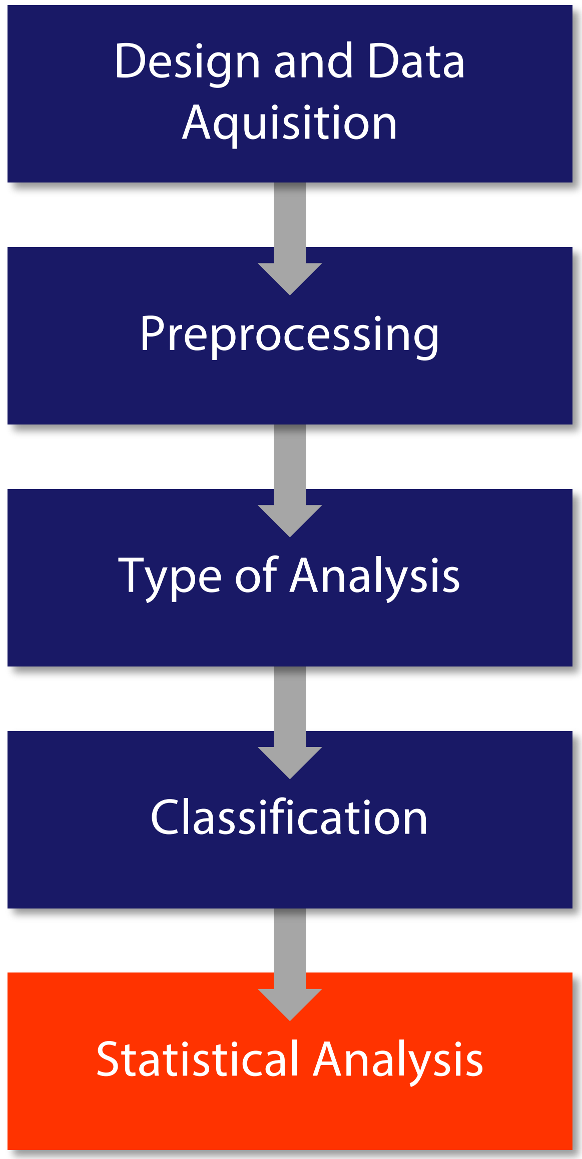
Session V: Statistics for MVPA

Martin Hebart

Laboratory of Brain and Cognition

NIMH

MVPA Workflow



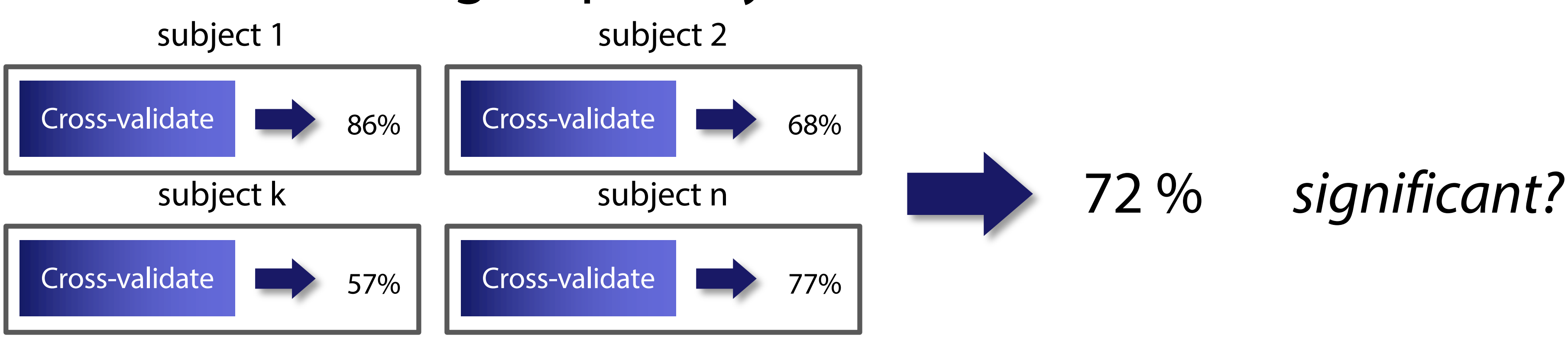
Two Levels of Statistical Analysis

"Decoding level"

Cross-validate → 86% accuracy *significant?*

- within-subject classification (condition A vs. condition B)
- between-subject classification (group A vs. group B)

"Second level" (group analysis)



Overview

"Decoding level" statistics

- Without cross-validation: Binomial test
- With cross-validation: Permutation test

Group-level statistics

- Parametric methods
- Non-parametric methods
- Why these methods are not quite correct

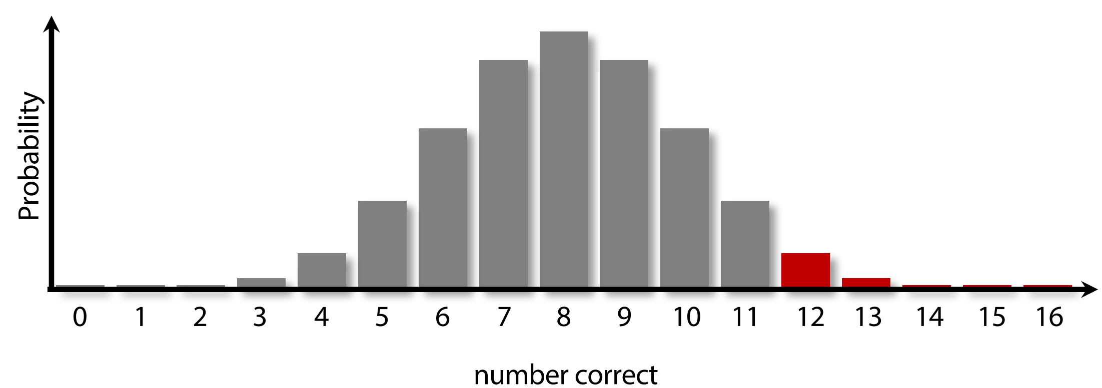
DECODING-LEVEL STATISTICS

Decoding Level: Binomial Test

Exact test: Chance level corresponds to fair toss of coin

- Null hypothesis: The observed accuracy has come about by chance (i.e. it comes from the null distribution)
- Alternative hypothesis: The observed accuracy did not come about by chance (it does not come from the null distribution)
- *p*-value = 1 – (probability to have head max x times)

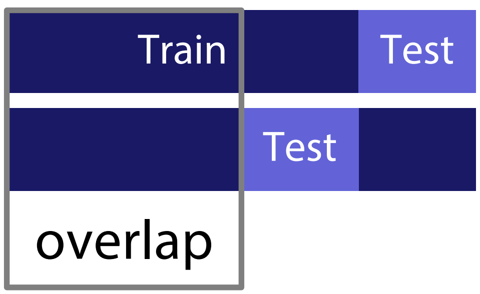
$$f(x|n,p\ ) = 1 - \sum_{x=0}^{n} \binom{n}{x} p^{x}(1 - p^{n-x})$$



Problem for Binomial Test in Cross-Validation

Assumption of binomial test: independent samples

- Valid only for independent test data
- Likely invalid for within-run analyses (are data in a run independent?)
- Definitely invalid for cross-validation: Assumption violated because same data is used multiple times during training

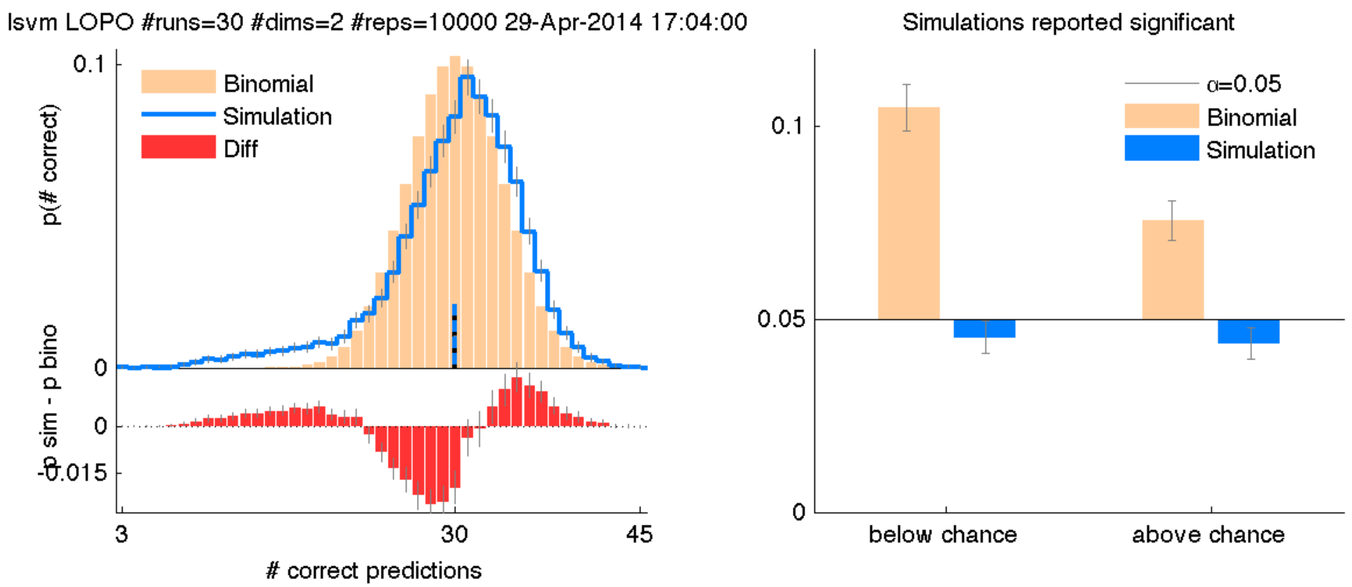


We almost always do cross-validation, i.e. we will almost *never* use a binomial test at the decoding-level
(and of course no test that assumes independent sampling, e.g. *t*-test, *z*-test, …)

Problem for Binomial Test in Cross-Validation

Cross-validated accuracies are *not* binomially distributed!

- Estimate of mean not positively biased (i.e. chance level correct)
- Estimate of variance different to binomial distribution



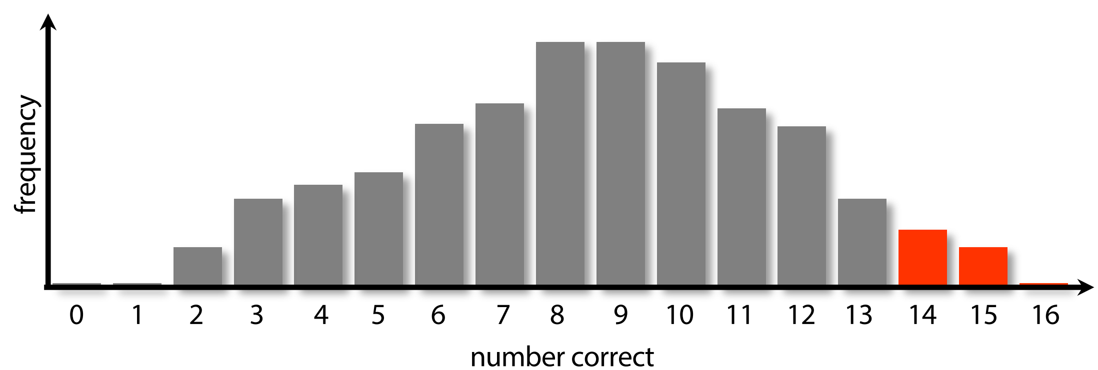
Görgen, Hebart et al (in prep), Noirhomme et al (2014) – Neuroimage Clin; Jamalabadi et al (2016) – HBM

Decoding Level: Permutation Test

Permutation test

- H0: There is no systematic relationship between the class labels and the data (i.e. data of all classes comes from one distribution)
- H1: There is some meaning in the assignment of labels to data
- Estimate null distribution from the data by permuting labels
- Under the H0, label assignment is meaningless, i.e. probability of observing an extreme result based on that distribution (e.g. accuracy) is small → extreme result means there is some meaning in labels

Example null distribution



Common Misunderstandings for Permutation Test

"Permutation tests are assumption free"

→ **Wrong!** No assumption of independence, but weaker assumption of exchangeability

"The nominal chance-level cannot be trusted, but permutation tests can establish an "empirical chance-level" that may be higher or lower than nominal chance"

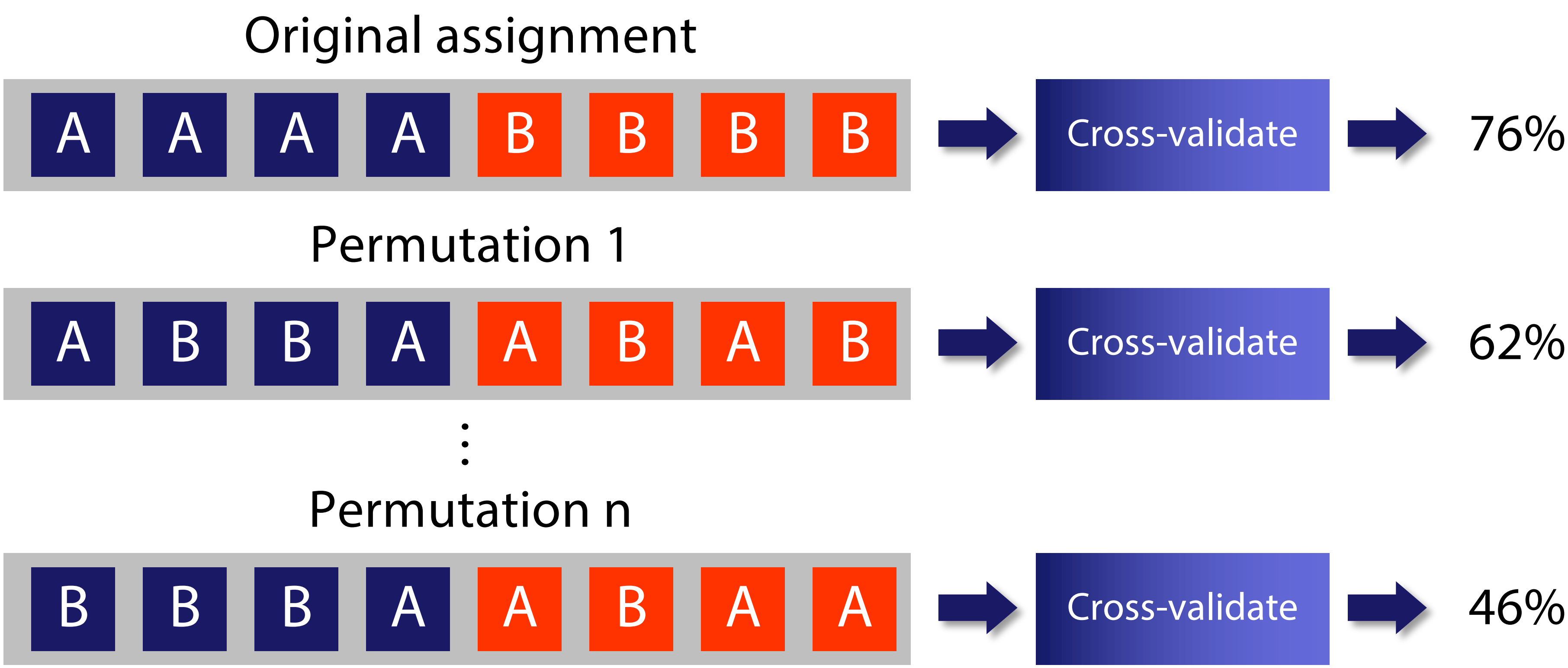
→ **Wrong!** True chance in cross-validation is not biased!

→ Goal of permutation test is <u>not</u> to establish empirical chance

→ If a confound biases null distribution, permutation test will only correct for it if accurately modeled

Decoding Level: Permutation Test

Typical procedure: Label permutation (i.e. random shuffling)



p-value: How often 76 % reached or exceeded by n permutations?

Etzel & Braver (2013) – PRNI

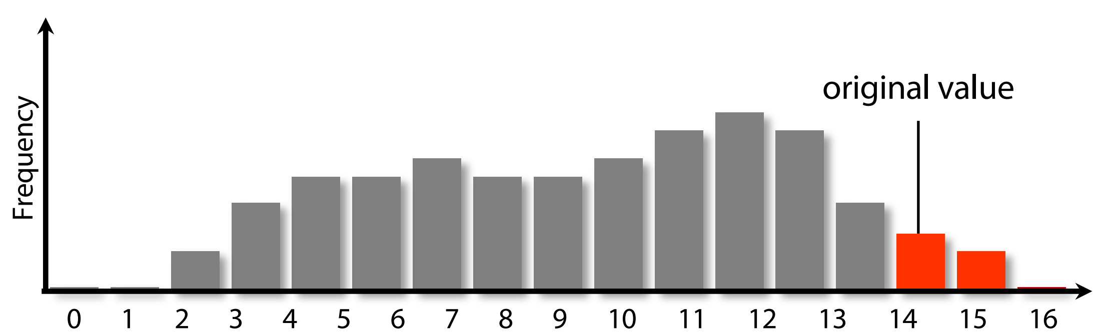
Decoding Level: Permutation Test

Procedure:

1. Calculate statistic (e.g. accuracy) using normal procedure
2. Permute labels and repeat same procedure
3. Repeat n times or until exhaustion

For exhaustive test: *p*-value is $\frac{k}{n}$ where *k* is number of permutations with equal or higher accuracy and *n* is number of all permutations (includes original result)

For non-exhaustive test (aka Monte-Carlo permutation test): $\frac{k+1}{n+1}$



Decoding Level: Permutation Test

Very common problem: Non-exchangeable samples

- Sequential dependencies (autocorrelation, correlated regressors) limit exchangeability *within* run
- Dependence *within run* limits exchangeability of labels *between runs*, i.e. do not permute labels across runs! (else you may predict run rather than class label)

→ Correct treatment: Retain leave-one-run-out (this is called block permutation)

→ Permute and then do everything the same as done originally

Hebart, Görgen, Allefeld (ongoing project)

Decoding Level: Permutation Test

Overview of versions in literature

- Random within-run permutation: valid when no sequential dependencies, else only full exchange (all labels 1 → -1 and vice versa)
- Permute labels randomly across runs → not valid in cross-validation
- Permute labels during cross-validation (i.e. no fixed assignment between data and labels) → not valid
- Only permute training labels in cross-validation → not valid
- Only permute test labels in cross-validation → not valid

Pereira et al (2011) – Neuroimage; Al-Rawi et al, 2012 – Neurocomputation; Stelzel et al (2012) – Neuroimage

Group Level (Second-Level) Statistics

Parametric tests

- *t*-test, ANOVA, etc.
- Mostly valid when statistic (e.g. accuracy) is not positively biased
- Larger variance at decoding level (through cross-validation) will translate to group-level, i.e. test rather conservative

Group Level Statistics

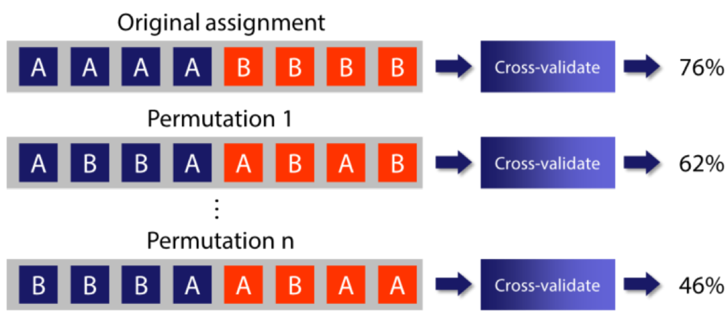
Permutation test

a) Sign permutation test
   - Alternative to e.g. *t*-test on group level
   - Cannot be used on biased results at decoding-level
   - Popular, because people trust permutation tests more

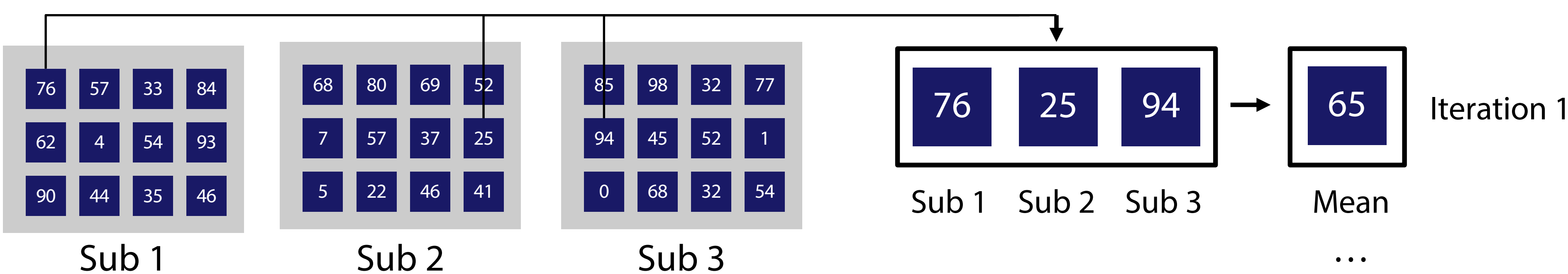
b) "Two-step" permutation test
   - Can be used on biased results at decoding-level
   - Much more computationally expensive

Group level Statistics: Two-step Permutation Test

- Step 1: Calculate a number of permutations at subject level



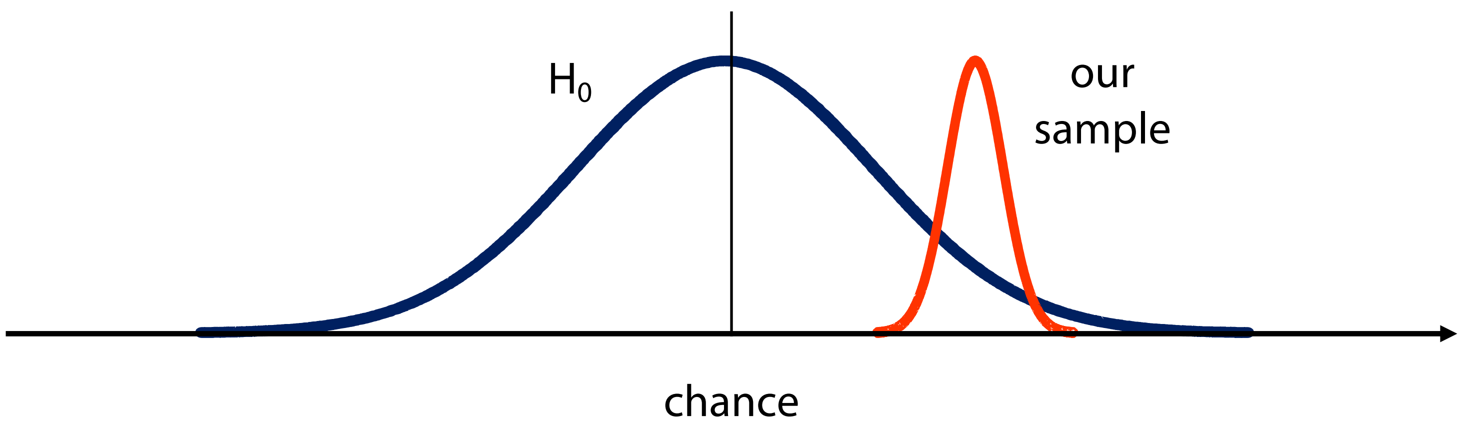
- Step 2: Draw for each subject one result from the pool of these permutation results (including the original result), calculate group-level outcome; repeat n times



Most Tests at Group Level Not Quite Correct

H_0 of group statistical analysis: Our sample belongs to the null distribution, i.e. the distribution of the statistic under the null model

- Random effects analysis (e.g. group *t*-test): Assumes decoding-level error is negligible



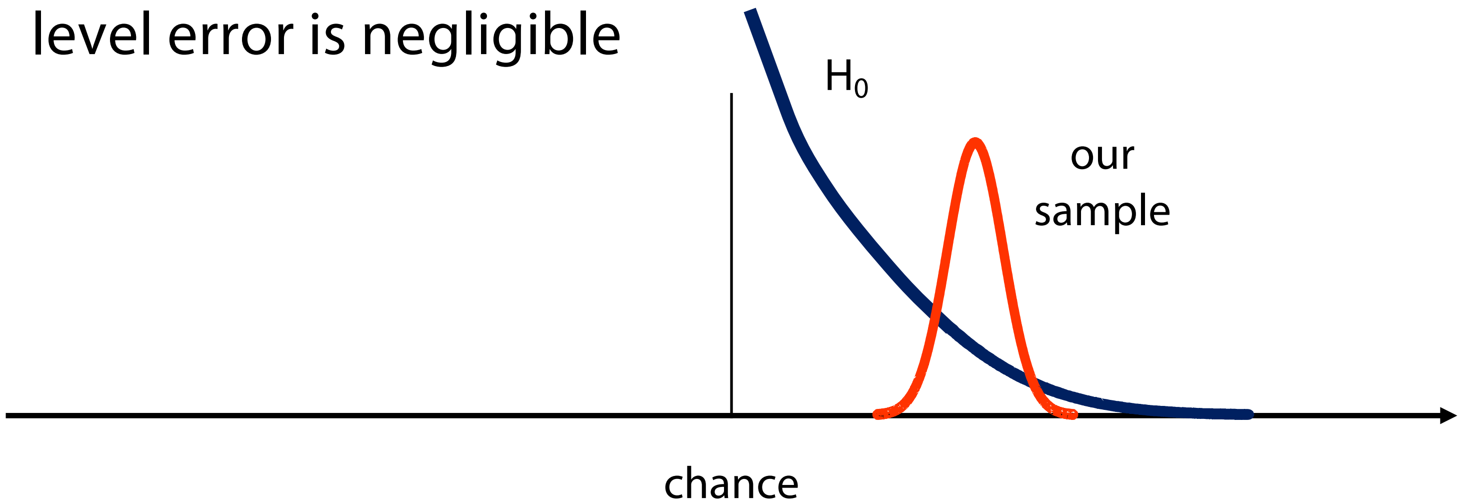
PROBLEM: There are no *true* effects smaller than chance, i.e. null-hypothesis is wrong

Allefeld et al (2016) – Neuroimage

Most Tests at Group Level Not Quite Correct

H_0 of group statistical analysis: Our sample belongs to the null distribution, i.e. the distribution of the statistic under the null model

- Random effects analysis (e.g. group *t*-test): Assumes decoding-level error is negligible



➡ Real H0? Real group-level chance? Unknown

Allefeld et al (2016) – Neuroimage

Most Tests at Group Level Not Quite Correct

Only null distribution symmetrical around chance is zero variance point hypothesis

→ random effects test collapses to fixed effects test

→ actual H_0: no subject carries an effect

→ actual H_1: at least one subject carries an effect

- Unclear how severe this effect is
- Alternative test: prevalence inference, i.e. do a majority of participants carry an effect? But less sensitive
- Requires permutation test within-participant

Allefeld et al (2016) – Neuroimage

Summary

Decoding-level:

- Many standard tests not valid for cross-validation designs (*t*-test, binomial test, wrong permutation tests)
- Permutation tests that respect data dependence recommended

Second-level:

- Classical t-tests and ANOVAs are ok, but slightly conservative
- Two-step permutation tests are possible alternative, good for biased results
- All these tests are likely not quite correct but test fixed effects hypothesis

Study Questions

Question 1: A colleague comes to you and asks for your expertise in decoding statistics. He has classified patients vs. matched controls using cross-validation and wants to carry out a statistical analysis at the decoding-level. What does he have to consider?

Question 2: A colleague comes to you and asks for your expertise in decoding statistics. He has classified patients vs. matched controls using cross-validation, but he has left out a separate test set of patients and controls that he has applied the classifier to. He has generated a classification accuracy and wants to test whether it is significant. What test do you recommend her to use?

Question 3 (difficult): You want to know if your subject has been classified above chance. You have done leave-one-run out cross-validation with one beta per condition per run and have a total of 8 runs. How many unique permutations are possible? What can you do to get more?