Recherche Zen
Séance 4 : Analyses

Carlos Ramisch and Manon Scholivet
Partly based on the course by Adeline Paiement

November 8, 2023

Plan

Expectation. . .

	dataset	metric1	metric2	metric3 [1]
SOTA system	DS1	82.3	75.9	48.0
Our system	DS1	**95.3**	**89.8**	**65.4**
SOTA system	DS2	67.7	65.2	56.8
Our system	DS2	**80.3**	**91.1**	**69.8**
SOTA system	DS3	77.6	74.1	92.8
Our system	DS3	**84.9**	**78.3**	**98.1**

1. Higher is better

Expectation. . .

	dataset	metric1	metric2	metric3 [1]
SOTA system	DS1	82.3	75.9	48.0
Our system	DS1	**95.3**	**89.8**	**65.4**
SOTA system	DS2	67.7	65.2	56.8
Our system	DS2	**80.3**	**91.1**	**69.8**
SOTA system	DS3	77.6	74.1	92.8
Our system	DS3	**84.9**	**78.3**	**98.1**

$\Longrightarrow$ Our system is better than state of the art ! 🎉

1. Higher is better

. . . Vs. reality !

	dataset	metric1	metric2	metric3
SOTA system	DS1	**82.3**	75.9	48.0
Our system	DS1	80.7	**76.2**	**50.4**
SOTA system	DS2	67.7	**65.2**	**56.8**
Our system	DS2	**67.9**	nan	49.6
SOTA system	DS3	77.6	**74.1**	92.8
Our system	DS3	**79.0**	**74.1**	**93.4**

. . . Vs. reality !

	dataset	metric1	metric2	metric3
SOTA system	DS1	**82.3**	75.9	48.0
Our system	DS1	80.7	**76.2**	**50.4**
SOTA system	DS2	67.7	**65.2**	**56.8**
Our system	DS2	**67.9**	nan	49.6
SOTA system	DS3	77.6	**74.1**	92.8
Our system	DS3	**79.0**	**74.1**	**93.4**

$\Longrightarrow$ Wake up and smell the coffee

Results analysis

- Identify overall trends
- Identify potential sources of problems (or bugs)
- Ensure conclusions are valid, claims are (statistically) sound

Experimental results

- Diversity of experiments $\implies$ diversity of results
  - $\rightarrow$ Task at hand
  - $\rightarrow$ Datasets
  - $\rightarrow$ Evaluation metrics
  - $\rightarrow$ ...
- This course : no silver bullet, rather a toolbox

Plan

Statistics

- A mathematical framework to analyse data
- Foundations : probability theory
- Statistical inference $\Longrightarrow$ data science, machine learning

  $\rightarrow$ Also : finances, health, biology, physics, social sciences, ...
- Identify trends, check hypotheses, measure correlations, ...

The problem with statistics

Finding good learning materials in statistics is hard

Too applied :

Too theoretical :

Weak Law of Large Numbers

The weak law of large numbers (cf. the strong law of large numbers) is a result in probability theory also known as Bernoulli's theorem. Let $X_1, ..., X_n$ be a sequence of independent and identically distributed random variables, each having a mean $\langle X_i \rangle = \mu$ and standard deviati σ. Define a new variable

$$X \equiv \frac{X_1 + \ldots + X_n}{n}.$$

Then, as $n \to \infty$, the sample mean $\langle x \rangle$ equals the population mean μ of each variable.

$$\begin{aligned}\langle X \rangle &= \left\langle \frac{X_1 + \ldots + X_n}{n} \right\rangle \\ &= \frac{1}{n}(\langle X_1 \rangle + \ldots + \langle X_n \rangle) \\ &= \frac{n\,\mu}{n} \\ &= \mu.\end{aligned}$$

In addition,

$$\begin{aligned}\mathrm{var}(X) &= \mathrm{var}\left(\frac{X_1 + \ldots + X_n}{n}\right) \\ &= \mathrm{var}\left(\frac{X_1}{n}\right) + \ldots + \mathrm{var}\left(\frac{X_n}{n}\right) \\ &= \frac{\sigma^2}{n^2} + \ldots + \frac{\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Therefore, by the Chebyshev inequality, for all $\epsilon > 0$,

$$P(|X - \mu| \geq \epsilon) \leq \frac{\mathrm{var}(X)}{\epsilon^2} = \frac{\sigma^2}{n\,\epsilon^2}.$$

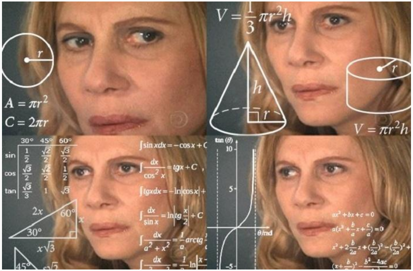
What usually happens

- A given statistical tool is used without (full) understanding
- Statistical tools applied because supervisor/reviewer asked
- Give up trying to understand, just use it as a blackbox

Truth be told : everyone hates statistics

Probability and statistics :

Difficult math, boring and totally useless, everyone hates it !



Truth be told : everyone hates statistics

Probability and statistics :

~~Difficult math, totally useless and so boring, everyone hates it !~~

- Difficult : mostly sums and products of fractions
- Boring : that's subjective, but yes, it may be boring
- Useless : definitely not ! The basis of empirical science

Truth be told : everyone hates statistics

Probability and statistics :

~~Difficult math, totally useless and so boring, everyone hates it !~~

- Yes, we may hate it, but we also need it !
  - → Knowing what we're doing can make us feel more at ease
  - → It is worth the effort of overcoming initial resistance

Probability and statistics :

A framework to model and reason in the presence of uncertainty

Truth be told : everyone hates statistics

Probability and statistics :

A framework to model and reason in the presence of uncertainty

We'll cover only what we absolutely need, promise.

Ready ? Let's go !

Wooclap

Wooclap time !

First things first

What is the difference between probability and statistics?

First things first

What is the difference between probability and statistics ?

Probability

- Mostly theoretical

  $\rightarrow$ Formal demonstrations

Statistics

- Manipulates data

  $\rightarrow$ Approximate probabilities

First things first

What is the difference between probability and statistics ?

Probability

- Mostly theoretical
  - → Formal demonstrations
- Notions we'll need :
  - → Random variable
  - → Probability distribution
  - → Normal distribution

Statistics

- Manipulates data
  - → Approximate probabilities
- Notions we'll need :
  - → Sampling, mean, variance
  - → Covariance, correlation
  - → Hypotheses testing

Random variable

- A random variable is a variable with no specific value
  - $\rightarrow$ It takes some value within a (known) set of possible values
  - $\rightarrow$ We are not interested in its actual value

Examples :

- A human's age takes values form 0 to 130 years
- The sea water temperature ranges from 0°C to 100°C
- A person's handedness can be righ-handed, left-handed, both

Random variable

Are the following (interesting) random variables ?

- 1. The number of tentacles of an octopus ?

Random variable

Are the following (interesting) random variables ?

- 1. The number of tentacles of an octopus ?

  $\longrightarrow$ No, always the same value

Random variable

Are the following (interesting) random variables?

- 2. An adult human's height in centimeters?

Random variable

Are the following (interesting) random variables?

- 2. An adult human's height in centimeters?

  → Yes, e.g. values from 50cm to 300cm

Random variable

Are the following (interesting) random variables ?

- 3. The distance between the Earth and the Moon ?

Random variable

Are the following (interesting) random variables ?

- 3. The distance between the Earth and the Moon ?

  $\rightarrow$ Yes, it actually varies from 363K to 406K km

Random variable

Are the following (interesting) random variables ?

- 4. A person's vote in the last presidential elections ?

Random variable

Are the following (interesting) random variables?

- 4. A person's vote in the last presidential elections?

  $\longrightarrow$ Yes, the values are the candidates/parties running

Random variable

Are the following (interesting) random variables ?

- 5. A person's opinion about how cute an octopus is ?

Random variable

Are the following (interesting) random variables ?

- 5. A person's opinion about how cute an octopus is ?

  $\rightarrow$ No, ill-defined, no closed set of possible values

  $\rightarrow$ Actually, everyone finds them cute ! ;-)

Random variable

Are the following (interesting) random variables ?

- 1. The number of tentacles of an octopus ? No
- 2. An adult human's height in centimeters ? Yes
- 3. The distance between the Earth and the Moon ? Yes
- 4. A person's vote in the last presidential elections ? Yes
- 5. A person's opinion about how cute an octopus is ? No

In short

- A variable is not random if its value is fixed / constant
- Random variables can have non-numerical values
- We need to be able to describe its set of possible values

  $\rightarrow$ The set may be infinite (e.g. real numbers)

Why do we need random variables ?

- Use their characteristics to understand the data
- Model features and evaluation metrics as random variables
- Basic block in probability and statistics
  - → People have been studying them for a while
  - → Statistical tools associated to them can be useful

Probability distributions

- Random variables are not interesting per se
- They come with probability distributions

Probability distribution

Given a random variable X :

- Each of its possible values $x_i \rightarrow$ number $p(x_i)$ between 0 and 1
  - $\rightarrow$ This number is called the probability of x_i
  - $\rightarrow$ $p(x_i)$ indicates how likely that value is
- The sum of $p(x_i)$ for all x_i values must be equal to 1
- The set of all $p(x_i)$ values form X's **probability distribution**

Expressing probabilities

$$P\{X = a\} = p(a) = 0.8$$

- X : The random variable that we're interested in
- a : The particular value of that random variable
- 0.8 : The probability that variable X takes value a

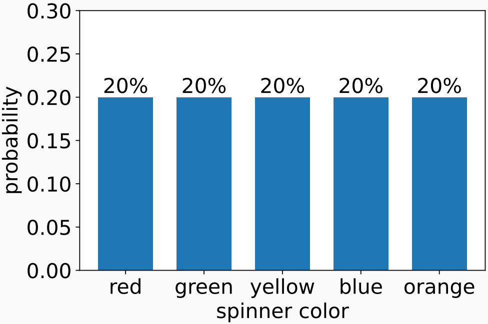
Expressing probabilities

$$P\{X = a\} = p(a) = 0.8$$

- X : The random variable that we're interested in
- a : The particular value of that random variable
- 0.8 : The probability that variable X takes value a
- Note : we shorten $P\{X = a\}$ as $p(a)$ if there is no ambiguity
- Note : the probability value 0.8 is often written 80%

Simple probability distributions

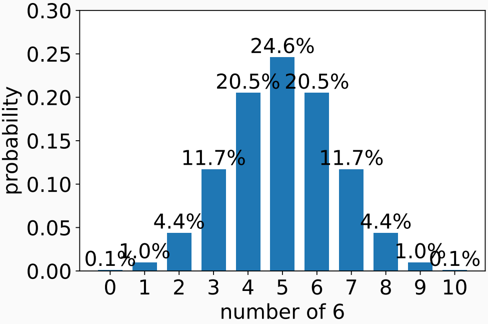
- X_1 : color of a 5-coloured spinner wheel



$$P\{X_1 = \text{red}\} = p(\text{green}) = \ldots = p(\text{orange}) = \frac{1}{5}$$

Simple probability distributions

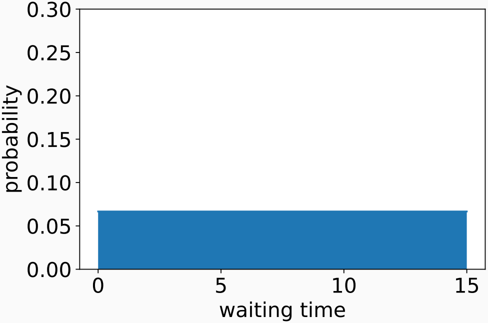
- X_2 : number of "face" when throwing a fair coin 10 times



$$p(1) = p(10) = \frac{1}{2}^{1} \times \frac{1}{2}^{9} = 0.001$$

Simple probability distributions

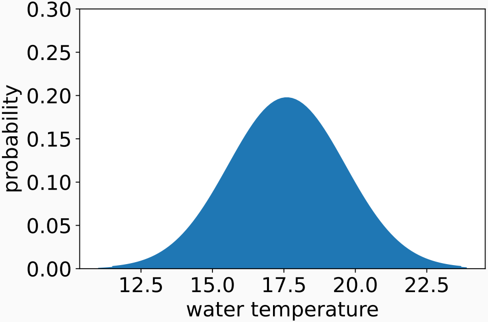
- X_3 : waiting time for a bus passing every 15min



$$P\{0 \leq X_3 < 5\} = \frac{5-0}{15} = 0.33$$

Simple probability distributions

- X_4 : sea water temperature in July in Marseille



$$P\{X_4 < 17.6\} = 0.5$$

Wooclap

Wooclap time !

Probability distribution or not ?

Which of the following are proper probability distributions ? Why ?

a)

x_i	$p(x_i)$
1	0.4
2	-0.2
3	0.8

b)

x_i	$p(x_i)$
0.4	0.4
0.35	0.35
0.25	0.25

c)

x_i	$p(x_i)$
-1	0.4
-2	0.2
-3	0.8

d)

x_i	$p(x_i)$
-1	0.4
0	0.2
1	0.2
2	0.1

Probability distribution or not ?

Which of the following are proper probability distributions ? Why ?

a)

x_i	$p(x_i)$
1	0.4
2	-0.2
3	0.8

No, $p(2) < 0$

b)

x_i	$p(x_i)$
0.4	0.4
0.35	0.35
0.25	0.25

Yes, sum=1

c)

x_i	$p(x_i)$
-1	0.4
-2	0.2
-3	0.8

No, sum > 1

d)

x_i	$p(x_i)$
-1	0.4
0	0.2
1	0.2
2	0.1

No, sum < 1

From probabilities to statistics

- Probability distributions are theoretical abstractions
  - $\rightarrow$ We often learn probabilities with toy examples
  - $\rightarrow$ In practice, X's "real" distribution is not accessible
- A sample is often used to estimate the probabilities
  - $\rightarrow$ Most of the time, probabilities are approximated
  - $\rightarrow$ Proportion in sample (%) $\rightarrow$ estimated probability

$$\frac{\text{count}(a)}{n} \approx P\{X = a\}$$

Random samples

- Randomly select a finite set of data points to study
  - → A set of sentences to translate
  - → A set of GPS positions to track
  - → A set of people to perform a task
  - → ...

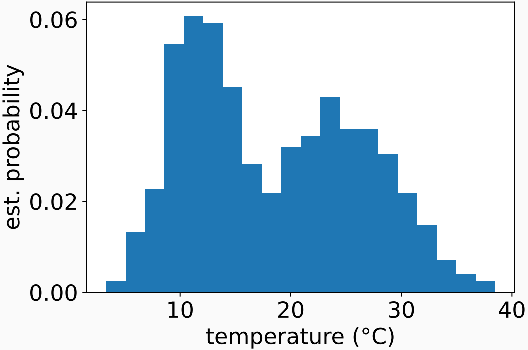
Source: https://www.thoughtco.com/purposive-sampling-3026727

Sampling : example

Daily temperature of a captor in a power plant

→ Sample size : 365 days

→ `[10.1, 14.0, 8.9, 6.7, 9.4, 10.3, ... 12.5, 15.3, 13.3]`



Estimated probability distribution = normalized histogram

Sampling : example

Jupyter notebook 1 & 2

1. Open the dataset using `pandas.read_csv()`
2. Explore the different columns and their values
3. Make a histogram of the `compositionality` column

   → This is an estimate of its distribution !

Compositionality dataset

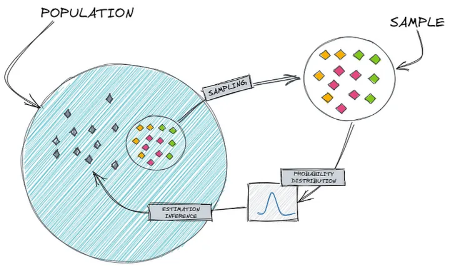
- *Is a dry run literally a run which is dry ?*

  → not at all ←0 - 1 - 2 - 3 - 4 - 5 → absolutely yes
- Compositionality score : average rating of 10-15 annotators
- Sample : 180 compounds in French

Source: https://aclanthology.org/J19-1001/

Why do we need samples ?

- A representative sample can inform us about the whole
  - → Full data not available, but sample findings can be generalised
  - → Infer properties of the (unknown) distribution
  - → Draw conclusions in the presence of uncertainty



Source: https://towardsdatascience.com/understanding-random-variables-and-probability-distributions-1ed1daf2e66

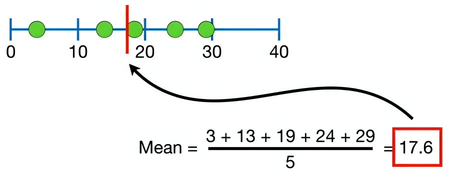
Descriptive statistics

- We can characterise our sample
  - $\longrightarrow$ Central tendency : mean
  - $\longrightarrow$ Dispersion : variance

Mean / average

- A single value at the center of the sample
  - $\rightarrow$ Summarise the whole data in a single number
- The arithmetic mean of a set of i.i.d. values $x_1 \dots x_n$:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$



Source: StatQuest : `https://www.youtube.com/watch?v=SzZ6GpcfoQY`

Wooclap

Wooclap time !

Mean / average quiz

- Is the mean a probability (value between 0 and 1) ?

Mean / average quiz

- Is the mean a probability (value between 0 and 1) ?

  $\rightarrow$ No, it depends on the values (arbitrary range)
- Is the value of the mean contained in the sample ?

Mean / average quiz

- Is the mean a probability (value between 0 and 1) ?

  $\rightarrow$ No, it depends on the values (arbitrary range)

- Is the value of the mean contained in the sample ?

  $\rightarrow$ No, it can be a new value, not contained in the sample

- Is the value of the mean always positive ?

Mean / average quiz

- Is the mean a probability (value between 0 and 1) ?

  $\rightarrow$ No, it depends on the values (arbitrary range)

- Is the value of the mean contained in the sample ?

  $\rightarrow$ No, it can be a new value, not contained in the sample

- Is the value of the mean always positive ?

  $\rightarrow$ No, e.g. if the variable only takes negative values

The larger the better

- The expected value of a (discrete) random variable :

$$E[X] = p(x_1)x_1 + p(x_2)x_2 + \ldots + p(x_n)x_n$$

- Sample mean $\overline{x}$ $\rightarrow$ normalised sum of n i.i.d. random variables

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- The law of large numbers states that $\overline{x} \rightarrow E[X]$ for large n

  $\rightarrow$ The (sample) mean $\overline{x}$ is an estimator of the expected value $E[X]$

The larger the better

- The expected value of a (discrete) random variable :

$$E[X] = p(x_1)x_1 + p(x_2)x_2 + \ldots + p(x_n)x_n$$

- Sample mean $\overline{x} \rightarrow$ normalised sum of n i.i.d. random variables

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

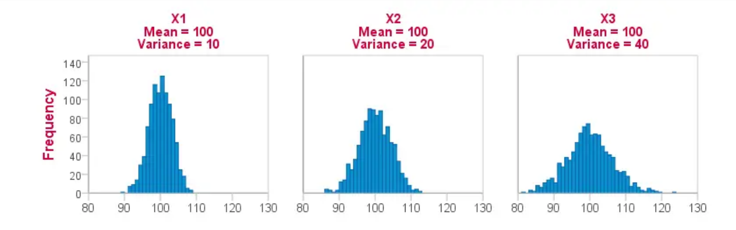
- The law of large numbers states that $\overline{x} \rightarrow E[X]$ for large n
  - $\rightarrow$ The (sample) mean $\overline{x}$ is an estimator of the expected value $E[X]$

The larger the sample, the better $\overline{x}$ approximates "true" mean $E[X]$

Data dispersion

- Mean does not take into account data dispersion

$$S_1 = [0] \implies \overline{S_1} = 0$$
$$S_2 = [-4,\ -4,\ 4,\ 4] \implies \overline{S_2} = 0$$
$$S_3 = [-6,\ -2,\ 1,\ 7] \implies \overline{S_3} = 0$$
$$S_4 = [-1500,\ 1500] \implies \overline{S_4} = 0$$



https://www.spss-tutorials.com/descriptive-statistics-one-metric-variable/

Getting to the variance

Idea 1 : average the difference between each value and the mean

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

- Calculate this amount for the sample `[-4, -4, 4, 4]`

Getting to the variance

Idea 1 : average the difference between each value and the mean

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

- Calculate this amount for the sample `[-4, -4, 4, 4]`

$$\frac{(-4-0)+(-4-0)+(4-0)+(4-0)}{4} = 0 \quad ☹$$

Getting to the variance

Idea 2 : average the absolute value of the $x_i - \overline{x}$ difference

$$\sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$$

- Calculate this amount for the sample `[-4, -4, 4, 4]`

Getting to the variance

Idea 2 : average the absolute value of the $x_i - \overline{x}$ difference

$$\sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$$

- Calculate this amount for the sample `[-4, -4, 4, 4]`

$$\frac{|-4-0| + |-4-0| + |4-0| + |4-0|}{4} = 4 \quad ☺$$

Getting to the variance

Idea 2 : average the absolute value of the $x_i - \overline{x}$ difference

$$\sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$$

- Calculate this amount for the sample `[-6, -2, 1, 7]`

Getting to the variance

Idea 2 : average the absolute value of the $x_i - \overline{x}$ difference

$$\sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$$

- Calculate this amount for the sample `[-6, -2, 1, 7]`

$$\frac{|-6-0| + |-2-0| + |1-0| + |7-0|}{4} = 4 \quad ☹$$

Moreover, absolute value is not differentiable at 0

This is inconvenient : `https://www.youtube.com/watch?v=sHRBg6BhKjI`

Getting to the variance

Idea 3 : average the squared differences $x_i - \overline{x}$

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

- Calculate this amount for the sample `[-4, -4, 4, 4]`

Getting to the variance

Idea 3 : average the squared differences $x_i - \overline{x}$

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

- Calculate this amount for the sample `[-4, -4, 4, 4]`

$$\frac{(-4-0)^2 + (-4-0)^2 + (4-0)^2 + (4-0)^2}{4} = 64 \quad ☺$$

Getting to the variance

Idea 3 : average the squared differences $x_i - \overline{x}$

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

- Calculate this amount for the sample `[-6, -2, 1, 7]`

Getting to the variance

Idea 3 : average the squared differences $x_i - \overline{x}$

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

- Calculate this amount for the sample `[-6, -2, 1, 7]`

$$\frac{(-6-0)^2 + (-2-0)^2 + (1-0)^2 + (7-0)^2}{4} = 90 \quad ☺$$

Source: Example adapted from

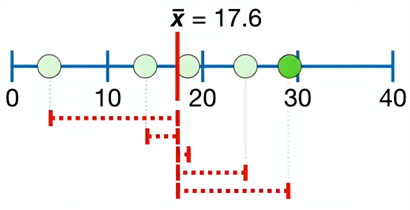
`https://www.mathsisfun.com/data/standard-deviation.html`

Variance

- Variance characterises the dispersion/spread of a distribution
  - → Intuition : average distance from the mean
  - → $(x_i - \overline{x})$ can be positive or negative $\implies$ square it !

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

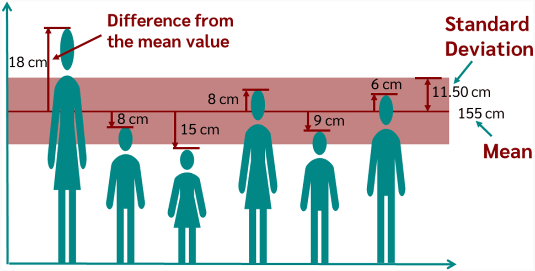
  - → Variance is always positive, differently from mean



Standard deviation

- Variance averages *squared* differences
  - $\rightarrow$ Its absolute value is hard to interpret
  - $\rightarrow$ Bring back to original value range $\rightarrow$ squared root
- The squared root of variance is called standard deviation

$$\sigma = \sqrt{Var(X)}$$



https://datatab.net/tutorial/dispersion-parameter

Estimated standard deviation

- Population standard deviation :

$$\sigma_X = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$$

- Sample standard deviation, unbiased estimator :

$$s_X = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$$

- Why ? `https://www.youtube.com/watch?v=sHRBg6BhKjI`

Estimated standard deviation

- Population standard deviation :

$$\sigma_X = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$$

- Sample standard deviation, unbiased estimator :

$$s_X = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$$

- Why ? https://www.youtube.com/watch?v=sHRBg6BhKjI

In practice, we only need s_X $\rightarrow$ Ensure your stats library does this !

Calculating mean and standard deviation

Jupyter notebook 3

1. Open dataset containing 180 compositionality scores
2. Use Pandas' `comp.describe()` to obtain a summary
3. Is the obtained standard deviation σ_X or s_X ?

One distribution to rule them all

The Normal distribution

$$P\{a < X < b\} = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} exp^{\frac{x-\mu}{\sigma}}$$

One distribution to rule them all

The **Normal** distribution

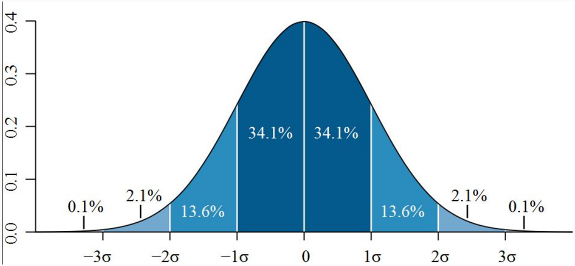
$$P\{a < X < b\} = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} exp^{\frac{x-\mu}{\sigma}}$$

Who cares !

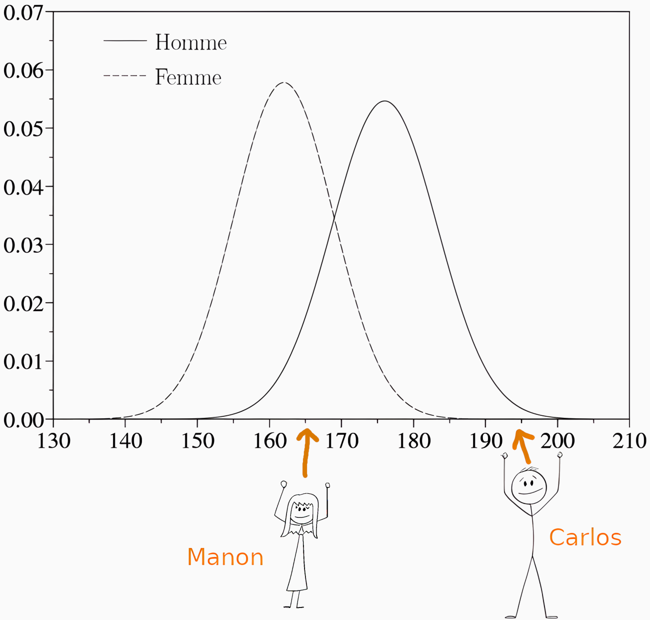
One distribution to rule them all

The Normal distribution

- Well known distribution for continuous random variables
- Probability density function is a symmetric bell-shaped curve
- Characterised by mean μ and standard deviation σ
    - $\rightarrow$ Bell centered around μ, narrower or wider according to σ
    - $\rightarrow$ 99% of probability between $\mu - 3\sigma$ and $\mu + 3\sigma$



Normal distribution : example

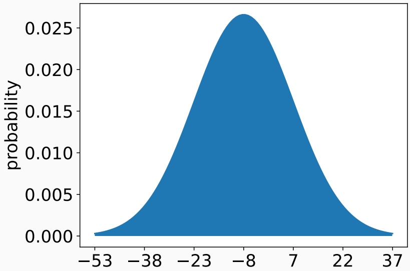


Wooclap

Wooclap time !

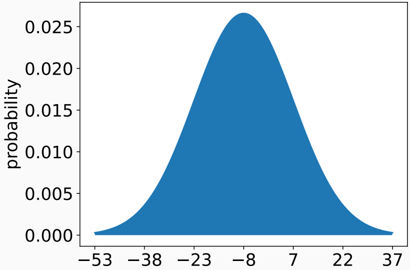
Who's that normal ?

1. What are the μ and σ parameters for the following curve ?



Who's that normal ?

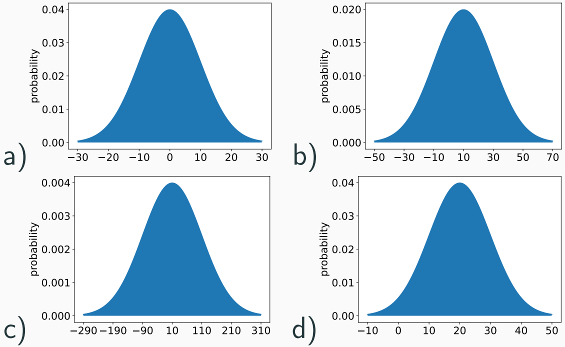
1. What are the μ and σ parameters for the following curve ?



$\mu = -8$ and $\sigma = 15$

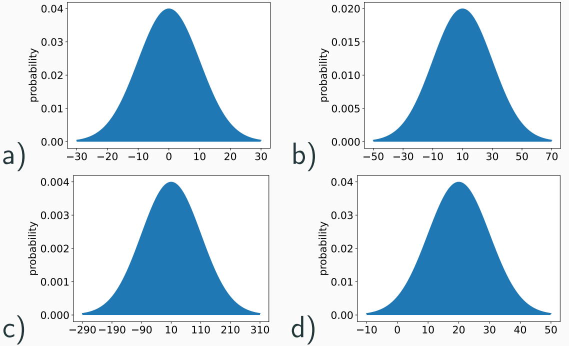
Who's that normal ?

1. What are the μ and σ parameters for the following curve ?
2. Which curve corresponds to $\mu = 10$ and $\sigma = 20$?



Who's that normal ?

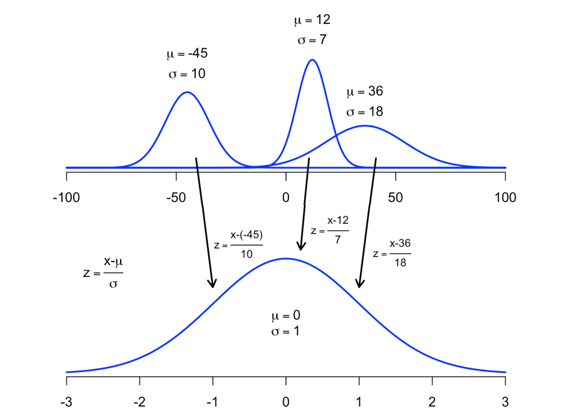
1. What are the μ and σ parameters for the following curve ?
2. Which curve corresponds to $\mu = 10$ and $\sigma = 20$?



curve b) – notice different heights

Standardization

- Calculate probability $\rightarrow$ integration (<o> aaaaah !)
  - $\rightarrow$ Normal is impossible to integrate analytically
- In practice :
  - $\rightarrow$ Standardize $z = \frac{x-\mu}{\sigma}$, then lookup table of $\Phi(a)$



Wooclap

Wooclap time !

The most famous probability distribution

Why is the normal distribution so important ?

The most famous probability distribution

Why is the normal distribution so important?

- Turns out most measurements are normally distributed
- Used in many statistical tools, e.g. hypothesis testing
- Plays a central role in describing estimated means

It's normal to be average

- Normalised sum of i.i.d. variables is normally distributed
  - → Even if the variables are not normally distributed !
- The mean $\overline{x}$ of a sample is normally distributed
  - → Comes in handy to analyse averaged values
- This is known as the central limit theorem
  - → Connects statistics and probability

Central limit theorem : example

Jupyter notebook 4 & 5

1. Build n random samples of size 30 from compositionality data
2. Calculate mean of each random sample, save values
3. Estimate sample mean's distribution with histogram

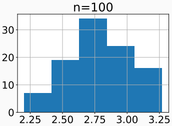
   $\longrightarrow$ What happens when n increases ?

Central limit theorem : example

Jupyter notebook 4 & 5

1. Build n random samples of size 30 from compositionality data
2. Calculate mean of each random sample, save values
3. Estimate sample mean's distribution with histogram

$\longrightarrow$ What happens when n increases ?

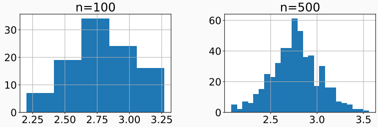


Central limit theorem : example

Jupyter notebook 4 & 5

1. Build *n* random samples of size 30 from compositionality data
2. Calculate mean of each random sample, save values
3. Estimate sample mean's distribution with histogram

   $\longrightarrow$ What happens when *n* increases ?

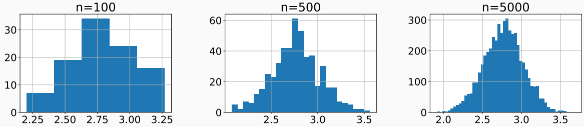


Central limit theorem : example

Jupyter notebook 4 & 5

1. Build n random samples of size 30 from compositionality data
2. Calculate mean of each random sample, save values
3. Estimate sample mean's distribution with histogram

$\longrightarrow$ What happens when n increases ?

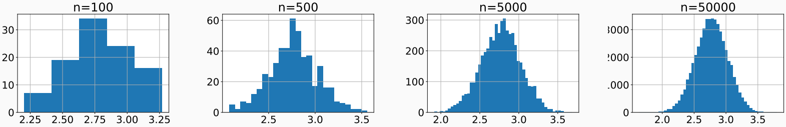


Central limit theorem : example

Jupyter notebook 4 & 5

1. Build *n* random samples of size 30 from compositionality data
2. Calculate mean of each random sample, save values
3. Estimate sample mean's distribution with histogram

   $\longrightarrow$ What happens when *n* increases ?



In short

- Random variables and probability distributions
  - → Theoretical model for features and metrics
  - → In practice, estimated using sampling
- Mean and standard deviation characterise the data
  - → Ensure your stats library divides by $n - 1$
- Normal distribution : bell shaped around the mean
  - → Useful to characterise values that are means

In short

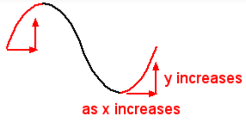
- Random variables and probability distributions
  - → Theoretical model for features and metrics
  - → In practice, estimated using sampling
- Mean and standard deviation characterise the data
  - → Ensure your stats library divides by $n-1$
- Normal distribution : bell shaped around the mean
  - → Useful to characterise values that are means

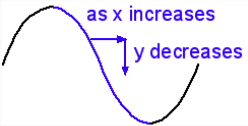
Now we're ready for the next steps !

Plan

Two random variables

- For the moment we looked at random variables one by one
- It may be interesting to look at two random variables X and Y
  - $\rightarrow$ They may influence each other
  - $\rightarrow$ They may be both influenced by similar factors
- How does X and Y vary together ?





Two variables : scatter plot

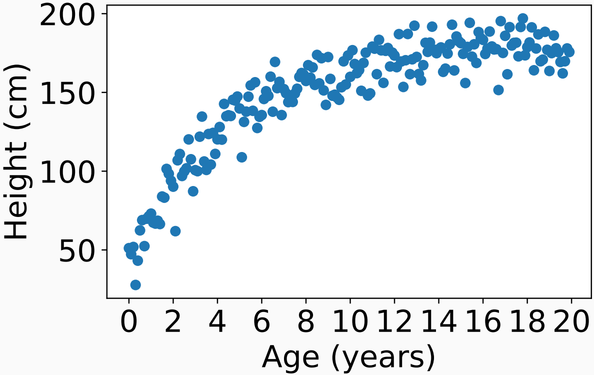
- Variable X on x-axis, variable Y on y-axis
- `plt.scatter(x,y)`
- The two variables are paired or aligned
  - $\rightarrow$ The sample consists of pairs of values
  - $\rightarrow$ Each value of X has a corresponding value of Y
  - $\rightarrow$ Both variables are numeric

Scatter plot example 1

A person's age (X) vs. height (Y)

Scatter plot example 1

A person's age (X) vs. height (Y)

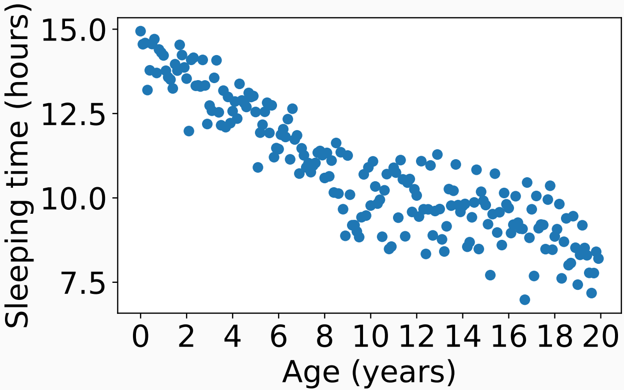


Scatter plot example 2

A person's age (X) vs. number of sleeping hours (Y)

Scatter plot example 2

A person's age (X) vs. number of sleeping hours (Y)

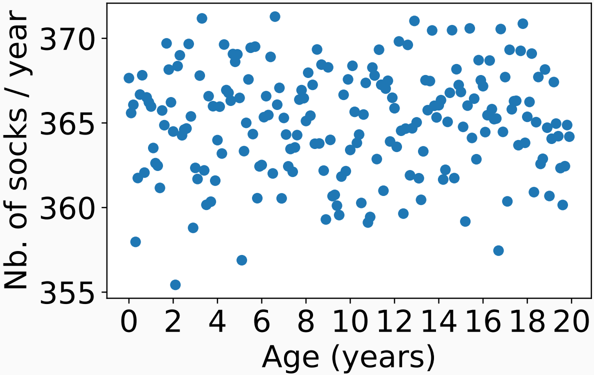


Scatter plot example 3

A person's age (X) vs. number of socks used per year (Y)

Scatter plot example 3

A person's age (X) vs. number of socks used per year (Y)



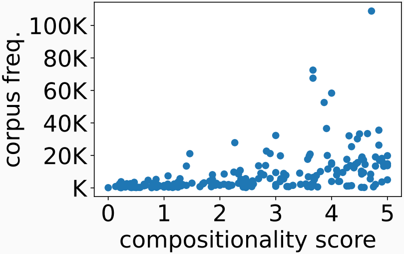
Example : compositionality and number of occurrences

Jupyter notebook 6 & 7

- Hypothesis : frequent compounds are less compositional
- What is the relation between compositionality and frequency ?

Jupyter notebook 6 & 7

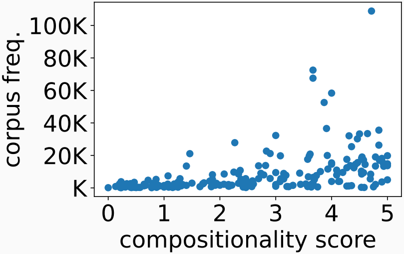
- Hypothesis : frequent compounds are less compositional
- What is the relation between compositionality and frequency ?



Example : compositionality and number of occurrences

Jupyter notebook 6 & 7

- Hypothesis : frequent compounds are less compositional
- What is the relation between compositionality and frequency ?



- Is there really something to see or are we over-interpreting ?

Quantifying relations

- It would be nice to be able to quantify the relation !

Quantifying relations

- It would be nice to be able to quantify the relation !

We will obtain such metric in two steps :

1. Covariance
   - $\rightarrow$ Not so easy to interpret
   - $\rightarrow$ Computational step towards calculating correlation
2. Correlation
   - $\rightarrow$ Much easier to interpret

Covariance : far from the mean

- Relation between each value x_i and the mean $\bar{x}$
- Relation between each value y_i and the mean $\bar{y}$

Covariance : far from the mean

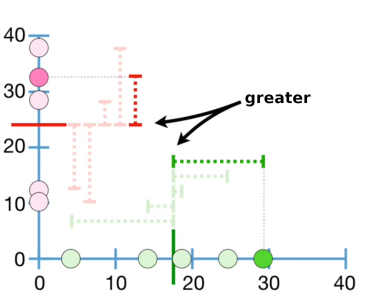
- Relation between each value x_i and the mean $\bar{x}$
- Relation between each value y_i and the mean $\bar{y}$

  $\rightarrow$ Does $x_i > \bar{x}$ imply $y_i > \bar{y}$?

  $\rightarrow$ Does $x_i < \bar{x}$ imply $y_i < \bar{y}$?

Covariance : far from the mean

- Relation between each value x_i and the mean $\bar{x}$
- Relation between each value y_i and the mean $\bar{y}$
  - $\rightarrow$ Does $x_i > \bar{x}$ imply $y_i > \bar{y}$?
  - $\rightarrow$ Does $x_i < \bar{x}$ imply $y_i < \bar{y}$?

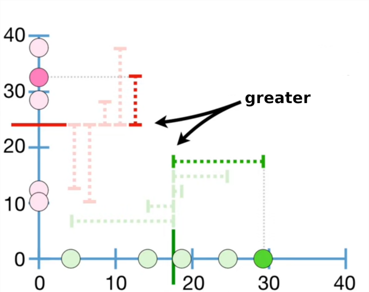


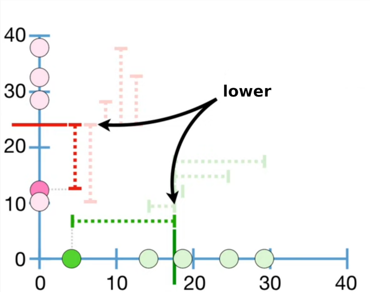
Covariance : far from the mean

- Relation between each value x_i and the mean $\bar{x}$
- Relation between each value y_i and the mean $\bar{y}$

  → Does $x_i > \bar{x}$ imply $y_i > \bar{y}$?

  → Does $x_i < \bar{x}$ imply $y_i < \bar{y}$?





Source: `https://www.youtube.com/watch?v=qtaqvPAeEJY`

Covariance : vary together

- Relation between each value x_i and the mean $\bar{x}$

  $\rightarrow x_i > \bar{x} \implies (x_i - \bar{x})$ positive

  $\rightarrow x_i < \bar{x} \implies (x_i - \bar{x})$ negative

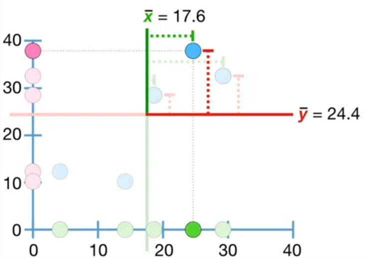
- Relation between each value y_i and the mean $\bar{y}$

  $\rightarrow y_i > \bar{y} \implies (y_i - \bar{y})$ positive

  $\rightarrow y_i < \bar{y} \implies (y_i - \bar{y})$ negative

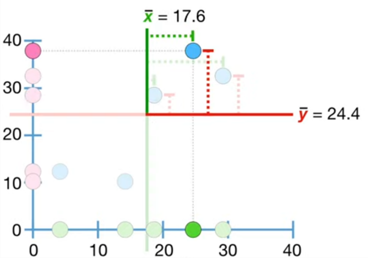
Covariance : vary together

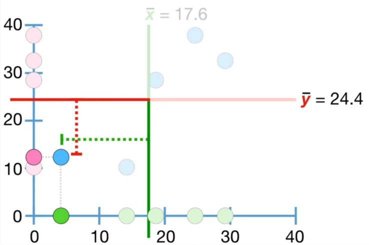
- Relation between each value x_i and the mean $\bar{x}$
  - $\rightarrow x_i > \bar{x} \implies (x_i - \bar{x})$ positive
  - $\rightarrow x_i < \bar{x} \implies (x_i - \bar{x})$ negative
- Relation between each value y_i and the mean $\bar{y}$
  - $\rightarrow y_i > \bar{y} \implies (y_i - \bar{y})$ positive
  - $\rightarrow y_i < \bar{y} \implies (y_i - \bar{y})$ negative



Covariance : vary together

- Relation between each value x_i and the mean $\bar{x}$
  - $\rightarrow x_i > \bar{x} \implies (x_i - \bar{x})$ positive
  - $\rightarrow x_i < \bar{x} \implies (x_i - \bar{x})$ negative
- Relation between each value y_i and the mean $\bar{y}$
  - $\rightarrow y_i > \bar{y} \implies (y_i - \bar{y})$ positive
  - $\rightarrow y_i < \bar{y} \implies (y_i - \bar{y})$ negative





Source: `https://www.youtube.com/watch?v=qtaqvPAeEJY`

Covariance : vary together

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are positive
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**

Covariance : vary together

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are positive
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are negative
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**

Covariance : vary together

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are positive
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are negative
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- $(x_i - \bar{x})$ is positive and $(y_i - \bar{y})$ is negative
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **negative**

Covariance : vary together

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are positive
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are negative
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- $(x_i - \bar{x})$ is positive and $(y_i - \bar{y})$ is negative
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **negative**
- $(x_i - \bar{x})$ is negative and $(y_i - \bar{y})$ is positive
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **negative**

Covariance : the formula

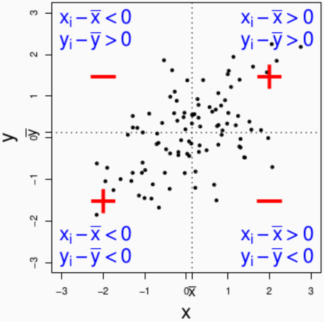
1. First calculate means $\bar{x}$ and $\bar{y}$
2. Then calculate the covariance as :

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

Covariance : the formula

1. First calculate means $\bar{x}$ and $\bar{y}$
2. Then calculate the covariance as :

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



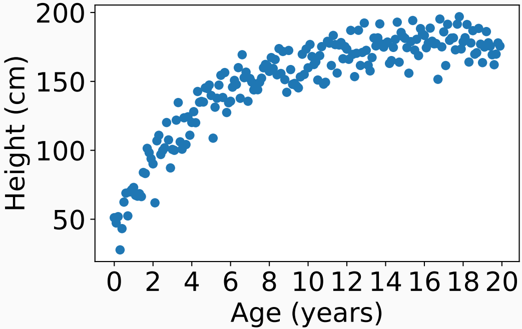
https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Wooclap

Wooclap time !

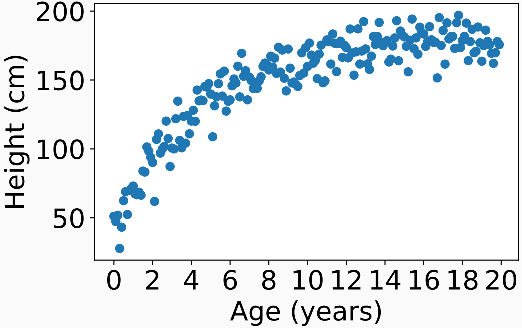
Exercise : guess the covariance

1. A person's age (X) vs. height (Y)



Exercise : guess the covariance

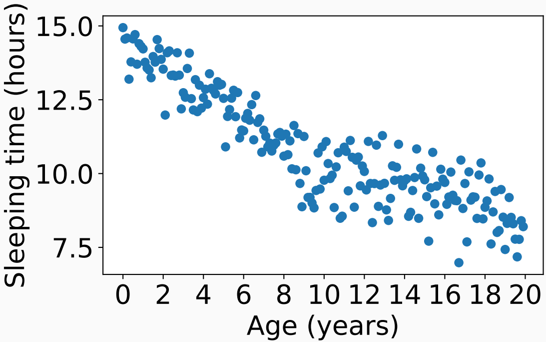
1. A person's age (X) vs. height (Y)



$$Cov(X, Y) = +180.9$$

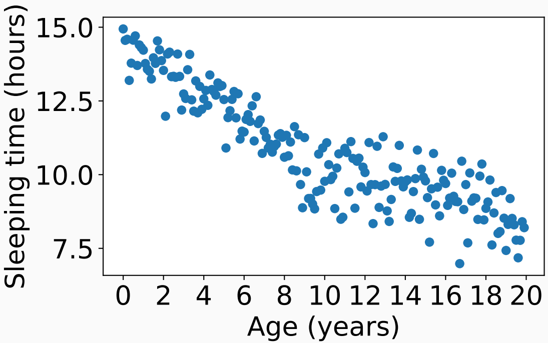
Exercise : guess the covariance

A person's age (X) vs. number of sleeping hours (Y)



Exercise : guess the covariance

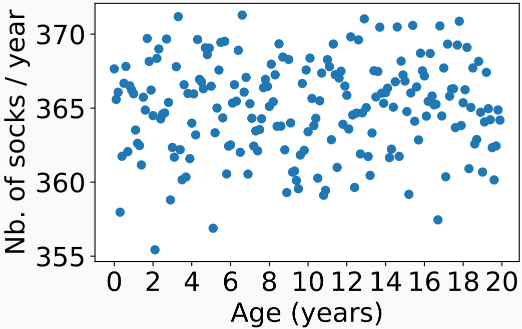
A person's age (X) vs. number of sleeping hours (Y)



$$Cov(X, Y) = -9.0$$

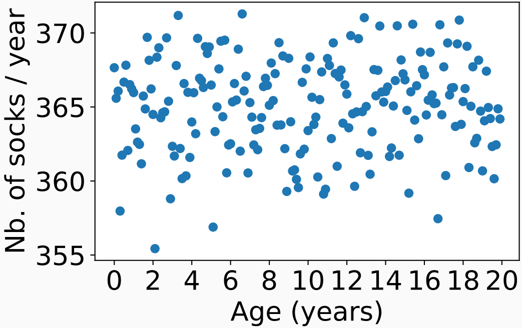
Exercise : guess the covariance

A person's age (X) vs. number of socks used per year (Y)



Exercise : guess the covariance

A person's age (X) vs. number of socks used per year (Y)



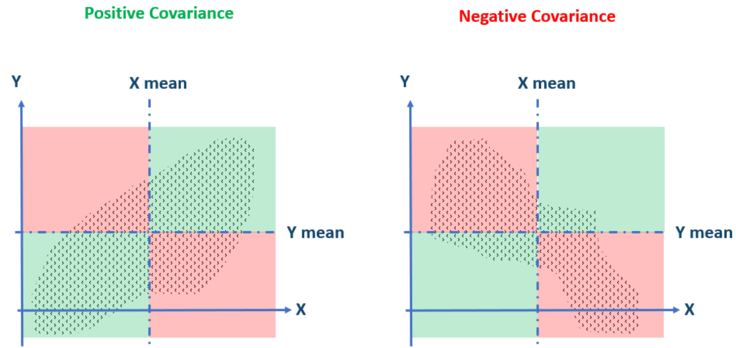
$$Cov(X, Y) = 0.77$$

Covariance is sensitive to unit

- What if X and Y have very different ranges ?
  - $\rightarrow$ For instance, X in cm, Y in km

Covariance is sensitive to unit

- What if X and Y have very different ranges ?
  $\rightarrow$ For instance, X in cm, Y in km
- Covariance is unbounded - ranges from $-\infty$ to $+\infty$
  $\rightarrow$ Indicates whether a linear relation exists, but not its strength



Covariance : it's a sign !

- Covariance is **positive**

  → Increasing X tends to make Y increase too
- Covariance is **negative**

  → Increasing X tends to make Y decrease
- Covariance is **zero**

  → Increasing X has no impact on Y

  → Increasing Y has no impact on X

What if...

- What if we could normalise covariance ?
- Can we get a measure that is bounded ?

Correlation coefficient (r)

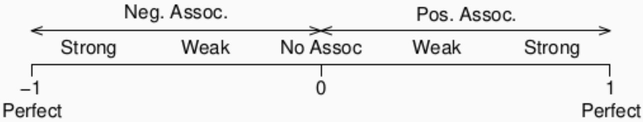
- Covariance can be normalised using X and Y's **variances**

$$r_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{s_X s_Y}$$

- Dividing by standard deviation puts both on same scale
- Also called Pearson or linear correlation

Correlation interpretation

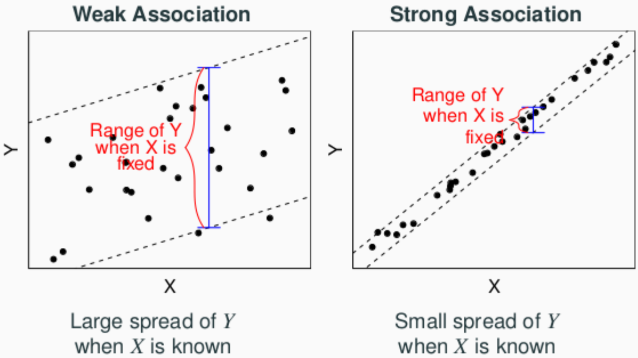
- Ranges from -1 to $+1$
  - $\rightarrow$ $r \approx +1$: strong positive association
  - $\rightarrow$ $r \approx -1$: strong negative association
  - $\rightarrow$ $r \approx 0$: weak/no linear relationship



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation and spread

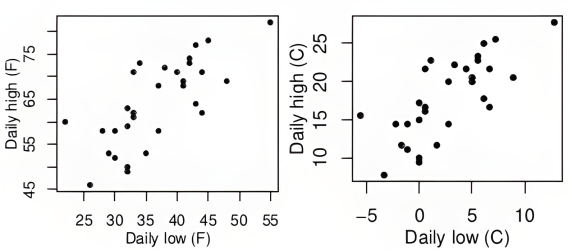
- Correlation tells how close or far from linear regression line
  - → Knowing x allows predicting y (and vice-versa)



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation is unit-less

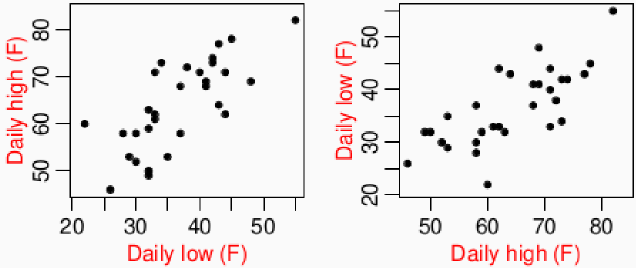
- Covariance is unbounded, depends on variable ranges
- Correlation allows comparing metrics with different ranges
  - → Example : max vs. min. temperature in Celsius or Farehnheit
  - → In both cases, correlation is the same : $r = 0.74$



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation is symmetric

- Correlation is symmetric
  - → Example : max vs. min. temperature or vice-versa
  - → In both cases, correlation is the same : $r = 0.74$



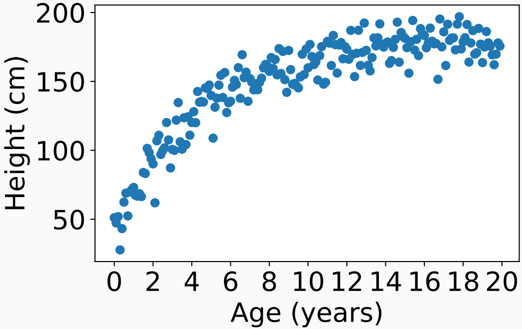
https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Wooclap

Wooclap time !

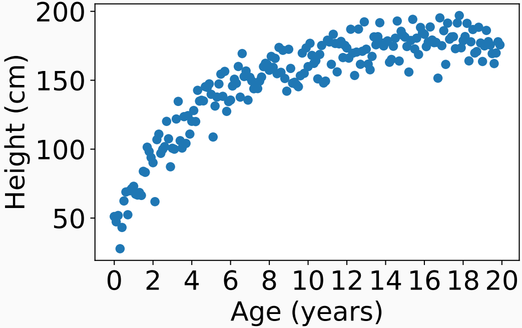
Exercise : guess the correlation

1. A person's age (X) vs. height (Y)



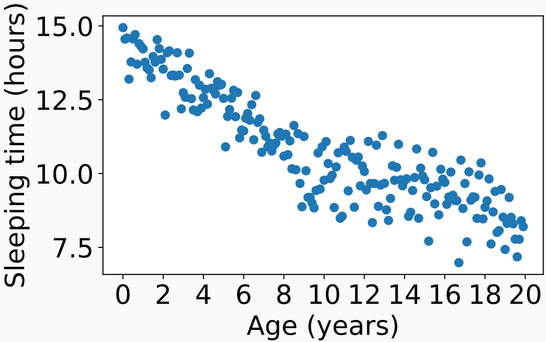
Exercise : guess the correlation

1. A person's age (X) vs. height (Y)



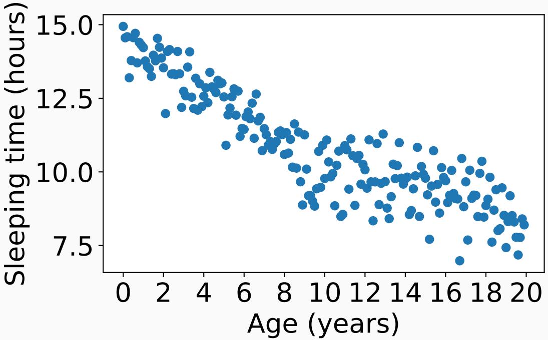
$$r(X, Y) = 0.85$$

A person's age (X) vs. number of sleeping hours (Y)



Exercise : guess the correlation

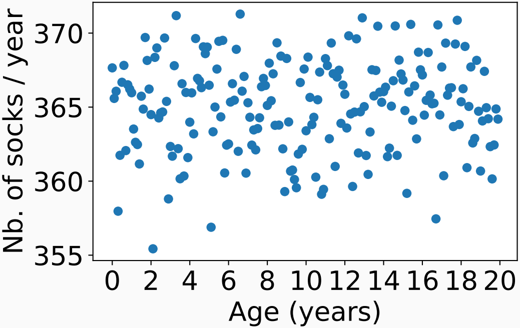
A person's age (X) vs. number of sleeping hours (Y)



$$r(X, Y) = -0.89$$

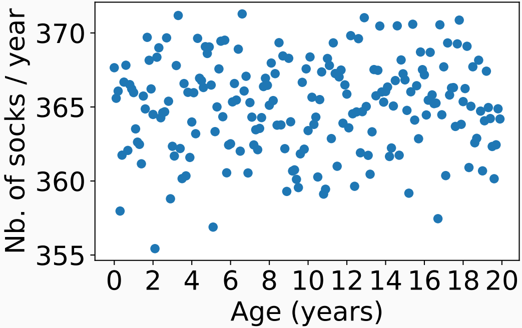
Exercise : guess the correlation

A person's age (X) vs. number of socks used per year (Y)



Exercise : guess the correlation

A person's age (X) vs. number of socks used per year (Y)



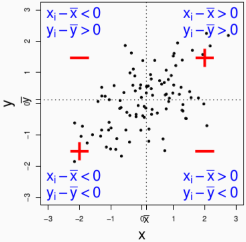
$$r(X, Y) = 0.04$$

Why dividing by standard deviations ?

$$r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X s_Y}$$

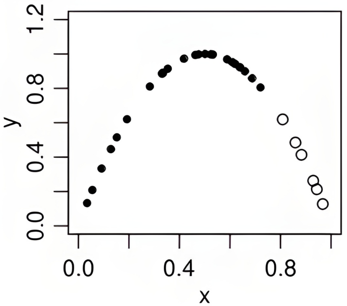
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_X}\right)\left(\frac{y_i - \overline{y}}{s_Y}\right)$$

- Similar to standardisation in normal distribution
  - $\longrightarrow$ Discounting the mean centers around zero
  - $\longrightarrow$ Dividing by standard deviation homogenizes width



Correlation shows linear association

- Correlation does not model non-linear association



r of all black dots = 0.803,
r of all dots = −0.019.
(black + white)

https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation of compositionality

Jupyter notebook 8

- Hypothesis : compositionality and frequency are correlated
  - $\rightarrow$ Frequency is better represented in logarithmic scale
- Does correlation change if frequency is in linear or log scale ?

Spearman's rank correlation

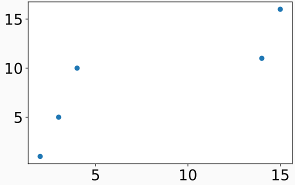
- The actual compared X and Y values may be irrelevant
  - $\rightarrow$ Does X rank items more or less in the same order as Y?
- Spearman's ρ : linear (Pearson) correlation between ranks
  - $\rightarrow$ Models monotonic relation

Spearman's rank correlation

- The actual compared X and Y values may be irrelevant
  $\rightarrow$ Does X rank items more or less in the same order as Y?
- Spearman's ρ : linear (Pearson) correlation between ranks
  $\rightarrow$ Models monotonic relation

Example :

```
x = [2,3,4,14,15]
y = [1,5,10,11,16]
```



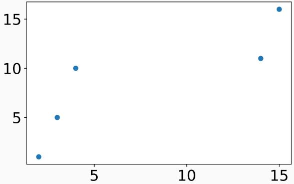
$\Longrightarrow$

Spearman's rank correlation

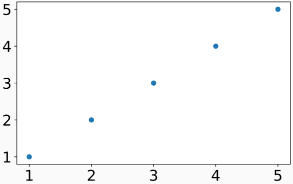
- The actual compared X and Y values may be irrelevant
  $\rightarrow$ Does X rank items more or less in the same order as Y?
- Spearman's ρ : linear (Pearson) correlation between ranks
  $\rightarrow$ Models monotonic relation

Example :

```
x = [2,3,4,14,15]
y = [1,5,10,11,16]
```



$\Longrightarrow$



Spearman correlation

- Obtain ranks rX_i for X in ascending order
- Obtain ranks rY_i for Y in ascending order
- Obtain difference between ranks $d_i = rX_i - rY_i$
- Calculate Spearman's rank correlation :

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

Spearman correlation

- Obtain ranks rX_i for X in ascending order
- Obtain ranks rY_i for Y in ascending order
- Obtain difference between ranks $d_i = rX_i - rY_i$
- Calculate Spearman's rank correlation :

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

- Alternatively, Pearson correlation between rX_i and rY_i

Spearman correlation : example

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

Source: https://en.wikipedia.org/wiki/Spearman_correlation

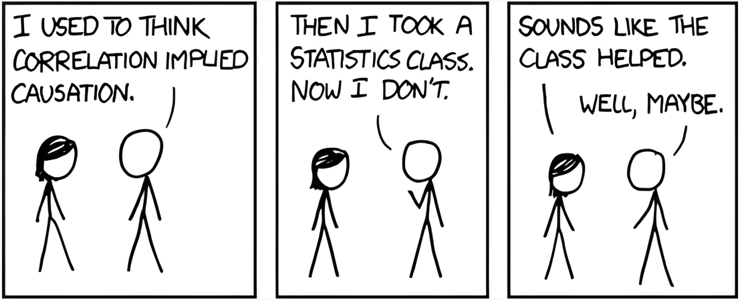
Pereson vs. Spearman of compositionality

Jupyter notebook 9 & 10

- Compare Pearson and Spearman correlation
  - $\rightarrow$ Compositionality vs. frequency
  - $\rightarrow$ Compositionality vs. log-frequency
- Compare manual implementation and scipy

Confounders

- Suppose X independent and Y dependent variables
- A confounder can influence both X and Y
- Correlation is not causation

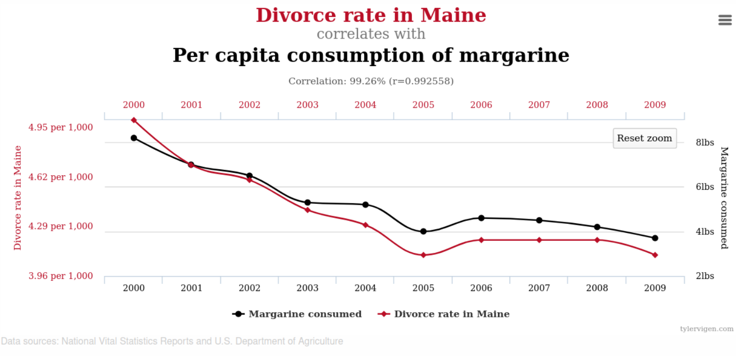


Source: `https://xkcd.com/552/`

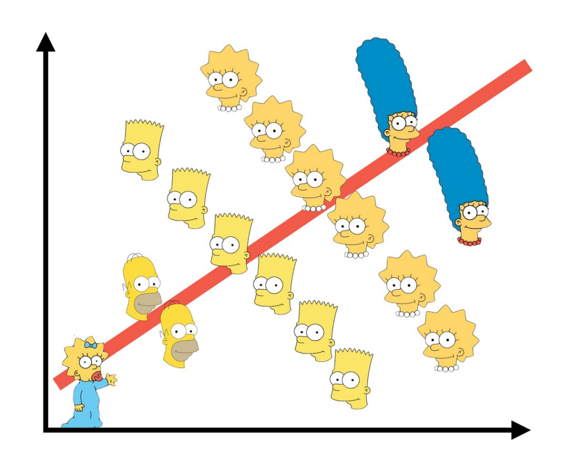
Spurious correlations

- Correlations can be found between unrelated variables
- Procrastinate : `https://www.tylervigen.com/spurious-correlations`

  → What possible confounders could explain these correlations ?



Simpson's paradox



https://www.arte.tv/fr/videos/107398-002-A/voyages-au-pays-des-maths/

Plan

Year 3000...

The Earth is finally a safe and pleasant place for humans again.

However, 1000 years of global warming released a dangerous bacteria from the permafrost.

The bacteria starts to infect human hosts, causing a mysterious disease.

Centuries in insipid watery ice made the bacteria obsessive about...

...vanilla ice-cream ! ♡

The illness is called

- Compulsive
- Obsessive
- Vanilla
- Ice-cream
- Disease

CHAOS ! !

The bacteria spreads rapidly, and infected humans start eating tons of vanilla ice-cream.

Milk prices rise to the stratosphere, ice-cream makers strike, diabetes and obesity break records...

Governments impose ice-cream lockdowns, interplanetary travel is forbidden, panic everywhere !

After months of an unprecedented crisis...

A lab finally announces a vaccine at phase 3 !

In phase 3, a vaccine is evaluated using an experiment called randomized control trial

Randomized control trial

Group A	Group B
Vaccine	Placebo

Average nb. ice-creams/day (ICD) :

- Group A : $ICD_A = 1.47$
- Group B : $ICD_B = 1.56$

Conclusion :

The vaccine works.
What a relief for humanity !

But. . . maybe humans forgot all about statistics ?

- Is the observed difference large enough ?
  - $ICD_A = 1.47$ ice/creams per day
  - $ICD_B = 1.56$ ice/creams per day

$$\delta = ICD_B - ICD_A = 0.09$$

- Maybe the sample is too small or biased

  $\rightarrow$ Affects our conclusion that vaccine (A) better than placebo (B) ?

But. . . maybe humans forgot all about statistics ?

- Is the observed difference large enough ?
  - $ICD_A = 1.47$ ice/creams per day
  - $ICD_B = 1.56$ ice/creams per day

$$\delta = ICD_B - ICD_A = 0.09$$

- Maybe the sample is too small or biased

  $\rightarrow$ Affects our conclusion that vaccine (A) better than placebo (B) ?

Given the samples, the metrics, and the experiment's conditions :
Probability of making a false claim assuming A $\neq$ B in general ?

$\rightarrow$ **p-value** !

System comparison

- Incremental research
  - State of the art or Baseline system B (placebo)
  - My own Awesome proposal system A (vaccin)
- How can I check whether A is **better** than B ?
- What's the probability of drawing a wrong conclusion ?

  → Ideally, very low, close to zero
- Methodological framework

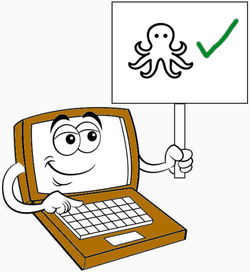
  → Take inspiration from health, biology, social siences

Comparison framework : example

- Our Baseline system classifies images
  - → Two categories : octopus or not octopus

Comparison framework : example

- Our Baseline system classifies images
  - → Two categories : octopus or not octopus

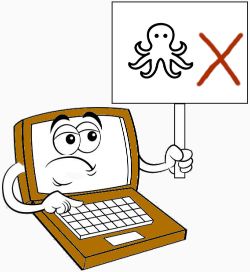


Comparison framework : example

- Our Baseline system classifies images
  → Two categories : octopus or not octopus

Comparison framework : example

- Our Baseline system classifies images
  → Two categories : octopus or not octopus



Comparison framework : example

- Our Baseline system classifies images

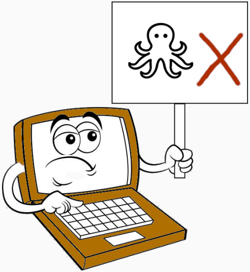
  → Two categories : octopus or not octopus

Comparison framework : example

- Our Baseline system classifies images

  $\longrightarrow$ Two categories : octopus or not octopus

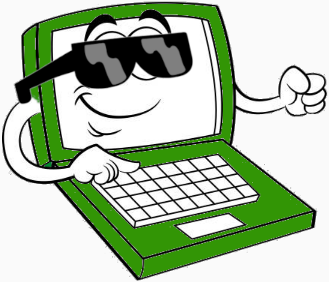
- Sometimes it makes mistakes



Comparison framework : example

- We developed an Awesome new system !
  - $\rightarrow$ E.g. the new system was trained on more data

$\Longrightarrow$

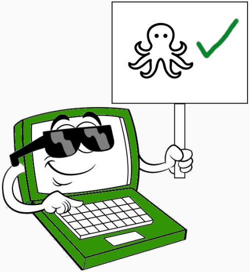


Comparison framework : example

- We developed an Awesome new system !
  - → E.g. the new system was trained on more data

Comparison framework : example

- We developed an Awesome new system !
  - $\rightarrow$ E.g. the new system was trained on more data
- It seems that it makes less mistakes $\Longrightarrow$ 🎉



Test set

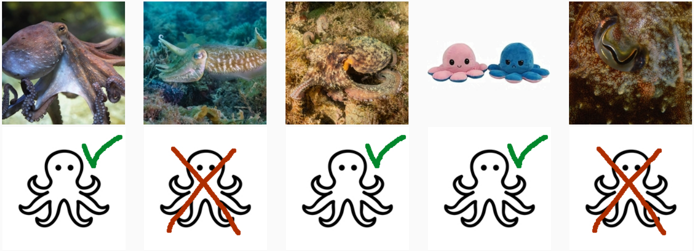
- Is A really better than B ?
  → Testing on a couple examples is not enough !
- Use a test set containing (x,y) pairs
  → x - sea animal images
  → y - gold/reference octopus / other labels
- The test set was not used to build the system

Test set : example

Images x selected to be in the held-out test set

$x \rightarrow$

$y \rightarrow$

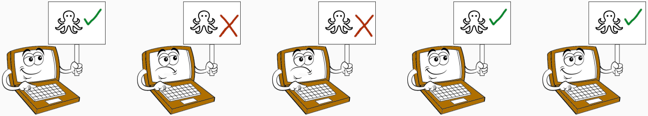


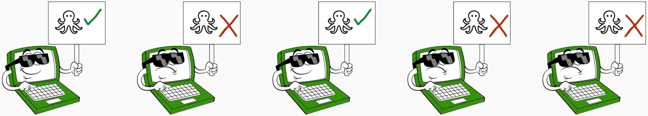
Reference/gold labels y considered true (e.g. annotated by humans)

System predictions

Both systems generate predictions $\hat{y}$ for test set instances x

$x \rightarrow$ 

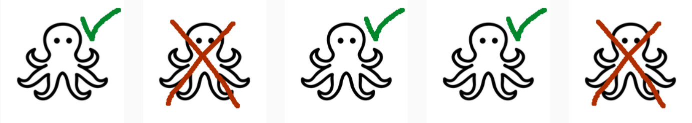
$\hat{y}_B(x) \rightarrow$ 

$\hat{y}_A(x) \rightarrow$ 

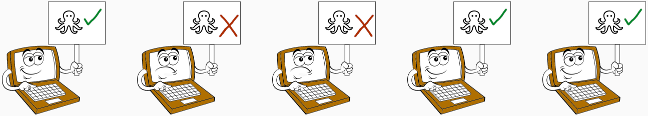
Evaluation metrics

Compare predictions $\hat{y}_B$ and $\hat{y}_A$ to reference y

$y \rightarrow$



$\hat{y}_B(x) \rightarrow$

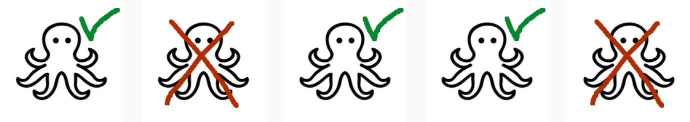


$$M(B, x, y) = \frac{3}{5} = 0.6$$

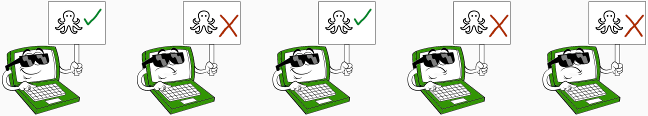
Evaluation metrics

Compare predictions $\hat{y}_B$ and $\hat{y}_A$ to reference y

$y \rightarrow$



$\hat{y}_A(x) \rightarrow$



$$M(A, x, y) = \frac{4}{5} = 0.8$$

Wooclap

Wooclap time !

System score comparison

- The accuracies of both systems are :

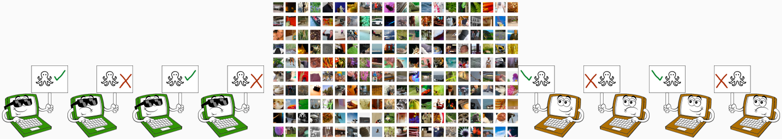
$$M(B, x, y) = \frac{3}{5} = 0.6$$
$$M(A, x, y) = \frac{4}{5} = 0.8$$

- It seems like A is better than B
- The difference (delta) is positive

$$\delta_{A-B}(x, y) = M(B, x, y) - M(A, x, y) = 0.8 - 0.6 = \boxed{0.2}$$

System comparison : example

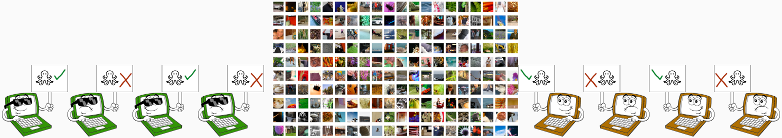
We obtained a much larger test set x',y'



We compare A and B again and obtain :

System comparison : example

We obtained a much larger test set x',y'

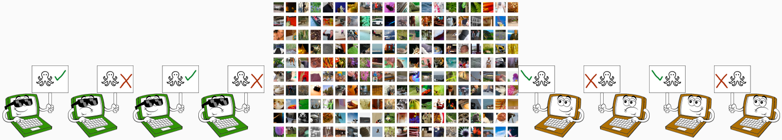


We compare A and B again and obtain :

$$\delta_{A-B}(x', y') = M(B, x', y') - M(A, x', y')$$
$$= 0.7612 - 0.7586$$
$$= \boxed{0.0026}$$

System comparison : example

We obtained a much larger test set x',y'



We compare A and B again and obtain :

$$\delta_{A-B}(x',y') = M(B,x',y') - M(A,x',y')$$
$$= 0.7612 - 0.7586$$
$$= \boxed{0.0026}$$

- Can we still affirm that A is better than B ?
- If we add or remove a couple of images, could the result flip ?

Interpretting delta

$$\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y)$$

- Delta allows us to translate the comparison into maths
  - $\rightarrow$ A better than B $\rightarrow$ $\delta_{A-B}(x, y) > 0$
  - $\rightarrow$ A equivalent to B $\rightarrow$ $\delta_{A-B}(x, y) = 0$
  - $\rightarrow$ A worse[2] than B $\rightarrow$ $\delta_{A-B}(x, y) < 0$
- In some disciplines, $\delta_{A-B}(x, y)$ is called effect

2. Yes, the old Baseline may beat the new Awesome system !

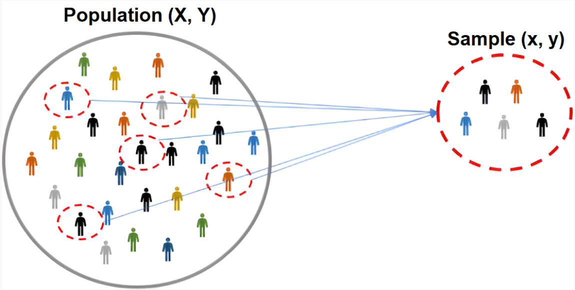
In short : maximise the effect !

1. We develop a system A supposed to be better than B
2. To verify this, we apply both systems to the same test set :
   - → Get output of system A on the test set (x, y)
   - → Get output of system B on the test set (x, y)
3. Calculate the evaluation metric $M(\cdot)$ for both outputs

$$\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y)$$

4. Large positive $\delta_{A-B}(x, y) \implies$ 🎉
5. In practice, $\delta_{A-B}(x, y)$ is often small 😑

Test sets as random samples

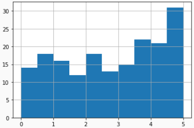


- Could the observed $\delta_{A-B}(x, y) > 0$ be due to chance ?
  - $\rightarrow$ (x, y) is a sample of joint random variables (X, Y)
  - $\rightarrow$ What effect/difference would be observed for sample (x', y') ?
- What is the probability that A is actually no better than B
  - $\rightarrow$ If we ever had access to the "real" distribution of (X, Y) ?

Effects as random variables

- We obtain a single $\delta_{A-B}(x, y)$ value
- This value depends on the test set (x, y), which is a sample
- We can see $\delta_{A-B}(x, y)$ as a sampled value of a random variable

$$\delta_{A-B}(X, Y) \rightsquigarrow$$



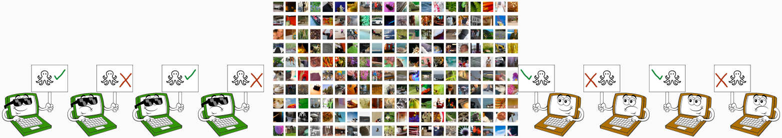
P-value

- **P-value** : probability of obtaining at least $\delta_{A-B}(x, y)$
  - When in reality, A is no better than B
- In short : p-value = probability that your conclusion is wrong !

Wooclap

Wooclap time !

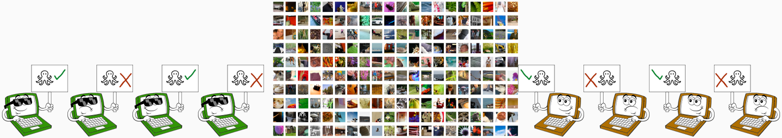
P-value : example



We have one value obtained on the large dataset (x', y')

$$\delta_{A-B}(x', y') = 0.0026$$

P-value : example



We have one value obtained on the large dataset (x', y')

$$\delta_{A-B}(x', y') = 0.0026$$

If we had all possible images of sea creatures X and their classes

$\rightarrow$ Imagine we have access to the real distribution $\delta_{A-B}(X, Y)$

- Probability of obtaining 0.0026 difference (or more)
- If A is actually no better than B

Hypothesis testing

- $H_0 : \delta_{A-B}(X, Y) \leq 0 \implies$ if true, then A not better than B
- $H_1 : \delta_{A-B}(X, Y) > 0$

- Goal : reject H_0

  $\rightarrow$ Conclusion : significant difference between the systems

Hypothesis testing and p-value

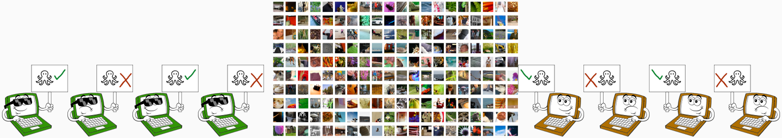
Remember

- $H_0 : \delta_{A-B}(X, Y) \leq 0$
- $H_1 : \delta_{A-B}(X, Y) > 0$

- **P-value** : probability of observing $\delta_{A-B}(x, y$ while H_0 true

  $\rightarrow$ Intuituion : if H_0 was true, large $\delta_{A-B}(x, y)$ are unlikely
- In mathematical notation :

$$\text{p-value} = P\{\delta_{A-B}(X, Y) \geq \delta_{A-B}(x, y) \mid H_0\}$$



$$\text{p-value} = P\{\delta_{A-B}(X, Y) \geq 0.0026 \mid \delta_{A-B}(X, Y) \leq 0\}$$

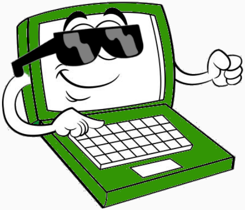
Estimate p-value, if small enough $\implies$ A better than B

Type I errors

- Type I error : false positive

  $\rightarrow$ Rejecting H_0 when it is actually true

Conclusion of the test :



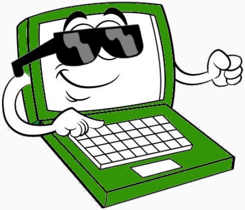
is better than

Reality : But it isn't better !

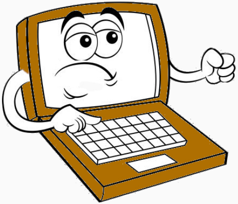
Type II errors

- Type II error : false negative
  - $\rightarrow$ Not rejecting H_0 when it is actually false

Conclusion of the test :



is not better than



Reality : But it is better !

Goal

- Probability of type-I error is upper bounded by α

  $\rightarrow$ α is called the significance level or threshold
- Probability of type-II error is as low as possible

  $\rightarrow$ Test power : ability to avoid type-II errors

Statistically significant result

$$\text{p-value} < \alpha \implies \text{statistically significant!}$$ 🎉

- p-value : probability of extreme outcome
- α : significance threshold

  $\rightarrow$ Usual "magic" value : $\alpha = 0.05$

Statistically significant result

$$\text{p-value} < \alpha \implies \text{statistically significant!}$$ 🎉

- p-value : probability of extreme outcome
- α : significance threshold

  $\rightarrow$ Usual "magic" value : $\alpha = 0.05$

The word significant should not be used to anything else

How can we estimate p-values ?

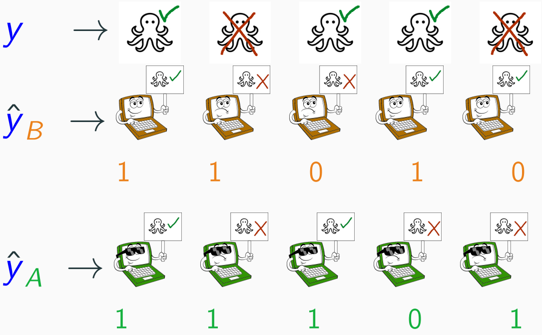
- P-value depends on $\delta_{A-B}(X, Y)$ probability distribution
- Which in turn depends on $M(A, x, y)$ and $M(B, x, y)$
  - $\rightarrow$ Remember : $M(\cdot)$ is our evaluation metric
- $M(\cdot)$'s distribution determines that of δ (if we're lucky)

$\Longrightarrow$ Study the probability distribution of $M(\cdot)$!

Wooclap

Wooclap time !

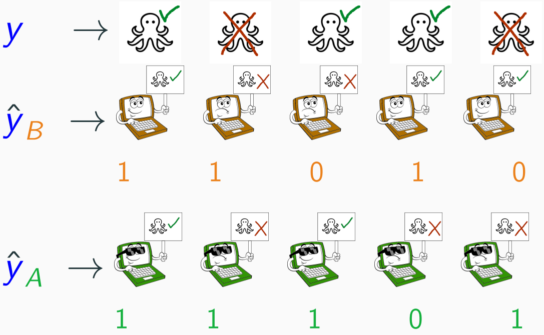
Accuracy is an average



$Acc_B = \frac{1+1+0+1+0}{5} = \frac{3}{5}$ $\qquad$ $Acc_A = \frac{1+1+1+0+1}{5} = \frac{4}{5}$

Accuracy is an average

Accuracy is an average



$$Acc_B = \frac{1+1+0+1+0}{5} = \frac{3}{5} \qquad Acc_A = \frac{1+1+1+0+1}{5} = \frac{4}{5}$$

Accuracy is an average

→ Normally distributed !

The t-test for paired samples

- T-test : hypothesis testing for normally distributed variables
- Based on Student's t distribution

  $\rightarrow$ Looks like normal distribution for large samples

$$\text{t-stat} = \frac{M(A, x, y) - M(B, x, y)}{SE/\sqrt{m}}$$

- m : size of the paired sample (x, y)
- SE : standard deviation of the difference $\hat{y}_A - \hat{y}_B$

The t-test for paired samples

- T-test : hypothesis testing for normally distributed variables
- Based on Student's t distribution

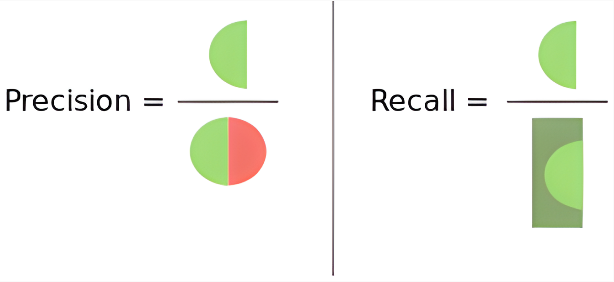
  → Looks like normal distribution for large samples

$$\text{t-stat} = \frac{M(A, x, y) - M(B, x, y)}{SE/\sqrt{m}}$$

- m : size of the paired sample (x, y)
- SE : standard deviation of the difference $\hat{y}_A - \hat{y}_B$

- P-value : check Student's t table, $m - 1$ degrees of freedom

Precision is not an average



- Recall ($\frac{tp}{tp+fn}$) can be seen as an average like accuracy
  - $\rightarrow$ $tp + fn$ does not depend on the system
- Precision ($\frac{tp}{tp+fp}$) cannot be seen as an average
  - $\rightarrow$ $tp + fp$ depends on the system
  - $\rightarrow$ System class distribution is unpredictable
- $\Longrightarrow$ F-score cannot be assumed to be normally distributed

Non parametric tests

- Problem of t-test : assumes $M(A, x, y) \sim$ normally distributed
- Other metrics :
  - Recall $R = \frac{tp}{tp+fn}$, $tp + fn$ constant
    $\rightarrow$ t-test OK ✓
  - Precision $P = \frac{tp}{tp+fp}$ depends on $tp + fp$, unknown distribution
    $\rightarrow$ t-test not OK ✗
  - F-score $2PR/(P + R)$ depends on P, unknown distribution
    $\rightarrow$ t-test not OK ✗

Parametric vs. non parametric

Many authors use the terms parametric vs. non parametric tests

- What does it mean ?
- Most of the time, by "parametric" we mean "the random variable normally distributed"

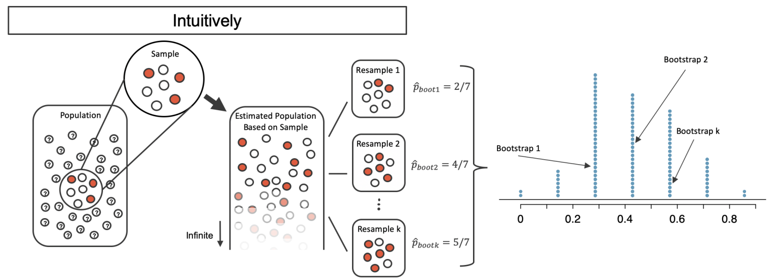
Non parametric tests

- Alternative : non parametric tests
  1. No sampling
     - Fast
     - Conservative, will not state A better than B for small δ (not powerful)
     - E.g. sign test, McNemar's test, Wilcoxon
  2. With sampling
     - Slow
     - Powerful, low type-II error probability
     - E.g. randomised approximaiton, bootstrap test

Source : Yeh (2000) `https://aclanthology.org/C00-2137/`

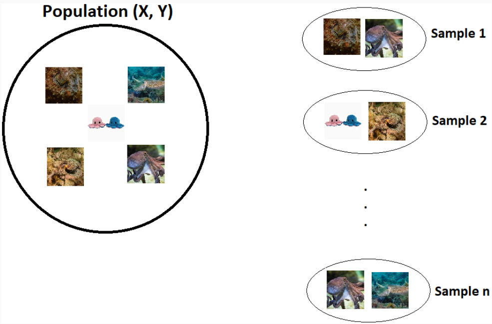
Bootstrap

Idea : estimate M distribution by random re-sampling in x, y



https://bookdown.org/gregcox7/ims_psych/foundations-bootstrapping.html

Bootstrap



$Acc(A, \text{Sample 1}) = \frac{2}{2}$

$Acc(B, \text{Sample 1}) = \frac{1}{2}$

$Acc(A, \text{Sample 2}) = \frac{1}{2}$

$Acc(B, \text{Sample 2}) = \frac{1}{2}$

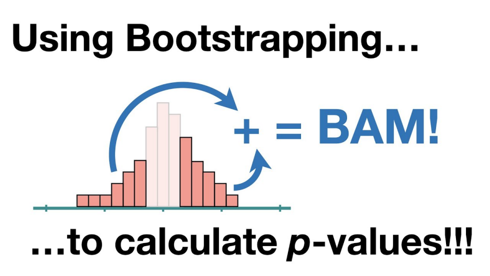
$Acc(A, \text{Sample n}) = \frac{2}{2}$

$Acc(B, \text{Sample n}) = \frac{2}{2}$

Bootstrap for significance

```
deltaobs = M(A,x,y) - M(B,x,y)  # delta on test set
R = 10000                       # 10k random samples
for i = 1 .. R :
  xs, ys = sample(x,y,m)        # with repetition
  deltasample = M(A,xs,ys) - M(B,xs,ys)
  if deltasample > 2 * deltaobs :
        r = r + 1
pvalue = r/R
```

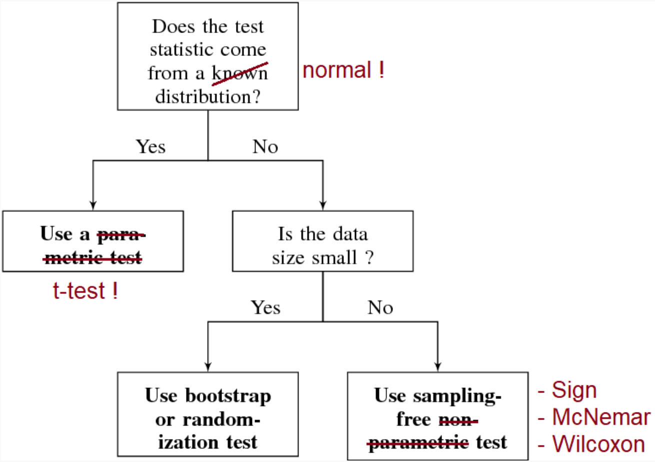
Why comparing with 2 × deltaobs ?



StatQuest with Josh Starmer

https://www.youtube.com/watch?v=N4ZQQqyIf6k

Which test to apply ?



Source: Dror et al. (2018) https://aclanthology.org/P18-1128/

Evaluation metric M distribution vs. test

- Parametric test ($M(A, x, y)$ from known distribution)
  - Paired Student's t-test
- Non-parametric tests ($M(A, x, y)$ from unknown distribution)
  - No sampling (less powerful)
    - Sign test
    - McNemar's test
    - Wilcoxon signed rank test
  - Sampling (computationally expensive)
    - Permutation (randomized approximation) test
    - Bootstrap test

Multiple comparisons

- Multiple comparisons : probability of false claims increases
- Bonferroni's correction
  - Divide significance level α by the number of tests N
- Replicability analysis (Dror et al. 2020)

P-hacking

A significant *p*-value can always be obtained

$\rightarrow$ As long as the sample is large enough

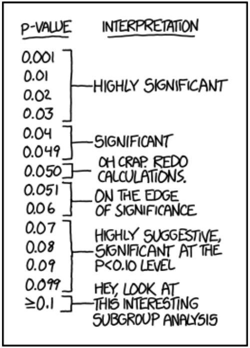
$\rightarrow$ `https://www.youtube.com/watch?v=HDCOUXE3HMM`

P-hacking

A significant *p*-value can always be obtained

→ As long as the sample is large enough

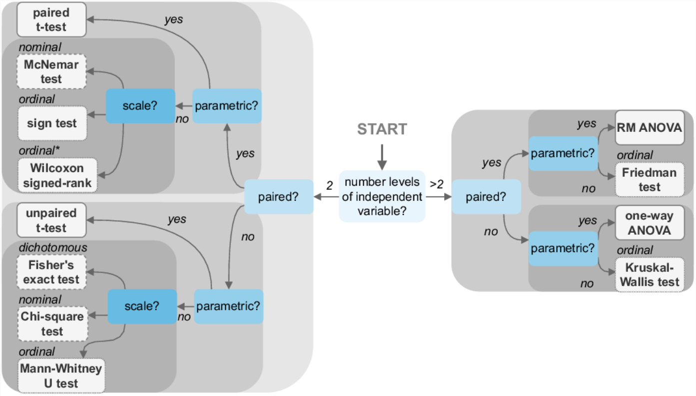
→ https://www.youtube.com/watch?v=HDCOUXE3HMM



Source: https://xkcd.com/1478/

Unpaired samples

- We only covered significance for paired samples
  - → Two systems A and B, same dataset items (x,y)
  - → Other tests for unpaired samples



 Source: https://doi.org/10.1017/S1351324922000535, thanks to Elie Antoine

Plan

Community's practice

NLP conferences (ACL) and journals (TACL)

General Statistics	**ACL '17**	**TACL '17**
Total number of papers	196	37
# papers that **do not** report significance	117	15
# papers that report significance	63	18
# papers that report significance but use the **wrong** statistical test	6	0
# papers that report significance but do not mention the test name	21	3

Source: Dror et al. 2018

Statistics libraries

- Visual : Excel, Libreoffice, ...
- Python : `matplotlib`, `numpy`, `scipy`, `sklearn`, ...
- R : multiple libraries including linear models
- Proprietary : Matlab, SPSS, ...

Error analysis

- Characterise the errors in our system's output
- Scripts to print characteristics of errors

  $\longrightarrow$ Frequency, length, resolution, predicted/gold class, ...

  $\longrightarrow$ Example : compounds predicted in wrongest positions
- Manual error annotation : taxonomies, guidelines

  $\longrightarrow$ Gain insight on most promising improvements

Interpretability analysis

Try to understand why systems generate a prediction

- Feature-based methods (SHAP, LIME)
  → Which parts of the inputs influence prediction ?
- Visualisation
  → Attention salience, 2-D projection (UMAP, t-SNE, topology)
- Adversarial examples, perturbations
  → Difficult minimal pairs

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Source: `https://homes.cs.washington.edu/~marcotcr/blog/lime/`

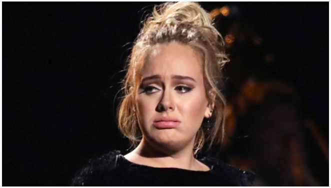
Leaderboards

- Remember Goodhart's law (metric $\neq$ objective)
- Beating state of the art is good
- Learning something interesting about the problem is better
- From time to time : remember the research question

Negative results

- Well designed hypotheses → interesting "negative" results
- Experiments require persistence and somea faith
- Source of frustration : publish or perish

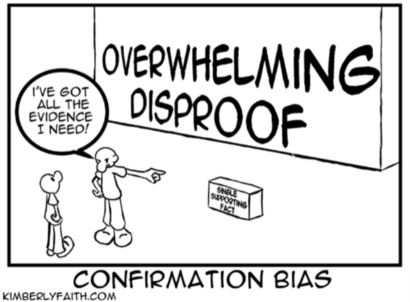
  → Is it a problem with my results or with the system ?
- Negative results are publishable if sound experimental design



Confirmation bias

- Tendency to favour interpretations that confirm initial beliefs
- May lead to cognitive dissonance, well studied in psychology
- Tip : try to demonstrate the opposite of the initial hypothesis

  $\rightarrow$ If you fail for long enough, maybe the initial hypothesis is true



Source: `https://moveyourcompanyforward.com/2020/11/03/four-ways-to-overcome-confirmation-bias/`

Sources

- Cours d'Adeline Paiement
- Statistical Significance Testing for NLP (Dror et al. 2020)
- https://bodo-winter.net/tutorials.html (thanks Leonardo Pinto Arata)
- Wikipedia
- Google images
- StatQuest Youtube : https://www.youtube.com/@statquest

Backup slides

Random variables : formal definition i

- Experiment : flip 3 different coins, note head (H) or tail (T)
- The sample space S contains all possible experiment outcomes

  $\rightarrow$ The subsets of S are called events E_i
- The random variable X denots the number of heads (H)
  - A variable whose exact value is unknown or irrelevant
  - We know (or estimate) its probability distribution $P\{X = x_i\}$

E_i	$\{HHH\}$	$\{THH, HTH, HHT\}$	$\{TTH, THT, HTT\}$	$\{TTT\}$
$P(E_i)$	1/8	1/8 + 1/8 + 1/8	1/8 + 1/8 + 1/8	1/8
X	0	1	2	3
$P\{X = x_i\}$	1/8	3/8	3/8	1/8

Random variables : formal definition ii

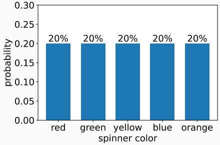
Formalisation

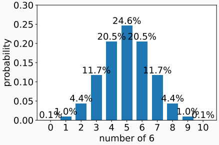
A random variable is a function $X : S \to \mathbb{R}$ such that :

1. Discrete random variable :
   - $\rightarrow$ Its set of possible values $X(S) = \{x_i, i \in \mathbb{N}^*\}$ is countable
   - $\rightarrow$ For all $x_i \in X(S)$: $\{X = x_i\} \Leftrightarrow \{e_i \in S | X(e_i) = x_i\} \in \mathcal{F}$
   - $\rightarrow$ $\mathcal{F}$ is the set of all possible events (subsets) of S
   - $\rightarrow$ $p(x_i) = P\{X = x_i\}$ is the probability mass function of X
2. Continuous random variable :
   - $\rightarrow$ $\forall$ value $x \in (-\infty, +\infty)$, $\forall$ interval $B \in \mathbb{R}$
   - $\rightarrow$ A non-negative function $P\{X \in B\} = \int_B f(x)\, dx$ exists
   - $\rightarrow$ $f(x)$ is the probability density function of X

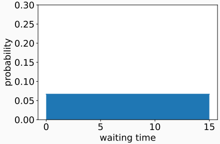
Types of probability distributions

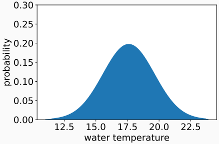
- Discrete random variables
  - $\rightarrow$ Bar graphic, finite set of values
  - $\rightarrow$ Probability at exact value $P\{X = a\}$





- Continuous random variables
  - $\rightarrow$ Line graphic, uncountable set of values (real numbers)
  - $\rightarrow$ Probability of interval $P\{a < X < b\}$





Random sample or i.i.d. variables ?

- Sampled items can be seen as n random variables $X_1 \ldots X_n$
  $\rightarrow$ For instance, tossing a coin n times
- We assume that all variables have the same distribution
- We assume that all items are independent[3]
- This is often stated as independent and identically distributed
  $\rightarrow$ The acronym i.i.d. is usually employed in probability

3. Formally : $\forall X_i \neq X_j, \forall a, b \in X_i(S) \quad P\{X_i = a | X_j = b\} = P\{X_i = a\}$

Random sample or i.i.d. variables ?

- Sampled items can be seen as n random variables $X_1 \dots X_n$

  $\rightarrow$ For instance, tossing a coin n times
- We assume that all variables have the same distribution
- We assume that all items are independent [3]
- This is often stated as independent and identically distributed

  $\rightarrow$ The acronym i.i.d. is usually employed in probability

Random sample = set of n values of i.i.d. variables $X_1 \dots X_n$

3. Formally : $\forall X_i \neq X_j, \forall a, b \in X_i(S) \quad P\{X_i = a | X_j = b\} = P\{X_i = a\}$

Correlation significance

- A simple transformation of r can be proved following a Student T distribution
- One can know quite straightforward if a correlation is significantly different from 0
- Most libraries provide this p-value by default
- More details : Dror et al. Significativity tests for NLP - M&C book

Kendall-tau correlation

- Rank correlation, distinguishes local/distant mismatches
  - → Ranking an item 5 instead of 3 is not too bad
  - → Ranking an item 58 instead of 3 is really bad
- Consider all possible pairs (x_i, x_j) and (y_i, y_j) with $i < j$
  - → If $x_i < x_j$ and $y_i < y_j \implies$ concordant
  - → If $x_i > x_j$ and $y_i > y_j \implies$ concordant
  - → Else, discordant pairs

$$\tau = \frac{\#(concordant\ pairs) - \#(discordant\ pairs)}{\#(total\ pairs)}$$
$$= 1 - \frac{2 \times \#(discordant\ pairs)}{\binom{n}{2}}$$

Example : https://www.statisticshowto.com/kendalls-tau/

Advanced data analysis

- Correlation works well for 2 numerical variables
- What if the variables are categorical ?
- What if we have more than 2 variables ?

Advanced data analysis

- Correlation works well for 2 numerical variables
- What if the variables are categorical ?
- What if we have more than 2 variables ?

Further statistical tools

- Information theory
- ANOVA
- Linear models
- Mixed models
- ...

Information theory

- Entropy : alternative view of variability/skewness
  - $\rightarrow H = -\sum p(x_i) \log p(x_i)$ $\rightarrow$ amount of uncertainty
  - $\rightarrow H = \max$ for uniform distribution (unpredictable)
  - $\rightarrow H = 0$ for highly skewed distribution (predictable)
- Other useful notions :
  - $\rightarrow$ Cross entropy
  - $\rightarrow$ Mutual information
  - $\rightarrow$ Kullbak-Leibler divergence (asymmetric)
  - $\rightarrow$ Jensen–Shannon divergence (symmetric)

Models for categorical variables

- ANOVA : Generalise t-test for more than 2 means
- Linear model : predict a linear regression slope
  - → Is the slope significantly different from zero ?
  - → Notation : pitch $\approx$ sex $+\varepsilon$
- Mixed model : more sophisticated for multiple factors