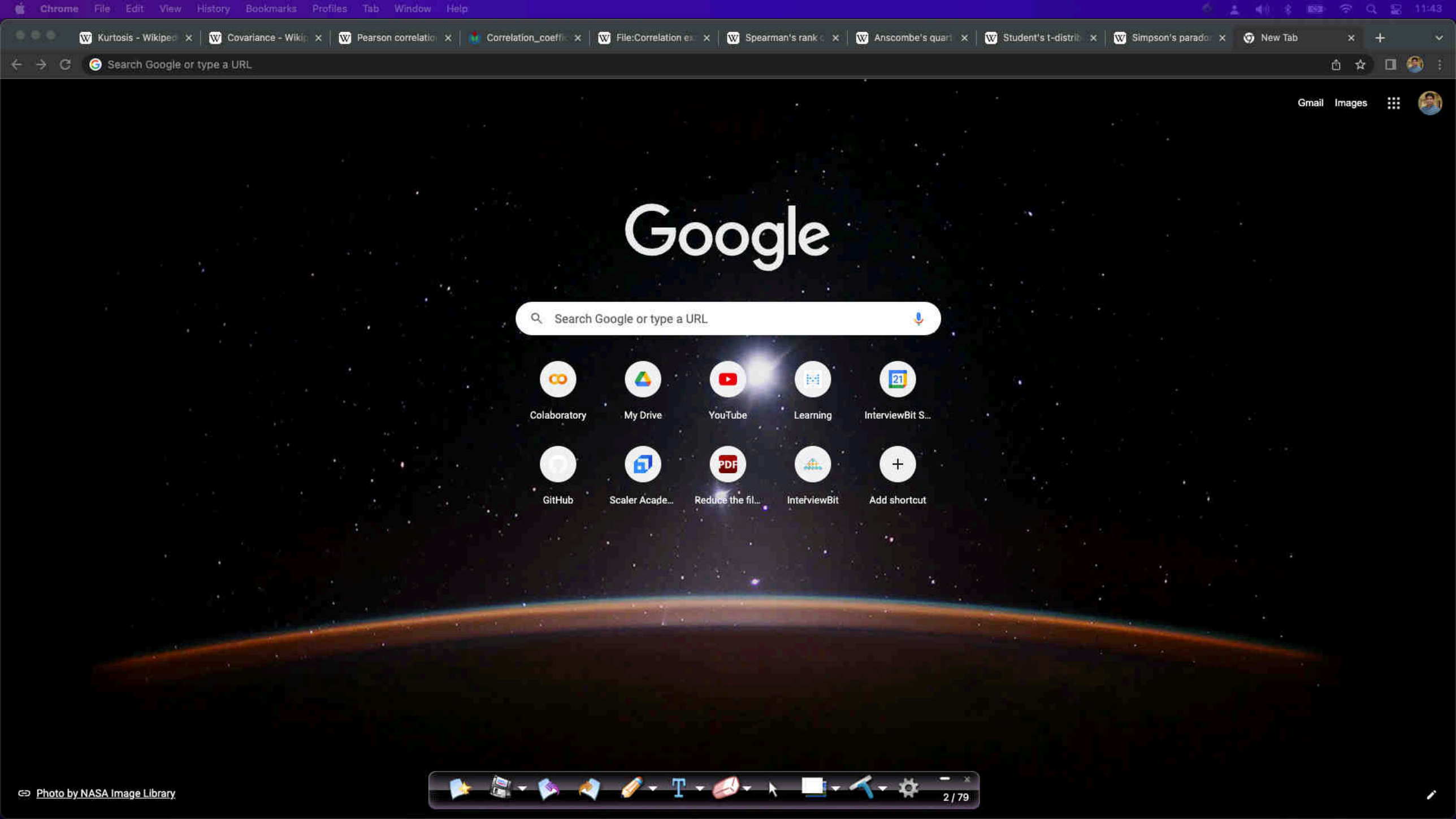
Topics:

- Types of ANOVA
- α, β, power & effect-size — Defns & terms
- Kurtosis ✓
- Covariance ✓
- { pearson correlation coefficient, Spearman rank corr. coefficient } ✓


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Kurtosis - Wikipe
Covariance - Wiki
Pearson correlatio
Correlation_coeffi
File:Correlation ex
Spearman's rank c
Anscombe's quart
Student's t-distrib
Simpson's parado
New Tab
Search Google or type a URL
Gmail
Images
Google
Search Google or type a URL
Colaboratory
My Drive
YouTube
Learning
InterviewBit S...
GitHub
Scaler Acade...
Reduce the fil...
InterviewBit
Add shortcut
Photo by NASA Image Library
2 / 79

✓ Programming → mock-interviews

✓ SQL → mock-interviews

[Foundational Math → "

"Framework"

ANOVA Variations

→ Tens of variations

learnt-

One way

2-way

Multiway ANOVA

72

indep. variable

drug-type → d_1, d_2, d_3

→ #days to recover

dependent var

dep.var = #days to recover

indep-var ✓

(2) → d_1, d_2, d_3

K, A, O

dep-var = #days to recover

indep-var → d_1, d_2, d_3

→ K, A, O

→ SA ; SEA

AA ; CAU

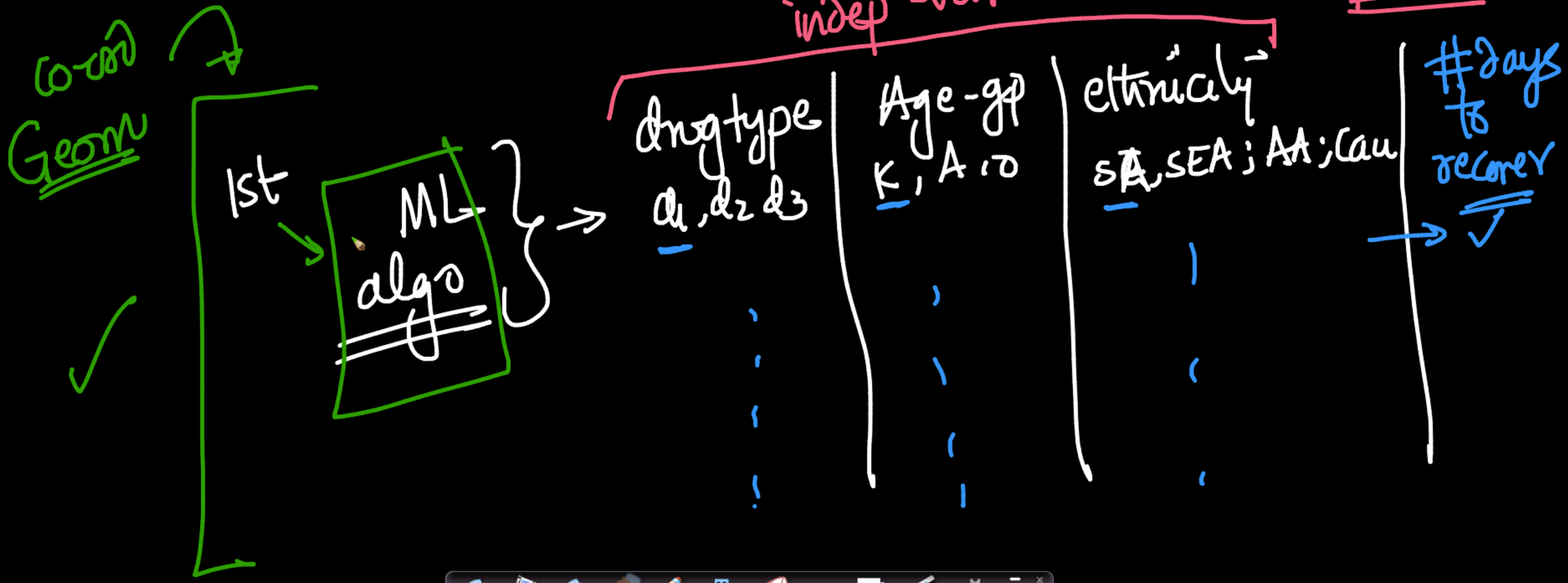
Multi-way ANOVA

special case

Linear Regr

Simplest ML

indep -var



MANOVA

multiple

dep-var : Salary

sq.feet of the home

indep.var

College-degree

[Hs, B, M, Phd]

One-way Manova

2-way MANOVA:

$\alpha = .5$

2 indep-var:

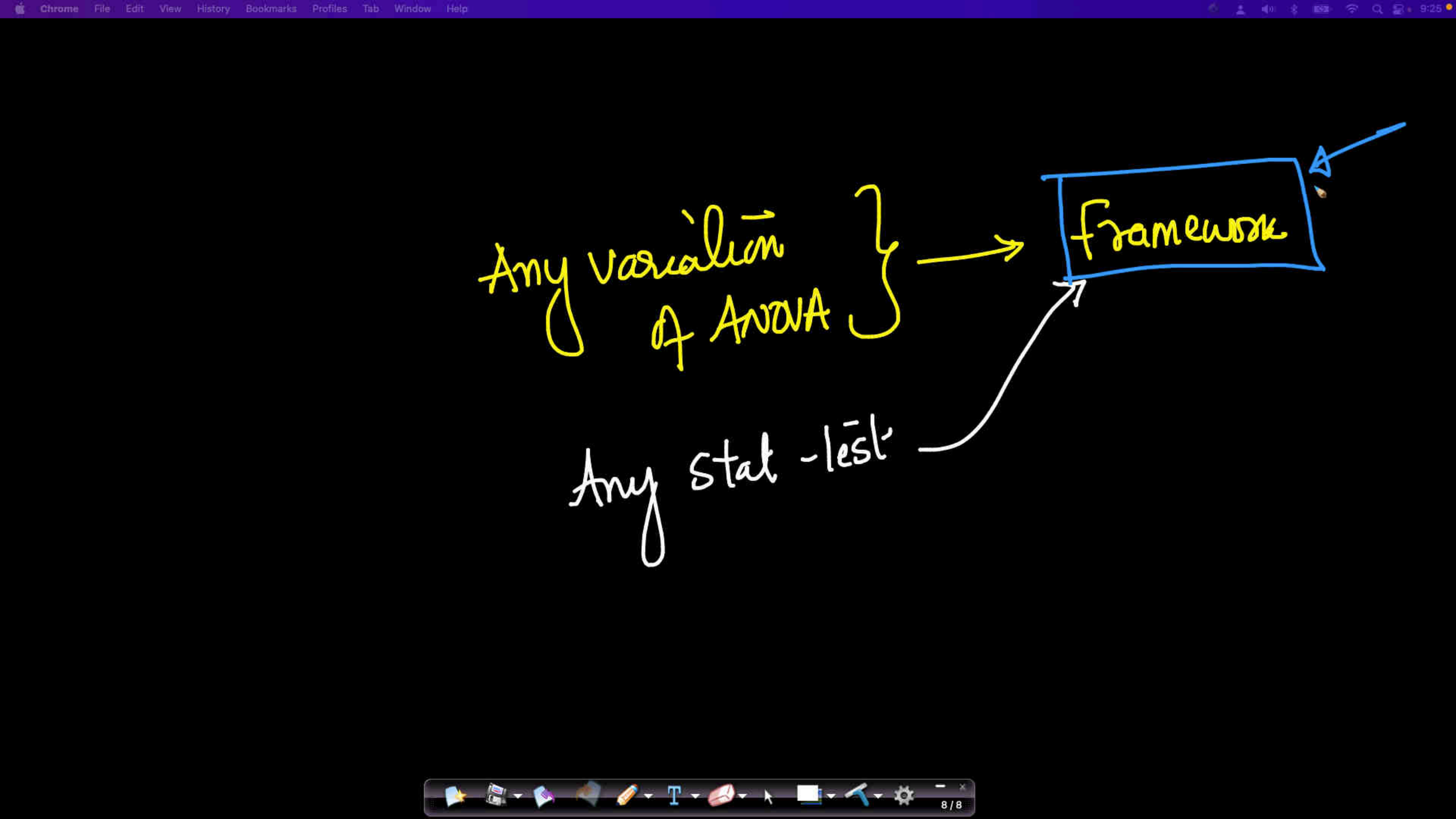
- College-degree [HS, B, M, PhD]
- IQ-score [L, M, H, VH]

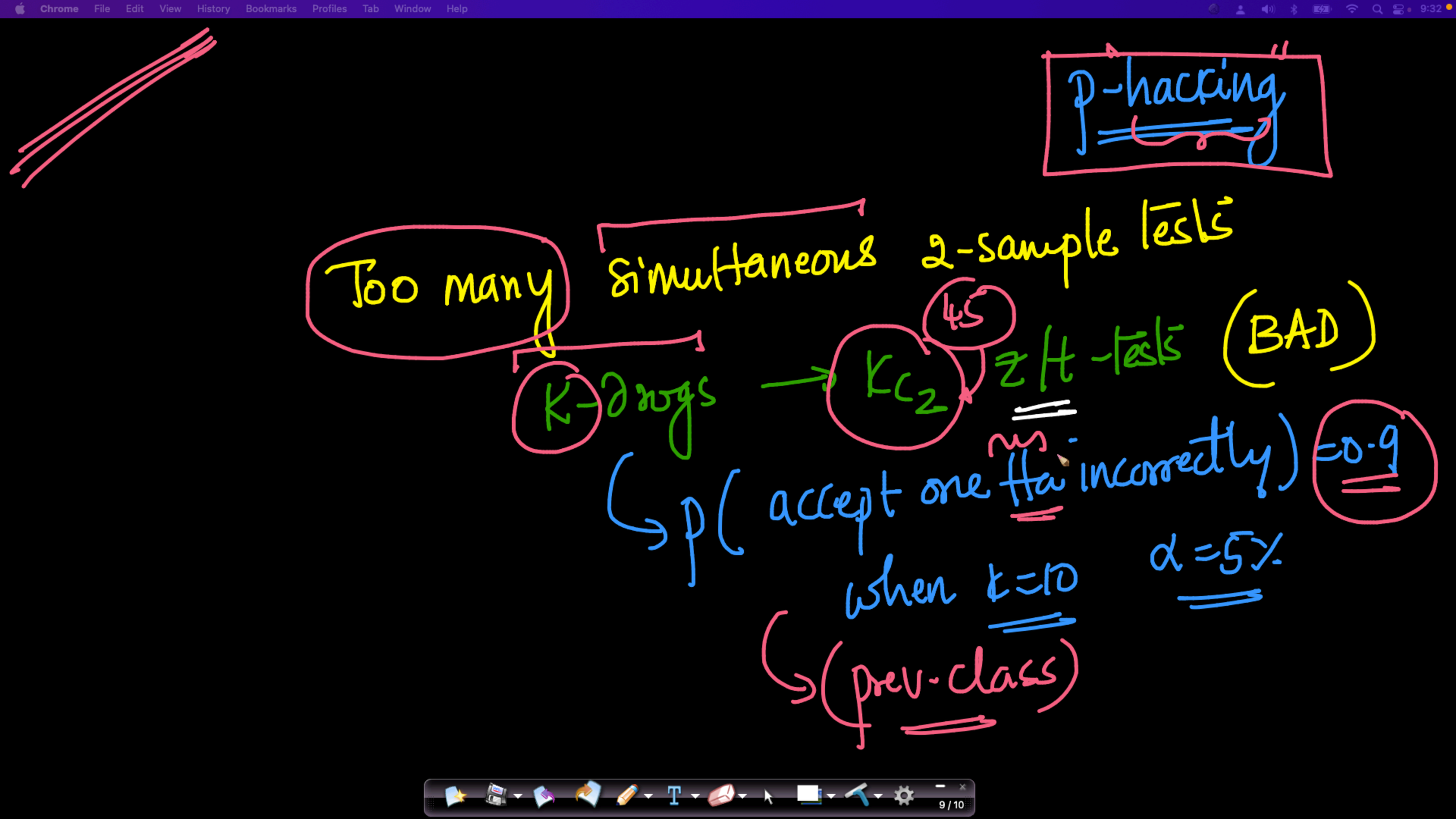
dep.var:

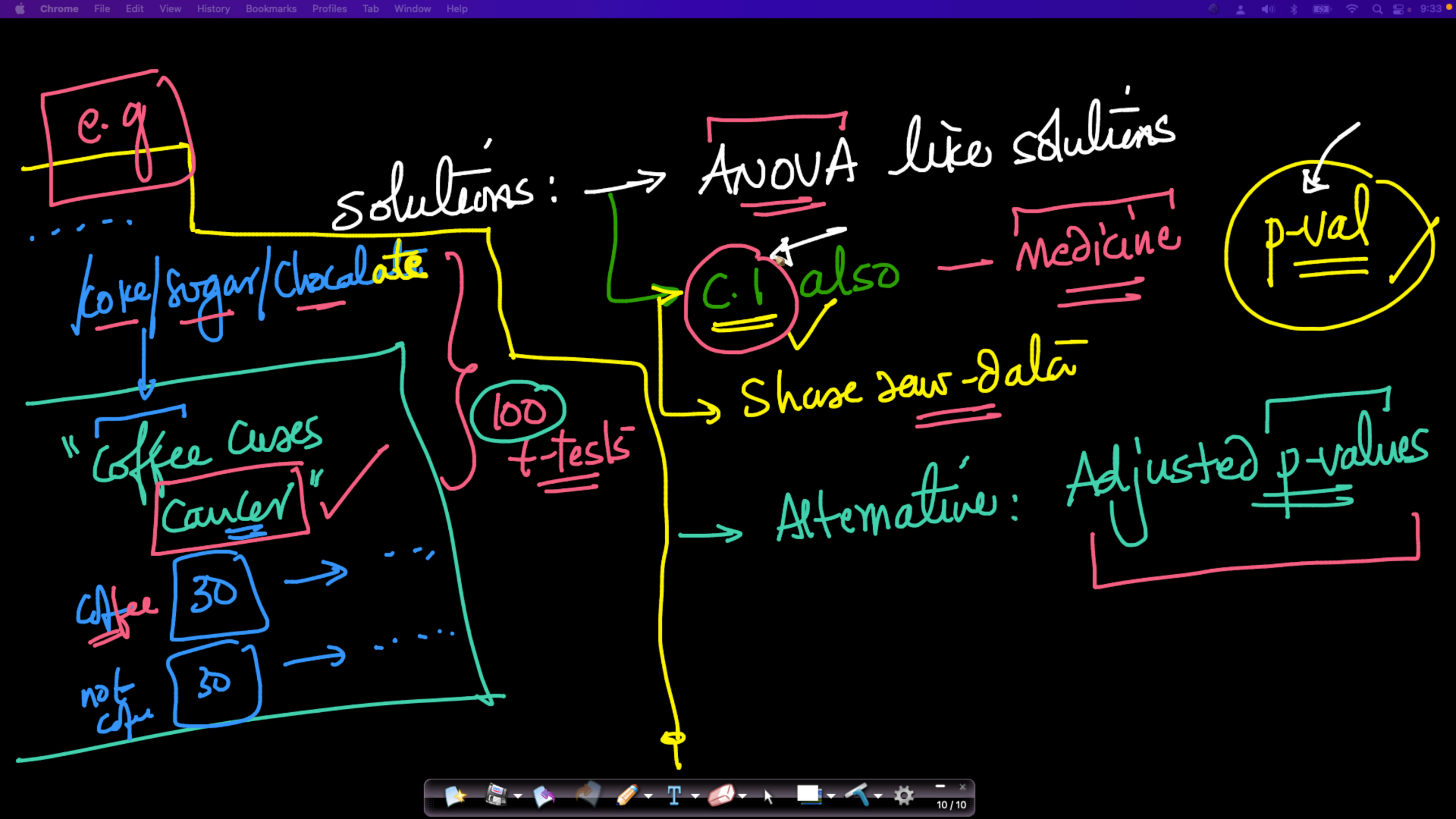
- Salary
- sq ft of the home

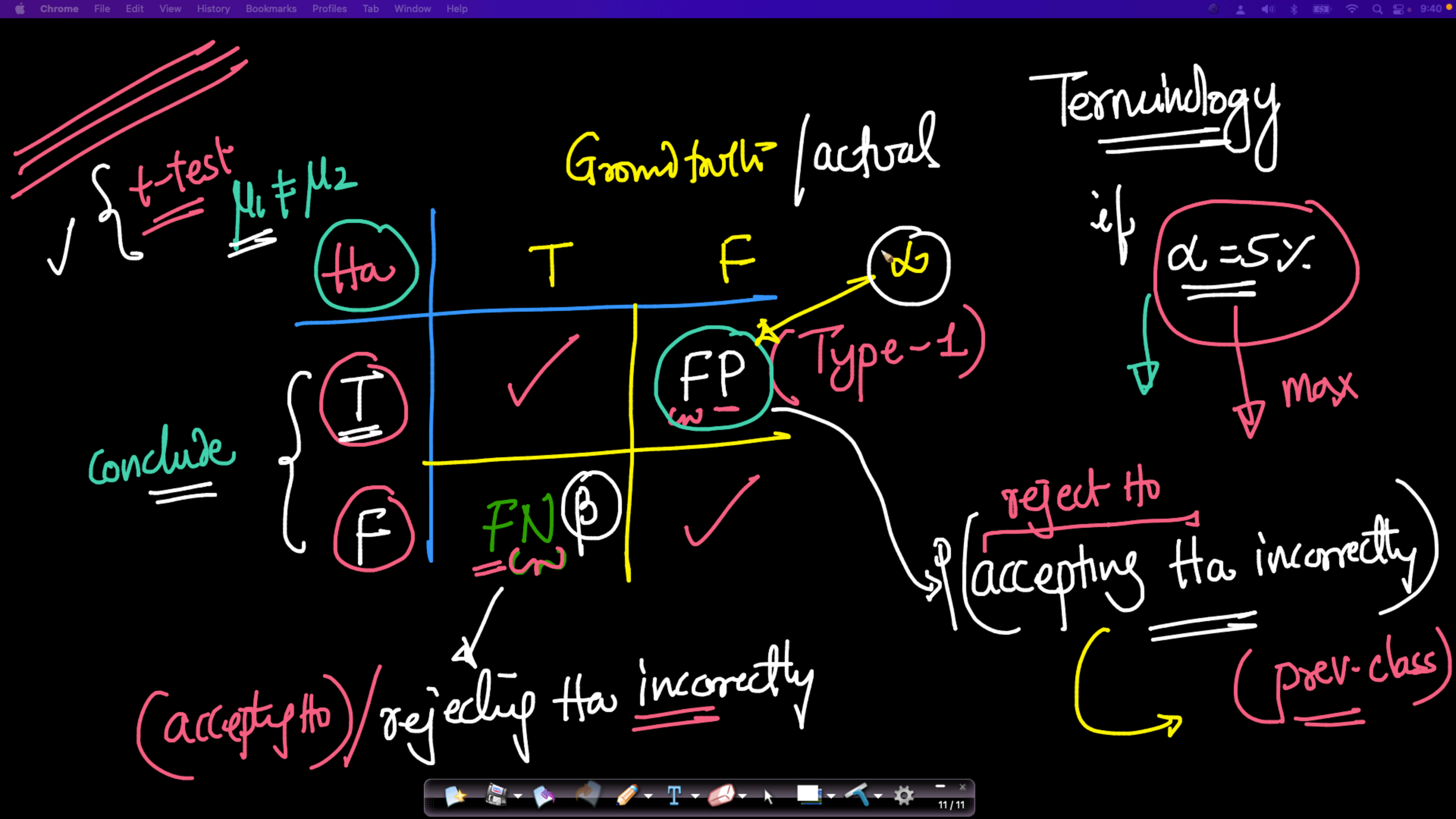
multiple-stat-tests:

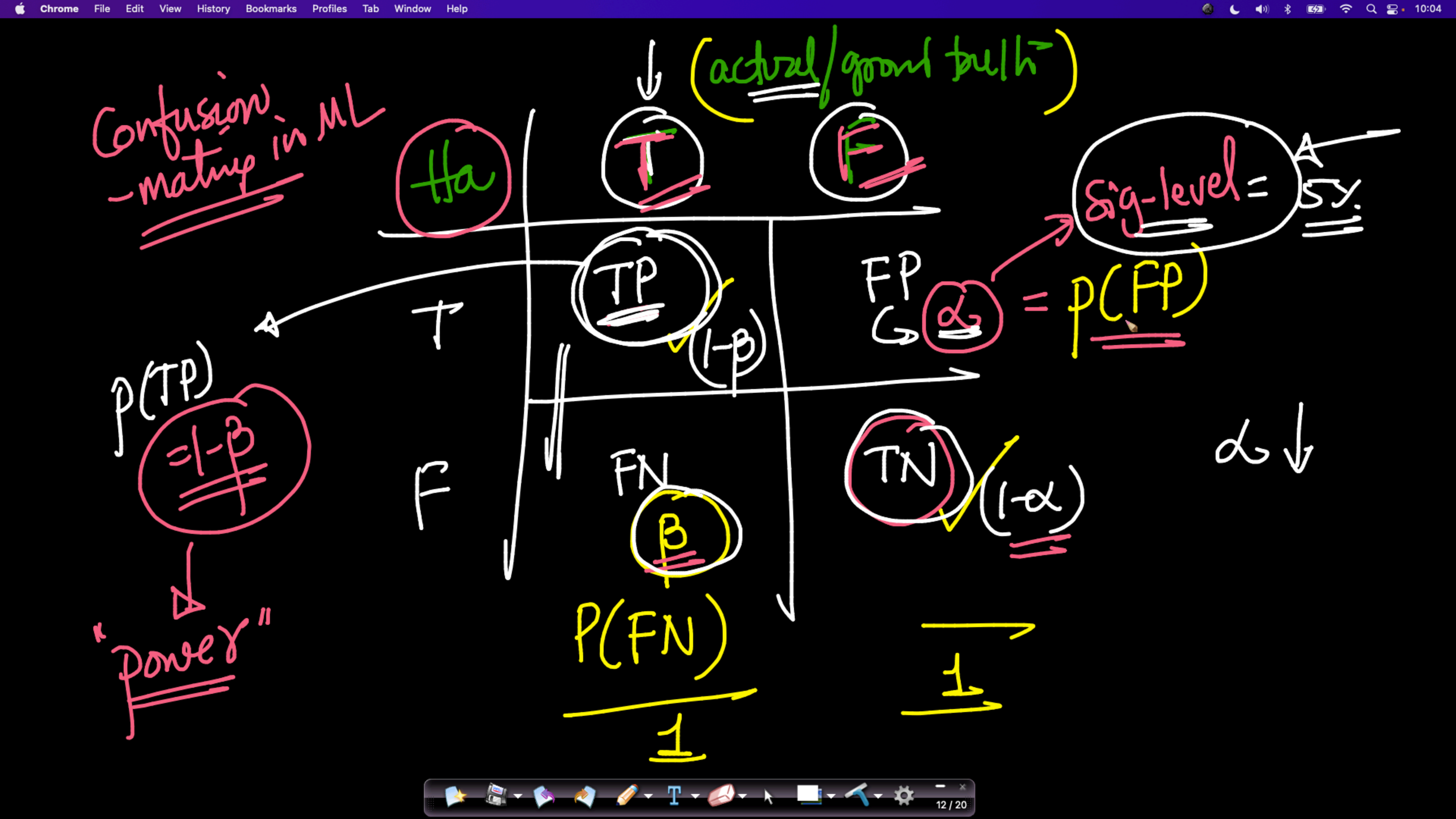
K_{C_2}

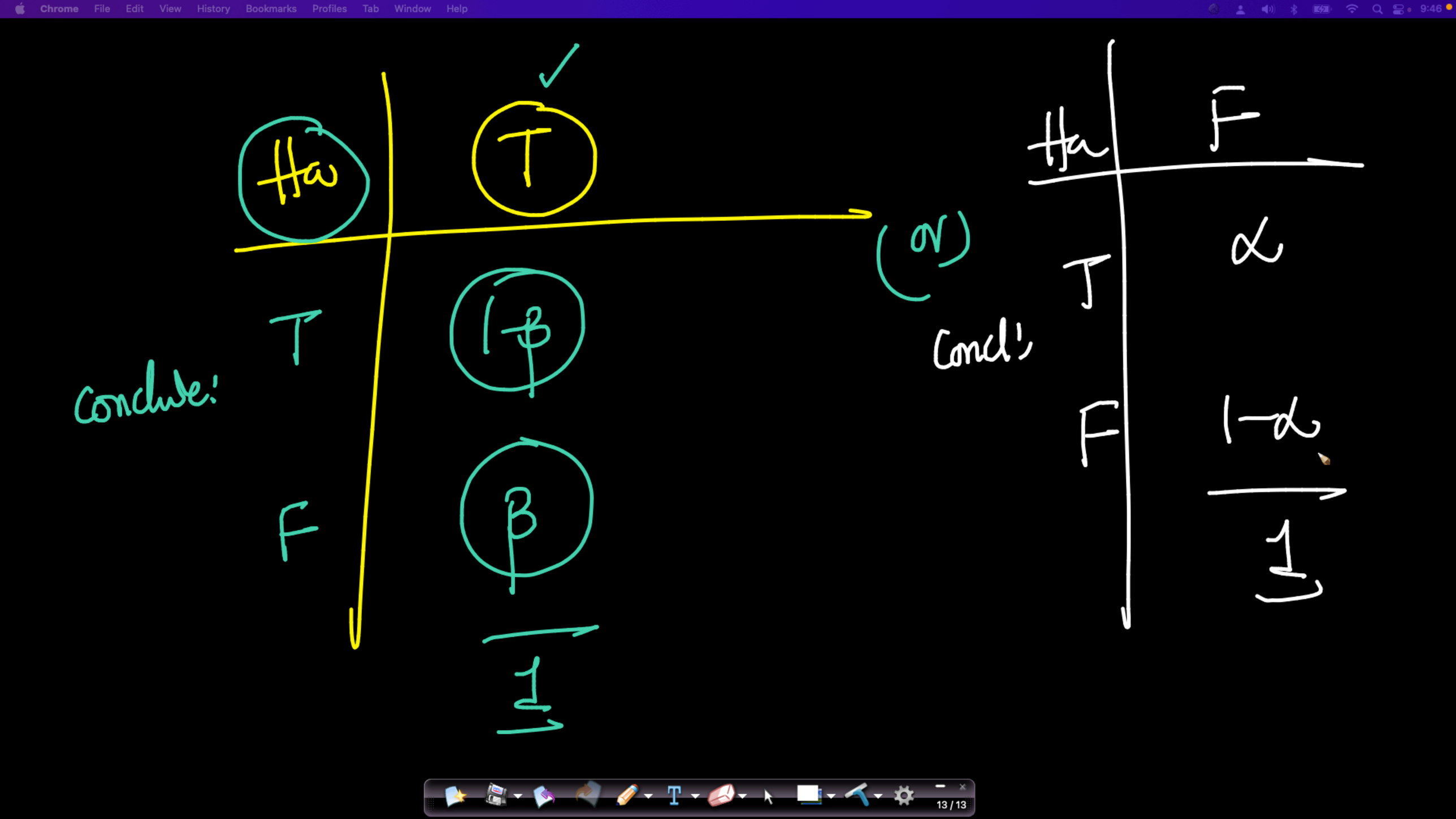

Any variation of ANOVA
Any stat -test
framework


p-hacking
Too many
simultaneous 2-sample tests
K-drugs
KC_2
45
2 H_0-tests (BAD)
↳ p(accept one H_0 incorrectly) = 0.9
when k=10
$\alpha = 5\%$
↳ (prev-class)


e.g
Coke/Sugar/Chocolate
Solutions:
ANOVA like solutions
C.I also
medicine
p-val
Share raw-data
"Coffee Causes Cancer"
100 t-tests
Alternative:
Adjusted p-values
coffee
30
not coffee
30


Terminology
Ground truth / actual
t-test
$\mu_1 \neq \mu_2$
Ha
T
F
T
F
FP
(Type-1)
FN
β
Conclude
if
α = 5%
max
reject Ho
(accepting Ho incorrectly
(prev. class)
(accepting Ho) / rejecting Ho incorrectly




Chrome File Edit View History Bookmarks Profiles Tab Window Help
9:46
Ha
T
1-β
β
T
F
Conclude:
1
(or)
Concl:
Ha
F
α
T
F
1-α
1
13 / 13

α = Sig-level

$1-\beta$ = power

Effect-size

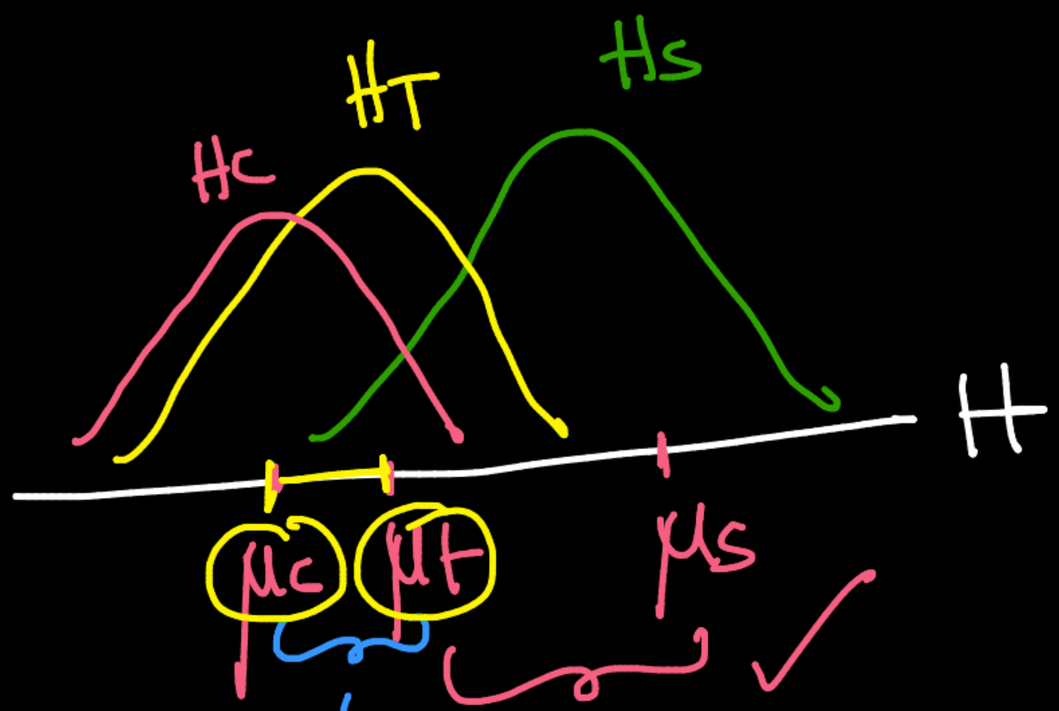
Hs: heights of Swedish people

Ht: " " Thailand

Hc: " " China

t-test:

(1) $\mu_c = \mu_t$



effect size is small

(TP)

↳ power like test ↓

Very close

(2) $\mu_t = \mu_s$

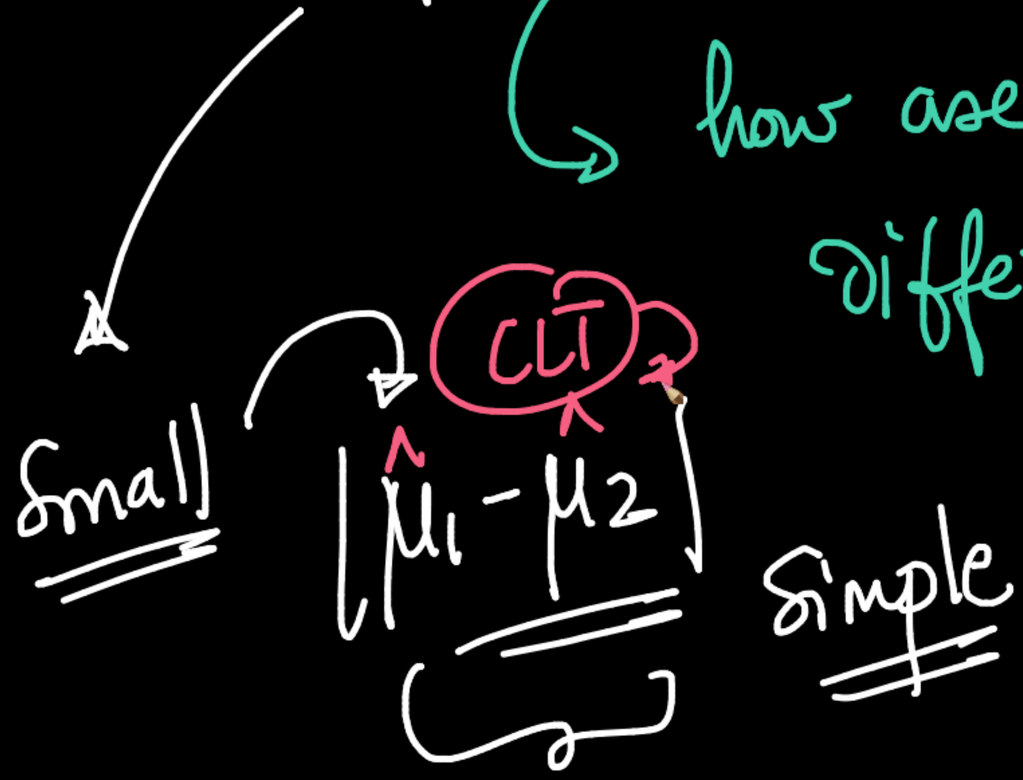
effect size is larger

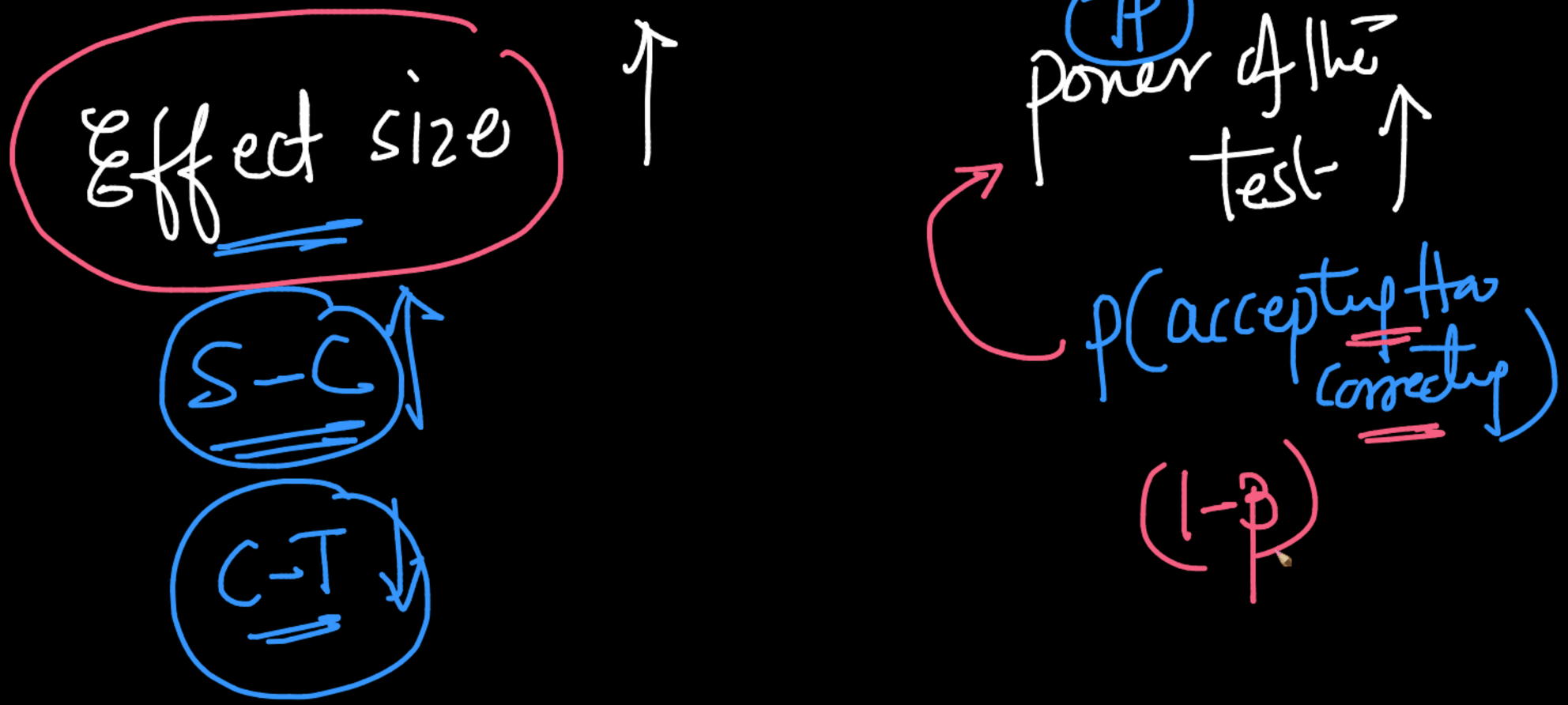
↓

Power of my test ↑

Effect-size in hyp-testing

how are we the underlying difference in means (in this context)




Effect size
S-C
C-T
TP
Power of the test
p(accepting Ho | Ho is correct)
(1-β)

T-test: → α, β

(problem/data)

TN

{ Effect-size : how different are the underlying populations

↳ v small → power ↓↓ → TP

Practical → increase n_1 & n_2

H_a

A/B Test

α = 0.1%

β = function of test ; n_1 , n_2 ↑ ...

power ↑

$1-\alpha$ ↑

Kurtosis

measure of tailedness

$$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4$$

$K_{normal} = 3$ (proven ...)

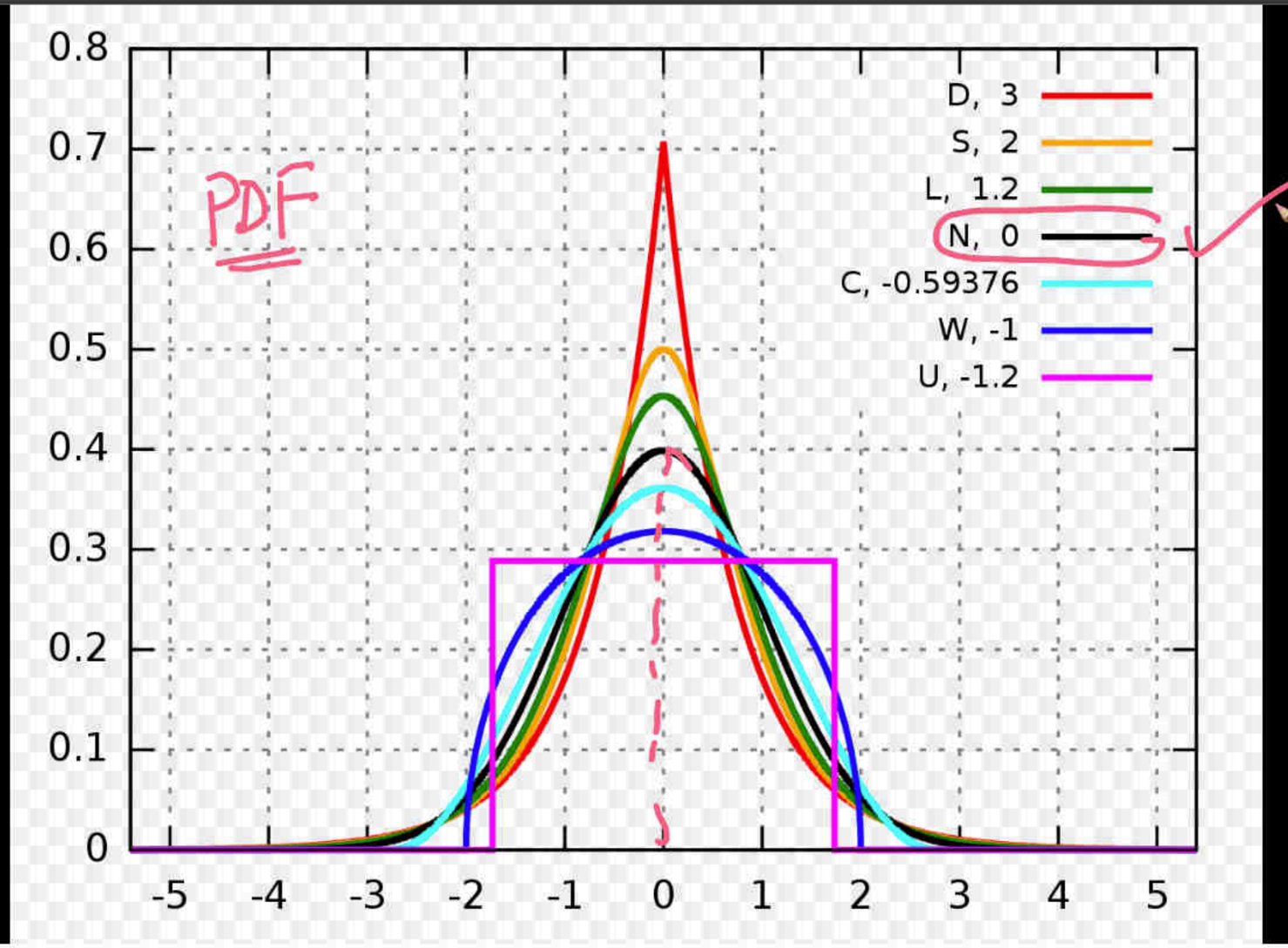
Mean: measure of central value

Var: measure of spread

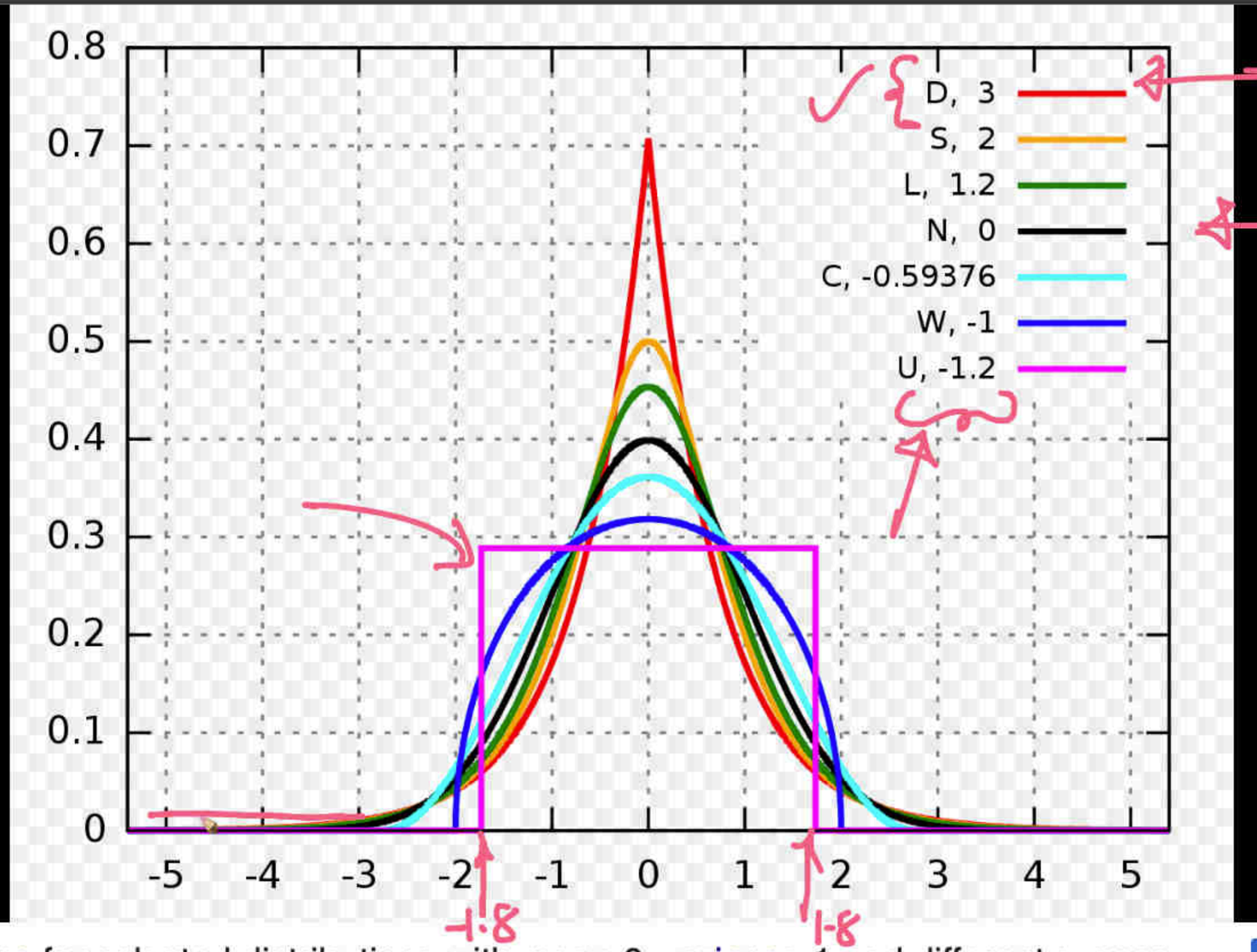
"Excess - kurtosis" = K − 3

↳ Normal = 0

$K_e > 0$ fatter tails than Gaussian

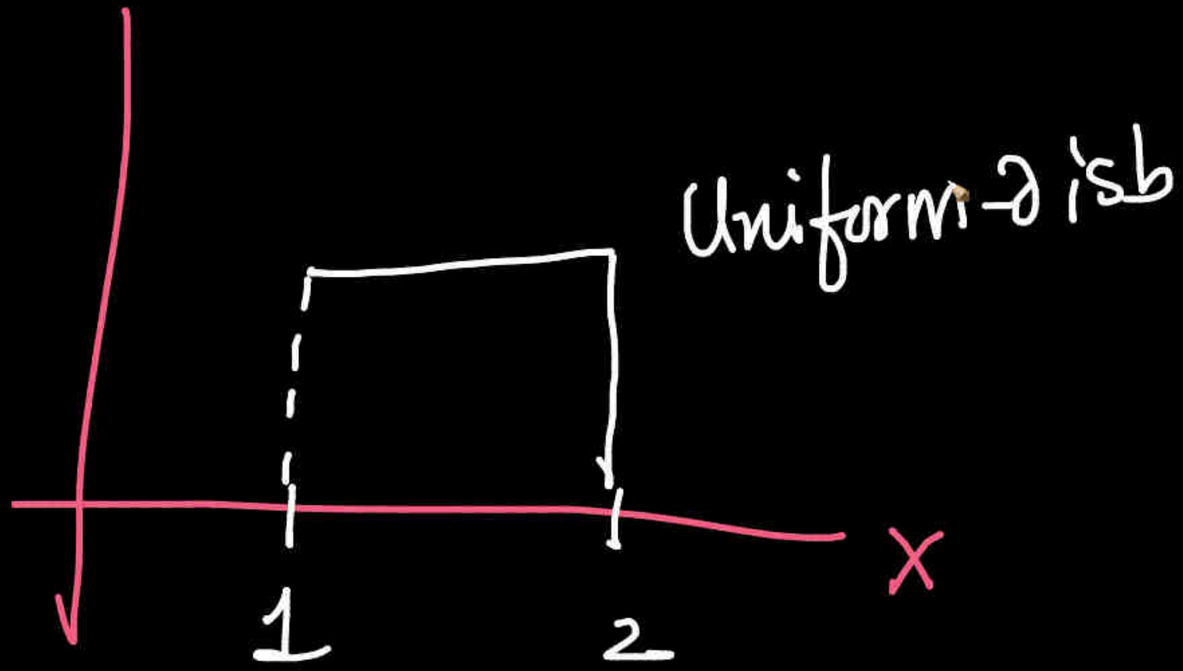


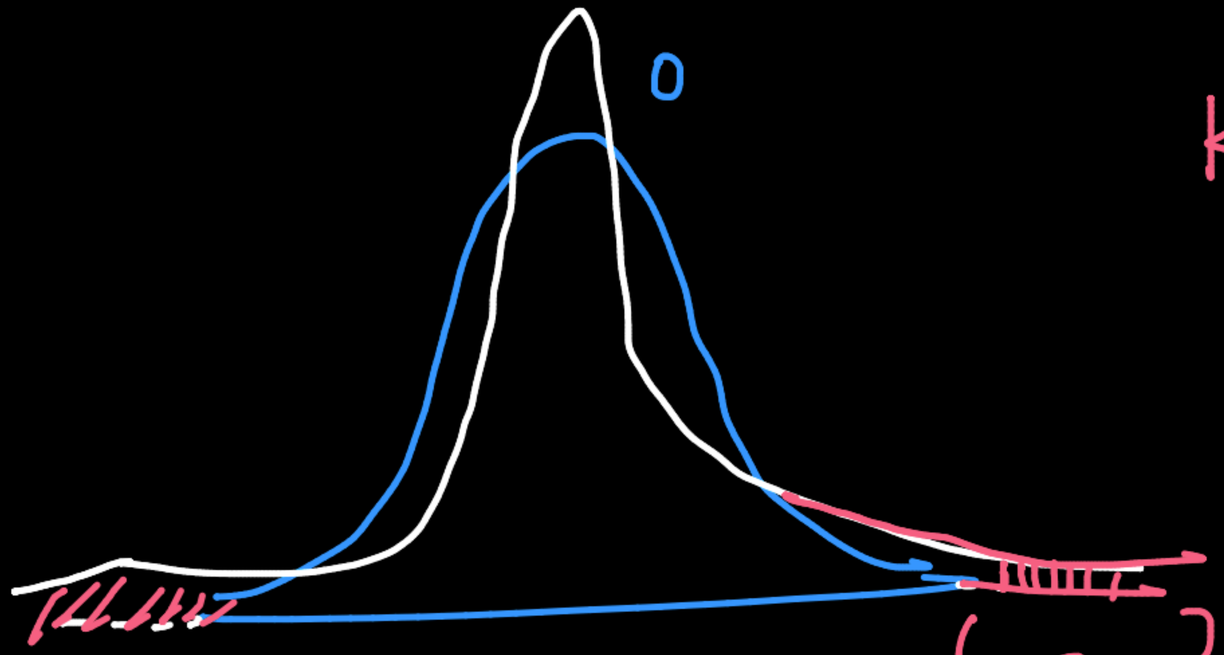
Probability density functions for selected distributions with mean 0, variance 1 and different excess kurtosis



Probability density functions for selected distributions with mean 0, variance 1 and different excess kurtosis

More details

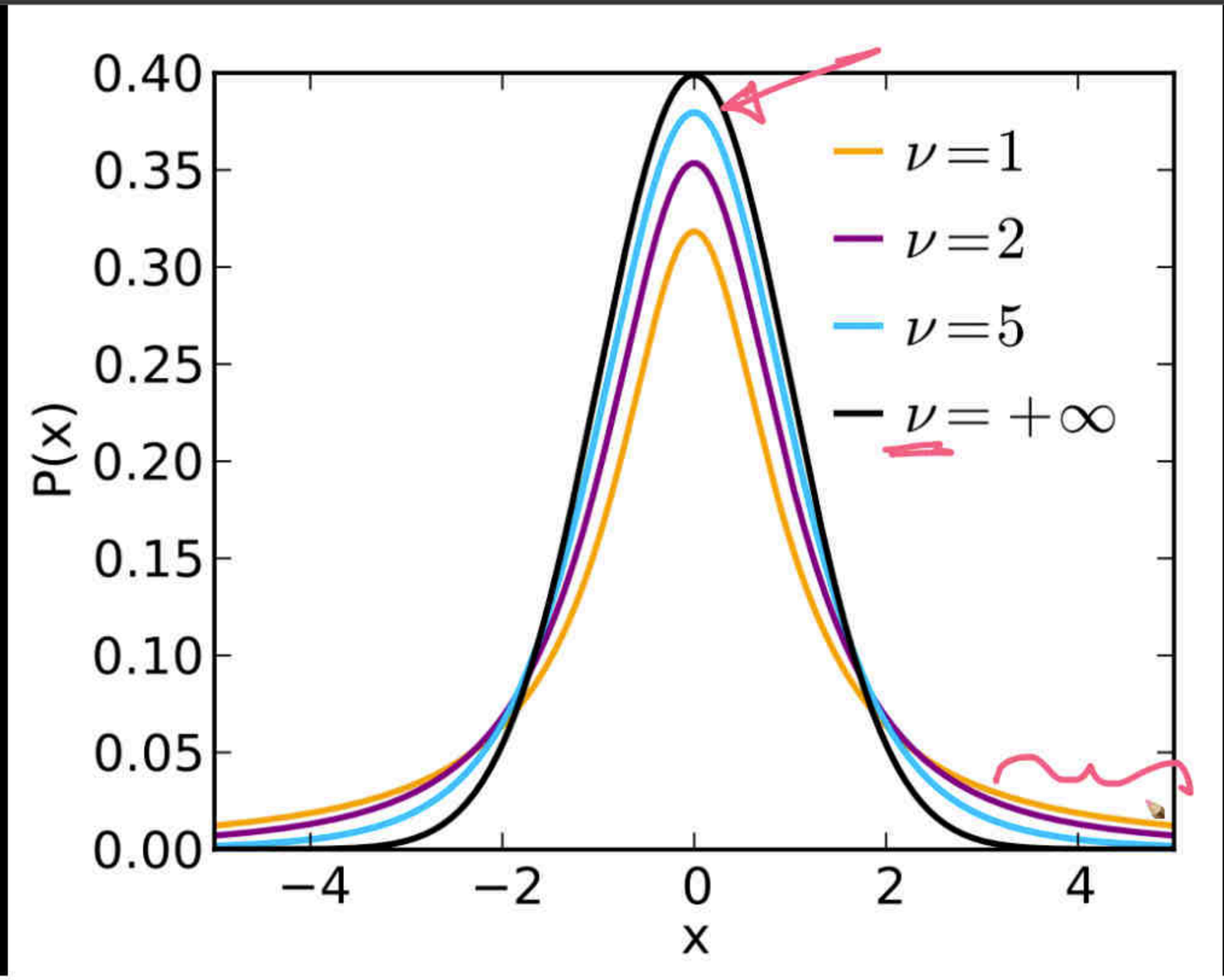




$K \uparrow \Rightarrow$ more chances of extreme values

can find more pts here

can find more pts

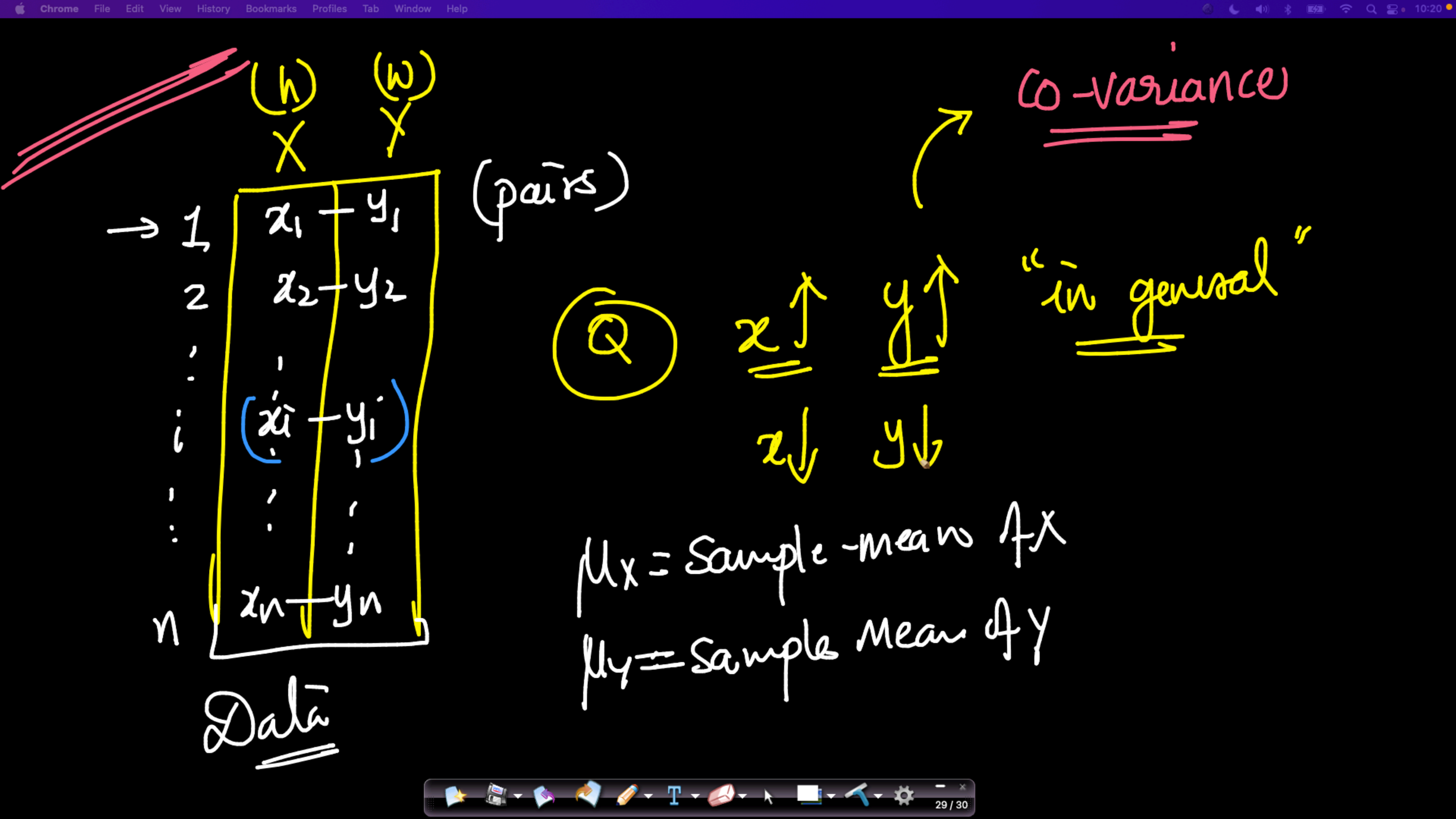


Plot of the density function for several members of the Student t family.

Skbkekas - Own work

$$\uparrow k = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4$$

$$\uparrow \text{var} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

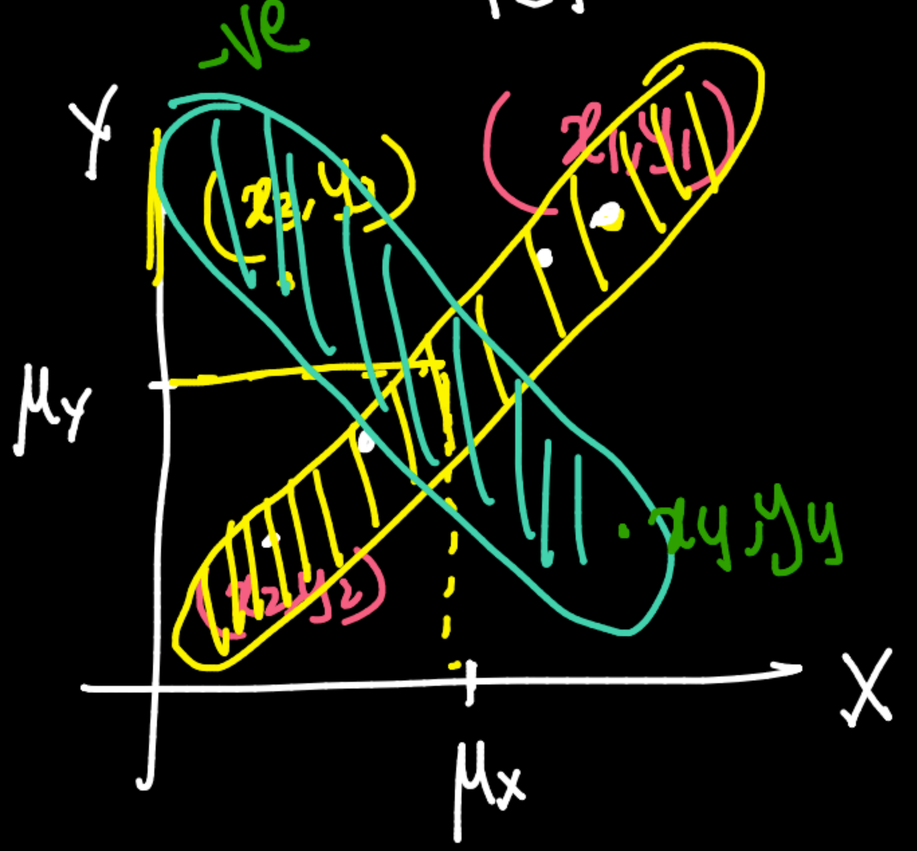

(h)
(w)
X
Y
(pairs)
1
2
i
n
x_1 y_1
x_2 y_2
x_i y_i
x_n y_n
Data
Co-variance
Q
x ↑ y ↑
x ↓ y ↓
"in general"
μ_x = Sample-mean of x
μ_y = Sample Mean of y

$$\text{Co-var}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$$\text{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)$$

$$\text{Cov}(X,X) = \text{Var}(X)$$

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$



(x_3, y_3) : (-ve) (+ve)

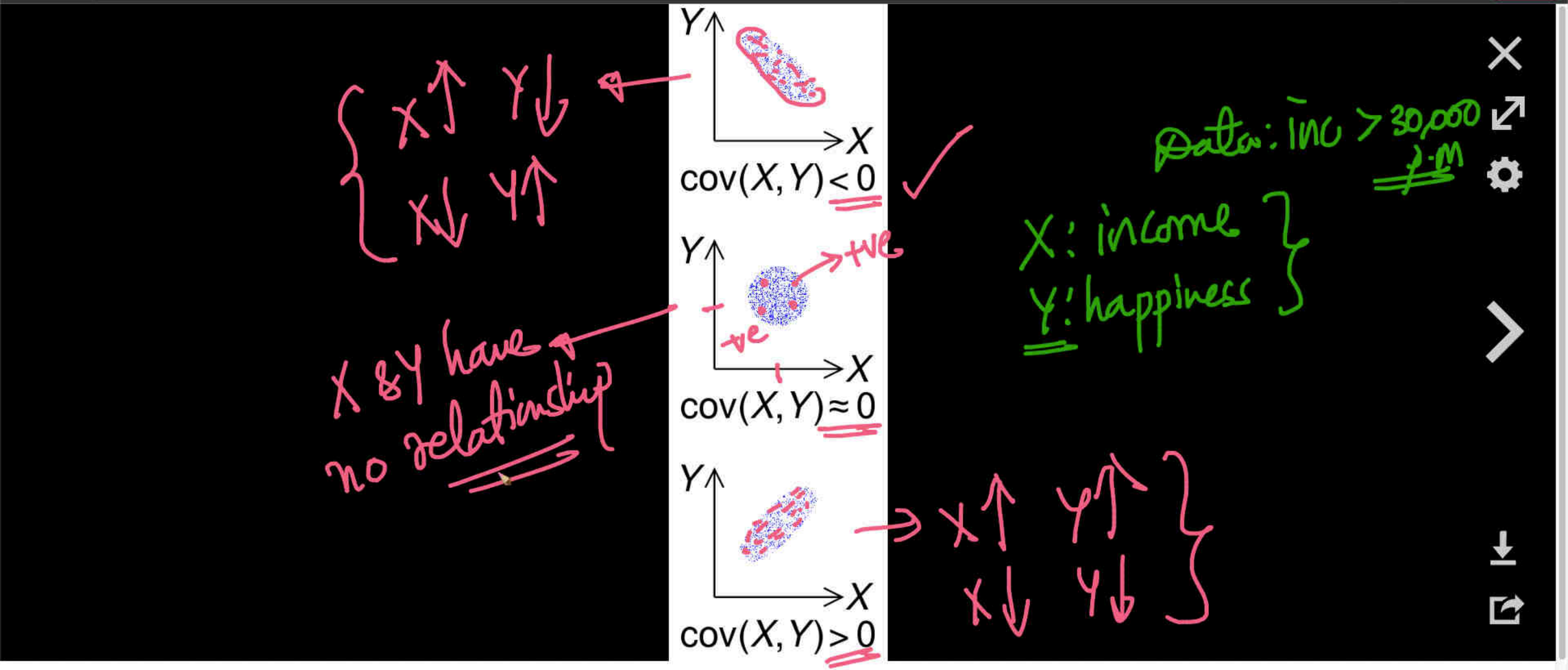
-ve

(x_1, y_1) : +ve · +ve

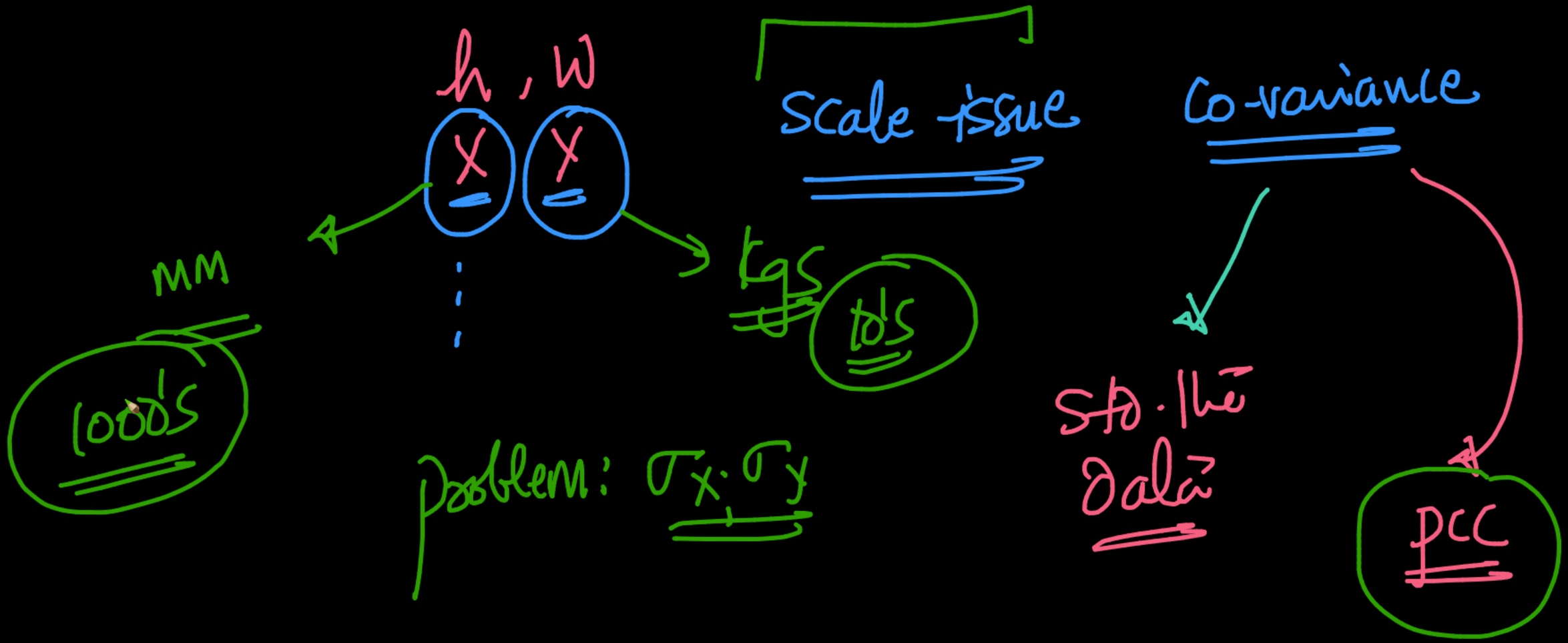
+ve

(x_2, y_2) : -ve · -ve

(+ve)



The sign of the covariance of two random variables X and Y


h , w
X
X
Scale issue
Co-variance
MM
1000's
kgs
lbs
Problem: $\sigma_x \cdot \sigma_y$
Std. like data
PCC

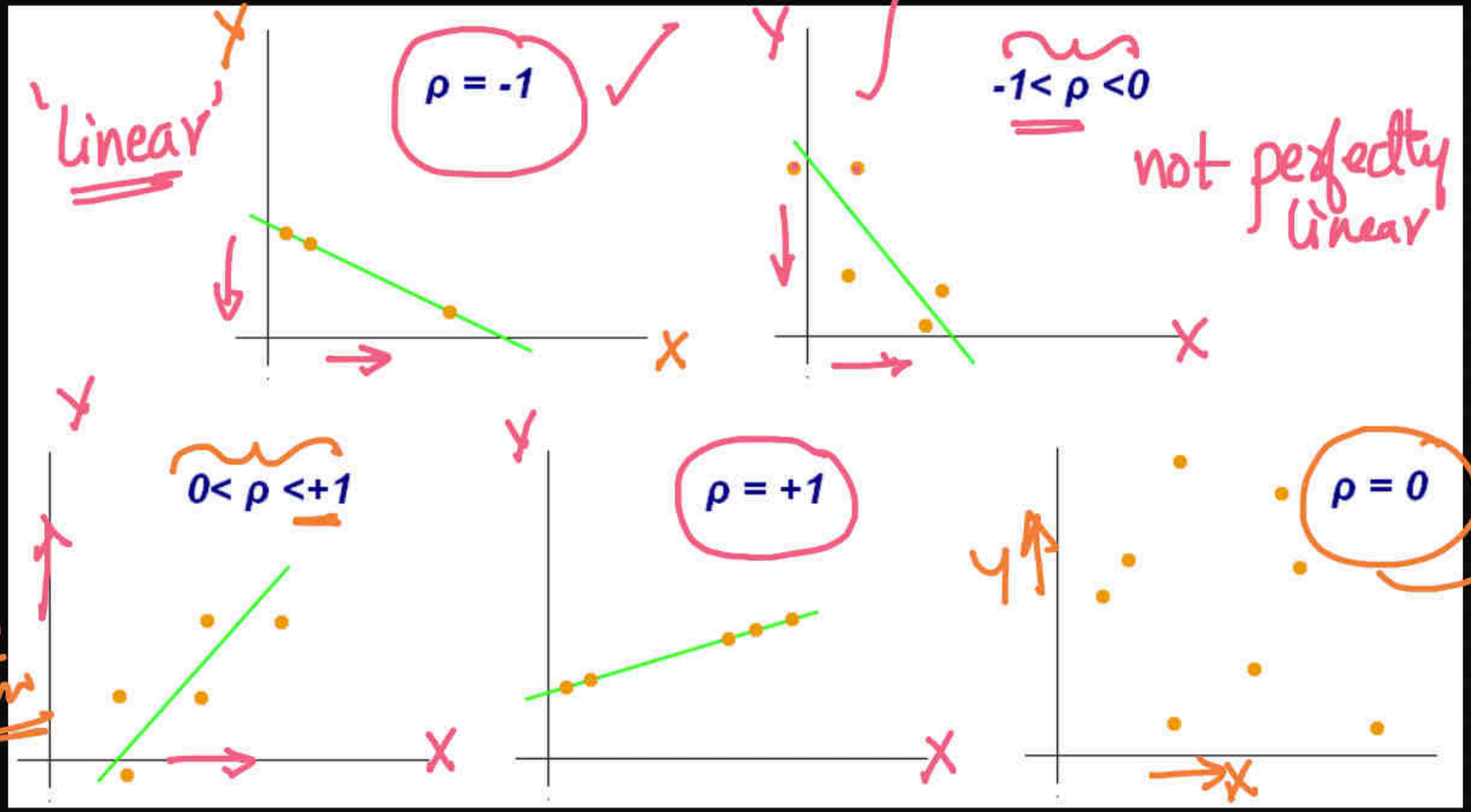
P.C.C

(Very widely)

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

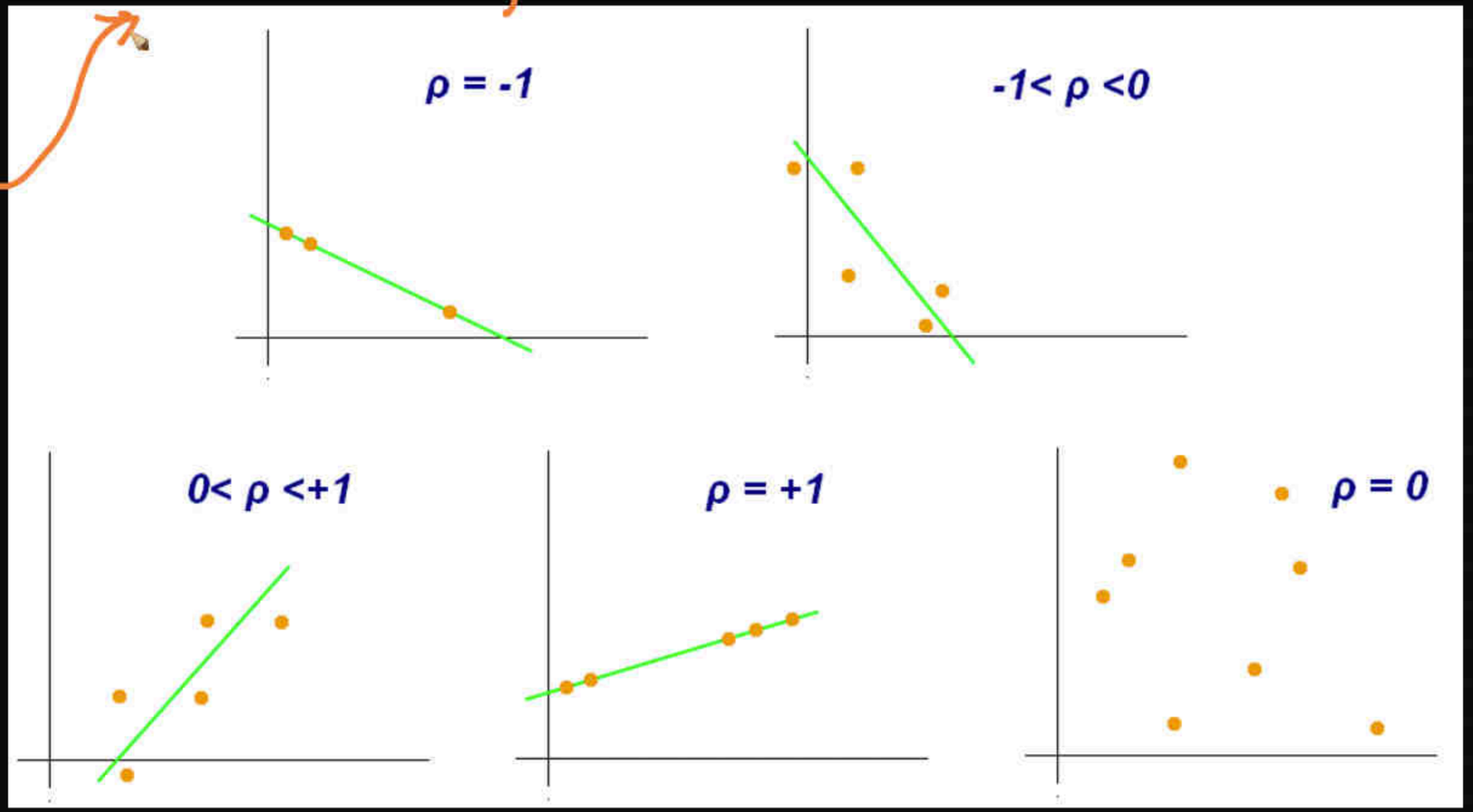
Corrects issues with std. dev

[to be continued....]



P.C.C → relationship is linear

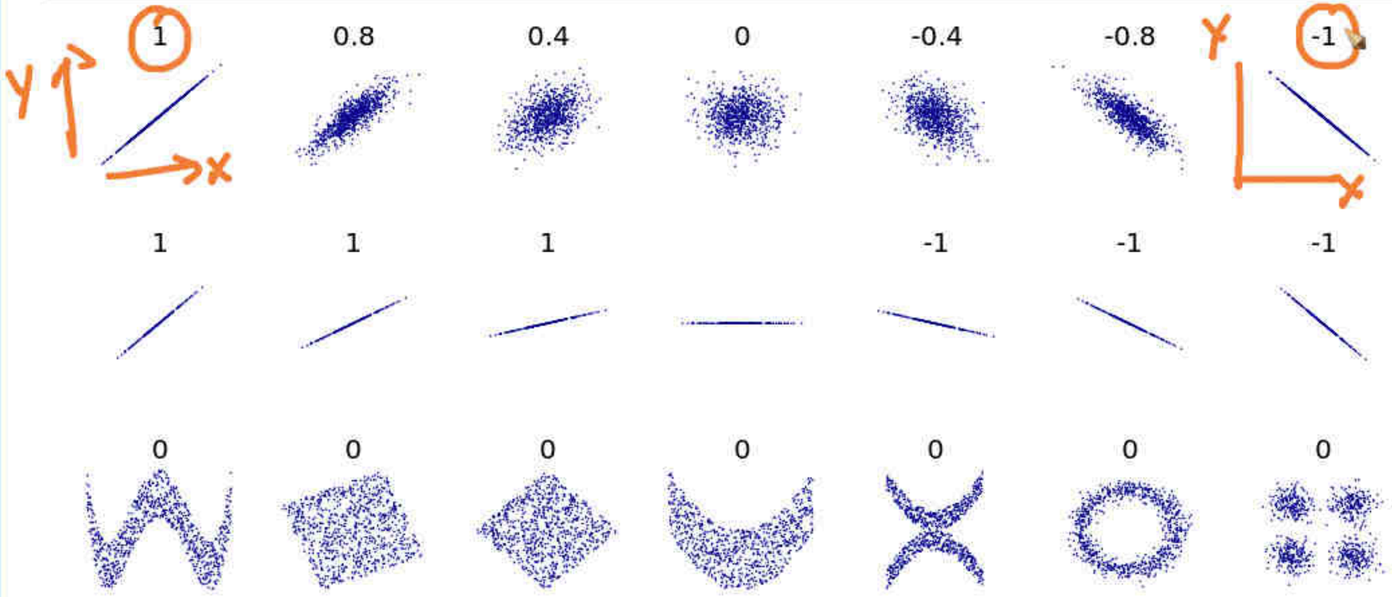
ρ



File:Correlation examples2.svg

From Wikipedia, the free encyclopedia

File | File history | File usage | Global file usage | Metadata



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

Open in Media Viewer

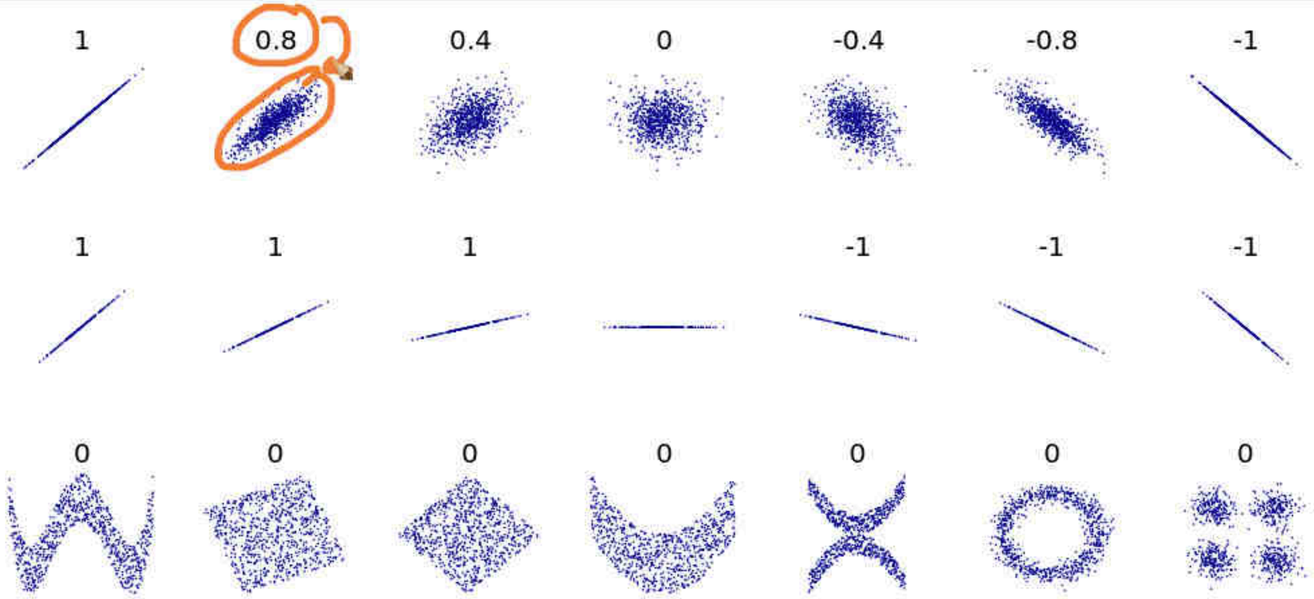
This description page there is shown below.

37 / 38

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia

File File history File usage Global file usage Metadata



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

Open in Media Viewer

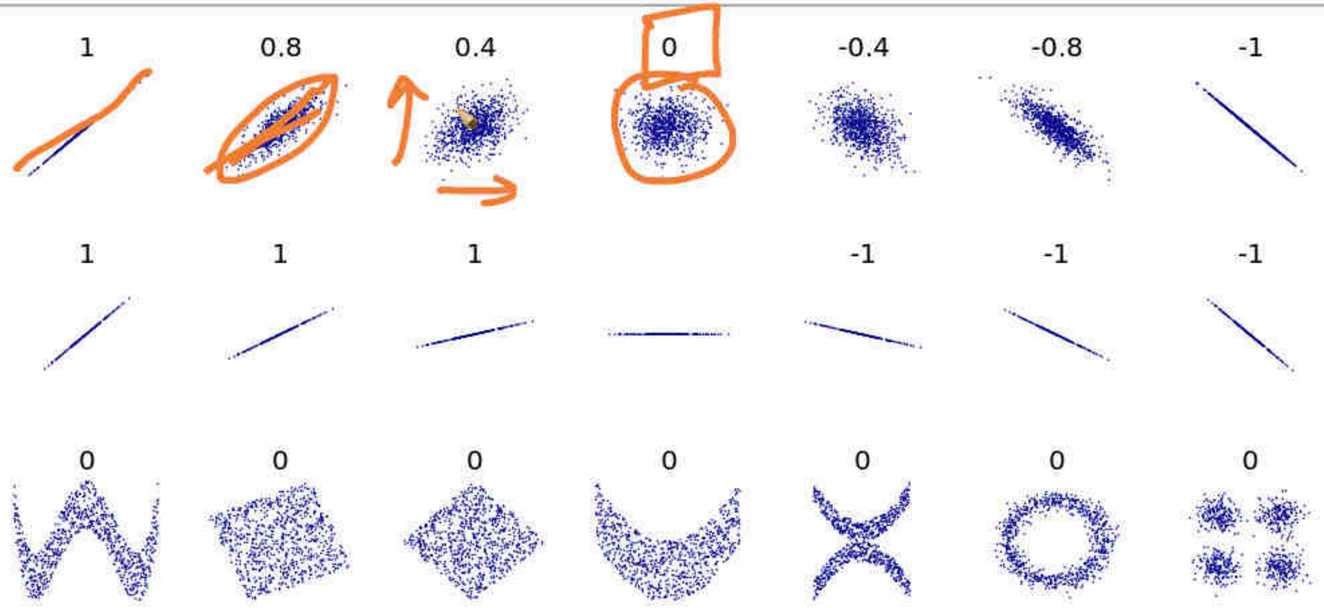
This description page there is shown below.

38 / 39

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia

File File history File usage Global file usage Metadata



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

Open in Media Viewer

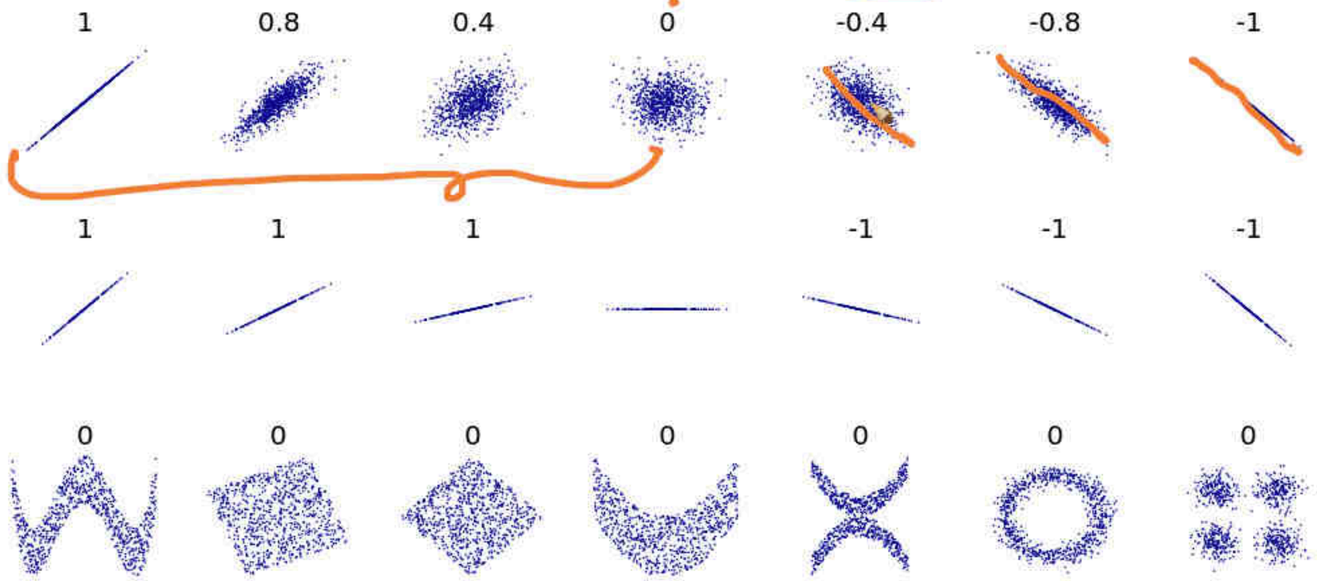
This description page there is shown below.

39 / 40

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia

File File history File usage Global file usage Metadata



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

Open in Media Viewer

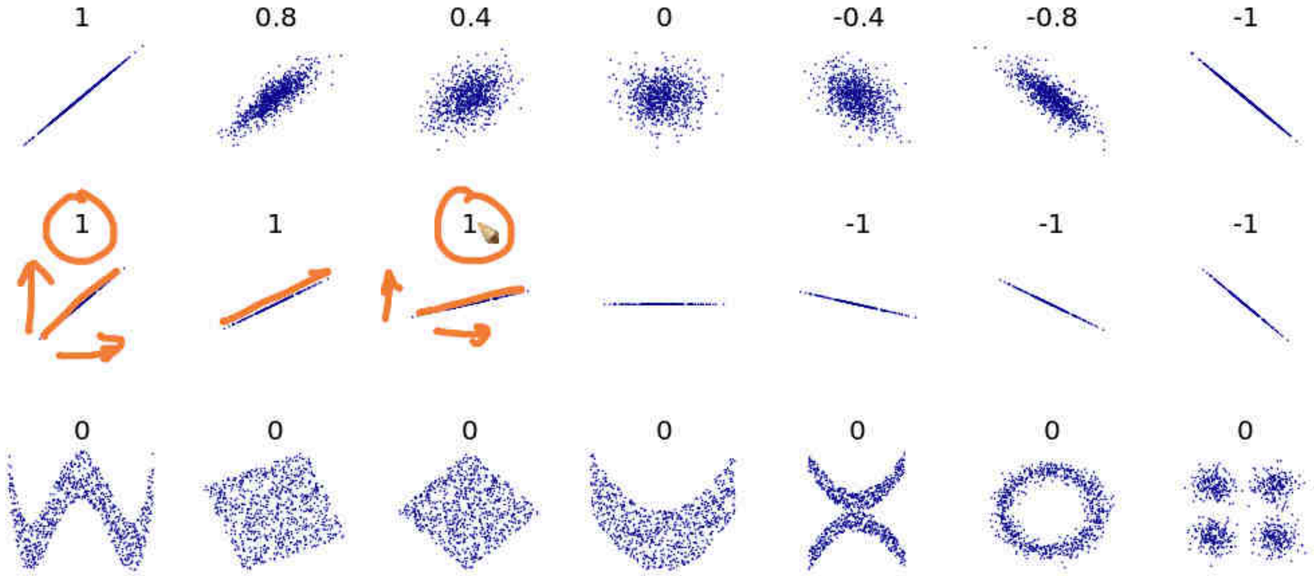
This description page there is shown below.

40 / 41

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia

File | File history | File usage | Global file usage | Metadata



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

Open in Media Viewer

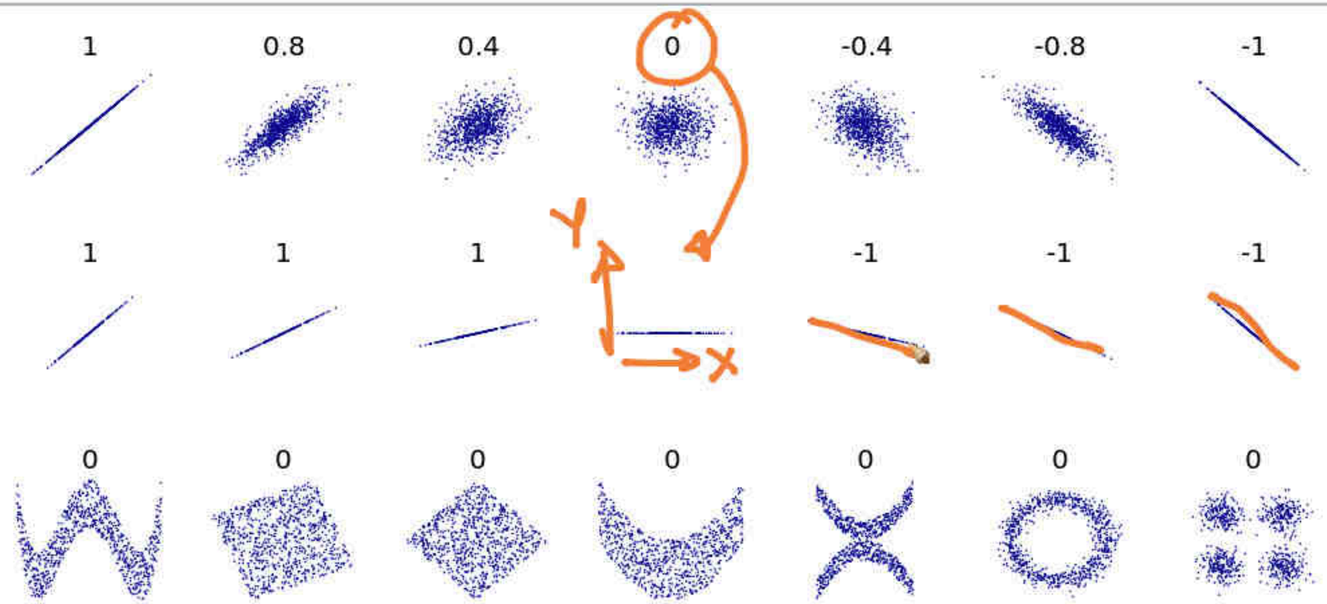
This description page there is shown below.

41 / 42

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia

File | File history | File usage | Global file usage | Metadata



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

Open in Media Viewer

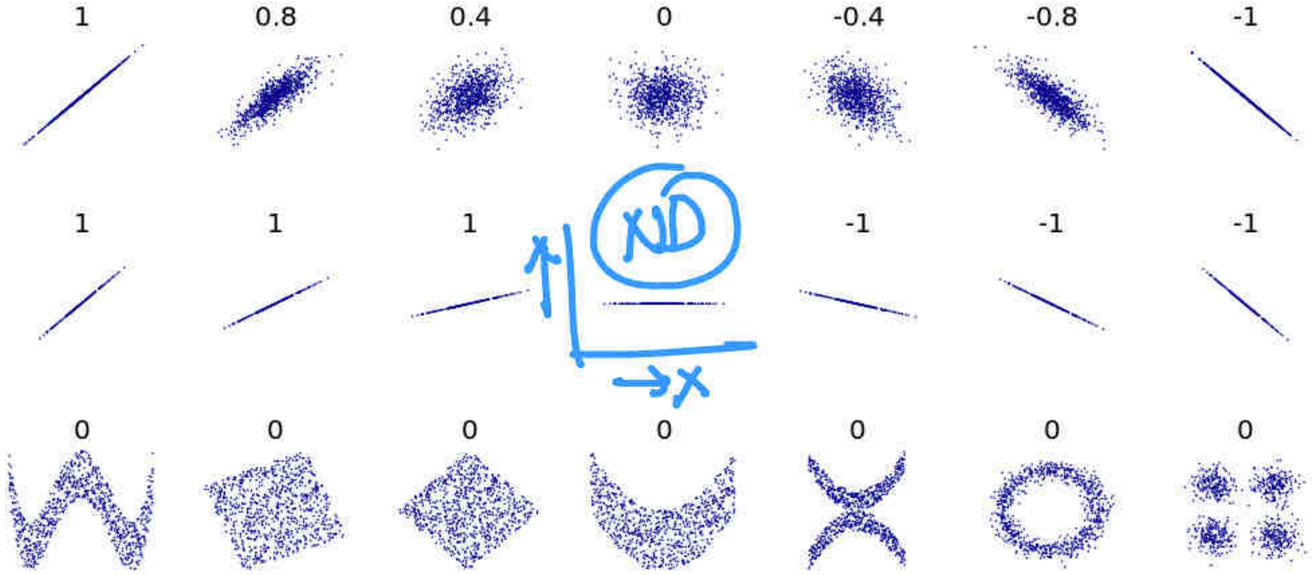
This description page there is shown below.

42 / 43

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia

File | File history | File usage | Global file usage | Metadata



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

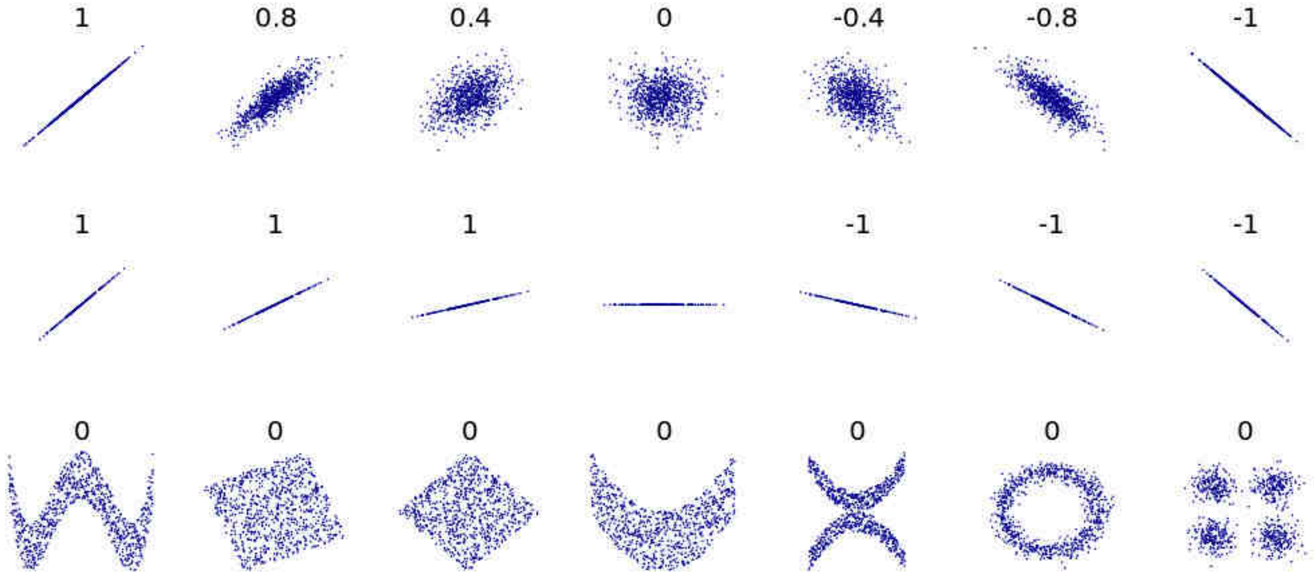
Open in Media Viewer

This description page there is shown below.

43 / 48

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

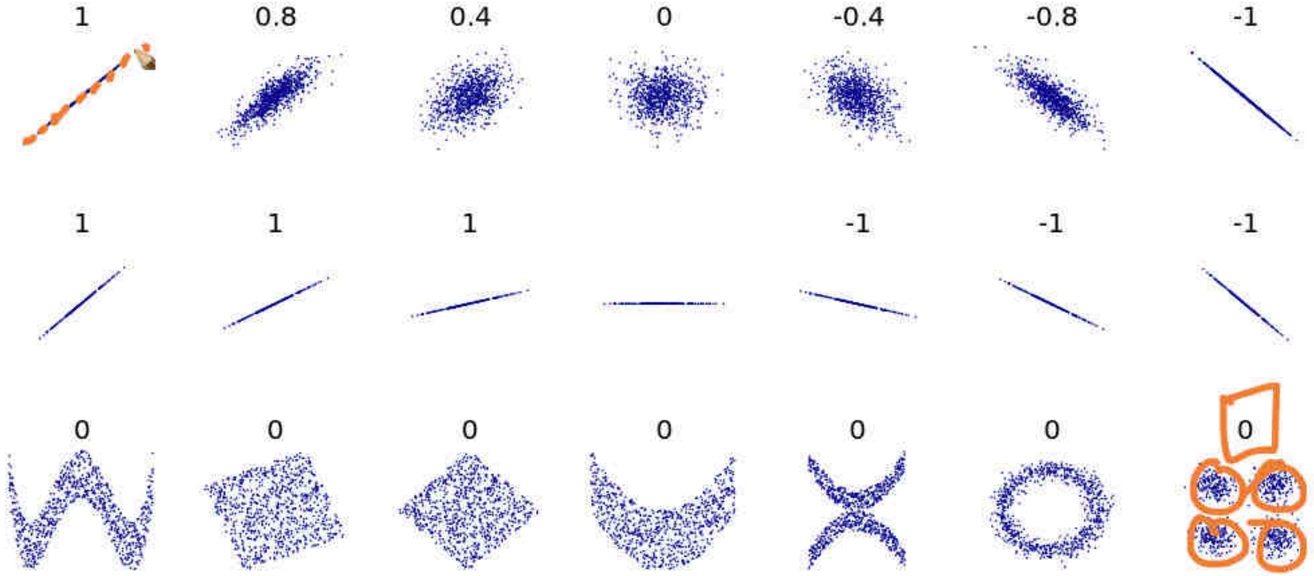
Open in Media Viewer

This description page there is shown below.

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia

File | File history | File usage | Global file usage | Metadata



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

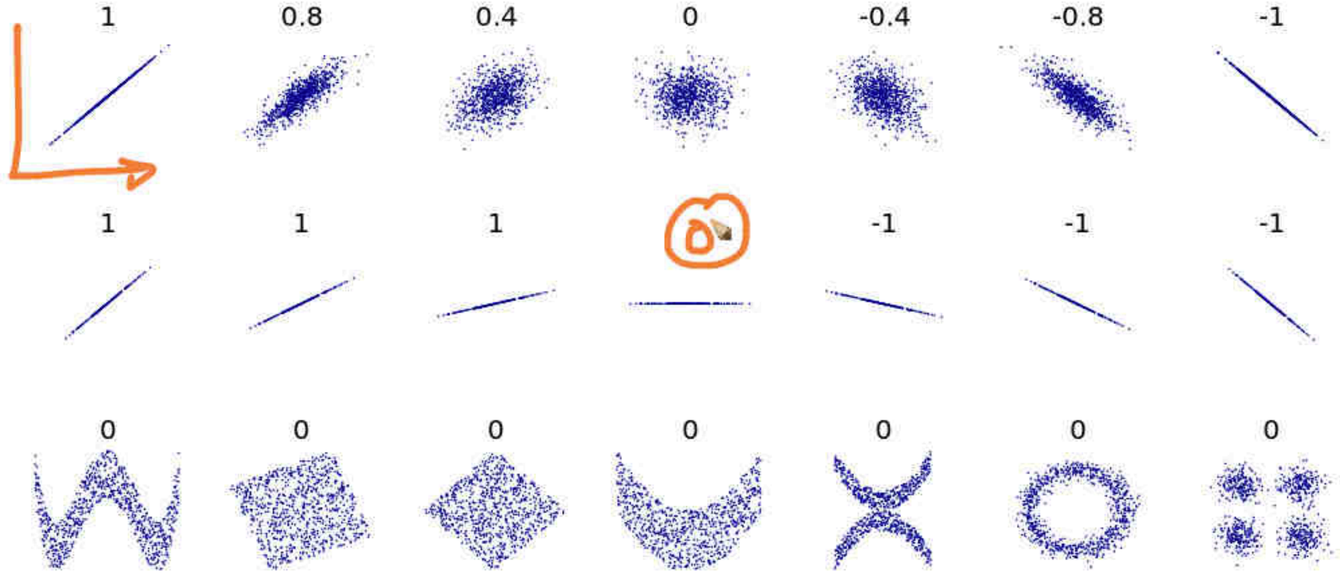
Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

Open in Media Viewer

This description page there is shown below.

File:Correlation examples2.svg

From Wikipedia, the free encyclopedia



Size of this PNG preview of this SVG file: 506 × 231 pixels. Other resolutions: 320 × 146 pixels | 640 × 292 pixels | 1,024 × 467 pixels | 1,280 × 584 pixels | 2,560 × 1,169 pixels.

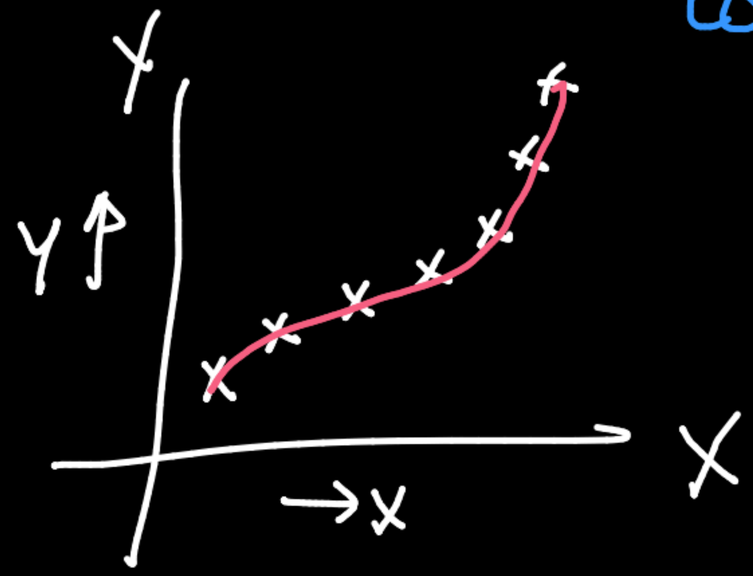
Original file (SVG file, nominally 506 × 231 pixels, file size: 2.18 MB)

Open in Media Viewer

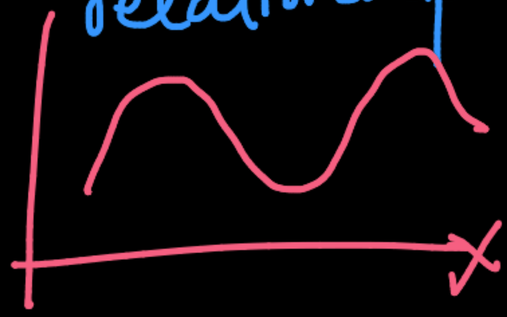
This description page there is shown below.

Linear correlation → PCC

monotonic relationship → hack on PCC



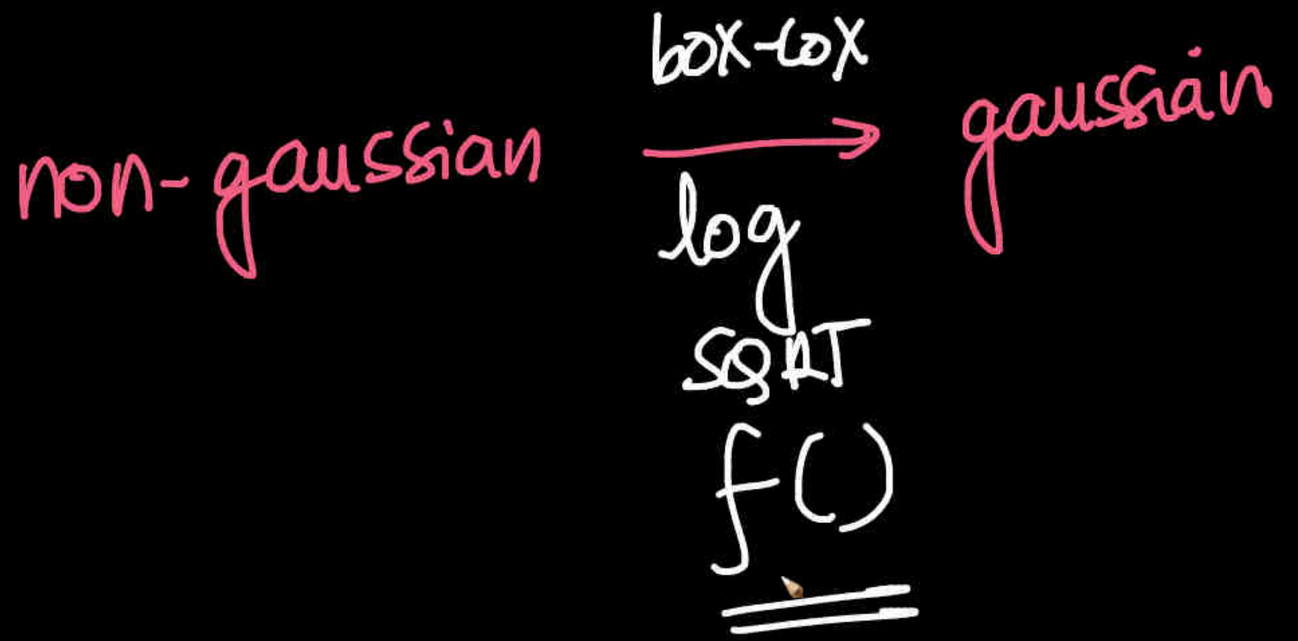
Complex non-linear relationship



→ ML-algos

GBDT / RF / DL

(later)


non-gaussian
box-cox
log
sqrt
f()
gaussian

Spearman- rank- correlation:

(monotonic relationship)

D:

X	Y
10	20
12	30
8	12
6	4
5	1

Sort by

Sort xi's:

5, 6, 8, 10, 12

1 2 3 4 5

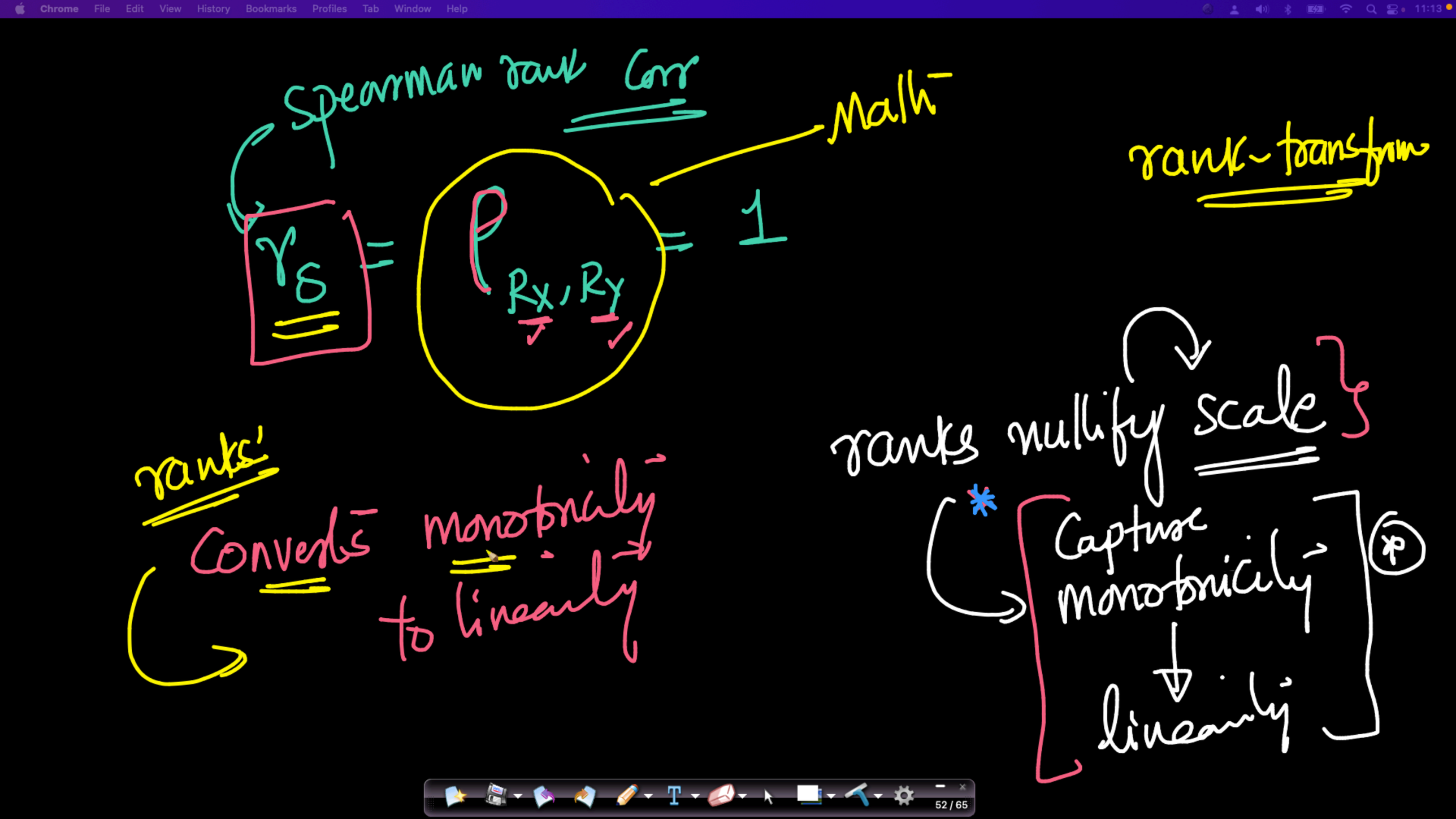
Sort yi's:

1, 4, 12, 20, 50

1 2 3 4 5

D'

Rx	Ry
4	4
5	5
3	3
2	2
1	1


Spearman rank Corr
Malh-
rank-transform
r_s = ρ_{R_X, R_Y} = 1
ranks' converts monotonicity to linearity
ranks nullify scale
Capture monotonicity ⇓ linearity
(*)

Spearman's rank correlation coefficient

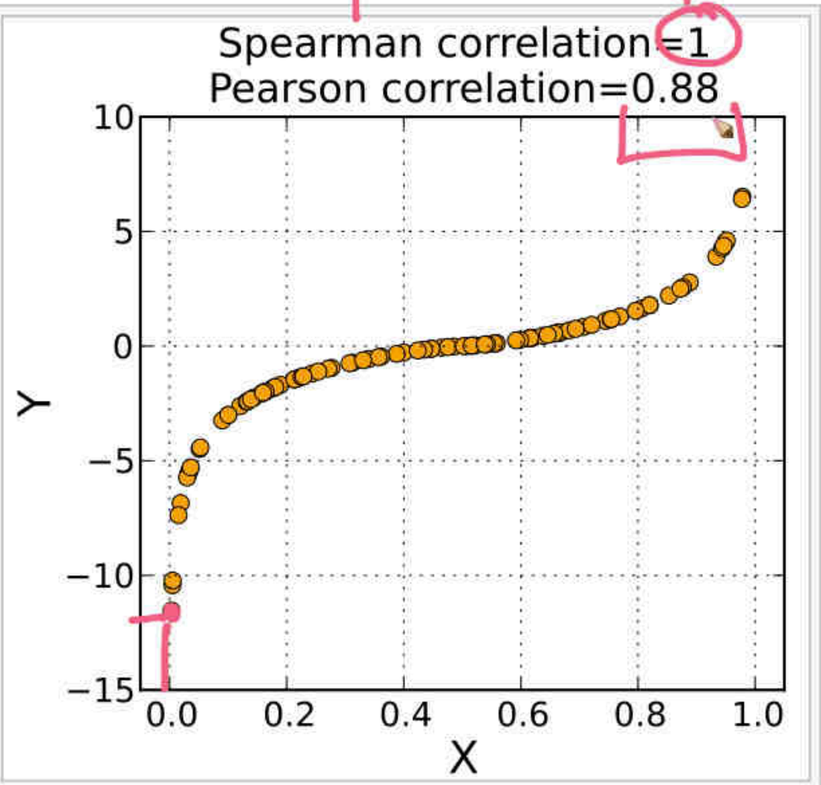
rank - transform

From Wikipedia, the free encyclopedia

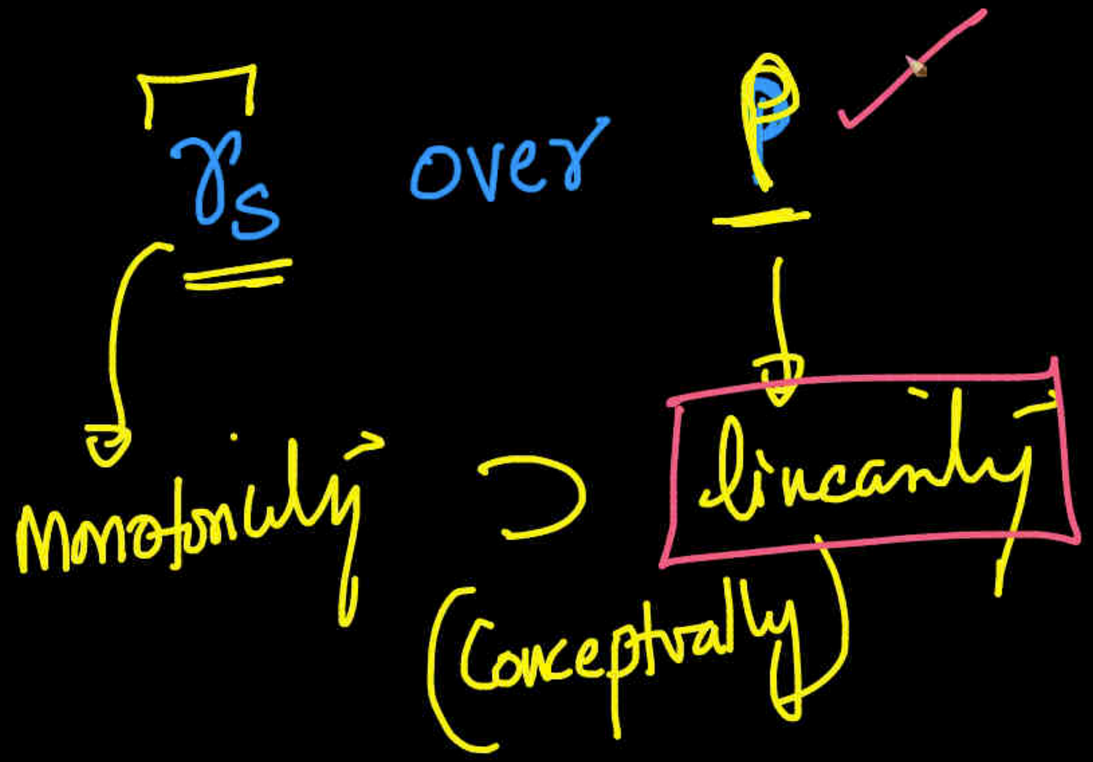
In statistics, **Spearman's rank correlation coefficient** or **Spearman's *ρ***, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. re



A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that l data points with greater *x* values than that of a ven data point will have greater *y* values as well. In

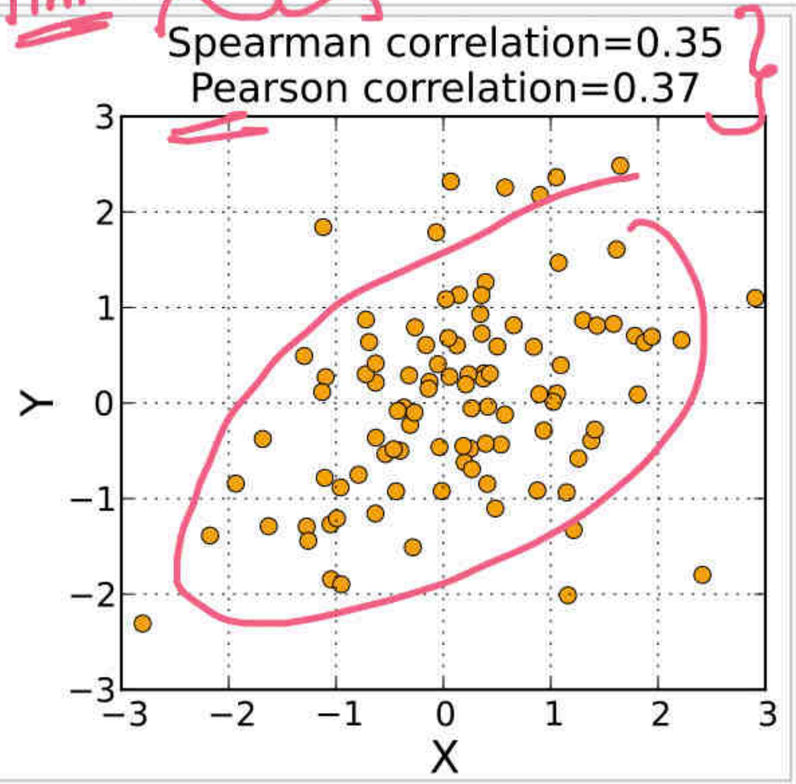

γ_s
over
$\mathbb{P}$
Monotonicity
$\supset$
linearity
(Conceptually)

observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.[1][2] Both Spearman's ρ and Kendall's τ can be formulated as special cases of a more general correlation coefficient.

Contents [hide]

contrast, this does not give a perfect Pearson correlation.

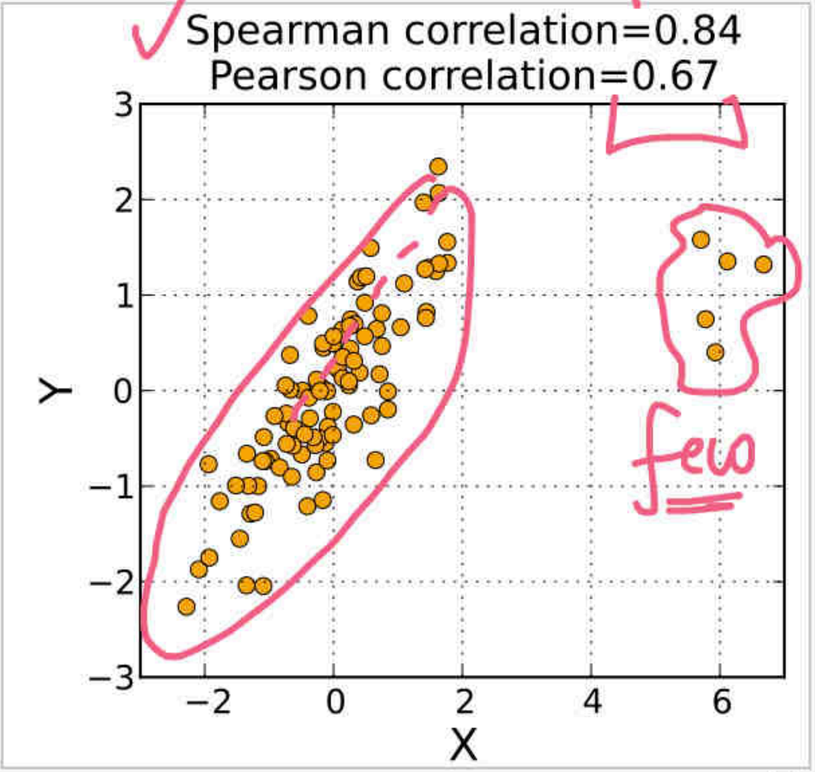


When the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.

Spearman correlation=0.84

Português
中文
18 more
Edit links

and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.



The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's ρ limits the outlier to the value of its rank.

Definition and calculation [edit]

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables.[3]

For a sample of size *n*, the *n* raw scores X_i, Y_i are converted to ranks $\mathrm{R}(X_i), \mathrm{R}(Y_i)$, and r_s is computed as

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}},$$

where

ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,

$\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))$ is the covariance of the rank variables,

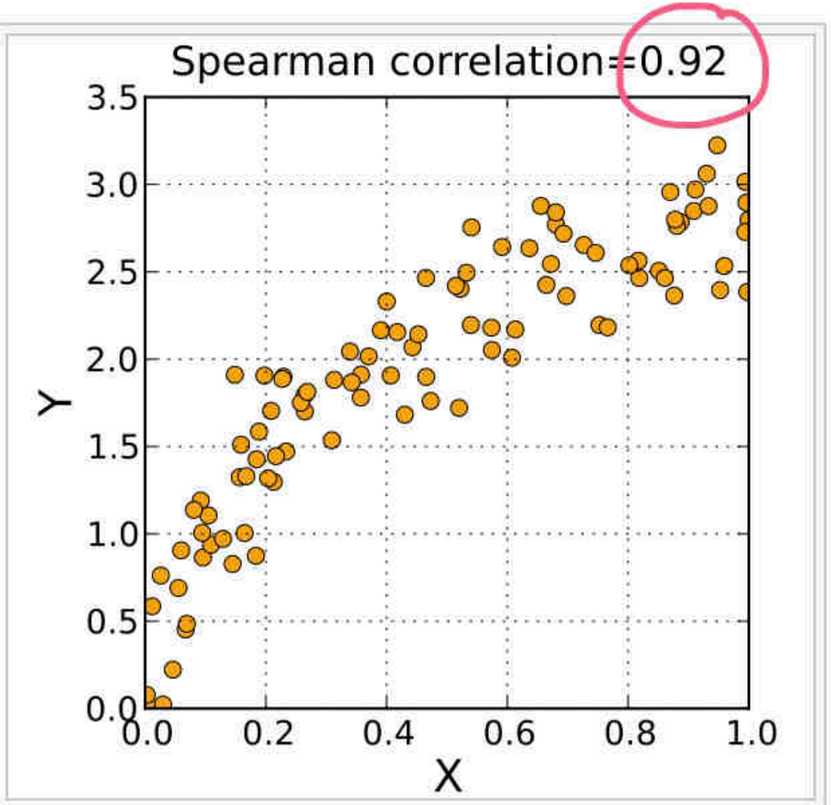
$\sigma_{\mathrm{R}(X)}$ and $\sigma_{\mathrm{R}(Y)}$ are the standard deviations of the rank variables.

Only if all *n* ranks

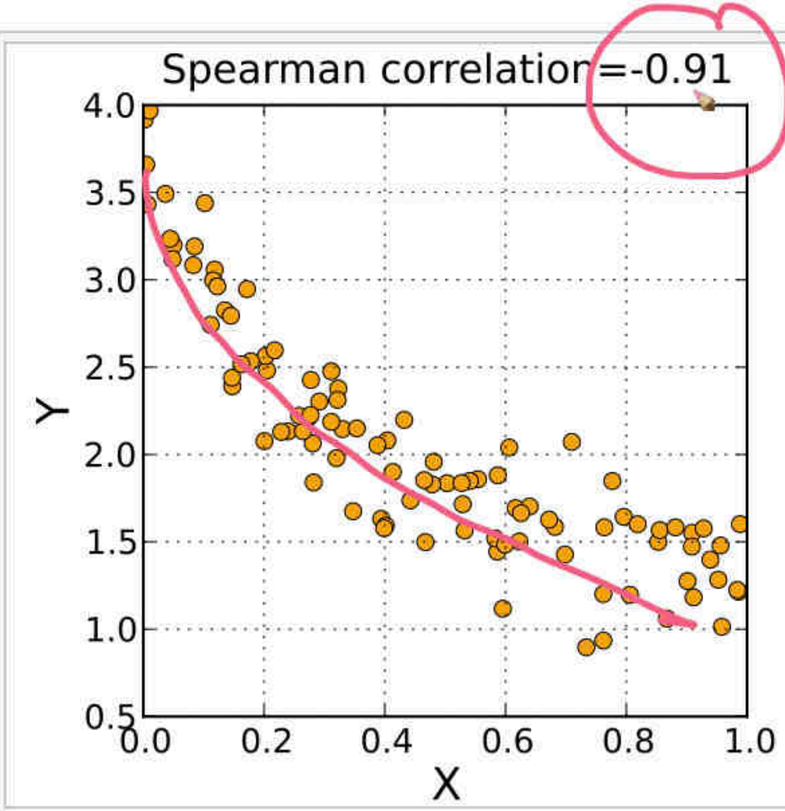
Interpretation [edit]

The sign of the Spearman correlation indicates the direction of association between *X* (the independent variable) and *Y* (the dependent variable). If *Y* tends to increase when *X* increases, the Spearman correlation coefficient is positive. If *Y* tends

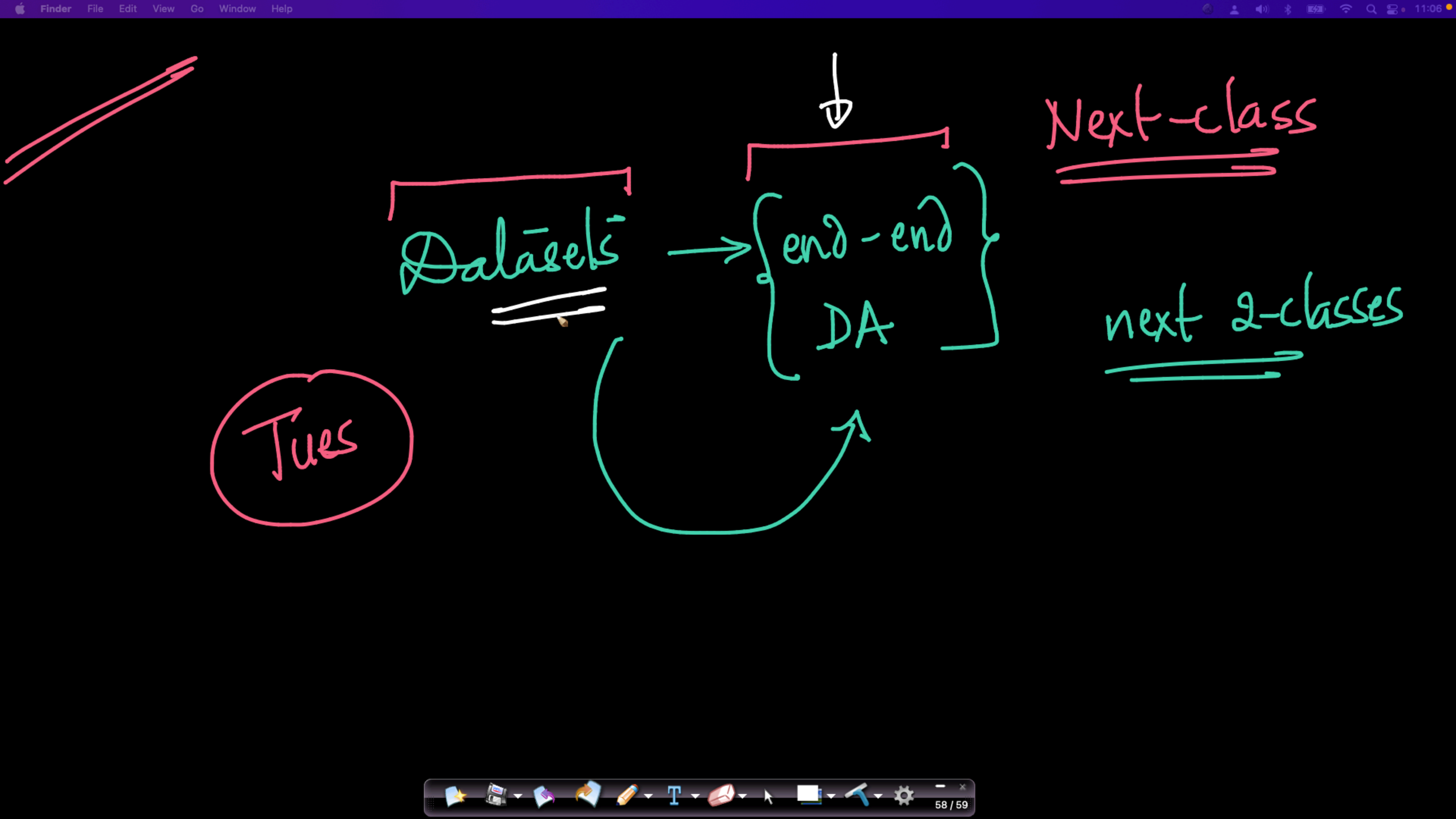
Positive and negative Spearman rank correlations

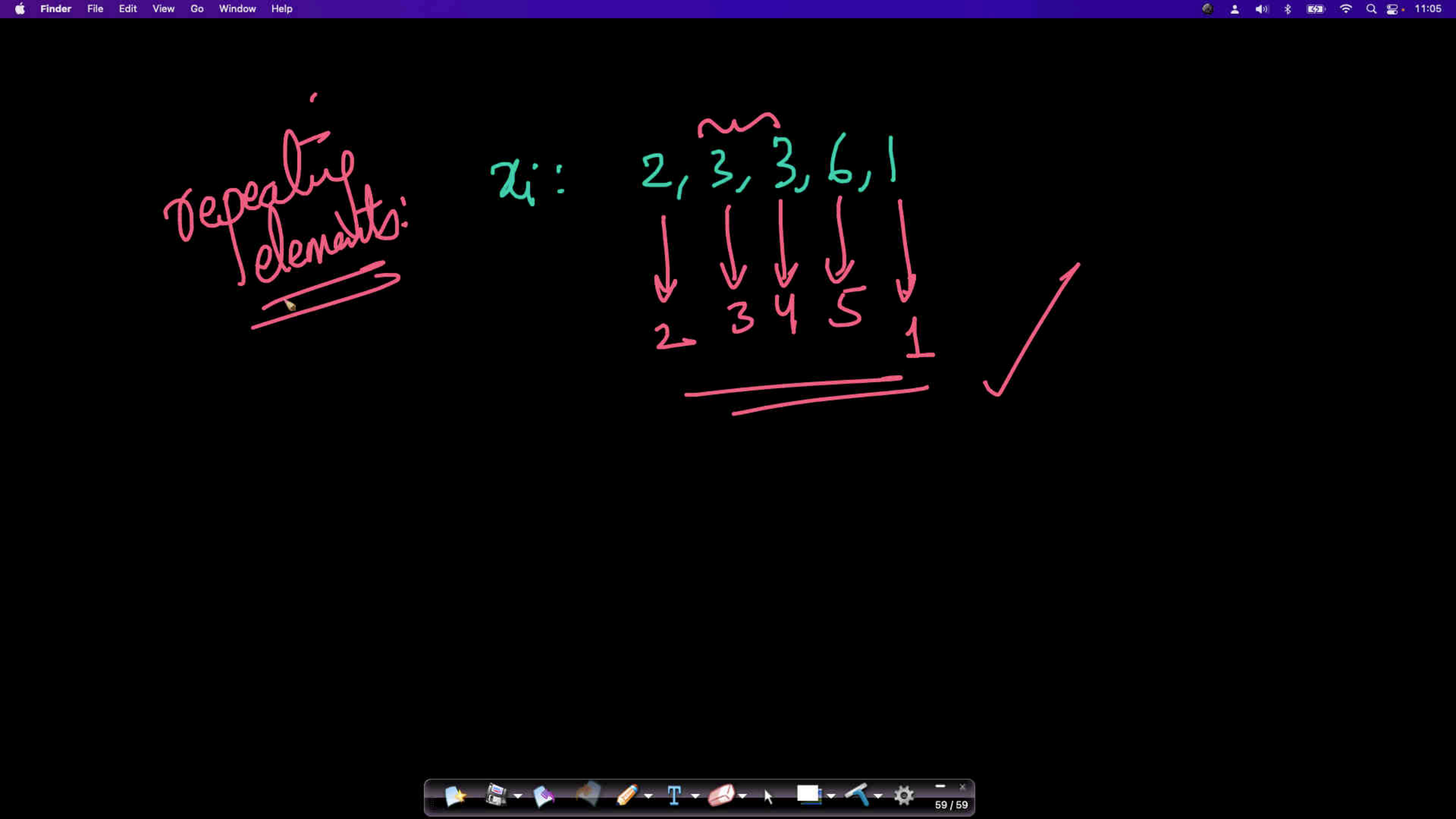


A positive Spearman correlation coefficient corresponds to an increasing monotonic trend between *X* and *Y*.



A negative Spearman correlation coefficient corresponds to a decreasing monotonic trend between *X* and *Y*.


Datasets
end-end
DA
Next-class
next 2-classes
Tues


Repeating elements:
x_i : 2, 3, 3, 6, 1
2 3 4 5 1

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
Deutsch
Español
Français
한국어
日本語
Português
Русский
Tiếng Việt
中文
11 more
Edit links

Data [edit]

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third g ould have a different regression line (a robust

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
Deutsch
Español
Français
한국어
日本語
Português
Русский
Tiếng Việt
中文
11 more
Edit links

Data [edit]

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third g ould have a different regression line (a robust

61 / 61

From Wikipedia, the free encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

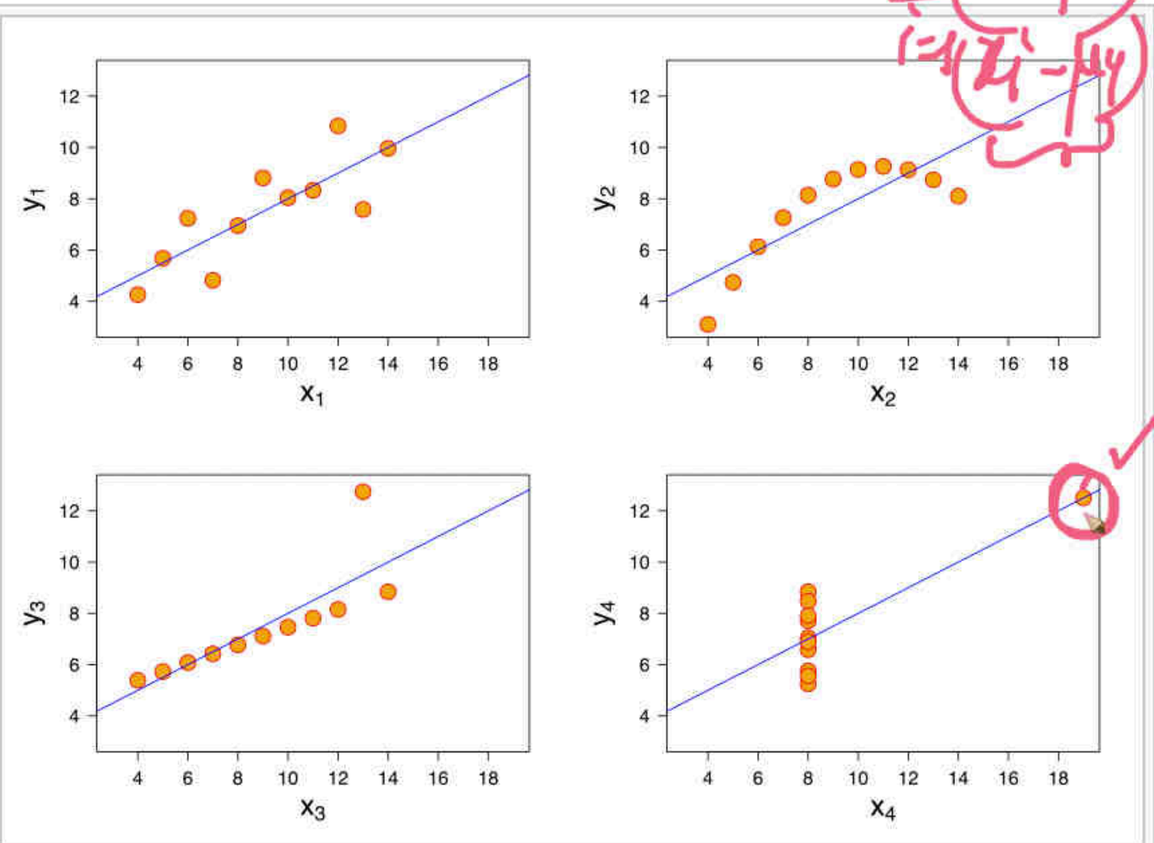
Contribute

Help
Learn to edit
Community portal
Recent changes
Upload file

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."[1]



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

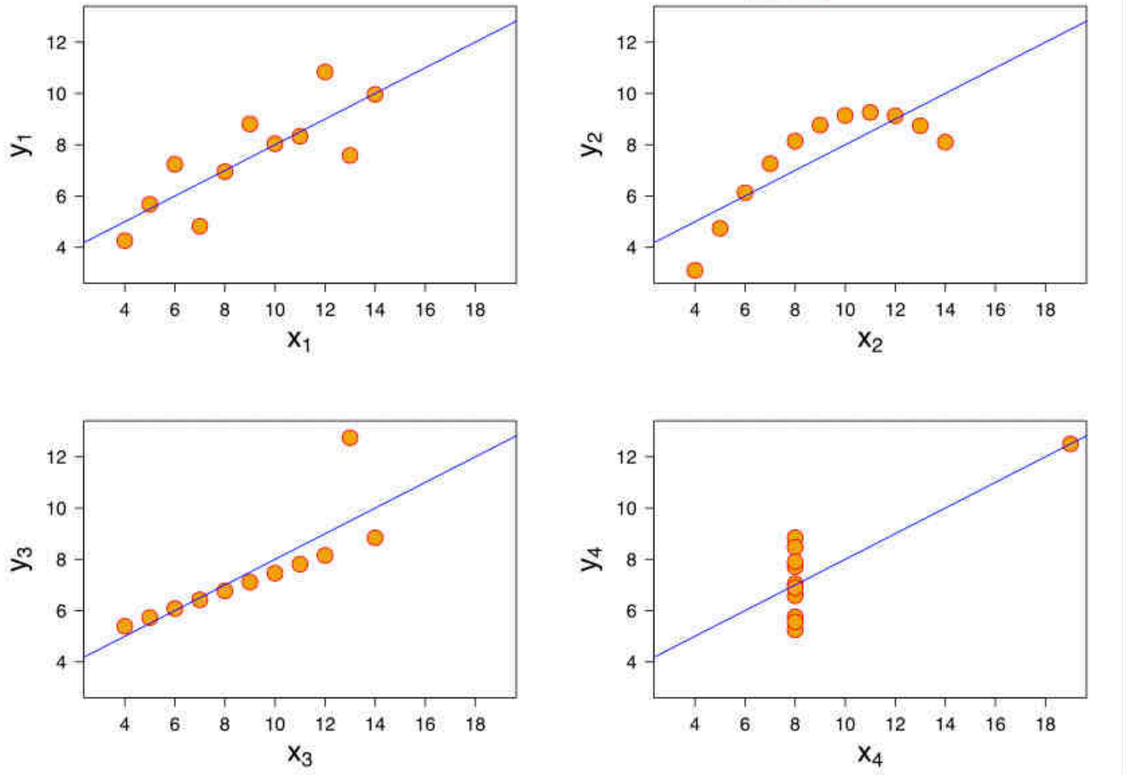
Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

From Wikipedia, the free encyclopedia

plot - datasaurus

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."[1]



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Contents [hide]

63 / 63

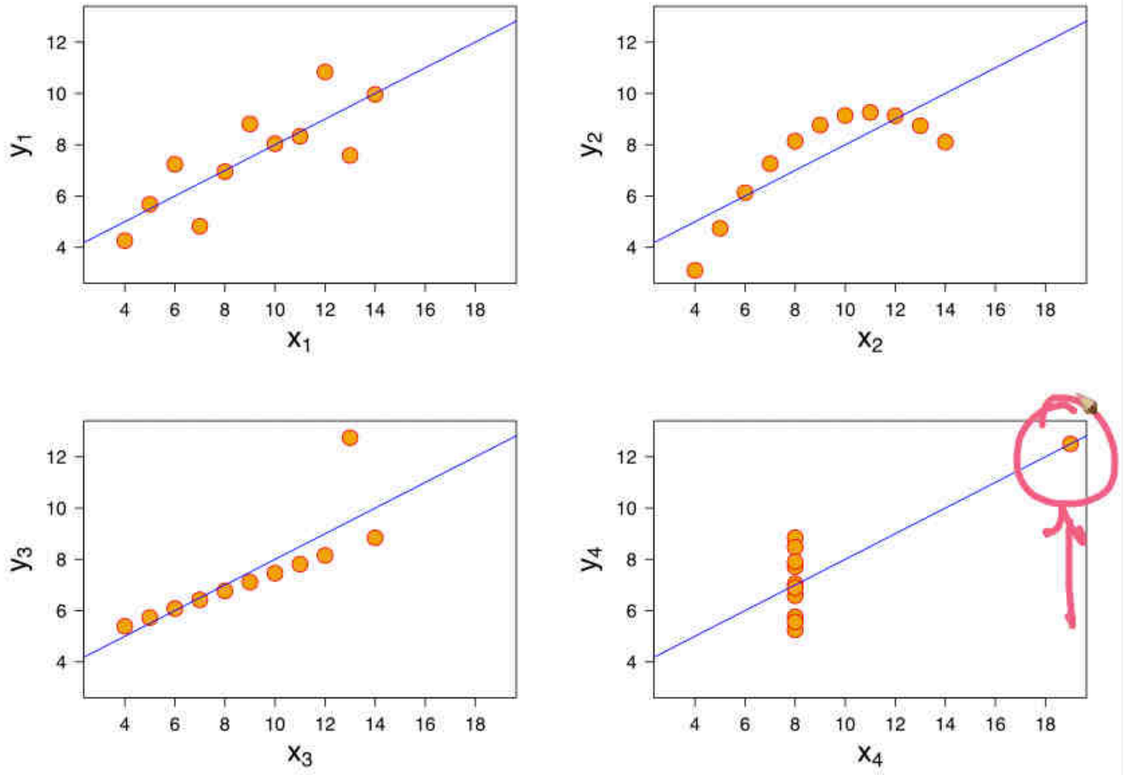
Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

From Wikipedia, the free encyclopedia

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."[1]



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Contents [hide]

64 / 64

SRCC = r_s → test/check for Monotonicity

ρ → " " linearity

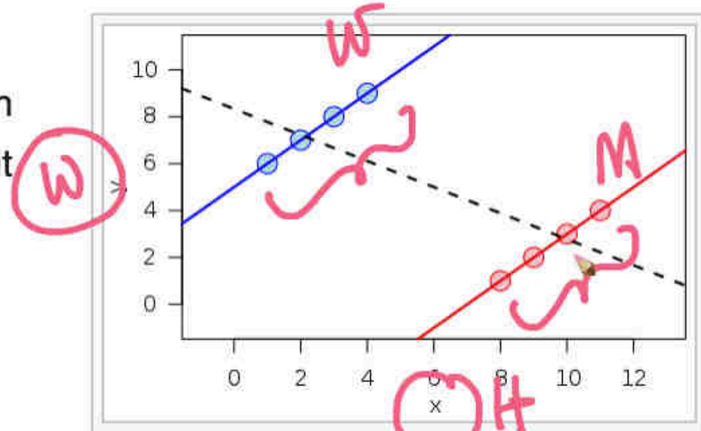
multi-collinearity → linear regression (ML)

Simpson's paradox

From Wikipedia, the free encyclopedia

See also: Misuse of statistics

Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined. This result is often encountered in social-science and medical-science statistics,[1][2][3] and is particularly problematic when frequency data is unduly given causal interpretations.[4] The paradox can be resolved when confounding variables and causal relations are appropriately addressed in the statistical modeling.[4][5] Simpson's paradox has been used to illustrate the kind of misleading results that the misuse of statistics can generate.[6][7]



Simpson's paradox for quantitative data: a positive trend (——, ——) appears for two separate groups, whereas a negative trend (- - - -) appears when the groups are combined.

Edward H. Simpson first described this phenomenon in a technical paper in 1951,[8] but the statisticians Karl Pearson et al., in 1899,[9] and Udny Yule, in 1903,[10] had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.[11] It is also referred to as **Simpson's reversal**, **Yule–S**

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes

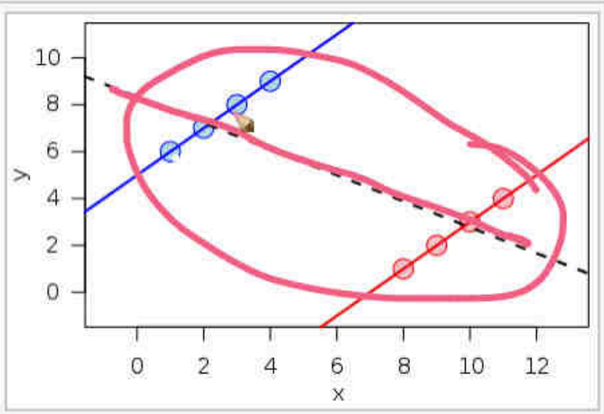
Simpson's paradox

From Wikipedia, the free encyclopedia

See also: Misuse of statistics

Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined. This result is often encountered in social-science and medical-science statistics,[1][2][3] and is particularly problematic when frequency data is unduly given causal interpretations.[4] The paradox can be resolved when confounding variables and causal relations are appropriately addressed in the statistical modeling.[4][5] Simpson's paradox has been used to illustrate the kind of misleading results that the misuse of statistics can generate.[6][7]

Edward H. Simpson first described this phenomenon in a technical paper in 1951,[8] but the statisticians Karl Pearson et al., in 1899,[9] and Udny Yule, in 1903,[10] had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.[11] It is also referred to as **Simpson's reversal**, **Yule–S**



Simpson's paradox for quantitative data: a positive trend (——, ——) appears for two separate groups, whereas a negative trend (- - - -) appears when the groups are combined.

Simpson's paradox

From Wikipedia, the free encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

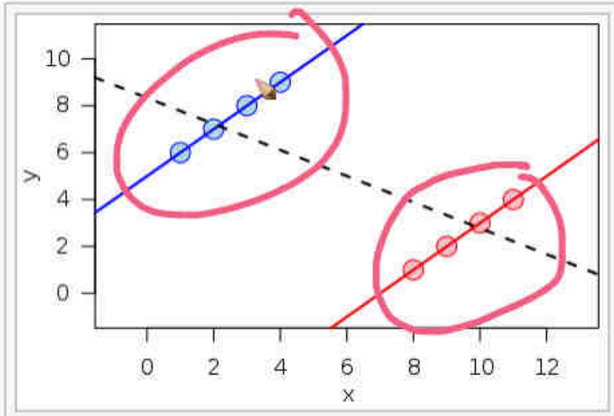
Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes

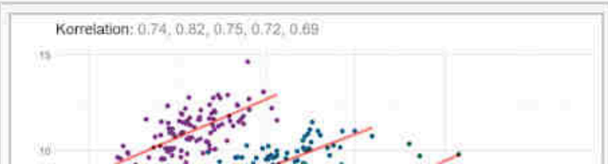
See also: Misuse of statistics

Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined. This result is often encountered in social-science and medical-science statistics,[1][2][3] and is particularly problematic when frequency data is unduly given causal interpretations.[4] The paradox can be resolved when confounding variables and causal relations are appropriately addressed in the statistical modeling.[4][5] Simpson's paradox has been used to illustrate the kind of misleading results that the misuse of statistics can generate.[6][7]

Edward H. Simpson first described this phenomenon in a technical paper in 1951,[8] but the statisticians Karl Pearson et al., in 1899,[9] and Udny Yule, in 1903,[10] had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.[11] It is also referred to as **Simpson's reversal**, **Yule–S**



Simpson's paradox for quantitative data: a positive trend (——, ——) appears for two separate groups, whereas a negative trend (- - - -) appears when the groups are combined.



Simpson's paradox

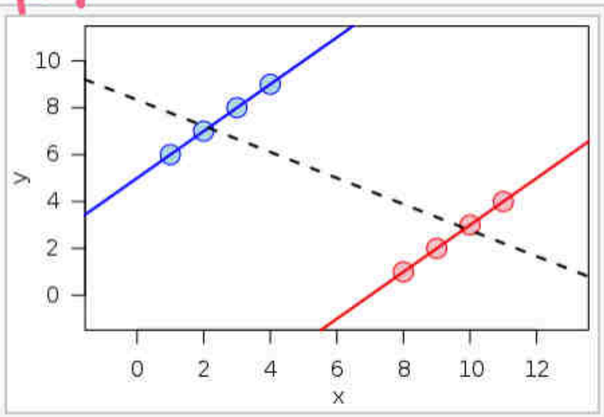
Cross-sections/ each gp may show diff beh as compared total data

From Wikipedia, the free encyclopedia

See also: Misuse of statistics

Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined. This result is often encountered in social-science and medical-science statistics,[1][2][3] and is particularly problematic when frequency data is unduly given causal interpretations.[4] The paradox can be resolved when confounding variables and causal relations are appropriately addressed in the statistical modeling.[4][5] Simpson's paradox has been used to illustrate the kind of misleading results that the misuse of statistics can generate.[6][7]

Edward H. Simpson first described this phenomenon in a technical paper in 1951,[8] but the statisticians Karl Pearson et al., in 1899,[9] and Udny Yule, in 1903,[10] had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.[11] It is also referred to as **Simpson's reversal**, **Yule–S** [12]



Simpson's paradox for quantitative data: a positive trend (——, ——) appears for two separate groups, whereas a negative trend (- - - -) appears when the groups are combined.

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes

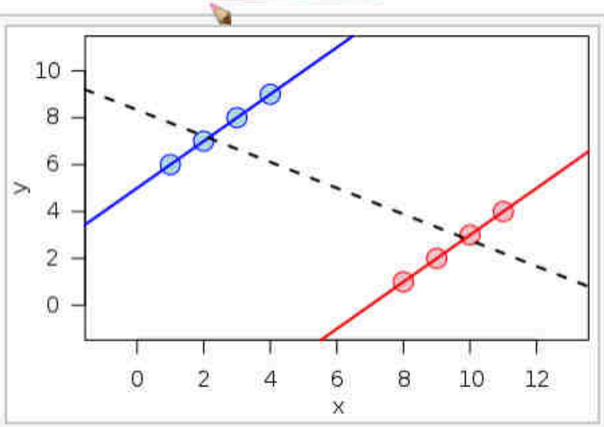
Simpson's paradox

From Wikipedia, the free encyclopedia

See also: Misuse of statistics

Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined. This result is often encountered in social-science and medical-science statistics,[1][2][3] and is particularly problematic when frequency data is unduly given causal interpretations.[4] The paradox can be resolved when confounding variables and causal relations are appropriately addressed in the statistical modeling.[4][5] Simpson's paradox has been used to illustrate the kind of misleading results that the misuse of statistics can generate.[6][7]

Edward H. Simpson first described this phenomenon in a technical paper in 1951,[8] but the statisticians Karl Pearson et al., in 1899,[9] and Udny Yule, in 1903,[10] had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.[11] It is also referred to as **Simpson's reversal**, **Yule–S**



Simpson's paradox for quantitative data: a positive trend (——, ——) appears for two separate groups, whereas a negative trend (- - - -) appears when the groups are combined.

Recent changes
Upload file

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export

Download as PDF
Printable version

In other projects

Wikimedia Commons

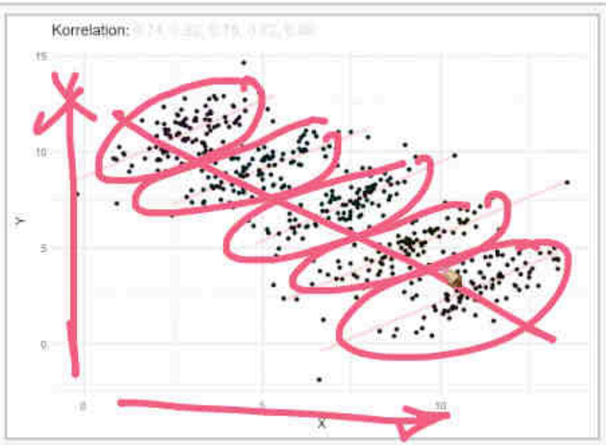
Languages

العربية
Deutsch
Español
Français
日本語
Português

Edward H. Simpson first described this phenomenon in a technical paper in 1951,[8] but the statisticians Karl Pearson et al., in 1899,[9] and Udny Yule, in 1903,[10] had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.[11] It is also referred to as **Simpson's reversal**, **Yule–Simpson effect**, **amalgamation paradox**, or **reversal paradox**.[12]

Contents [hide]

combined.



Visualization of Simpson's paradox on data resembling real-world variability indicates that risk of misjudgment of true causal relationship can be hard to spot

Recent changes
Upload file

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export

Download as PDF
Printable version

In other projects

Wikimedia Commons

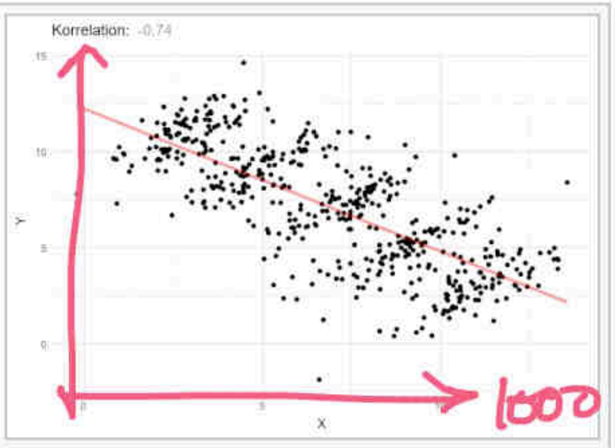
Languages

العربية
Deutsch
Español
Français
日本語
Português

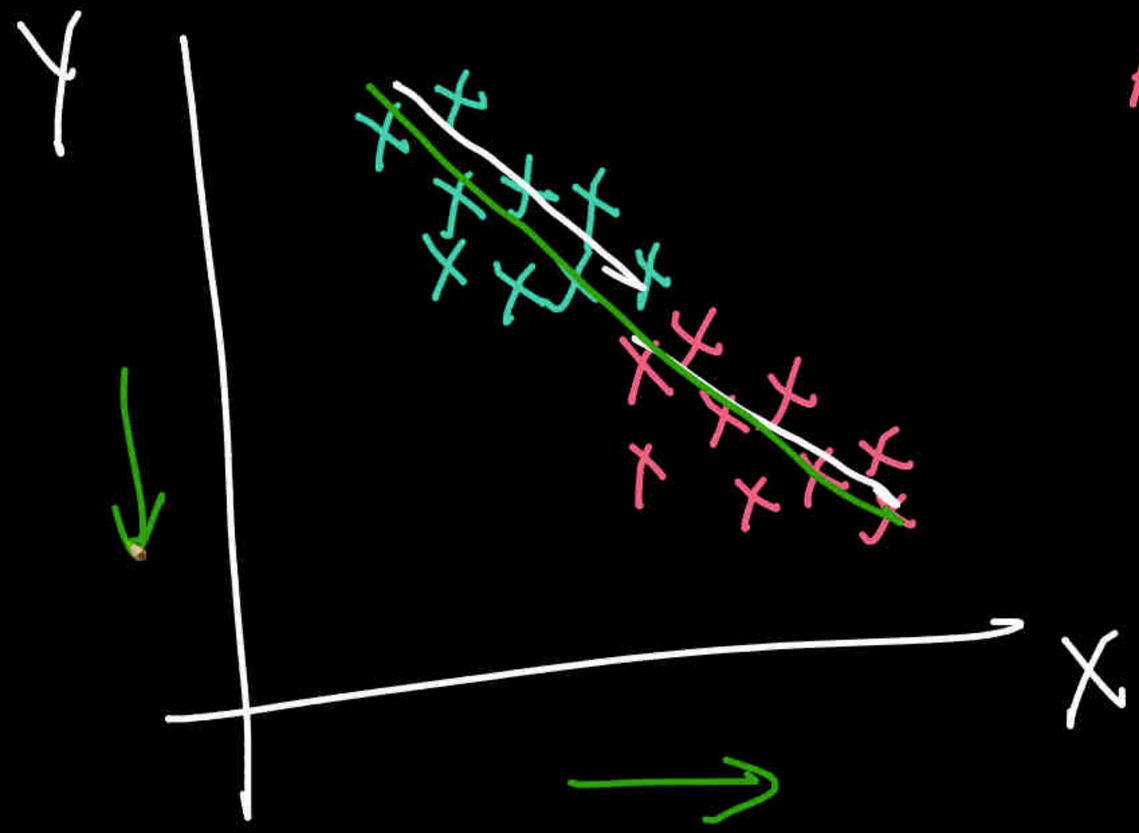
Edward H. Simpson first described this phenomenon in a technical paper in 1951,[8] but the statisticians Karl Pearson et al., in 1899,[9] and Udny Yule, in 1903,[10] had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.[11] It is also referred to as **Simpson's reversal**, **Yule–Simpson effect**, **amalgamation paradox**, or **reversal paradox**.[12]

Contents [hide]

combined.



Visualization of Simpson's paradox on data resembling real-world variability indicates that risk of misjudgment of true causal relationship can be hard to spot



ANOVA
↓
Ha ✓

(plot the data)

Sizes: M, S, L, XL (gp)

height

Task! . . .

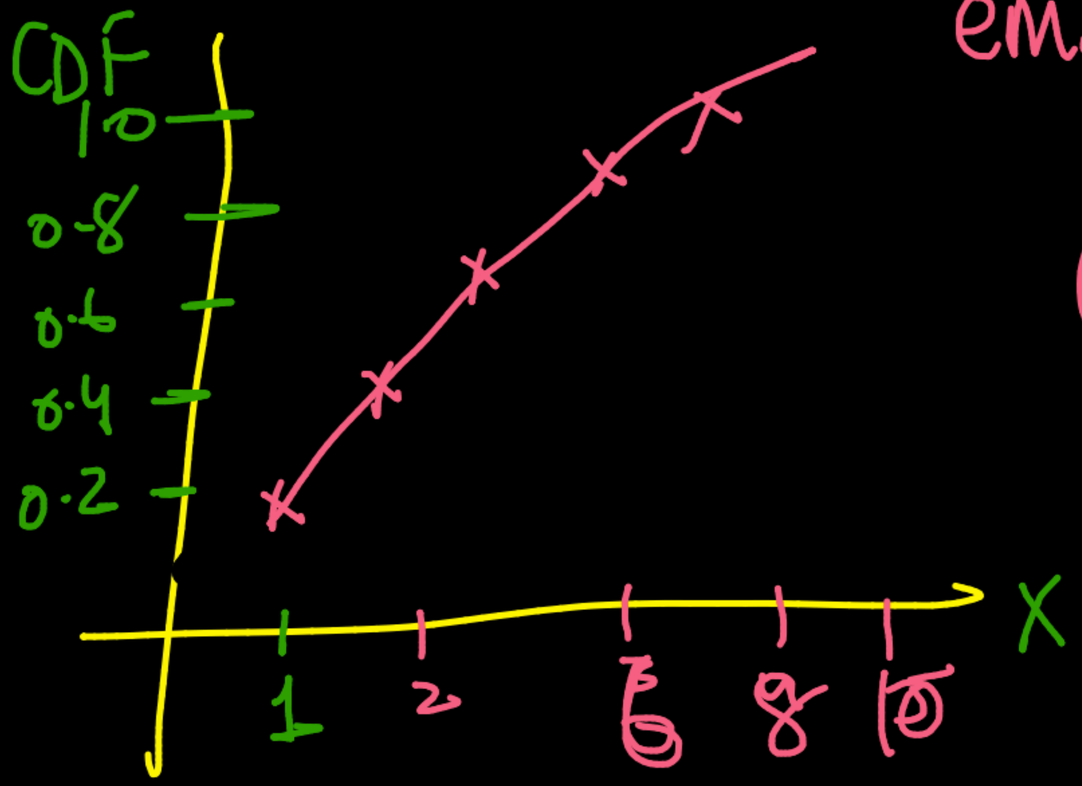
Q8A

10, 8, 6, 2, 1

$x_1, x_2 \ldots x_n \longrightarrow$ sort $\longrightarrow$

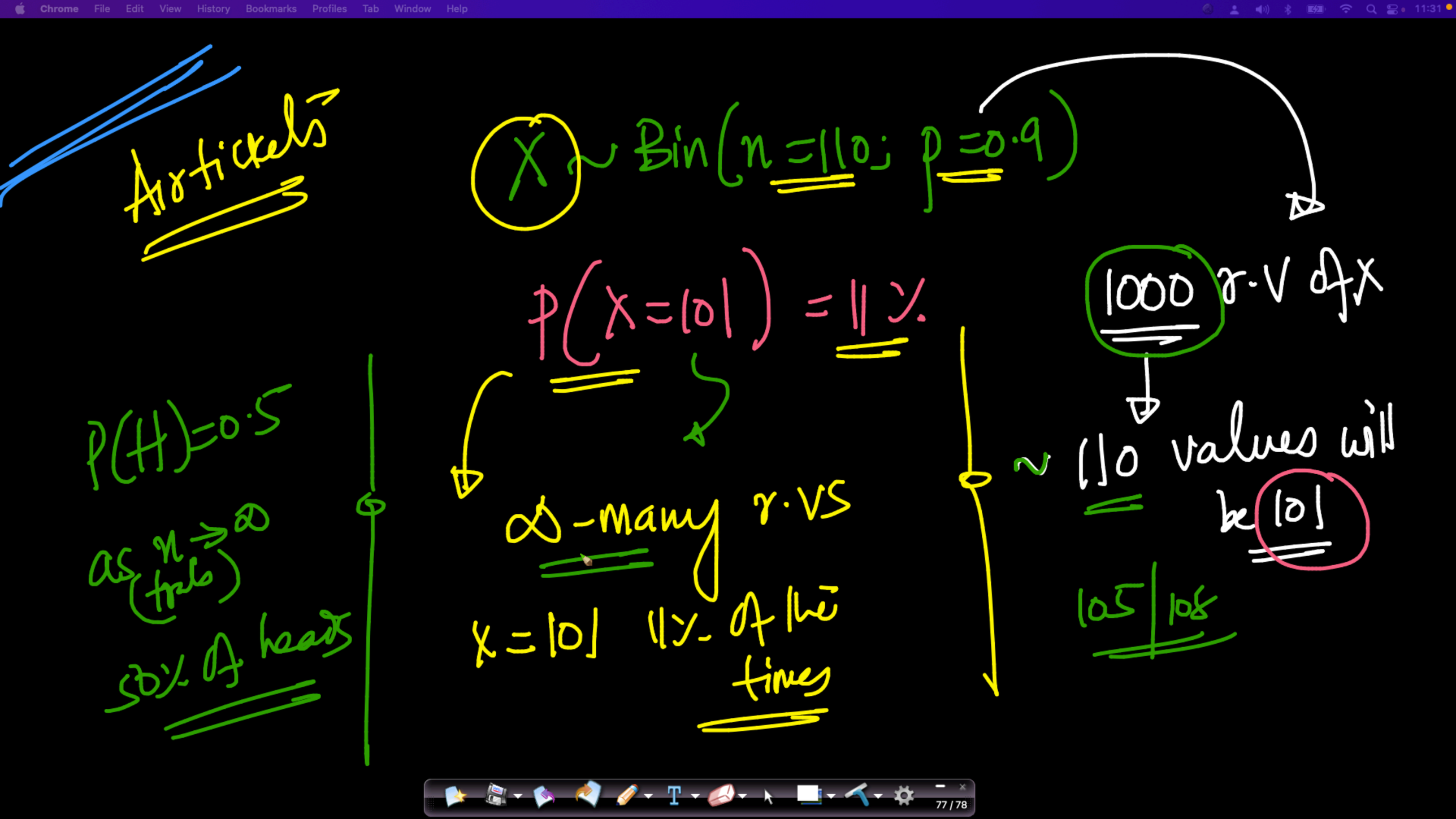
$n = 5$

1, 2, 6, 8, 10



emperical CDF

(KS-Test)


Articles
X ~ Bin(n = 110; p = 0.9)
P(X = 101) = 11%
1000 r.v of X
~ 110 values will be 101
105 | 108
P(H) = 0.5
as n → ∞ (trials)
50% of heads
∞-many r.vs
X = 101 11% of the times

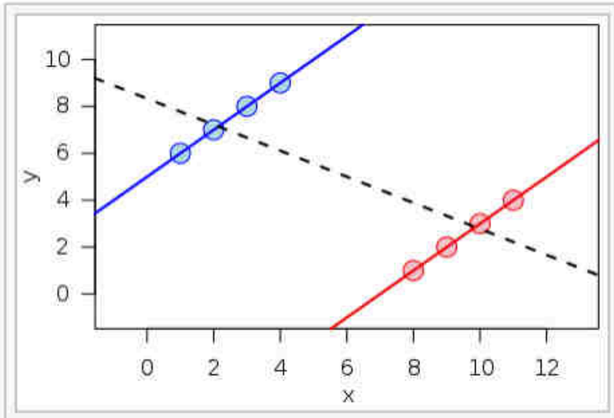
Simpson's paradox

From Wikipedia, the free encyclopedia

See also: Misuse of statistics

Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined. This result is often encountered in social-science and medical-science statistics,[1][2][3] and is particularly problematic when frequency data is unduly given causal interpretations.[4] The paradox can be resolved when confounding variables and causal relations are appropriately addressed in the statistical modeling.[4][5] Simpson's paradox has been used to illustrate the kind of misleading results that the misuse of statistics can generate.[6][7]

Edward H. Simpson first described this phenomenon in a technical paper in 1951,[8] but the statisticians Karl Pearson et al., in 1899,[9] and Udny Yule, in 1903,[10] had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.[11] It is also referred to as **Simpson's reversal**, **Yule–Simpson effect**, **amalgamation paradox**, or **reversal paradox**.[12]



Simpson's paradox for quantitative data: a positive trend (——, ——) appears for two separate groups, whereas a negative trend (- - - -) appears when the groups are combined.

Contents [hide]