END: ...

Topics:

- Misc: (case-study)
  - Z-score
  - Feature Transforms
  - Box-cox Transform
  - Extended-Bayes Thm
- KS-test ✓
- T-test ✓
- A/B Testing in real-world

- Z-proportions test

- Framework
- Z-test

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:11

(Standard-score)

observations X: → any distribution

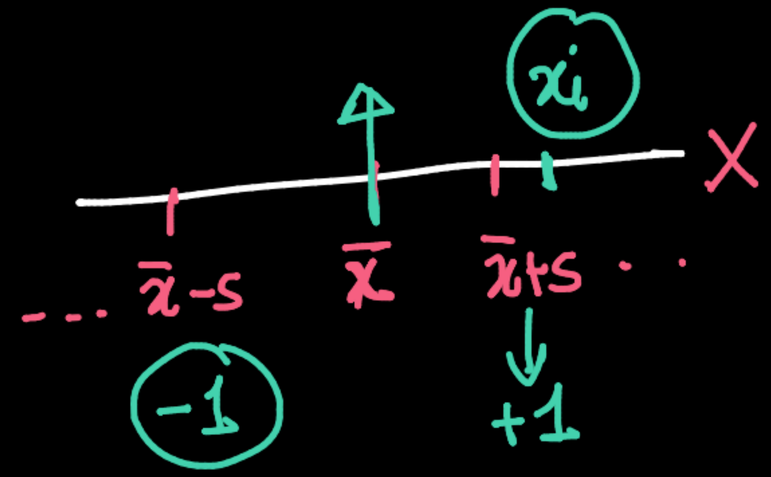
[$x_1, x_2 \dots x_n$: sample

x_i → z-score:

sto-dev that x_i is at in the dist of X

mean = $\bar{x}$

sto-dev = s



z-score

z-dist

↓

N(0, 1)

mean = 0

1 sto dev above mean = +1

1 sto dev below mean = -1

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

✓ Standardization

$$\text{Z-score}(x_i) = \left\{ \frac{x_i - \bar{x}}{s} \right\} \checkmark$$

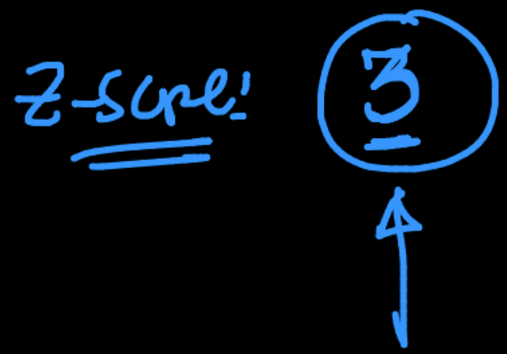
any new-term → Google search away

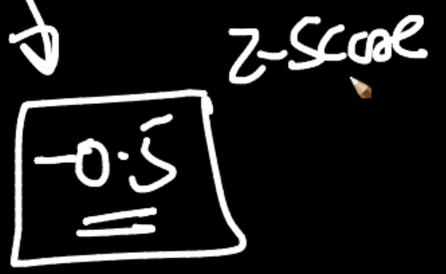
Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:11

recovery times

medicine 1 → X

ith patient : 16.5 days





Z-score is -ve :

$$\frac{x_i - \bar{x}}{s} < 0 \quad \Rightarrow \quad x_i < \bar{x}$$

($\bar{x}$ → mean)

"Feature" Transforms:

non-gaussian

X: $x_1, x_2, \ldots x_i \ldots x_n$: delivery times of Amazon

↳ right-skewed ; could be log-normal (let)

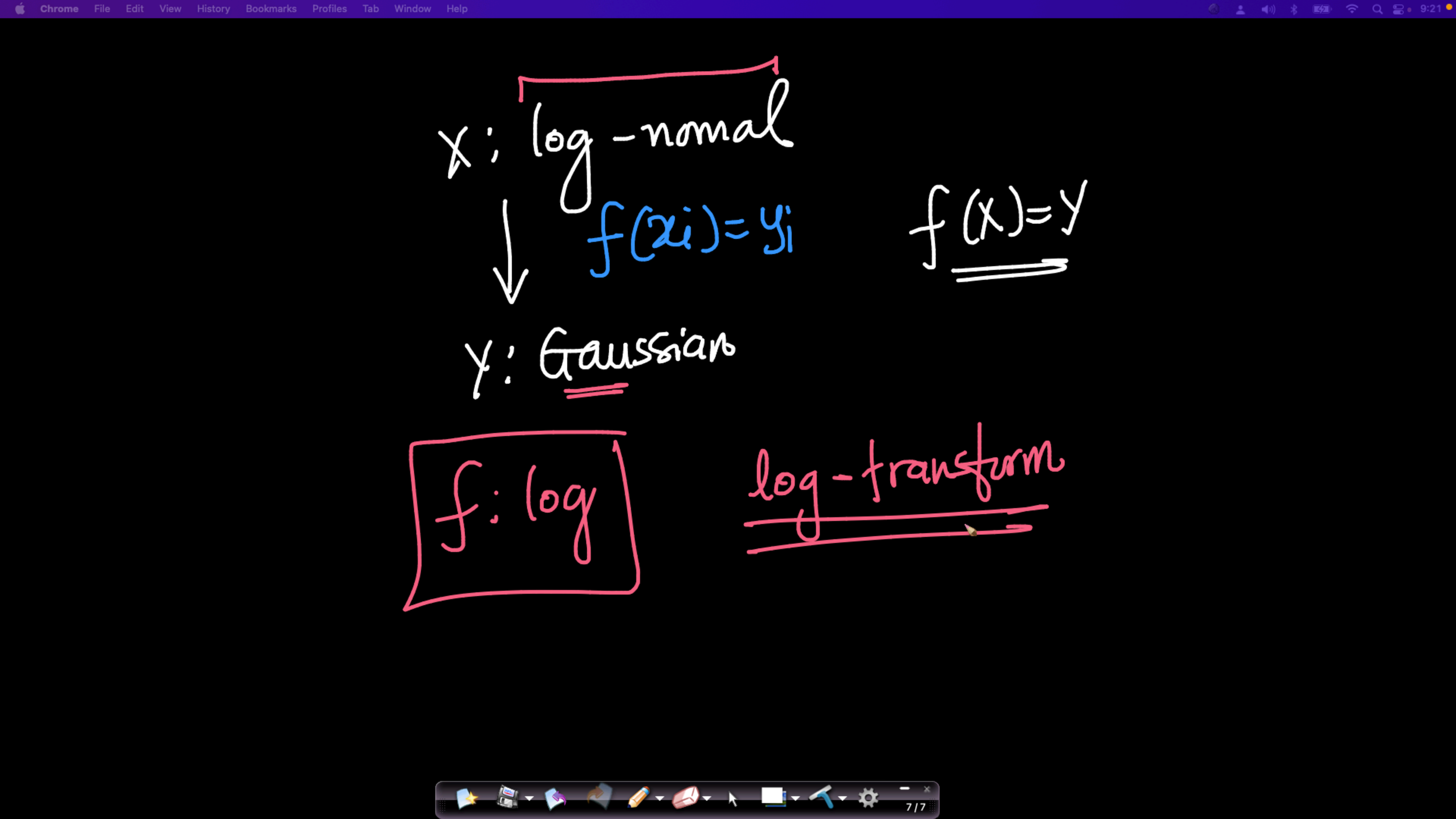
Stats

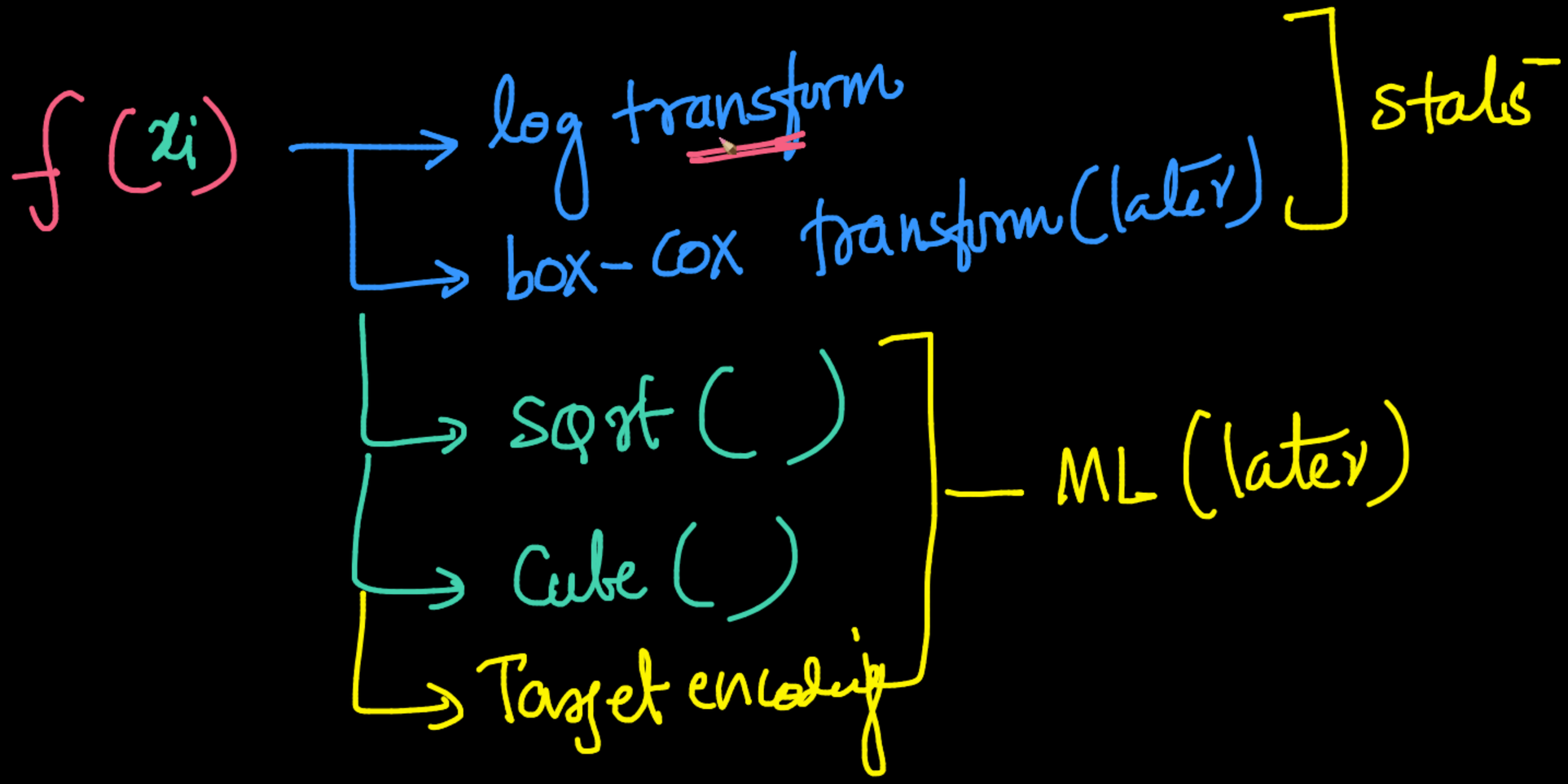
$f(\,)$

$f(x_i) = y_i$

Y: $y_1, y_2 \ldots y_i \ldots y_n$

$Y \sim$ Gaussian


x: log-nomal
f(xi)=yi
f(x)=y
y: Gaussian
f: log
log-transform


f(xi)
log transform
box-cox transform (later)
stats
sqrt ( )
Cube ( )
Target encoding
ML (later)

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:26

popular transforms

Task!

non-Gaussian → Gaussian.

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

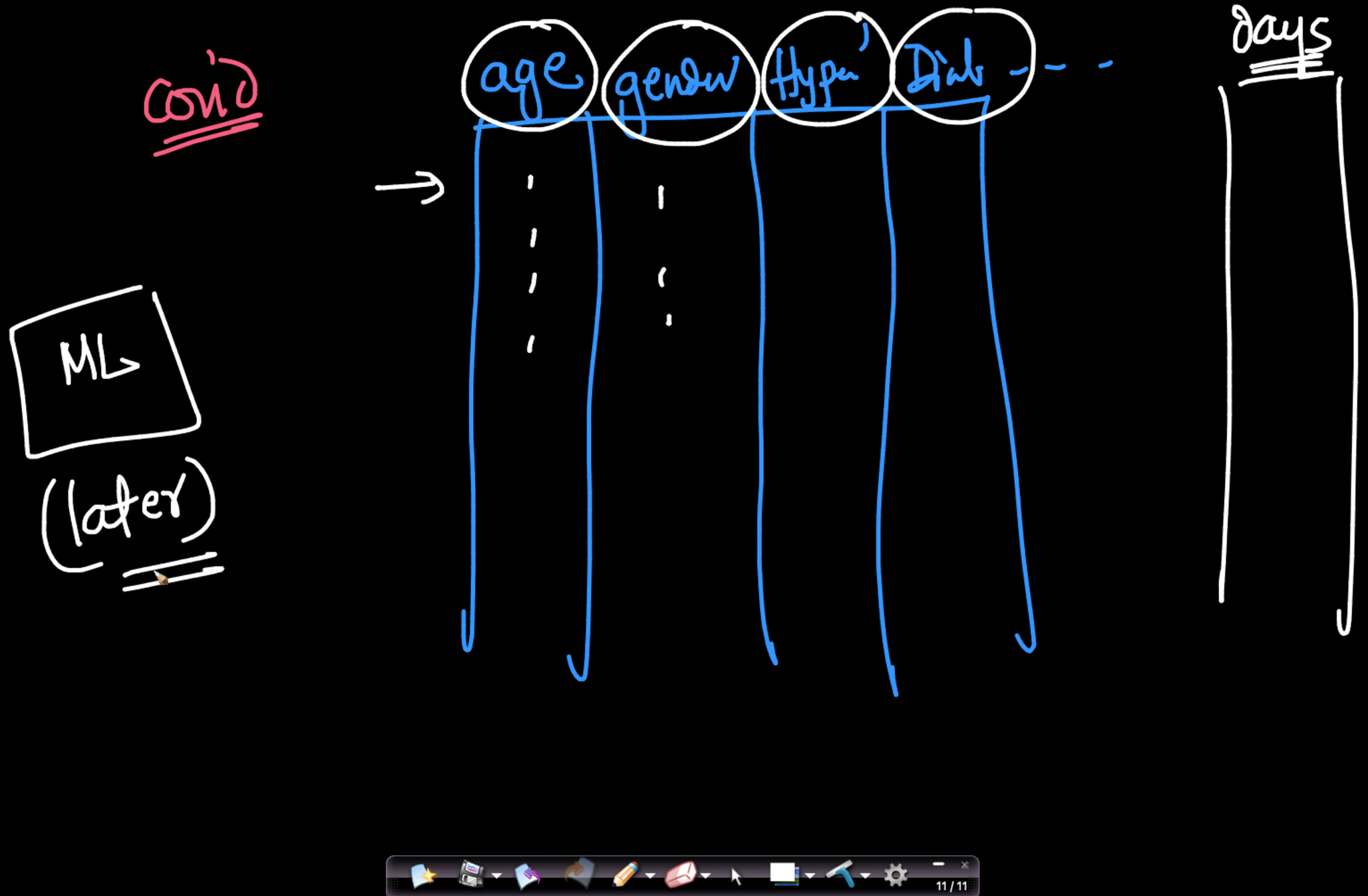
X → $\sqrt{\left(\frac{x_i - \mu}{\sigma}\right)}$ → $\sqrt{f(x_i) = y_i}$

mean = 0 ; stddev = 1

distb may not change

→ min-max [0, 1] → $\left(\frac{x_i - \min x}{\max x - \min x}\right)$ → $f(x_i)$ ✓

[its not gaurantee to convert to Gaussian


covid
age
gendw
Hypn'
Diab
days
ML
(later)

scipy-stats

box-cox: → generalization of log-transform

$x_1, x_2 \ldots x_i \ldots x_n$

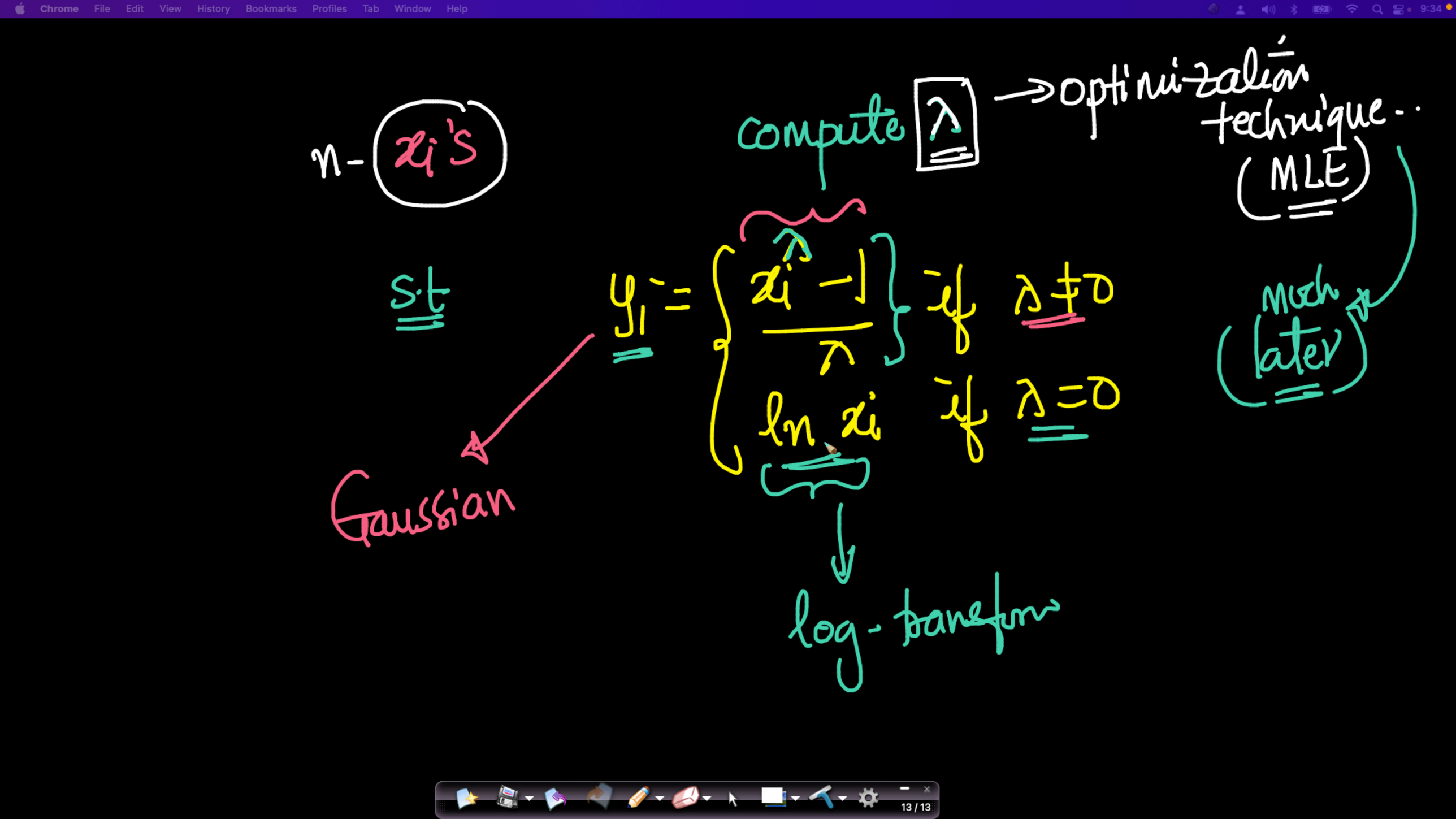
- tries its best
- not guaranteed to always work

conditions

$x_i > 0$

$y_1, y_2 \ldots y_i \ldots y_n$ ~ Gaussian

→ QQ-plot (or) KS test (later)


n - xi's
s.t
compute λ
→ optimization technique..
(MLE)
much later
yi = { (xi^λ - 1)/λ } if λ ≠ 0
ln xi if λ = 0
Gaussian
log-transform

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:38

Prog-2

$[x_1 \; x_2 \; - \; \bar{x} \; - \; x_n]$

$\downarrow f(x)$

$y_1 \; y_2 \; - \; \bar{y} \; y_i \; - \; y_n$ ✓

k-grps ✓

ANOVA (next classes)

↓

[Obs to be Gaussian dist]

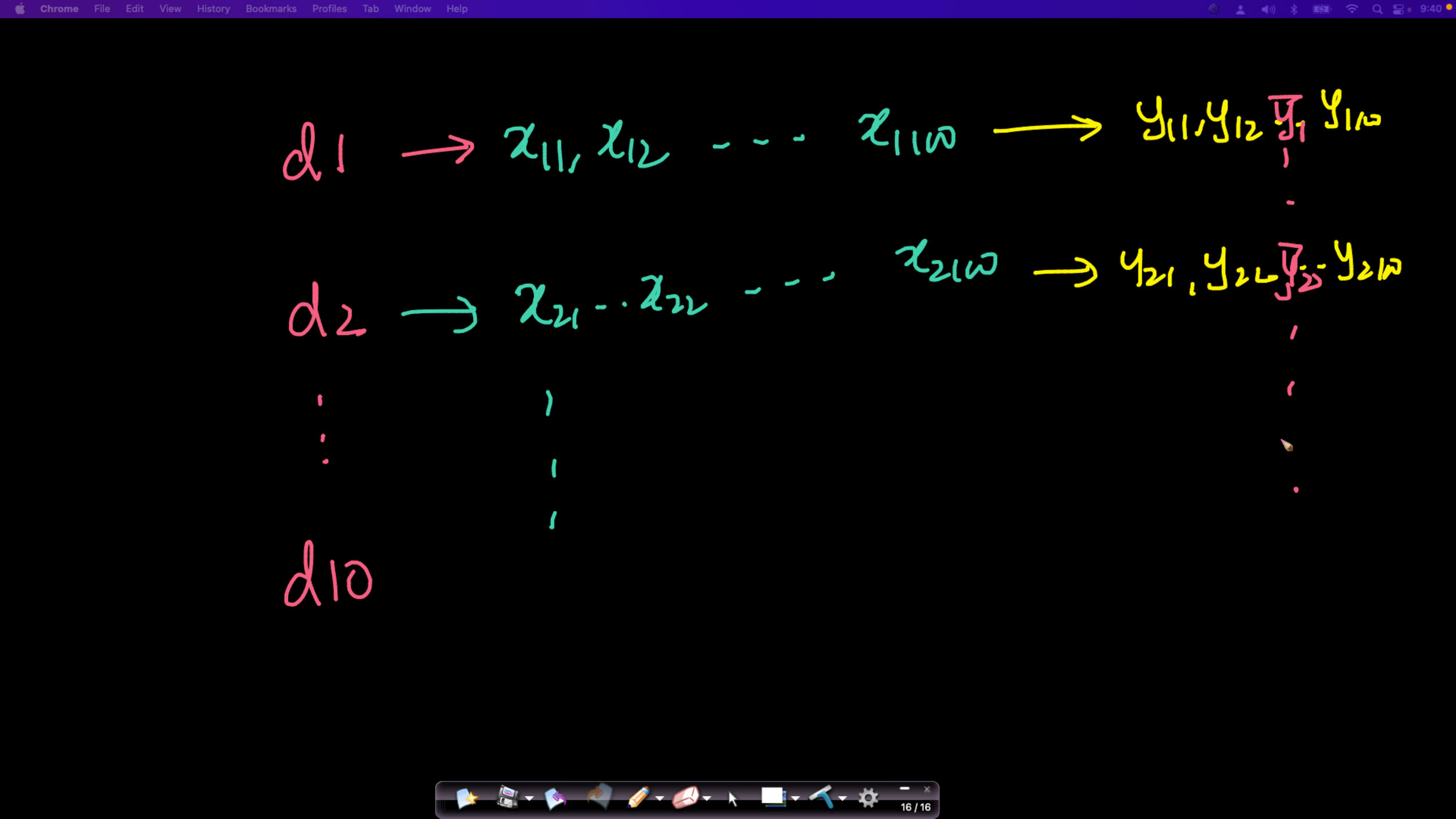
$\lambda \neq 0$

$$y_i = \frac{x_i^{\lambda} - 1}{\lambda}$$

$$f_\lambda(x_i) = y_i$$

$$\left[\exp\left\{ \frac{\log(\lambda y_i + 1)}{\lambda} \right\} \right] = x_i$$

$$f^{-1}(y_i) = x_i$$


d1 → x11, x12 - - - x1100 → y11, y12 ȳ1 y1100
d2 → x21 - - x22 - - - x2100 → y21, y22 ȳ22 y2100
d10

Chrome File Edit View History Bookmarks Profiles Tab Window Help

boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Kolmogorov–Smirnov test - Wi | KStest_Ttest.ipynb - Colaborat

colab.research.google.com/drive/1HBSL2rle-pokrz50QIhgWGuZiUaxAeUe#scrollTo=33F37cJcR0gr

+ Code + Text

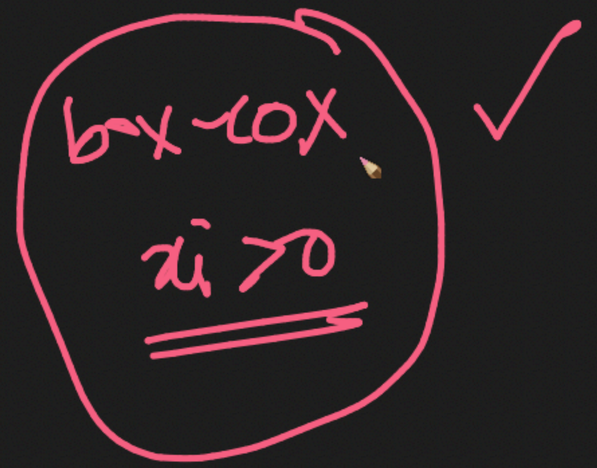
RAM
Disk

[1] 2s

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is depre
  import pandas.util.testing as tm
```

[ ]

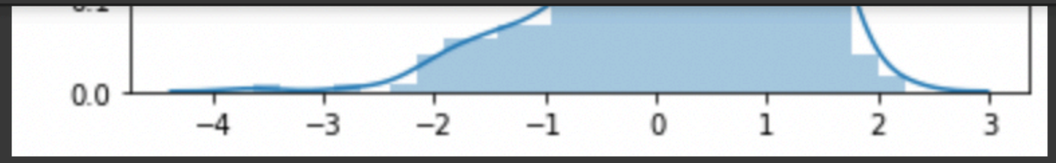
```
# gather some data
data =[ 0.04177737,  0.97977259,  1.19684675,  0.75969411,  0.2772351 ,
        1.20400739,  1.19512711, -1.33315966,  0.47241401,  0.58453053,
        0.21167461,  0.87106215, -0.56663286,  0.3702523 ,  0.72724427,
        0.41126015,  0.33358864,  0.72878097,  0.69929305,  0.72581333,
        1.67334826, -1.54572083, -1.22840893,  0.47103287,  0.895276  ,
        0.16538052, -0.43575904,  1.62784202,  0.98340417,  0.90482144,
       -0.47914975,  0.71812022,  1.14243   , -0.04393411,  1.24946471,
       -0.8699551 ,  1.60196517,  1.00140898,  1.48233878, -0.37088602,
       -0.0954339 ,  1.2969551 ,  0.0457524 , -0.06486335,  0.43257115,
       -0.18945797,  0.46525944,  0.12974487, -0.10501035,  0.94060547,
       -1.57714093,  0.24292938,  0.68759359,  0.24113398,  0.74353881,
        0.0129037 ,  0.47936105, -0.0596165 ,  0.3300311 , -0.19409805,
       -2.15213968, -0.9169724 ,  1.40476752,  0.74067023,  0.36119747,
        1.04507563, -0.54692221,  0.65000261,  0.5359208 ,  0.40091749,
        0.16959609,  0.43828974,  1.69191812, -0.40588725,  0.52772481,
        0.2410331 ,  1.8226663 , -1.36677194,  0.41745297,  0.94050797,
        1.15797033,  0.13883716,  0.9648131 ,  0.71495948,  1.73284151,
        0.9571359 ,  0.38785663,  0.41290029,  1.10291874,  0.41268708
```



17 / 17

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:44
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Kolmogorov–Smirnov test - Wi | KStest_Ttest.ipynb - Colaborat
colab.research.google.com/drive/1HBSL2rle-pokrz50QIhgWGuZiUaxAeUe#scrollTo=e3U8mXjRR7uL
Update
+ Code + Text
RAM
Disk

[4] 0s



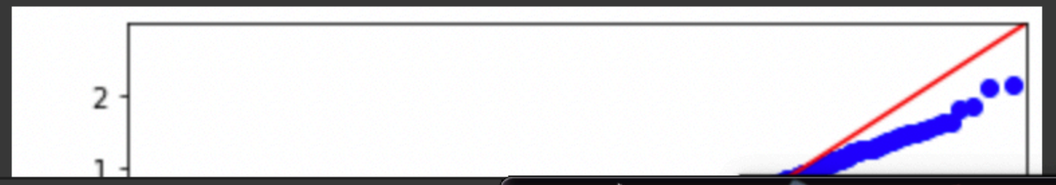

```
#Z-score
i=100
xi= x[i]
print(xi)

print ("zscore:", (xi-np.mean(x))/np.std(x))
```

```
-1.59860382
zscore: -2.341243992536608
```

$z_{100} = -1.598$

$Z\text{score}_{100} = -2.34$

```
# qq plot of x vs normal
sm.qqplot(x, stats.norm, fit=True, line="45")
```

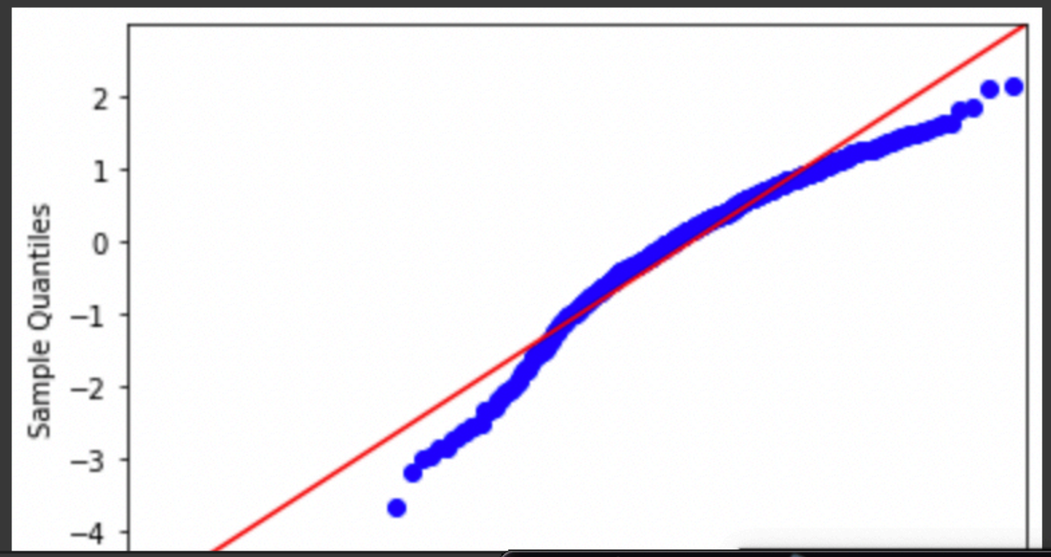


18 / 18

```
xi= x[i]
print(xi)

print ("zscore:", (xi-np.mean(x))/np.std(x))
```

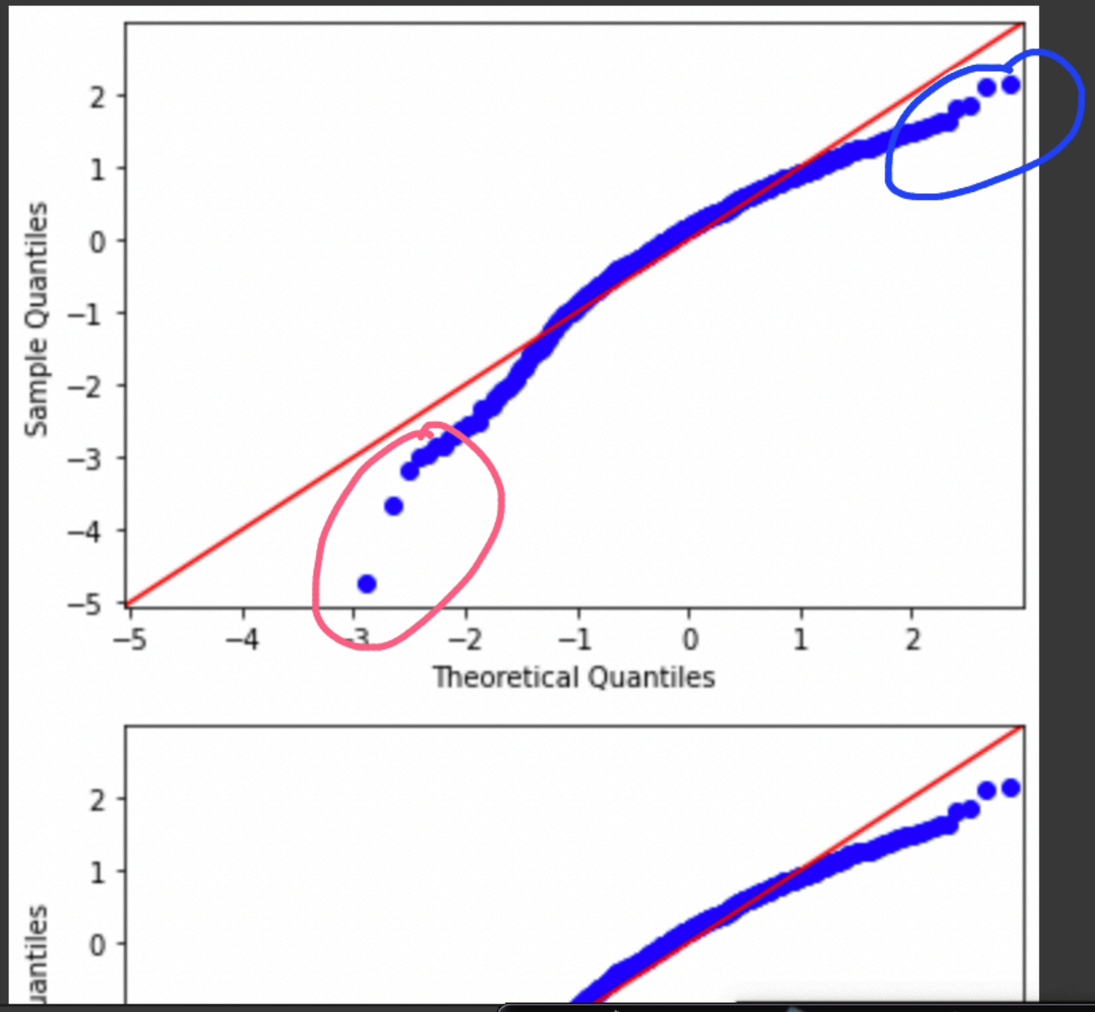
```
-1.59860382
zscore: -2.341243992536608
```

```
# qq plot of x vs normal
sm.qqplot(x, stats.norm, fit=True, line="45")
```



Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:45
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Kolmogorov–Smirnov test - Wi | KStest_Ttest.ipynb - Colaborat
colab.research.google.com/drive/1HBSL2rle-pokrz50QIhgWGuZiUaxAeUe#scrollTo=e3U8mXjRR7uL
Update
+ Code + Text
RAM
Disk

```
[ ]  # qq plot of x vs normal
     sm.qqplot(x, stats.norm, fit=True, line="45")
```



x ≠ Gaussian

20 / 20

Theoretical Quantiles

```
np.min(x)
```

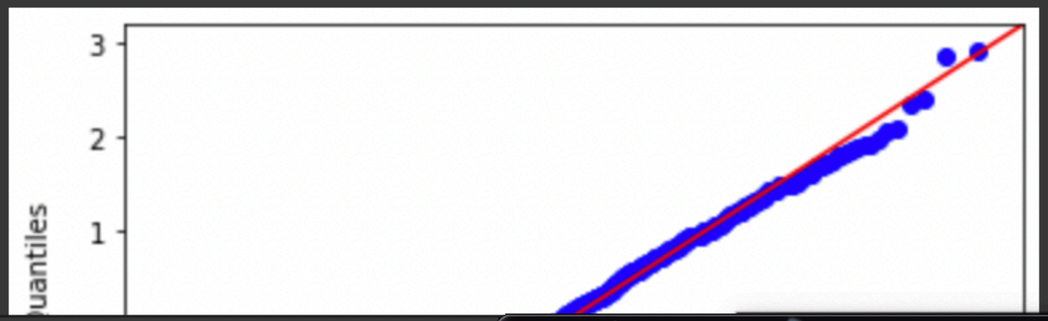
```
-3.63123097
```

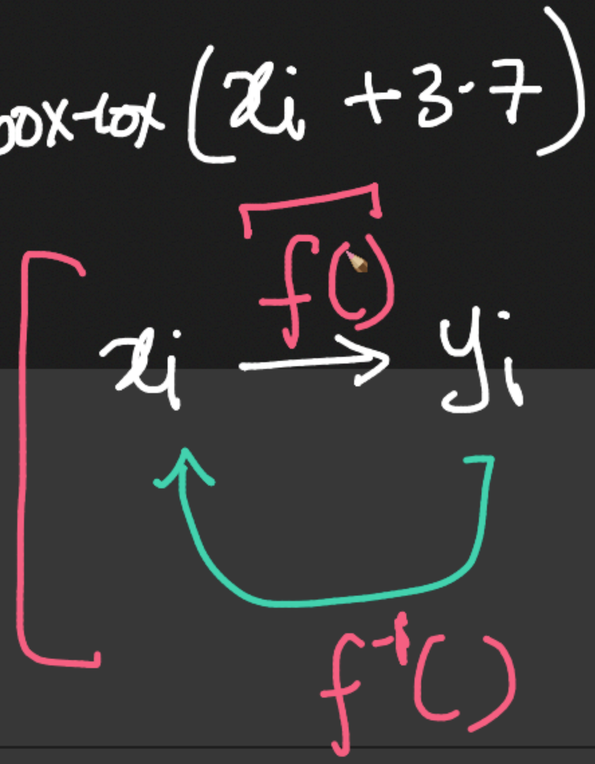
Min x = -3.63...

```
# box cox transform
x1 = x + 3.7
xt, l = stats.boxcox(x1,); # returns x_trasnfomred and lambda
print("lambda :" + str(l))

# check if xt is gaussian or not using QQ-Plot
sm.qqplot(xt, stats.norm, fit=True, line="45")
```

box-cox (xi + 3.7)

```
lambda :2.2233087629443724
```





Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:48


```
[ ] np.min(x)
```

```
-3.63123097
```

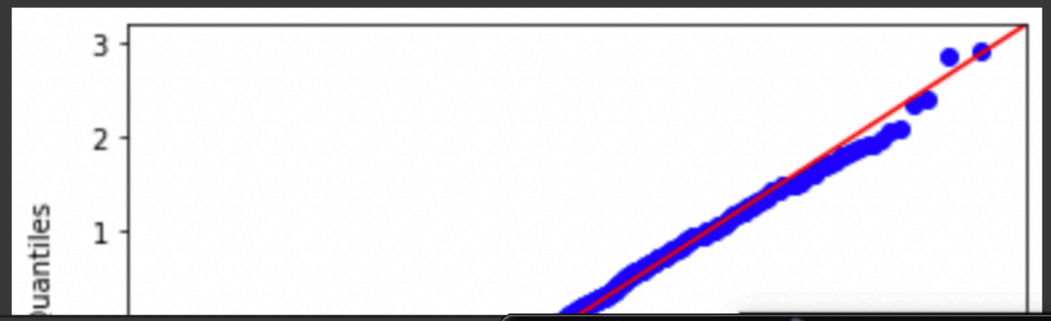
$x_i + 3.7 \quad \forall i = 1 \to n$

$\downarrow$ box-cox

$y_i \quad \forall i : 1 \to n \;\&\; l$

```
# box cox transform
x1 = x + 3.7
xt, l = stats.boxcox(x1,); # returns x_trasnfomred and lambda
print("lambda :" + str(l))

# check if xt is gaussian or not using QQ-Plot
sm.qqplot(xt, stats.norm, fit=True, line="45")
```

```
lambda :2.2233087629443724
```



$\lambda \neq 0 \qquad \lambda = 2.223$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:49

boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Kolmogorov–Smirnov test - Wi | KStest_Ttest.ipynb - Colaborat

colab.research.google.com/drive/1HBSL2rle-pokrz50QIhgWGuZiUaxAeUe#scrollTo=Lg17ejIFSxH2

Update

+ Code + Text

RAM Disk

Theoretical Quantiles

```
[ ] np.min(x)
```

```
-3.63123097
```

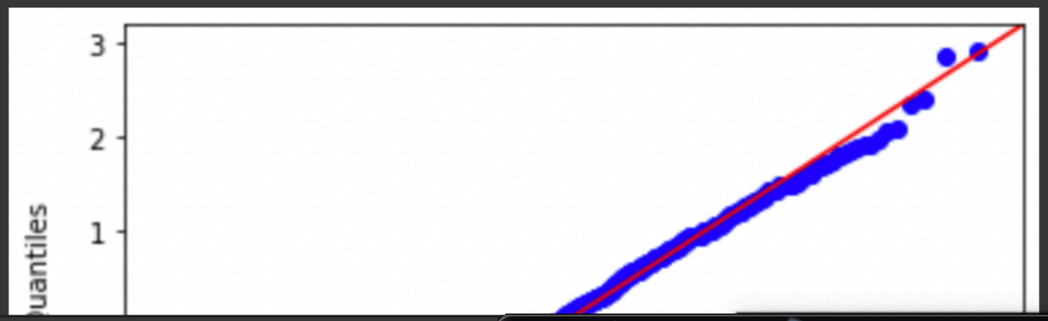
$$y_i = \frac{x_i^{\lambda} - 1}{\lambda} \quad \text{if } \lambda \neq 0$$

```
# box cox transform
x1 = x + 3.7
xt, l = stats.boxcox(x1,); # returns x_trasnfomred and lambda
print("lambda :" + str(l))

# check if xt is gaussian or not using QQ-Plot
sm.qqplot(xt, stats.norm, fit=True, line="45")
```

$= x_{t_i}$

$$y_i = \frac{(x_i + 3.7)^{2.223} - 1}{2.223}$$

```
lambda :2.2233087629443724
```



23 / 23

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:49

boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Kolmogorov–Smirnov test - Wi | KStest_Ttest.ipynb - Colabora

colab.research.google.com/drive/1HBSL2rle-pokrz50QIhgWGuZiUaxAeUe#scrollTo=Lg17ejIFSxH2

Update

+ Code + Text

RAM
Disk

Theoretical Quantiles

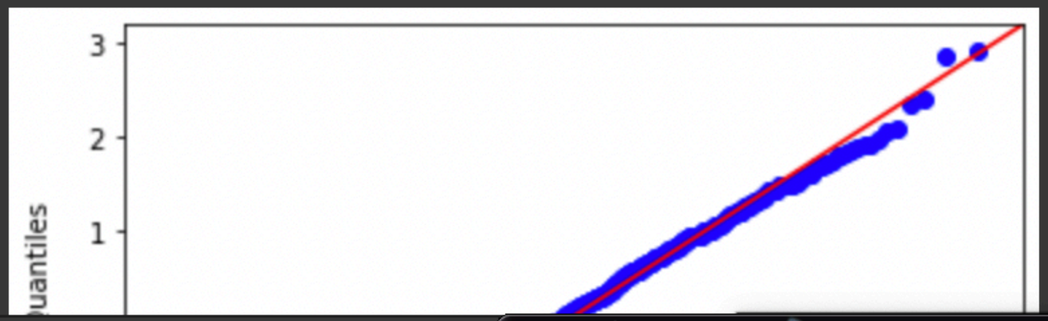
```
[ ]  np.min(x)
```

```
-3.63123097
```

```
# box cox transform
x1 = x + 3.7
xt, l = stats.boxcox(x1,); # returns x_trasnfomred and lambda
print("lambda :" + str(l))

# check if xt is gaussian or not using QQ-Plot
sm.qqplot(xt, stats.norm, fit=True, line="45")
```

box-cox tries its best

```
lambda :2.2233087629443724
```



24 / 24

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:50

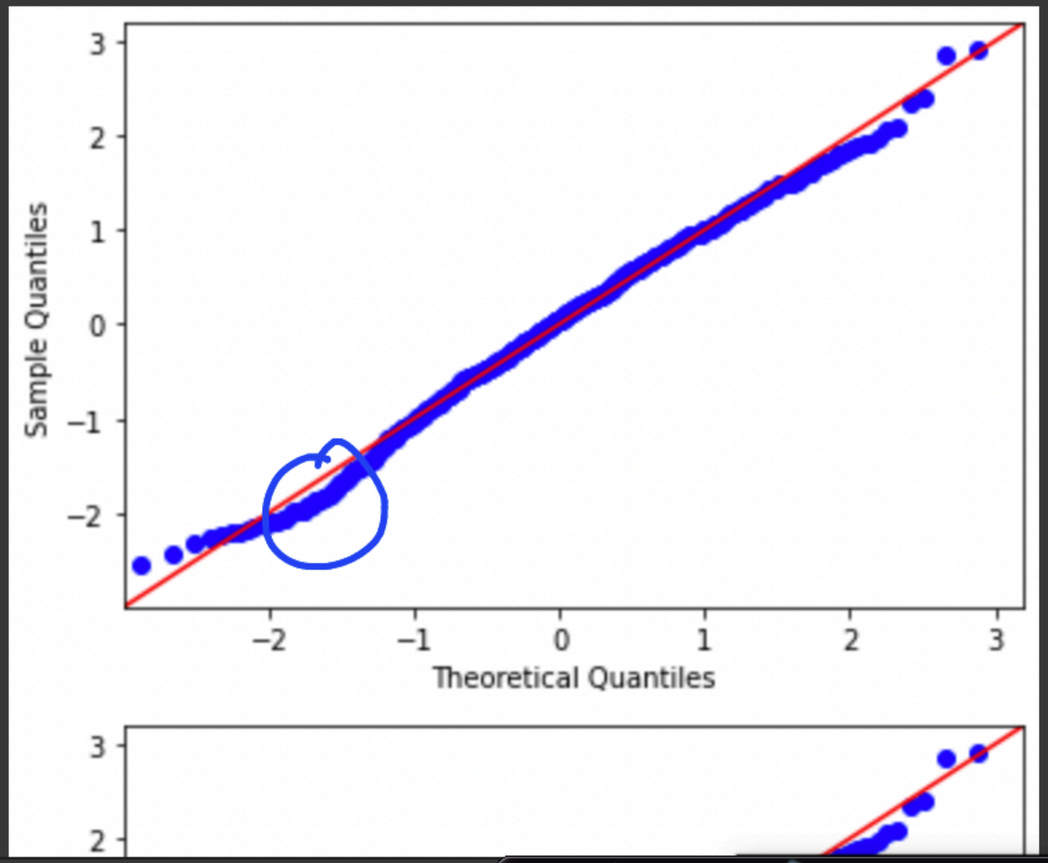
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Kolmogorov–Smirnov test - Wi | KStest_Ttest.ipynb - Colaborat

colab.research.google.com/drive/1HBSL2rle-pokrz50QIhgWGuZiUaxAeUe#scrollTo=Lg17ejIFSxH2 Update

+ Code + Text RAM Disk

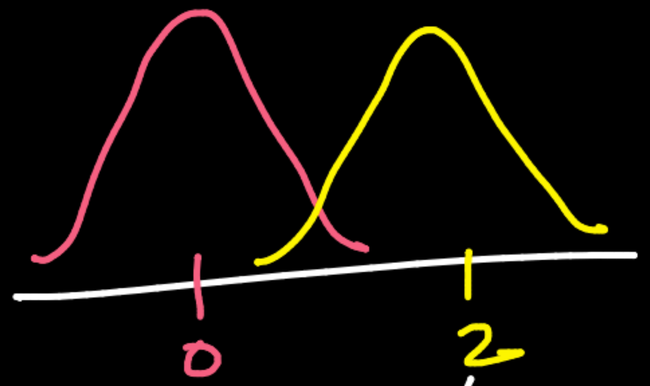
```
print("lambda :" + str(l))

# check if xt is gaussian or not using QQ-Plot
sm.qqplot(xt, stats.norm, fit=True, line="45")
```

lambda :2.2233087629443724



$x_t^i = y_i \sim$ Gaussian

$x_i + \alpha$

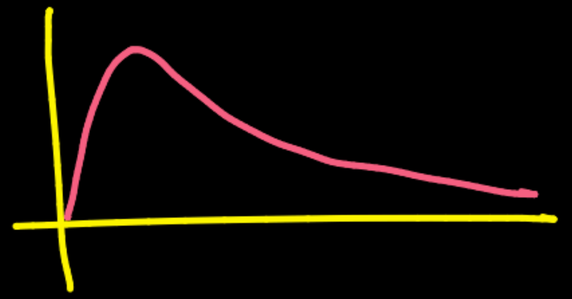


Shifting

$X \sim N(0, 1)$

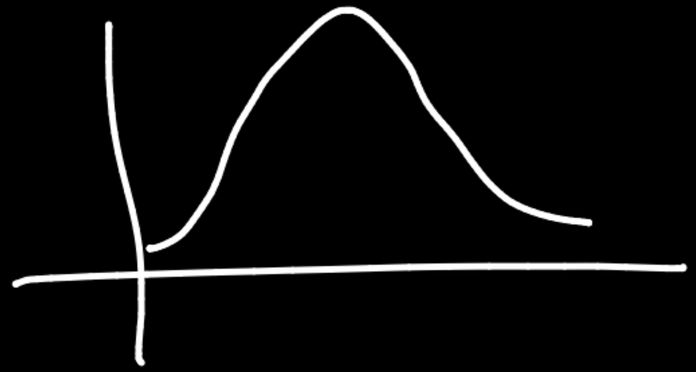
$X_i + 2 \sim N(2, 1)$

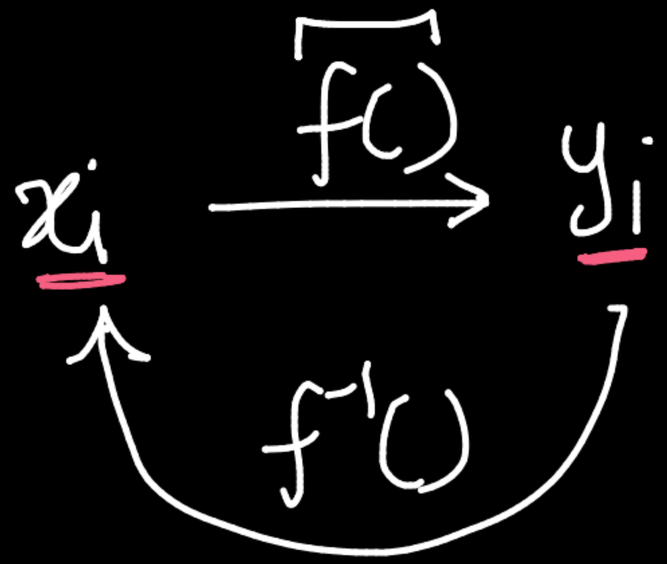
log-normal



log

Gaussian





no loss info

if f^{-1} doesnot exist → loss of info →...

box-cox

non-gaussian

↳ Gaussian

$$y_i = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \lambda = 0 \end{cases}$$

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

$|x_i|$: avoid it

x^2 : non-invertible

$$\begin{cases} x_1 = -2 \cdot 3 \\ x_2 = 2 \cdot 3 \end{cases}$$

box-cox $(x_i + 3.7)$

Compositions: $g(f(x))$

$\sqrt{\log(x_i + 3.7) + x_i^2}$

Stats ⟶ lots of the Math proofs... ✓

ANOVA

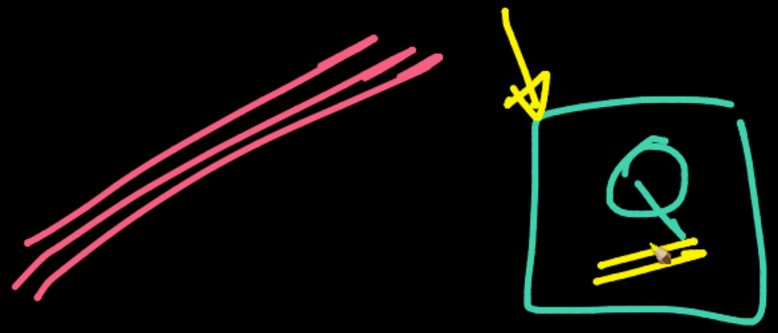
Z-test ⟶ population μ & σ are finite

Compare means

CLT ~ Means are Gaussian

ML: advanced: — GBDT / RF } — don't Care
DL

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:05



Q

Can I use CLT to convert non-gaussian data to Gaussian?

$M_j \sim$ Gaussian

CLT

$x_1, x_2 \ldots x_i \ldots x_n$
(non-gaussian)

f(·)

$y_1 \ldots y_i \ldots y_n$
(Gaussian)

can't

✓
- $M_1 \leftarrow S1$
- $M_2 \leftarrow S2$
- $M_K \quad S_K$

CLT is not a transformalism

$P(A|B)$ ✓

conditional Probs

$$P(A \mid \overline{B,C}) = P(A \mid B \cap C)$$

algebra

$$= \frac{P(A \cap B \cap C)/P(C)}{P(B \cap C)/P(C)} \quad \text{if } P(C) \neq 0$$

$$= \frac{P(A \cap B \mid C)}{P(B \mid C)} = \frac{P(A, B \mid C)}{P(B \mid C)}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:15

Bayes!

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$$\left.\begin{array}{l} P(B) \neq 0 \\ P(A) \neq 0 \end{array}\right\}$$

Extended Bayes!

$$\frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Question:

$$P(A \mid B, C) = \frac{P(A, B, C)}{P(B, C)}$$

$A, B = A \cap B$

Conditional Prob & ∩

$$= \frac{P(B \mid A, C)\, P(A, C)}{P(B \mid C)\, P(C)}$$

$$= \frac{P(B \mid A, C)\, P(A \mid C)\, P(C)}{P(B \mid C)\, P(C)}$$

$A, B, C \equiv A \cap B \cap C$

$$p(A, B, C) = p(B, \underbrace{A, C}) = p(B \mid \underbrace{A, C})\, p(A, C)$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:17

10:22

✓ [ks-test
t-test
A/B-testing
z-poop]

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:25

Z-Test:

2-sample

Med1: x_{11} - - - $x_{1,100}$ ← n_1

$m_1 = \bar{x}_1$

Med2: x_{21} - - - $x_{2,120}$ ← n_2

$m_2 = \bar{x}_2$

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim Z(0,1)$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:28

1-sample

batch

$H: h_1 \dots h_{100} = n$

$\bar{h}$ = mean-height

$$\begin{cases} \mu = 158\,cm \\ \sigma = 20\,cm \end{cases}$$

$$T = \frac{\bar{h} - \mu}{\sigma/\sqrt{n}} \sim Z(0,1)$$

σ → Sample-std-dev $\sim \sigma$ if n is large

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

$X: x_1, x_2 \cdots x_n$

g_s X Gaussian → Normality testing: AD

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:42

Framework:

Data (1)

KS-test

Data:

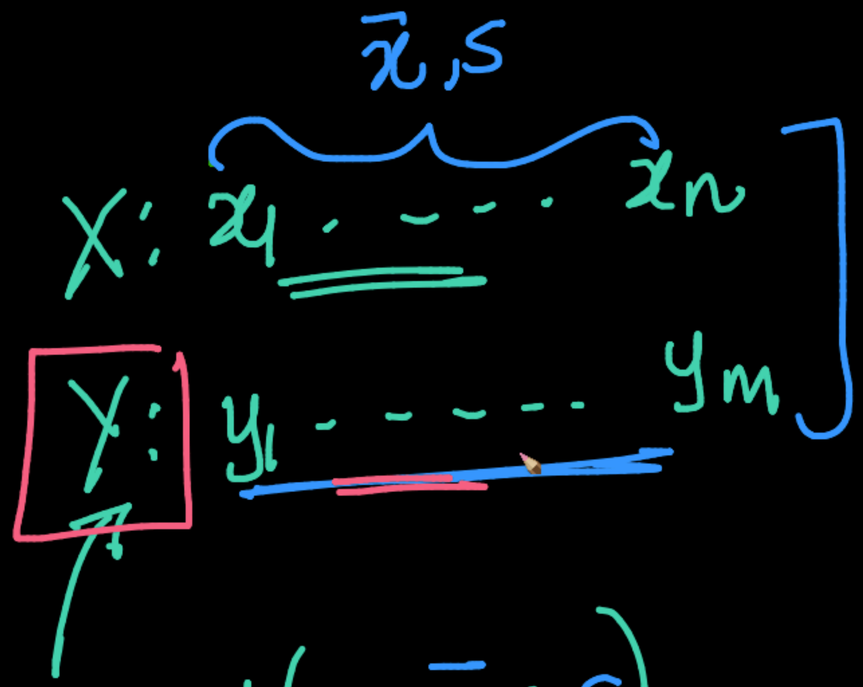
$$\begin{cases} X: & x_1, x_2 \cdots x_i \cdots x_n \\ Y: & y_1, y_2 \cdots \cdots y_m \end{cases}$$

[2-sample]

n & m are not too small

Task: disb of X same as disb of Y

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

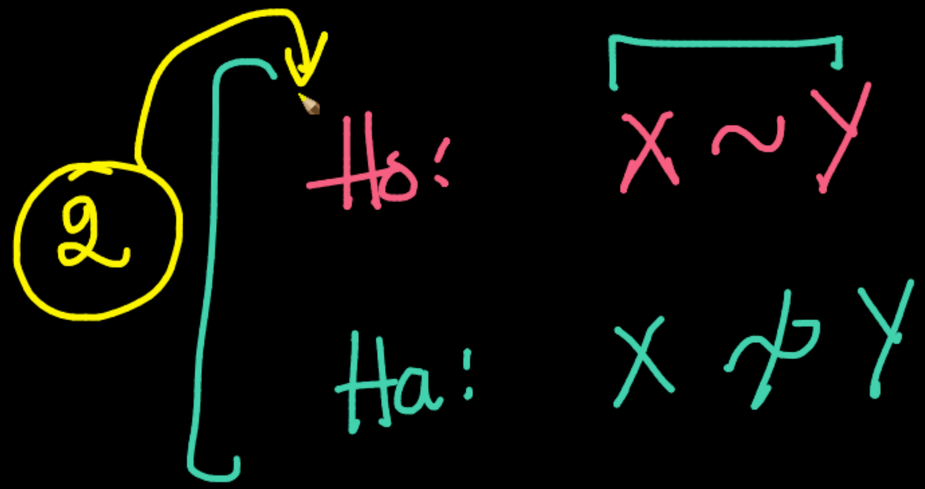


Gs X Gaussian

State-rvs

Normal($\mu = \bar{x}, \sigma = s$)

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:42



have same dist

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:47

4 Sensible

T_{KS} = Maximal Supremum | gap b/w CDFs

abs

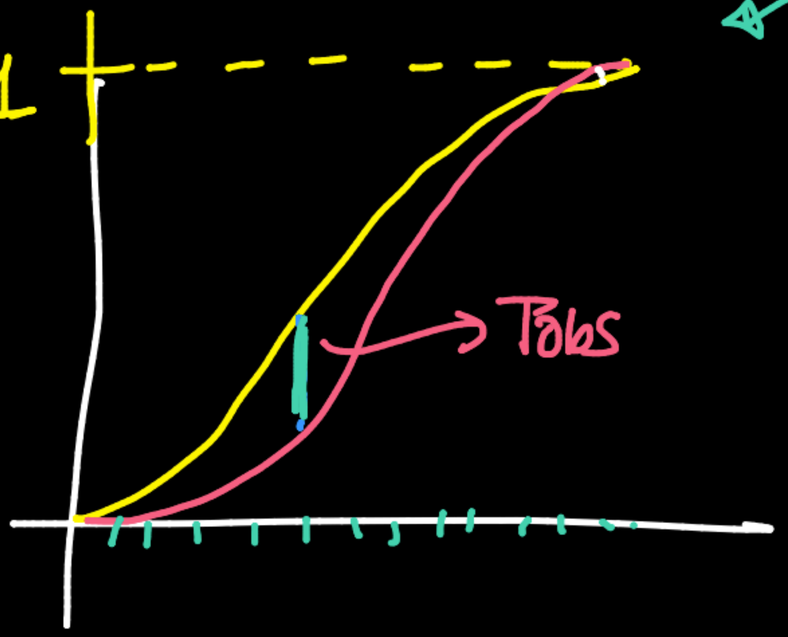
→ QQ-plot

→ Compare CDFs

3

Under H_0: $T_{KS} \to 0$ ✓

CDF are close to each other



Under H_a: T_{KS}: large +ve val

T_{KS} under H_0

$\hookrightarrow K_0$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:36
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Kolmogorov–Smirnov test - Wi | KStest_Ttest.ipynb - Colabora | Normality test - Wikipedia
en.wikipedia.org/wiki/Kolmogorov–Smirnov_test#Kolmogorov_distribution
Update

Kolmogorov distribution [edit]

The Kolmogorov distribution is the distribution of the random variable

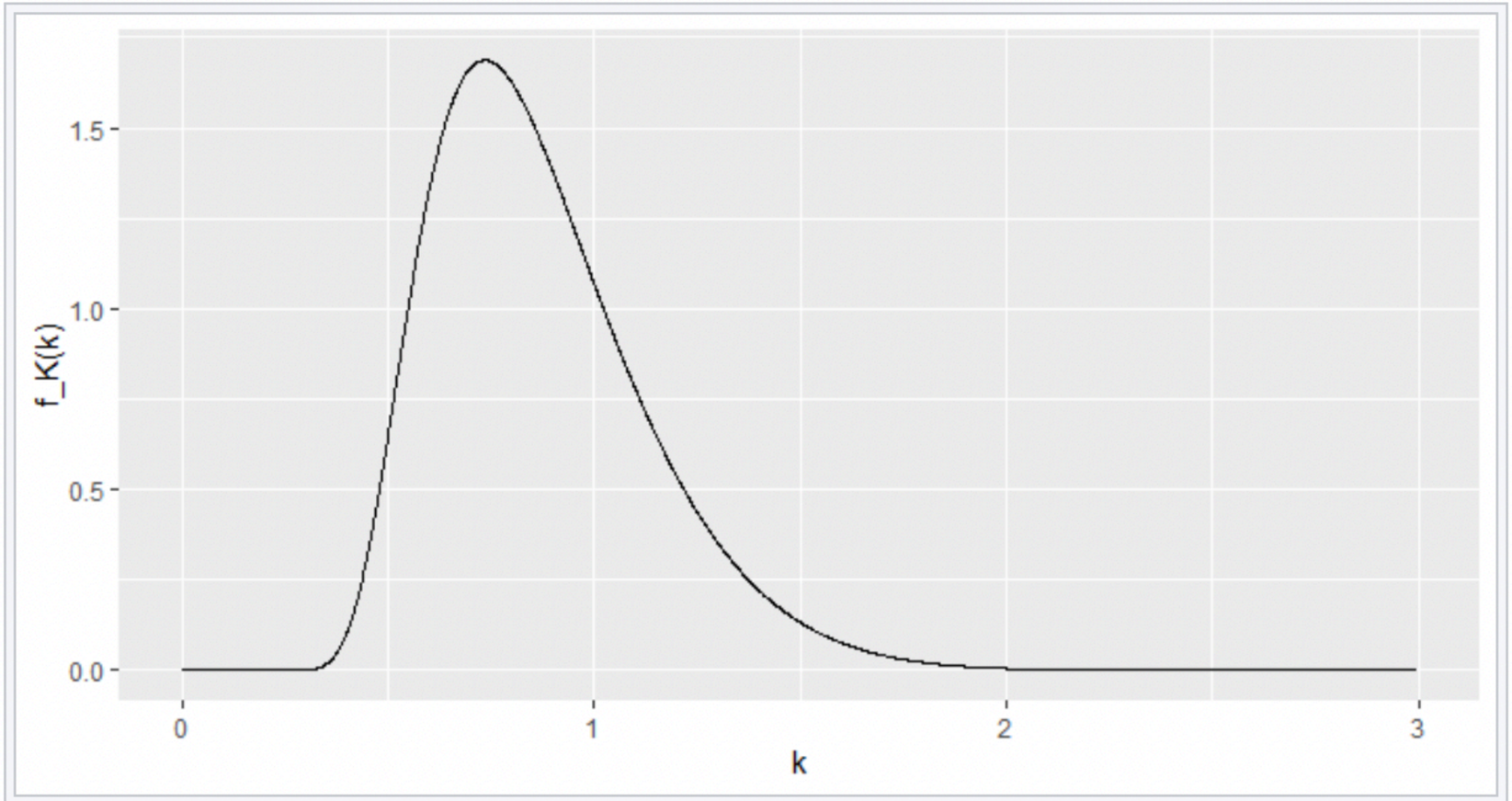


Illustration of the Kolmogorov distribution's PDF.

$$K = \sup_{t \in [0,1]} |B(t)|$$

where $B(t)$ is the Brownian bridge. The cumulative distribution function of K is given by[3]

49 / 49

such as the Anderson–Darling test statistic) to properly reject the null hypothesis.

Kolmogorov distribution [edit]

The Kolmogorov distribution is the distribution of the random variable

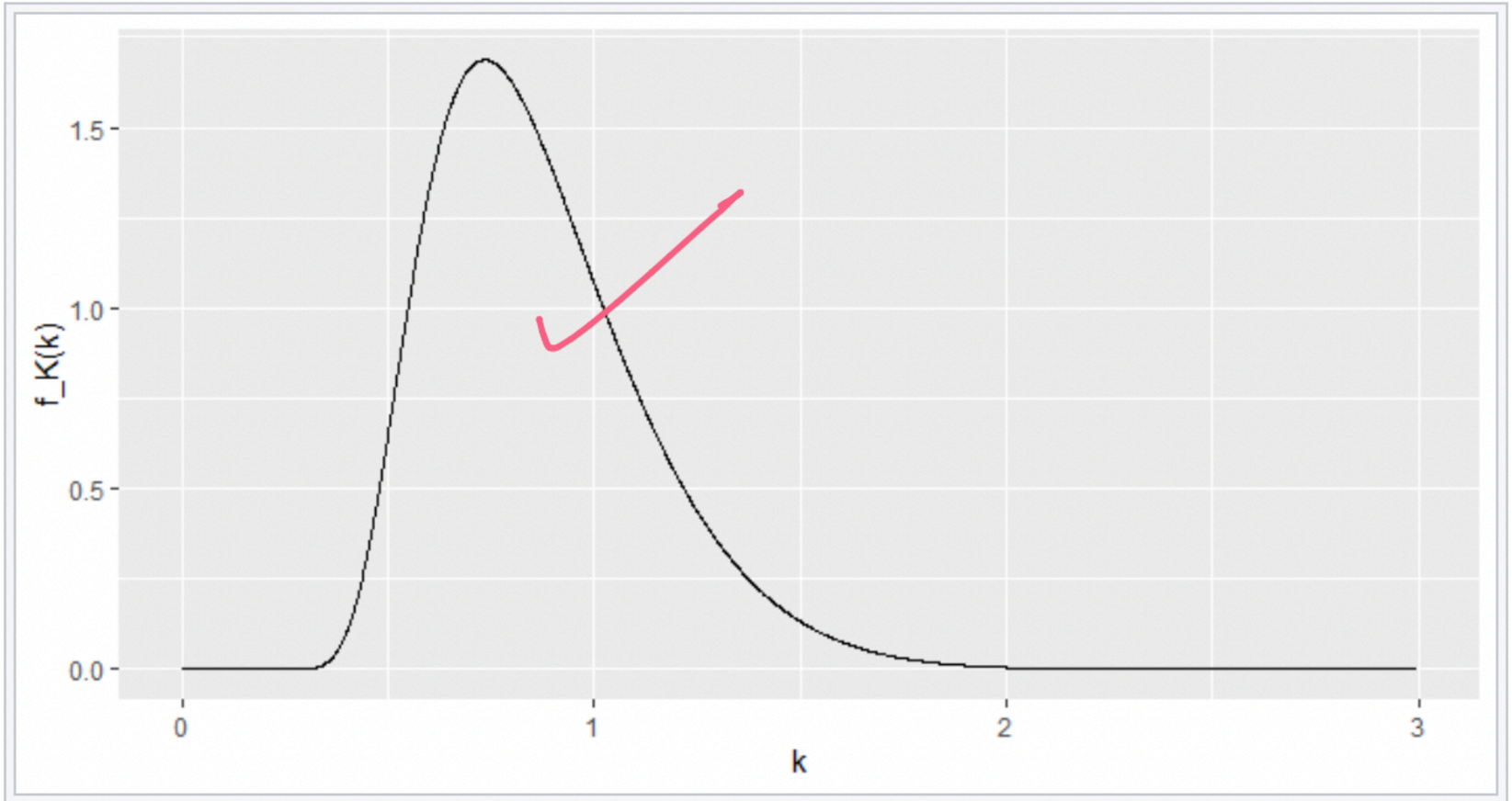


Illustration of the Kolmogorov distribution's PDF.

$$K = \sup_{t \in [0,1]} |B(t)|$$

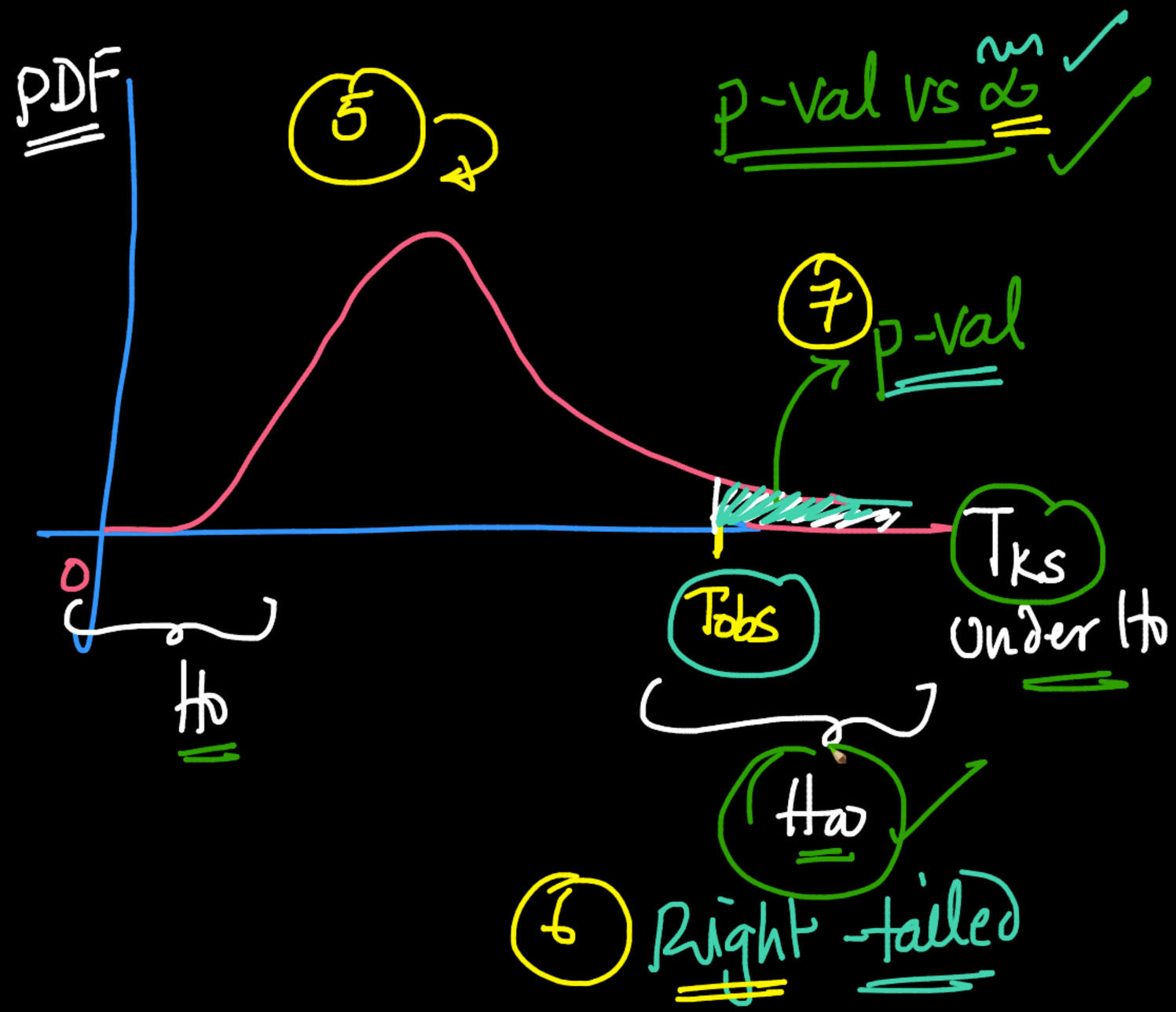
Physics

Brownian Motion ✓

KS-test

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:48

$T_{obs} =$



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

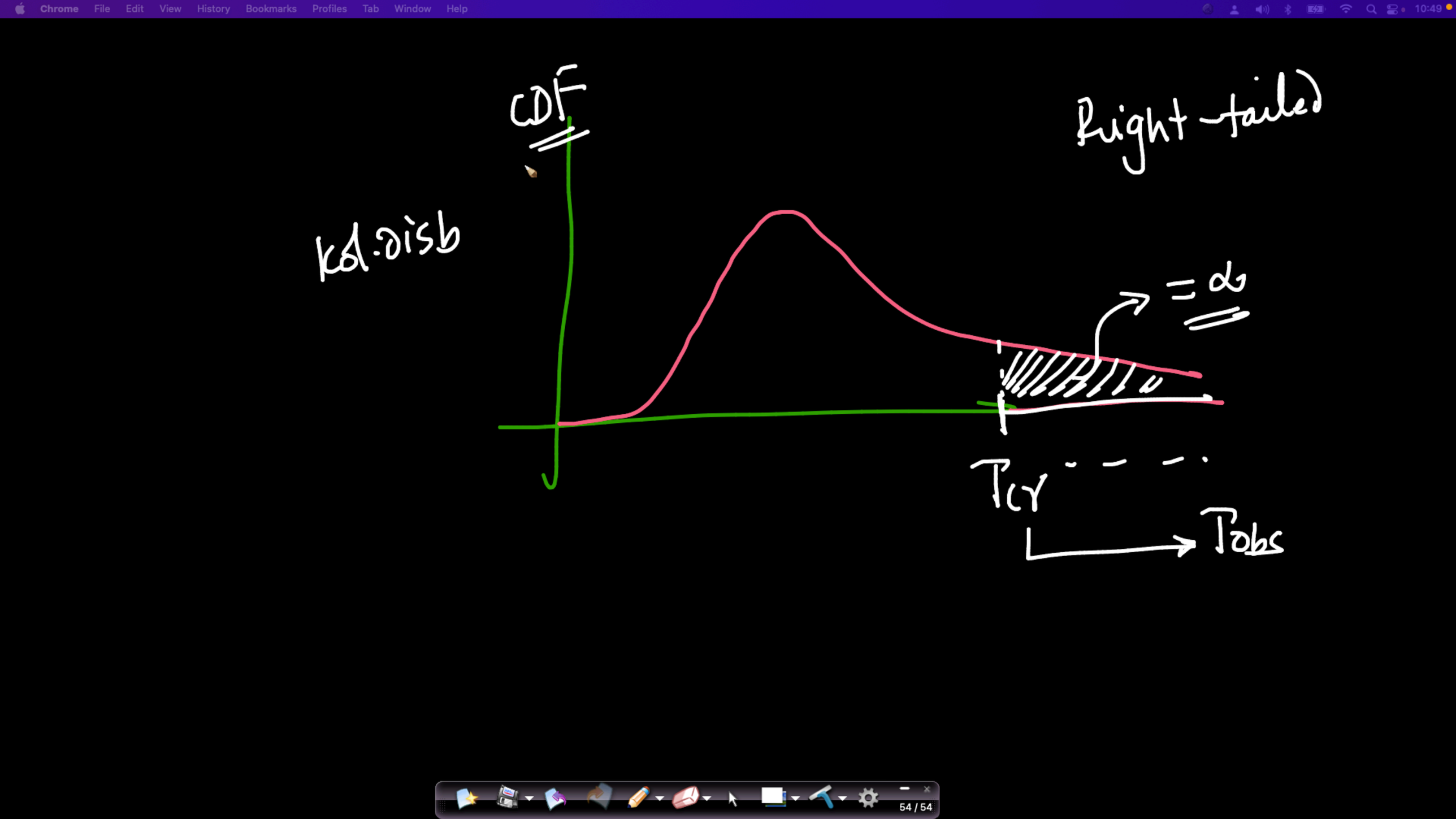
KS-test → parametric → some assumptions about pop
KS-test → non-param

$T_{KS} \sim$ Kol-Dist

{ any assumption about the pop. dist of X or Y

Z-test :- pop finite μ & σ → param

$T_z \sim Z(0,1)$ X


CDF
Right-tailed
kd.disb
= α
Tcr
Tobs

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:51

boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colabora | Normality test - Wikipedia

colab.research.google.com/drive/1lWuS8AxULBBoId2GlfBKUeT1RSj3fFAJ#scrollTo=MmvmI3kRVi0d

Update

+ Code + Text

RAM Disk

KS-Test

Z-test: Compare means

Med1: } Compare -2 sb

Med2: }

[1] 1s

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

0s

```
# recovery times of patients who took medicine-1
r1 = [8.82420842, 7.47774471, 7.55712098, 7.98131439, 6.82771606,
      7.48566433, 9.15385732, 5.84040502, 8.26124313, 8.4728876 ,
      6.82582186, 7.00490974, 8.43423058, 6.72099932, 6.97495982,
      5.93748053, 5.40707847, 6.16385557, 6.71421056, 4.42396183,
      6.87285228, 8.00313581, 6.69035041, 7.83622942, 8.70984957,
      5.56284584, 9.08093437, 4.98165193, 7.67769408, 6.04738478,
      7.64921582, 7.31051639, 6.74463303, 7.27356973, 8.16787232,
      6.90990965, 7.06439167, 6.62921957, 6.08283539, 6.2458137 ,
      8.65173634, 5.76080646, 6.20573219, 8.91561004, 6.22560201,
      5.67542104, 6.97412435, 8.31354697, 8.14172701, 8.26099345,
      7.87612791, 6.24835109, 9.95324783, 6.59504627, 6.17365145,
      6.05676895, 7.23030223, 7.71311809, 7.37163804, 5.69798738,
```

55 / 55

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:52

boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colabora | Normality test - Wikipedia

colab.research.google.com/drive/1lWuS8AxULBBoId2GlfBKUeT1RSj3fFAJ#scrollTo=wVXdQQi1WSwT

Update

+ Code + Text

RAM Disk

[6] n1

0s

```
100
```

med1 ; - - -

med2 ; - -

[7] n2

0s

```
120
```

```
#2-sample KS Test
stats.ks_2samp(d1, d2)
```

```
Ks_2sampResult(statistic=0.3233333333333333, pvalue=1.516338798324135e-05)
```

p-value = 0.00001516 < 0.01 = alpha

=> Reject H0 (r1 and r2 same distribution)

=> Accept Ha (the distributions of r1 and r2 are different)

```
plt.grid()
a = plt.hist(d1, bins=100, cumulative=True, label='CDF', density=True, histtype='step')
```

56 / 56

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

boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colabora | Normality test - Wikipedia

colab.research.google.com/drive/1lWuS8AxULBBoId2GlfBKUeT1RSj3fFAJ#scrollTo=wVXdQQi1WSwT

Update

+ Code + Text

RAM Disk

0.01

α = 1%

α = 0.1%

0.001

[6] n1

0s

100

[7] n2

0s

120

```
#2-sample KS Test
stats.ks_2samp(d1, d2)
```

Tks

1.5X10^-5

```
Ks_2sampResult(statistic=0.3233333333333333, pvalue=1.5163387983241 e-05)
```

p-value = 0.00001516 < 0.01 = alpha

=> Reject H0 (r1 and r2 same distribution) ✓

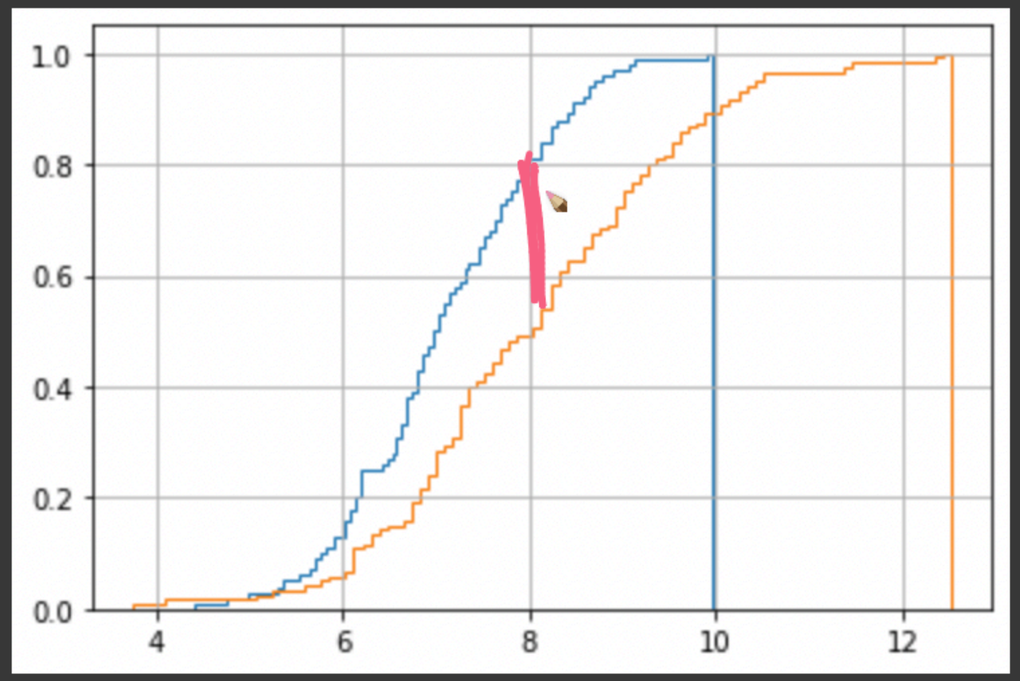
=> Accept Ha (the distributions of r1 and r2 are different)

p-val < 0.001

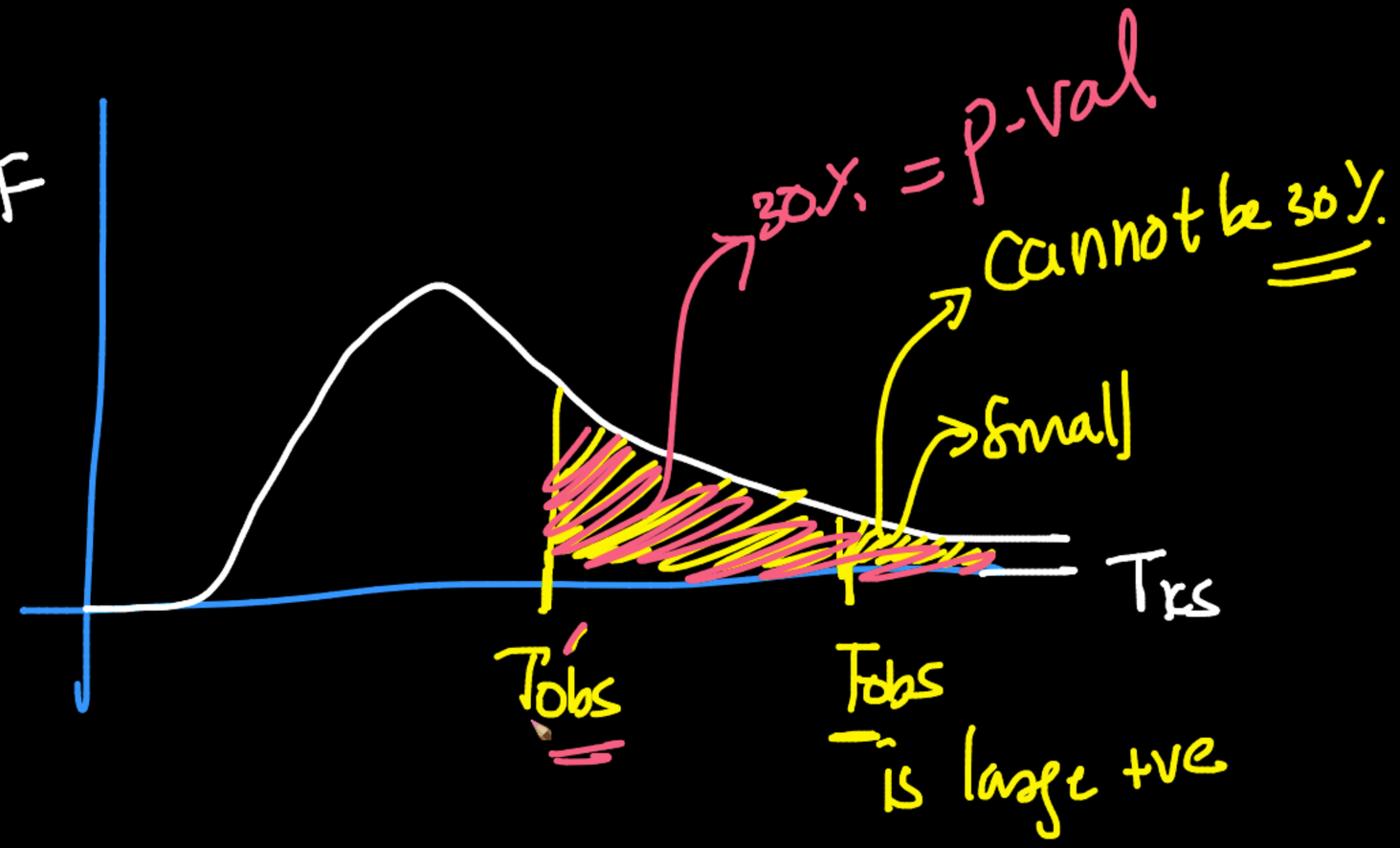
```
[ ] plt.grid()
    a = plt.hist(d1, bins=100, cumulative=True, label='CDF', density=True, histtype='step')
```

57 / 57

```
plt.grid()
a = plt.hist(d1, bins=100, cumulative=True, label='CDF', density=True, histtype='step')
b = plt.hist(d2, bins=100, cumulative=True, label='CDF', density=True, histtype='step')
plt.show()
```



Double-click (or enter) to edit



$\alpha = 0.1\%$

p-val > α

$\Rightarrow$ cannot Ho

$\Rightarrow$ accept Ho

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:59
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colaborat | Kolmogorov–Smirnov test - Wi
en.wikipedia.org/wiki/Kolmogorov–Smirnov_test#Kolmogorov_distribution
Update

Kolmogorov distribution [edit]

The Kolmogorov distribution is the distribution of the random variable

Normal(0,1)

(let)

✓ $x_1 \cdots x_{100}$

✓ $y_1 \cdots y_{100}$

samples

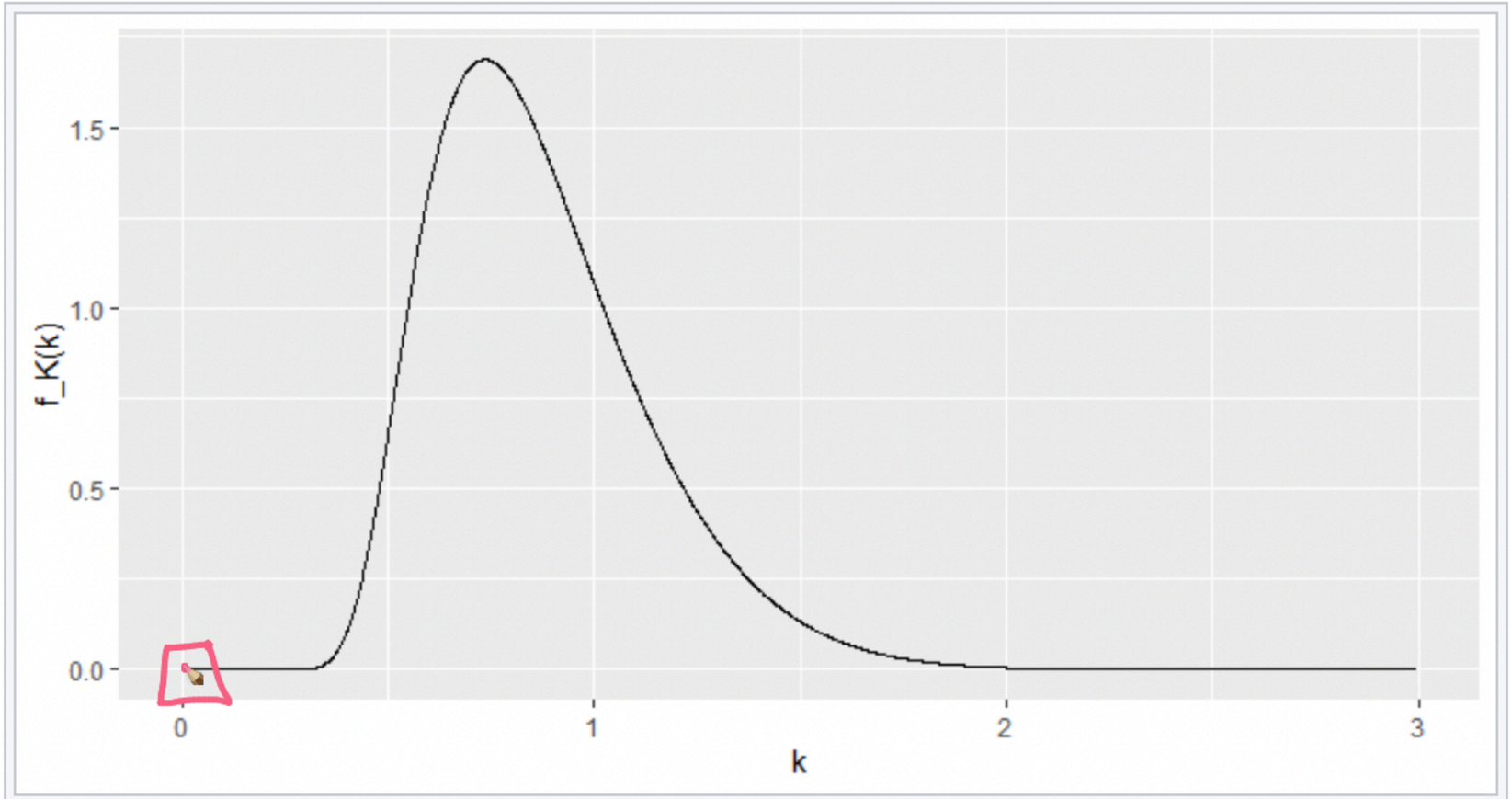


Illustration of the Kolmogorov distribution's PDF.

$$K = \sup_{t \in [0,1]} |B(t)|$$

where $B(t)$ is the Brownian bridge. The cumulative distribution function of K is given by[3]

60 / 60

Chrome File Edit View History Bookmarks Profiles Tab Window Help
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colaborat | Kolmogorov–Smirnov test - Wi
en.wikipedia.org/wiki/Kolmogorov–Smirnov_test#Kolmogorov_distribution
Update

Kolmogorov distribution [edit]

The Kolmogorov distribution is the distribution of the random variable

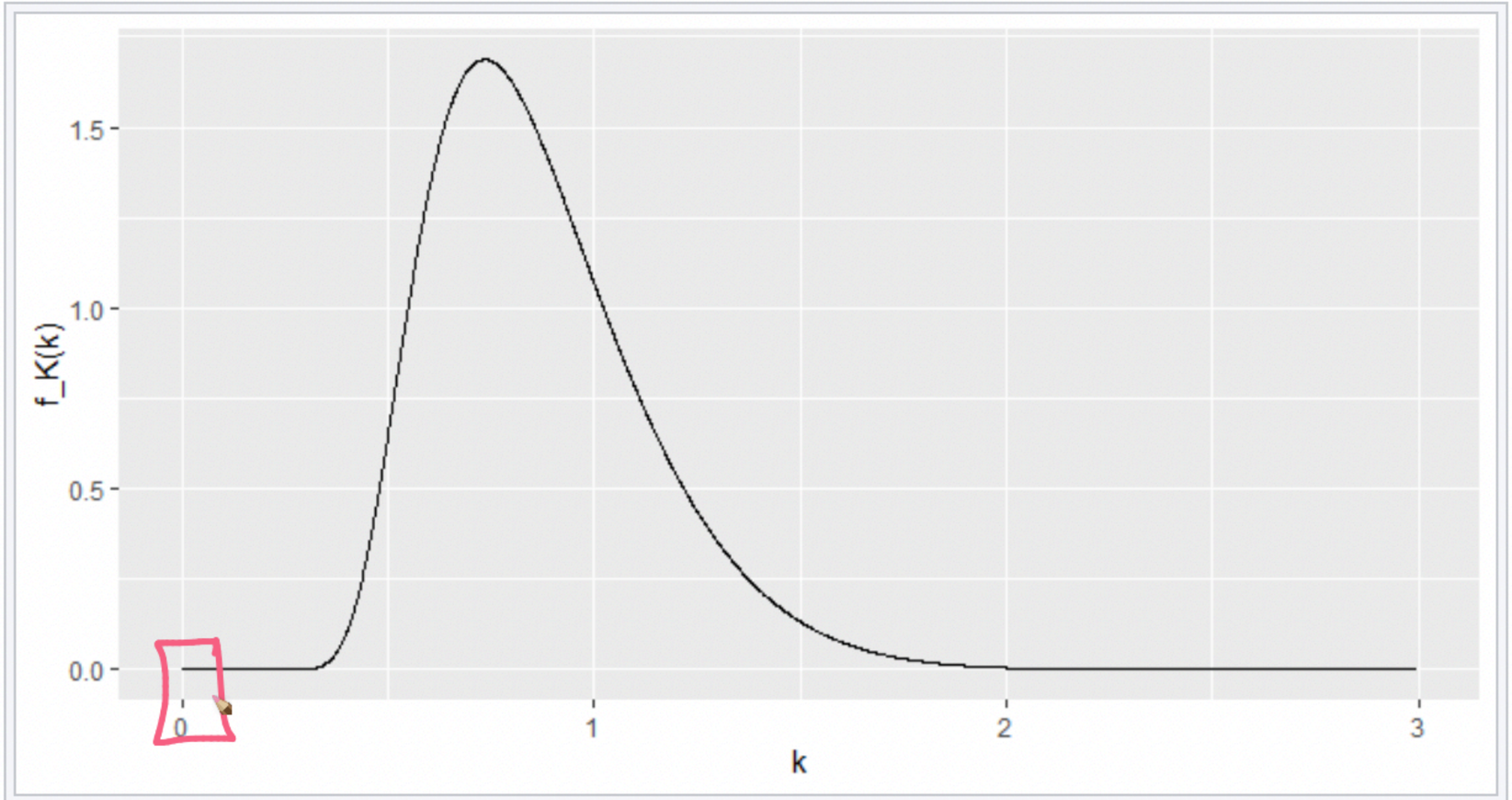
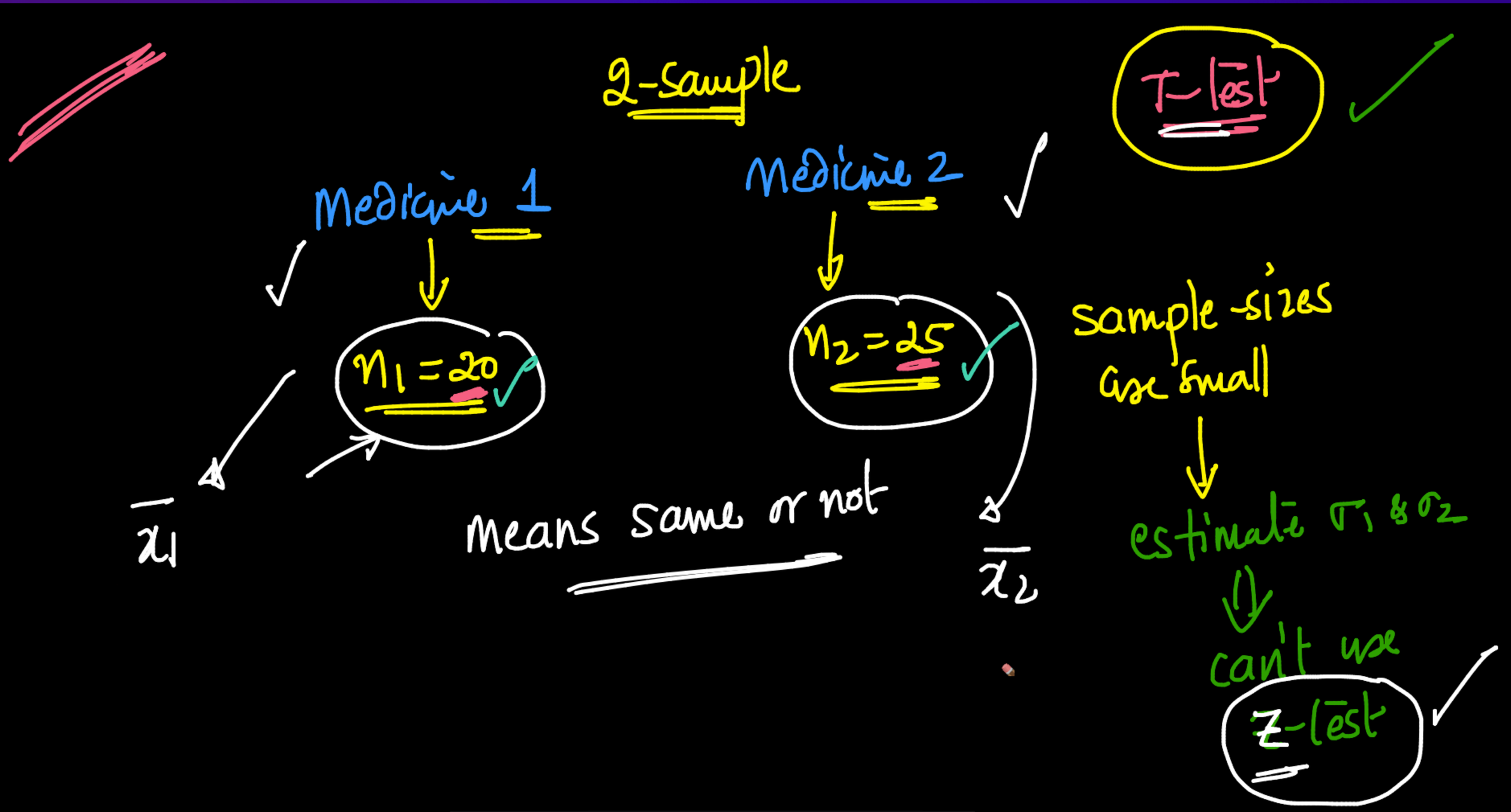


Illustration of the Kolmogorov distribution's PDF.

$$K = \sup_{t \in [0,1]} |B(t)|$$

where $B(t)$ is the Brownian bridge. The cumulative distribution function of K is given by[3]

61 / 61


2-sample
T-test
Medicine 1
Medicine 2
$n_1 = 20$
$n_2 = 25$
$\bar{x}_1$
$\bar{x}_2$
Means same or not
sample sizes are small
estimate σ_1 & σ_2
can't use
Z-test

Framework:

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

[real-world: n_1 & n_2 are small

↳ extreme medical expts ...] → surveys

→ high-cost ← physics

Assumptions:

T-test

1. sample-means $\bar{x}_1$ & $\bar{x}_2$ → Gaussian distb (proof)
   - ↳ CLT ⇒ population μ & σ are finite
2. obs are random and indep..
3. σ_1 & σ_2 are not known/estimatable because n_1 & n_2 are small

n_1 & n_2 < 30 ⇐

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:15

T-dist

$T(\,)$ ← $T_t =$

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Sensible

s_1^2 & s_2^2 are sample sta-dev ✓

$$T_z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$Z(0,1)$ ←

σ_1^2, σ_2^2 → pop. sta-dev → given ✓

↘ estimate them well

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

$$T_t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

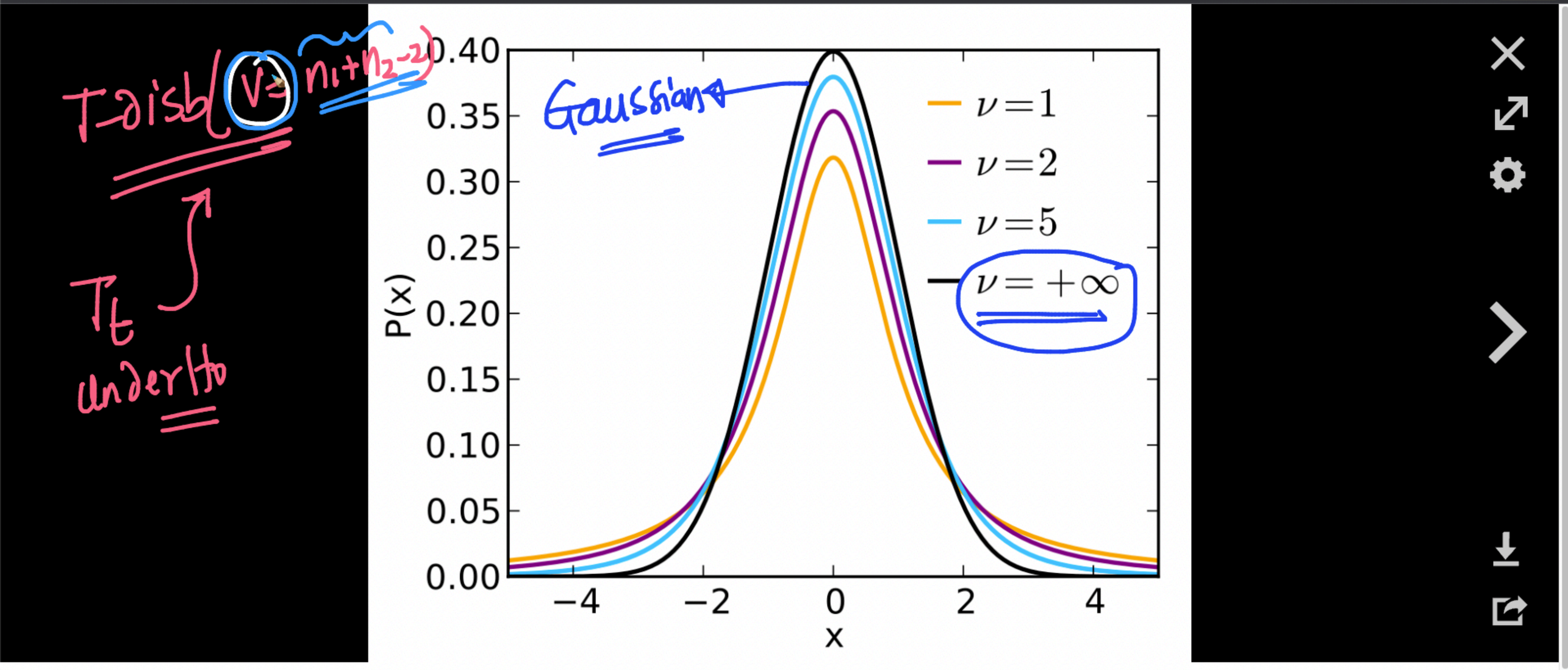
$\sim T\text{-dist}(v)$ — DoF

$v = n_1 + n_2 - 2$

(Pooled)

66 / 66

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:18
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colaborat | Student t pdf - Student's t-dist
en.wikipedia.org/wiki/Student%27s_t-distribution#/media/File:Student_t_pdf.svg
Update



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

More details
Skbkekas - Own work
67 / 67
CC BY 3.0

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

✓ $\nu = (n_1 + n_2 - 2)$ ↑ $\Rightarrow$ larger-sample sizes

$T_t \rightarrow$ Gaussian

$s_1^2 \approx \sigma_1^2$

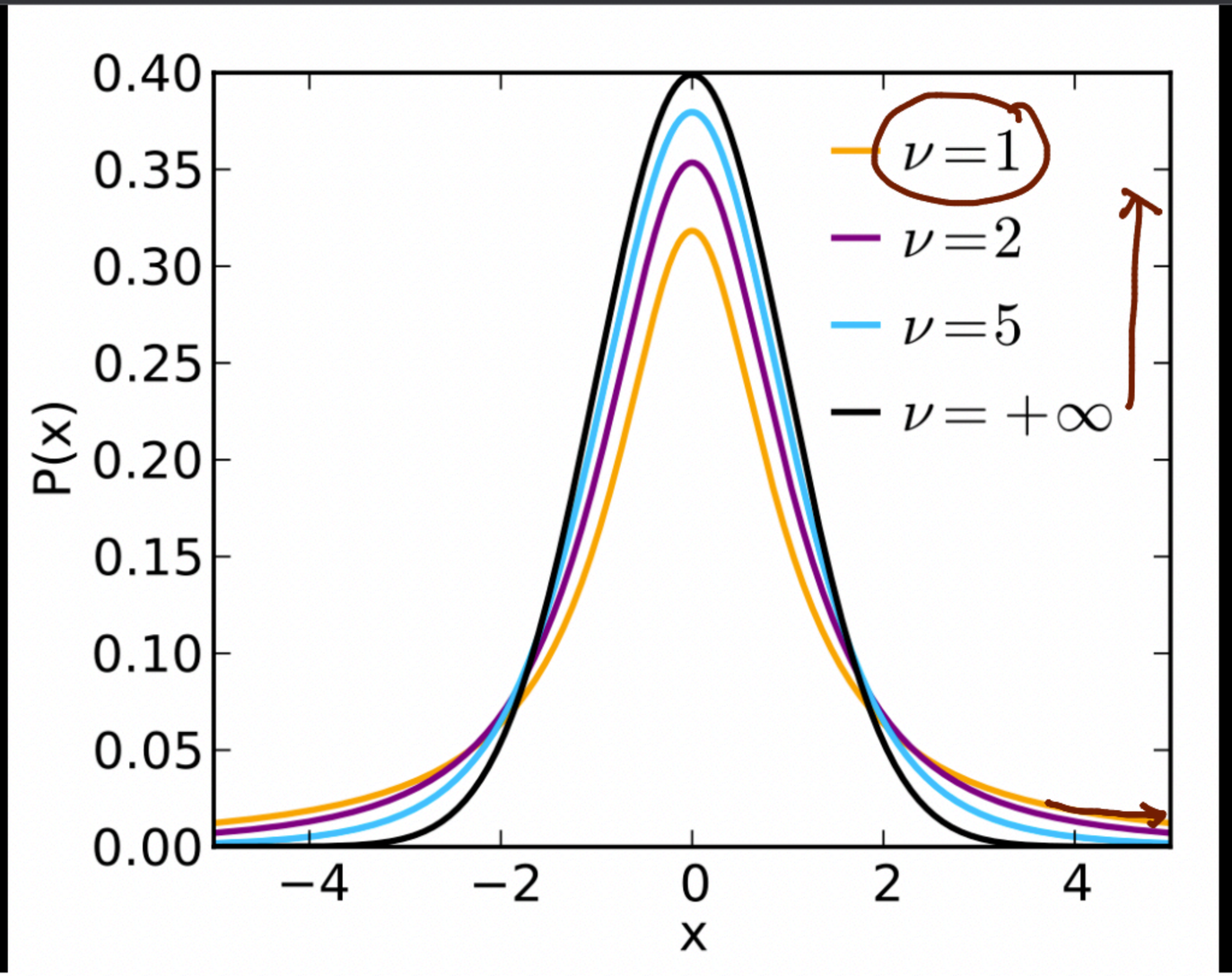
$s_2^2 \approx \sigma_2^2$

$T_t \approx T_z \rightarrow$ Gaussian

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:20
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colabora | Student t pdf - Student's t-dist
en.wikipedia.org/wiki/Student%27s_t-distribution#/media/File:Student_t_pdf.svg
Update

20 to
$v = n_1 + n_2 - 2$

$v = 28$



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

More details

Skbkekas - Own work

69 / 69

CC BY 3.0

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:20
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colaborat | Student's t-distribution - Wikip
en.wikipedia.org/wiki/Student%27s_t-distribution
Update

In other projects

Wikimedia Commons

Languages

العربية
Deutsch
Español
Français
Jawa
日本語
Português
Русский
中文

23 more

Edit links

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]

1 History and etymology
2 How Student's distribution arises from sampling
3 Definition
3.1 Probability density function
3.2 Cumulative distribution function
3.3 Special cases
4 How the *t*-distribution arises
4.1 Sampling distribution
4.2 Bayesian inference
5 Characterization
5.1 As the distribution of a test statistic
5.1.1 Deriva
5.2 As a ma

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 $1 < \nu \leq 2$, otherwise undefined
Skewness	0 for $\nu > 3$, otherwise undefined
Ex. kurtosis	$\frac{6}{\nu-4}$ for $\nu > 4$, ∞ for $2 < \nu \leq 4$, otherwise undefined
ntropy	$\frac{\nu+1}{2}\left[\psi\left(\frac{1+\nu}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right]$

70 / 70

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

Under H_0: $T_t \sim T\text{-}Dist(V = n_1 + n_2 - 2)$

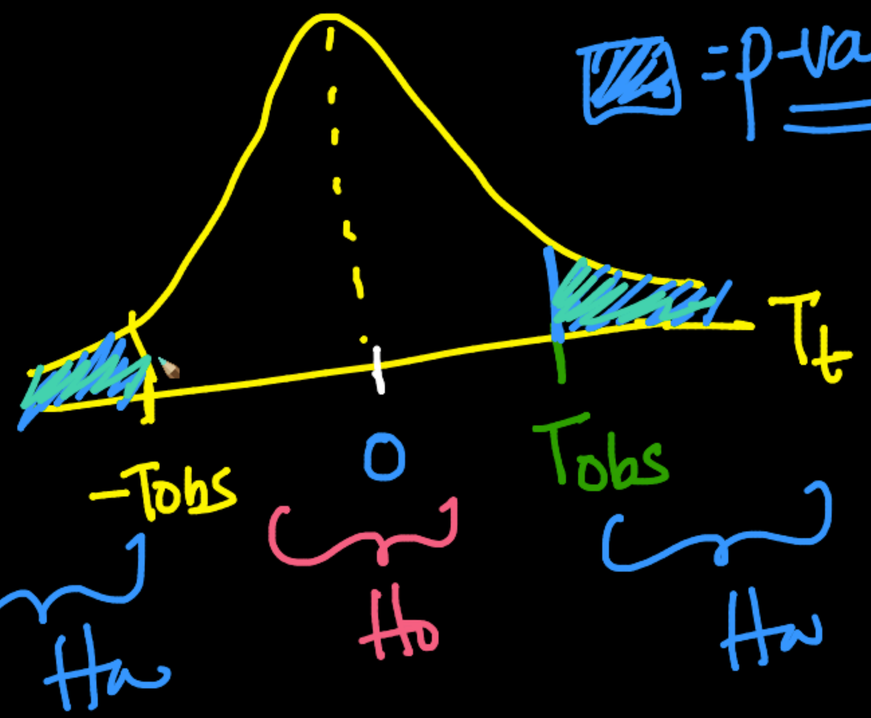
↙ Sensible as it is similar to T_z

2-sided or 1-sided

observed data

$$T_t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Symmetric T-dist^b $(V = n_1 + n_2 - 2)$



Under H_0: $T_t \to 0$

Under H_a: $T_t \to$ large +/-ve values

2-sided

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

p-val vs α

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:27

- Z-proportions
- χ^2 test
- ANOVA

Code → Youtube - dalā

Z-test
T-test
χ^2 test

{ 1-ad → ~~~
2-ad → ~~~

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:29
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colaborat | Normal distribution - Wikipedia
en.wikipedia.org/wiki/Normal_distribution
Update

Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

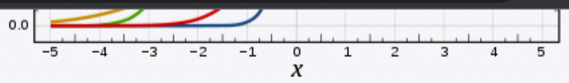
Languages
العربية
Español
Français
हिन्दी
मराठी
Português
தமிழ்
اردو
中文
60 more

is partly due to the central limit theorem. It states that, under some conditions, the average of many samples (observations) of a random variable with finite mean and variance is itself a random variable—whose distribution converges to a normal distribution as the number of samples increases. Therefore, physical quantities that are expected to be the sum of many independent processes, such as measurement errors, often have distributions that are nearly normal.[4]

Moreover, Gaussian distributions have some unique properties that are valuable in analytic studies. For instance, any linear combination of a fixed collection of normal deviates is a normal deviate. Many results and methods, such as propagation of uncertainty and least squares parameter fitting, can be derived analytically in explicit form when the relevant variables are normally distributed.

A normal distribution is sometimes informally called a **bell curve**.[5] However, many other distributions are bell-shaped (such as the Cauchy, Student's *t*, and logistic distributions).

The univariate probability distribution is generalized for vectors in the multivariate normal distribution and for matrices in the matrix normal distributio



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 \in \mathbb{R}_{>0}$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2p-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
MAD	$\sigma\sqrt{2/\pi}$
Skewness	0
Ex. kurtosis	0
Entropy	$\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}$
MGF	$\exp(\mu t + \sigma^2 t^2/2)$

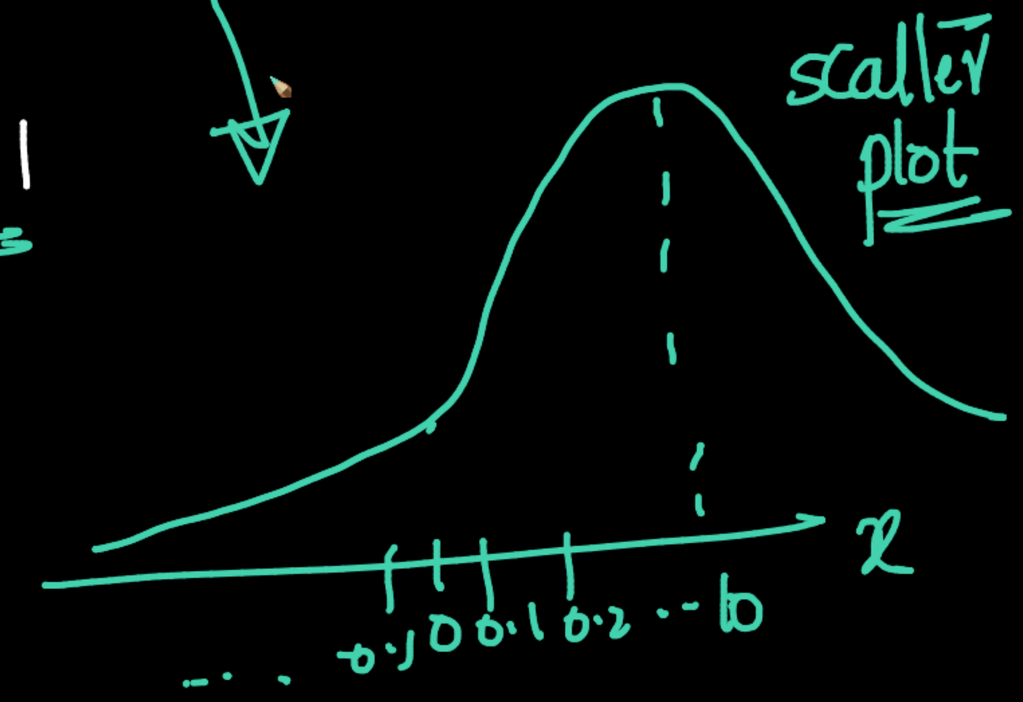
75 / 76

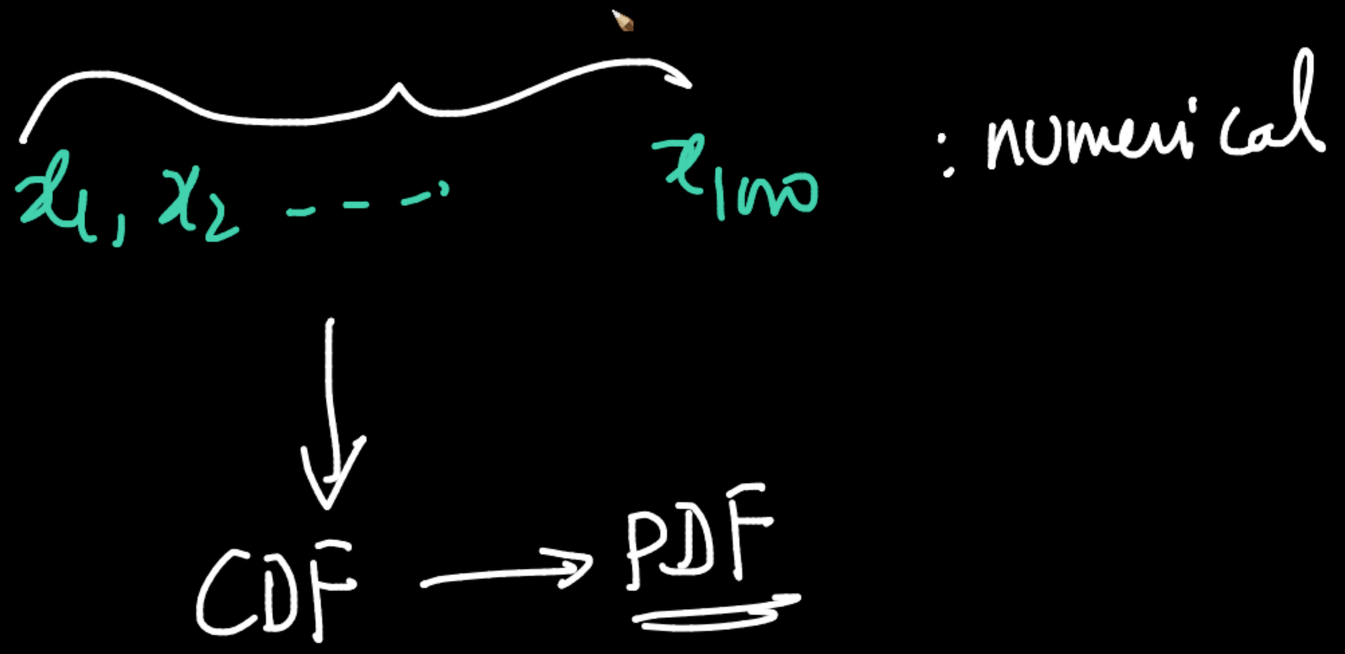
Theoritical dist

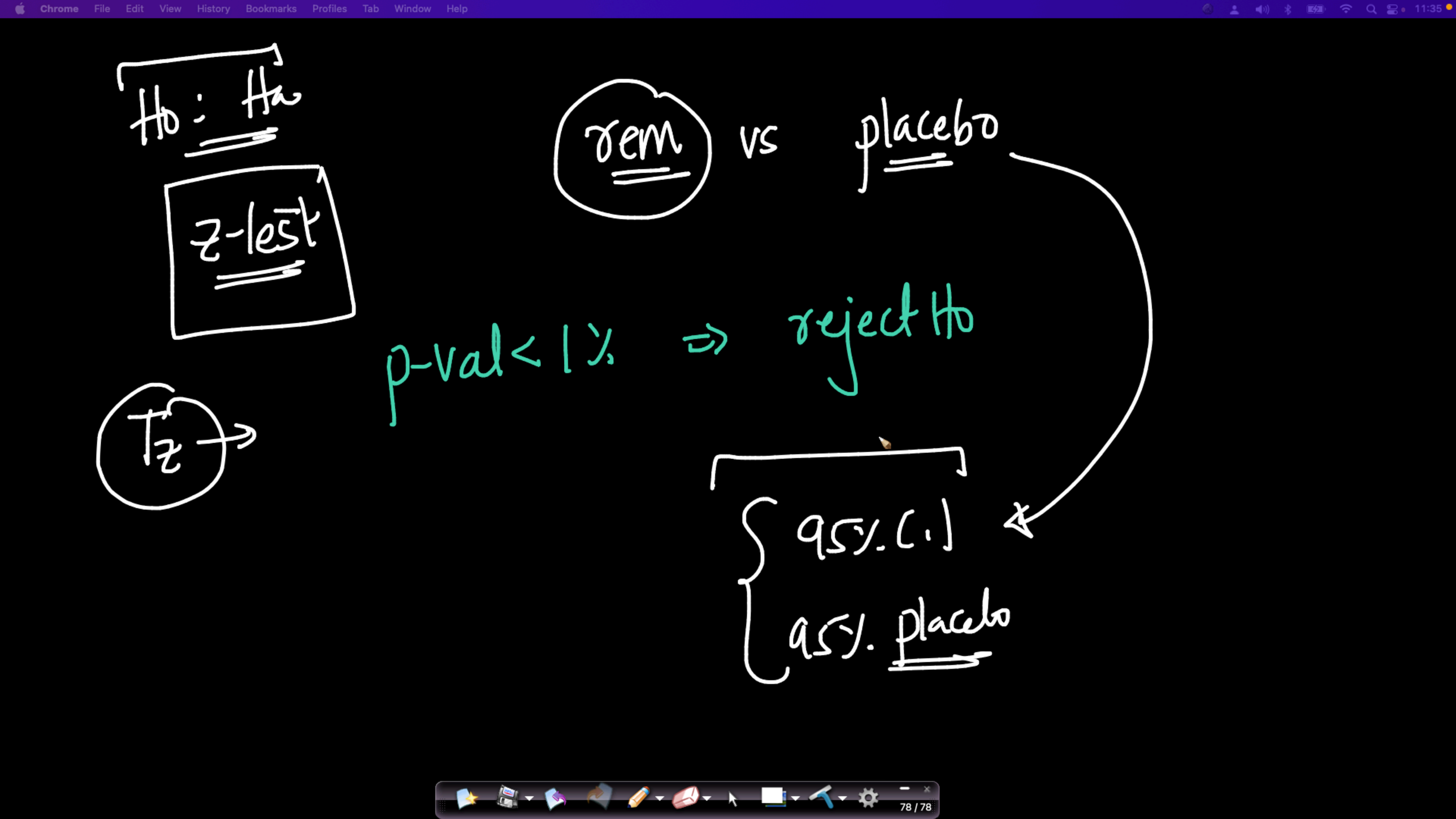
pdf

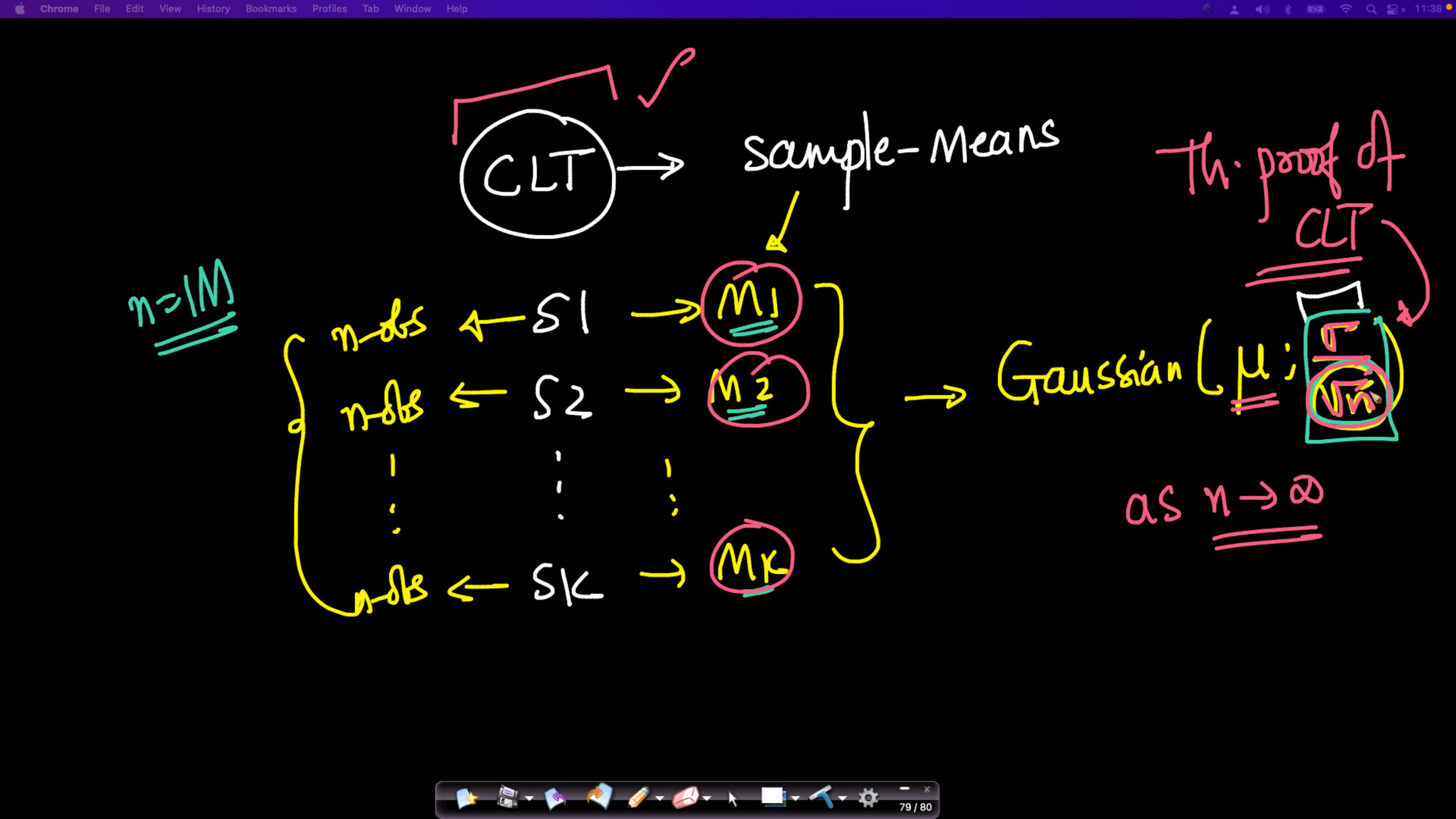
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-10}{1}\right)^2\right\}$$

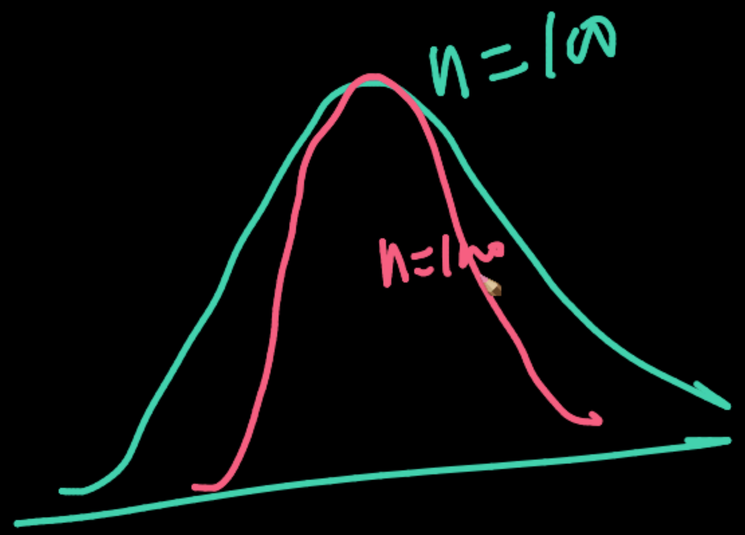
$\mu = 10$; $\sigma = 1$




$x_1, x_2 \cdots$
x_{100}
: numerical
CDF
PDF


Ho : Ha
z-test
rem vs placebo
p-val < 1% ⇒ reject Ho
Tz
95% CI
95% placebo


CLT
sample-Means
Th. proof of CLT
n=1M
n_obs
S1
M1
n_obs
S2
M2
n_obs
Sk
Mk
Gaussian (μ;
as n → ∞
79 / 80


n=10
n=10

box-cox

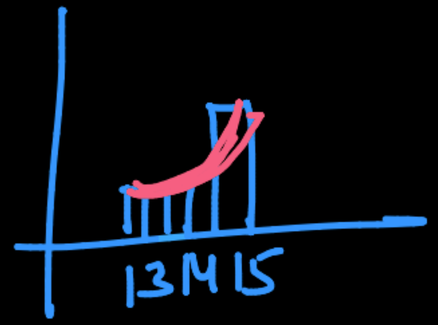
$$y_i = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases}$$

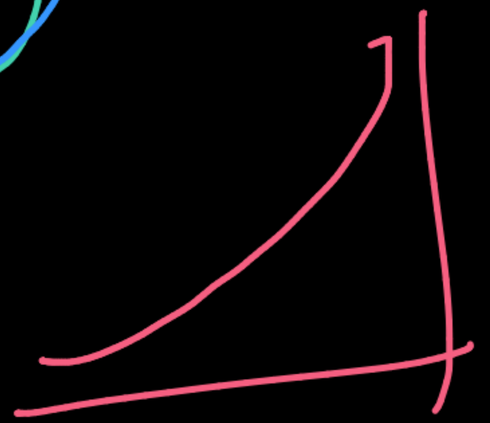
power-law

log-normal

✓ x_i's : 13, 14, 15, 15

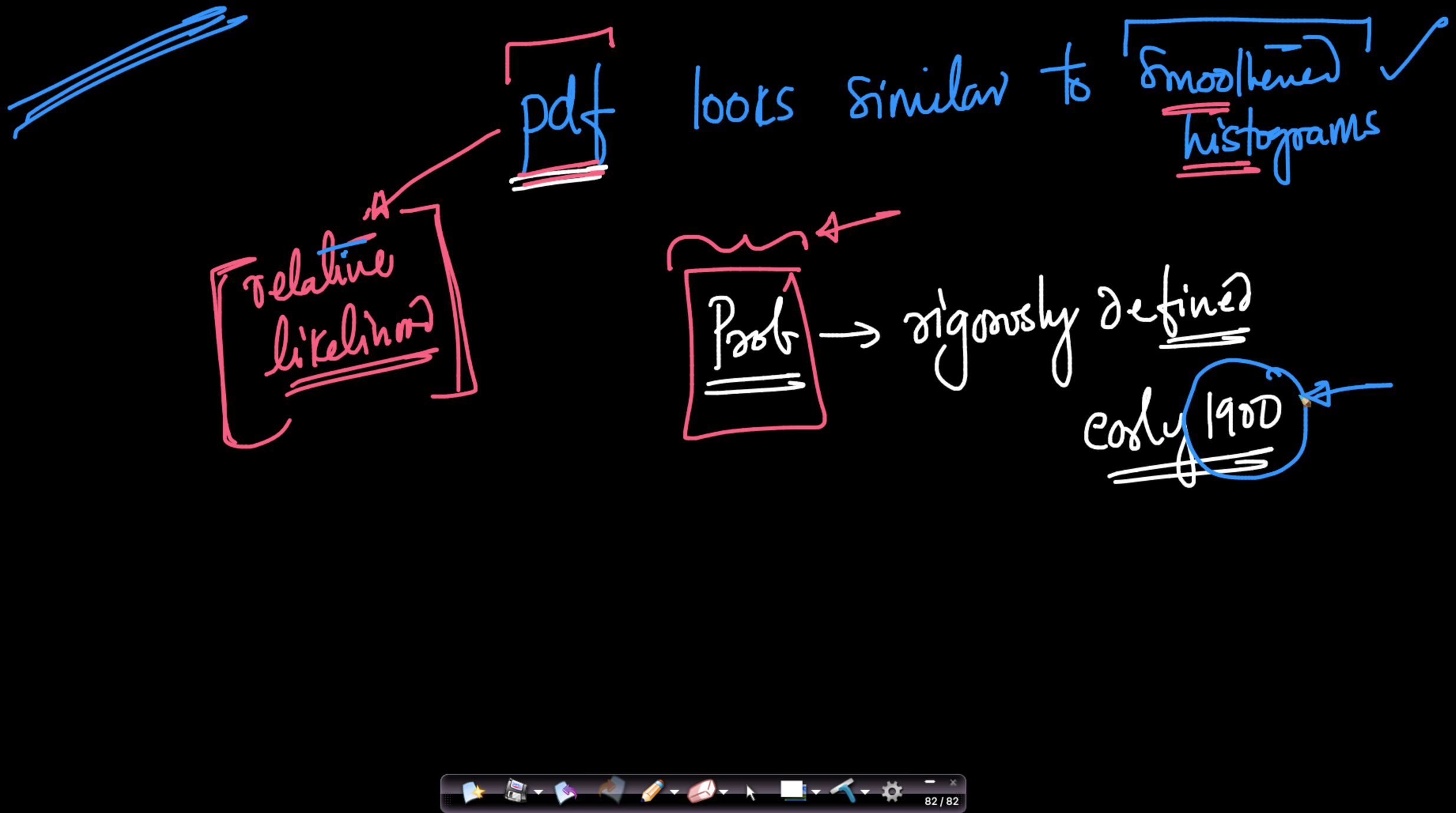
y_i's :





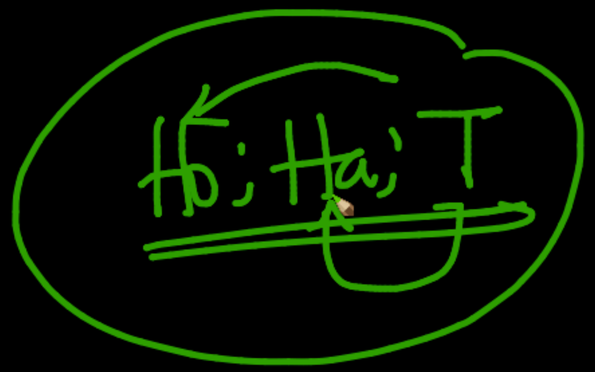
KS-test(s) & plot

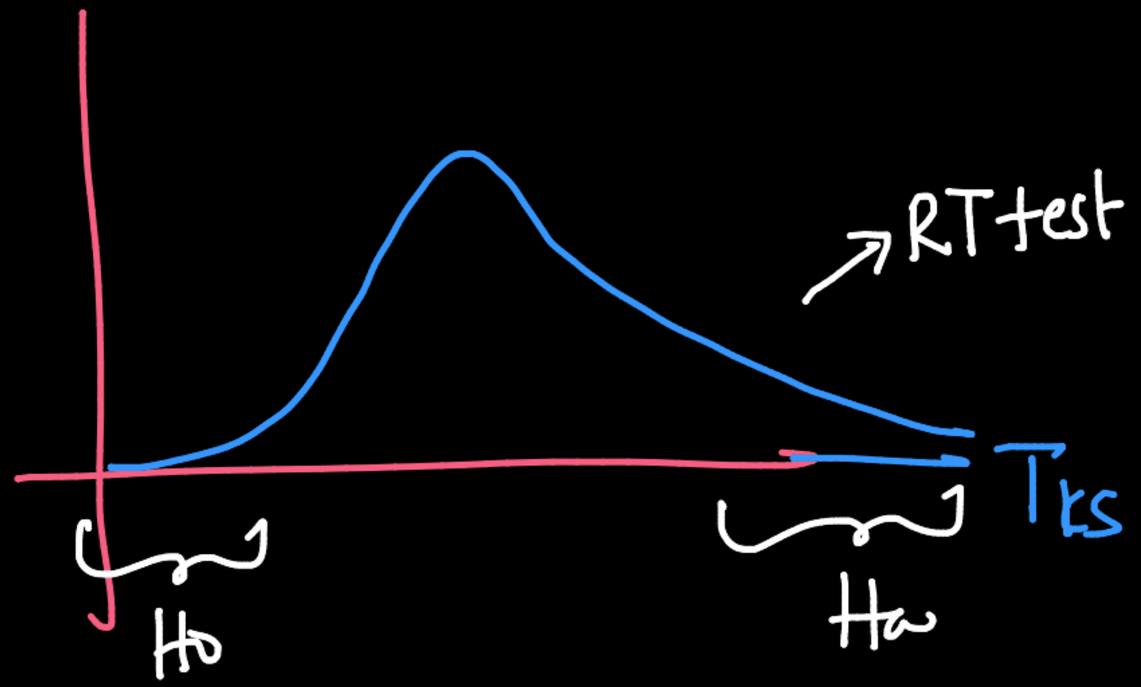
81 / 81


pdf
looks similar to
Smoothened
histograms
relative
likelihood
Prob
→ rigorously defined
early 1900

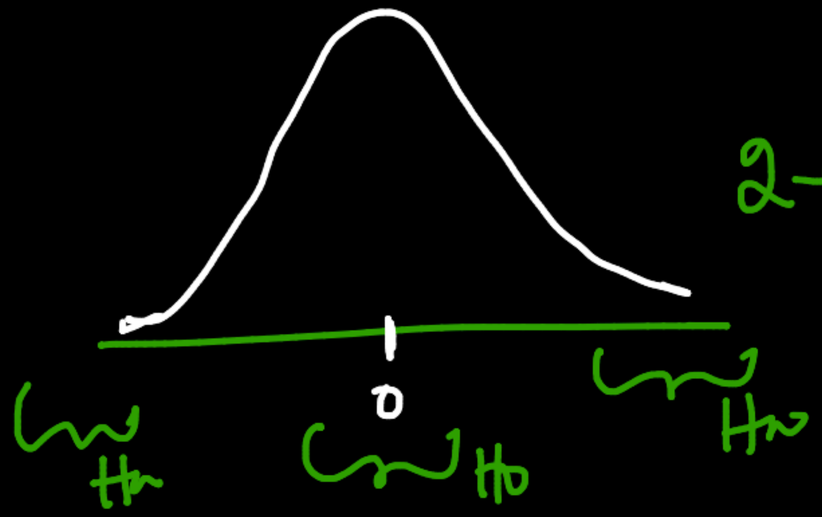
Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:49

T_{RS}





T_t



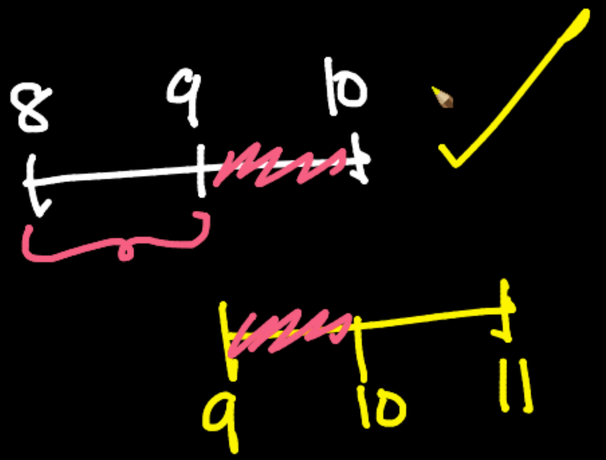
2-sided test-

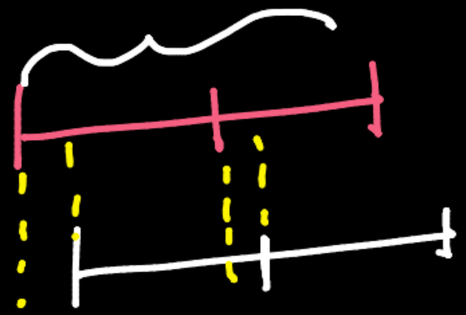
83 / 83

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

✓ Med 1: Mean rec. time

Med 2: Mean sec. time





→ no conclusion

Chrome File Edit View History Bookmarks Profiles Tab Window Help
boxcox and z-score.ipynb - Co | scipy.stats.boxcox — SciPy v1. | Anderson–Darling test - Wikip | KStest_Ttest.ipynb - Colabora | Power transform - Wikipedia
en.wikipedia.org/wiki/Power_transform#Yeo–Johnson_transformation
Update

not be a substantive problem in many applications.[6][7]

Box–Cox transformation [edit]

The one-parameter Box–Cox transformations are defined as

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases}$$

and the two-parameter Box–Cox transformations as

$$y_i^{(\lambda)} = \begin{cases} \dfrac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0, \\ \ln(y_i + \lambda_2) & \text{if } \lambda_1 = 0, \end{cases}$$

as described in the original article.[8][9] Moreover, the first transformations hold for $y_i > 0$, and the second for $y_i > -\lambda_2$.[8]

The parameter λ is estimated using the profile likelihood function and using goodness-of-fit tests.[10]

Confidence interval [edit]

Confidence interva y constructed using Wilks's theorem on the profile likelihood function to find all the possible values of λ that fulfill the following restriction:[11]

$$\ln\left(L(\lambda)\right) > \ln\left(L(\hat{\lambda})\right) - \frac{1}{2}\chi^2_{1,\alpha};$$

85 / 85