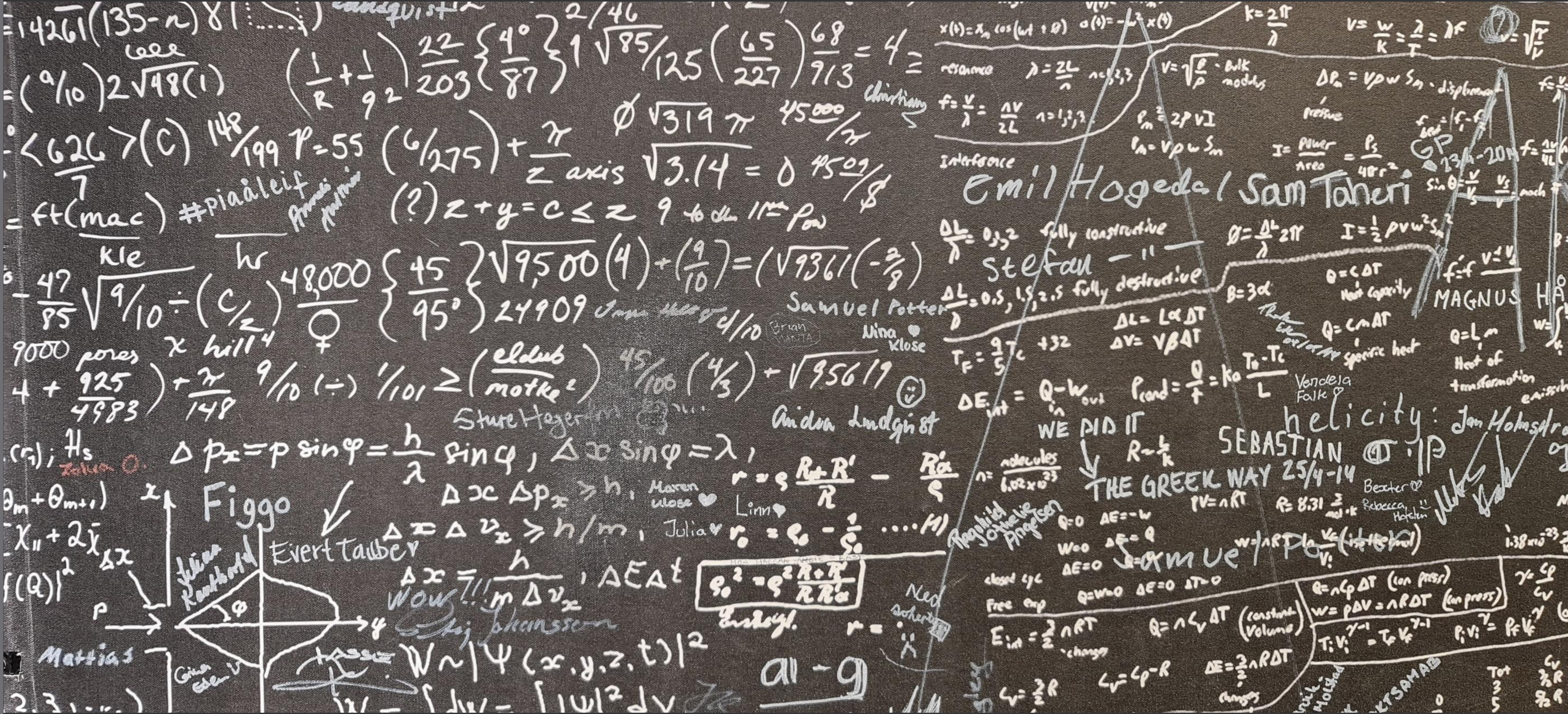



[Photo by Thomas T on Unsplash](https://unsplash.com) (retrieved 2023-06-20)

# A few notions from statistics

# Why statistics skills are vital

1. Measure

Icons from Flaticon (retrieved 2023-06-20)

# Why statistics skills are vital

Icons from Flaticon (retrieved 2023-06-20)

1. Measure

2. Describe

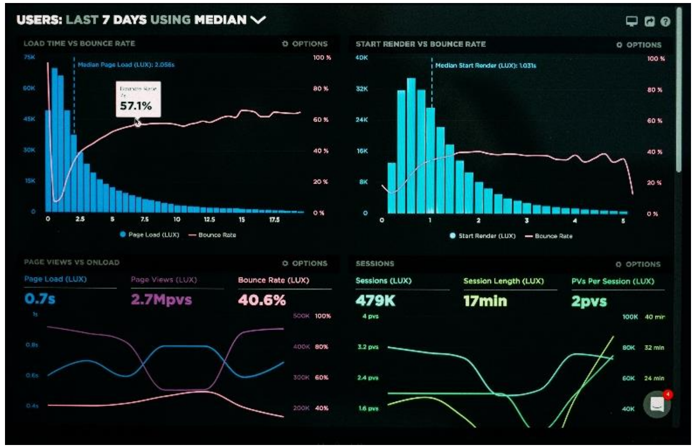



Photo by Luke Chesser on Unsplash (retrieved 2023-05-13)

# Why statistics skills are vital

1. Measure

Icons from Flaticon (retrieved 2023-06-20)

2. Describe

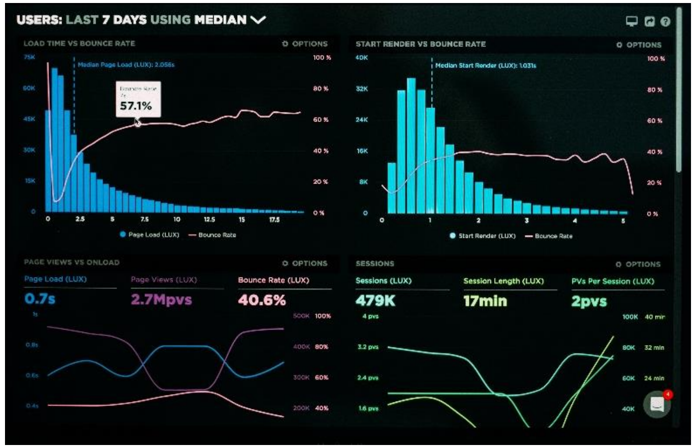


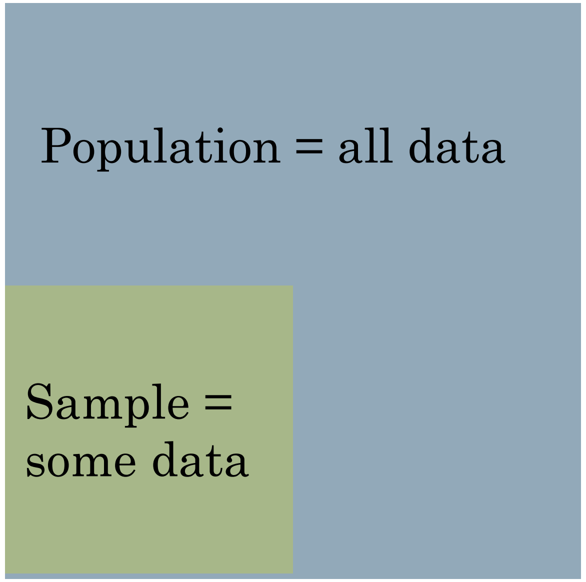

Photo by Luke Chesser on Unsplash (retrieved 2023-05-13)

3. Occam's razor

*If two models are equally good, use the simpler one.*

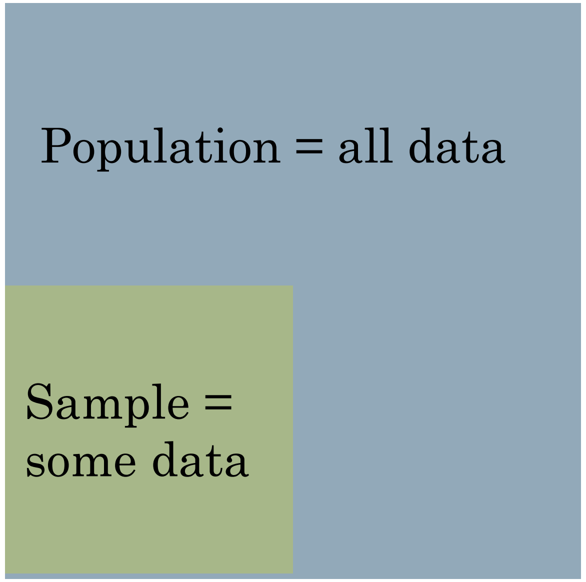
# Statistics - basics

Population vs sample



# Statistics - basics

Population vs sample

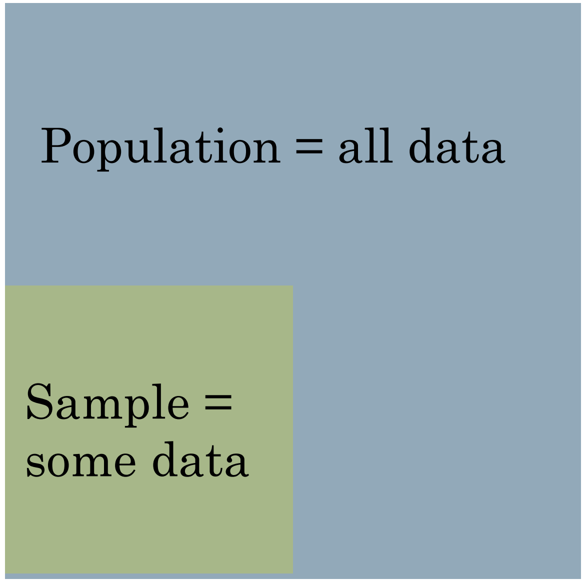



In Data Science, sampling can be a challenge:

- **iid** for ML → <u>***i**ndependent, **i**dentically **d**istributed*</u>
- Be representative
    - avoid biases
    - cover the issue

# Statistics - basics

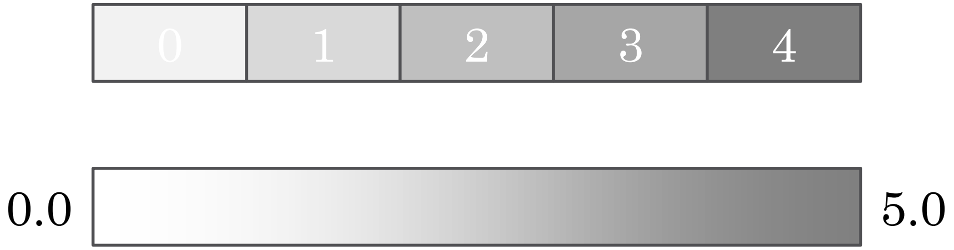
Population vs sample




In Data Science, sampling can be a challenge:

- **iid** for ML → ***i***<u>*ndependent,*</u> <u>***i****dentically* ***d****istributed*</u>
- Be representative
    - avoid biases
    - cover the issue

Discrete vs. continuous

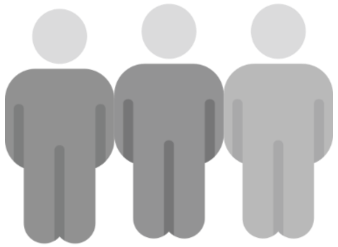


# Statistics - basics

(Probability) distributions

Icon by Flaticon (retrieved 2023-05-19)

# Statistics - basics



Icon by Flaticon (retrieved 2023-05-19)

(Probability) distributions

# Statistics - basics

- Mode
- Median
- Mean

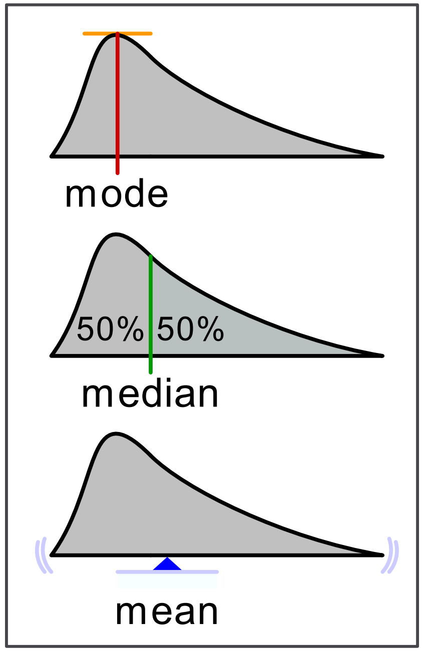
- Most frequent number
- The "middle" number
- $\mu = \frac{1}{N}\sum_i^N x_i$

# Statistics - basics

- Mode
- Median
- Mean

- Most frequent number
- The "middle" number
- $\mu = \frac{1}{N}\sum_{i}^{N} x_i$

Wikimedia (retrieved 2023-05-07)

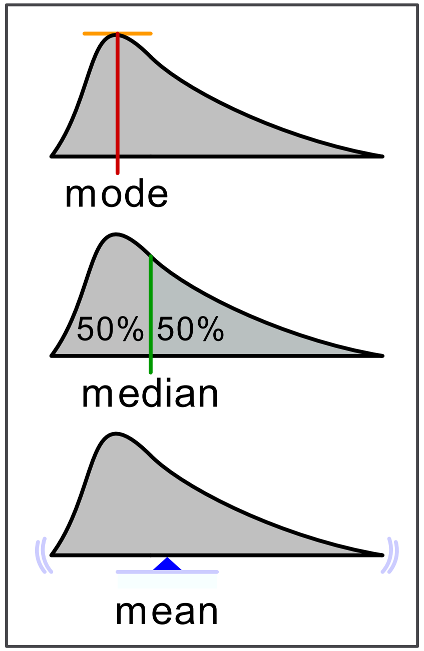


# Statistics - basics

- Mode
- Median
- Mean
- $k^{th}$ Quantile

- Most frequent number
- The “middle” number
- $\mu = \frac{1}{N}\sum_{i}^{N} x_i$
- The number covering $k^{th}$ part of the values

Wikimedia (retrieved 2023-05-07)

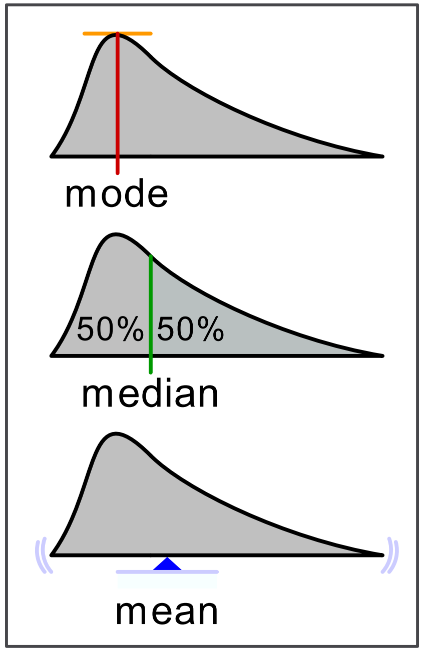



[1, 2, 4, 5, 7, 11, 17, 29, 29, 29, 30, 32, 65, 107, 125]

# Statistics - basics

- Mode
  - Most frequent number
- Median
  - The "middle" number
- Mean
  - $\mu = \frac{1}{N}\sum_i^N x_i$

Wikimedia (retrieved 2023-05-07)




- $k^{th}$ Quantile
  - The number covering $k^{th}$ part of the values

[1, 2, 4, 5, 7, 11, 17, 29, 29, 29, 30, 32, 65, 107, 125]

- Standard deviation
  - $\sigma = \sqrt{\frac{1}{N}\sum_i^N (x_i - \mu)^2}$

Variance = $\sigma^2$

- Z-score
  - $z_i = \frac{x_i - \mu}{\sigma}$ , standardized: $z_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

# Statistics - basics

**Hypothesis testing**

Null hypothesis $\boldsymbol{H_0}$: the ‘base case’ / ‘status quo’ - what if it’s just randomness?

Alternative hypothesis $\boldsymbol{H_1}$: the ‘test case’ - what if it’s a thing?

**p-value**: smallest probability for an observation if $H_0$ is true; typically 0.05

# Statistics - basics

**Hypothesis testing**

Null hypothesis $\boldsymbol{H_0}$: the 'base case' / 'status quo' - what if it's just randomness?

Alternative hypothesis $\boldsymbol{H_1}$: the 'test case' - what if it's a thing?

**p-value**: smallest probability for an observation if $H_0$ is true; typically 0.05

Question: Does drinking *Red Einstein* impact your $ spending?

# Statistics - basics

**Hypothesis testing**

Null hypothesis $\boldsymbol{H_0}$: the 'base case' / 'status quo' - what if it's just randomness?

Alternative hypothesis $\boldsymbol{H_1}$: the 'test case' - what if it's a thing?

**p-value**: smallest probability for an observation if $H_0$ is true; typically 0.05

Question: Does drinking *Red Einstein* impact your $ spending?

1. Define $H_0$ and $H_1$

- $H_0$: *Red Einstein* consumption $k$ does not change your spending results $r$
- Collect results $r_A$ from group $A$ with $k > 0$
- Collect results $r_B$ from group $B$ with $k = 0$
- $H_0$: $\mu_A = \mu_B$
- $H_1$: $\mu_A \neq \mu_B$

# Statistics - basics

**Hypothesis testing**

Null hypothesis $\boldsymbol{H_0}$: the 'base case' / 'status quo' - what if it's just randomness?

Alternative hypothesis $\boldsymbol{H_1}$: the 'test case' - what if it's a thing?

**p-value**: smallest probability for an observation if $H_0$ is true; typically 0.05

Question: Does drinking *Red Einstein* impact your $ spending?

1. Define $H_0$ and $H_1$
2. Assume that $H_0$ is true, calculate the z-score for your observation.

# Statistics - basics

**Hypothesis testing**

Null hypothesis $\boldsymbol{H_0}$: the 'base case' / 'status quo' - what if it's just randomness?

Alternative hypothesis $\boldsymbol{H_1}$: the 'test case' - what if it's a thing?

**p-value**: smallest probability for an observation if $H_0$ is true; typically 0.05

Question: Does drinking *Red Einstein* impact your $ spending?

1. Define $H_0$ and $H_1$
2. Assume that $H_0$ is true, calculate the z-score for your observation.
3. Look up the probability related to the z-score in the Z table (e.g. in ztable.net)

# Statistics - basics

**Hypothesis testing**

Null hypothesis $\boldsymbol{H_0}$: the ‘base case’ / ‘status quo’ - what if it’s just randomness?

Alternative hypothesis $\boldsymbol{H_1}$: the ‘test case’ - what if it’s a thing?

**p-value**: smallest probability for an observation if $H_0$ is true; typically 0.05

Question: Does drinking *Red Einstein* impact your $ spending?

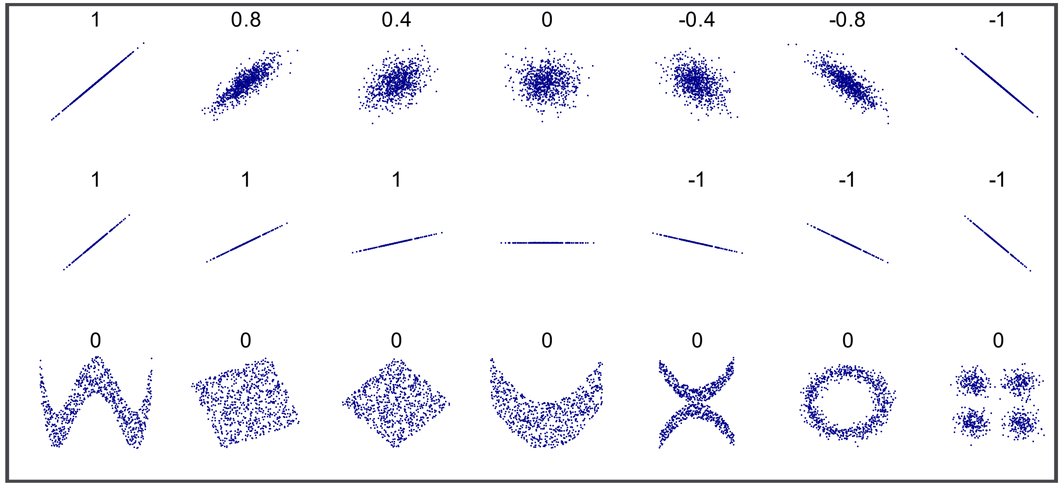
1. Define $H_0$ and $H_1$
2. Assume that $H_0$ is true, calculate the z-score for your observation.
3. Look up the probability related to the z-score in the Z table (e.g. in ztable.net)
4. If $p <$ p-value, reject $H_0$

# Statistics - basics

## Correlation coefficients

Pearson correlation

$$\rho_{x,y} = \frac{1}{N\ \sigma_x \sigma_y} \sum_{i}^{N} (x_i - \mu_x)^2 \left(y_i - \mu_y\right)^2 = \frac{cov(x,y)}{\sigma_x \sigma_y}$$



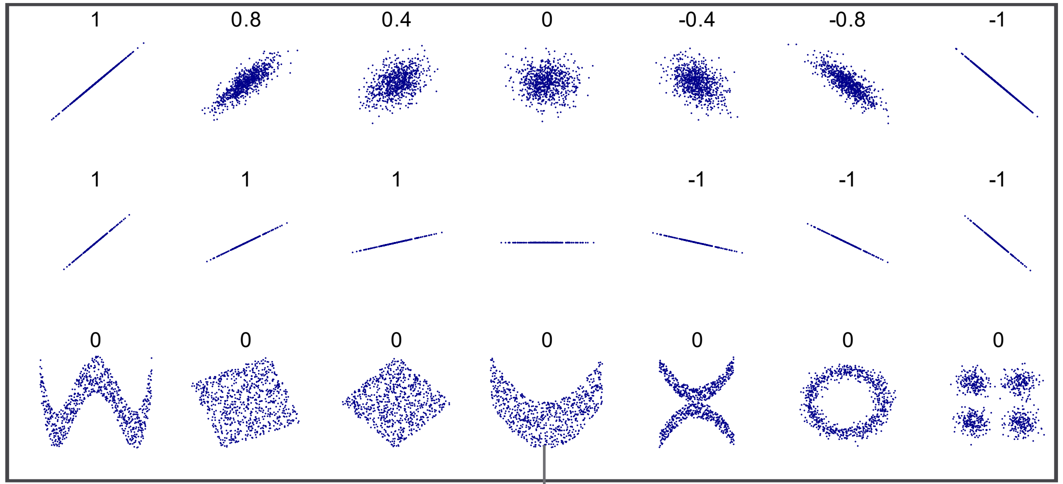

Wikimedia (retrieved 2023-05-07)

# Statistics - basics

**Correlation coefficients**

Pearson correlation

$$\rho_{x,y} = \frac{1}{N\,\sigma_x\sigma_y}\sum_{i}^{N}(x_i - \mu_x)^2\left(y_i - \mu_y\right)^2 = \frac{cov(x,y)}{\sigma_x\sigma_y}$$




Wikimedia (retrieved 2023-05-07)

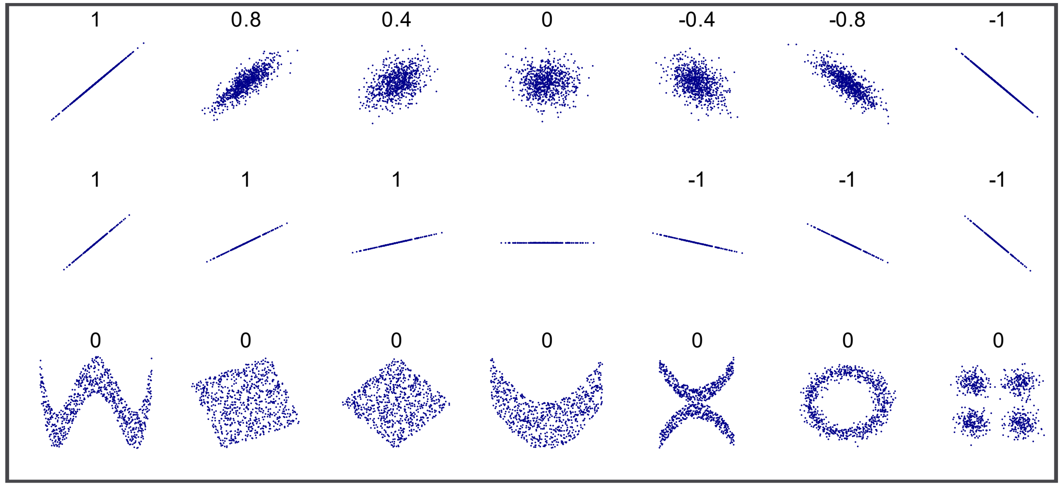
needs axis transformation

# Statistics - basics

## Correlation coefficients

Pearson correlation

$$\rho_{x,y} = \frac{1}{N\,\sigma_x \sigma_y} \sum_{i}^{N} (x_i - \mu_x)^2 \left(y_i - \mu_y\right)^2 = \frac{cov(x,y)}{\sigma_x \sigma_y}$$




[Wikimedia](Wikimedia) (retrieved 2023-05-07)

general problem with means

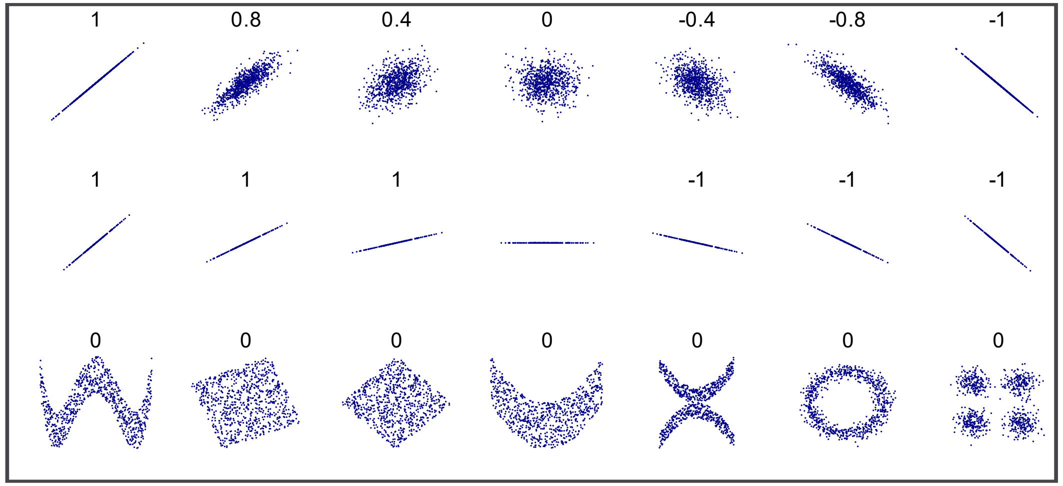
**Be careful with sample vs population!**

# Statistics - basics

## **Correlation coefficients**

Pearson correlation

$$\rho_{x,y} = \frac{1}{N\,\sigma_x \sigma_y} \sum_{i}^{N} (x_i - \mu_x)^2 \left(y_i - \mu_y\right)^2 = \frac{cov(x,y)}{\sigma_x \sigma_y}$$




Wikimedia (retrieved 2023-05-07)

$\rho_{x,y}$ is susceptible to outliers!

# Statistics - basics

**Correlation coefficients**

Pearson correlation

$$\rho_{x,y} = \frac{1}{N\,\sigma_x \sigma_y} \sum_{i}^{N} (x_i - \mu_x)^2 \left(y_i - \mu_y\right)^2 = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Spearman's rank

$$\rho_{R(x),R(y)} = \frac{cov(R(x),R(y))}{\sigma_{R(x)}\sigma_{R(y)}}$$

# Statistics - basics

**Correlation coefficients**

Pearson correlation

$$\rho_{x,y} = \frac{1}{N\,\sigma_x\sigma_y}\sum_{i}^{N}(x_i-\mu_x)^2\left(y_i-\mu_y\right)^2 = \frac{cov(x,y)}{\sigma_x\sigma_y}$$

Spearman's rank

$$\rho_{R(x),R(y)} = \frac{cov(R(x),R(y))}{\sigma_{R(x)}\sigma_{R(y)}}$$

$R(x_i)$ is the *rank* of $x_i$based on the value of $x_i$

For example:

$x$ = [1, 5, 2, 4, 7, 11, 17, 29, 29, 29, 32, 30, 65]

$R(x)$ = [1, 4, 2, 3, 5, 6, 7, 8, 8, 8, 10, 9, 11]

# Statistics - basics

**Correlation coefficients**

Pearson correlation

$$\rho_{x,y} = \frac{1}{N\,\sigma_x\sigma_y}\sum_i^N (x_i - \mu_x)^2 (y_i - \mu_y)^2 = \frac{cov(x,y)}{\sigma_x\sigma_y}$$

Spearman's rank

$$\rho_{R(x),R(y)} = \frac{cov(R(x),R(y))}{\sigma_{R(x)}\sigma_{R(y)}}$$

$R(x_i)$ is the *rank* of $x_i$ based on the value of $x_i$

For example:

$x = [1, 5, 2, 4, 7, 11, 17, 29, 29, 29, 32, 30, 65]$

$R(x) = [1, 4, 2, 3, 5, 6, 7, 8, 8, 8, 10, 9, 11]$

Ties? → Use Kendall $\tau$

# Statistics - basics

**Correlation coefficients**

Pearson correlation

$$\rho_{x,y} = \frac{1}{N\,\sigma_x\sigma_y}\sum_{i}^{N}(x_i-\mu_x)^2\left(y_i-\mu_y\right)^2 = \frac{cov(x,y)}{\sigma_x\sigma_y}$$

Spearman's rank

$$\rho_{R(x),R(y)} = \frac{cov(R(x),R(y))}{\sigma_{R(x)}\sigma_{R(y)}}$$

**Correlation is not causation!**

- Confounding variables
- Directionality problem
- "*Dumb luck*"