An introduction to computational analysis methods

Melanie Ganz, PhD
Cyril Pernet, PhD

Intended Learning Outcomes

- Being able to discuss the notions of statistical usefulness, validity and replicability
- Being able to explain differences among resampling techniques

Statistical thinking

Statistical thinking

a systematic way of thinking about how we describe the world and use data make decisions and predictions

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.” - H.G. Wells

What can we use statistics for?

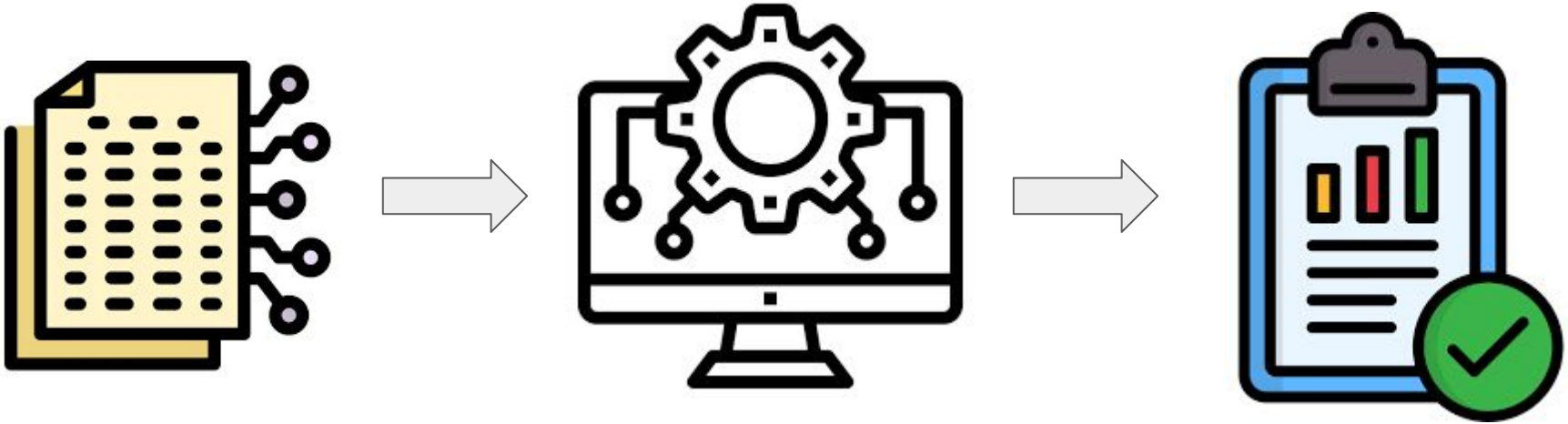
Describe: The world is complex and we often need to describe it in a simplified way that we can understand.

Decide: We often need to make decisions based on data, usually in the face of uncertainty.

Predict: We often wish to make predictions about new situations based on our knowledge of previous situations.

A typical setup for data analysis

A typical data analysis cycle

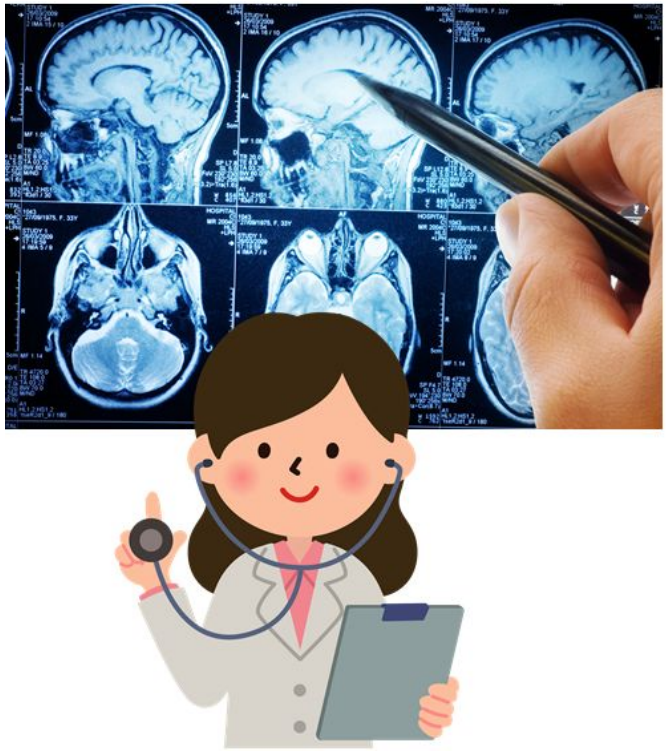


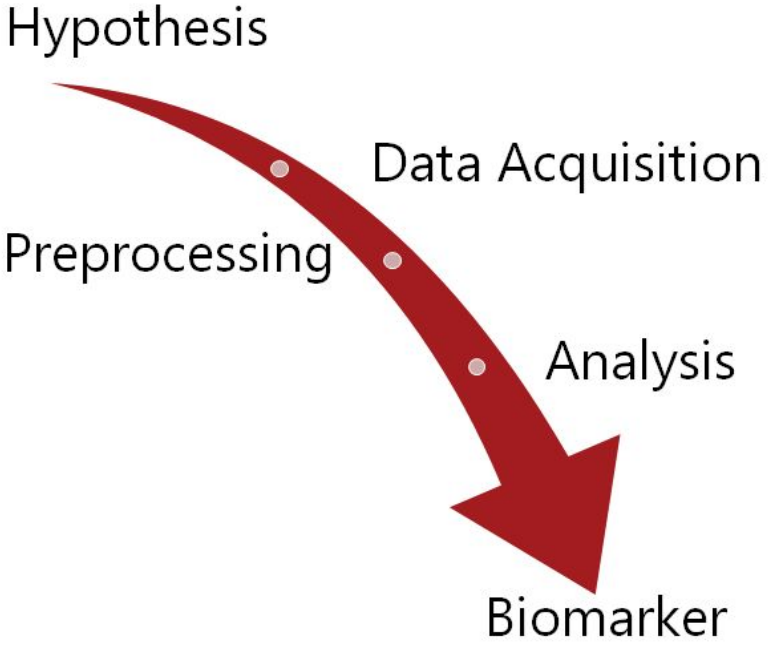
Data

Algorithm

Result

A neuroimaging example





Let’s focus on the results

Please head over to Padlet:

https://ucph.padlet.org/melanieganzbenjaminsen/validity2025

How do we know if the result of an analysis is valid?
How do we know if a result of an analysis is useful?
How do we know if the results of an analysis is reliable?

How do we know if the result of an analysis is reliable?

Often we talk about **a measure being reliable**.

Then we refer to the consistency of our measurements. One common form of reliability, known as “test-retest reliability”, measures how well the measurements agree if the same measurement is performed twice.

But how do we transfer this to a result of analysis?

(Definition taken from https://statsthinking21.github.io/statsthinking21-core-site/working-with-data.html)

How do we know if the result of an analysis is reliable?

To me a result is reliable if it is

Reproducible: A result is reproducible when the same analysis steps performed on the same dataset consistently produces the same answer.

Replicable: A result is replicable when the same analysis performed on different datasets produces qualitatively similar answers.

(The definitions here are following The Turing Way, https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html)

How do we know if the result of an analysis is valid?

Validity is often used wrt measurements. We want our measurements to be valid — that is, we want to make sure that we are actually measuring the construct that we think we are measuring.

But how do we transfer this to a result of analysis?

(Definition again following https://statsthinking21.github.io/statsthinking21-core-site/working-with-data.html)

How do we know if the result of an analysis is valid?

To me a result is valid if it is

Robust: A result is robust when the same dataset is subjected to different analysis workflows to answer the same research question and a qualitatively similar or identical answer is produced. Robust results show that the work is not dependent on the specificities of the workflow chosen to perform the analysis.

(The definitions here are following The Turing Way, but are slightly modified https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html)

How do we know if the result of an analysis is useful?

To me a result is useful if it is

Generalisable: Combining replicable and robust findings allow us to form generalisable results.

(The definitions here are following The Turing Way, https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html)

Now how can we ensure this?

Reproducible - save your code and data, so you can rerun it

Replicable - share your code with others so they can replicate it on their data

Robust - test the statistical robustness on your own data -> cross-validation, boot strapping, permutation testing

Generalisable - test your result out in a new data set

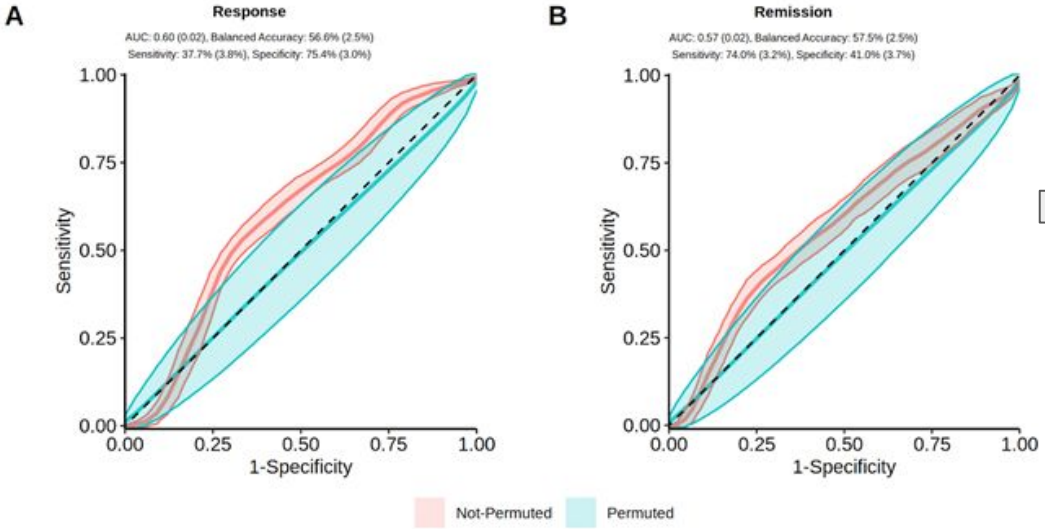
Back to a neuroimaging example

NeuroImage: Clinical
Volume 36, 2022, 103224

Generalizability of treatment outcome prediction in major depressive disorder using structural MRI: A NeuroPharm study

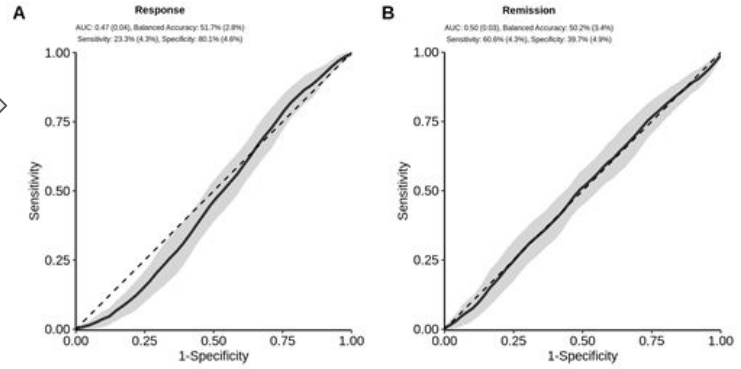
Vincent Beliveau [a b], Ella Hedeboe [a], Patrick M. Fisher [a], Vibeke H. Dam [a], Martin B. Jørgensen [a c d], Vibe G. Frokjaer [a c d], Gitte M. Knudsen [a c], Melanie Ganz [a e]

Neuropharm study



In sample

EMBARC study



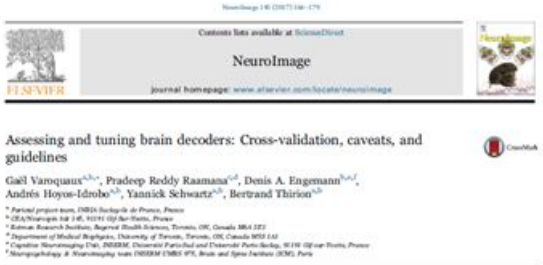
Out of sample

Cross-validation

Limited amounts of data, so we cannot afford an independent hold out dataset

Cross-validation

Standard tool in ML



https://www.sciencedirect.com/science/article/pii/S105381191630595X

Figure 1: **Cross-validation**: the data is split multiple times into a train set, used to train the model, and a test set, used to compute predictive power.

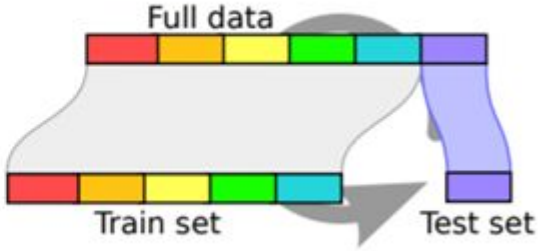
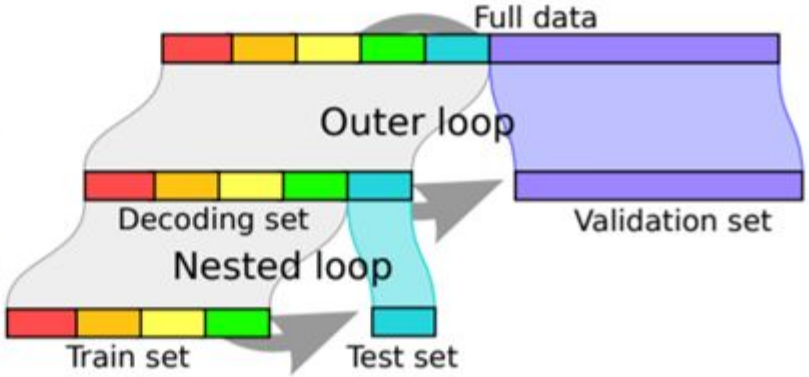


Figure 2: **Nested cross-validation**: two cross-validation loops are run one inside the other.



Cross-validation nuggets

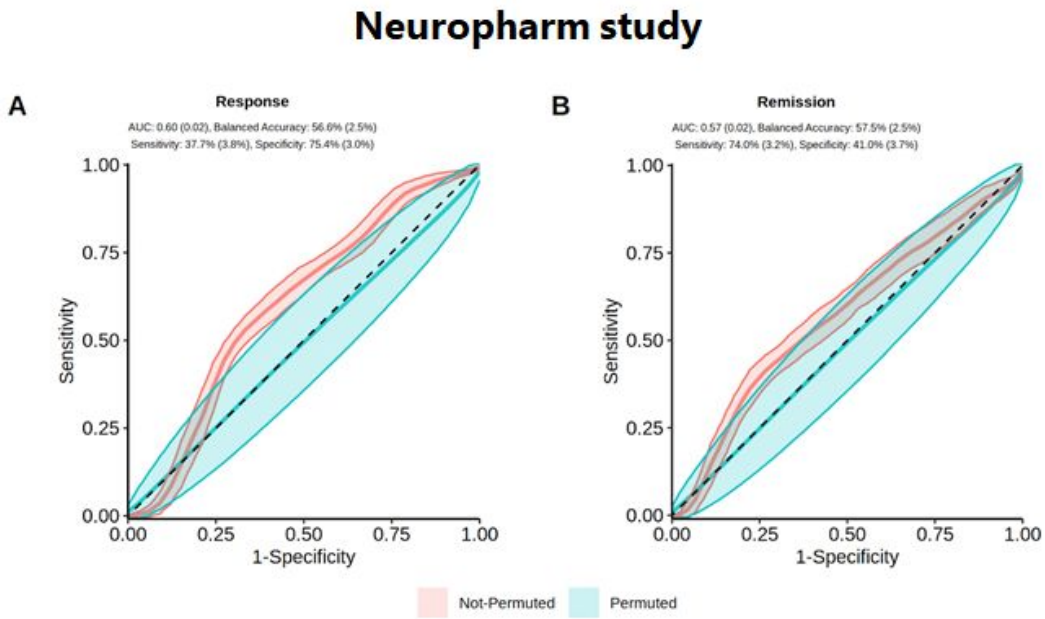
- Cross-validation with 5 or 10-folds is preferable to leave-one-out
- Randomized cross-validation will yield errors bars around your estimate
- Cross-validation performance will always still be positively biased, so overestimate generalization performance (!)

(From https://www.sciencedirect.com/science/article/pii/S105381191630595X)

Back to a neuroimaging example

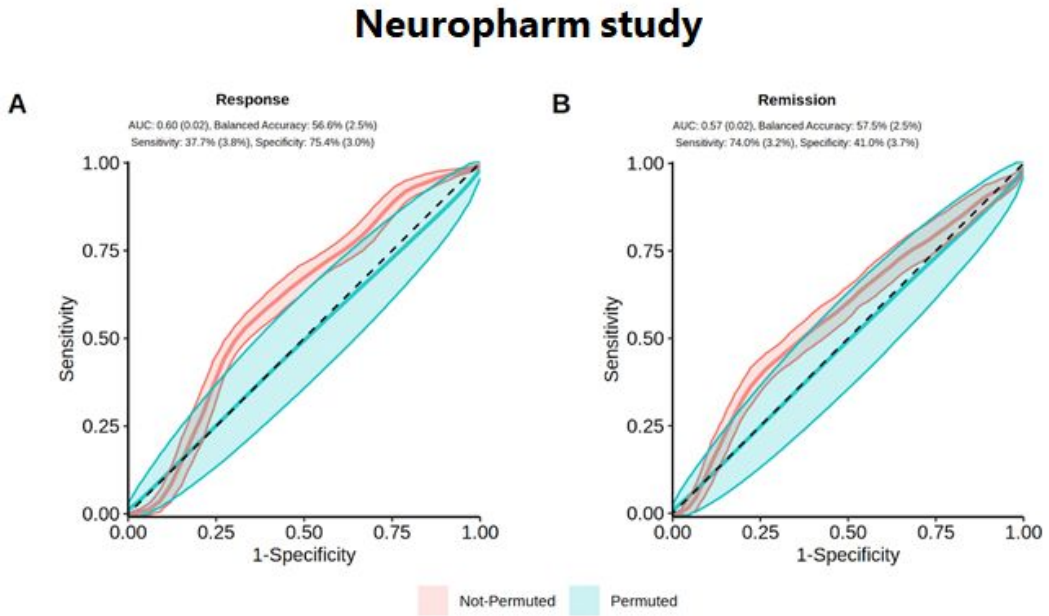
stratified nested cross-validation with a repeated cross-validation with 5-fold and 25 repeats as outer loop and a 5-fold cross-validation as nested loop for hyperparameters optimization

-> just yields the red curves



In sample

But how do we know our red curve is better than random?



In sample

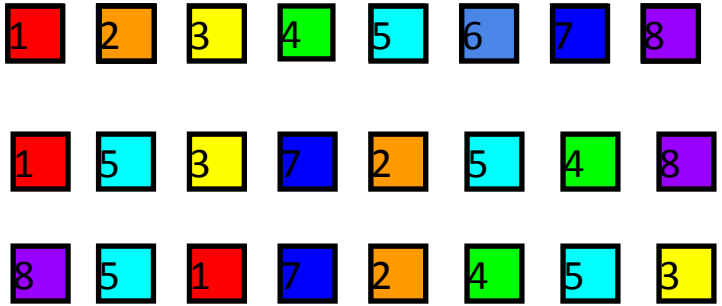
Bootstrap

Lots of data, so we have a hold out test set

What is a bootstrap?

resampling of observations with replacement

e.g. here of the numbers 1-8



(Efron & Tibshirani, “An introduction to the bootstrap”)

Basic behind the bootstrap

We create surrogate samples by resampling our data with replacement

We compute our statistic on each of those “bootstrap” samples

We then use the distribution of the statistic across those samples as a surrogate for its sampling distribution

-> We can estimate the variance in performance on a hold out set by bootstrapping it

Confidence intervals with a bootstrap

From A. Field, An Adventure In Statistics

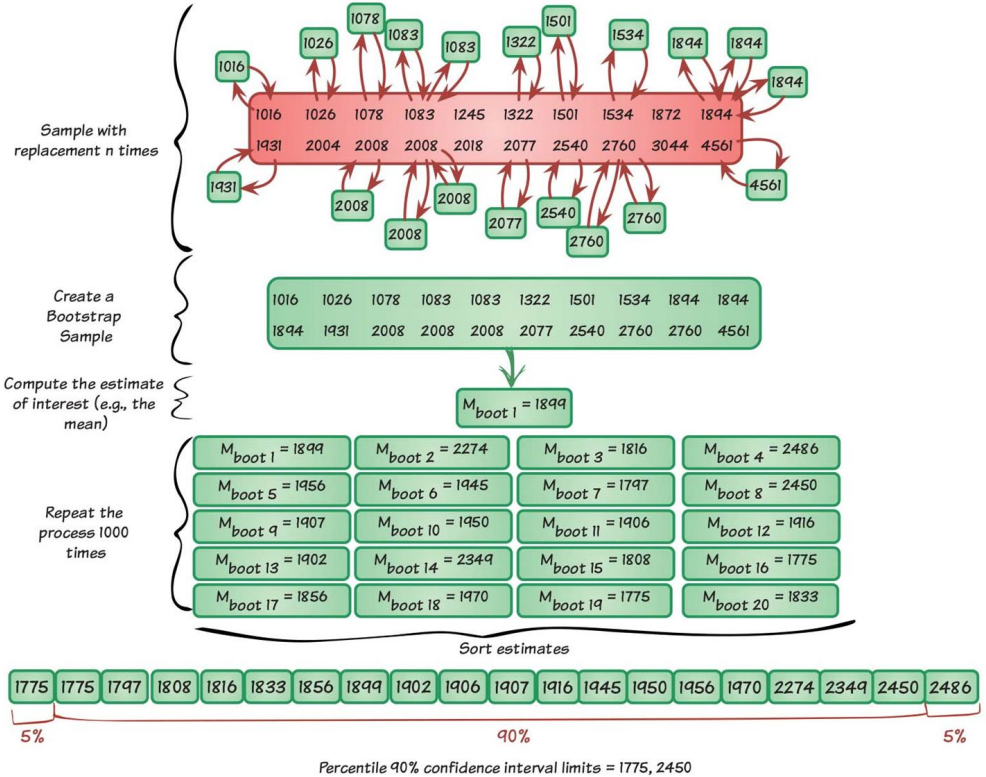


Figure 9.10 Illustration of the percentile bootstrap

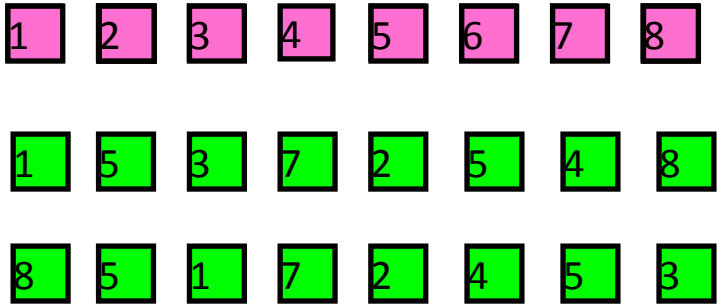
Permutations

Now onto estimations of the null distribution

What is a permutation?

Shuffling of observations

e.g. here of the numbers 1-8



Or shuffling of labels when e.g. performing classification

(Phillip Good, “Permutation, Parametric and Bootstrap Tests of Hypotheses”)

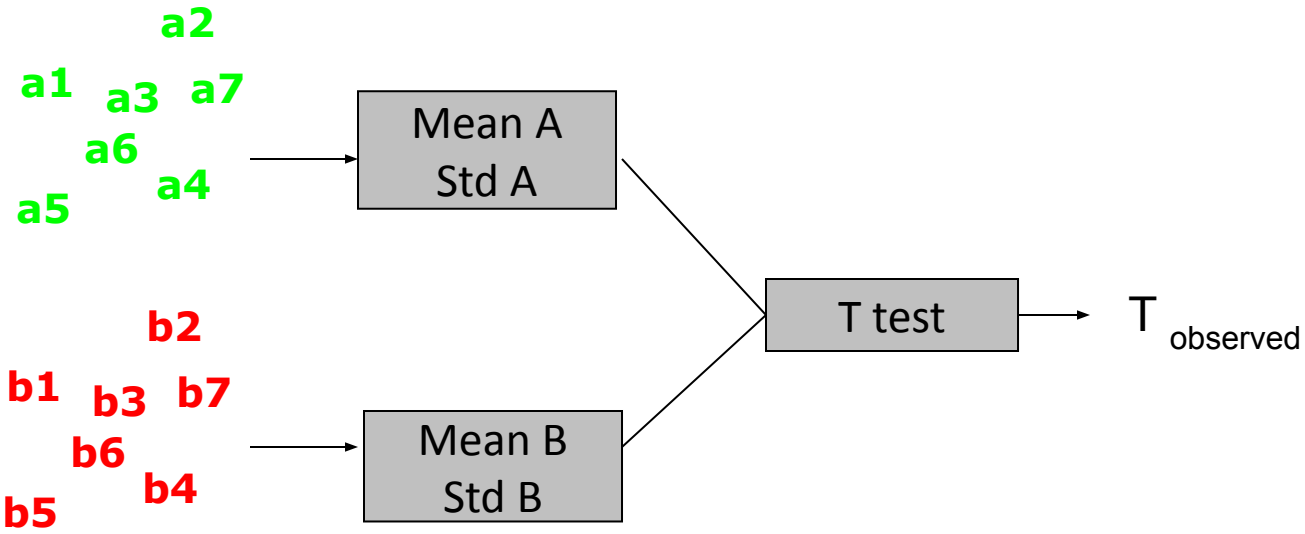
How can this be used in statistics?

Permuting numbers or labels breaks the associations we want to test for

- whether it is the difference in a statistical test between two samples
- or the null performance of a classifier

The outcome of our analysis on the permuted samples yields us a “null distribution” - it gives us the outcome and a variance of the outcome that we then can compare to our “real” analysis

Application to a 2 sample t-test: permutation



Application to a 2 sample t-test: permutation

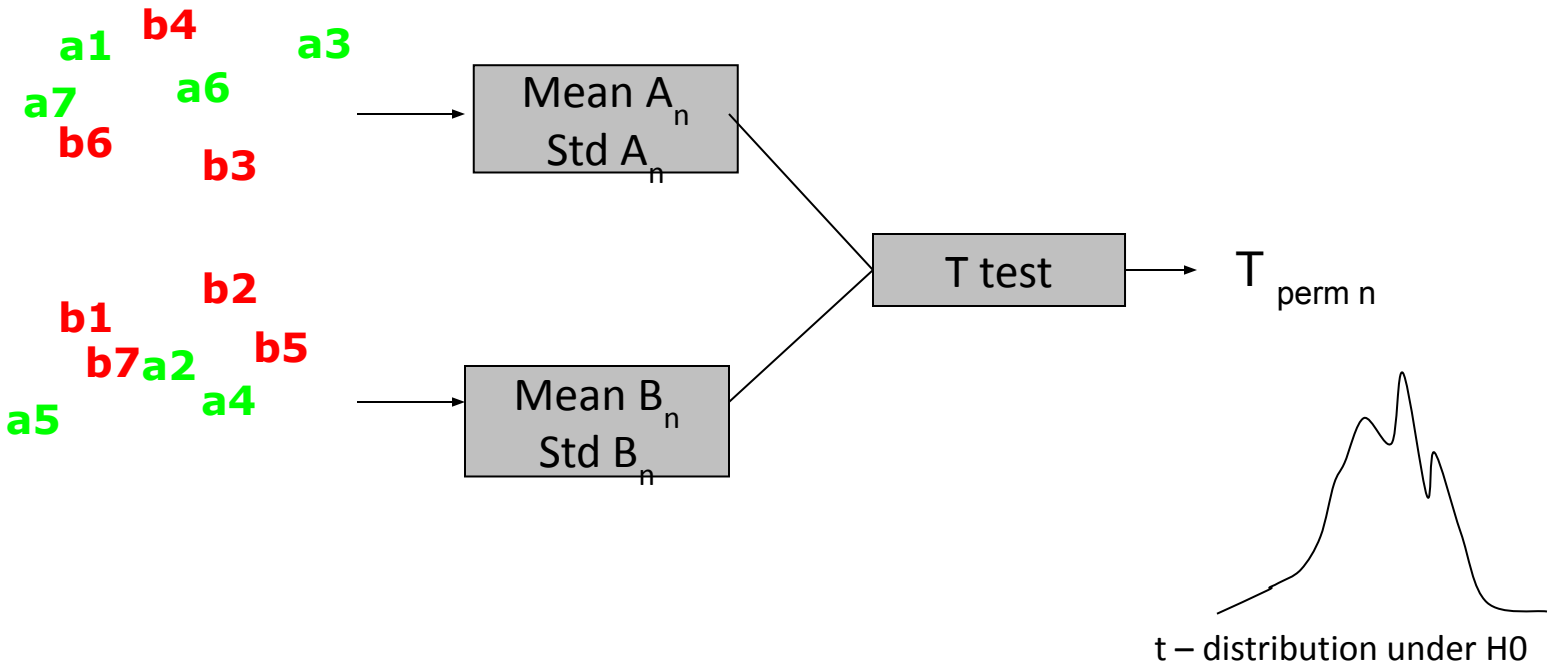
(1) Shuffle observations- If there is no effect, grouping is irrelevant

(2) Compute estimate e.g. mean difference

(3) Repeat (1) & (2) b times

(4) Get p-value

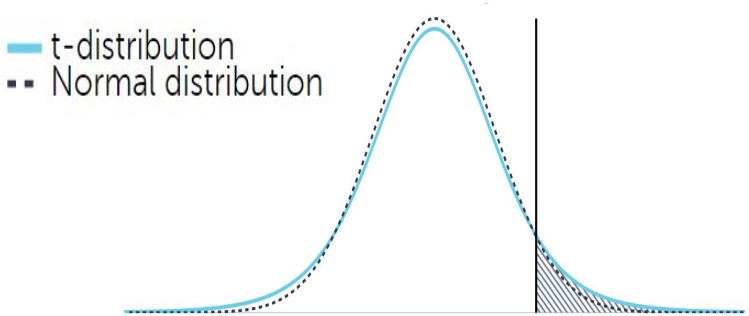
Application to a 2 sample t-test: permutation



Application to a 2 sample t-test: permutation

<u>What is the p value of the sample</u>

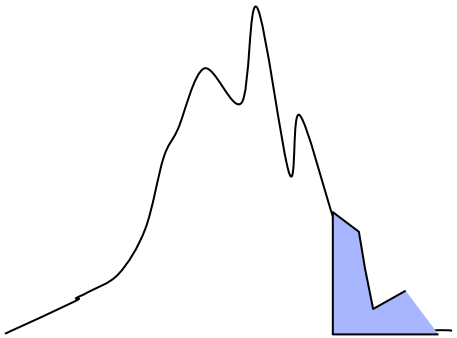
p(Obs≥t|H0) □ cumulative probability



area under the curve for T obs = p value
Significance = point of T critical

<u>What is the p value of the sample</u>

p(Obs≥t|H0) □ cumulative probability

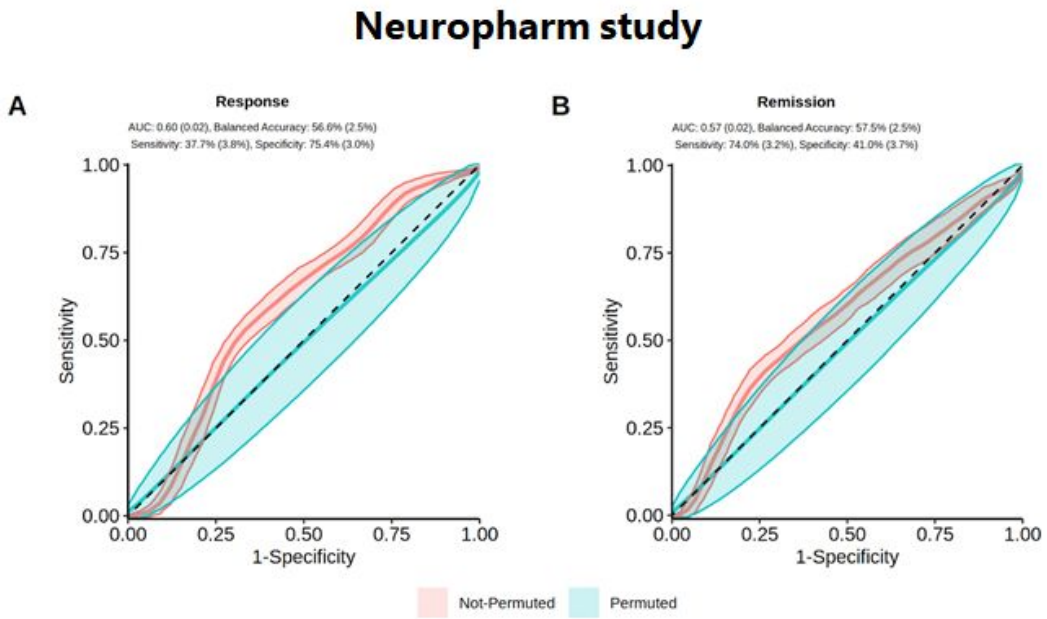


area under the curve for T obs = p value
Significance = percentile of the empirical t distribution
□ Theoretical T assumes data normality, we don't

Back to a neuroimaging example

The statistical significance of the AUCs were empirically derived using a null distribution estimated from 1,000 permutations for the elastic net and random forest and 100 permutations for the boosted trees and SVM.

-> yields the cyan curves



In sample

Generalizability

Why is generalizability important?

- If we really think we have found a “new” relation, we should test it on unseen data

- The methods we have presented test the performance on our data as well as we can, but this still does not guarantee the same conclusion on new data!

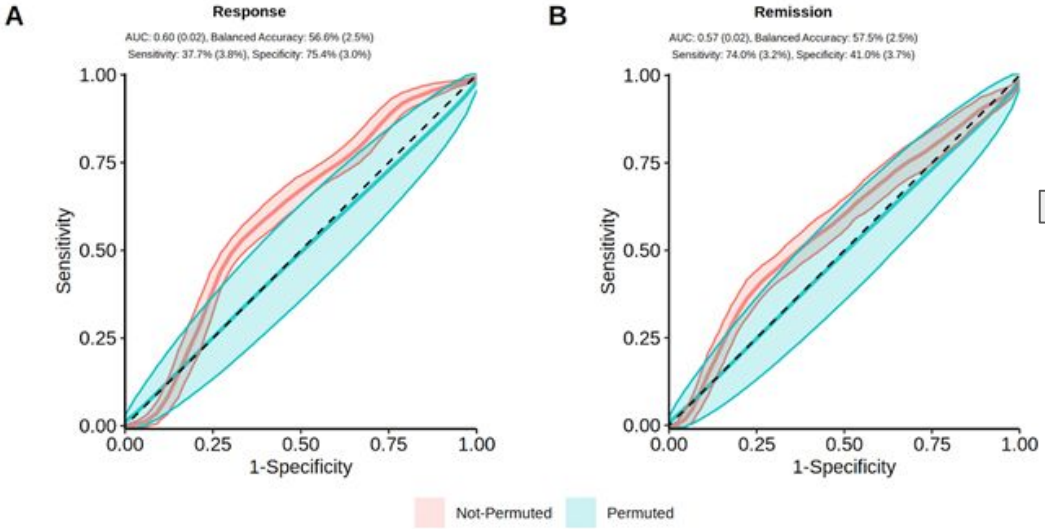
Back to a neuroimaging example

NeuroImage: Clinical
Volume 36, 2022, 103224

Generalizability of treatment outcome prediction in major depressive disorder using structural MRI: A NeuroPharm study

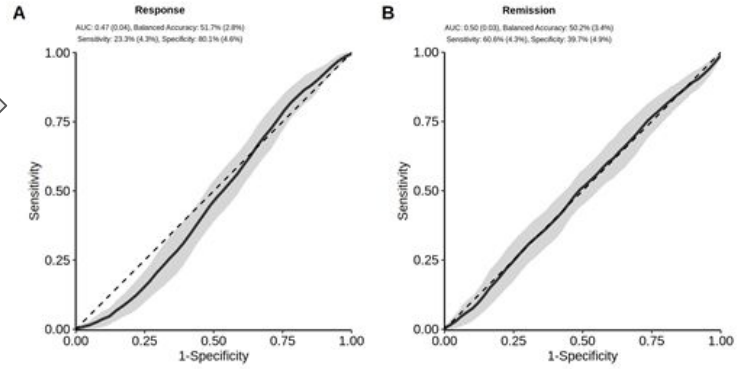
Vincent Beliveau [a b], Ella Hedeboe [a], Patrick M. Fisher [a], Vibeke H. Dam [a], Martin B. Jørgensen [a c d], Vibe G. Frokjaer [a c d], Gitte M. Knudsen [a c], Melanie Ganz [a e]

Neuropharm study



In sample

EMBARC study



Out of sample

If you want to further improve your statistical practices…

Improving your statistical inferences

https://www.coursera.org/learn/statistical-inferences

About this course: This course aims to help you to draw better statistical inferences from empirical research. First, we will discuss how to correctly interpret p-values, effect sizes, confidence intervals, Bayes Factors, and likelihood ratios, and how these statistics answer different questions you might be interested in. Then, you will learn how to design experiments where the false positive rate is controlled, and how to decide upon the sample size for your study, for example in order to achieve high statistical power. Subsequently, you will learn how to interpret evidence in the scientific literature given widespread publication bias, for example by learning about p-curve analysis. Finally, we will talk about how to do philosophy of science, theory construction, and cumulative science, including how to perform replication studies, why and how to pre-register your experiment, and how to share your results following Open Science principles.

Taught by: Daniel Lakens, Associate Professor

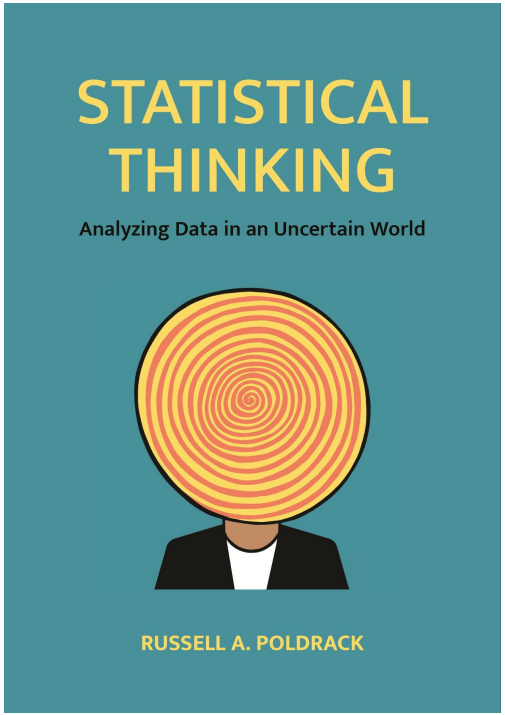
Department of Human-Technology Interaction

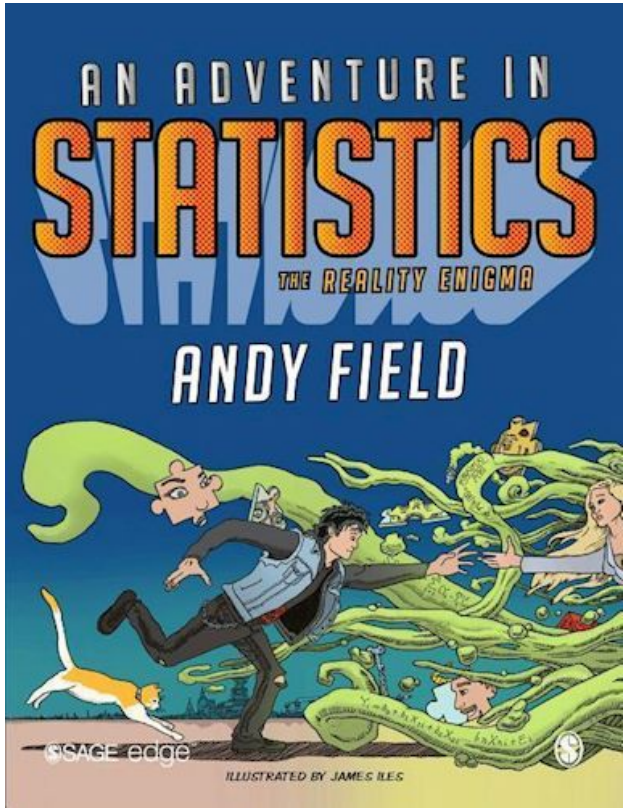
Intended Learning Outcomes

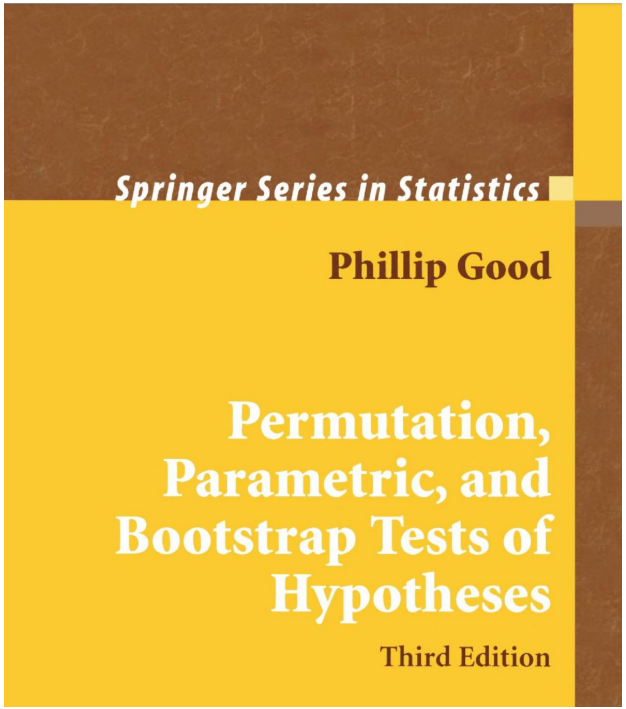
- Being able to discuss the notions of statistical usefulness, validity and replicability
- Being able to explain differences among resampling techniques

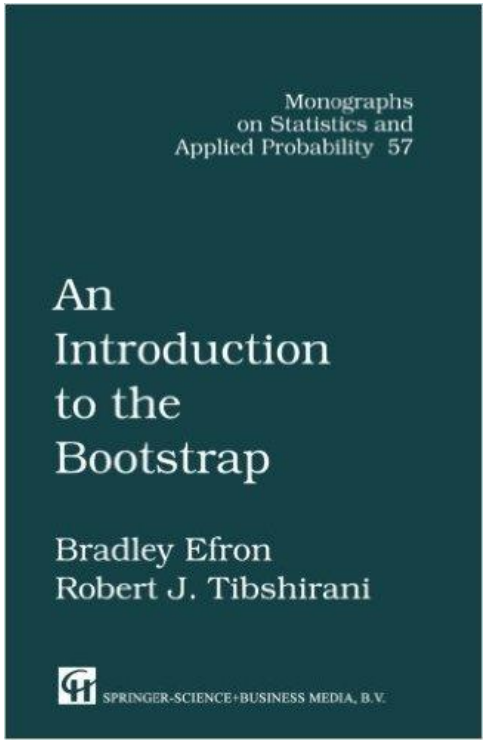
References

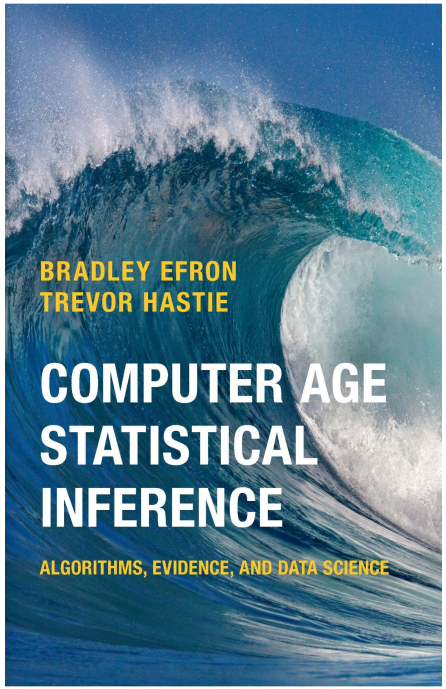
https://github.com/statsthinking21/statsthinking21-core

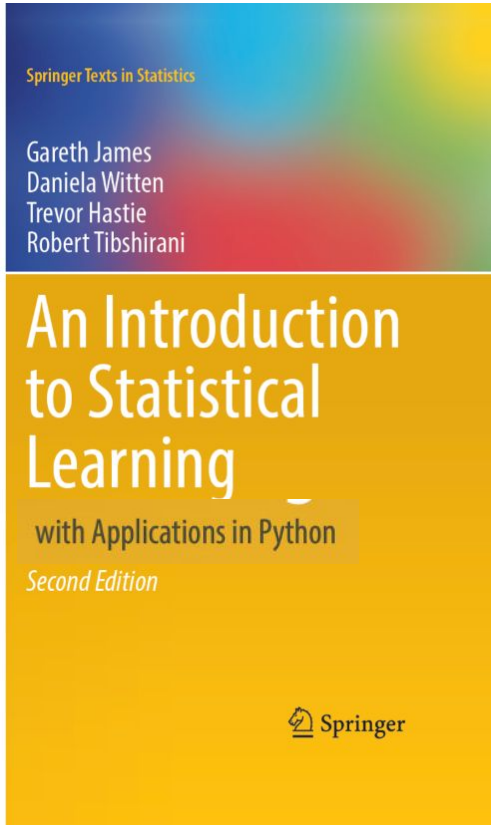












https://www.statlearning.com/