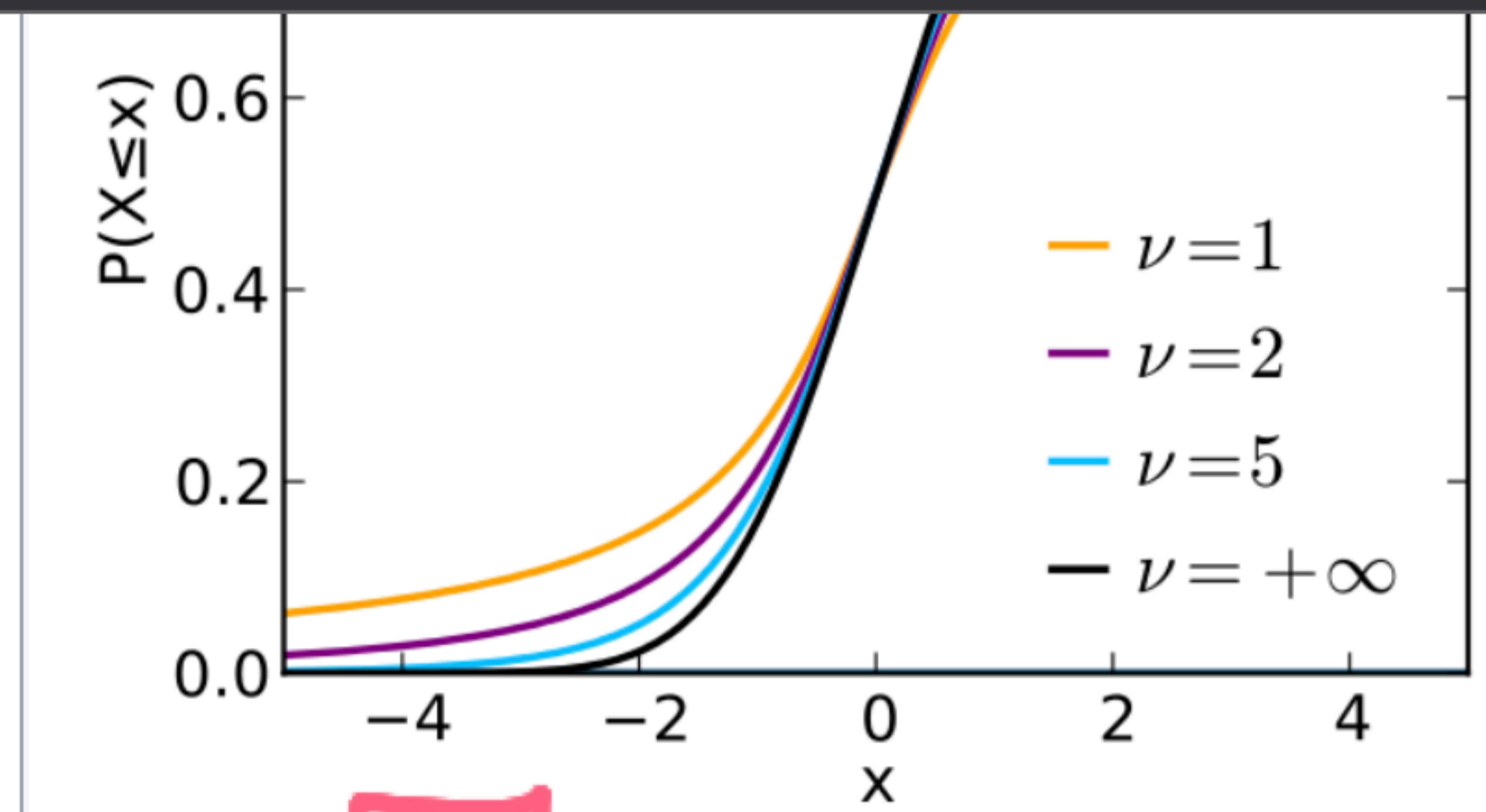
In other projects
Wikimedia Commons

Languages
العربية
Deutsch
Español
Français
Jawa
日本語
Português
Русский
中文
24 more
Edit links

construct a confidence interval for the true mean.

The *t*-distribution is symmetric and bell-shaped, like the normal distribution. However, the *t*-distribution has heavier tails, meaning that it is more prone to producing values that fall far from its mean. This makes it useful for understanding the statistical behavior of certain types of ratios of random quantities, in which variation in the denominator is amplified and may produce outlying values when the denominator of the ratio falls close to zero. The Student's *t*-distribution is a special case of the generalized hyperbolic distribution.

Contents [hide]



Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty, \infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$ where ${}_2F_1$ is the hypergeometric function
Mean	0 for $\nu > 1$, otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \le 2$, otherwise undefined
	3, otherwise undefined
Kurtosis	> 4, ∞ for $2 < \nu \le 4$, otherwise

76 / 76

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:13

$$\left\{ T_{obs} = \frac{m_1 - m_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \right.$$

$$\underbrace{T \sim t\text{-dist}(\nu)}$$

$$\nu \longrightarrow n_1 + n_2 - 2$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:51

t-test

n_1 & n_2 are small

$$T_{obs} = \frac{M_1 - M_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

t-dist $(\underbrace{n_1+n_2-2}_{\nu})$

z-test

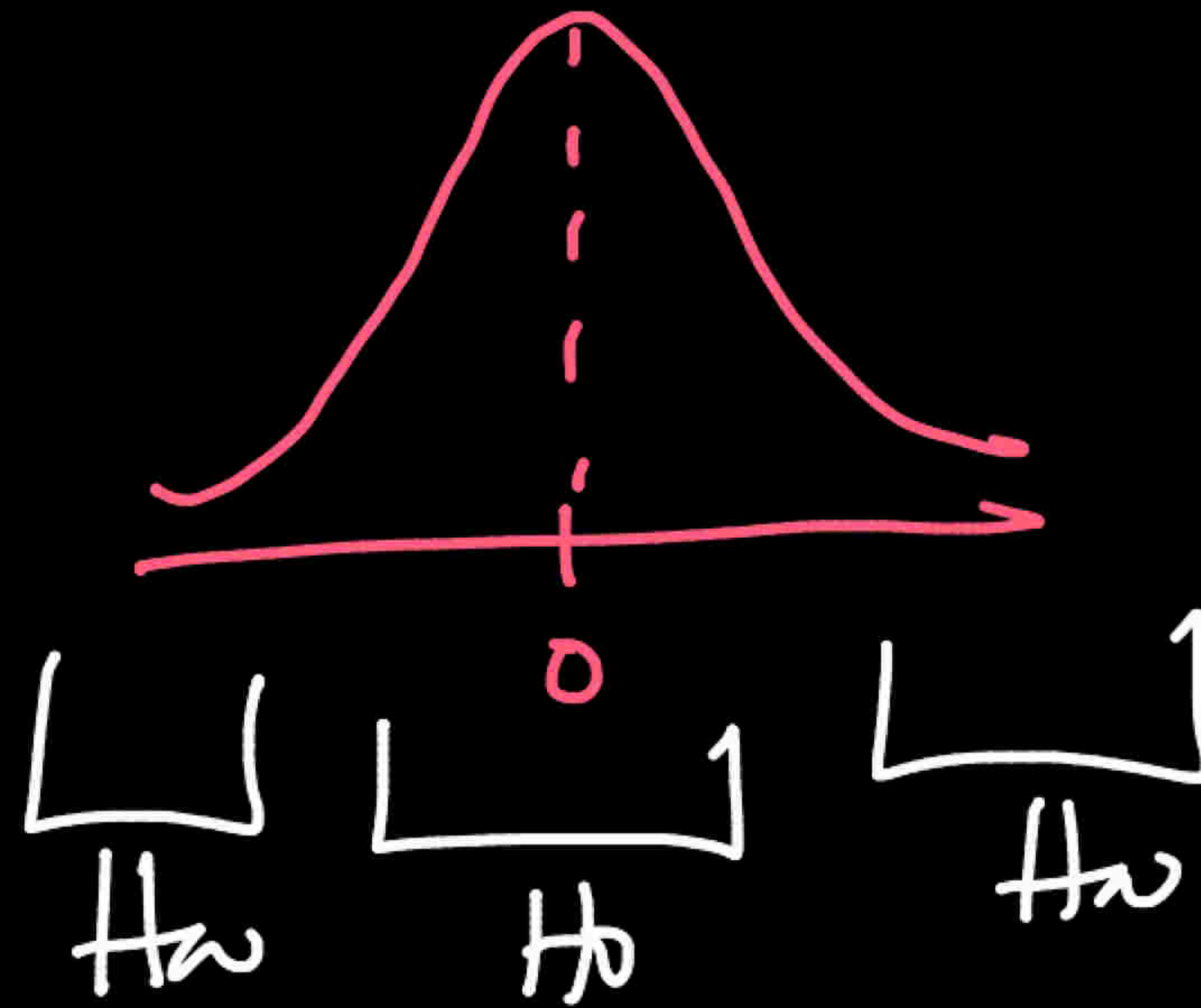
n_1 & n_2 are large

$$T_{obs} = \frac{M_1 - M_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

z-dist $(0,1)$

④ 2-sided or 1-sided

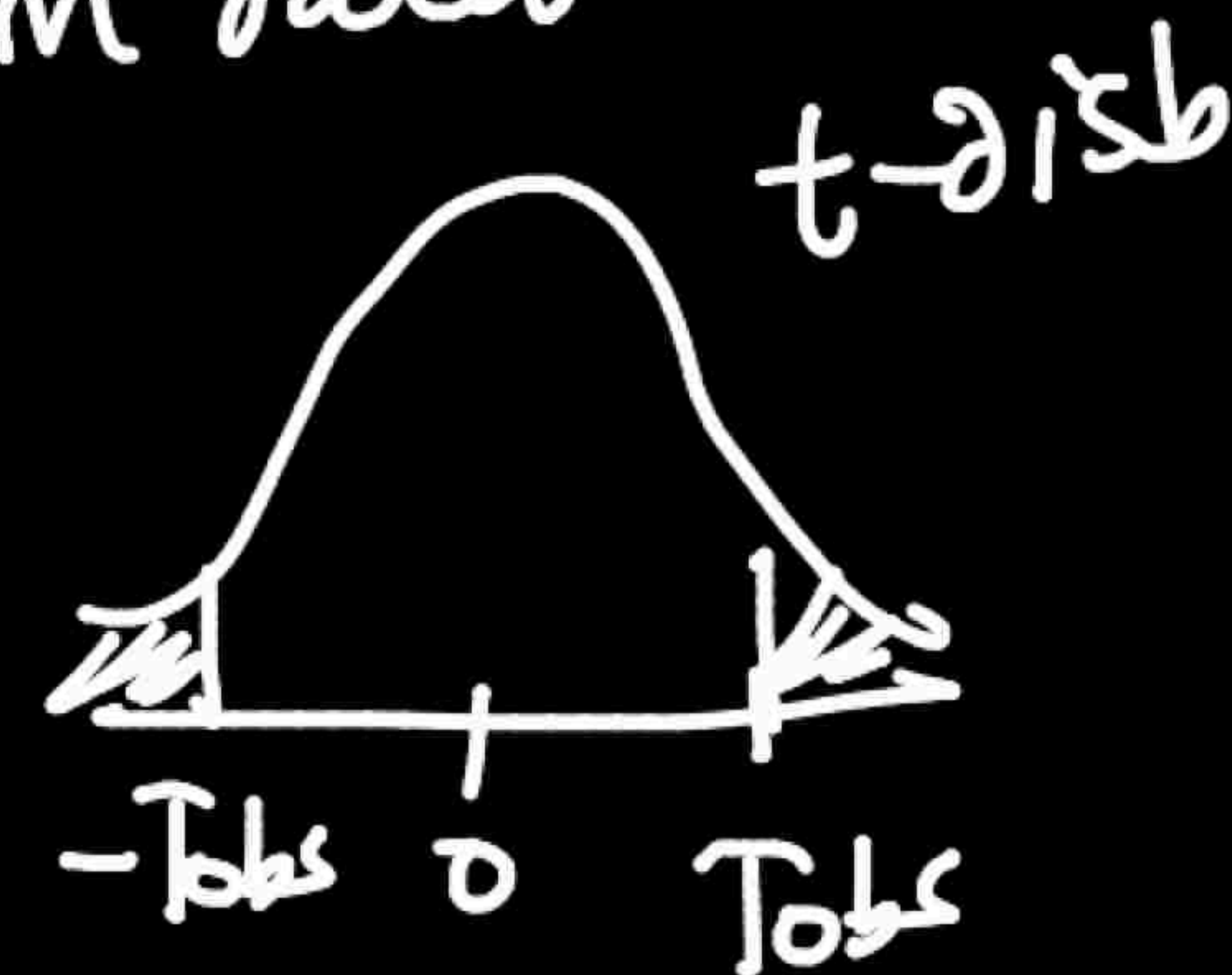
$$T_{obs} = \frac{M_1 - M_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:18

(5) T_{obs} from data

(6) p-val



(7) α-value

p-val vs α-value

80 / 80

T-Test

n1 n2

```
#T-Test
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
```

```
0.555666548844524
```

```
#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)
```

81 / 81

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:19

KStest_Ttest.ipynb - Colabora | Median test - Wikipedia | Student's t-distribution - Wikip

colab.research.google.com/drive/1lWuS8AxULBBold2GIfBKUeT1RSj3fFAJ#scrollTo=5Xplol8wht8k

+ Code + Text Reconnect

T-Test

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

19958

M1 - M2

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

0.555666548844524

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

Ttest_indResult(statistic=15.9634913022092, pvalue=5.438408586231319e-57)

82 / 82

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:21

$$df = n_1 + n_2 - 2 = \underline{19998}$$

t-test

z-test

$$\boxed{\begin{matrix} 5.4 \times 10^{-57} \\ 2.4 \times 10^{-57} \end{matrix}} < \alpha : \underline{5\%}$$

83 / 84

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:21

T-test

large samples

z-test

small samples

(FAIL)

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:24

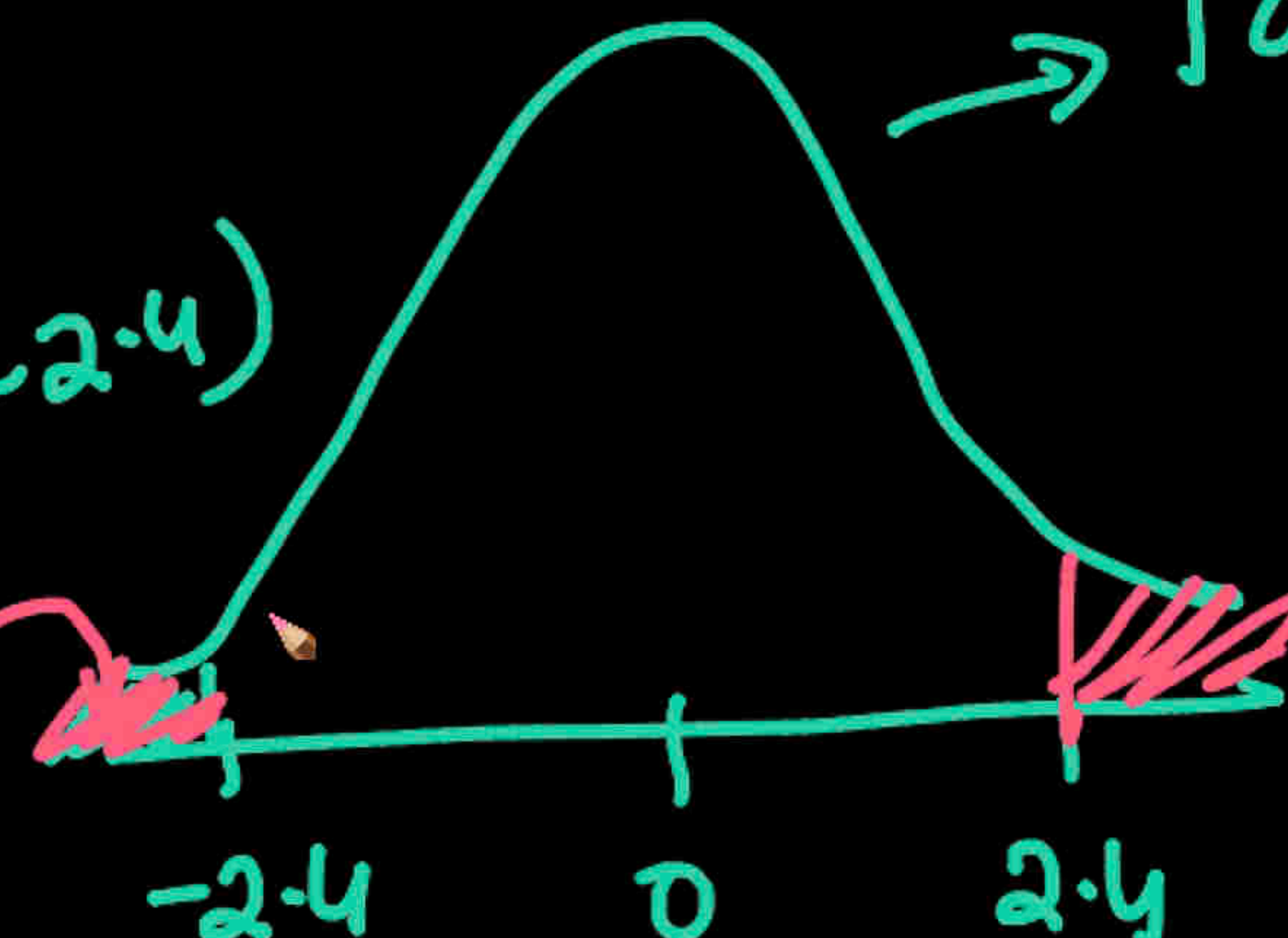
Q

$$dof = n_1 + n_2 - 2 = 50$$

$T_{obs} = -2{\cdot}4$

→ Tdisb(v=50)

CDF-Tdisb(v=50; -2·4)



85 / 85

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

Q

t-dist $\rightarrow$ z-dist

as dof $\rightarrow \infty$ ✓

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:28

z-test

2-sample

$$T_{obs} = \frac{m_1 - \tilde{M}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(e.g)

$$T_{obs} = \frac{M_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

1-sample

$\{x_1 \ldots x_{100}\}$

Indians

87 / 87

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:39

Q

χ^2-test ✓

Compare means → z/t test ✓

dist → KS test ✓

proportions → $\left\{ \frac{5 \text{ click}}{100} \right.$ ✓

$\left\{ \begin{array}{l} \text{means} \\ \text{3 or more} \end{array} \right.$ → ANOVA

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:41

z-test : $T_{obs} \sim Z(0,1)$

T-test : $T_{obs} \sim t(\nu)$

KS test : max gap $\sim \mathrm{Kol}(\ldots)$

$\vdots$

$$\left.\begin{array}{l} Z(0,1) \\ t(\nu) \end{array}\right\}$$

89 / 89

Q

$M_1, M_2 \ldots$

bootstrapping

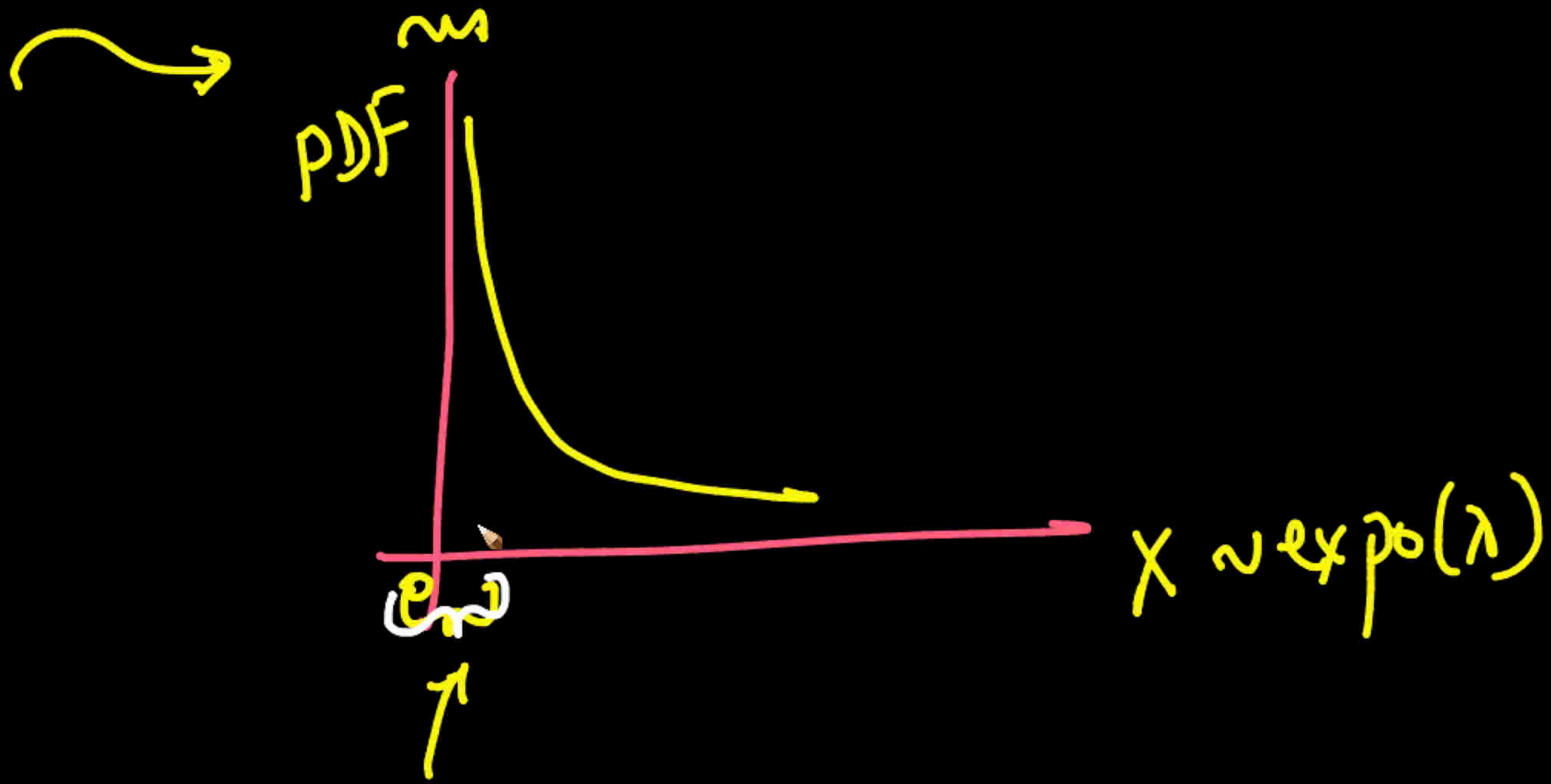
↓

std. error

$$\sigma \approx \boxed{\frac{\sigma}{\sqrt{n}}} \times \sqrt{n}$$

as $n \to \infty$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:46



91 / 92

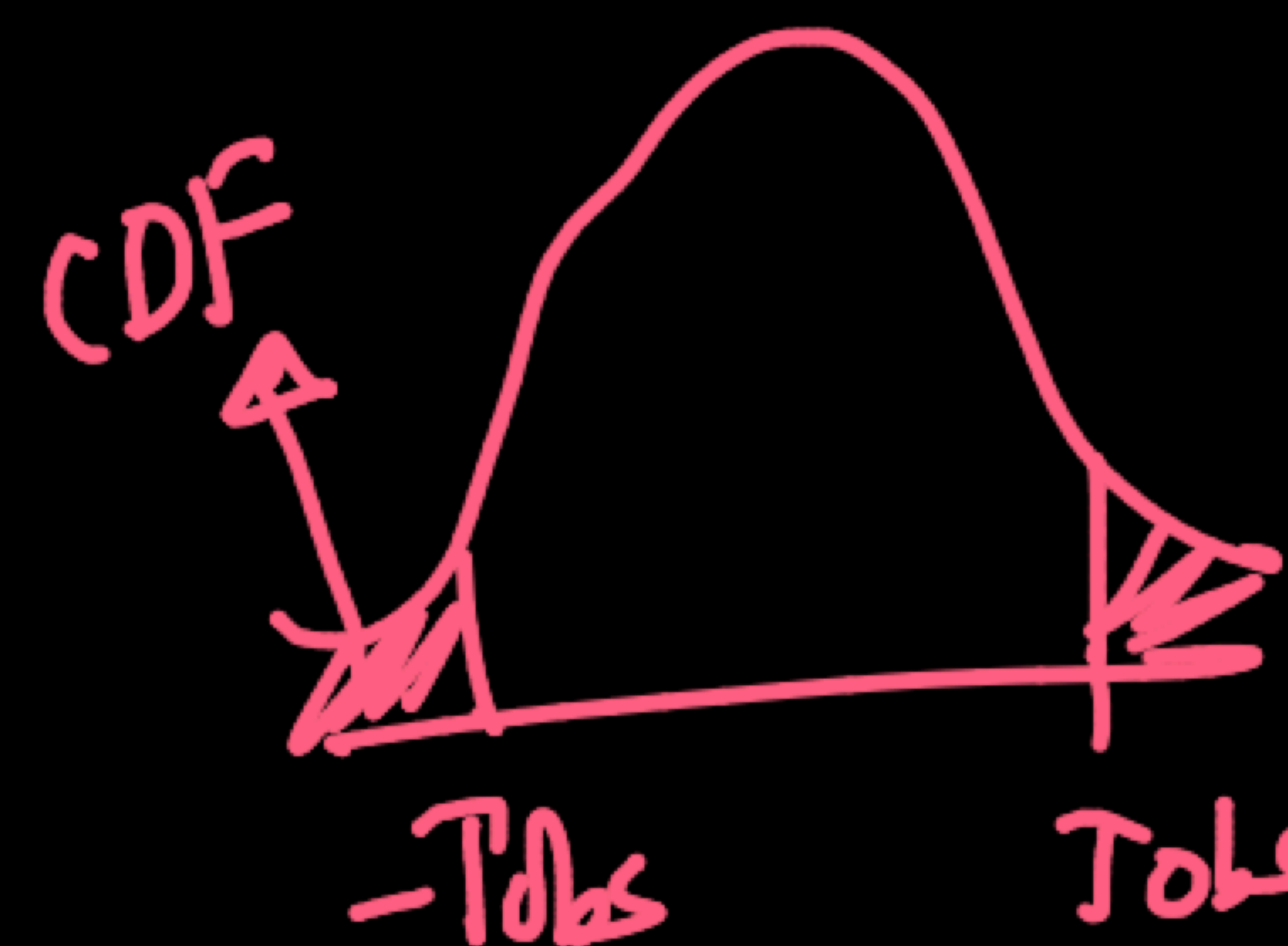
Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:52

T-test

right tailed

$\nu = n_1 + n_2 - 2$

$$T_{obs} = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



92 / 92

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:58

Q

χ^2-test (later)

obs	exp
1800	1.88×10^2
	188

93 / 93

Chrome File Edit View History Bookmarks Profiles Tab Window Help 12:01

Q

$X : x_1 \; x_2 \; . \; . \; x_n$

$E(X)$

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sum_{x} x \cdot \overbrace{P(X=x)}$$

PMF

94 / 94

Chrome File Edit View History Bookmarks Profiles Tab Window Help 12:04

Task

Youtube e.g

compare Means

Task

Cleaned & removed outliers

IQR / percentile

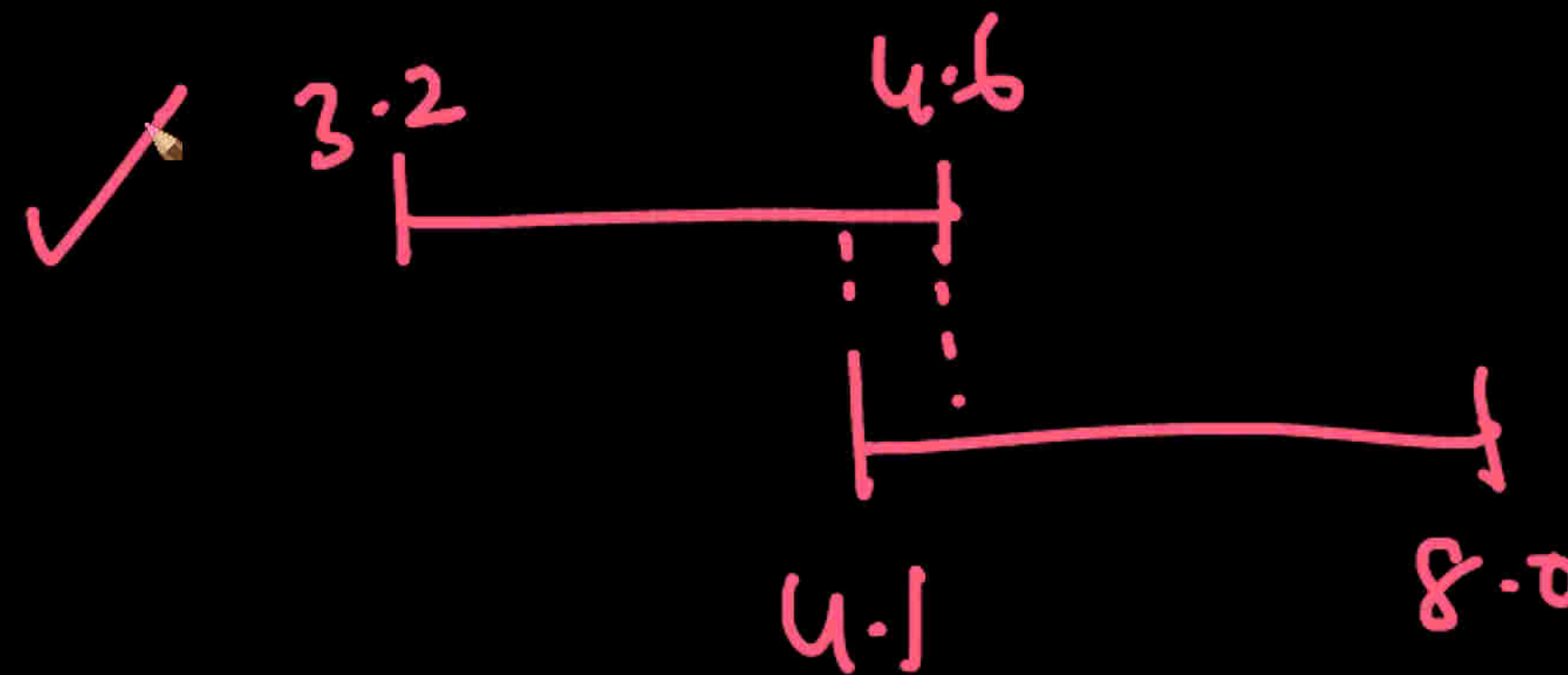
z-test ✓

95 / 96

Task $\longrightarrow$ what series of tech do I need to apply

Q

95% C.I of pop. Mean



Chrome File Edit View History Bookmarks Profiles Tab Window Help 12:05

n_1

n_2

Chrome File Edit View History Bookmarks Profiles Tab Window Help 12:11
KStest_Ttest.ip | Median test - W | scipy.stats.ttes | scipy.stats.ttes | scipy.stats.ttes | Student's t-tes | ttest.pdf
math.arizona.edu/~jwatkins/ttest.pdf
ttest.pdf 1 / 6 150%

The t Test

April 3-8, 2008

Again, we begin with independent normal observations $X_1.\dots,X_n$ with unknown mean μ and unknown variance σ^2. The likelihood function

99 / 99