Topics:

1. ANOVA - code → "play" → experiment

2. Chi-square Goodness of fit → χ^2 - test of indep...

3. Correlation & Covariance

4. Pearson Correlation - Coeff — Lots of diagrams

5. Spearman Rank CC — more diagrams

+ Problems


Ops

$$\mu = \sum_{i=1}^{n} p_i x_i.$$

(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.)

The variance of a collection of n equally likely values can be written as

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

where μ is the average value. That is,

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

The variance of a set of n equally likely values can be equivalently expressed, without directly referring to the mean, in terms of squared deviations of all pairwise squared distances of points from each other:[2]

$$\operatorname{Var}(X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i - x_j)^2 = \frac{1}{n^2}\sum_{i}\sum_{j>i}(x_i - x_j)^2.$$

Absolutely continuous random variable [edit]

If the random variable X has a probability density function $f(x)$, and $F(x)$ is the corresponding cumulative distribution function, then

$$\operatorname{Var}(X) = \sigma^2 = \int_{\mathbb{R}} (x - \mu)^2 f(x)\,dx$$

$$= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\mu\int_{\mathbb{R}} x f(x)\,dx + \mu^2 \int_{\mathbb{R}} f(x)\,dx$$

$$\mu = \sum_{i=1}^{n} p_i x_i.$$

(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.)

μ: pop-mean

The variance of a collection of n equally likely values can be written as

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

X: $x_1, x_2 \ldots x_n$

where μ is the average value. That is,

$\bar{X}$ Sample mean

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

The variance of a set of n equally likely values can be equivalently expressed, without directly referring to the mean, in terms of squared deviations of all pairwise squared distances of points from each other:[2]

$$\operatorname{Var}(X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i - x_j)^2 = \frac{1}{n^2}\sum_{i}\sum_{j>i}(x_i - x_j)^2.$$

Absolutely continuous random variable [edit]

If the random variable X has a probability density function $f(x)$, and $F(x)$ is the corresponding cumulative distribution function, then

$$\operatorname{Var}(X) = \sigma^2 = \int_{\mathbb{R}} (x-\mu)^2 f(x)\,dx$$

$$= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\mu\int_{\mathbb{R}} x f(x)\,dx + \mu^2\int_{\mathbb{R}} f(x)\,dx$$

The variance of a collection of n equally likely values can be written as

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

where μ is the average value. That is,

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

The variance of a set of n equally likely values can be equivalently expressed, without directly referring to the mean, in terms of squared deviations of all pairwise squared distances of points from each other:[2]

$$\operatorname{Var}(X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i - x_j)^2 = \frac{1}{n^2}\sum_{i}\sum_{j>i}(x_i - x_j)^2.$$

Absolutely continuous random variable [edit]

If the random variable X has a probability density function $f(x)$, and $F(x)$ is the corresponding cumulative distribution function, then

$$\begin{aligned}\operatorname{Var}(X) = \sigma^2 &= \int_{\mathbb{R}} (x-\mu)^2 f(x)\,dx \\ &= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\mu\int_{\mathbb{R}} x f(x)\,dx + \mu^2\int_{\mathbb{R}} f(x)\,dx \\ &= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu\int_{\mathbb{R}} x\,dF(x) + \mu^2\int_{\mathbb{R}} dF(x) \\ &= \int_{\mathbb{R}}\end{aligned}$$

variance can also be thought of as the covariance of a random variable with itself:

$$\operatorname{Var}(X) = \operatorname{Cov}(X, X).$$

The variance is also equivalent to the second cumulant of a probability distribution that generates X. The variance is typically designated as $\operatorname{Var}(X)$, or sometimes as $V(X)$ or $\mathbb{V}(X)$, or symbolically as σ_X^2 or simply σ^2 (pronounced "sigma squared"). The expression for the variance can be expanded as follows:

$$\begin{aligned}
\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\
&= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\
&= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\
&= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2
\end{aligned}$$

$x_1 \dots x_n$

$\frac{1}{n}\sum_{i=1}^{n} x_i^2$

sample-mean

In other words, the variance of X is equal to the mean of the square of X minus the square of the mean of X. This equation should not be used for computations using floating point arithmetic, because it suffers from catastrophic cancellation if the two components of the equation are similar in magnitude. For other numerically stable alternatives, see Algorithms for calculating variance.

Discrete random variable [edit]

If the generator of random variable X is discrete with probability mass function $x_1 \mapsto p_1, x_2 \mapsto p_2, \ldots, x_n \mapsto p_n$, then

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2,$$

where μ is the expected value. That is,

$$\mu = \sum_{i=1}^{n} p_i x_i.$$

Basic properties [edit]

Variance is non-negative because the squares are positive or zero:

$$\operatorname{Var}(X) \geq 0.$$

The variance of a constant is zero.

$$\operatorname{Var}(a) = 0.$$

Conversely, if the variance of a random variable is 0, then it is almost surely a constant. That is, it always has the same value:

$$\operatorname{Var}(X) = 0 \iff \exists a : P(X = a) = 1.$$

Variance is invariant with respect to changes in a location parameter. That is, if a constant is added to all values of the variable, the variance is unchanged:

$$\operatorname{Var}(X + a) = \operatorname{Var}(X).$$

If all values are scaled by a constant, the variance is scaled by the square of that constant:

$$\operatorname{Var}(aX) = a^2 \operatorname{Var}(X).$$

The variance of a sum of two random variables is given by

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y),$$

$$\operatorname{Var}(aX - bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) - 2ab \operatorname{Cov}(X, Y),$$

where $\operatorname{Cov}(X, Y)$ is the covariance.

Linear combinations [edit]

In general, for the sum of N random variables $\{X_1, \ldots, X_N\}$, the variance becomes:

$$\operatorname{Var}\left(\sum_{i=1}^{N} X_i\right)$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:09

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficient | Spearman's rank correlation c | New Tab

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=luJpH2kXGWpU

+ Code + Text Reconnect

Playing with ANOVA

```
[ ] from scipy.stats import f_oneway
    from scipy.stats import norm
    import seaborn as sns
    import matplotlib.pyplot as plt
```

.ipynb

```
[ ] G1 = norm.rvs(loc=10.0, scale=2.0, size=40)
    G2 = norm.rvs(loc=10.3, scale=2.1, size=40)
    G3 = norm.rvs(loc=9.8, scale=2.3, size=50)
```

```
sns.distplot(G1, hist=False)
sns.distplot(G2, hist=False)
sns.distplot(G3, hist=False)
plt.grid()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
```

8 / 8

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:09

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficient | Spearman's rank correlation c | New Tab

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=luJpH2kXGWpU

+ Code + Text Reconnect

Playing with ANOVA

one-way ANOVA

```
[ ] from scipy.stats import f_oneway
    from scipy.stats import norm
    import seaborn as sns
    import matplotlib.pyplot as plt
```

```
[ ] G1 = norm.rvs(loc=10.0, scale=2.0, size=40)
    G2 = norm.rvs(loc=10.3, scale=2.1, size=40)
    G3 = norm.rvs(loc=9.8, scale=2.3, size=50)
```

```
sns.distplot(G1, hist=False)
sns.distplot(G2, hist=False)
sns.distplot(G3, hist=False)
plt.grid()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
```

9 / 9

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficient | Spearman's rank correlation c | New Tab

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=luJpH2kXGWpU

+ Code + Text Reconnect

Playing with ANOVA

Simulation

G1, G2, G3

$N(\mu_1, \sigma_1)$ $N(\mu_2, \sigma_1)$ $N(\mu_3, \sigma_3)$

```
[ ] from scipy.stats import f_oneway
    from scipy.stats import norm
    import seaborn as sns
    import matplotlib.pyplot as plt
```

```
[ ] G1 = norm.rvs(loc=10.0, scale=2.0, size=40)
    G2 = norm.rvs(loc=10.3, scale=2.1, size=40)
    G3 = norm.rvs(loc=9.8, scale=2.3, size=50)
```

```
[ ] sns.distplot(G1, hist=False)
    sns.distplot(G2, hist=False)
    sns.distplot(G3, hist=False)
    plt.grid()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
```

10 / 10

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficient | Spearman's rank correlation c | New Tab

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=luJpH2kXGWpU

+ Code + Text Reconnect

Playing with ANOVA

```
[ ]  from scipy.stats import f_oneway
     from scipy.stats import norm
     import seaborn as sns
     import matplotlib.pyplot as plt
```

$n = m \times k$

```
[ ]  G1 = norm.rvs(loc=10.0, scale=2.0, size=40)
     G2 = norm.rvs(loc=10.3, scale=2.1, size=40)
     G3 = norm.rvs(loc=9.8, scale=2.3, size=50)
```

μ_1 σ_1 μ_2, σ_2 μ_3, σ_3

```
[ ]  sns.distplot(G1, hist=False)
     sns.distplot(G2, hist=False)
     sns.distplot(G3, hist=False)
     plt.grid()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
```

11 / 11

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

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficient | Spearman's rank correlation c | New Tab

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=luJpH2kXGWpU

+ Code + Text Reconnect

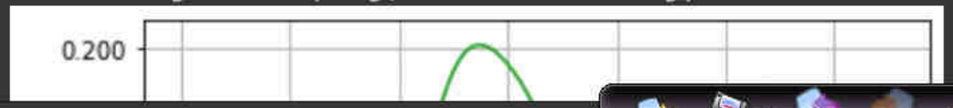
Playing with ANOVA

```
[ ] from scipy.stats import f_oneway
    from scipy.stats import norm
    import seaborn as sns
    import matplotlib.pyplot as plt
```

```
[ ] G1 = norm.rvs(loc=10.0, scale=2.0, size=40)
    G2 = norm.rvs(loc=10.3, scale=2.1, size=40)
    G3 = norm.rvs(loc=9.8, scale=2.3, size=50)
```

```
sns.distplot(G1, hist=False)
sns.distplot(G2, hist=False)
sns.distplot(G3, hist=False)
plt.grid()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
```



12 / 12

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

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficient | Spearman's rank correlation c | New Tab

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=luJpH2kXGWpU

+ Code + Text Reconnect

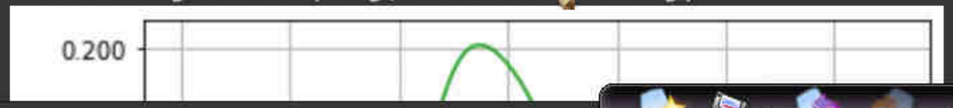
Playing with ANOVA

```
[ ] from scipy.stats import f_oneway
    from scipy.stats import norm
    import seaborn as sns
    import matplotlib.pyplot as plt
```

```
[ ] G1 = norm.rvs(loc=10.0, scale=2.0, size=40)
    G2 = norm.rvs(loc=10.3, scale=2.1, size=40)
    G3 = norm.rvs(loc=9.8, scale=2.3, size=50)
```

```
    sns.distplot(G1, hist=False)
    sns.distplot(G2, hist=False)
    sns.distplot(G3, hist=False)
    plt.grid()
```

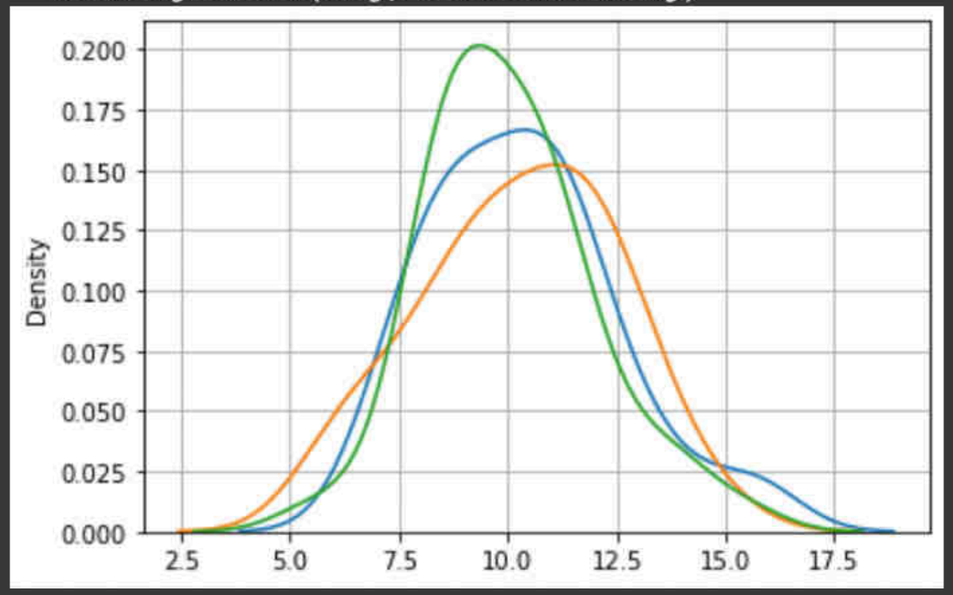
```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
```



13 / 13

```
[ ] sns.distplot(G3, hist=False)
    plt.grid()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
```



Q

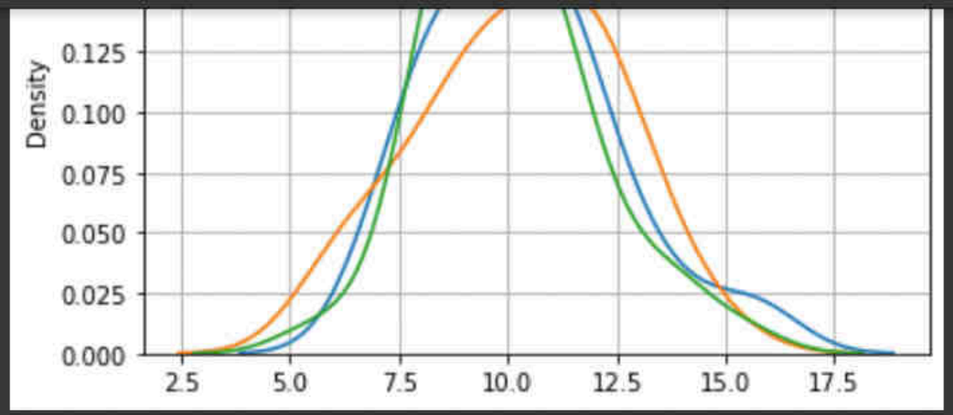
$H_0: \mu_1 = \mu_2 = \mu_3$

H_a: otherwise

```
[ ] f_oneway(G1, G2, G3)
```

+ Code + Text Reconnect

[ ]



p-val < α

< α = 5%

[ ] `f_oneway(G1, G2, G3)`

```
F_onewayResult(statistic=17.22314794729301, pvalue=2.408534447184806e-07)
```

[ ]

Double-click (or enter) to edit

[ ]

#

15 / 15

Chrome File Edit View History Bookmarks Profiles Tab Window Help

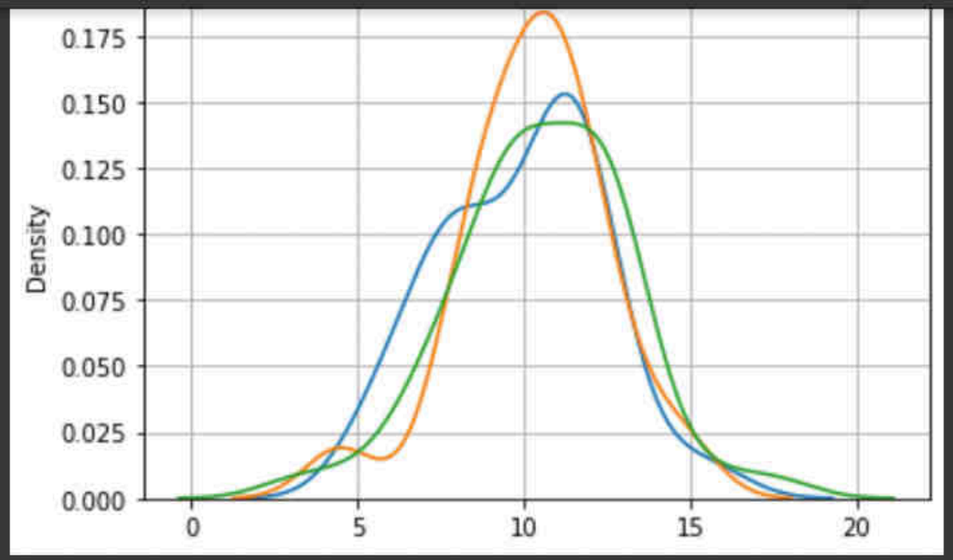
Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficien | Spearman's rank correlation c | New Tab | scipy.stats.f_oneway — SciPy

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=jWfSD1HNGa2a

+ Code + Text

RAM
Disk

[63]



$49.68 > \alpha = 5\%$

p-val

[64] f_oneway(G1, G2, G3)

```
F_onewayResult(statistic=0.7034160770793035, pvalue=0.4968095155468297)
```

Double-click (or enter) to edit

[ ]

#

16 / 16

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficient | Spearman's rank correlation c | New Tab | scipy.stats.f_oneway — SciPy

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=jWfSD1HNGa2a

+ Code + Text

RAM
Disk

Playing with ANOVA

[61]
```
from scipy.stats import f_oneway
from scipy.stats import norm
import seaborn as sns
import matplotlib.pyplot as plt
```

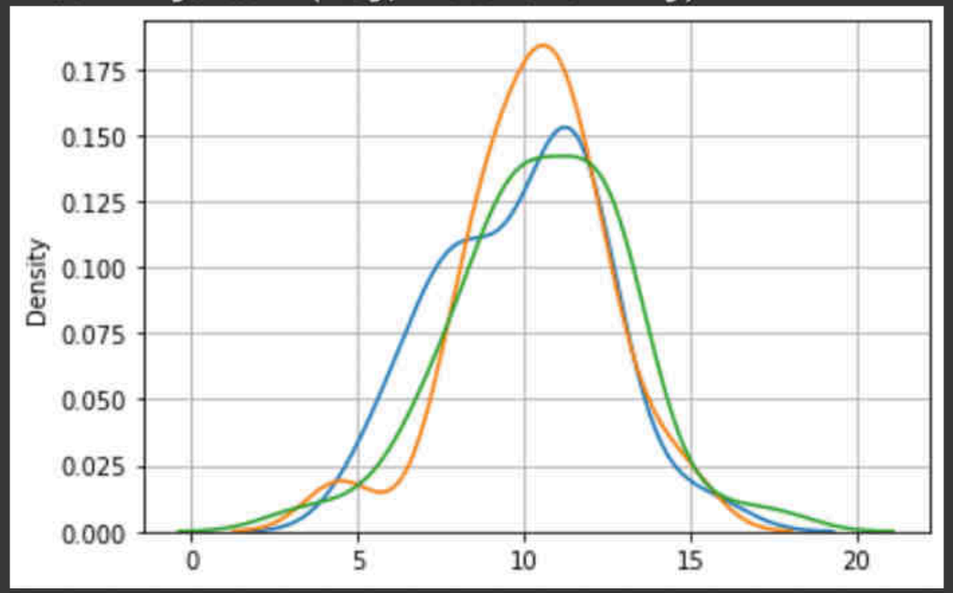
[62]
```
G1 = norm.rvs(loc=10.0, scale=2.0, size=40)
G2 = norm.rvs(loc=10.3, scale=2.1, size=40)
G3 = norm.rvs(loc=9.8, scale=2.3, size=50)
```

[63]
```
sns.distplot(G1, hist=False)
sns.distplot(G2, hist=False)
sns.distplot(G3, hist=False)
plt.grid()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
```

17 / 17

```
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
```



accept Ho

```
[64] f_oneway(G1, G2, G3)
```

```
F_onewayResult(statistic=0.7034160770793035, pvalue=0.4968095155468297)
```

Double-click (or enter) to edit

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:20

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficien | Spearman's rank correlation c | New Tab | scipy.stats.f_oneway — SciPy

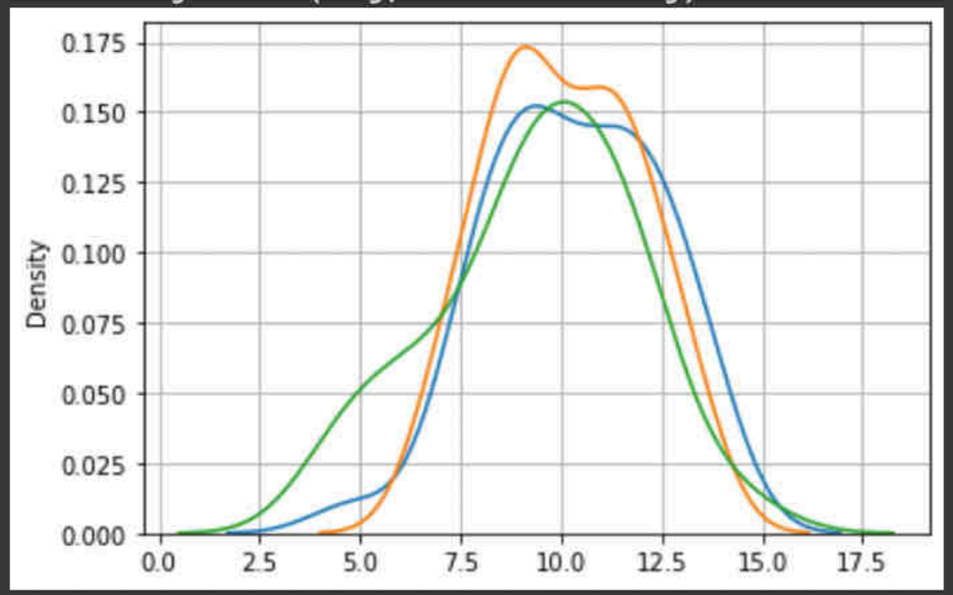
colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=jFHU0LHiGK2i

+ Code + Text

RAM Disk

[69]

```
warnings.warn(msg, FutureWarning)
```



✓ [accept Ho

① → p-val = 49.68%

② → p-val = 43.02%

③ → p-val = 10.44%

```
f_oneway(G1, G2, G3)
```

```
F_onewayResult(statistic=2.2994023688564447, pvalue=0.10448111249878583)
```

Double-click (or enter) to edit

[ ]

19 / 19

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:23

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficien | Spearman's rank correlation c | New Tab | scipy.stats.f_oneway — SciPy

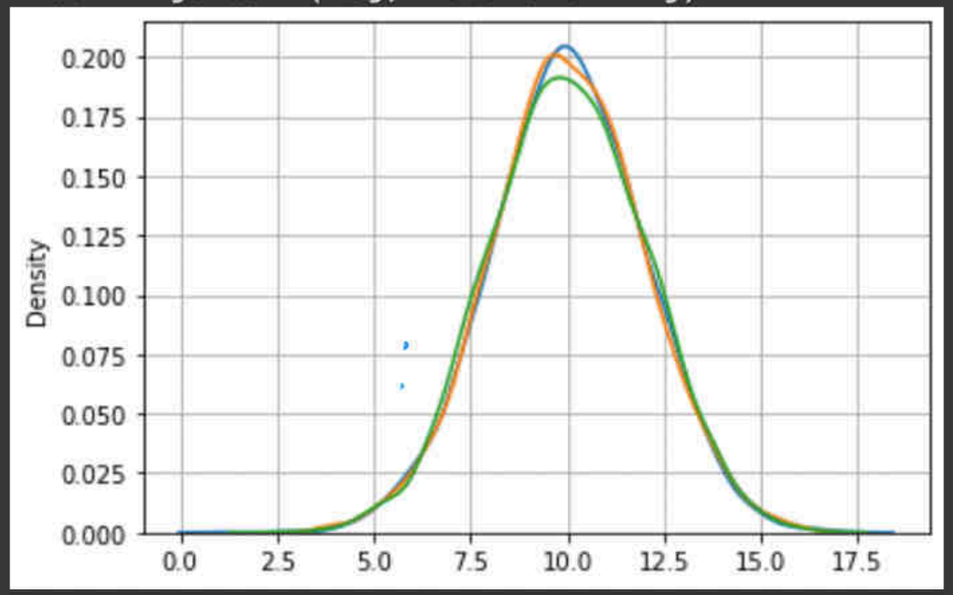
colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=jFHU0LHiGK2i

+ Code + Text

RAM Disk

[81]

```
warnings.warn(msg, FutureWarning)
```



M1 M2 M3

① with 40,40,50 → 43%, 45%, 10%.

② p-val = 88.60%; 82.7%

97.74%

4000, 4000, 5000

```
f_oneway(G1, G2, G3)
```

```
F_onewayResult(statistic=0.022831023765828056, pvalue=0.9774276710606851)
```

Double-click (or enter) to edit

[ ]

20 / 20

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:24

Variance - Wikipedia | ANOVA & Chi-Square.ipynb - C | Covariance - Wikipedia | Pearson correlation coefficient | Spearman's rank correlation c | New Tab | scipy.stats.f_oneway — SciPy

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=jFHU0LHiGK2i

+ Code + Text RAM Disk

Playing with ANOVA

```
[61] from scipy.stats import f_oneway
     from scipy.stats import norm
     import seaborn as sns
     import matplotlib.pyplot as plt
```

```
[80] G1 = norm.rvs(loc=10.0, scale=2.0, size=4000)
     G2 = norm.rvs(loc=10.0, scale=2.01, size=4000)
     G3 = norm.rvs(loc=10.0, scale=2.03, size=5000)
```

```
     sns.distplot(G1, hist=False)
     sns.distplot(G2, hist=False)
     sns.distplot(G3, hist=False)
     plt.grid()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
```

① as sample sizes ↑ ANOVA can detect small differences in μ_1, μ_2 & μ_3

② Variance in p-values depends on sample size

21 / 21

Playing with ANOVA

```
[61] from scipy.stats import f_oneway
     from scipy.stats import norm
     import seaborn as sns
     import matplotlib.pyplot as plt
```

```
[80] G1 = norm.rvs(loc=10.0, scale=2.0, size=4000)
     G2 = norm.rvs(loc=10.0, scale=2.01, size=4000)
     G3 = norm.rvs(loc=10.0, scale=2.03, size=5000)
```

```
[81] sns.distplot(G1, hist=False)
     sns.distplot(G2, hist=False)
     sns.distplot(G3, hist=False)
     plt.grid()
```

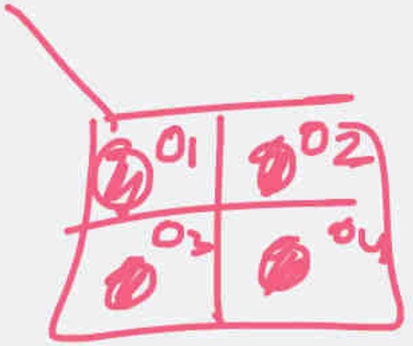
```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated func
```

together

- theory → concrete 'proofs'
- Simulation → help you build intuition

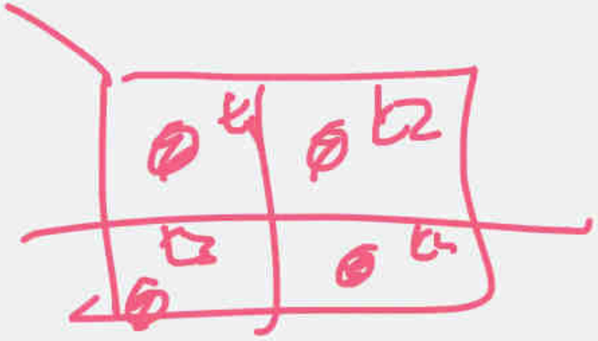
Likelihood	105.357	1	0.000

The resulting chi-square statistic is 102.596 with a p-value of .000. The 2X2 table also includes the expected values. Remember the chi-square statistic is comparing the expected values to the observed values from Donna's study. The results of the chi-square indicate this difference (observed – expected is large). Thus, Donna can reject the null hypothesis that entrepreneurialism and geographic location are independent and she can conclude that Entrepreneurialism levels depend on geographic location.



Conditions for Using the Chi-Square Test

Exercise caution when there are small expected counts. Minitab will give a count of the number of cells that have expected frequencies less than five. Some statisticians hesitate to use the chi-square test if more than 20% of the cells have expected frequencies below five, especially if the p-value is small and these cells give a large contribution to the total chi-square value.



Caution!

Sometimes researchers will categorize quantitative data (e.g., take height measurements and categorize as 'below average,' 'average,' and 'above average.") Doing so results in a loss of information - one cannot do the reverse of taking the categories and reproducing the raw quantitative measurements. Instead of categorizing, the data should be analyzed using quantitative methods.

« 6.1 - Chi-Square Test for Independence

6.3 - Risk, Relative Risk and Odds »

more the Sample size ↑

↓

more confident we are about accepting Ho or Ha

⇓ less power to errors

α : FP-error

↓ β

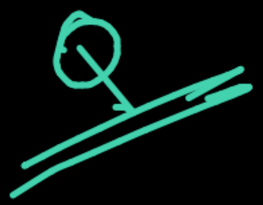
1−β = power ↑

- Markov ineq
- Chebyshev ineq

Threshold

p-val $> \alpha = 5\%$ $\rightarrow$ accept H_0

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



(1)

G_1

G_2

G_3

G_4

Populations are normal

Samples v.close ~ Normal

(40)

(2) Th: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

Pr: $s_1^2 \approx s_2^2 \approx s_3^2$ ✓

Q

$$J_{ANOVA} = f = \frac{\frac{1}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2}$$

'k'

$k-1$

$n-k$

n

$f \sim Fdist(k-1, n-k)$

{Task:

Hyp ↷

[#learners attending classes is "roughly" the same across the 5 days..]

X:

Data:

M → 1 → 100

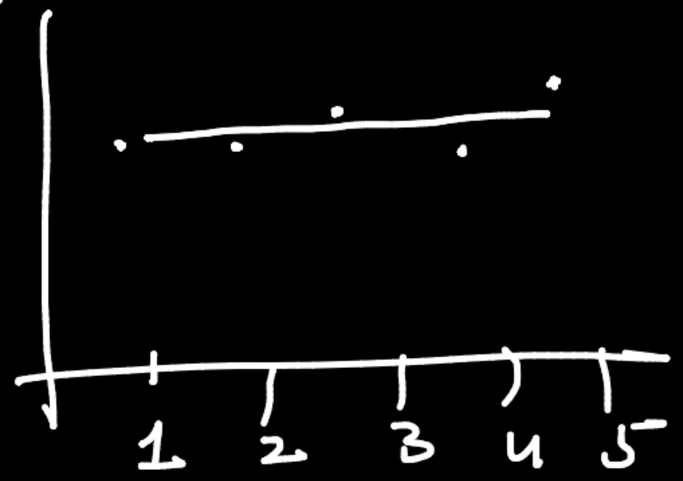
T → 2 → 98

W → 3 → 101

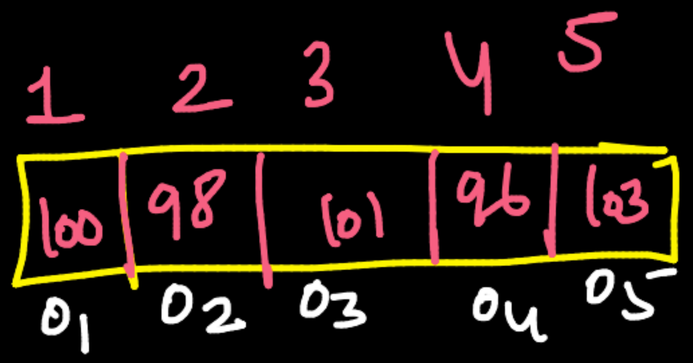
Th → 4 → 96

Fri → 5 → 103

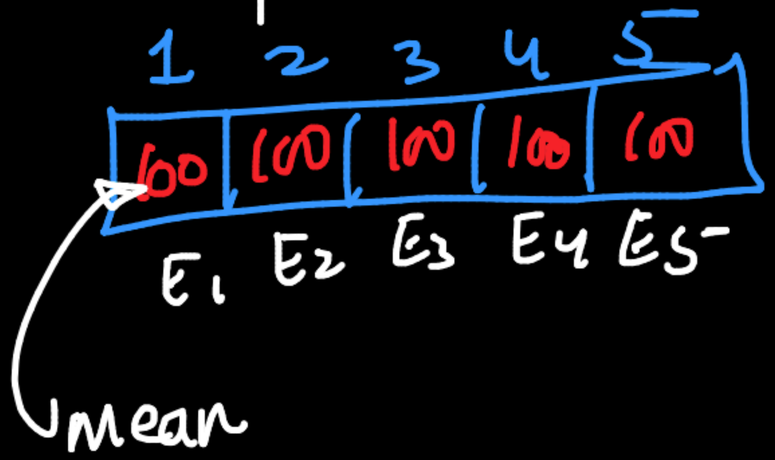
hist



Obs



exp Under H_0



$H_0: a_1 = a_2 = a_3 = a_4 = a_5$ → Uniform dist

H_1: otherwise

$$T_{\chi^2} = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2 \text{ distribution } (6-1)$$

χ^2-test: Goodness of fit

Uniform distribution

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Variance - Wikipe | ANOVA & Chi-Squ | Covariance - Wiki | Pearson correlatio | Spearman's rank c | New Tab | scipy.stats.f_one | Chi-Square test: F | Chi Square Analys | 6.2 - Chi-Square T

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=So7E5K2_K7MA

+ Code + Text

RAM
Disk

[ ] ↳ 4 cells hidden

Chi-Square Goodness of fit

```
from scipy.stats import chisquare
from scipy.stats import poisson
import math
```

0s

```
[41] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
```

0s

```
[42] freq_exp= (p_x_i*100).round()
     freq_exp
```

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

0s

```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

0s

```
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
  terms = (f_obs_float - f_exp)**2 / f_exp
```

33 / 33

Data ~ dist ?

- KS-test → Compares CDFs
  - ↑ continuous r.v
- χ^2 - Good ness of fit → discrete r.v

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

Variance - Wikipe | ANOVA & Chi-Squ | Covariance - Wiki | Pearson correlatio | Spearman's rank c | New Tab | scipy.stats.f_one | Chi-Square test: F | Chi Square Analys | 6.2 - Chi-Square T

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=So7E5K2_K7MA

+ Code + Text RAM Disk

Chi-Square Goodness of fit

→ Footbal goals

$\lambda = 2.5$

```
from scipy.stats import chisquare
from scipy.stats import poisson
import math
```

```
[41] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
```

```
[42] freq_exp= (p_x_i*100).round()
     freq_exp
```

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

```
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
  terms = (f_obs_float - f_exp)**2 / f_exp
Power_divergenceResult(statistic=nan, pvalue=nan)
```

35 / 35

Chi-Square Goodness of fit

$X \sim Poisson(\lambda = 2.5)$

```
from scipy.stats import chisquare
from scipy.stats import poisson
import math
```

$P(X=0) \rightarrow$

$P(X=1) \rightarrow$

```
[41] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
```

```
[42] freq_exp= (p_x_i*100).round()
     freq_exp
```

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

```
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
  terms = (f_obs_float - f_exp)**2 / f_exp
Power_divergenceResult(statistic=nan, pvalue=nan)
```

Chrome File Edit View History Bookmarks Profiles Tab Window Help

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=M9BPzgqPL2N0

+ Code + Text

RAM
Disk

```
[83] from scipy.stats import chisquare
     from scipy.stats import poisson
     import math
```

```
[85] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
     p_x_i
```

```
array([8.20849986e-02, 2.05212497e-01, 2.56515621e-01, 2.13763017e-01,
       1.33601886e-01, 6.68009429e-02, 2.78337262e-02, 9.94061650e-03,
       3.10644266e-03, 8.62900738e-04, 2.15725184e-04])
```

```
freq_exp= (p_x_i*100).round()
freq_exp
```

0 1 2 3 4 5 6 7 8 9 10

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

```
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
  terms = (f_obs_float - f_exp)**2 / f_exp
Power_divergenceResult(statistic=nan, pvalue=nan)
```

$P(X=2) \times 100 =$

37 / 37

```
[83] from scipy.stats import chisquare
     from scipy.stats import poisson
     import math
```

```
[85] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
     p_x_i
```

```
array([8.20849986e-02, 2.05212497e-01, 2.56515621e-01, 2.13763017e-01,
       1.33601886e-01, 6.68009429e-02, 2.78337262e-02, 9.94061650e-03,
       3.10644266e-03, 8.62900738e-04, 2.15725184e-04])
```

```
freq_exp= (p_x_i*100).round()
freq_exp
```

0 1 2 3 4 - - - 10

freq-exp

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

```
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
  terms = (f_obs_float - f_exp)**2 / f_exp
Power_divergenceResult(statistic=nan, pvalue=nan)
```

Chi-Square Goodness of fit

discrete r.v

```
[83] from scipy.stats import chisquare
     from scipy.stats import poisson
     import math
```

```
[85] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
     p_x_i
```

```
array([8.20849986e-02, 2.05212497e-01, 2.56515621e-01, 2.13763017e-01,
       1.33601886e-01, 6.68009429e-02, 2.78337262e-02, 9.94061650e-03,
       3.10644266e-03, 8.62900738e-04, 2.15725184e-04])
```

```
freq_exp= (p_x_i*100).round()
freq_exp
```

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

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

Variance - Wikipe | ANOVA & Chi-Squ | Covariance - Wiki | Pearson correlatio | Spearman's rank | New Tab | scipy.stats.f_one | Chi-Square test: | Chi Square Analys | 6.2 - Chi-Square

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=M9BPzgqPL2N0

+ Code + Text RAM Disk

```
[83] import math
```

```
[85] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
     p_x_i
```

```
array([8.20849986e-02, 2.05212497e-01, 2.56515621e-01, 2.13763017e-01,
       1.33601886e-01, 6.68009429e-02, 2.78337262e-02, 9.94061650e-03,
       3.10644266e-03, 8.62900738e-04, 2.15725184e-04])
```

```
freq_exp= (p_x_i*100).round()
freq_exp
```

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

0 1 2 3 4 5 6 7 8 9 10

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

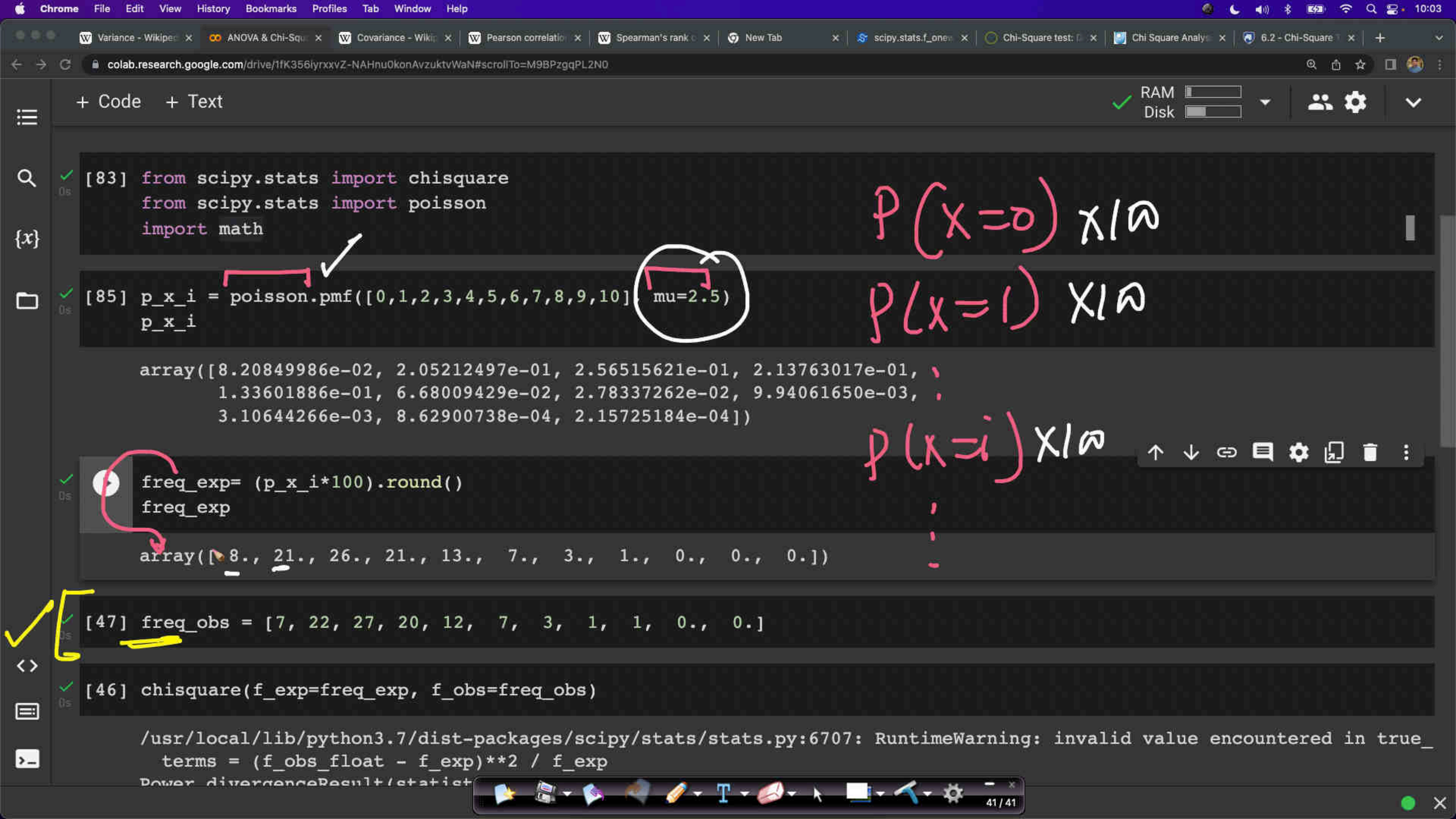
```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

```
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
  terms = (f_obs_float - f_exp)**2 / f_exp
Power_divergenceResult(statistic=nan, pvalue=nan)
```

```
[52] freq_exp = [ 8., 21., 26., 21., 13.,  7.,  3.,  1.]
```

Task: Can we conclude that #goals p.m ~ Poisson ($\lambda = 2.5$)?

40 / 40


Chrome File Edit View History Bookmarks Profiles Tab Window Help
colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=M9BPzgqPL2N0
+ Code + Text
RAM
Disk
[83] from scipy.stats import chisquare
from scipy.stats import poisson
import math
[85] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
p_x_i
array([8.20849986e-02, 2.05212497e-01, 2.56515621e-01, 2.13763017e-01,
1.33601886e-01, 6.68009429e-02, 2.78337262e-02, 9.94061650e-03,
3.10644266e-03, 8.62900738e-04, 2.15725184e-04])
freq_exp= (p_x_i*100).round()
freq_exp
array([8., 21., 26., 21., 13., 7., 3., 1., 0., 0., 0.])
[47] freq_obs = [7, 22, 27, 20, 12, 7, 3, 1, 1, 0., 0.]
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
terms = (f_obs_float - f_exp)**2 / f_exp
P(X=0) x1a
P(x=1) x1a
P(x=i) x1a
41 / 41

```
[83] from scipy.stats import chisquare
     from scipy.stats import poisson
     import math
```

```
[85] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
     p_x_i
```

```
array([8.20849986e-02, 2.05212497e-01, 2.56515621e-01, 2.13763017e-01,
       1.33601886e-01, 6.68009429e-02, 2.78337262e-02, 9.94061650e-03,
       3.10644266e-03, 8.62900738e-04, 2.15725184e-04])
```

```
freq_exp= (p_x_i*100).round()
freq_exp
```

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

```
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
  terms = (f_obs_float - f_exp)**2 / f_exp
```

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

Variance - Wikipe | ANOVA & Chi-Squ | Covariance - Wiki | Pearson correlatio | Spearman's rank c | New Tab | scipy.stats.f_one | Chi-Square test: F | Chi Square Analys | 6.2 - Chi-Square T

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=M9BPzgqPL2N0

+ Code + Text RAM Disk

```
[83] import math
```

```
[85] p_x_i = poisson.pmf([0,1,2,3,4,5,6,7,8,9,10], mu=2.5)
     p_x_i
```

```
array([8.20849986e-02, 2.05212497e-01, 2.56515621e-01, 2.13763017e-01,
       1.33601886e-01, 6.68009429e-02, 2.78337262e-02, 9.94061650e-03,
       3.10644266e-03, 8.62900738e-04, 2.15725184e-04])
```

$$T = \sum_{i=0}^{10} \frac{(Oi - Ei)^2}{Ei}$$

```
freq_exp= (p_x_i*100).round()
freq_exp
```

```
array([ 8., 21., 26., 21., 13.,  7.,  3.,  1.,  0.,  0.,  0.])
```

```
[47] freq_obs = [7, 22, 27, 20, 12,  7,  3,  1,  1,  0.,  0.]
```

```
[46] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

```
/usr/local/lib/python3.7/dist-packages/scipy/stats/stats.py:6707: RuntimeWarning: invalid value encountered in true_
  terms = (f_obs_float - f_exp)**2 / f_exp
Power_divergenceResult(statistic=nan, pvalue=nan)
```

```
[52] freq_exp = [ 8., 21., 26., 21.,
```

43 / 43

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

Variance - Wikipe | ANOVA & Chi-Squ | Covariance - Wiki | Pearson correlatio | Spearman's rank c | New Tab | scipy.stats.f_one | Chi-Square test: F | Chi Square Analys | 6.2 - Chi-Square T

colab.research.google.com/drive/1fK356iyrxxvZ-NAHnu0konAvzuktvWaN#scrollTo=M9BPzgqPL2N0

+ Code + Text

RAM
Disk

```
[52] freq_exp = [ 8., 21., 26., 21., 13.,  7.,  3.,  1.]
     freq_obs = [7, 22, 27, 20, 12,  8,  3,  1]
```

≈ 1 - 5% = 0.05

```
[53] chisquare(f_exp=freq_exp, f_obs=freq_obs)
```

```
Power_divergenceResult(statistic=0.47847985347985345, pvalue=0.9995215264692844)
```

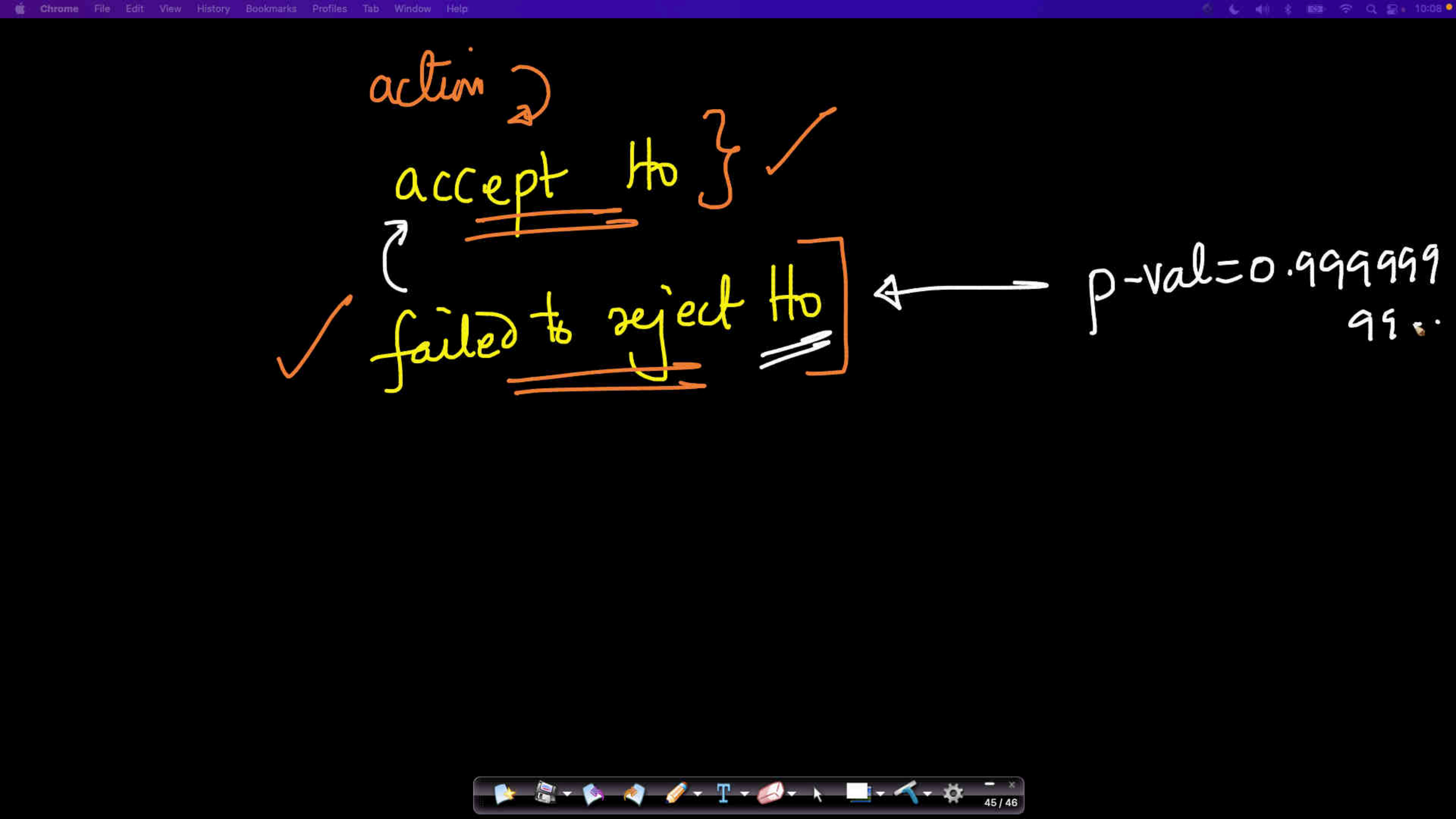
```
[ ]
```

Double-click (or enter) to edit

H0 ✓

#goals P.M ~ Pois(λ=2.5)

44 / 44


action
accept Ho
failed to reject Ho
p-val=0.999999

Q

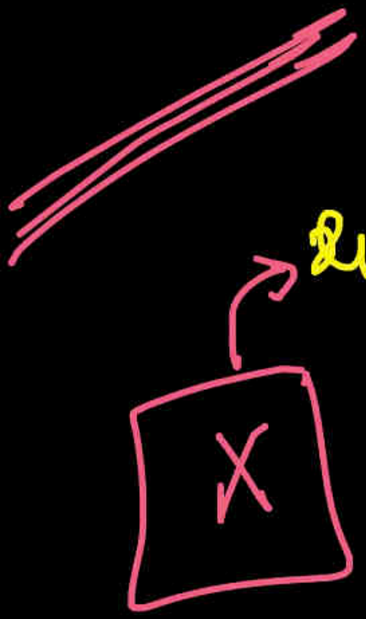
Cont - r.V

χ^2 good of fit } → KS-test

↓

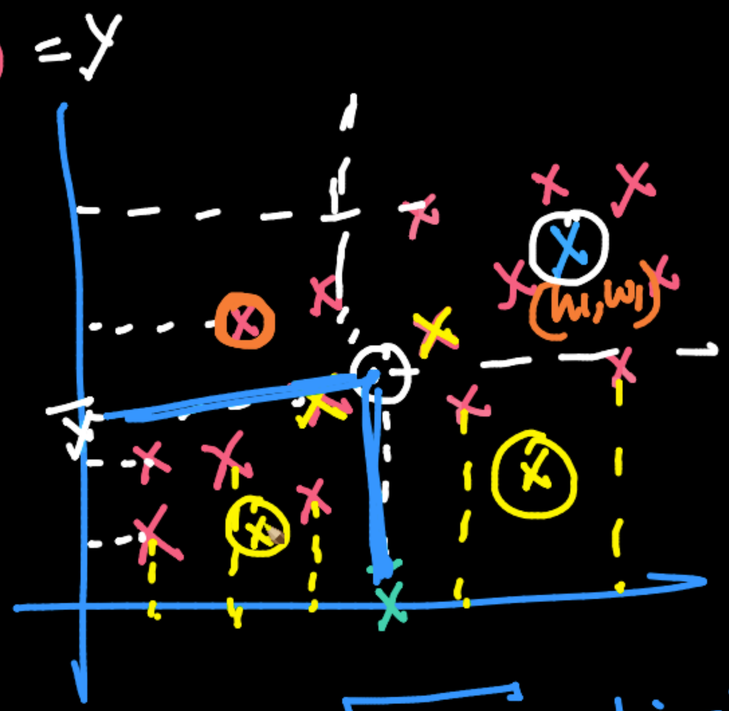
AD-test }

Covariance



$$\text{Variance} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

	H (cm)	W (kgs)
1	x_1	y_1
2	x_2	y_3
3	x_3	y_3
⋮		
n	x_n	y_n
	X	Y
	$\bar{x}$	$\bar{y}$



$$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$h_i - \bar{h}$: +ve

$w_i - \bar{w}$: +ve

avg

Case 1:

$1 \rightarrow (h_1, \omega_1)$

$h_1 > \bar{h}$

$\omega_1 > \bar{\omega}$

Cov ↑

Case 2:

$h_2 < \bar{h}$

-ve

$\omega_2 > \bar{\omega}$

+ve

Cov ↓

Case 3:

$h_2 > \bar{h}$

+ve

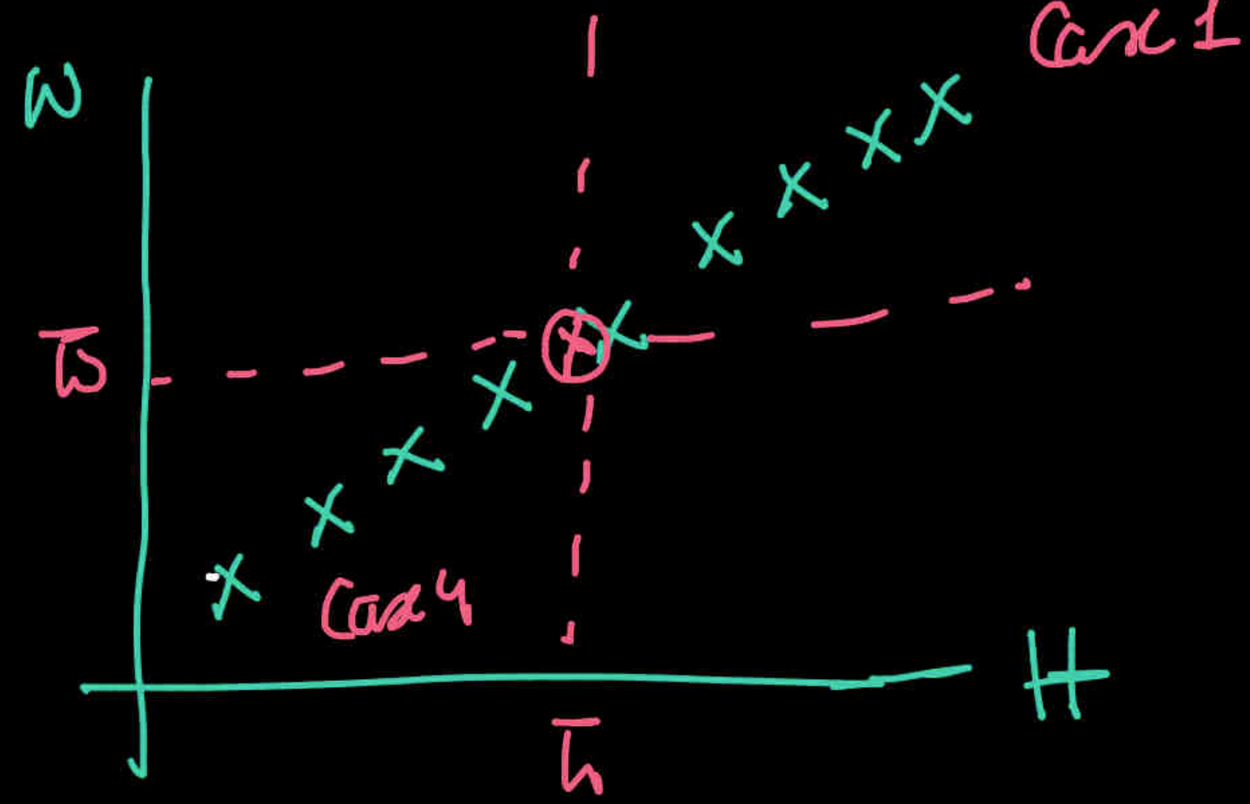
$\omega_2 < \bar{\omega}$

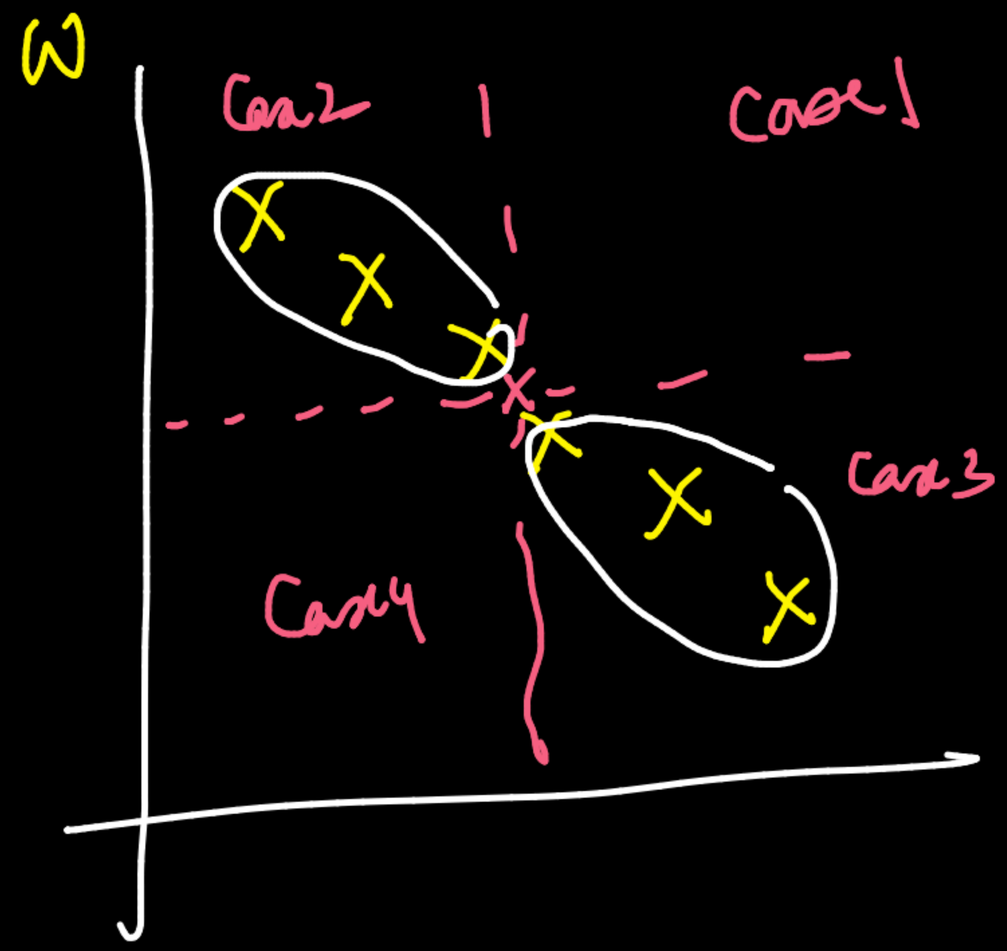
-ve

Cov ↓

Case 4: $h_1 < \bar{h}$ $\omega_1 < \bar{\omega}$ Cov ↑

Cov. high

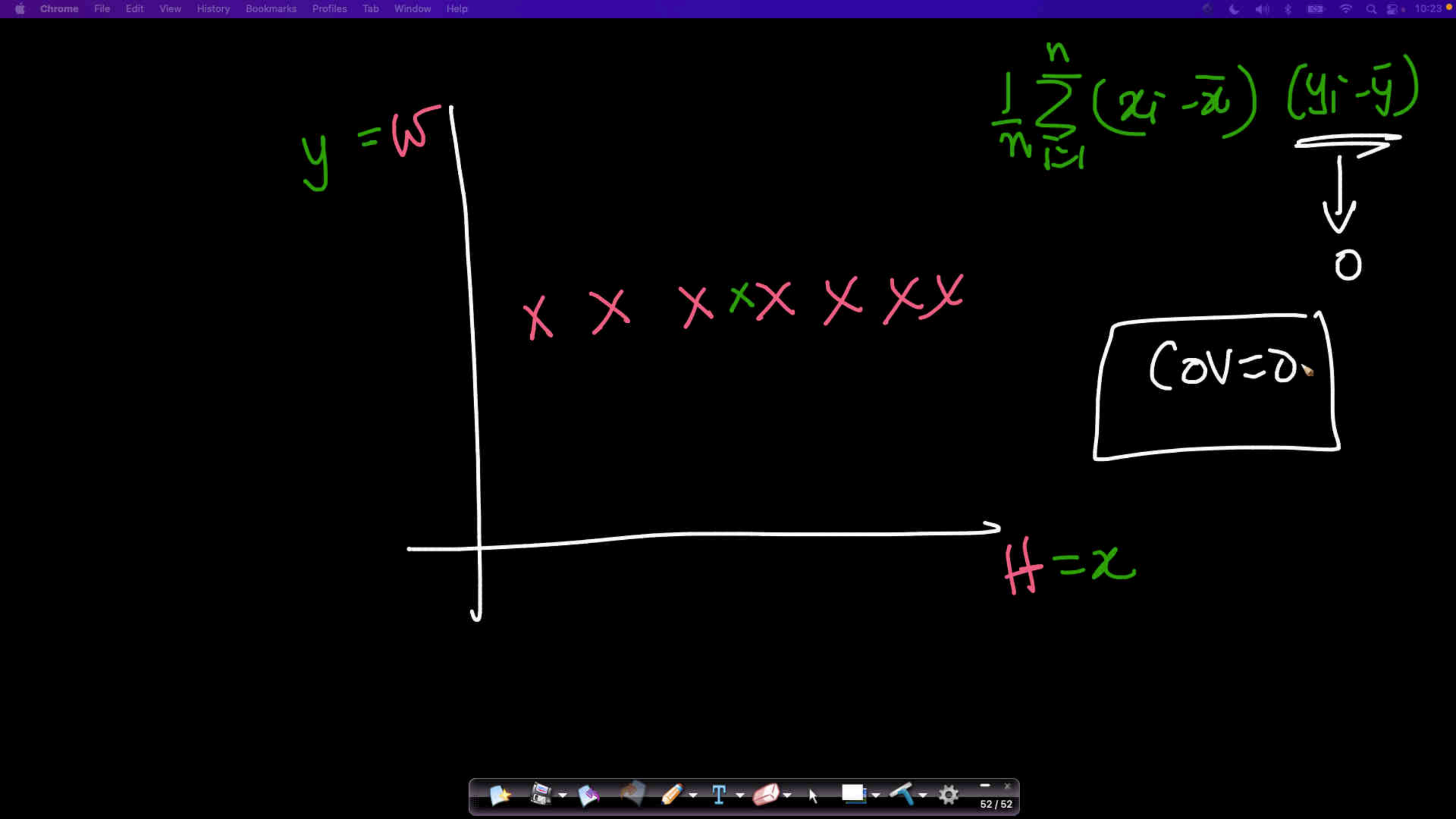


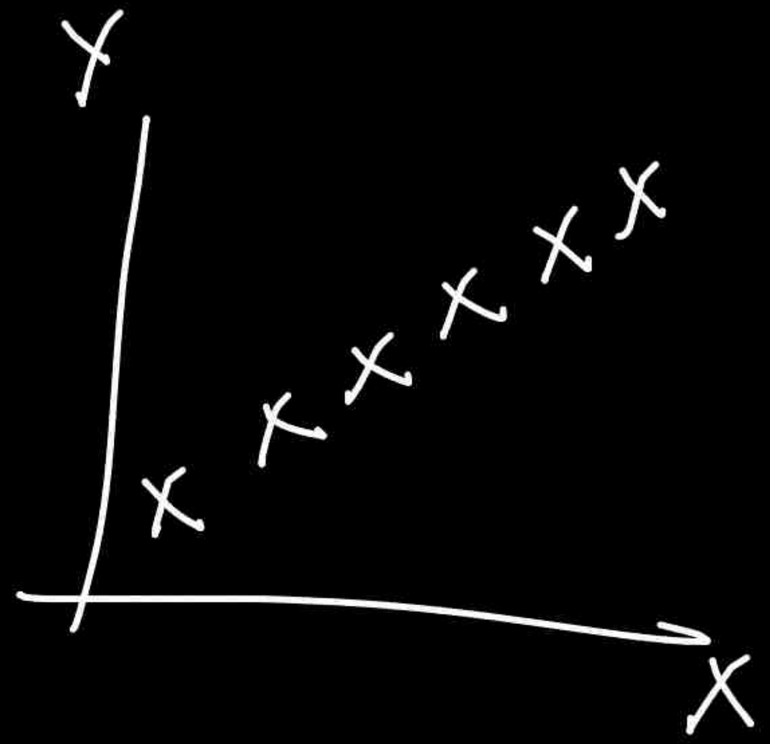


Cov ↓ ↓

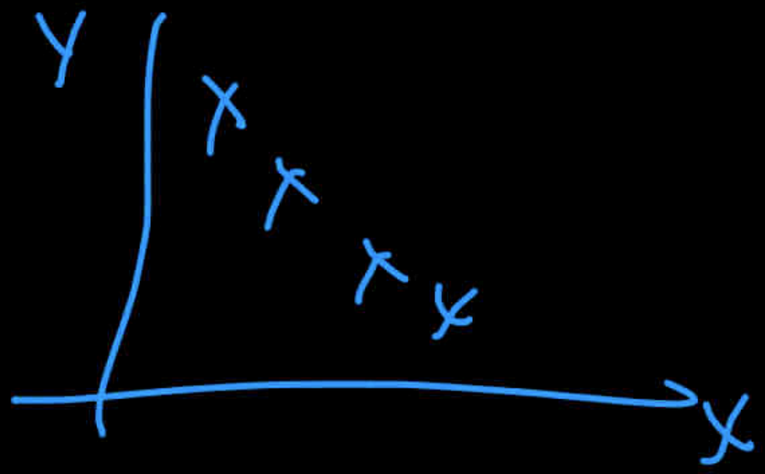
↓

large -ve val

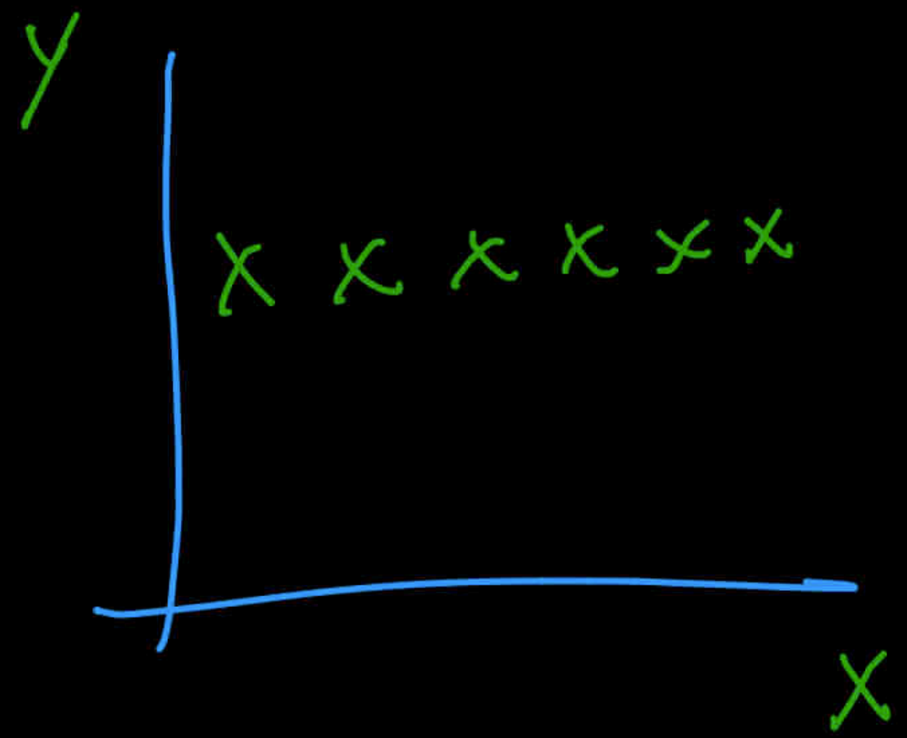

$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
0
Cov=0
y = w
H = x



Cov (X,Y) ↑ : large +ve

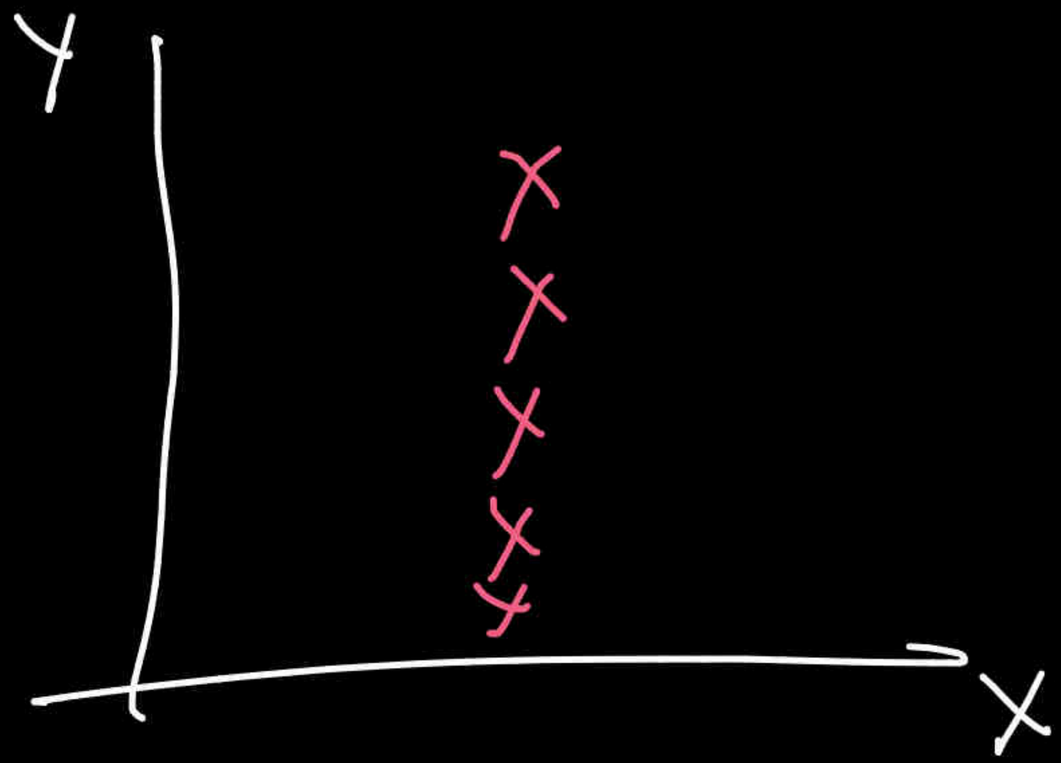


Cov (X,Y) ↓ : large -ve

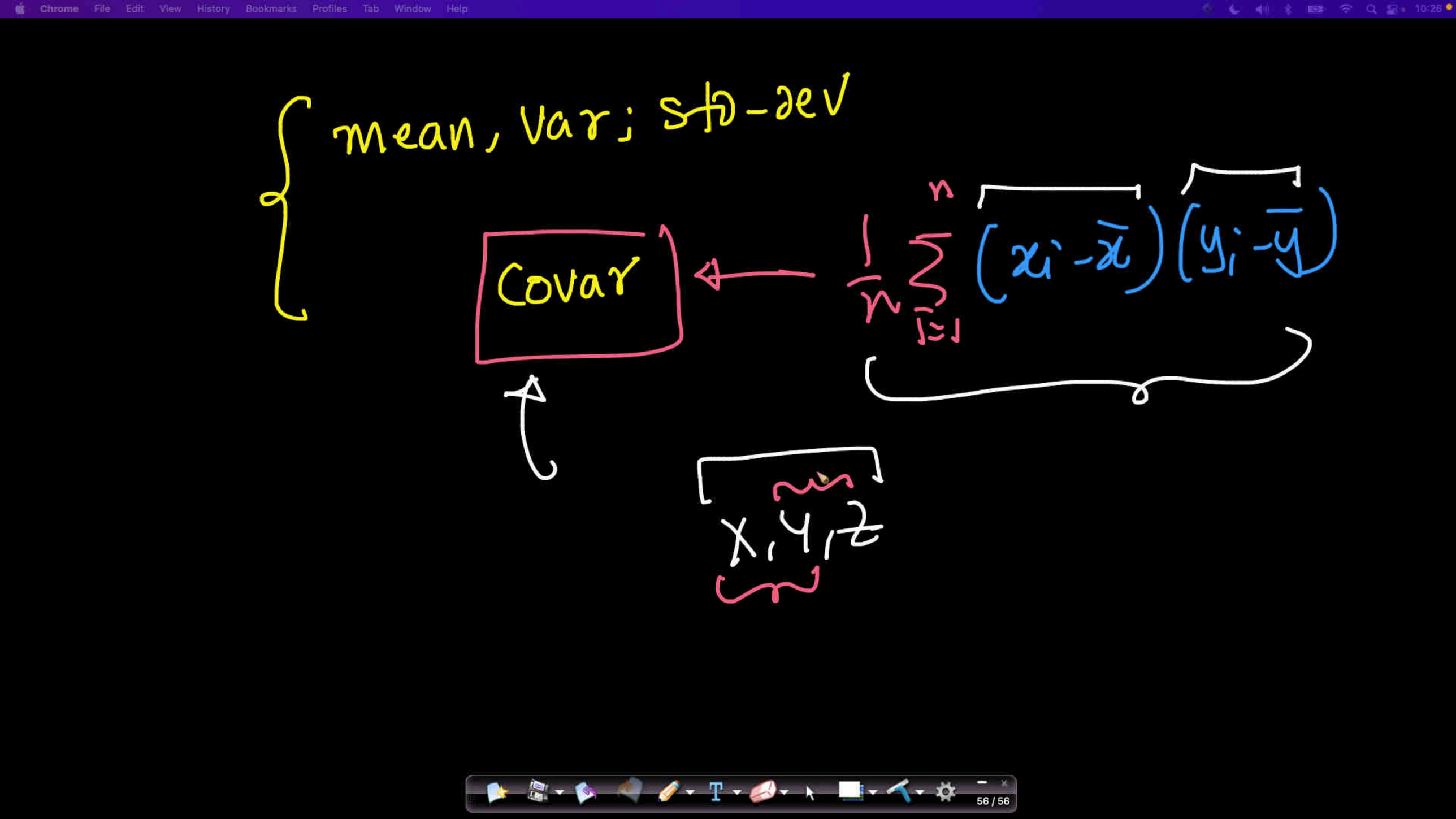


Co-var

→ Cov(X,Y) = 0



$$\underline{\underline{Cov}}(X,Y)=0$$


mean, Var; std-dev
Covar
$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
X, Y, Z

intuitive

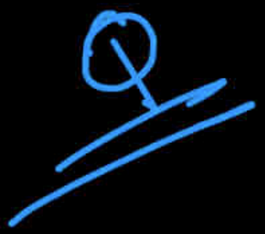
$X\uparrow \Rightarrow Y\uparrow$ → Cov : +ve

$X\uparrow \Rightarrow Y\downarrow$ → Cov : -ve

$X\uparrow \Rightarrow Y\rightarrow$

$Y\uparrow \Rightarrow X\rightarrow$

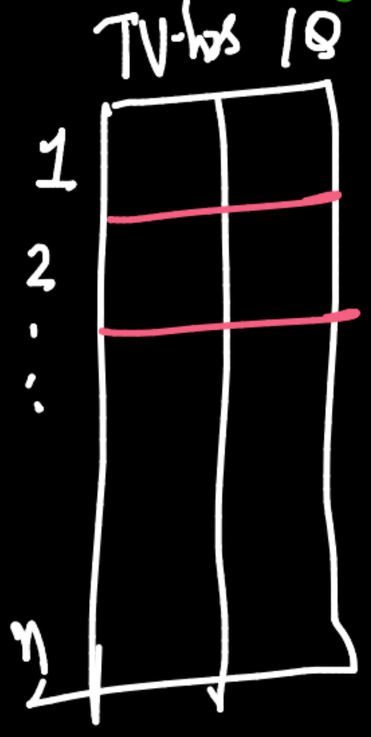
} → Cov : Zero

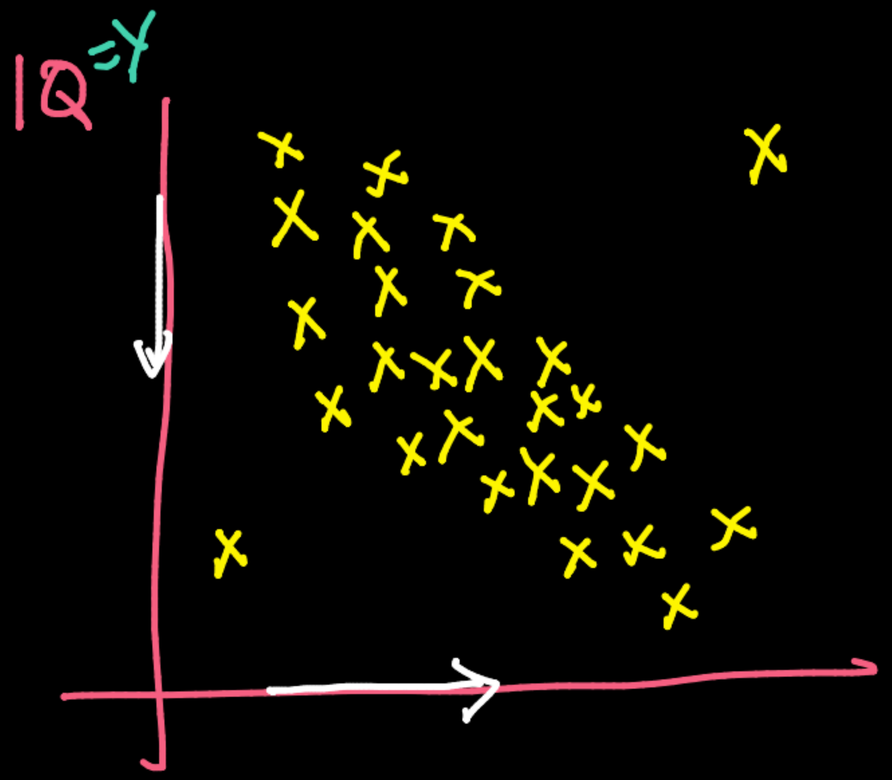


$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot(y_i-\bar{y})^2$$

$cov(x,y)$ → correlation

a. Cov > 0
b. Cov < 0
c. Cov ≈ 0





Cov(X,Y) < 0

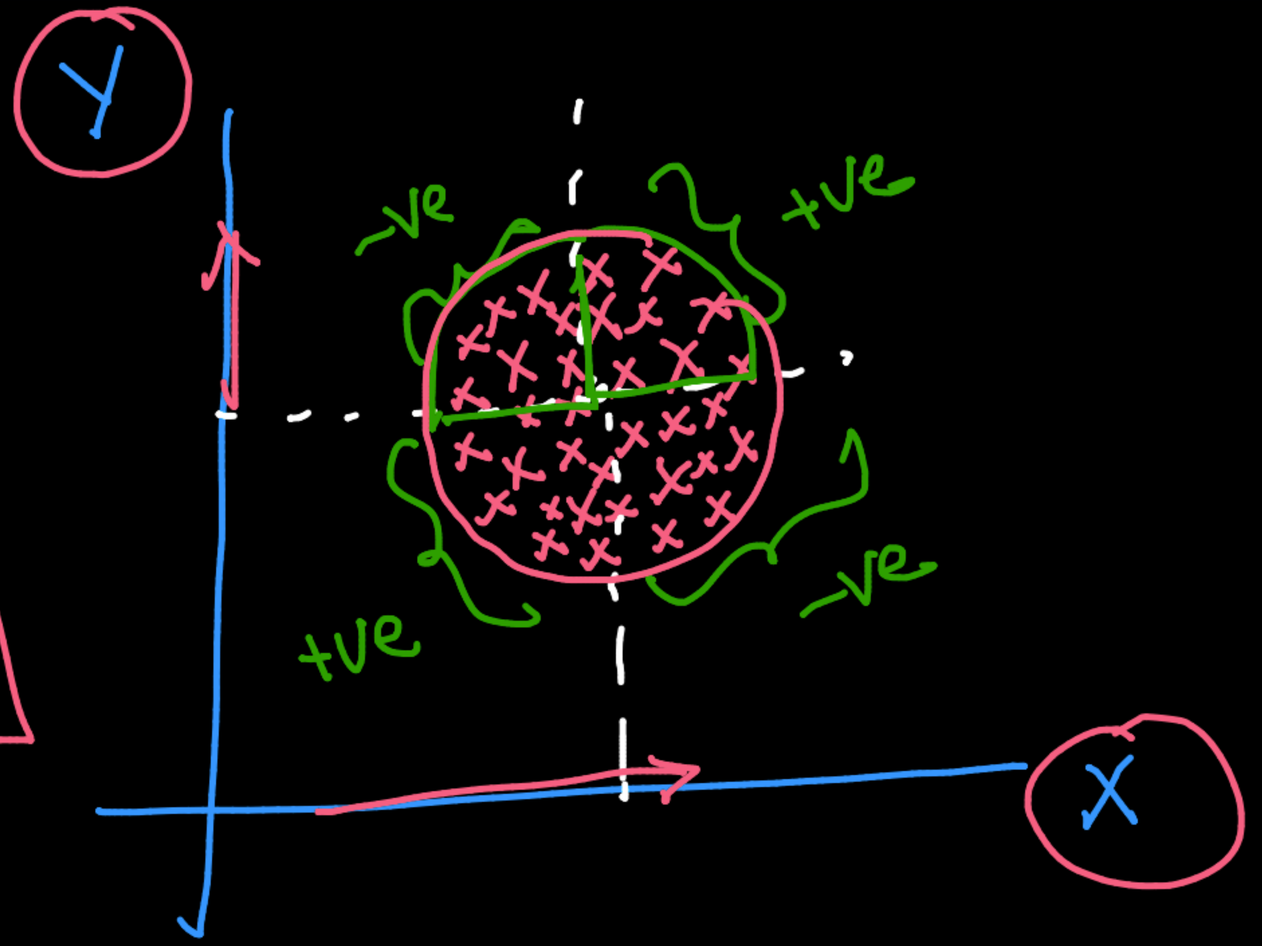
↳ Task: as #hrsTV ↑ does IQ ↓

Q

a: Cov(X,Y) > 0

b: Cov(X,Y) < 0

c: Cov(X,Y) ≈ 0 ✓



$$\mathrm{Cov}(x, x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})$$

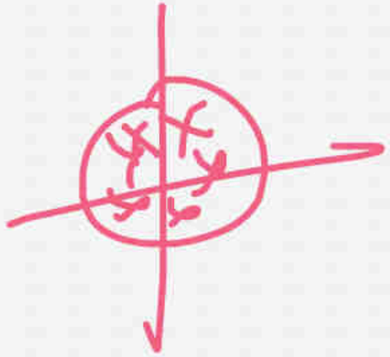
$$\hookrightarrow \mathrm{Var}(x)$$

Hoeffding's covariance identity [edit]

A useful identity to compute the covariance between two random variables X, Y is the Hoeffding's covariance identity:[7]

$$\operatorname{cov}(X,Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} \left(F_{(X,Y)}(x,y) - F_X(x)F_Y(y)\right)\, dx\, dy$$

where $F_{(X,Y)}(x,y)$ is the joint cumulative distribution function of the random vector (X,Y) and $F_X(x), F_Y(y)$ are the marginals.



Uncorrelatedness and independence [edit]

Main article: Correlation and dependence

Random variables whose covariance is zero are called uncorrelated.[4]:p. 121 Similarly, the components of random vectors whose covariance matrix is zero in every entry outside the main diagonal are also called uncorrelated.

If X and Y are independent random variables, then their covariance is zero.[4]:p. 123[8] This follows because under independence,

$$\operatorname{E}[XY] = \operatorname{E}[X] \cdot \operatorname{E}[Y].$$

The converse, however, is not generally true. For example, let X be uniformly distributed in $[-1, 1]$ and let $Y = X^2$. Clearly, X and Y are not independent, but

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \operatorname{cov}\left(X, X^2\right) \\
&= \operatorname{E}\left[X \cdot X^2\right] - \operatorname{E}[X] \cdot \operatorname{E}\left[X^2\right] \\
&= \operatorname{E}\left[X^3\right] - \operatorname{E}[X]\operatorname{E}\left[X^2\right] \\
&= 0 - 0 \cdot \operatorname{E}[X^2] \\
&= 0.
\end{aligned}$$

In this case, the relation[illegible] ce are measures of linear dependence between two random variables. This example shows that if two random variables are uncorrelated, that does not in general imply that they are

Hoeffding's covariance identity [edit]

A useful identity to compute the covariance between two random variables X, Y is the Hoeffding's covariance identity:[7]

$$\operatorname{cov}(X,Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} \left(F_{(X,Y)}(x,y) - F_X(x)F_Y(y)\right)\, dx\, dy$$

where $F_{(X,Y)}(x,y)$ is the joint cumulative distribution function of the random vector (X,Y) and $F_X(x), F_Y(y)$ are the marginals.

Uncorrelatedness and independence [edit]

Main article: Correlation and dependence

Random variables whose covariance is zero are called uncorrelated.[4]:p. 121 Similarly, the components of random vectors whose covariance matrix is zero in every entry outside the main diagonal are also called uncorrelated.

If X and Y are independent random variables, then their covariance is zero.[4]:p. 123[8] This follows because under independence,

$$\operatorname{E}[XY] = \operatorname{E}[X] \cdot \operatorname{E}[Y].$$

The converse, however, is not generally true. For example, let X be uniformly distributed in $[-1, 1]$ and let $Y = X^2$. Clearly, X and Y are not independent, but

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \operatorname{cov}\left(X, X^2\right) \\
&= \operatorname{E}\left[X \cdot X^2\right] - \operatorname{E}[X] \cdot \operatorname{E}\left[X^2\right] \\
&= \operatorname{E}\left[X^3\right] - \operatorname{E}[X]\operatorname{E}\left[X^2\right] \\
&= 0 - 0 \cdot \operatorname{E}[X^2] \\
&= 0.
\end{aligned}$$

In this case, the relation[illegible]ce are measures of linear dependence between two random variables. This example shows that if two random variables are uncorrelated, that does not in general imply that they are

$$\operatorname{cov}(X,Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} \left(F_{(X,Y)}(x,y) - F_X(x)F_Y(y)\right)\,dx\,dy$$

where $F_{(X,Y)}(x,y)$ is the joint cumulative distribution function of the random vector (X,Y) and $F_X(x), F_Y(y)$ are the marginals.

Uncorrelatedness and independence [edit]

Main article: Correlation and dependence

Random variables whose covariance is zero are called uncorrelated.[4]:p. 121 Similarly, the components of random vectors whose covariance matrix is zero in every entry outside the main diagonal are also called uncorrelated.

If X and Y are independent random variables, then their covariance is zero.[4]:p. 123[8] This follows because under independence,

$$\operatorname{E}[XY] = \operatorname{E}[X] \cdot \operatorname{E}[Y].$$

The converse, however, is not generally true. For example, let X be uniformly distributed in $[-1, 1]$ and let $Y = X^2$. Clearly, X and Y are not independent, but

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \operatorname{cov}\left(X, X^2\right) \\
&= \operatorname{E}\left[X \cdot X^2\right] - \operatorname{E}[X] \cdot \operatorname{E}\left[X^2\right] \\
&= \operatorname{E}\left[X^3\right] - \operatorname{E}[X]\operatorname{E}\left[X^2\right] \\
&= 0 - 0 \cdot \operatorname{E}[X^2] \\
&= 0.
\end{aligned}$$

In this case, the relationship between Y and X is non-linear, while correlation and covariance are measures of linear dependence between two random variables. This example shows that if two random variables are uncorrelated, that does not in general imply that they are independent. However, if two variables are jointly normally distributed (but not if they are merely individually normally distributed), uncorrelatedness *does* imply independence.

Relationship to inn

$f_{\mathbb{R}}\, f_{\mathbb{R}}$

where $F_{(X,Y)}(x, y)$ is the joint cumulative distribution function of the random vector (X, Y) and $F_X(x), F_Y(y)$ are the marginals.

Uncorrelatedness and independence [edit]

Main article: Correlation and dependence

Random variables whose covariance is zero are called uncorrelated.[4]:p. 121 Similarly, the components of random vectors whose covariance matrix is zero in every entry outside the main diagonal are also called uncorrelated.

If X and Y are independent random variables, then their covariance is zero.[4]:p. 123[8] This follows because under independence,

$$\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y].$$

The converse, however, is not generally true. For example, let X be uniformly distributed in $[-1, 1]$ and let $Y = X^2$. Clearly, X and Y are not independent, but

$$\begin{aligned}
\operatorname{cov}(X, Y) &= \operatorname{cov}\left(X, X^2\right) \\
&= \mathrm{E}\left[X \cdot X^2\right] - \mathrm{E}[X] \cdot \mathrm{E}\left[X^2\right] \\
&= \mathrm{E}\left[X^3\right] - \mathrm{E}[X]\,\mathrm{E}\left[X^2\right] \\
&= 0 - 0 \cdot \mathrm{E}[X^2] \\
&= 0.
\end{aligned}$$

In this case, the relationship between Y and X is non-linear, while correlation and covariance are measures of linear dependence between two random variables. This example shows that if two random variables are uncorrelated, that does not in general imply that they are independent. However, if two variables are jointly normally distributed (but not if they are merely individually normally distributed), uncorrelatedness *does* imply independence.

Relationship to inner products

Many of the properties of covariance can be extracted elegantly by observing that it satisfies similar properties to those of an inner product:

10:47

Q

wiki

If X & Y are indep then Cov(X,Y) = 0 ✓

If Cov(X,Y) = 0 then X & Y need "not" be independent

↓

both

where $F_{(X,Y)}(x, y)$ is the joint cumulative distribution function of the random vector (X, Y) and $F_X(x), F_Y(y)$ are the marginals.

Uncorrelatedness and independence [edit]

Main article: Correlation and dependence

Random variables whose covariance is zero are called uncorrelated.[4]:p. 121 Similarly, the components of random vectors whose covariance matrix is zero in every entry outside the main diagonal are also called uncorrelated.

If X and Y are independent random variables, then their covariance is zero.[4]:p. 123[8] This follows because under independence,

$$\operatorname{E}[XY] = \operatorname{E}[X] \cdot \operatorname{E}[Y].$$

The converse, however, is not generally true. For example, let X be uniformly distributed in $[-1, 1]$ and let $Y = X^2$. Clearly, X and Y are not independent, but

$$\begin{aligned}
\operatorname{cov}(X, Y) &= \operatorname{cov}\left(X, X^2\right) \\
&= \operatorname{E}\left[X \cdot X^2\right] - \operatorname{E}[X] \cdot \operatorname{E}\left[X^2\right] \\
&= \operatorname{E}\left[X^3\right] - \operatorname{E}[X]\operatorname{E}\left[X^2\right] \\
&= 0 - 0 \cdot \operatorname{E}[X^2] \\
&= 0.
\end{aligned}$$

$Y = X^2$

In this case, the relationship between Y and X is non-linear, while correlation and covariance are measures of linear dependence between two random variables. This example shows that if two random variables are uncorrelated, that does not in general imply that they are independent. However, if two variables are jointly normally distributed (but not if they are merely individually normally distributed), uncorrelatedness *does* imply independence.

Relationship to inn

Correlation

$x \uparrow \quad y \uparrow\downarrow \rightarrow$

$$\begin{cases} Cov(x,y) > 0 \\ Cov(x,y) < 0 \\ Cov(x,y) \approx 0 \end{cases}$$

$-\infty$ to ∞

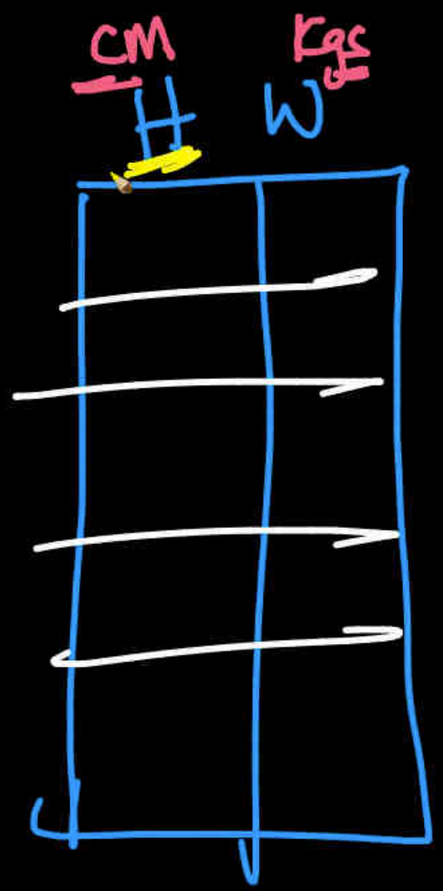
Pearson - CC (x,y)

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

sort of like standardization

$[-1 \quad +1]$

independent of the units of measurement



$Cov(H, W) > 0$

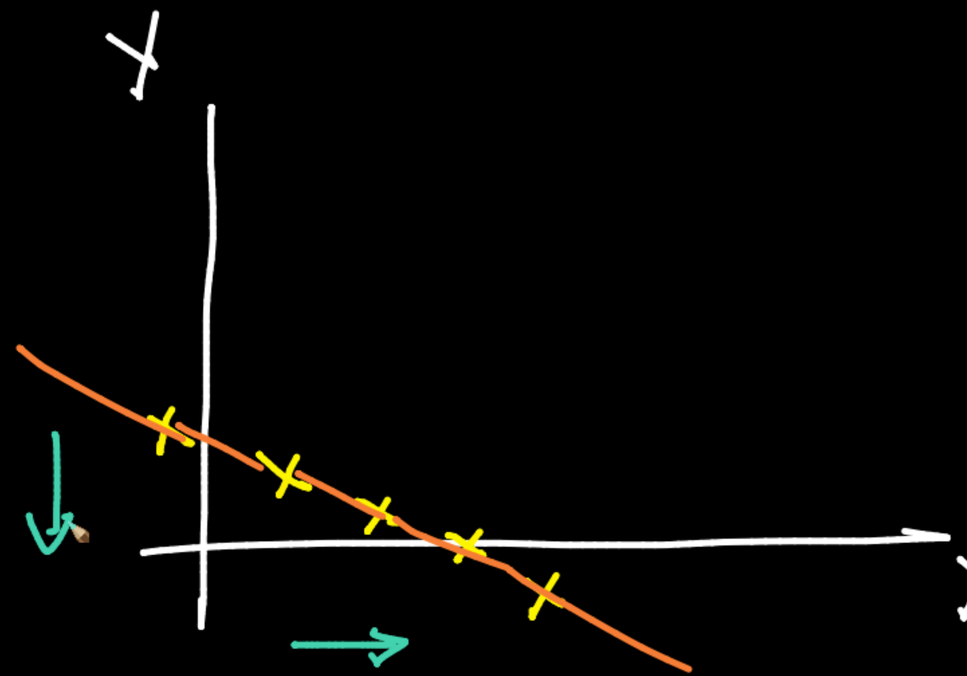
$Cov(HL, HS) > 0$

$$-1 \leq \rho_{X,Y} = \frac{\overbrace{Cov(X,Y)}}{\sigma_X \sigma_Y} < 1$$

Q

need not be

$$\frac{Cov(H_{cm}, W_{kg})}{\sigma_{H_{cm}}\ \sigma_{W_{kg}}} \neq = \frac{Cov(H_{ft}, W_{lbs})}{\sigma_{H_{ft}}\ \sigma_{W_{lbs}}}$$



Q

$\rho_{x,y} =$ -1 ✓

~~close to -1~~

~~-ve not -1~~

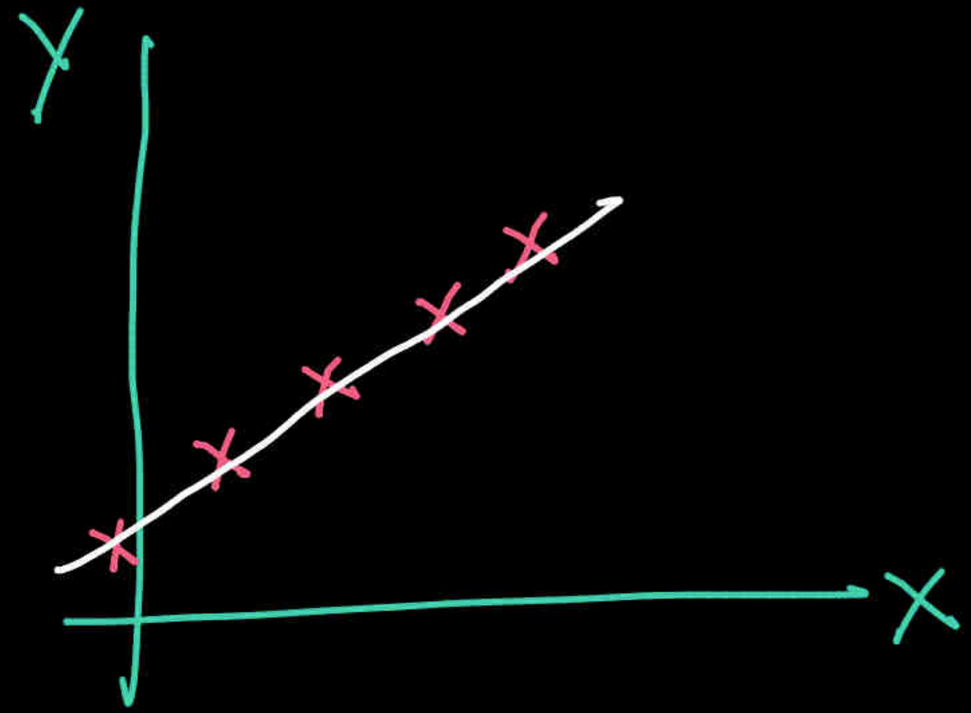
$\rho_{x,y}$

X -1

X & Y are perfectly inv - linearly related

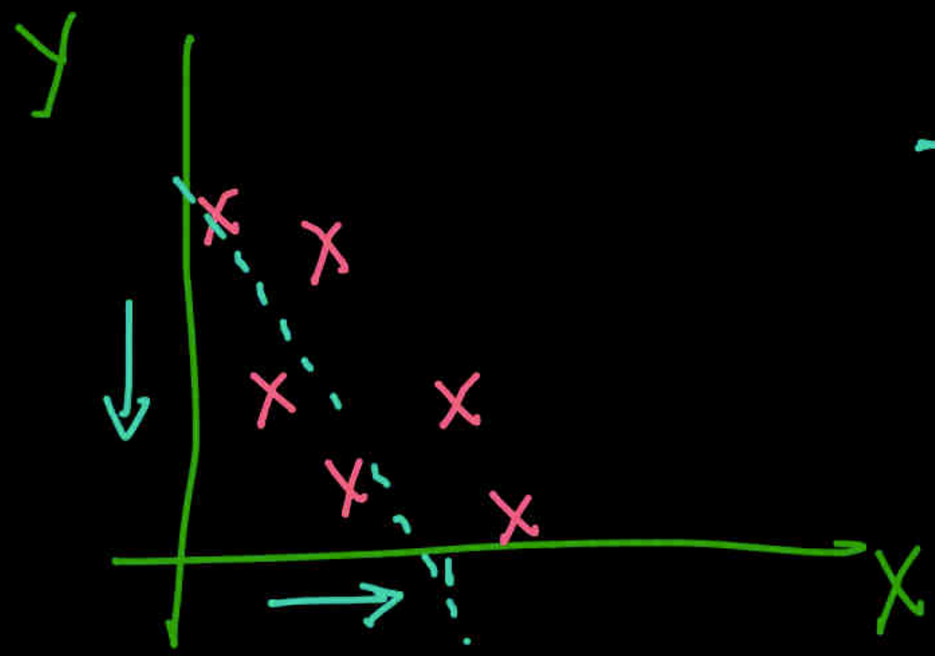
$\Rightarrow$ there exists a line

$$\rho_{cc} = \rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$



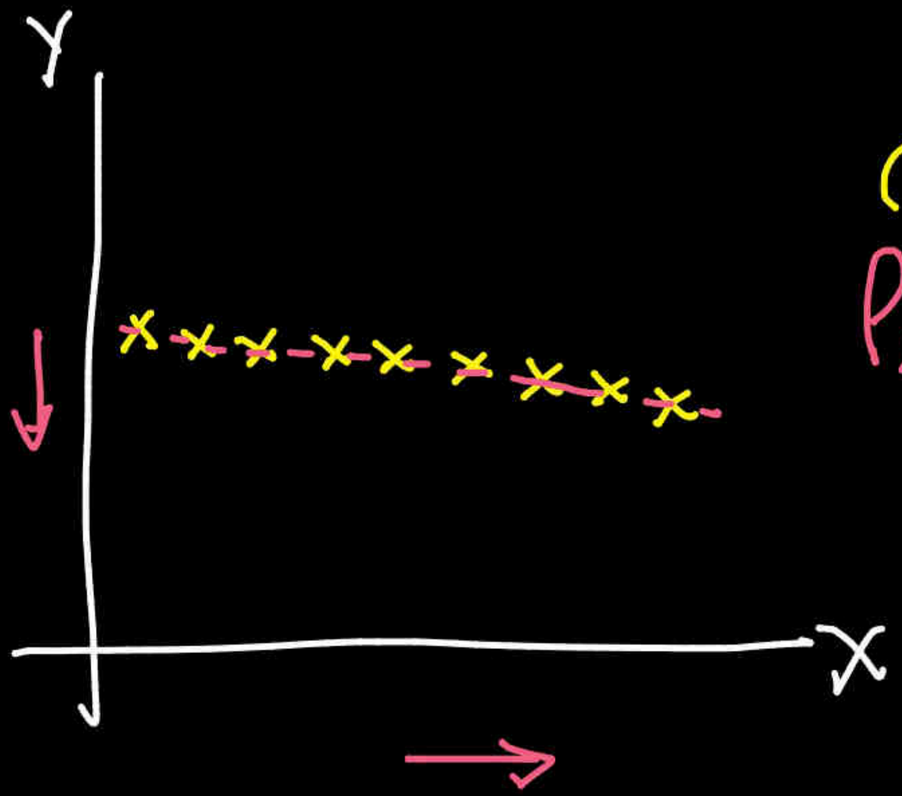
$$\rho_{X,Y} = +1$$

X & Y are perfectly linearly related



$$-1 < \rho_{X,Y} < 0$$

Q

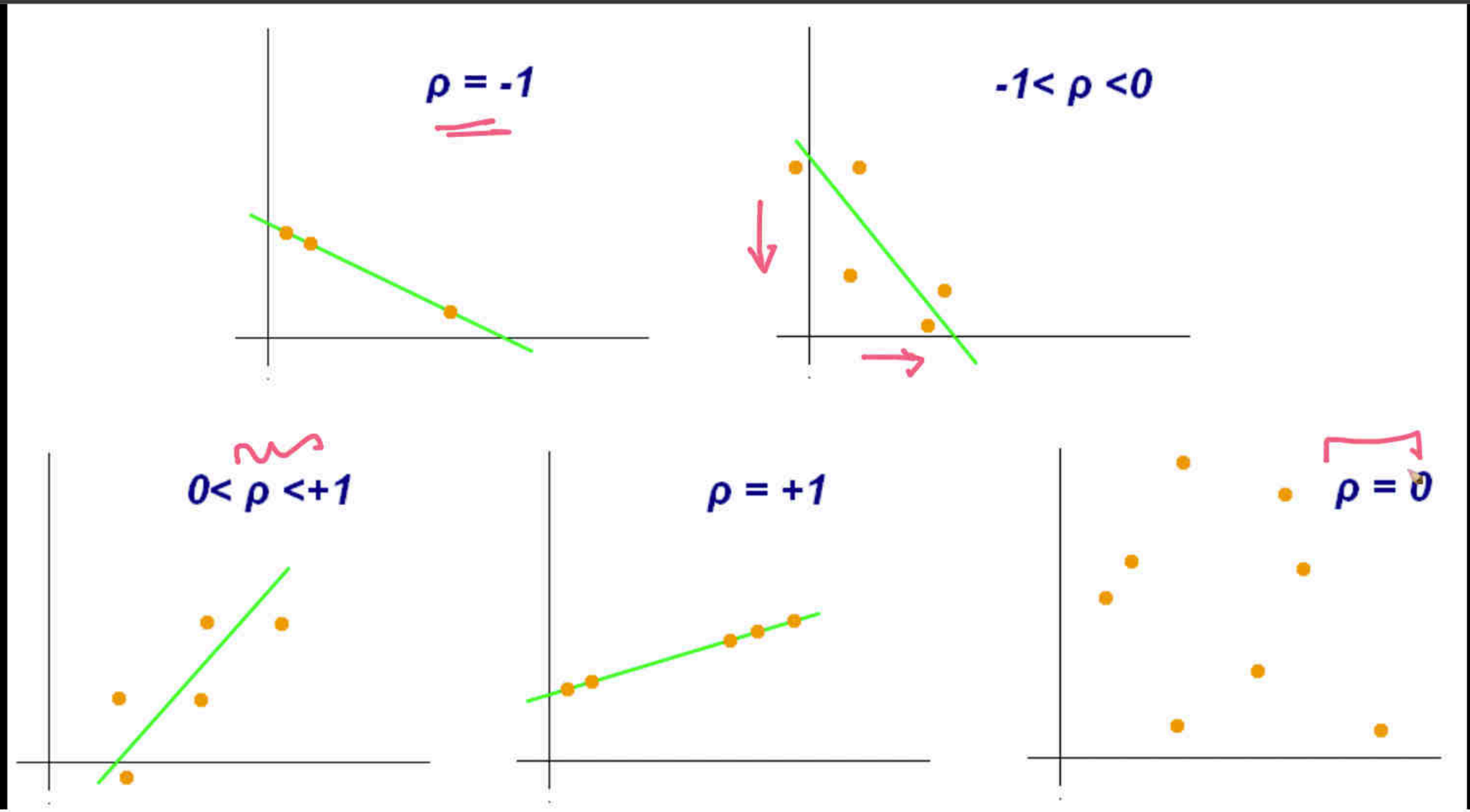


→ not measure the rate of change

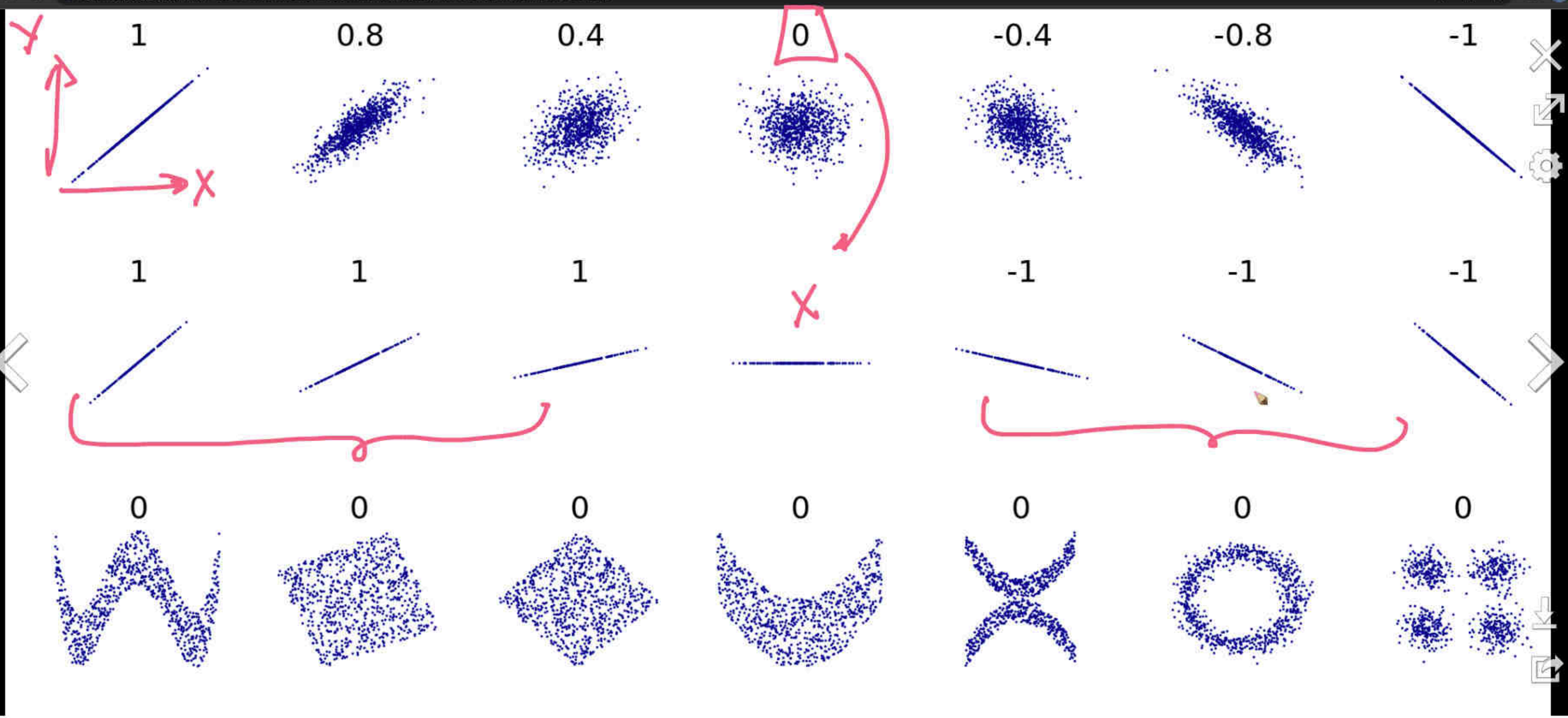
$\rho_{XY} = -1$


Q
Y
X
+1 > ρx,y ≥ 0

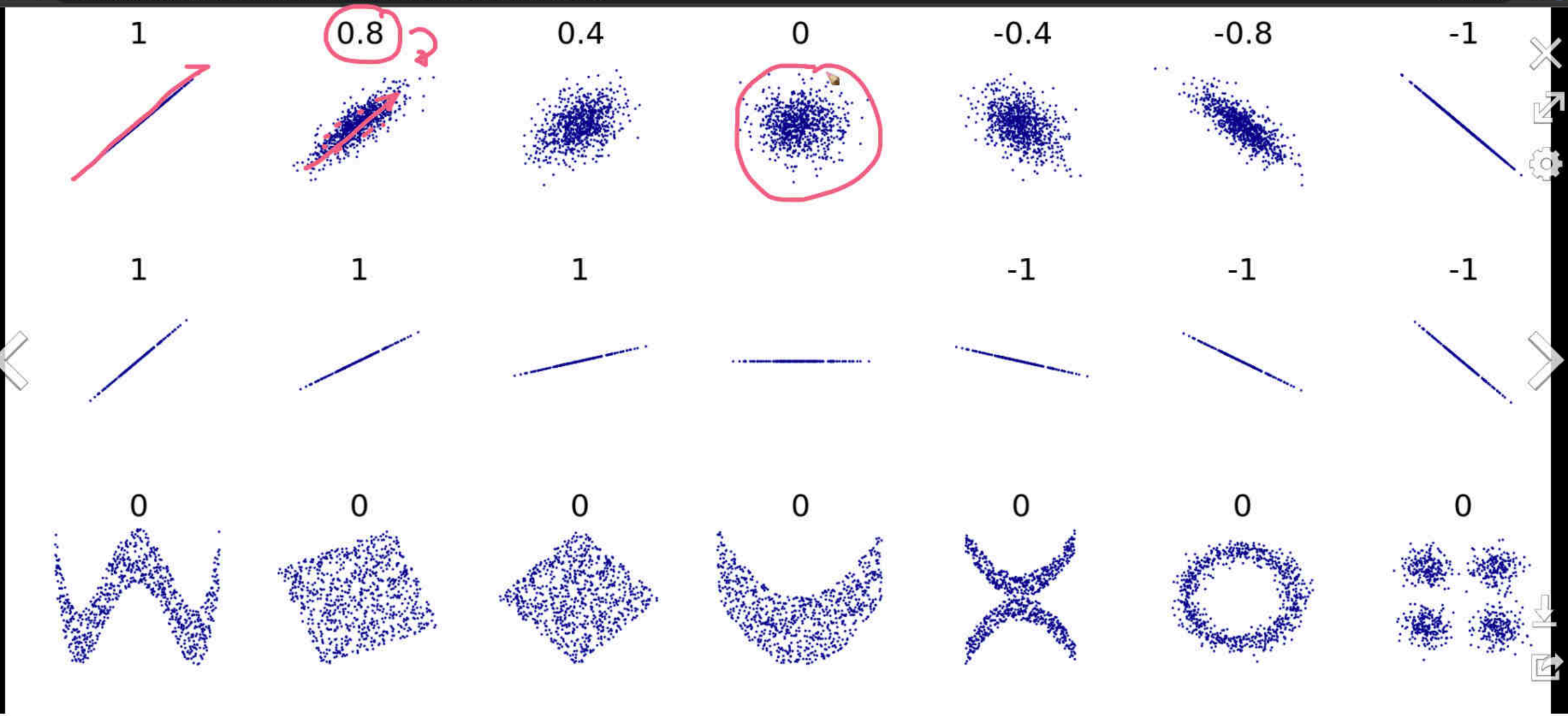

78 / 78



Examples of scatter diagrams with different values of correlation coefficient (ρ)



Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many ...

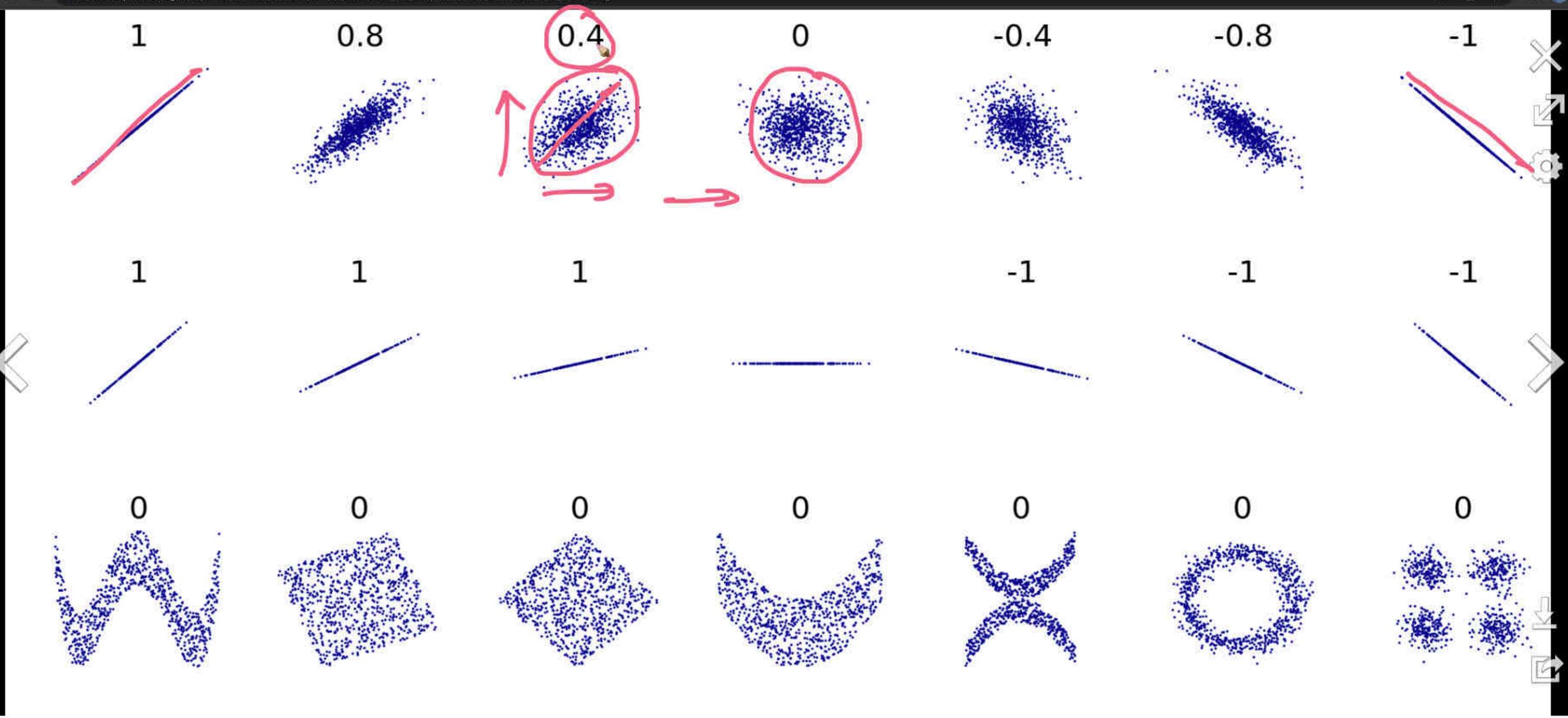


Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many ...

More details

DenisBoigelot, original uploader was Imagecreator

CC0 view terms

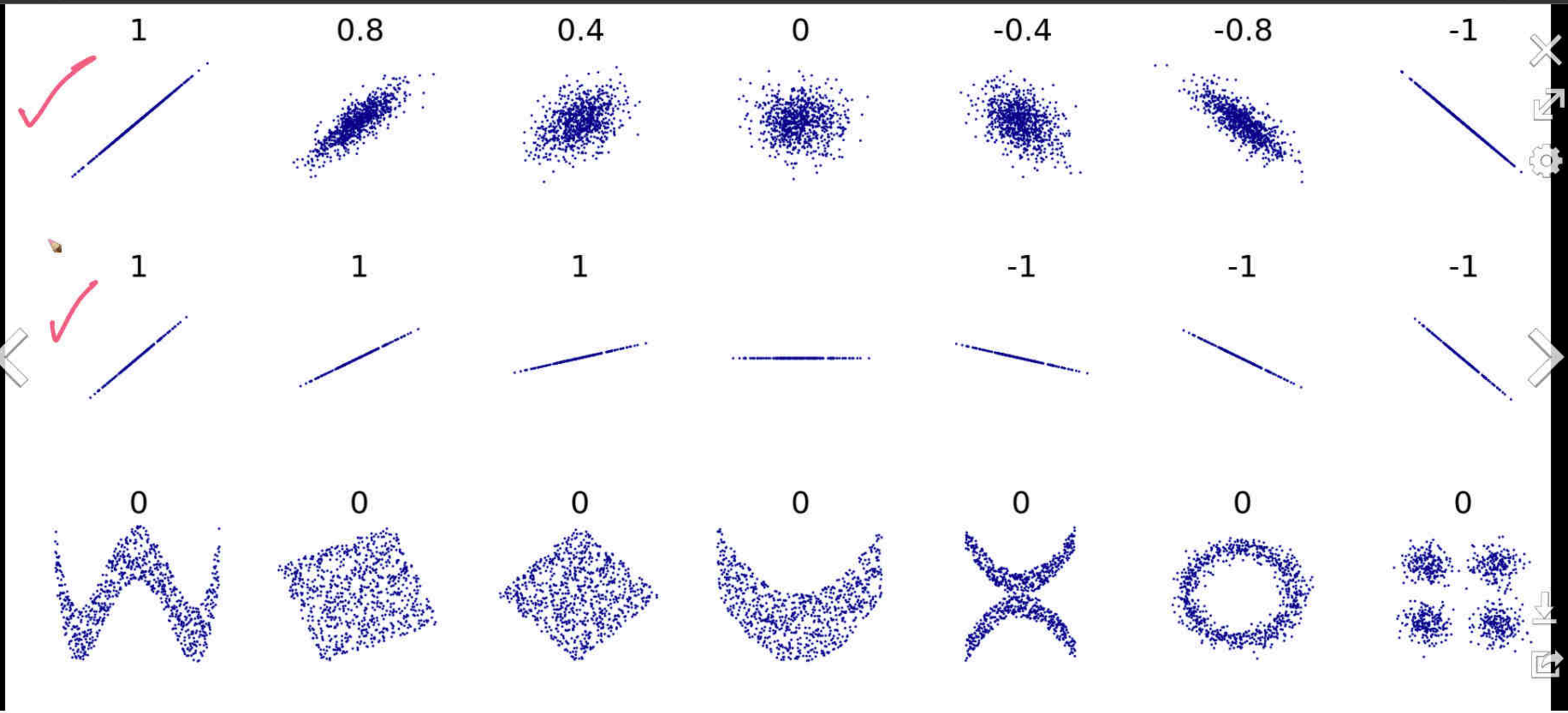


Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many ...

More details

DenisBoigelot, original uploader was Imagecreator

CC0 view terms

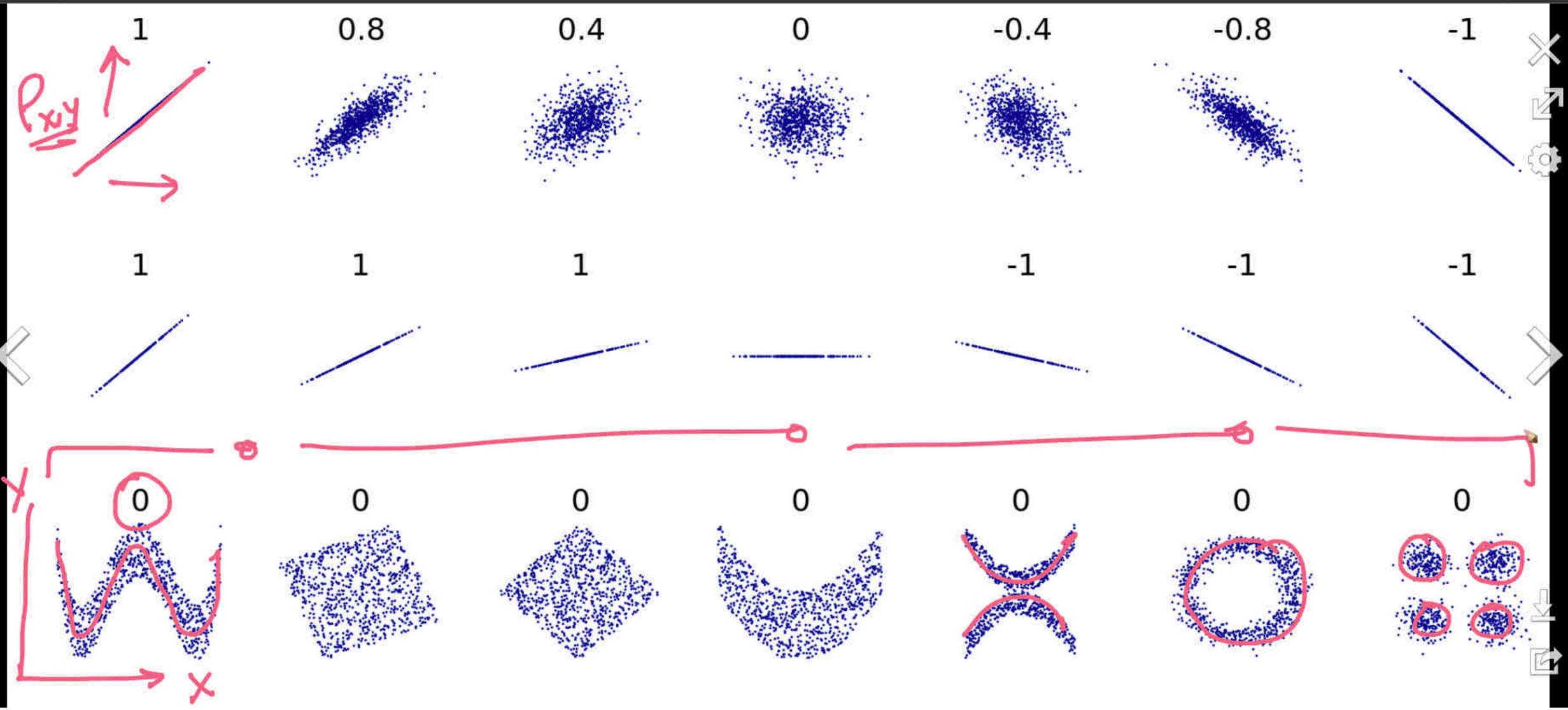


Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many ...

More details

DenisBoigelot, original uploader was Imagecreator

CC0 view terms

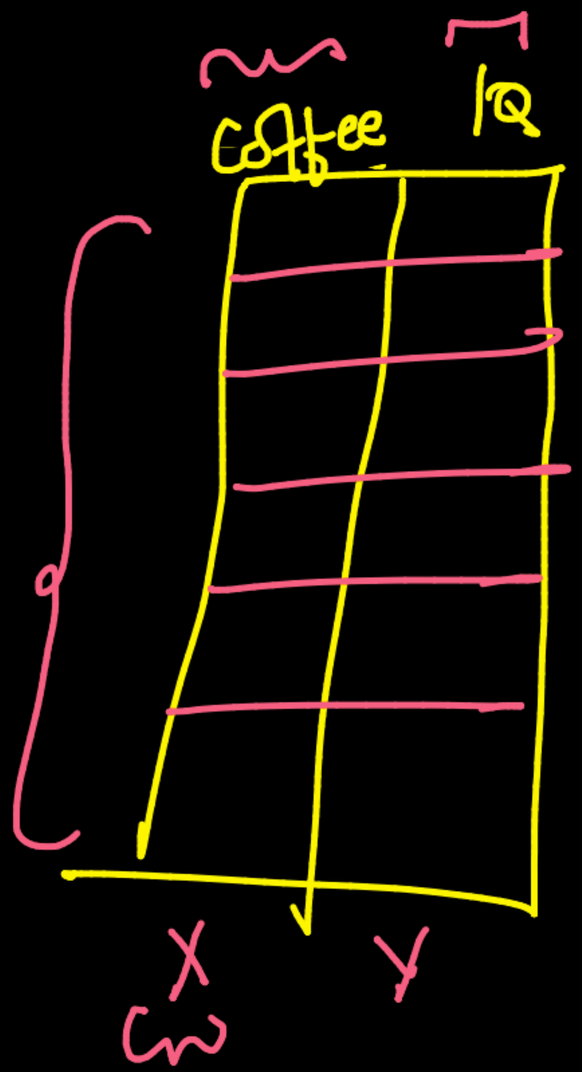


Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many ...

More details

DenisBoigelot, original uploader was Imagecreator

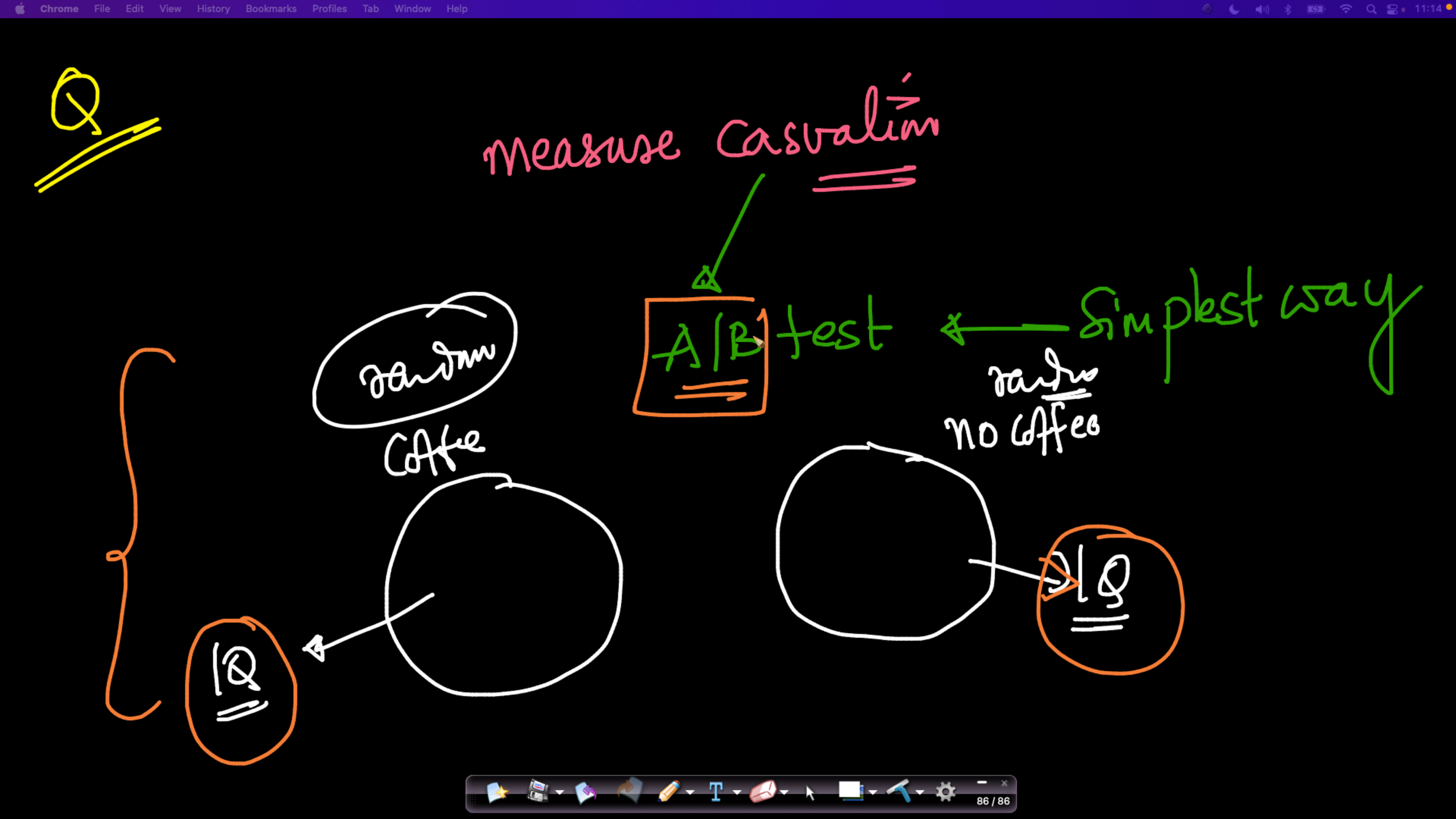
CC0 view terms



$\rho_{x,y} > 0$

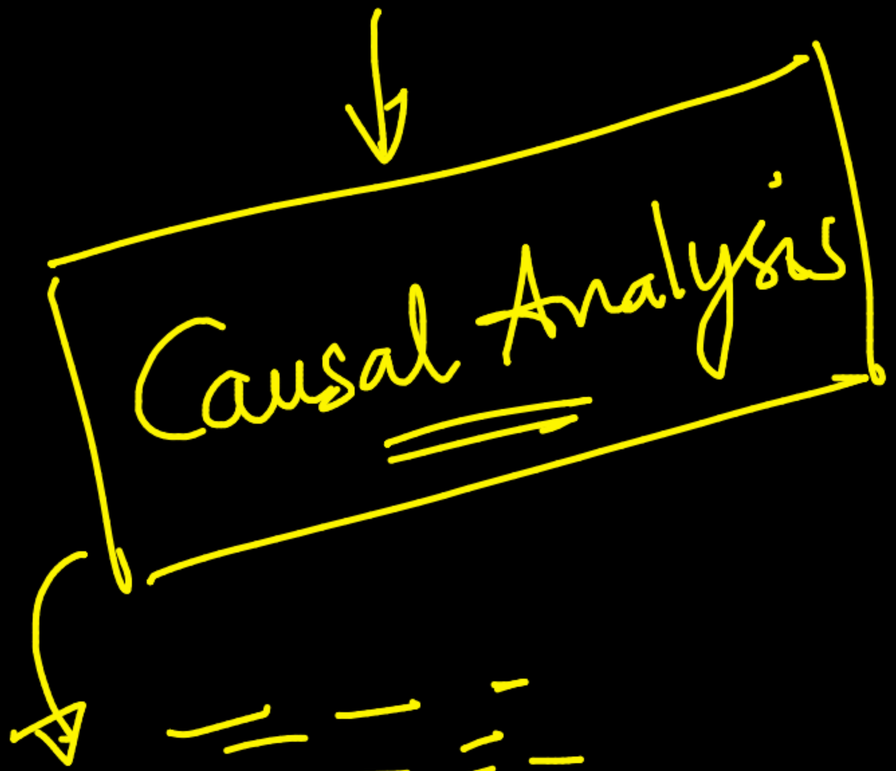
1. Correlation is +ve
2. Coffee increases your IQ

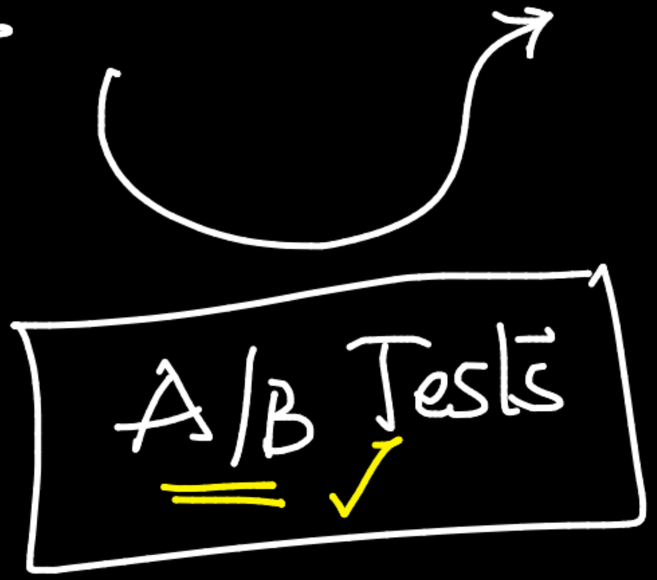
Causation


Measure Casualim
A/B test
Simplest way
random
coffee
random
no coffee
IQ
91Q

Chrome File Edit View History Bookmarks Profiles Tab Window Help

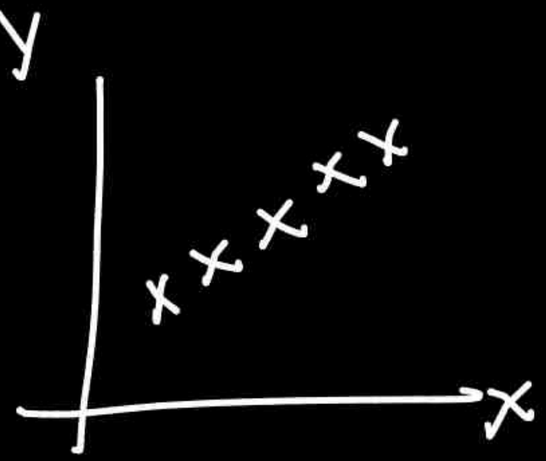
does reminder reduce #Days to recover



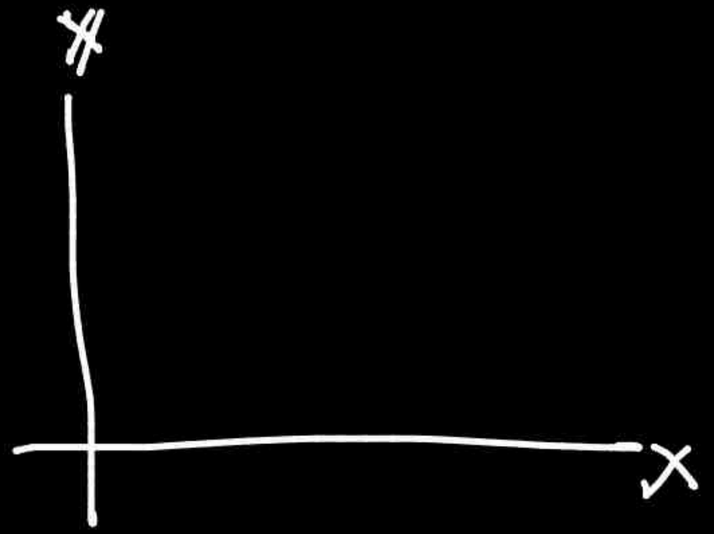


87 / 87

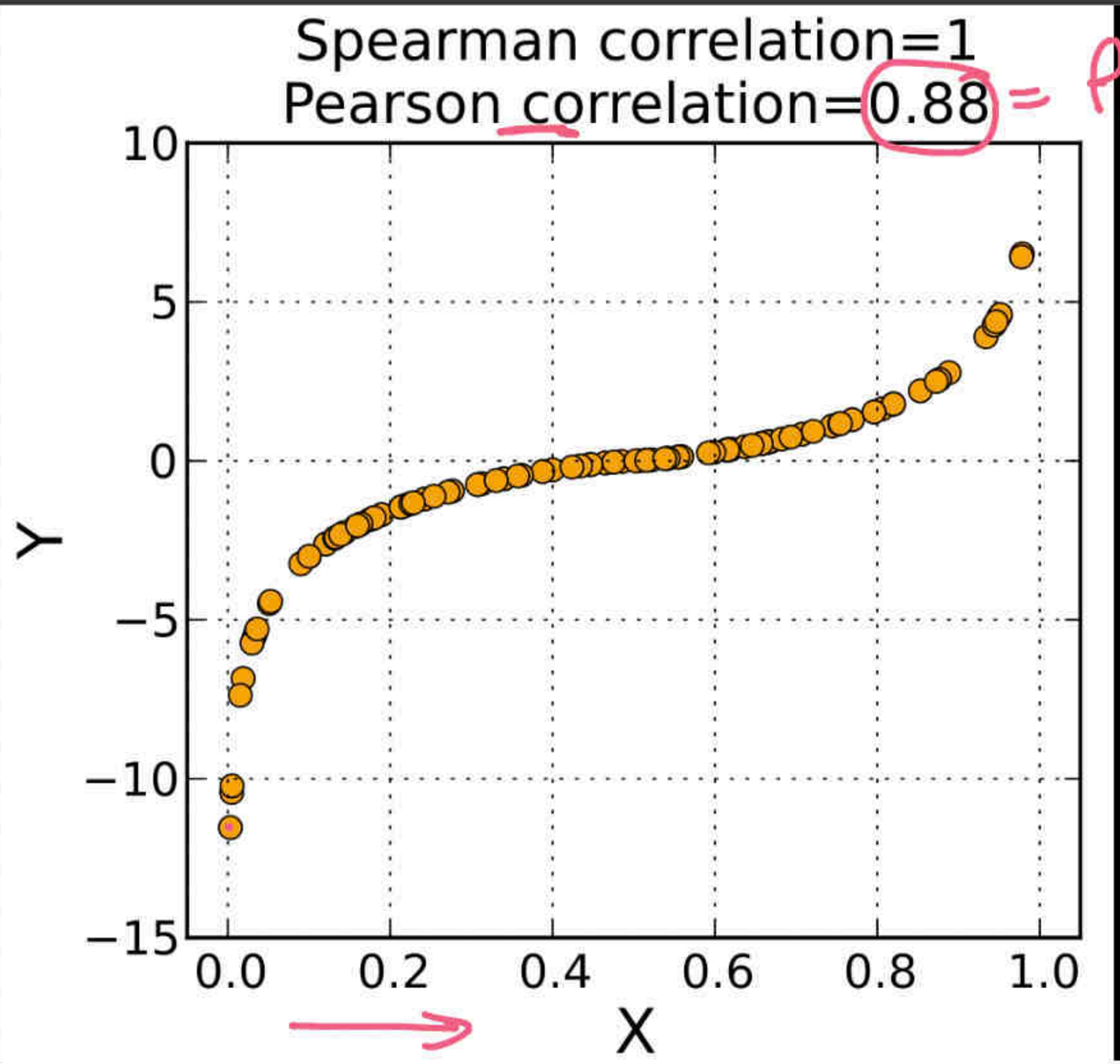
$\rho_{x,y} = 1$



monotonic relationship
need not be linear



$\gamma = 1$



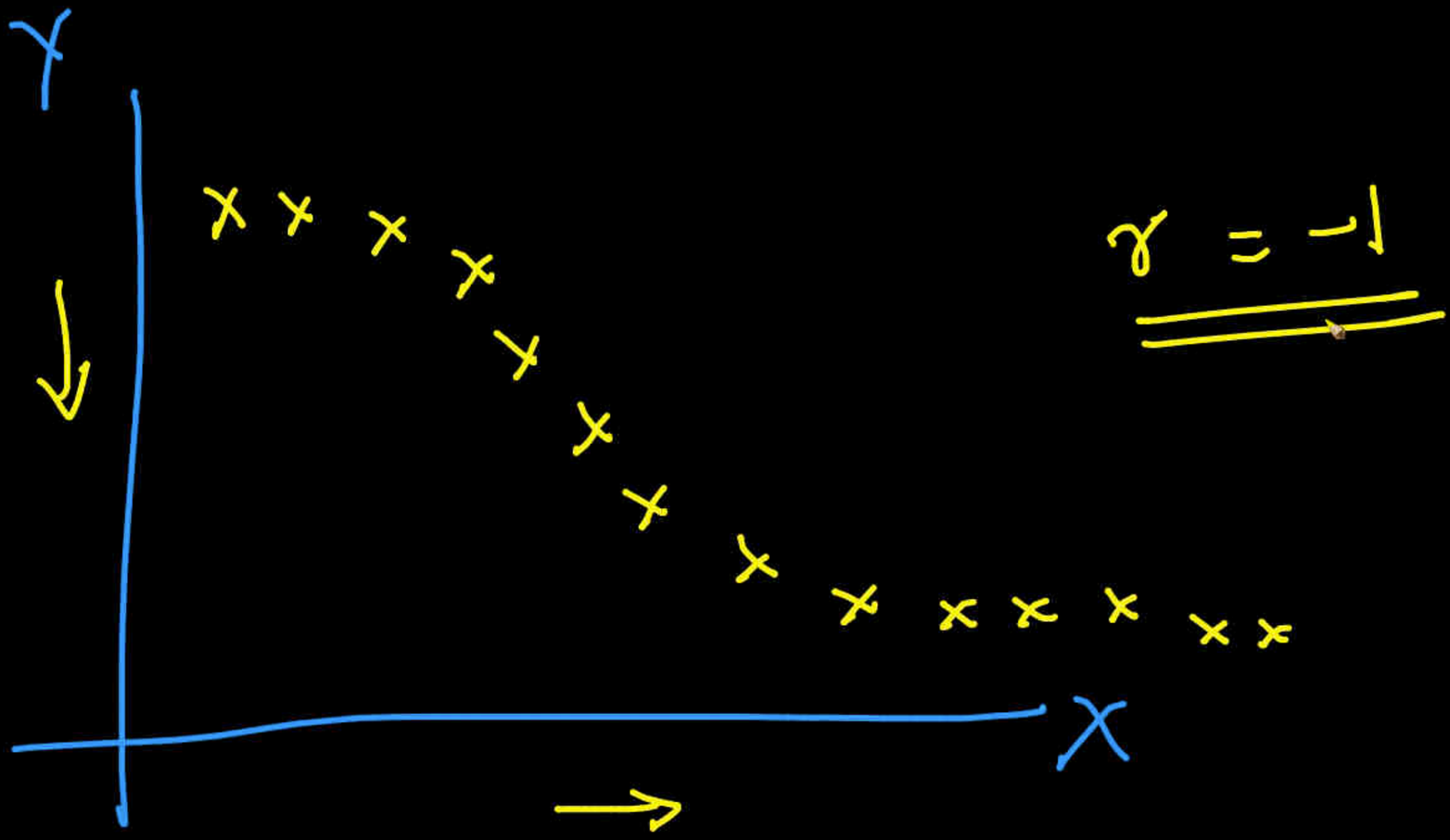
$= \rho_{X,Y}$

monotonic

need not be linear

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 all data points with greater *x* values than that of a given data point will ha ...

More details


Y
X
γ = -1

Spearman Rank CC.

spent in front of TV per week.[*citation needed*]

IQ, X_i	Hours of TV per week, Y_i
106	7
100	27
86	2
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Firstly, evaluate d_i^2. To do so use the following steps, reflected in the table below.

1. Sort the data by the first column (X_i). Create a new column x_i and assign it the ranked values 1, 2, 3, ..., *n*.
2. Next, sort the data by the second column (Y_i). Create a fourth column y_i and similarly assign it the ranked values 1, 2, 3, ..., *n*.
3. Create a fifth column d_i to hold the differences between the two rank columns (x_i and y_i).
4. Create one final column d_i^2 to hold the value of column d_i squared.

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86					

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	–4	16
99	28	3	8	–5	25
100	27	4	7	–3	9
101	50	5	10	–5	25
103	29	6	9	–3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

With d_i^2 found, add them to find $\sum d_i^2 = 194$. The value of *n* is 10. These values can now be substituted back into the equation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

to give

$$\rho = 1 - \frac{6 \times 194}{10(10^2 - 1)},$$

which evaluates to $\rho =$ -distribution).

With d_i^2 found, add them to find $\sum d_i^2 = 194$. The value of n is 10. These values can now be substituted back into the equation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

to give

$$\rho = 1 - \frac{6 \times 194}{10(10^2 - 1)},$$

which evaluates to $\rho = -29/165 = -0.175757575...$ with a *p*-value = 0.627188 (using the *t*-distribution).

That the value is close to zero shows that the correlation between IQ and hours spent watching TV is very low, although the negative value suggests that the longer the time spent watching television the lower the IQ. In the case of ties in the original values, this formula should not be used; instead, the Pearson correlation coefficient should be calculated on the ranks (where ties are given ranks, as described above).

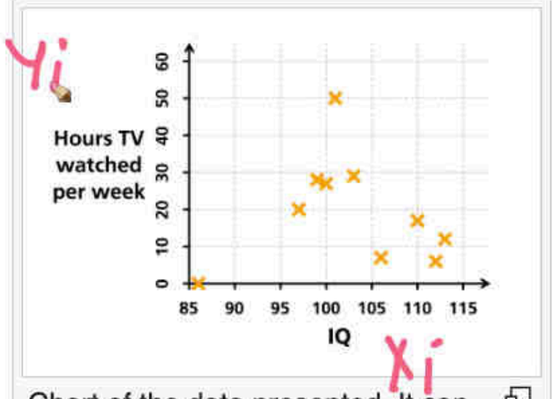


Chart of the data presented. It can be seen that there might be a negative correlation, but that the relationship does not appear definitive.

Determining significance [edit]

One approach to test whether an observed value of ρ is significantly different from zero (r will always maintain $-1 \le r \le 1$) is to calculate the probability that it would be greater than or equal to the observed r, given the null hypothesis, by using a permutation test. An advantage of this approach is that it automatically takes into account the number of tied data values in the sample and the way they are treated in computing the rank correlation.

Another approach parallels the use of the Fisher transformation in the case of the Pearson product-moment correlation coefficient. That is, confidence intervals and [illegible]ut using the Fisher transformation:

1. Sort the data by the first column (X_i). Create a new column x_i and assign it the ranked values 1, 2, 3, ..., *n*.
2. Next, sort the data by the second column (Y_i). Create a fourth column y_i and similarly assign it the ranked values 1, 2, 3, ..., *n*.
3. Create a fifth column d_i to hold the differences between the two rank columns (x_i and y_i).
4. Create one final column d_i^2 to hold the value of column d_i squared.

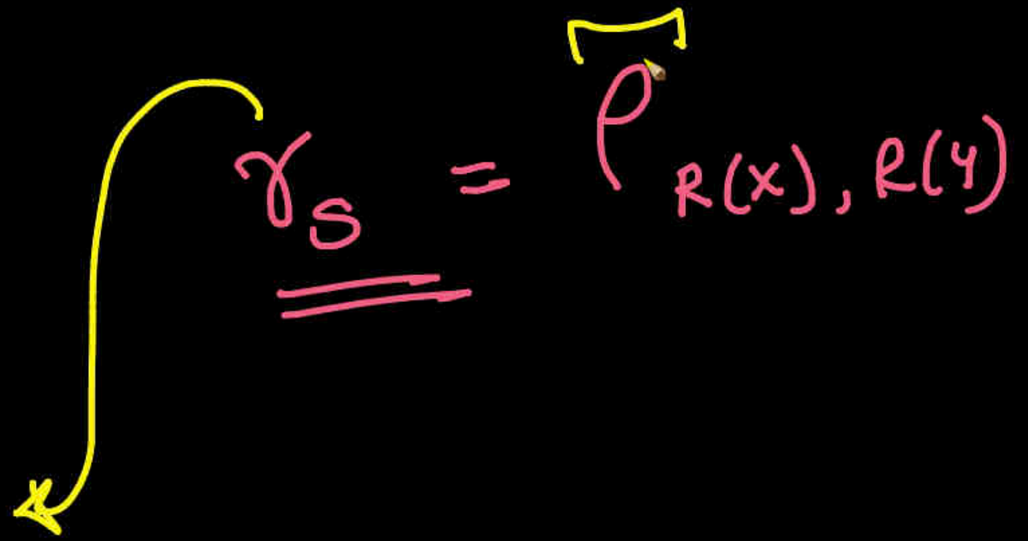
IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	–4	16
99	28	3	8	–5	25
100	27	4	7	–3	9
101	50	5	10	–5	25
103	29	6	9	–3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

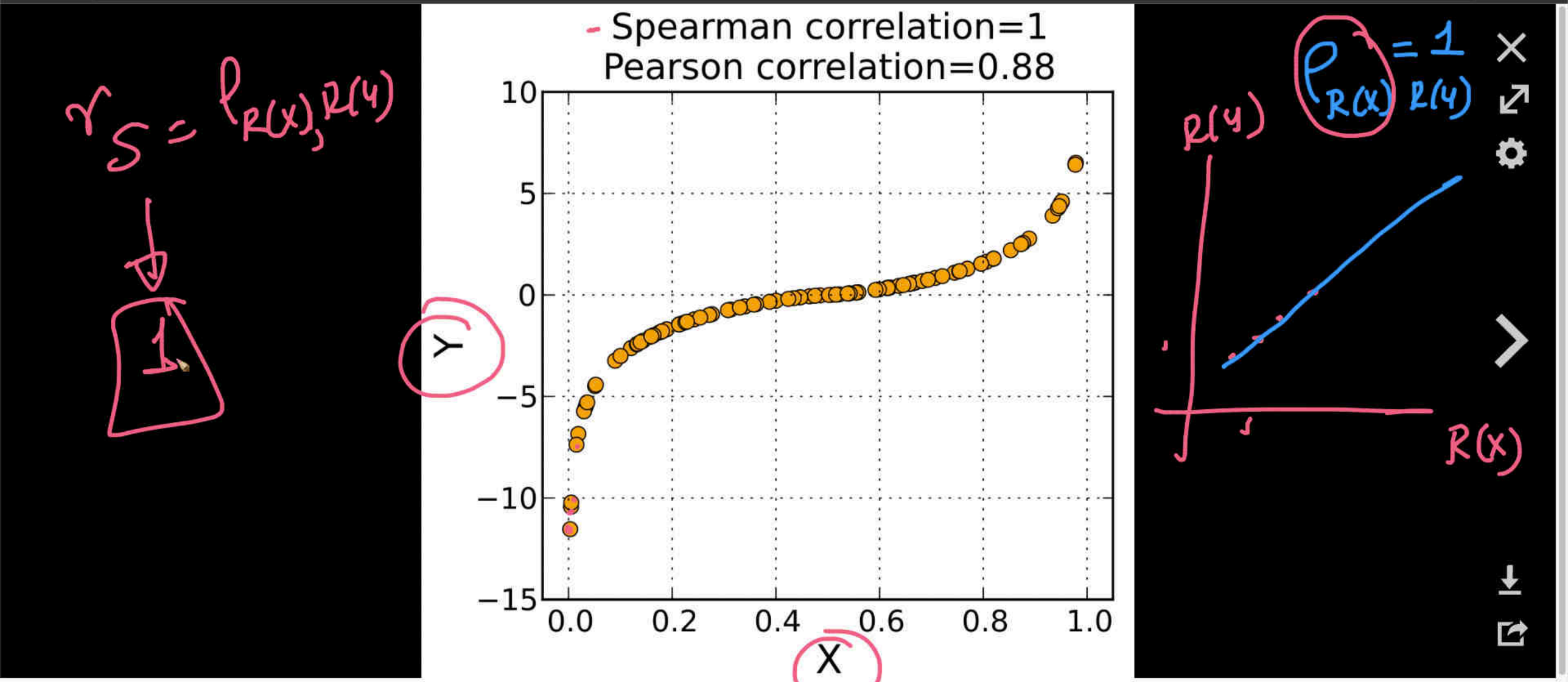
$\rho_{x,y}$

With d_i^2 found, add them to find $\sum d_i^2 = 194$. The value of *n* is 10. These values can now be substituted back into the equation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

to give



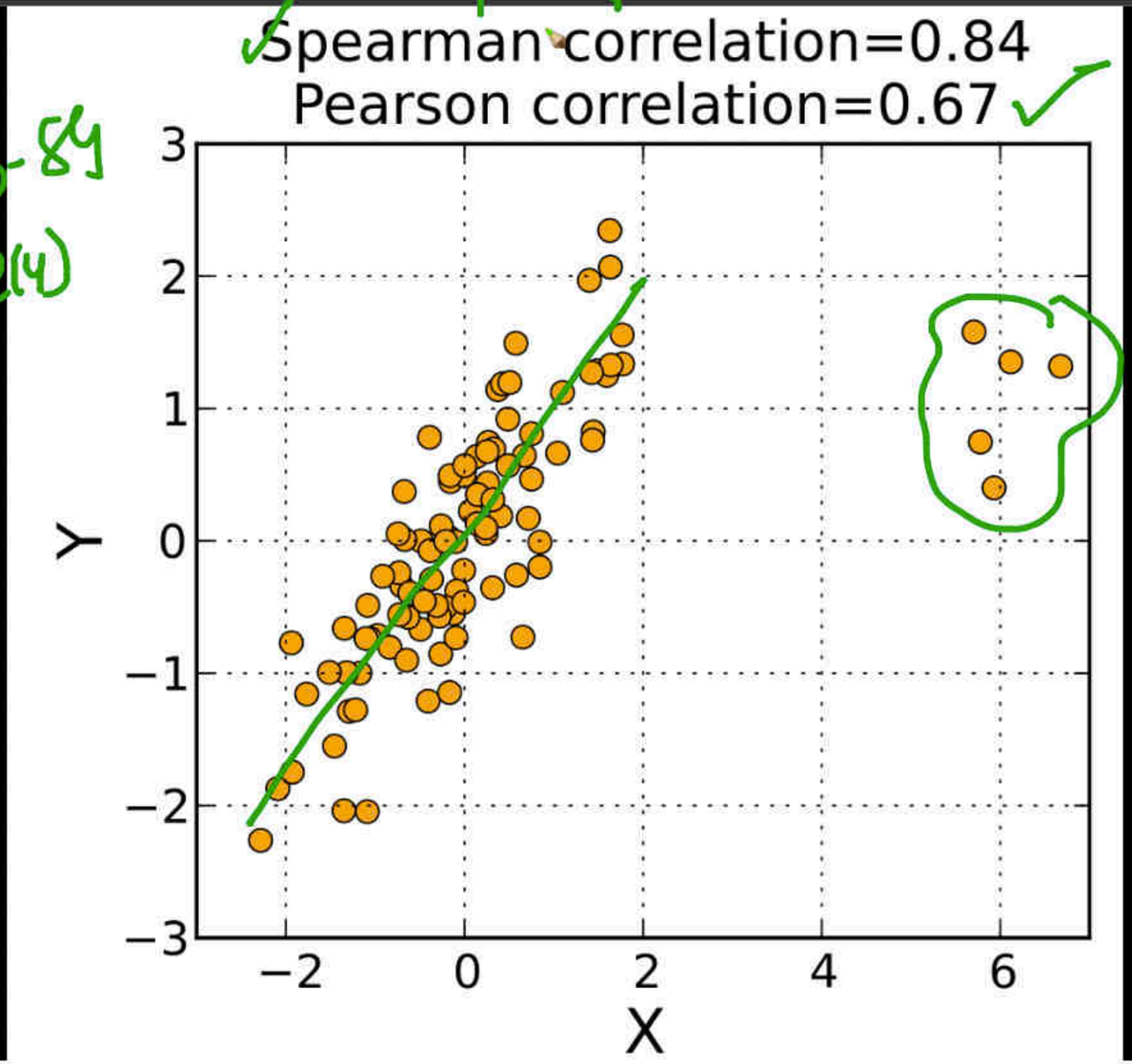


A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater *x* values than that of a given data point will ha ...

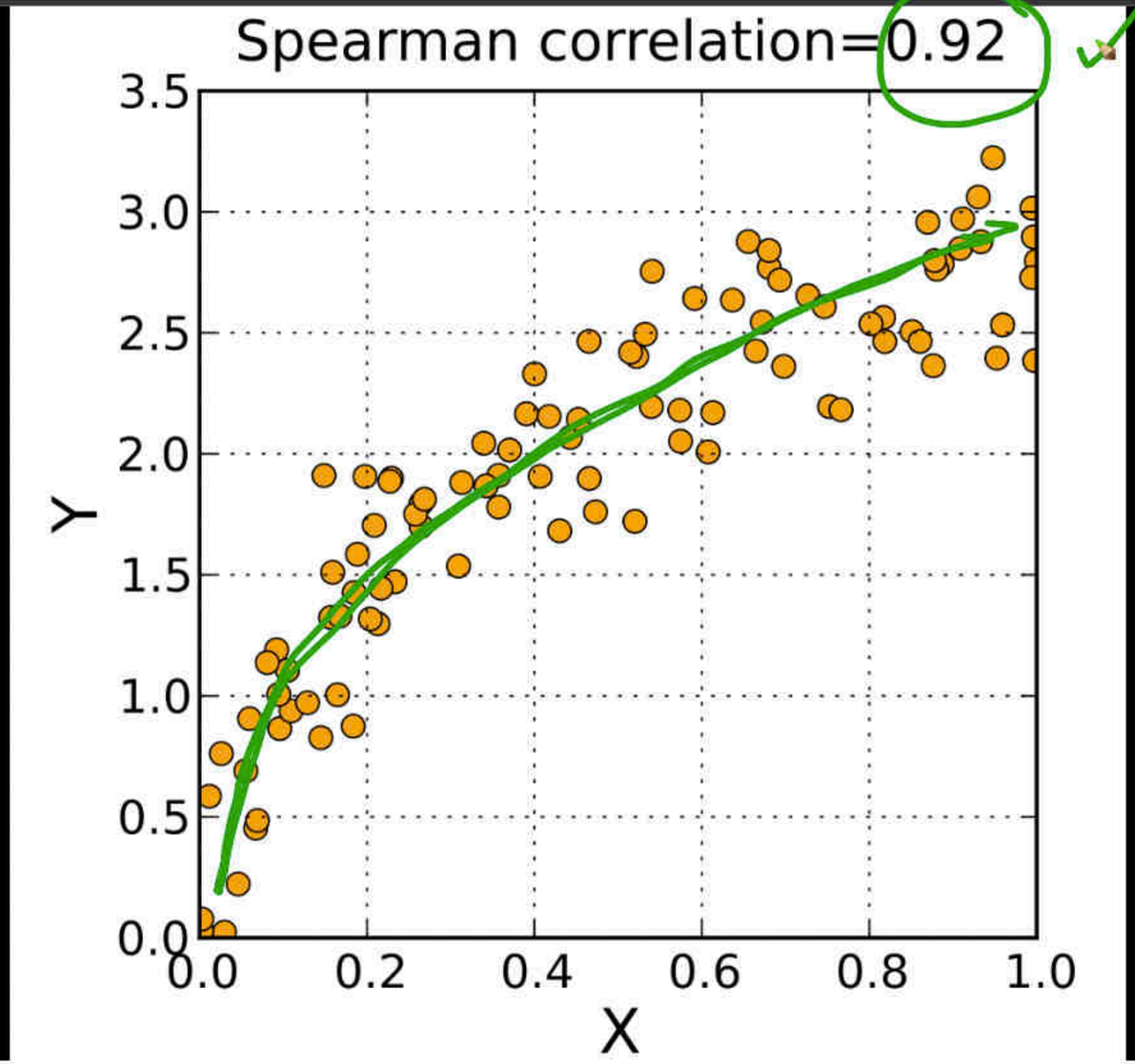
More details

$\rho_{X,Y} = 0.67$

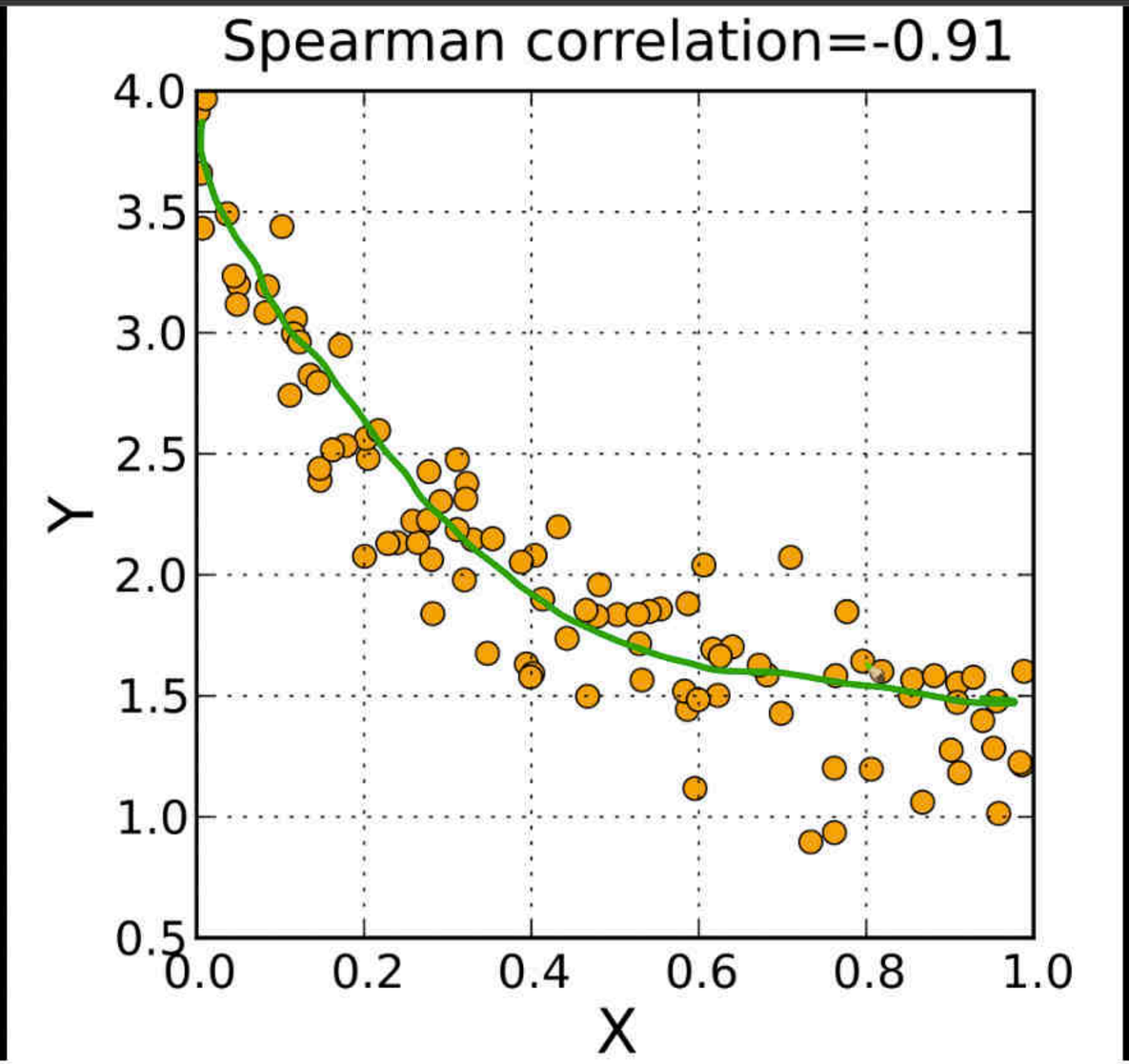
$r_{X,Y} = \rho_{R(X),R(Y)} = 0.84$



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.

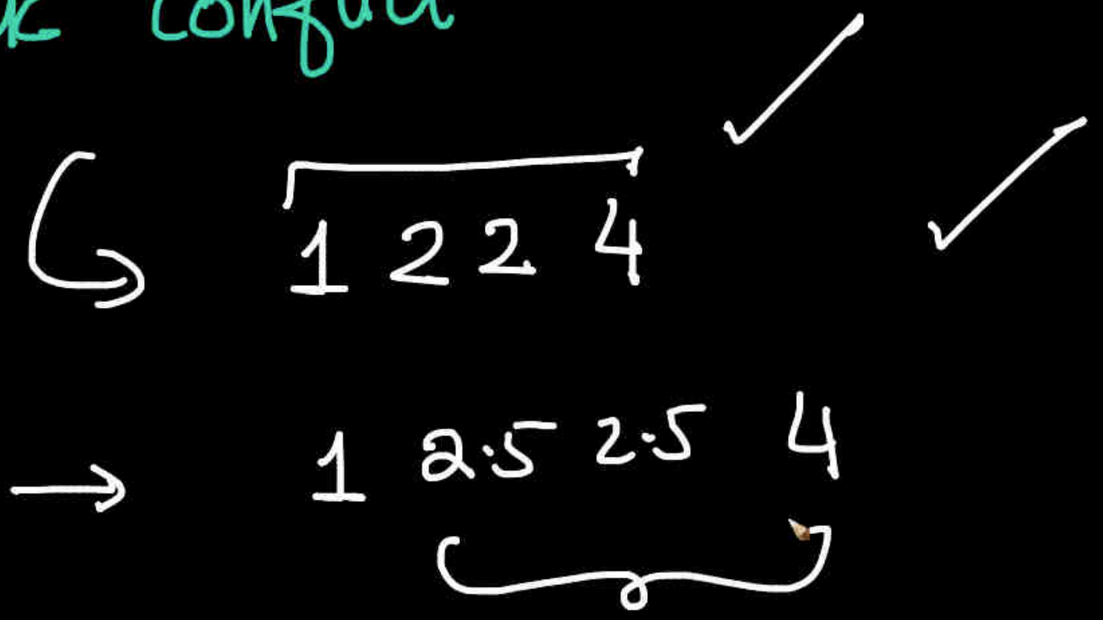


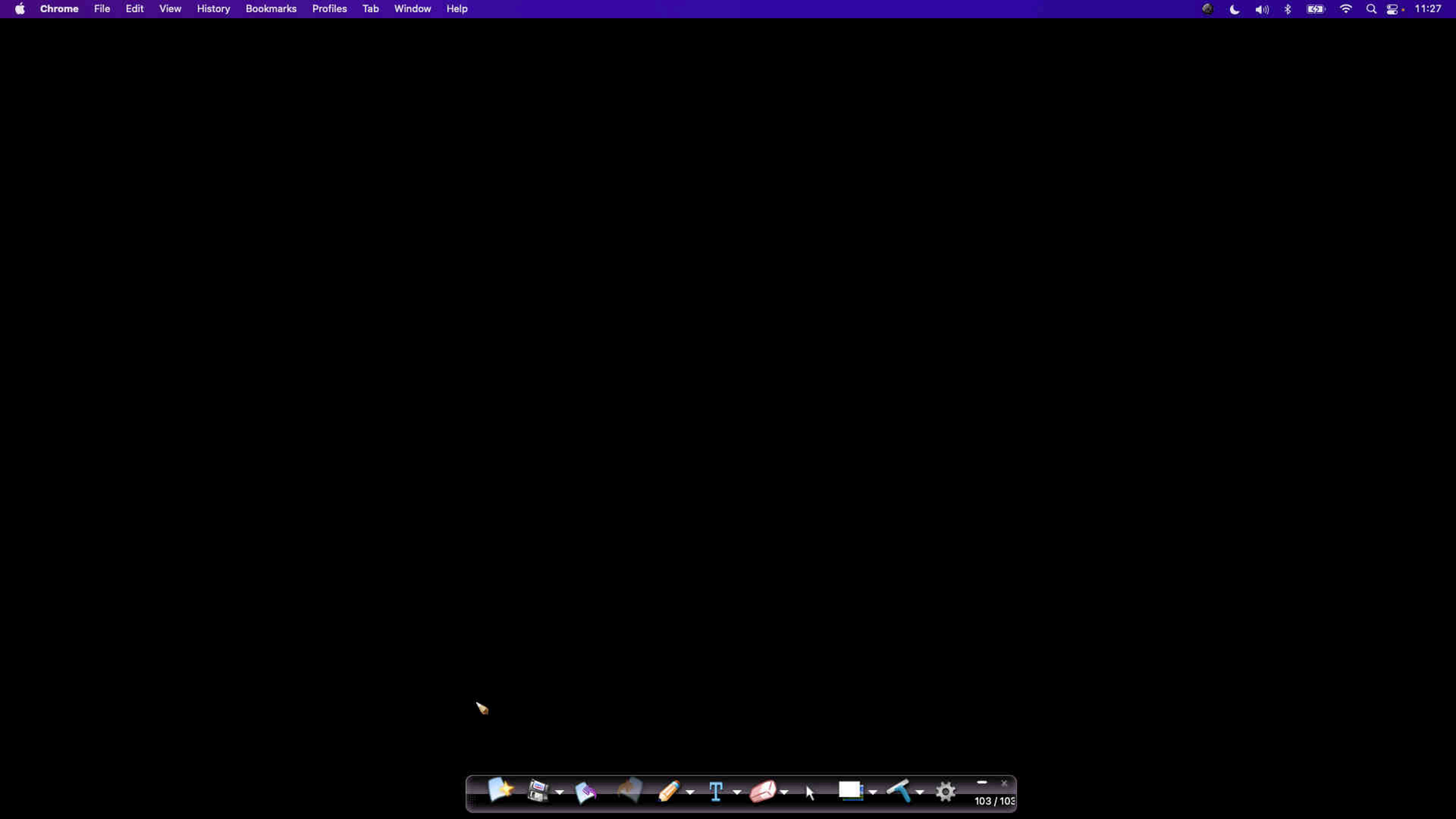
A positive Spearman correlation coefficient corresponds to an increasing monotonic trend between *X* and *Y*.



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

rank conflict



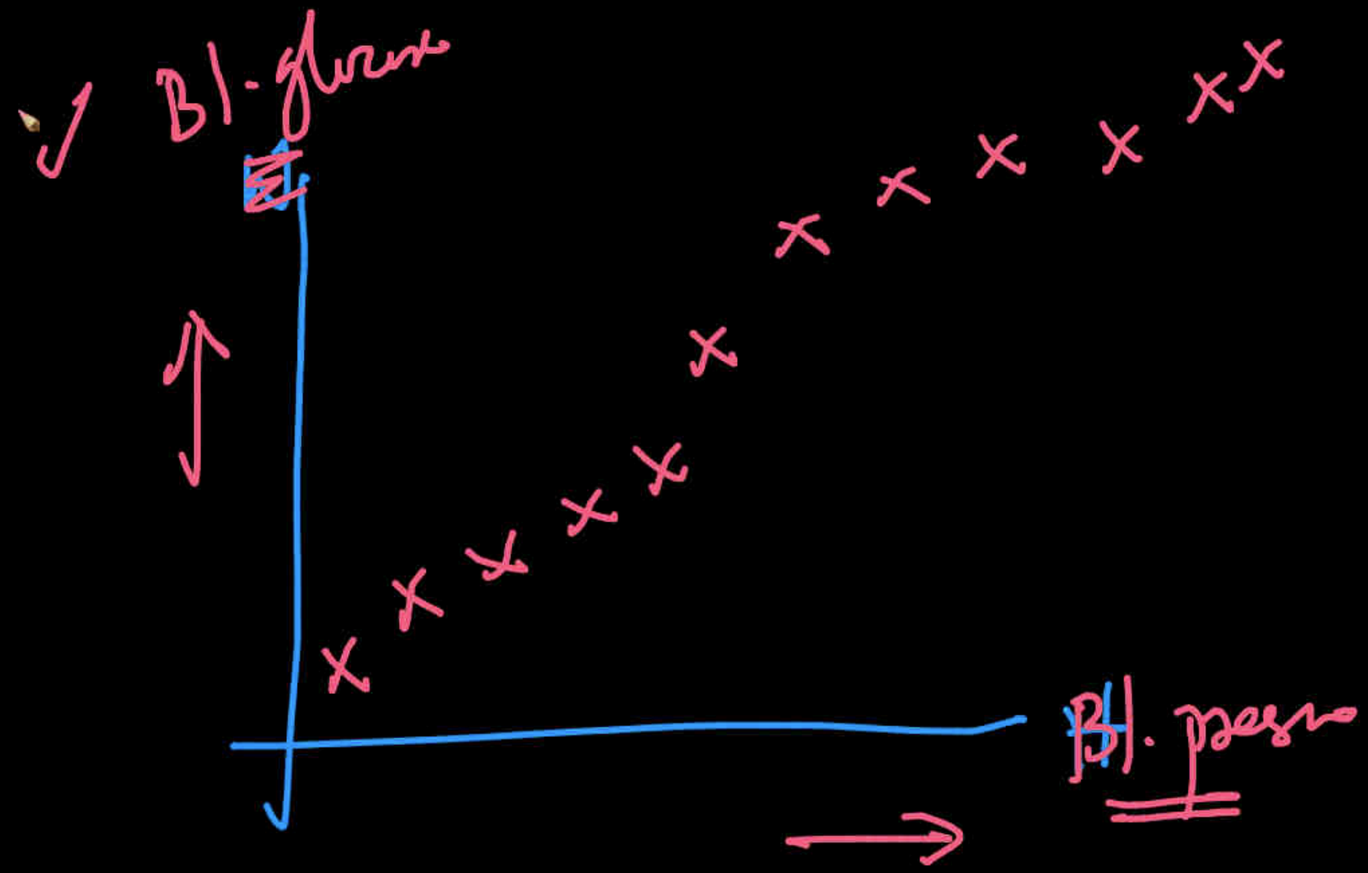

Chrome File Edit View History Bookmarks Profiles Tab Window Help
11:27
103 / 103

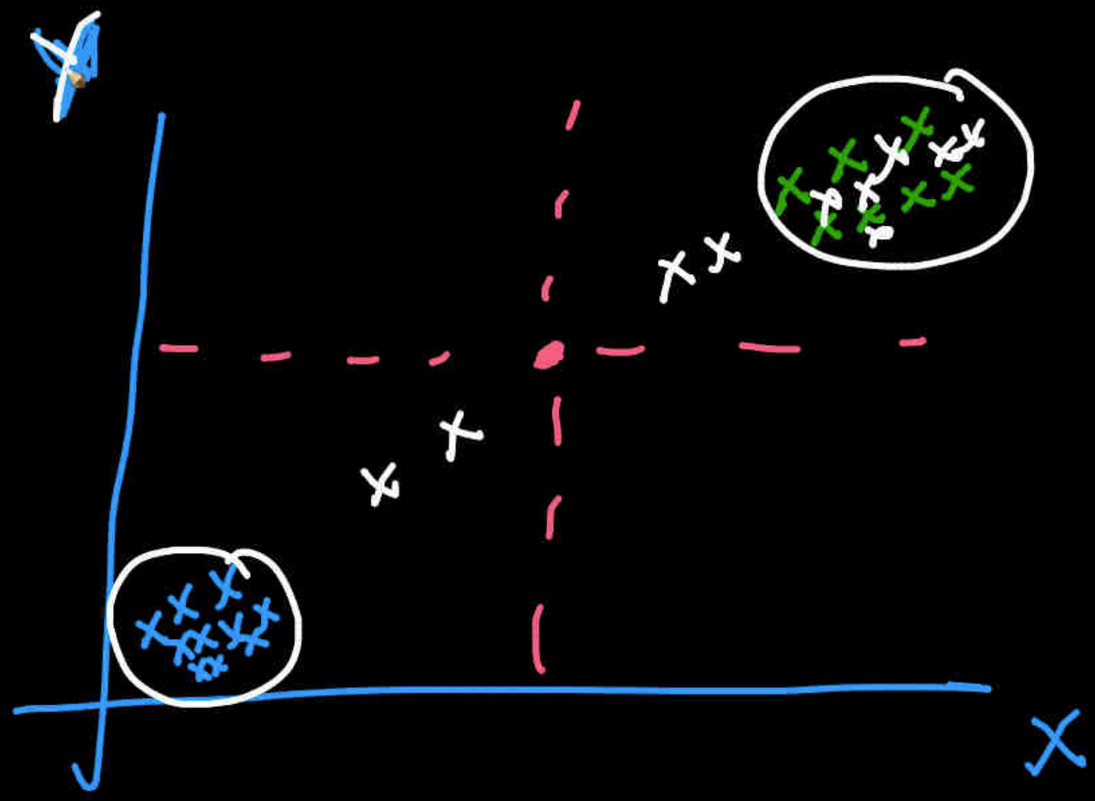
$-\infty \leq Cov(X,Y) \leq \infty \longrightarrow$ As X changes do Y change?

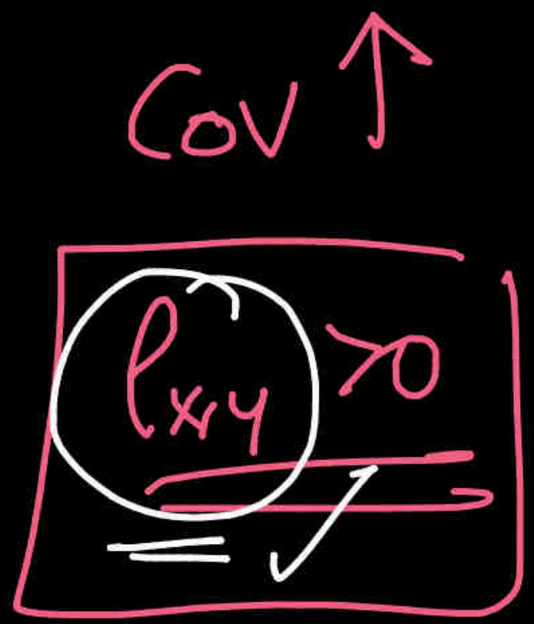
$-1 \leq \rho_{X,Y} \leq +1 \longrightarrow$ linear relationship

✓ $-1 \leq \gamma_{X,Y} \leq +1 \longrightarrow$ monotonicity

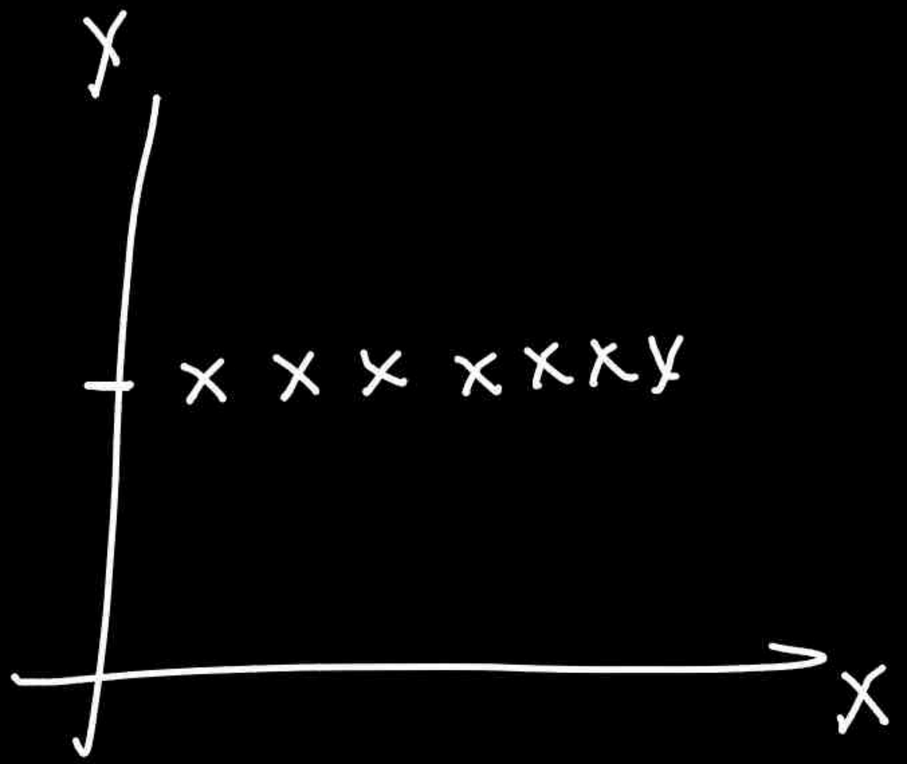
↑ less prone to outliers


Bl. glucose
Bl. pressure


Y
X


Cov ↑
$\rho_{XY} > 0$

Q



$$\frac{Cov}{\sigma_x \sigma_y} = \frac{0}{\sigma_y \sigma_x} = \text{Undefi}$$

D

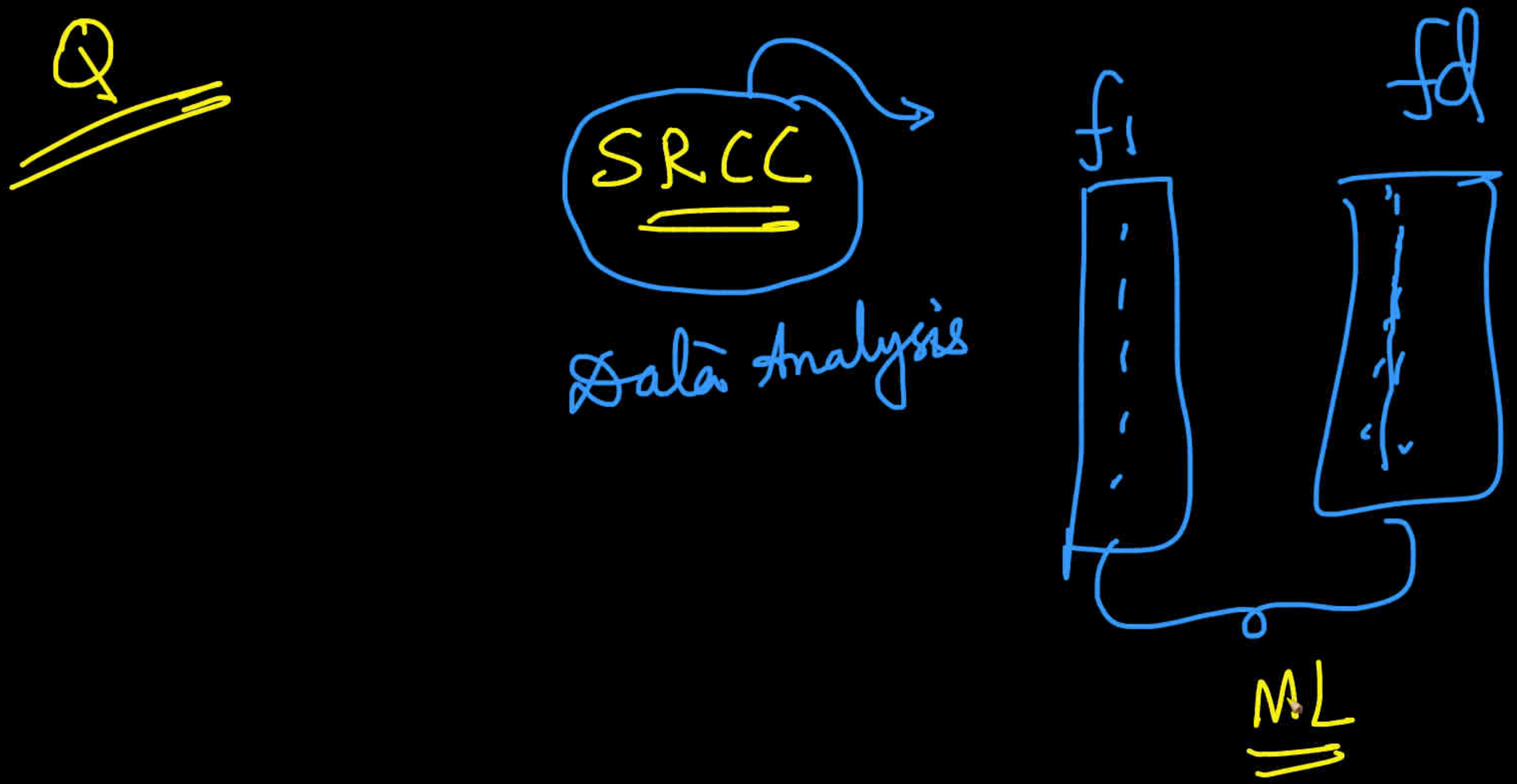
$r_s = +1$

$y = x^3$

$y = x^2 + x + 2$

$y = \sqrt{x}$

Linear-reg (ML)


SRCC
Data Analysis
f1
fd
ML

$$X \sim \text{distb}(\text{params})$$

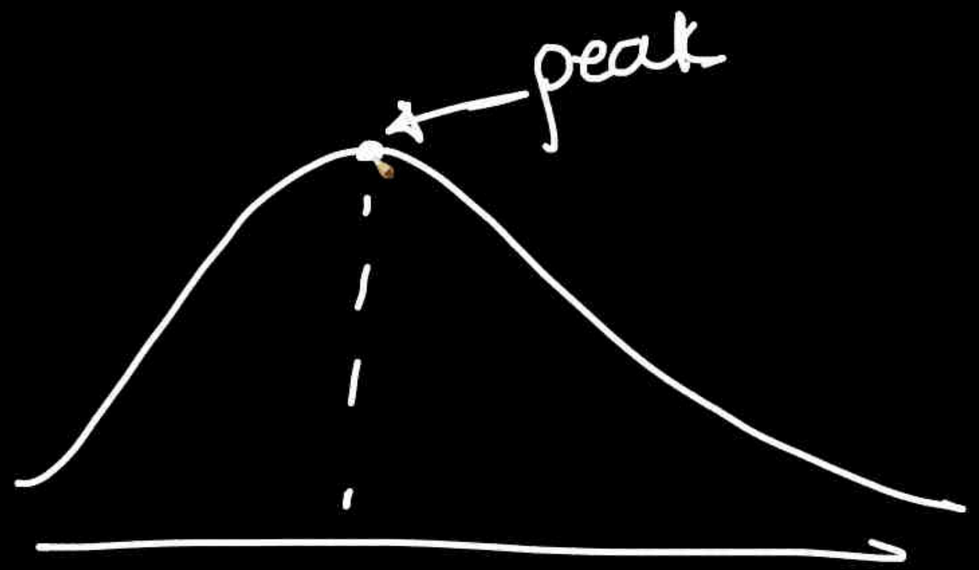
discrete

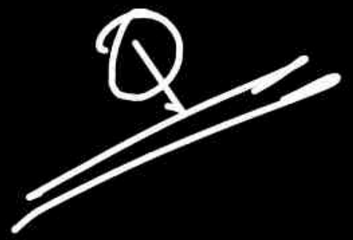
$$E(X) = \sum_{i=0}^{\infty} \underbrace{P(X=i)}_{\text{PMF}} \cdot i$$

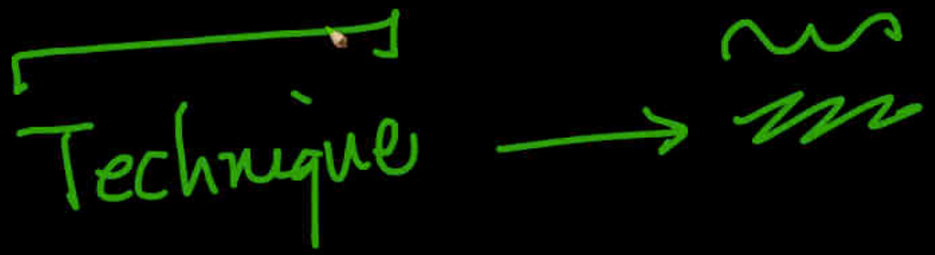
Cont:

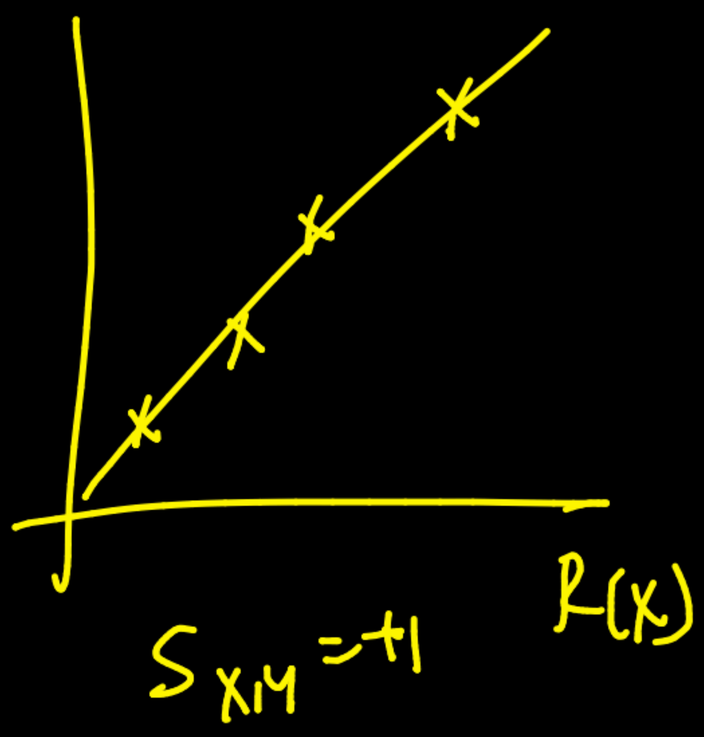
$$E(X) = \int_{x} \underbrace{f(x)}_{\text{PDF}} \cdot x \cdot dx$$

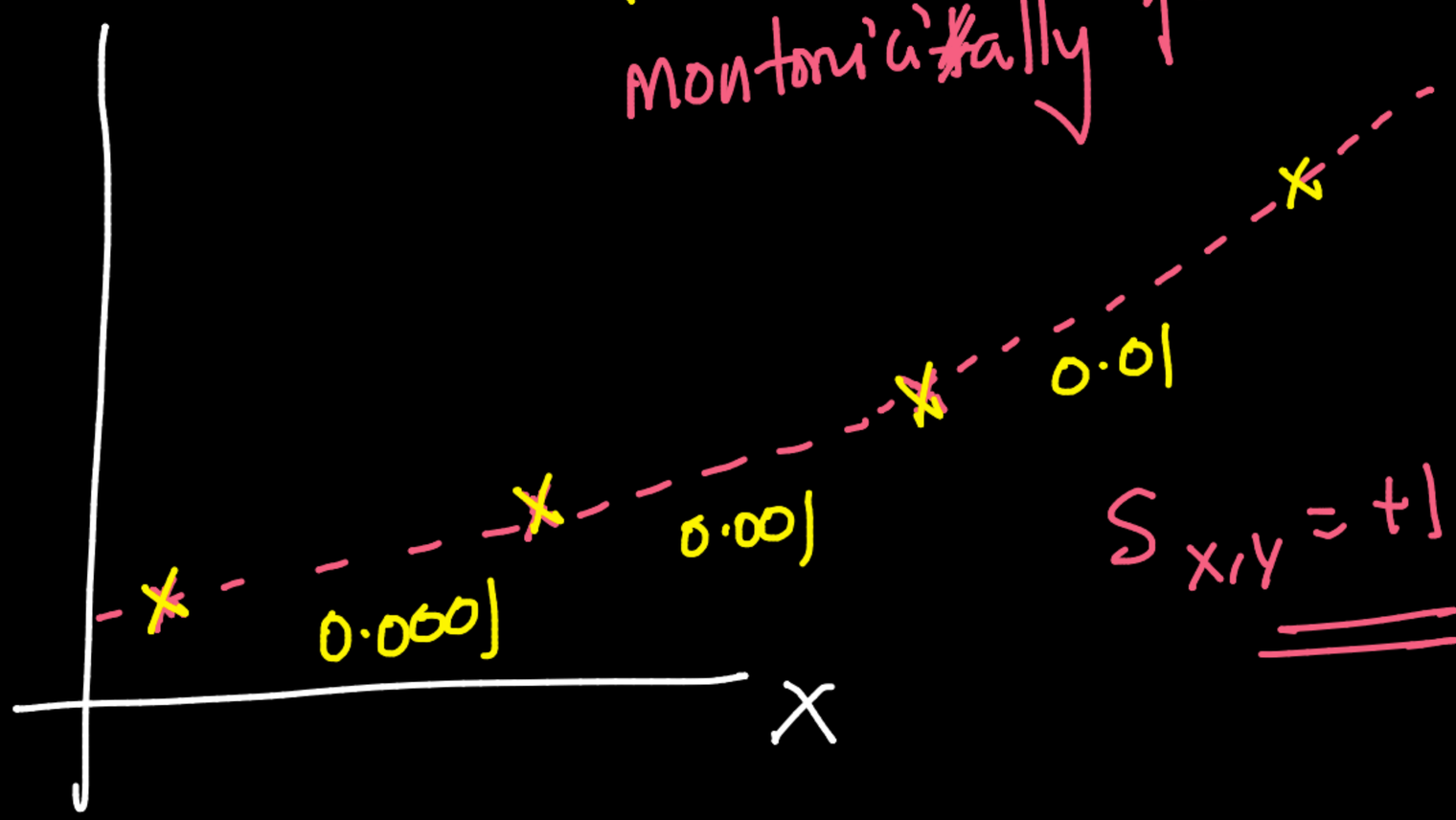
mode →






Technique





monotonically ↑

$S_{x,y} = +1$

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

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

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

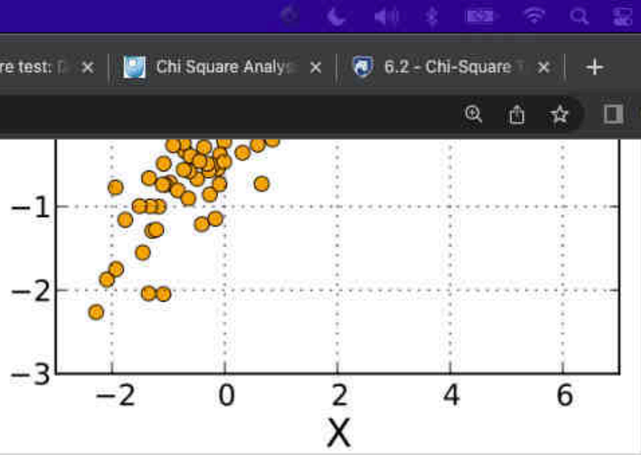
Only if all *n* ranks are *distinct integers*, it can be computed using the popular formula

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where

$d_i = \mathrm{R}(X_i) - \mathrm{R}(Y_i)$ is the difference between the two ranks of each observation,

n is the number of observations.



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.

[Proof] [show]

Identical values are usually[4] each assigned fractional ranks equal to the average of their positions in the ascending order of the values, which is equivalent to averaging over all possible permutations.

If ties are present in the data set, the simplified formula above yields incorrect results: Only if in both variables all ranks are distinct, then $\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)} = \operatorname{Var}(\mathrm{R}(X)) = \operatorname{Var}(\mathrm{R}(Y)) = (n^2 - 1)/12$ (calculated according to biased variance). The first equation — normalizing by the standard deviation — may be used even when ranks are normalized to [0, 1] ("relative ranks") because it is insensitive both to translation and linear scaling.

The simplified method should also not be used in cases where the data set is truncated; that is, when the Spearman's correlation coefficient