DATA SCIENCE COURSE

REMINDERS FOR A GREAT PROJECT!

Marc Santolini
Liubov Tupikina
7 Nov 2023
LPI Paris

marc.santolini@cri-paris.org

video: Kim Albrecht

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answer?)
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with random expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest — this should be done! (your name + mentor on google docs)
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answer?)
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with random expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answer?)
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with random expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

can be personal motivation, can be scientific interest, can be societal impact…

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question y
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with random expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

what are scientific articles about it? what methods have been deployed? what results obtained?

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answer?)
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with random expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

be specific!

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answ
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with random expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

describe the data you found, what format it is in, how it was obtained (source, method, reliability), how it allows to tackle the research question (including limitations, i.e. it allows to answer in a specific context)

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to a
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with rand
  expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

make the data tidy
keep only variables / categories of interest
look for outliers, artefacts? (e.g. errors in data collection process)
describe missing data (remove? impute?)

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you in
- **Extract data** (how did you get it? crawl/parse/dow
- **Clean data** (convert data type, remove outliers etc
- **Analyse data** (answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison v
  expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspec

Exploratory Data Analysis (EDA)

How is my data distributed?
Is it normal, heavy-tailed?
Should I log-transform?
Do I have outliers?
Should I remove them?
Do I have missing data?
Should I infer (impute) them?
Do I have very similar variables?
Should I reduce the dimensionality?
Do I have several peaks (multi-modality)?
Should I create sub-categories?
etc …
etc …

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answer?)
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with rand
  expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

scatter plot and correlation
boxplot and Mann-Whitney test
barplots and t-tests
always put labels, large font size,
legend, useful colors

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answer?)
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with random expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

effect size: how big is the effect? e.g:

➡ t-value in t test

➡ *Z* score when comparing observed value to random distribution

➡ AUC for Mann-Whitney

p-value: what is the probability it would occur at random?

$p<0.05$ is “significant”

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answer?)
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with random expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspecti

report what is the final answer to you research questions, positive or negative

About the project

Foundations for a full-fledged project:

- **Find a dataset** of interest
- **Motivate** the study (why is it important?)
- **Background** (what was explored before?)
- **Research question(s)** (what is the question you intend to answer?)
- **Extract data** (how did you get it? crawl/parse/download…)
- **Clean data** (convert data type, remove outliers etc)
- **Analyse data** (EDA, answer questions quantitatively)
- **Visualize data** (plots with labels, error bars etc)
- **Significance statement** (effect size, comparison with rand
  expectation, p-values) - including negative results!
- **Conclusion** (answer the research question)
- **Discussion** (what are limitations, what are perspectives)

be humble and highlight the limitations in the data, analysis (typically, what things you did not account for)

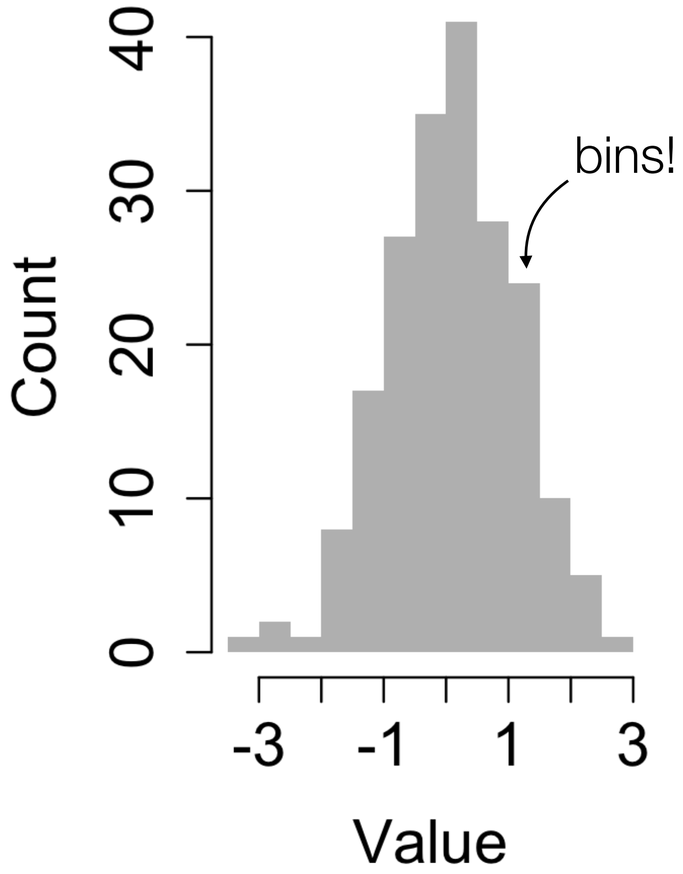
open up to next steps

some supporting material as a reminder

Representing a quantitative variable

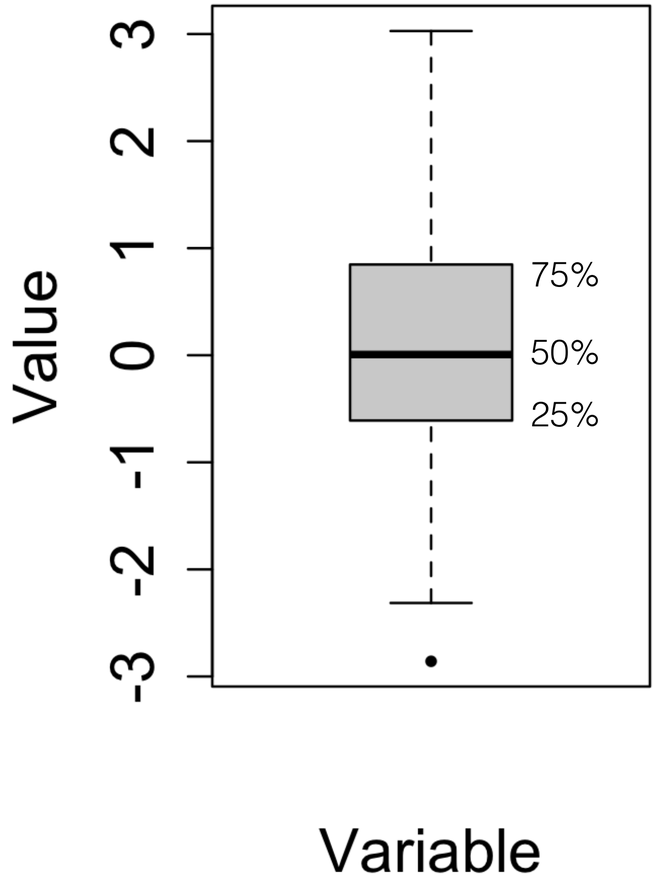
Histogram

‘parametric’
mean, standard deviation

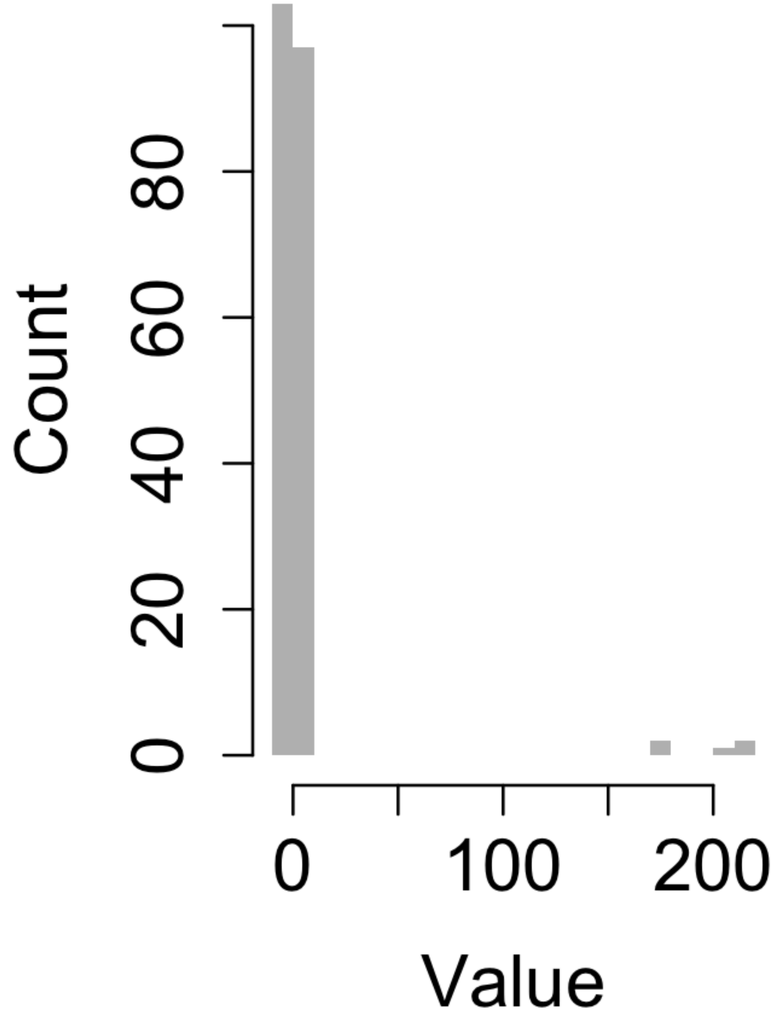


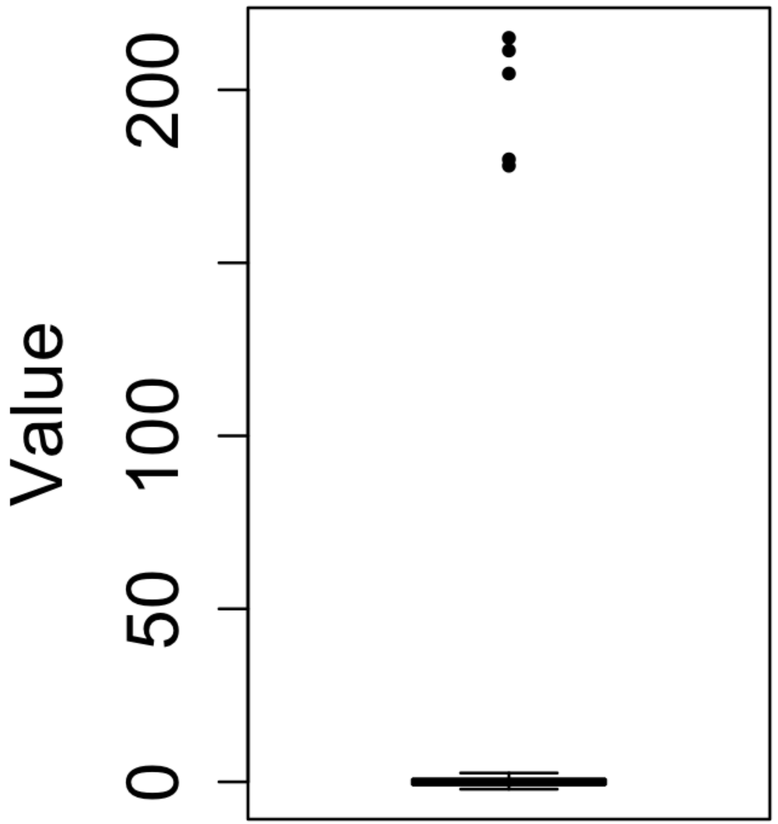
Boxplot

‘non-parametric’
median, inter-quantile range…

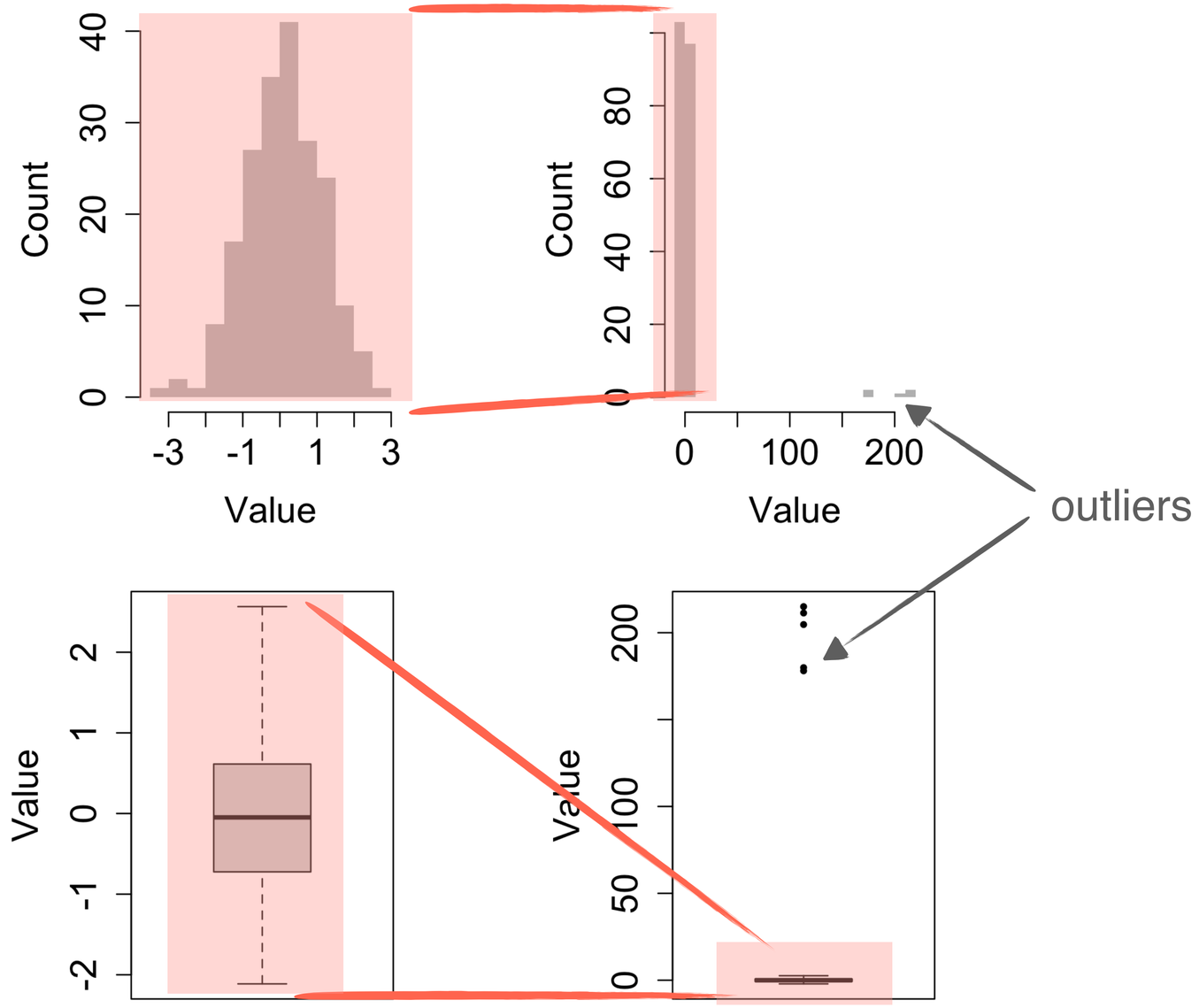


What happened here???

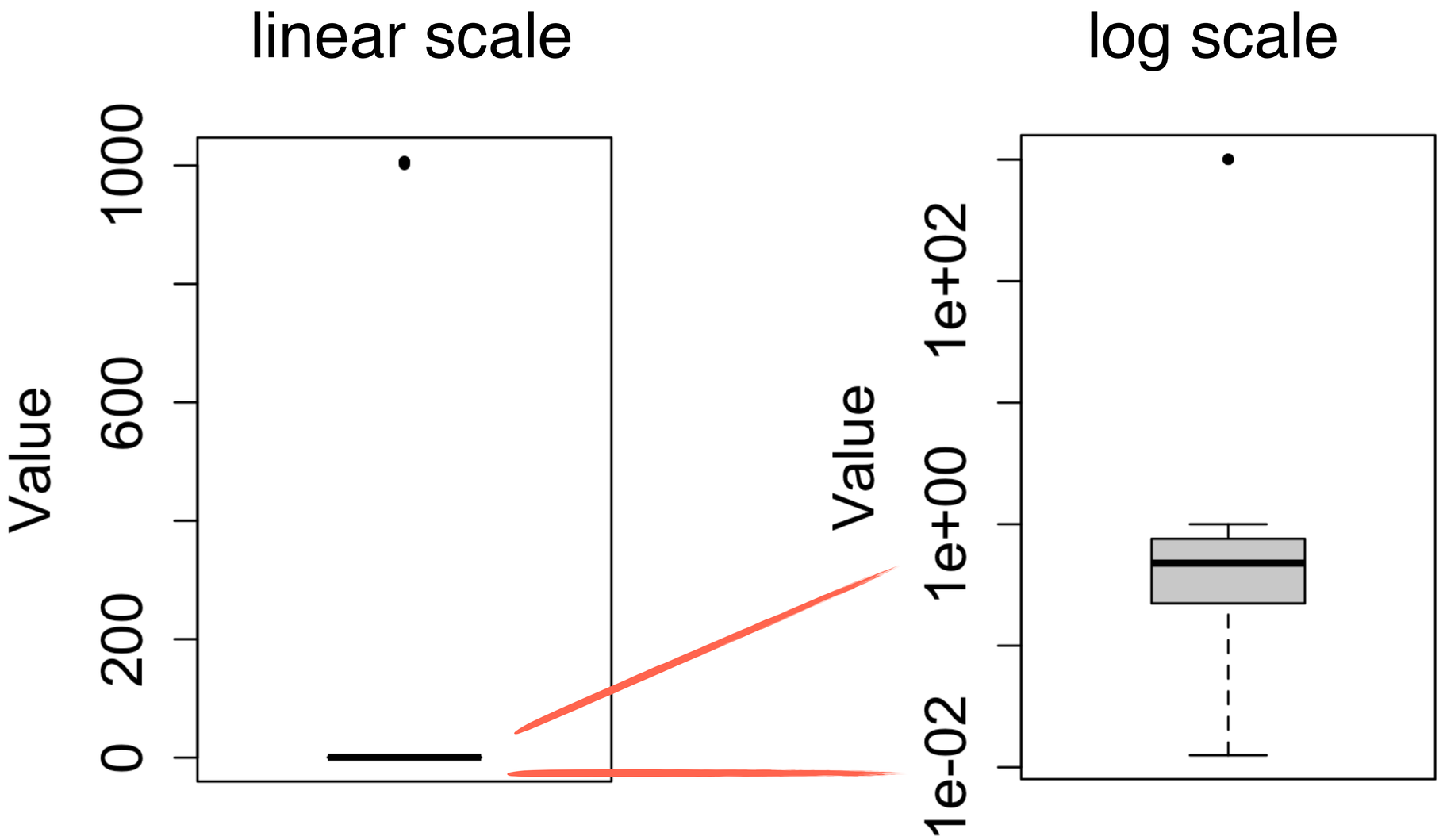




The importance of data cleaning…

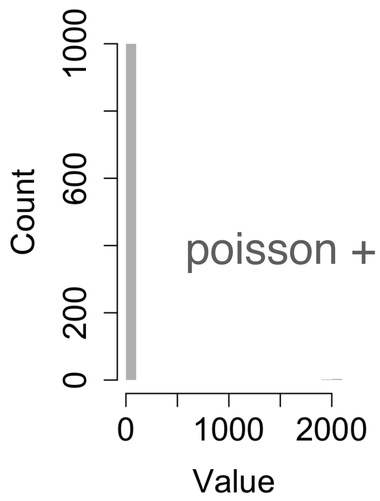


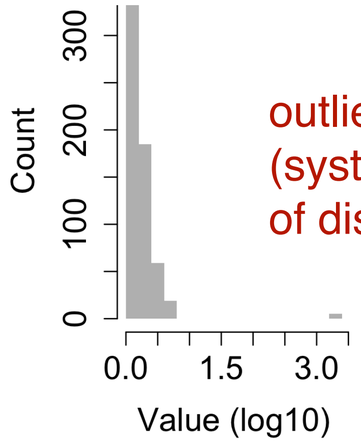
Expands small values!



Extreme events are usually not outliers, but part of the system

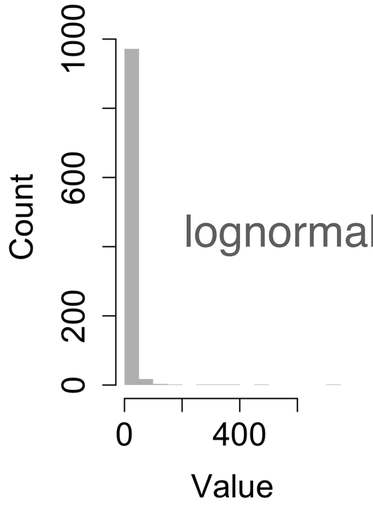
20th century statistics

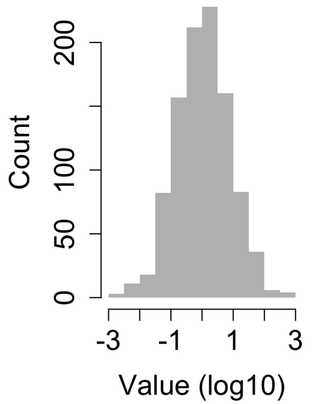




outliers!
(system is a mixture
of distributions)

21th century statistics:



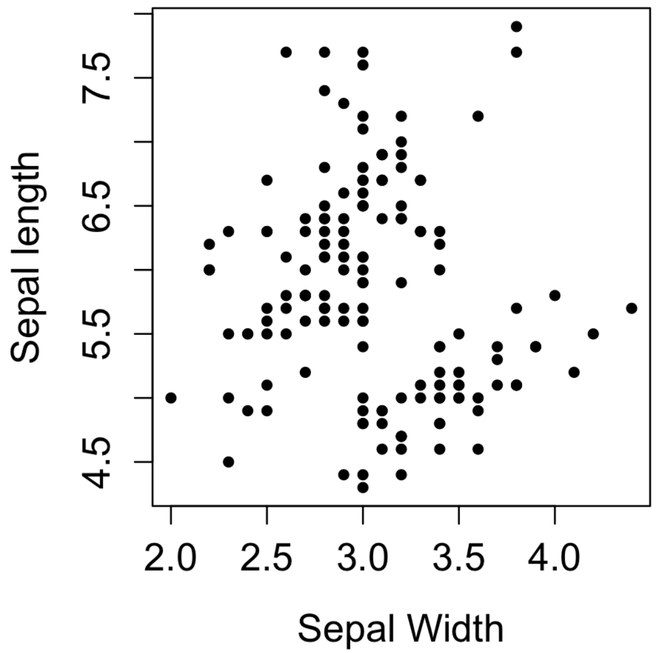


no outliers!
(one law describes
the whole system)

Different ways to represent variables

Scatterplot

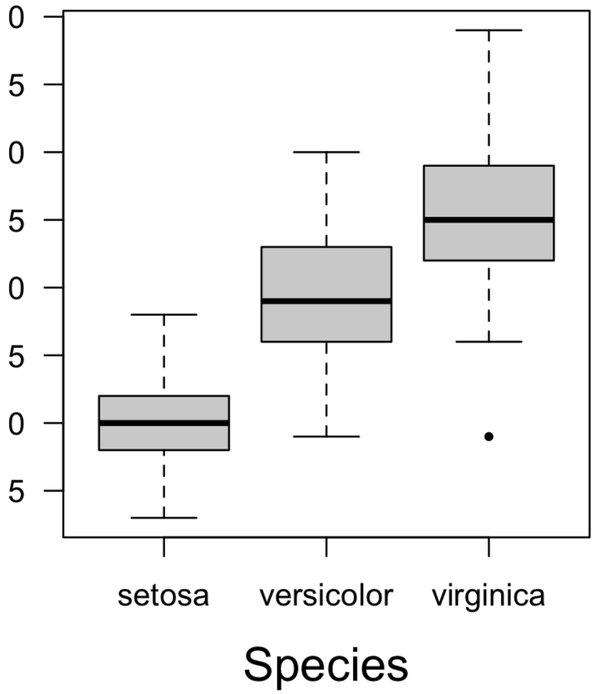
(compares quantitative variables)



How do x and y co-vary?

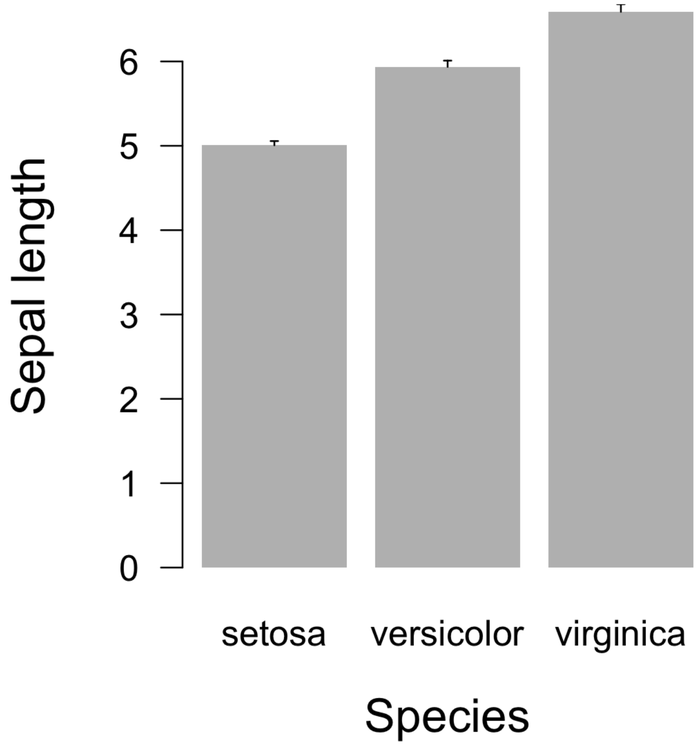
Boxplot

(compares **quantiles** across categories)



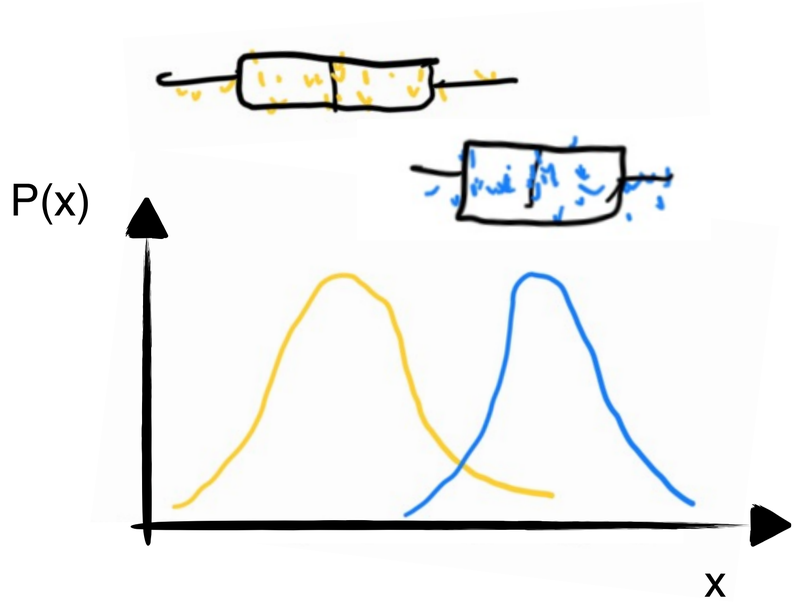
Barplot

(compares **means** acrosscategories)



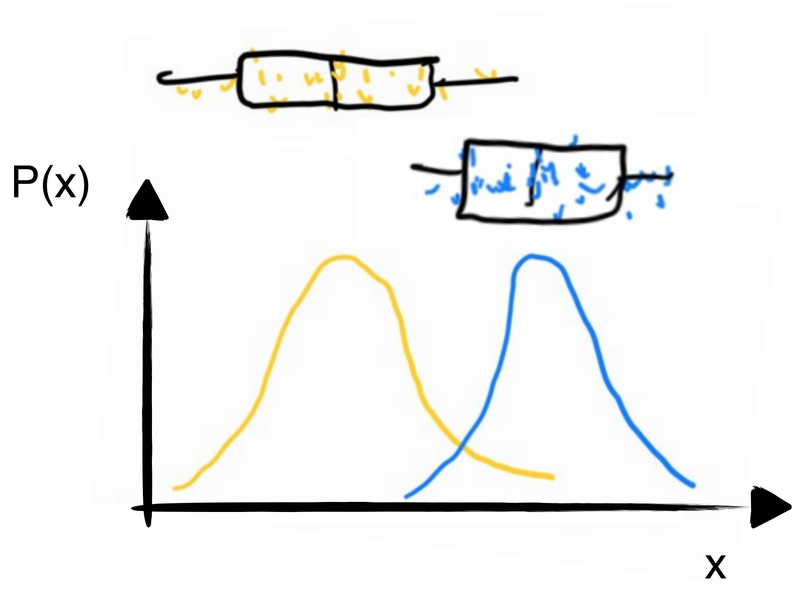
How do categories differ?

Comparing data distributions



How similar are
these distributions?

Comparing data distributions



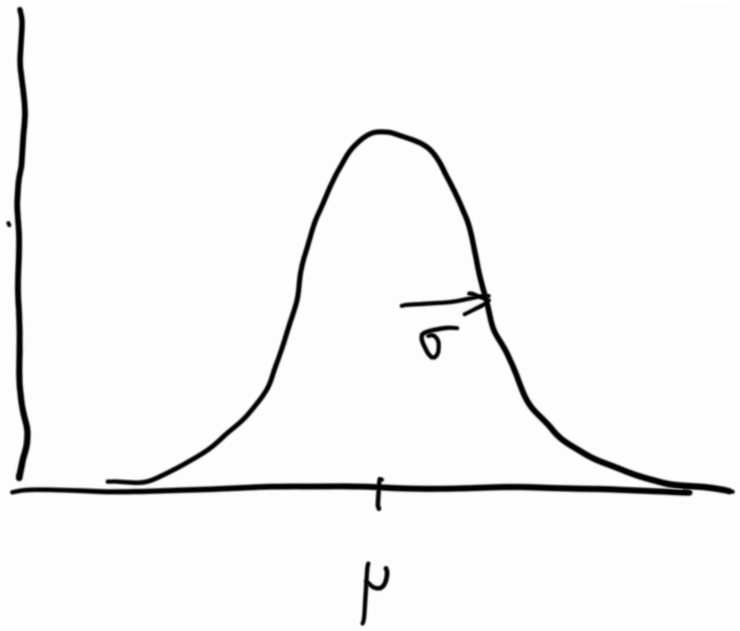
How similar are
these distributions?

1. Are the **means** different?

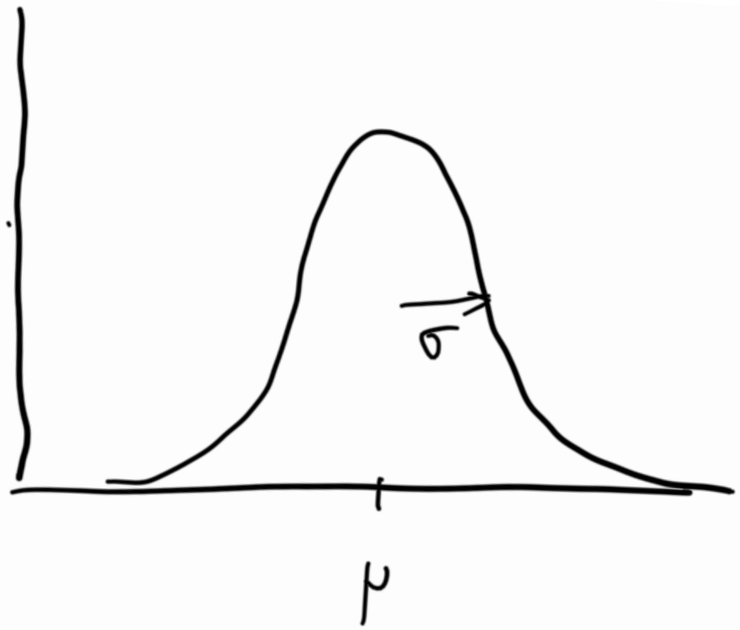
or

2. Are the **ranks** different?

Mean and standard deviation



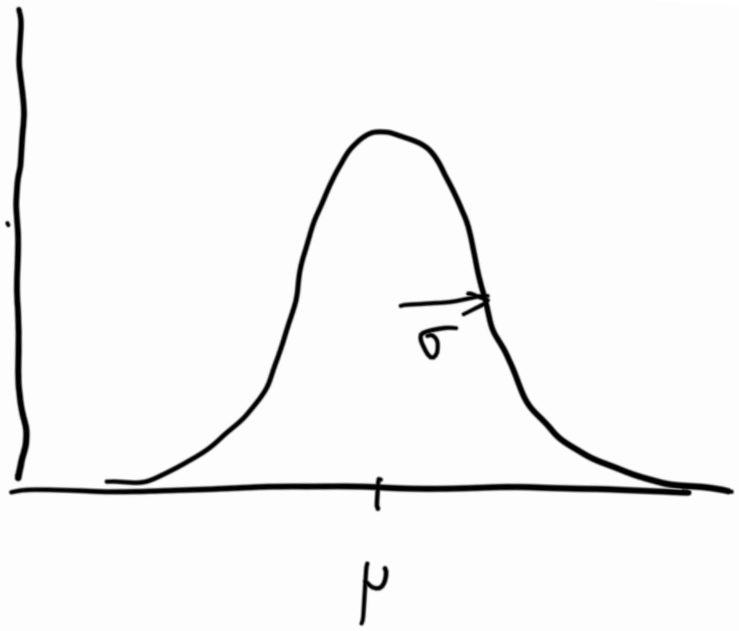
Mean and standard deviation



Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x$$

Mean and standard deviation



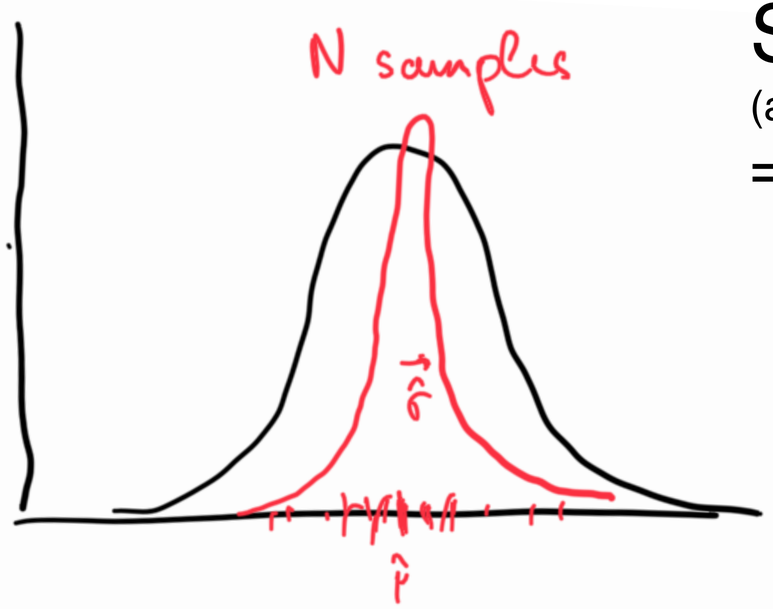
Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x$$

Standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Standard error

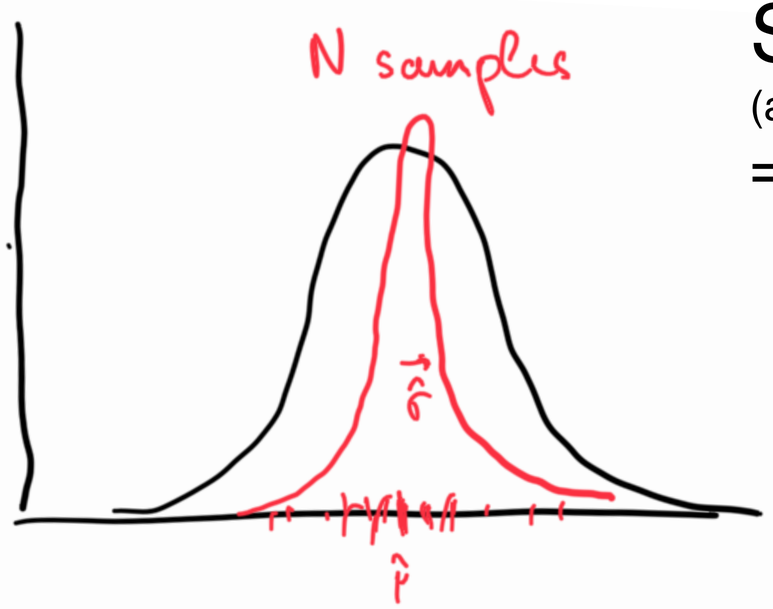


Standard error

(also Standard Error of the Mean SEM)

= what is the error when estimating the mean?

Standard error



$$\hat{\mu} \sim \mathcal{N}(\mu, \hat{\sigma})$$

Standard error

(also Standard Error of the Mean SEM)

= what is the error when estimating the mean?

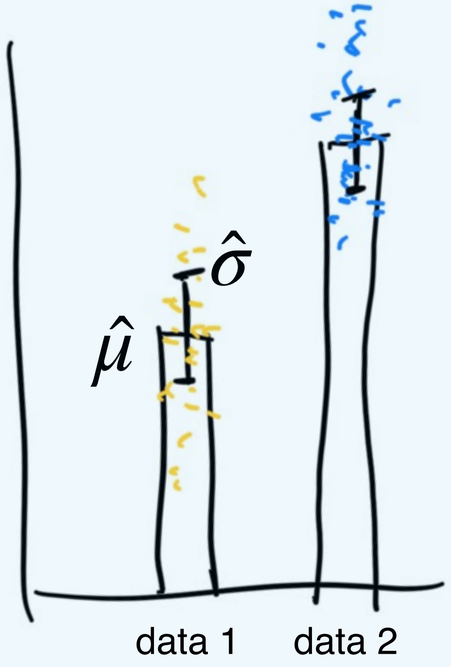
$$\hat{\sigma} = \frac{\sigma}{\sqrt{N}}$$

this comes from the “Central Limit Theorem”

Comparing means of samples

Student t-test

parametric



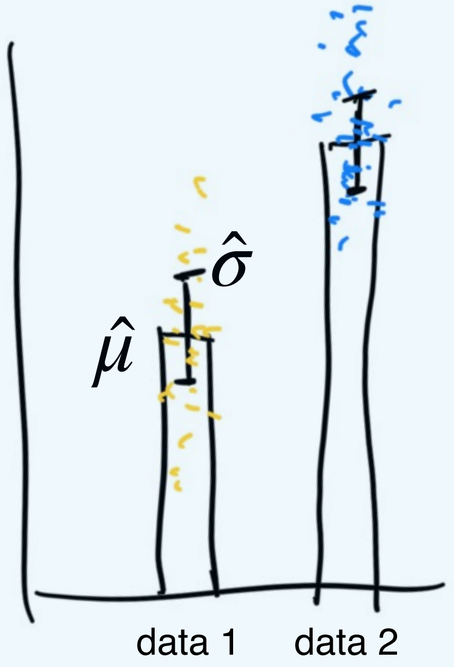
Are the means significantly far?

(The t-test is invalid for small samples from non-normal distributions, but it is valid for large samples as long there is a finite variance)

Comparing means of samples

Student t-test

parametric



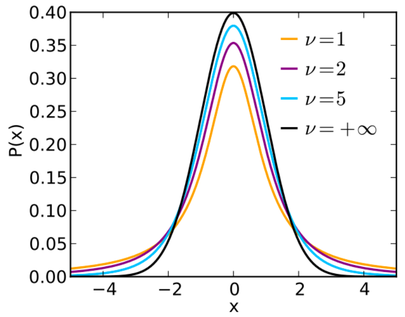
Are the means significantly far?

Two sample Student t-test

t-value

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}_1{}^2 + \hat{\sigma}_2{}^2}}$$

p-value



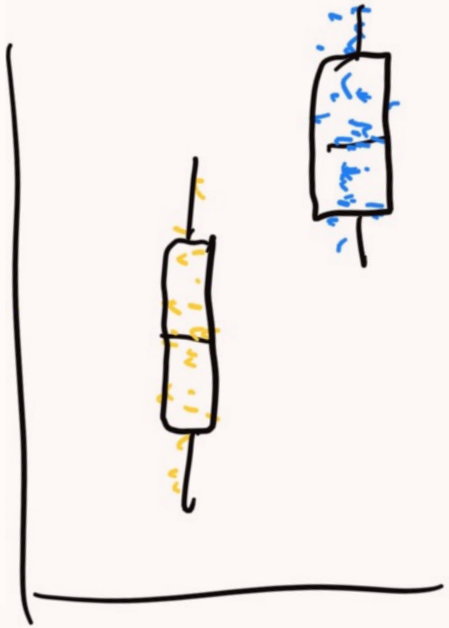
$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$P(t > t_{observed})$$

Comparing full distributions

Mann-Whitney U test
(or Wilcoxon signed-rank test)

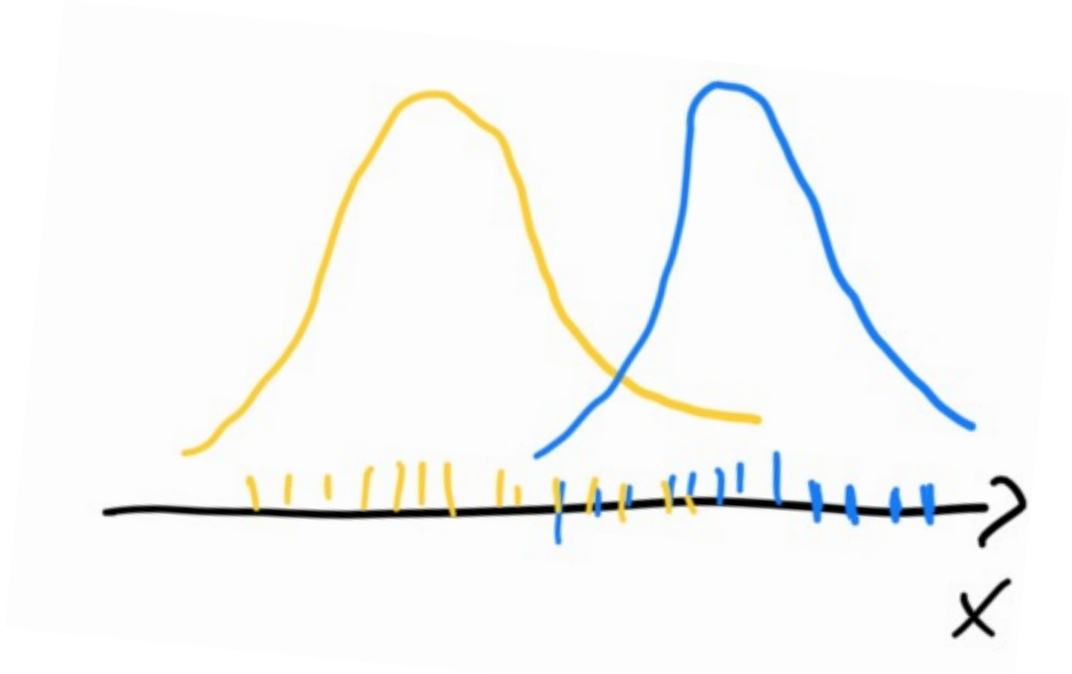
non-parametric



Are the ranks
different?

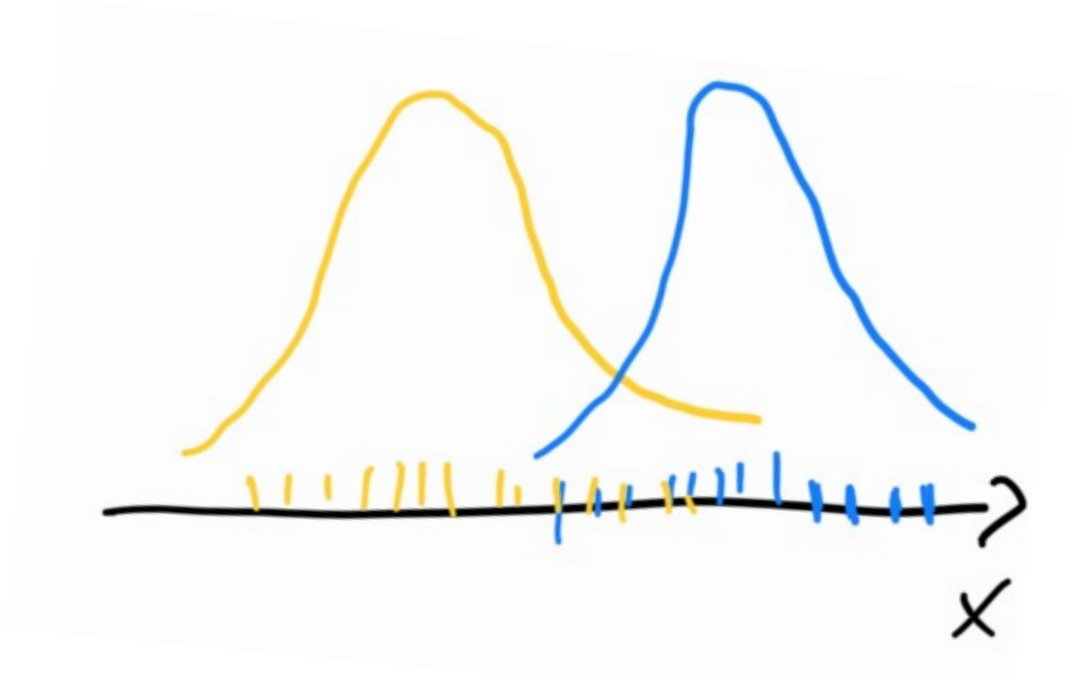
takes care of outliers!

Classification



How well can we separate the distributions?

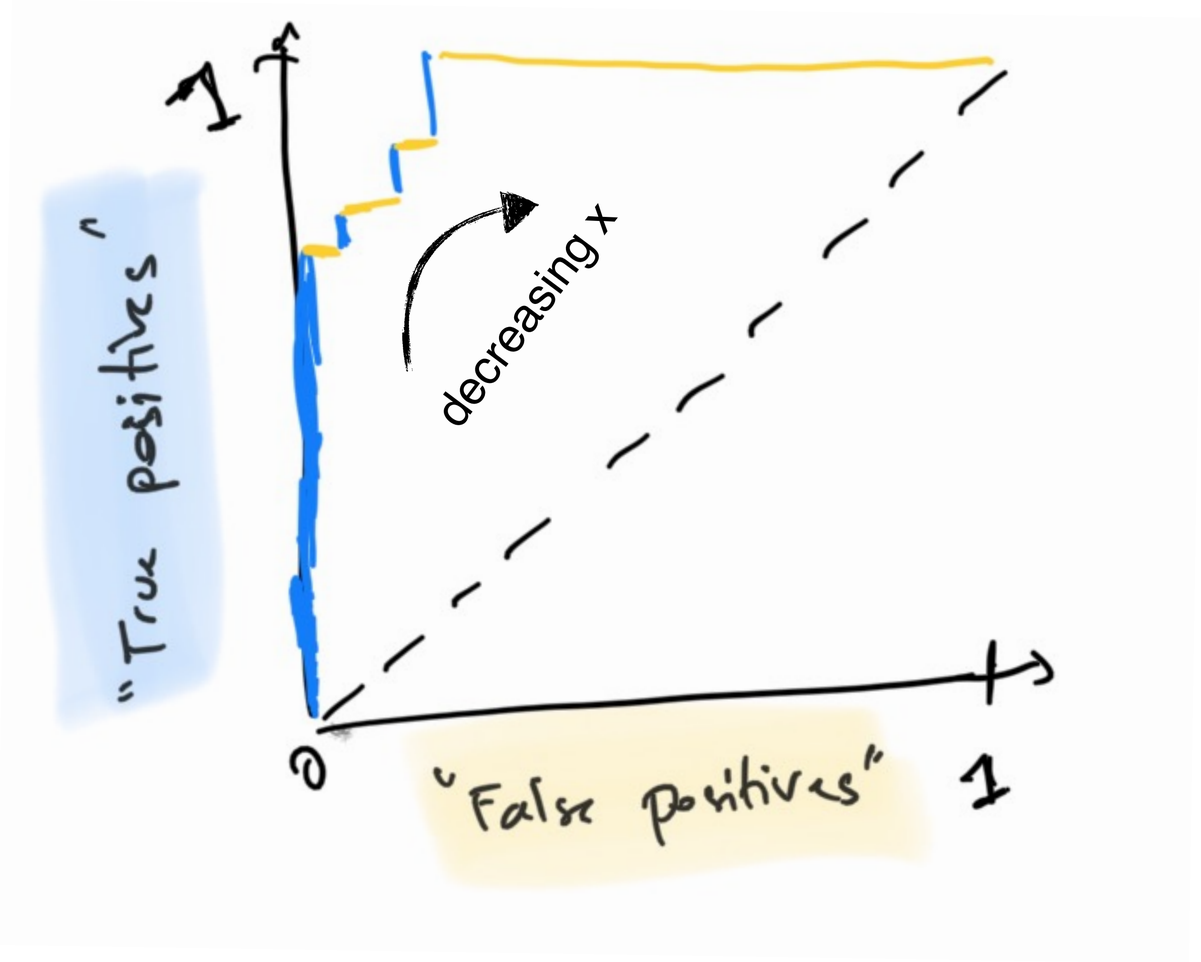
The ROC curve to assess the quality of classification



How well can we separate the distributions?

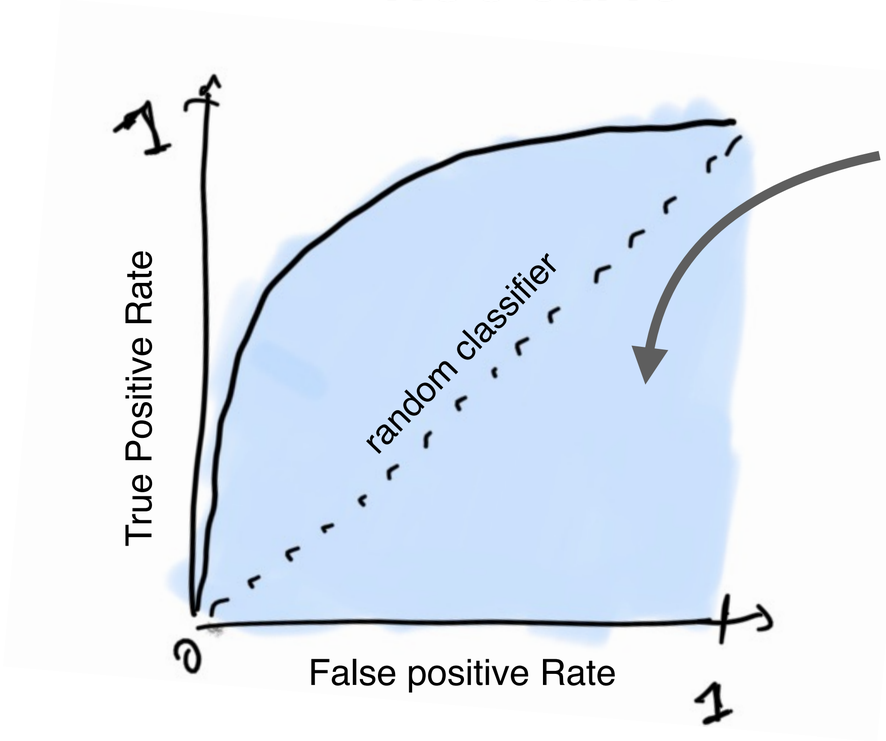
ROC curve

Receiver operating characteristic



The AUC quantifies the classification

ROC curve



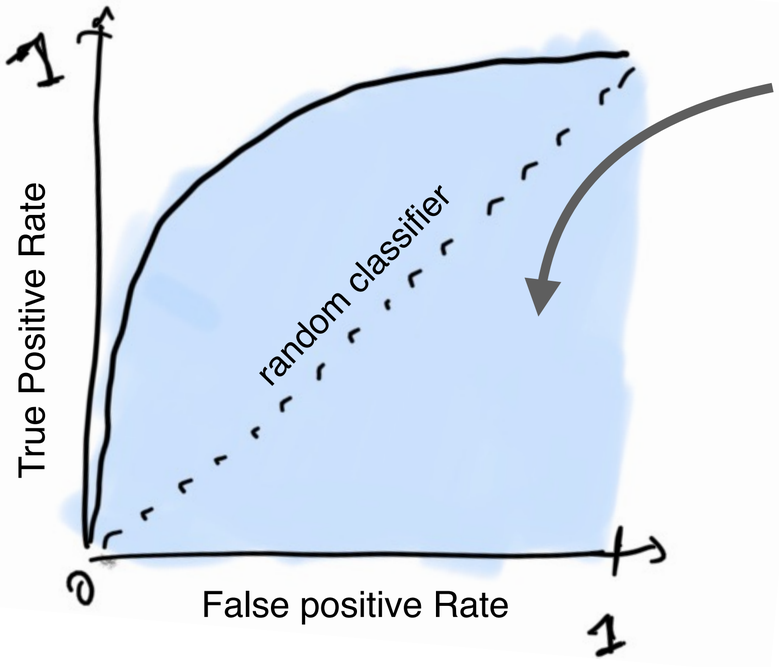
AUC

Area under the ROC curve
random = 0.5, max = 1

The AUC is the probability that a randomly chosen positive instance ranks higher than a randomly chosen negative one

Mann Whitney test gives the significance

ROC curve



AUC

Area under the ROC curve
random = 0.5, max = 1

The AUC is the probability that a randomly chosen positive instance ranks higher than a randomly chosen negative one

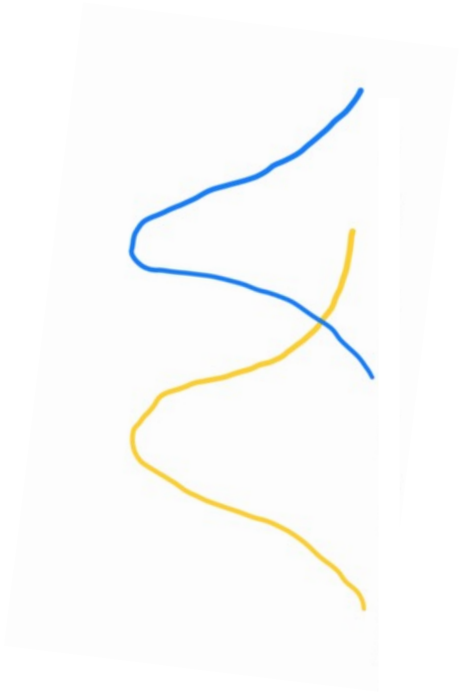
Mann-Whitney U test

(also Wilcoxon rank-sum test)

Tests the **significance of the AUC value** by testing the hypothesis that

$$P(X > Y) \neq P(Y > X)$$

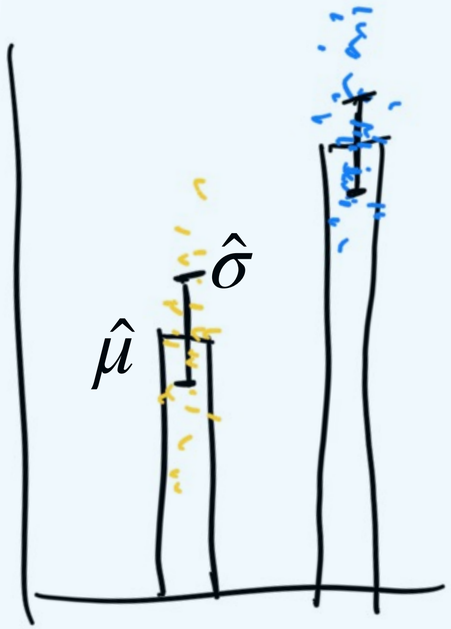
Summary



How **similar** are these data?

Student t-test

parametric

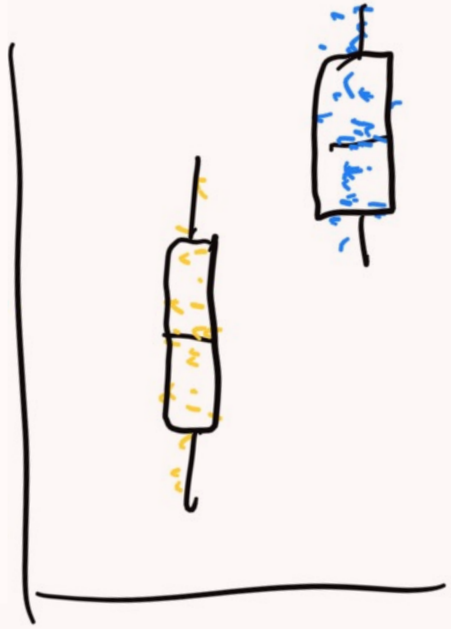


Are the **means** different?

sensitive to outliers

Wilcoxon U test

non-parametric



Are the **distributions** different?

not sensitive to outliers

Another important distribution

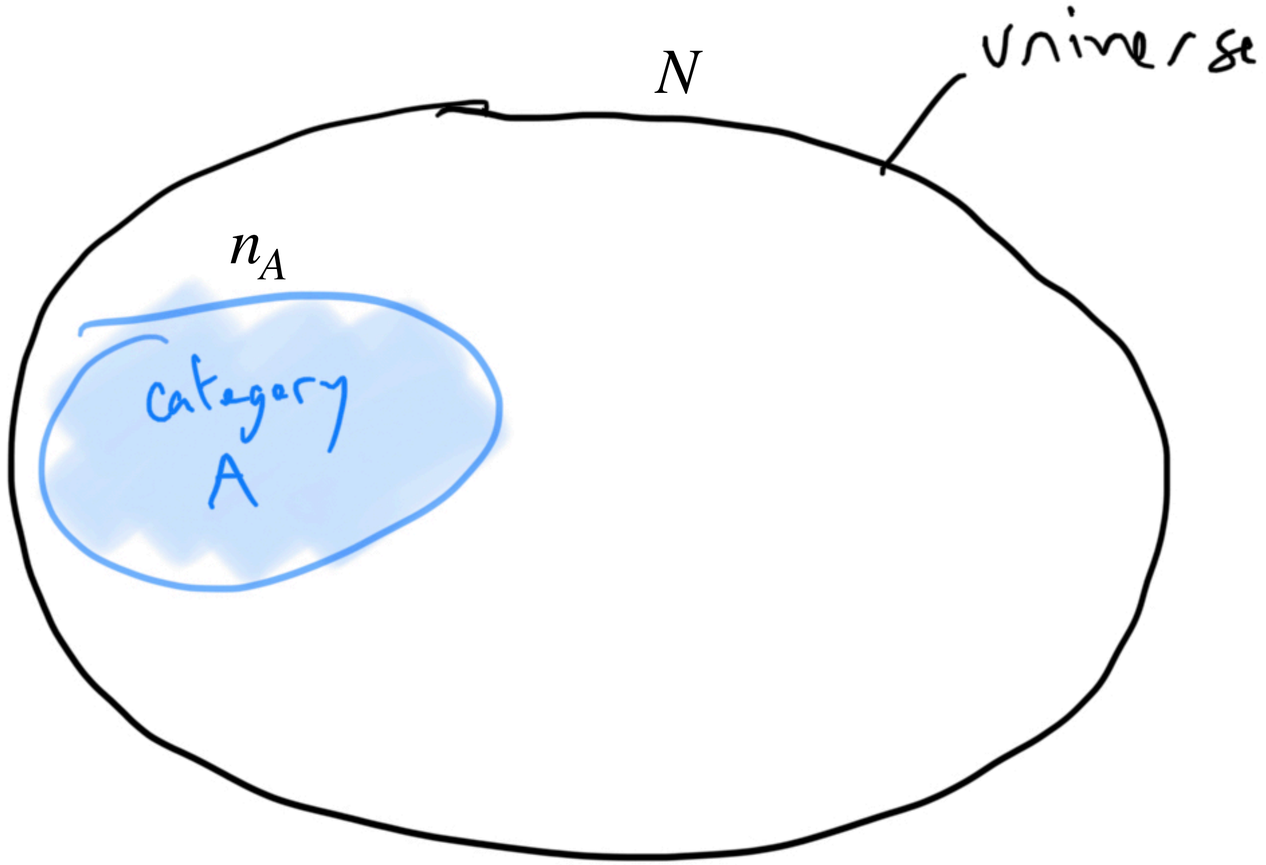
Suppose we have a set of categorical variables, for example

- articles annotated with research fields
- genes annotated with diseases or gene ontology
- individuals with their country of origin

How do we know whether some categorical variables are enriched in this set?

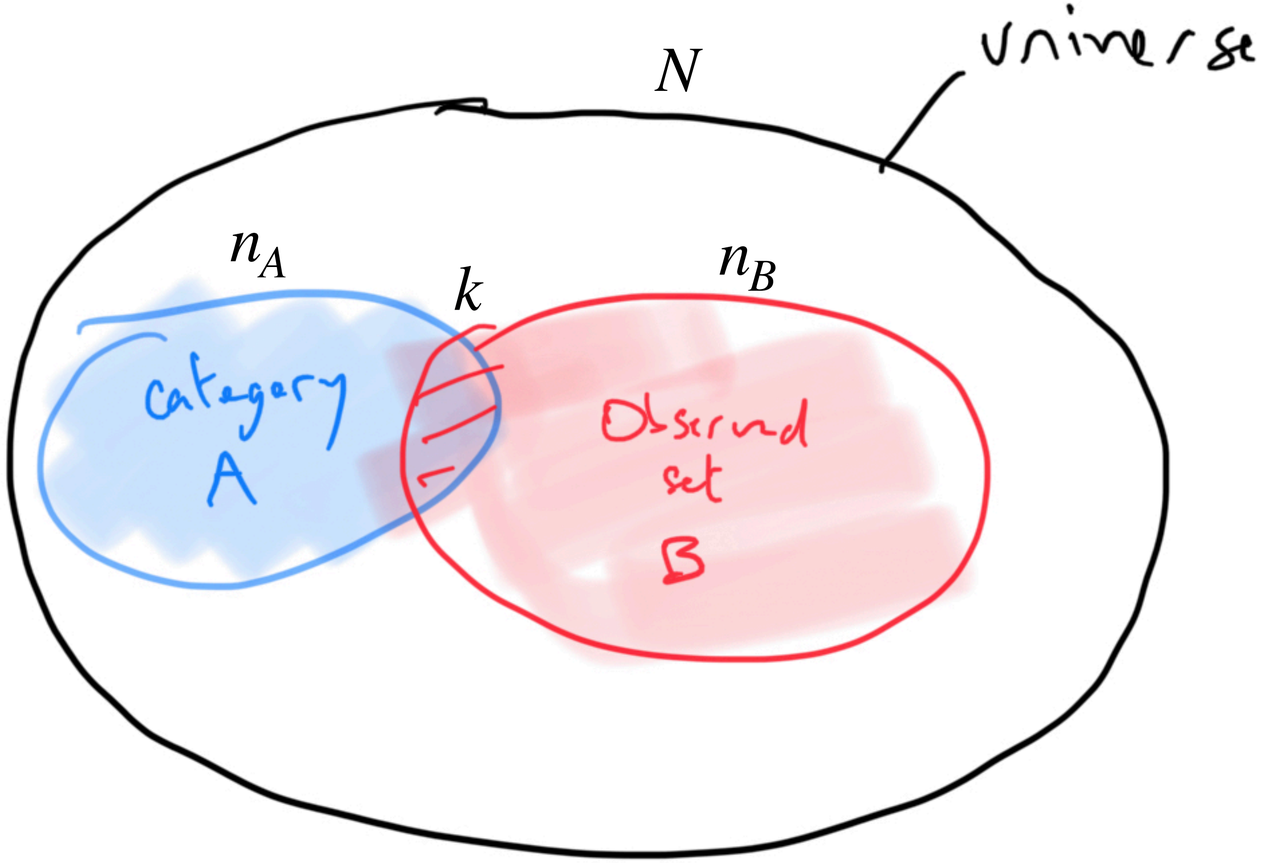
The hypergeometric distribution

We have a “universe” with all existing elements



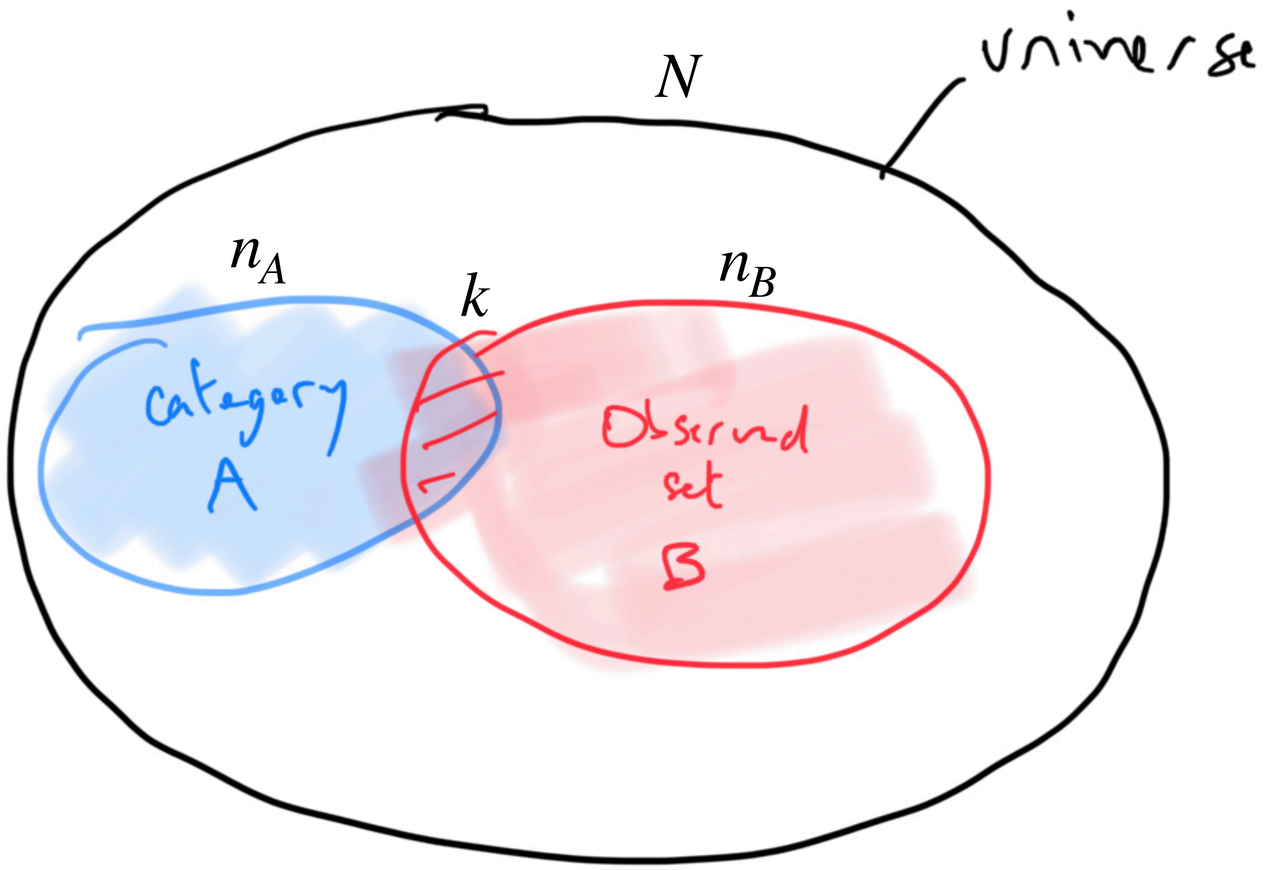
The hypergeometric distribution

We observed elements, some of which are associated with category A. How surprising is it to observe that many of them?



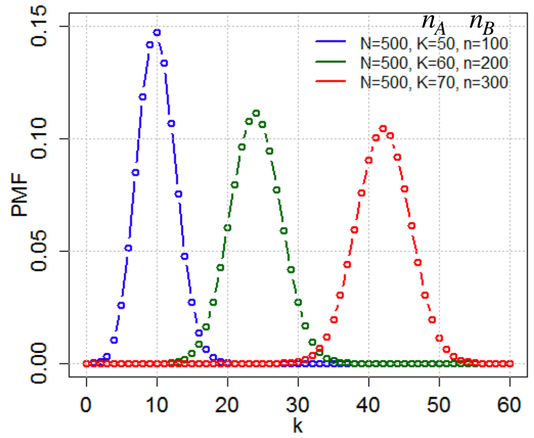
The hypergeometric distribution

We observed elements, some of which are associated with category A. How surprising is it to observe that many of them?



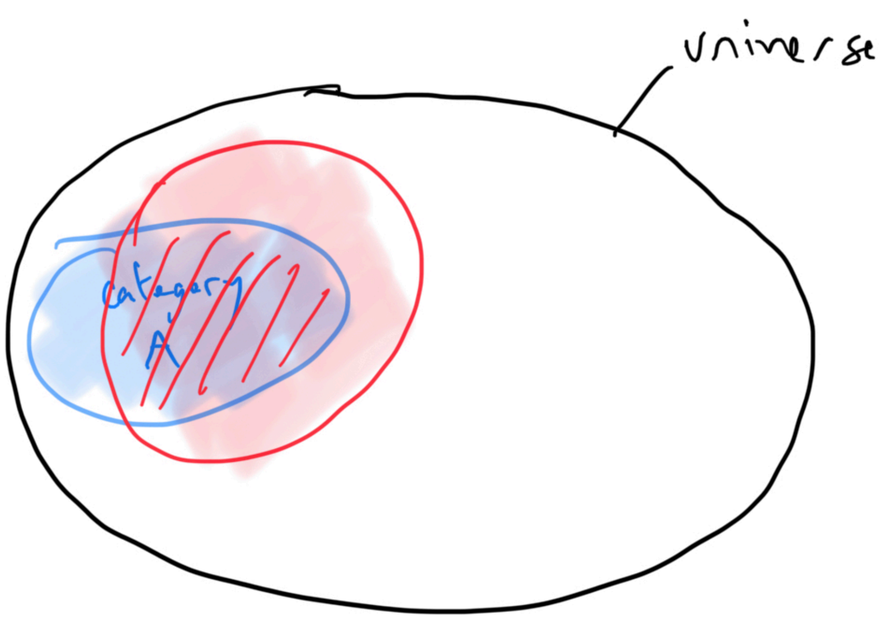
Hypergeometric distribution

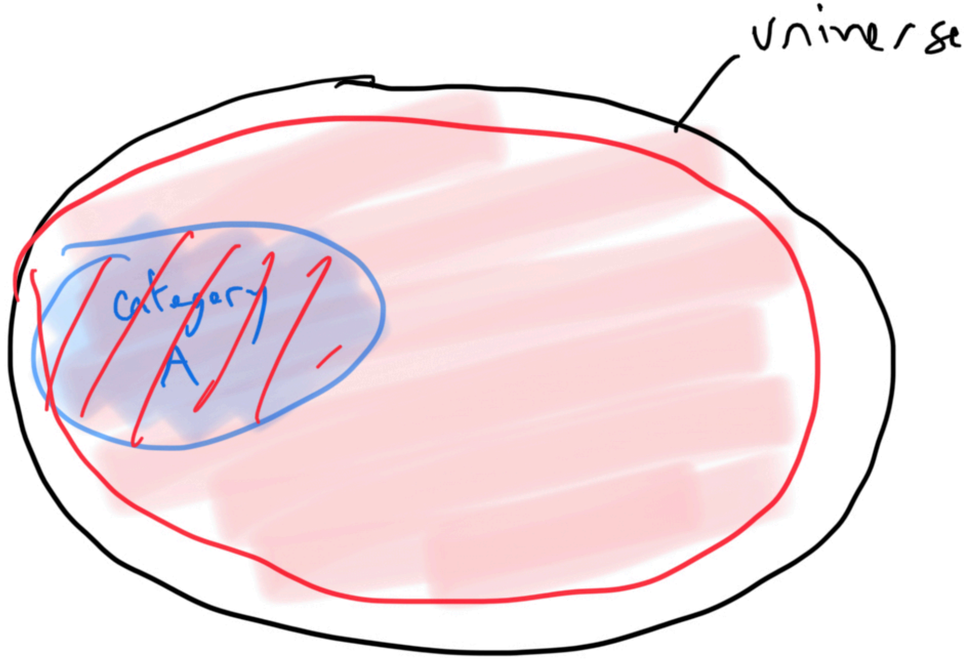
$$P(K=k)=\frac{\binom{n_A}{k}\binom{N-n_A}{n_B-k}}{\binom{N}{n_B}}$$



The hypergeometric distribution

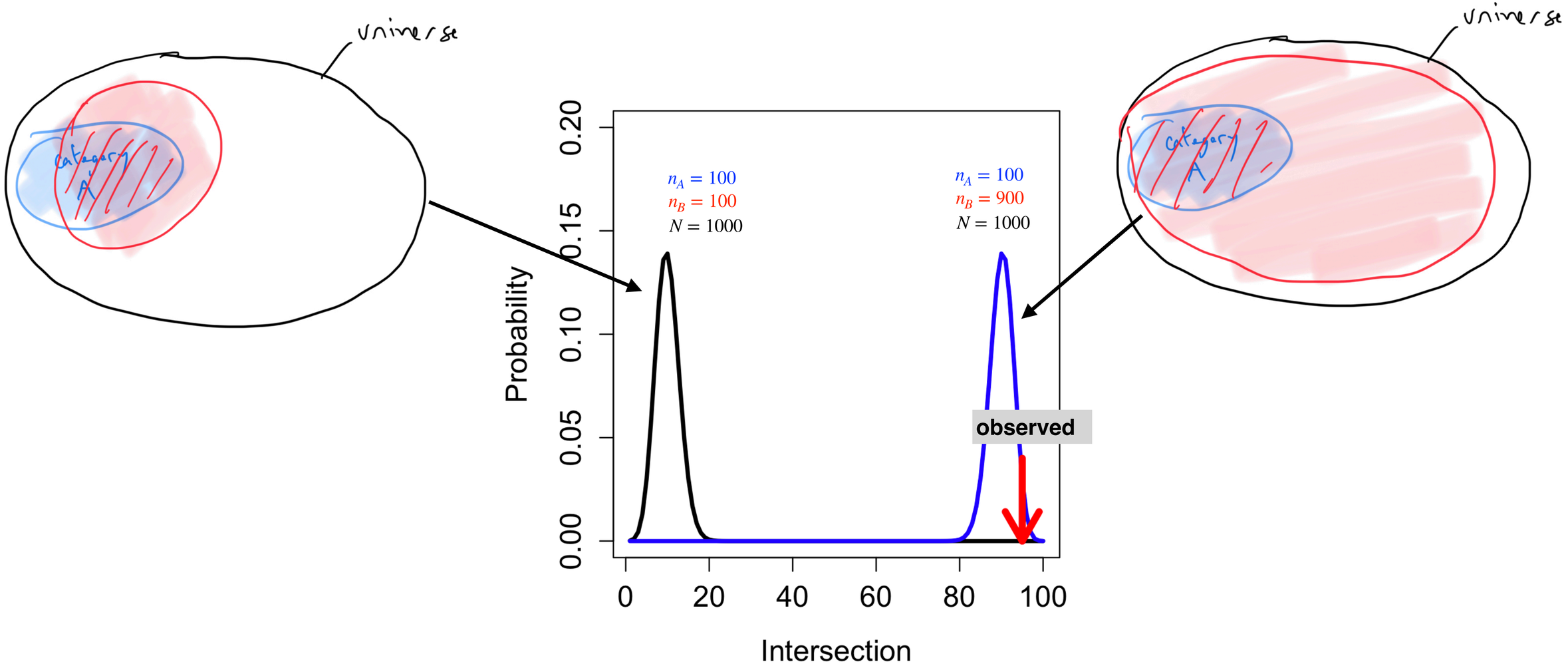
Which one is more significant than the other?





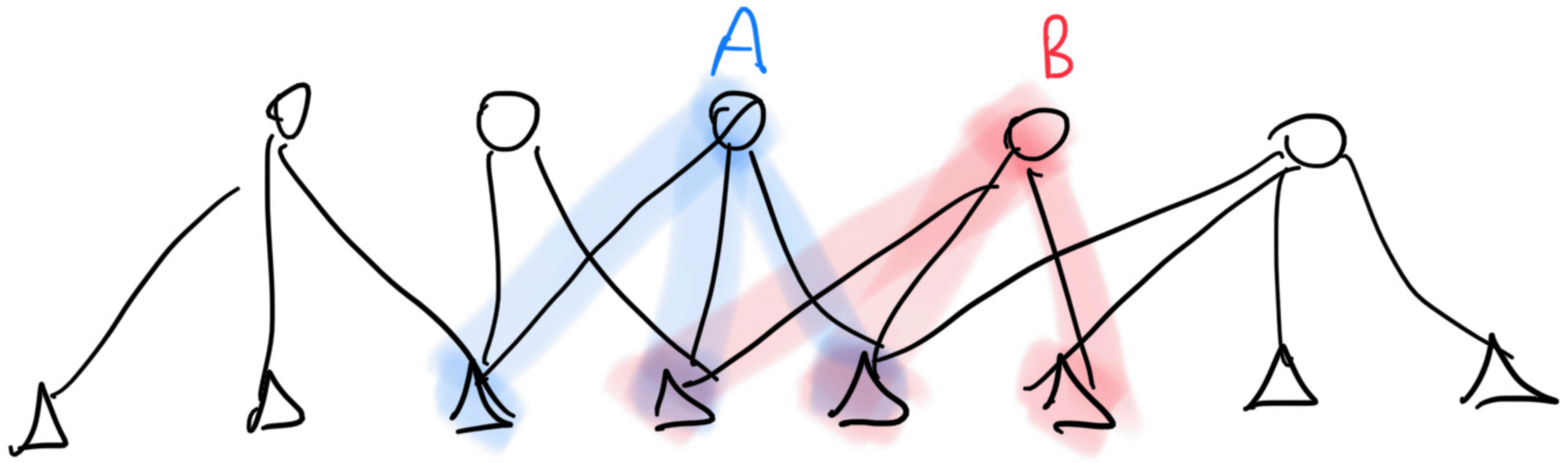
The hypergeometric distribution

Which one is more significant than the other?



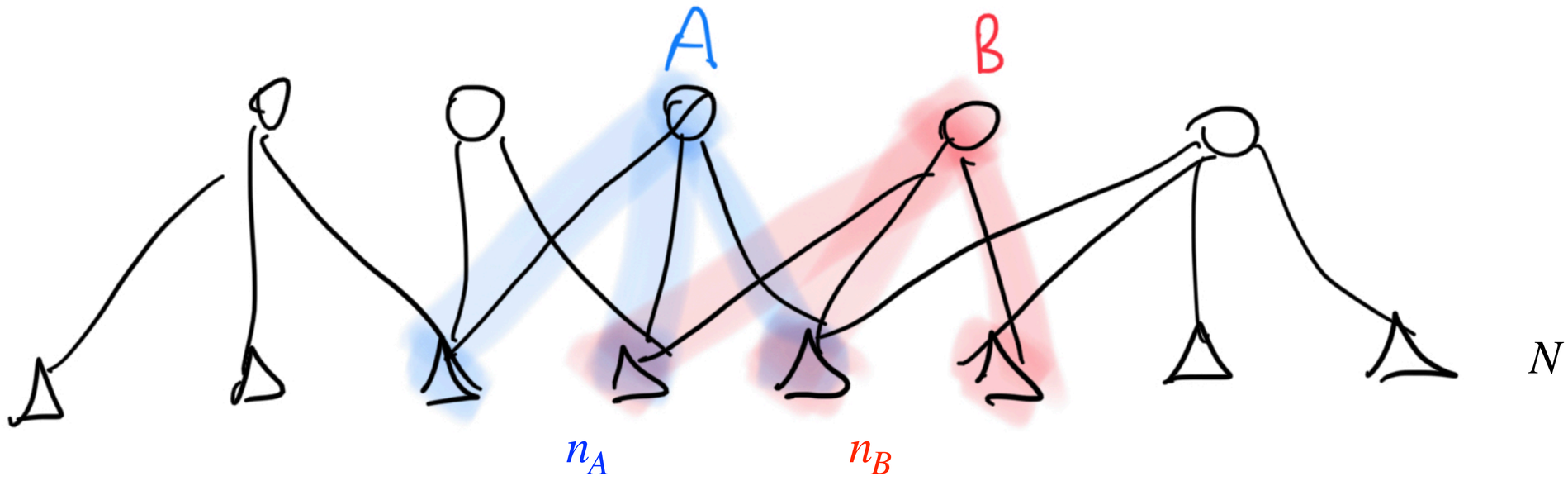
The hypergeometric distribution

A network example: how similar are two nodes in a bipartite network?

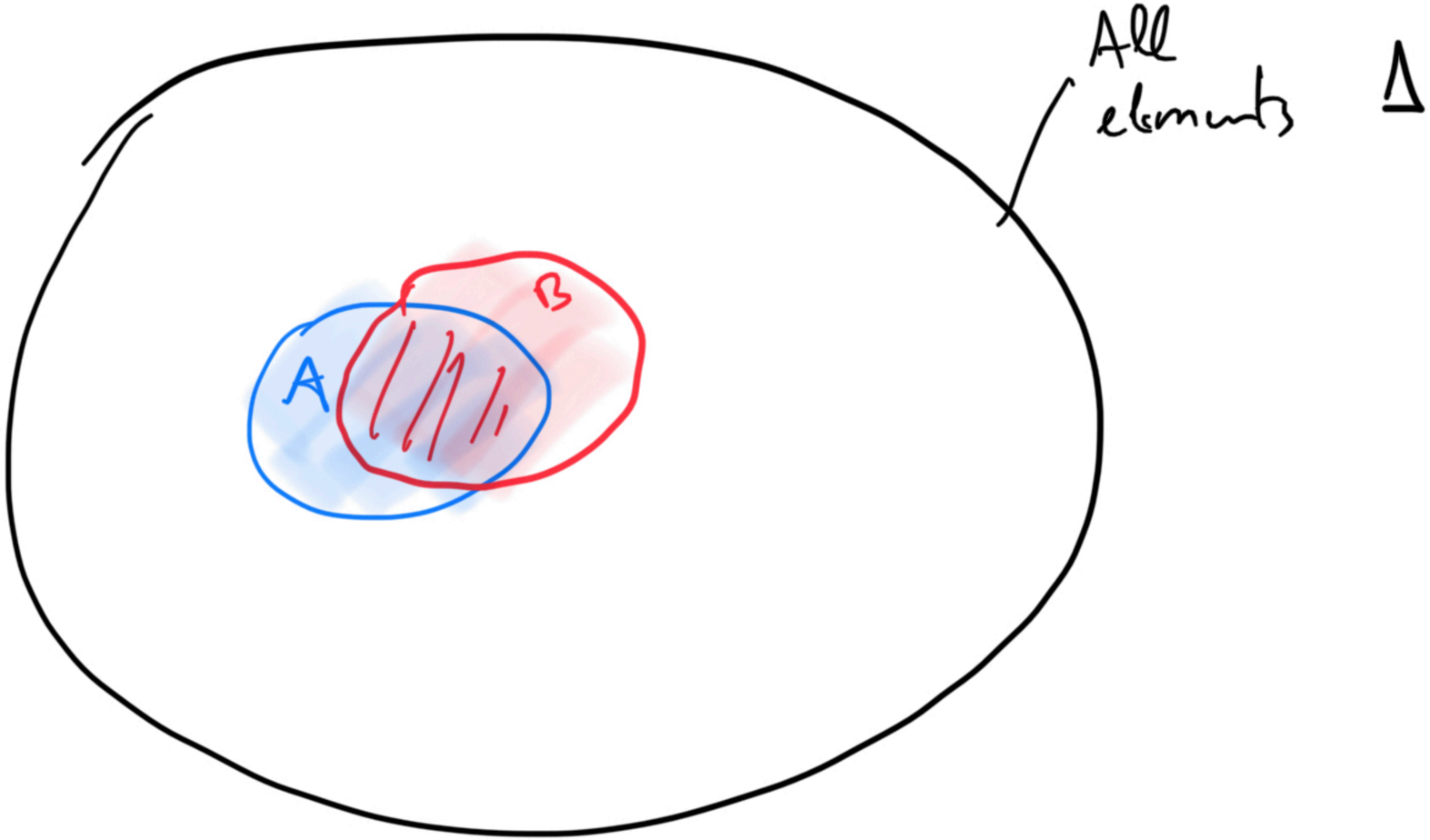


The hypergeometric distribution

A relevant example: how similar are two nodes in a bipartite network?



The hypergeometric distribution



Example: backbone extraction

Application: filter a co-occurence network

Research fields linked by how many papers mention them together. Problem: some fields co-occur with all others —> need to filter only *significant* associations.

Filtering based on hypergeometric p-value:

Edge weight: $-\log(p)$

Filter only edges with $p < 0.01$

Constructing tags co-occurrence network

We construct an undirected, weighted tag co-occurrence network where the edge weights indicates the similarity between the fields corresponding the tags. Let N are the total number of articles published, K the number of articles using field tag i and n the number of articles using field tag j with k articles that use both i, j. Then integrating Eq 5

$$p_{ij} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{5}$$

for k or more articles yields the hypergeometric p-value p_v that the two fields have at least this number of co-occurrences given the number of times they each have occurred. Note that here lower p-values indicate stronger similarity. As such, we define the edge weight W_{ij} between fields i and j as $-log_{10}(p_v)$. Edges corresponding to $p_v > 0.01$ are filtered out.

Singh et al, PLoS ONE 2022