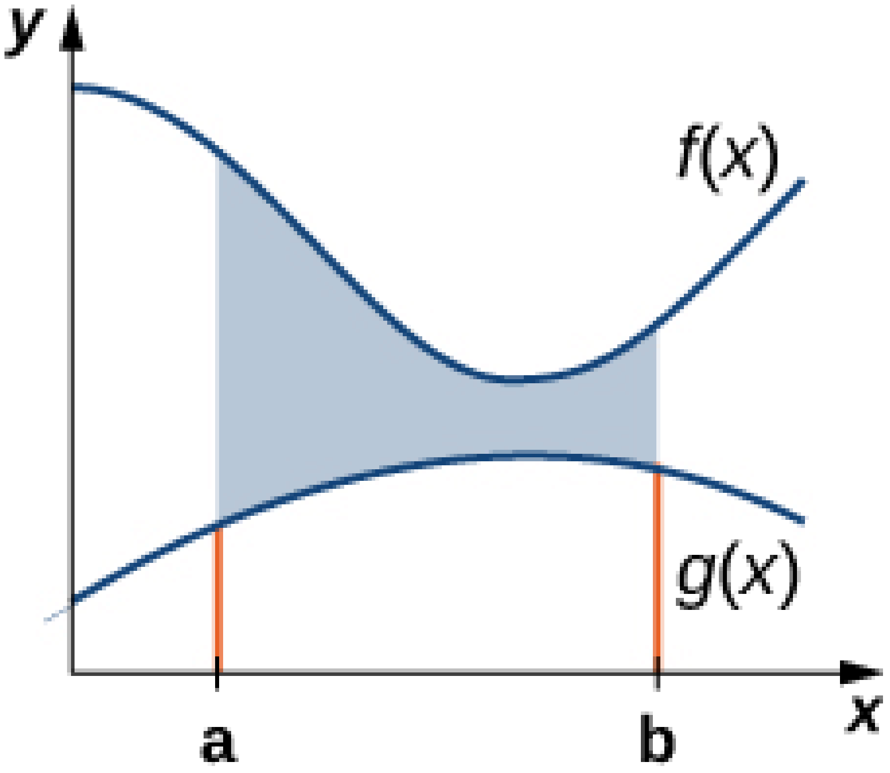
University of Wisconsin
SCHOOL OF MEDICINE
AND PUBLIC HEALTH

Permutation Test

Moo K. Chung
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

www.stat.wisc.edu/~mchung

Motivating problem



Test the equivalence of two functional data

Then compute *p*-value

How?

Permutation resampling

Path enumeration for exact probability computation

p-value computation in statistical inference

Data augmentation in deep learning

References

Hayasaka, S. Nichols, T.E. Validating cluster size inference: random field and permutation methods NeuroImage 20:2343-2356 http://www.sciencedirect.com/science/article/pii/S1053811903005020

Chung, M.K. et al. 2019. Rapid Acceleration of the permutation test via transpositions, *International Workshop on Connectomics in NeuroImaging*, Lecture Notes in Computer Science 11848:42-53.

Permutation Test

R.A. Fisher invented the method in 1935 in The Design of Experiments. This is the exact procedure for computing p-value.

Does not assume any statistical distribution- nonparametric.

Requires permutation invariance: exchangability under the equivalence of null hypothesis: two groups.

- There may be a situation we cannot permute.

Wide use of permutation test

Google scholar 888,000 papers.

One of the most widely used method in brain imaging. Why?

- Exact and easy to understand. No need to study statistical models.

Deep learning – Permutation invariant resampling: 14,000 deep papers related to permutations

Permutation group

$$\mathbf{x} = (x_1, x_2, \cdots, x_m) \qquad \mathbf{y} = (y_1, y_2, \cdots, y_n)$$

$$(\mathbf{x}, \mathbf{y}) = (x_1, \cdots, x_m, y_1, \cdots, y_n)$$

$\pi(\mathbf{x}, \mathbf{y}) \in \mathbb{S}_{m+n}$ Permutation group of order *m*+*n*

$$(x_1, y_1, x_3, \cdots, x_m, x_2, y_2, \cdots, y_n)$$

Number of permutations $\binom{m+n}{m} = \frac{(m+n)!}{m!n!}$

Permutability

$$\mathbf{x} = (x_1, x_2, \cdots, x_m) \qquad \mathbf{y} = (y_1, y_2, \cdots, y_n)$$

$$(1,3) = (2,4)$$

Null hypothesis: $f(\mathbf{x}) = f(\mathbf{y})$ $\qquad 2 = 3$

$$(\mathbf{x}, \mathbf{y}) = (x_1, \cdots, x_m, y_1, \cdots, y_n)$$

$$\pi \quad (x_1, y_1, x_3, \cdots, x_m, x_2, y_2, \cdots, y_n)$$

Permutability: $f(\pi(\mathbf{x})) = f(\pi(\mathbf{y}))$

Number of permutations in permuting group labels

Number of permutations in literature

Fisher 1935, The Design of Experiment

$$\binom{8}{4} = 70$$

Thompson et al. 2001, Nature Neuroscience

$$\binom{40}{20} = 1.34 \cdot 10^{11}$$

Nichols et al. 2002, Human Brain Mapping

4279 citations

$$\binom{6}{3} = 20 \qquad AAAAAA|BBB$$

test

Serious computational bottleneck in brain imaging

1) Need to permute million voxels.
2) Compute the statistic for each permutation

Thompson et al. 2001 used supercomputer:
1million permutations from

$\binom{40}{20} = 1.34 \cdot 10^{11}$ hundred billion permutations

→ Transposition test (online test) bypass the bottleneck
Used masterfully by a previous student
http://arxiv.org/pdf/2012.00675.pdf

Hypothesis testing via permutat

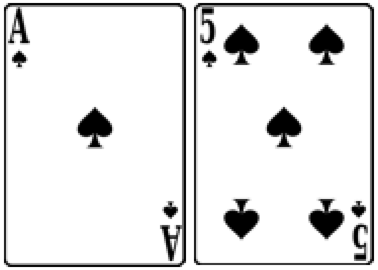
1. Set up reasonable null and and alternate hypothesis

2. Set up a function (statistic) that measures the strength of claims

3. Compute the probability about the statistic in the sample space generated by the permutation test

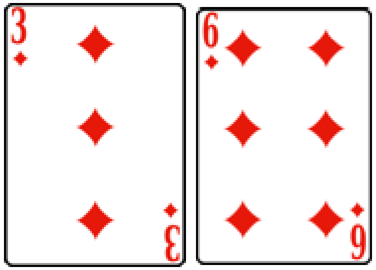
Sample space

1 5



Player 1

3 6



Player 2

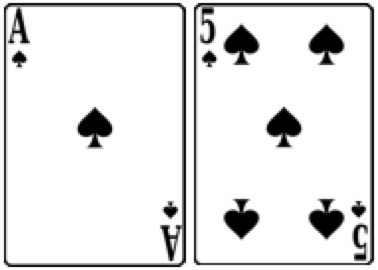
Observed data

What is the likelihood (probability) of this event (observed data) happening?

$$\binom{4}{2} = 4 * 3/2 = 6$$

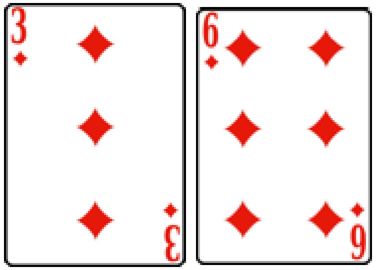
Sample space = all possible permu

1 5



Player 1

3 6



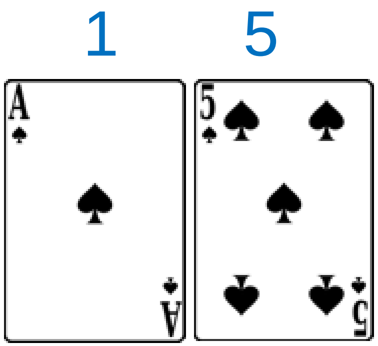
Player 2

Observed data

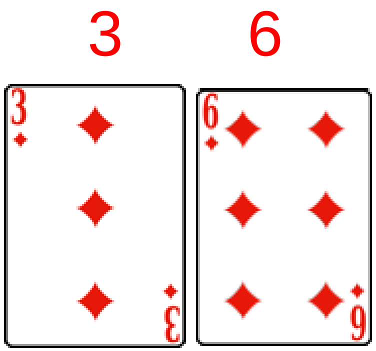
1	5	3	6
1	3	5	6
3	5	1	6
1	6	3	5
6	5	3	1
6	3	1	5

Generate the sample space using only observations

How to design a two-sample test?



Player 1



Player 2

Observed data

What is the likelihood (probability) of Player 1 and Player 2 have the same card power?

How to formulate the problem?

Denote X_j and Y_j as the values of cards for Player 1 and 2 respectively.

Power = sum of card values: $\sum_{j=1}^{m} X_j$ $\sum_{j=1}^{m} Y_j$

Test statistic – one sided

$$\sum_{j=1}^{m} X_j > \sum_{j=1}^{m} Y_j$$ → Player 1 has better cards

Test statistic (distance): $$d = \sum_{j=1}^{m} X_j - \sum_{j=1}^{m} Y_j$$

Observed distance=-3

1 5 3 6

$H_0 : d = 0$ → Player 1 and 2 have similar cards

$vs.$

$H_1 : d \geq 0$ → Player 1 has better cards

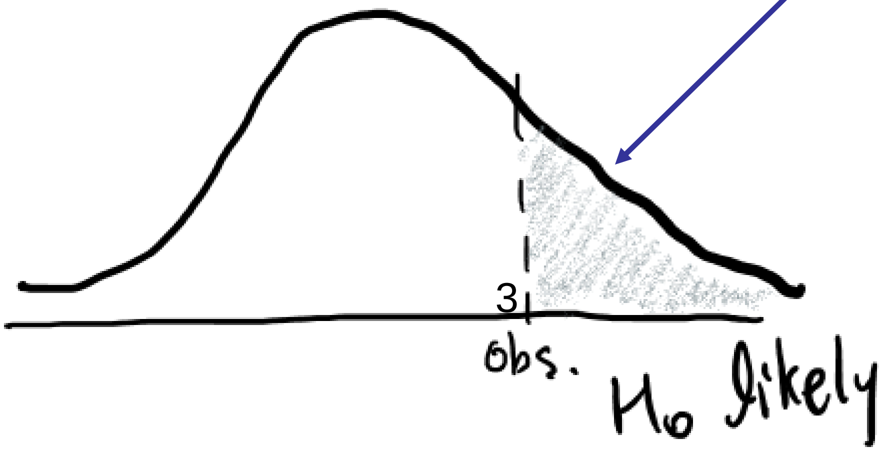
$$H_1 : d \leq 0 \leftarrow Player\ 2\ has\ better\ cards$$

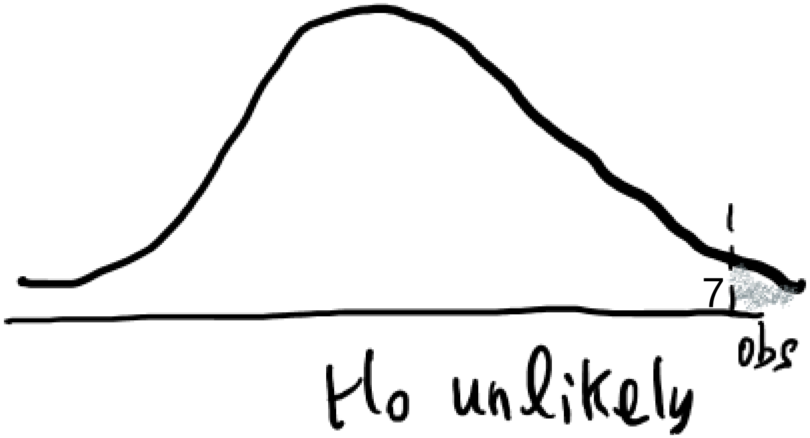
Distribution of test statistic

$$d = \sum_{j=1}^{m} X_j - \sum_{j=1}^{m} Y_j$$

1	5	3	6	Observation = -3
1	5	3	6	-3
1	3	5	6	-7
3	5	1	6	1
1	6	3	5	-1
6	5	3	1	7
6	3	1	5	3

$$P(d \geq -3) = 4/6$$

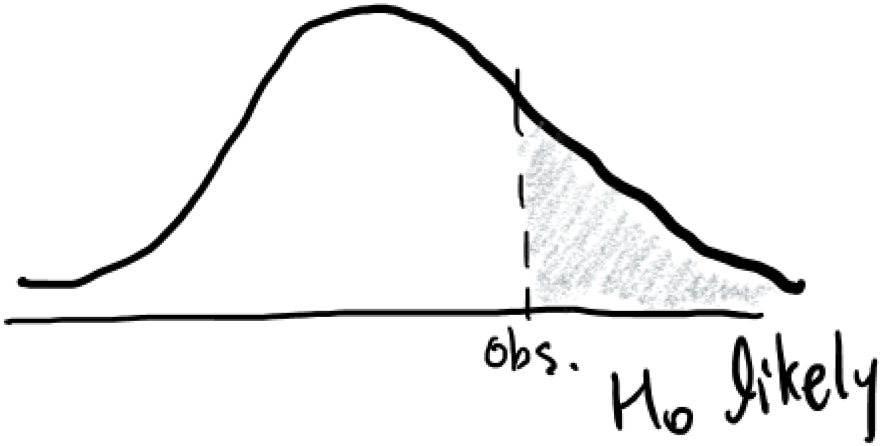


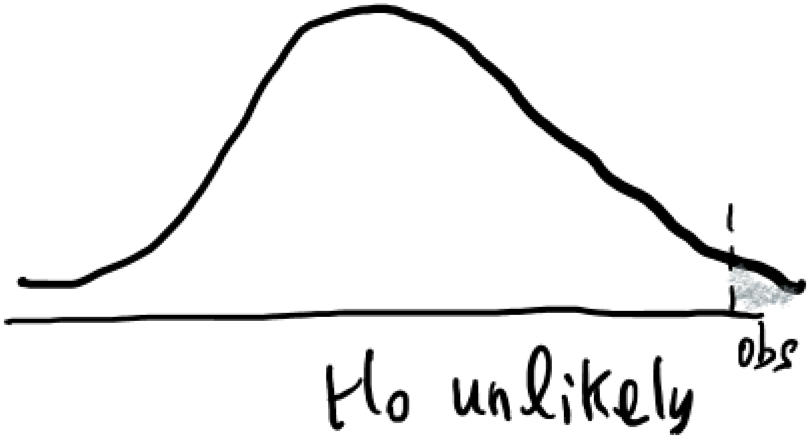


$$H_0 : d = 0$$

$$vs.$$

$$H_1 : d > 0$$





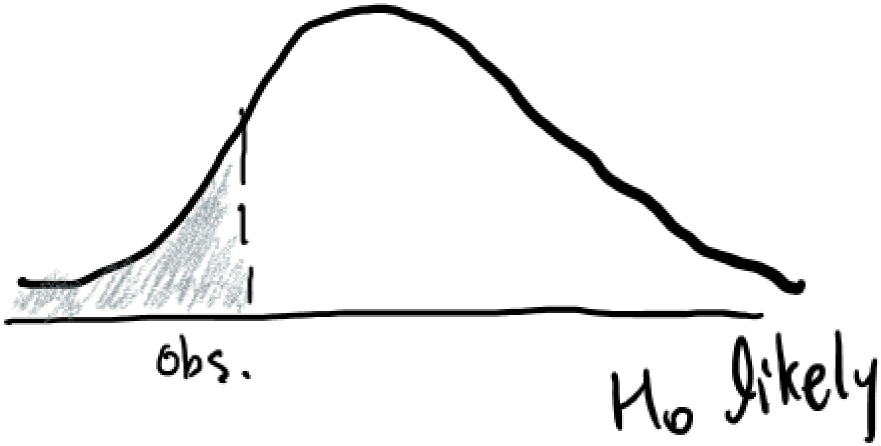
Definition: $$pvalue = P(d \geq observation)$$

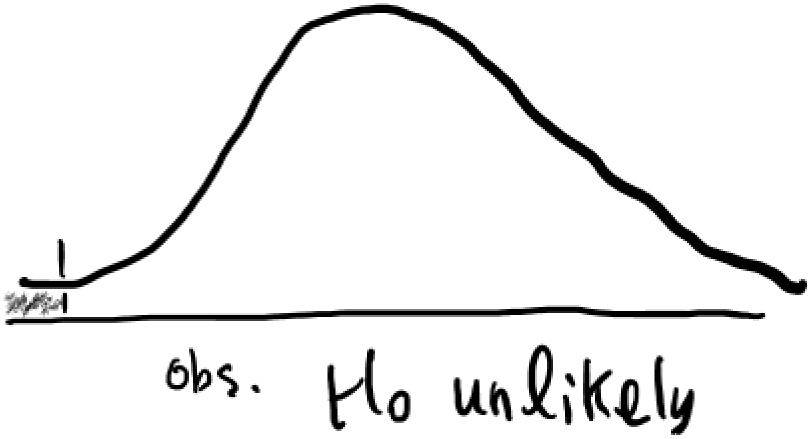
Interpret *p*-values as continuous indices of the strength of claim (H_0) or alternate claim (H_1)

$H_0 : d = 0$

$vs.$

$H_1 : d < 0$

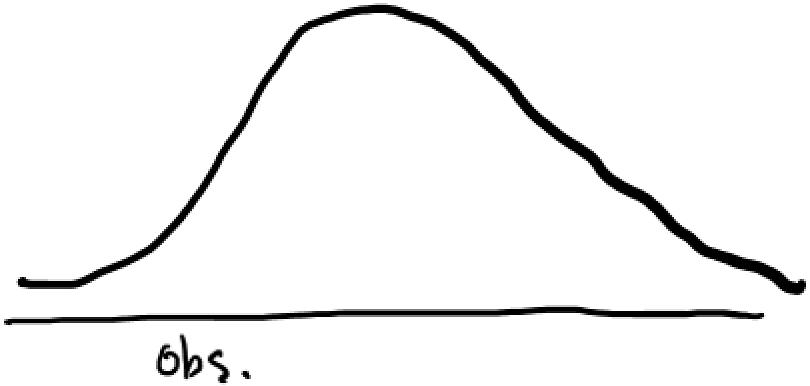


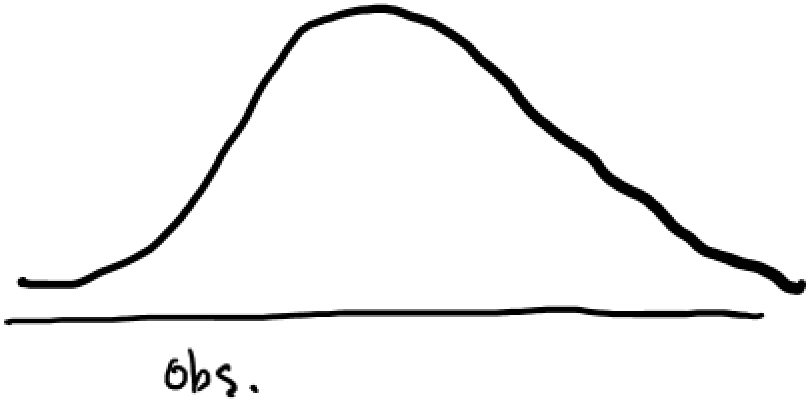


Definition: $pvalue = P(d \leq observation)$

Test statistic – two sided

$$H_0 : d = 0$$
$$vs.$$
$$H_1 : d \neq 0$$





$$P(|d - observation| > \epsilon)$$

$$pvalue = P(d \leq -|observation|) + P(d \geq |observation|)$$

p-value in the permutation test

$$\mathbf{x} = (x_1, x_2, \cdots, x_m) \qquad \mathbf{y} = (y_1, y_2, \cdots, y_n)$$

$$(\mathbf{x}, \mathbf{y}) = (x_1, \cdots, x_m, y_1, \cdots, y_n)$$

$$\pi(\mathbf{x}, \mathbf{y}) \in \mathbb{S}_{m+n}$$

observation

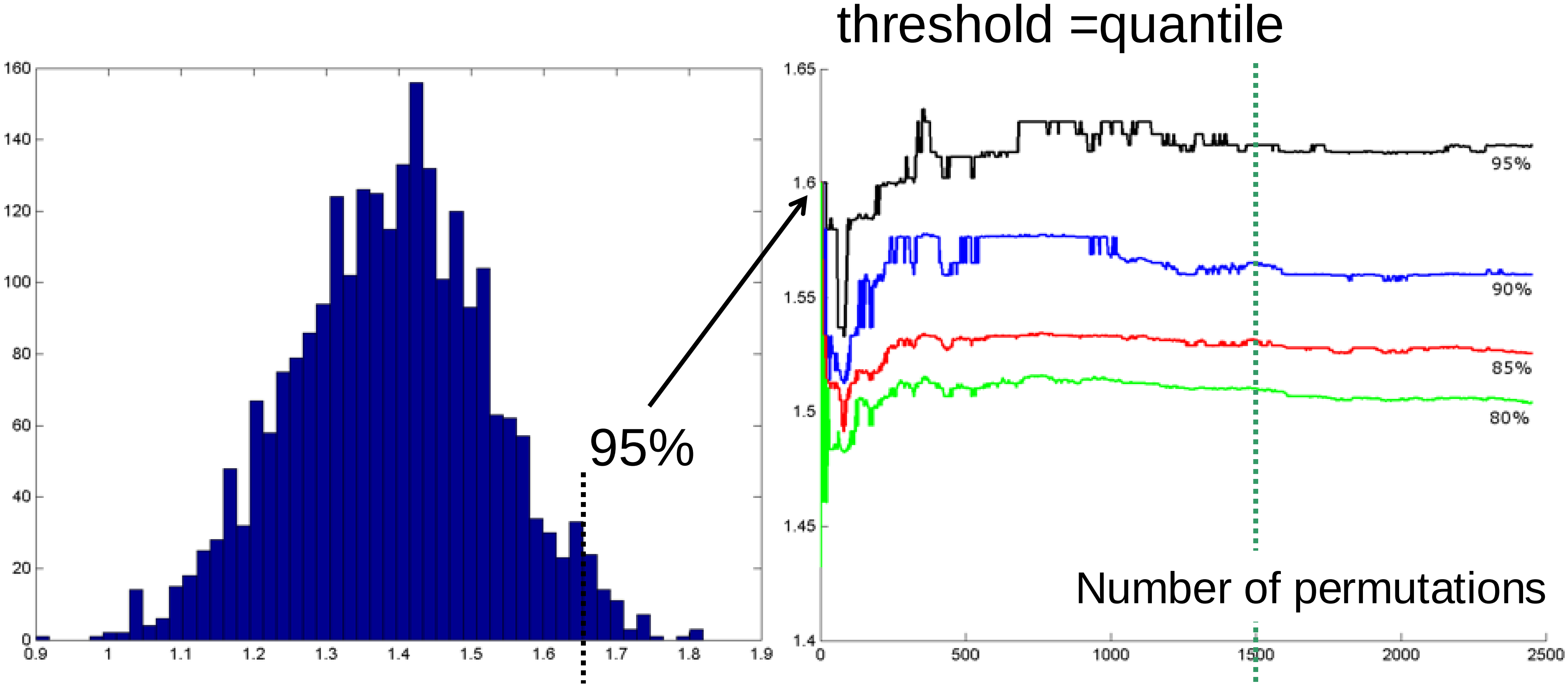
$$p\text{-value} = \frac{1}{(m+n)!} \sum_{\pi \in \mathbb{S}_{m+n}} \mathcal{I}\Big(f(\mathbf{x}, \mathbf{y}) \leq f(\pi(\mathbf{x}), \pi(\mathbf{y})) \Big)$$

If you work out symmetry,
you can reduce this to
m+n choose m

Test statistic
over
permutations

Exercise. Set up an iterative computational procedure

How to numerically check the convergence of permutation test

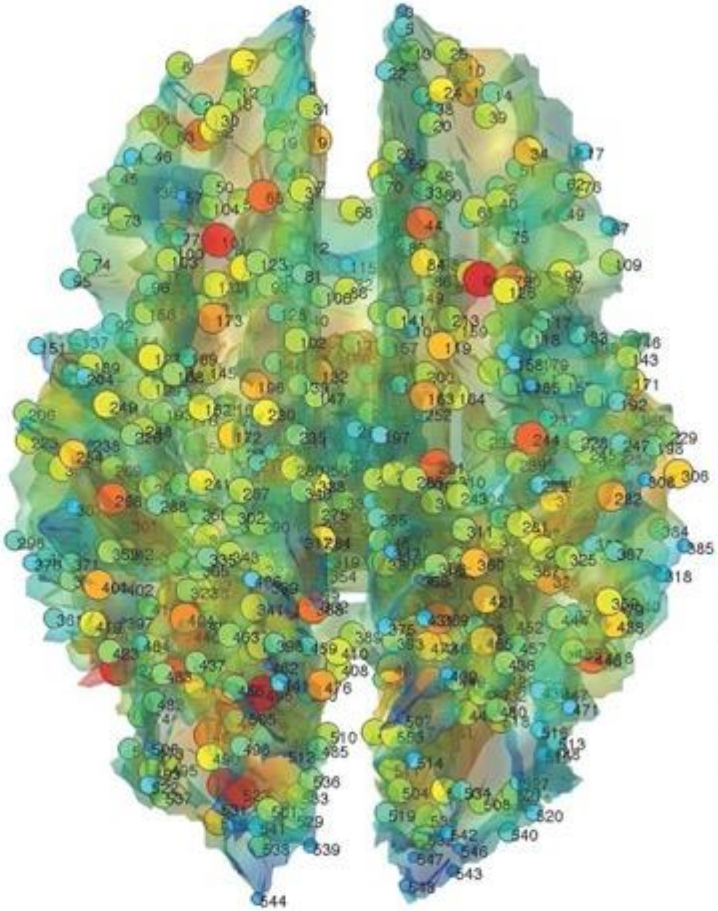


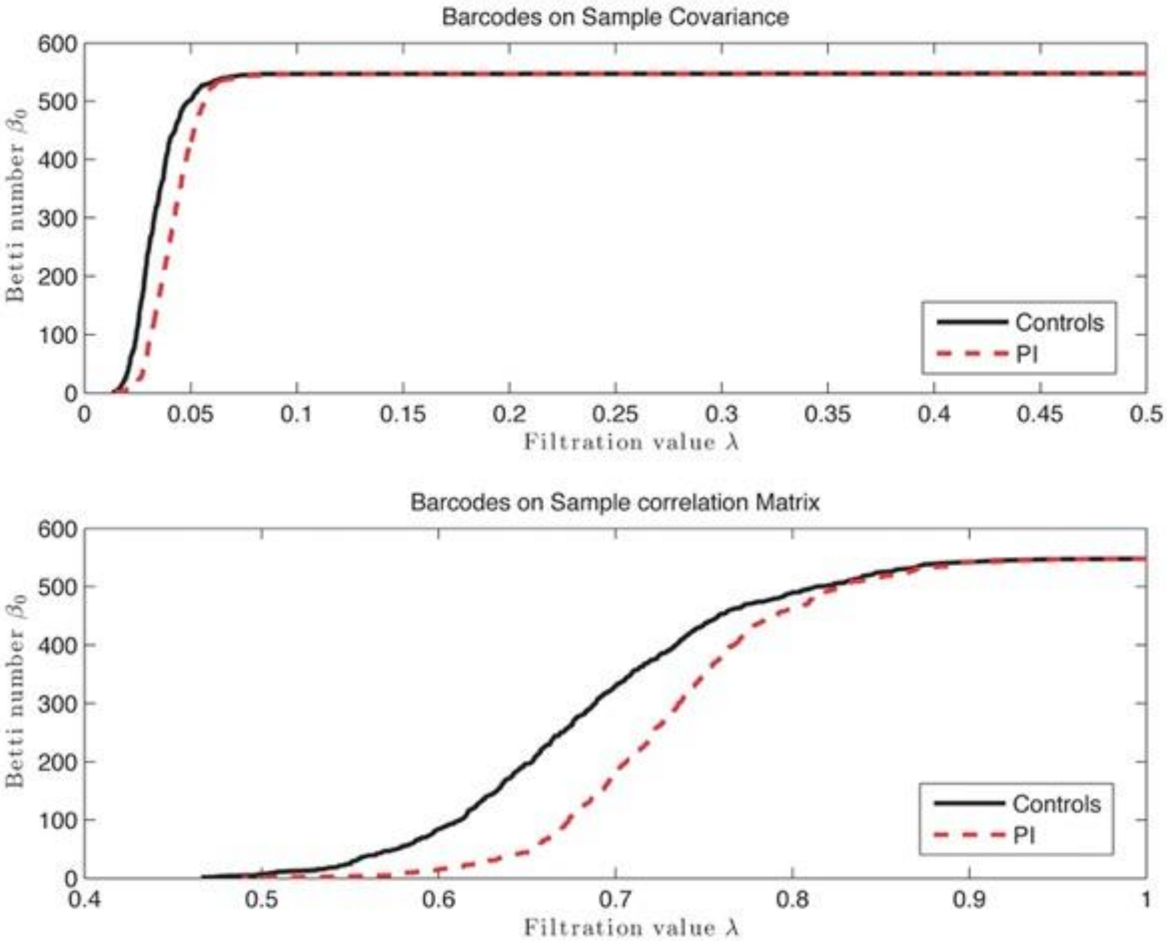
threshold =quantile

95%

Number of permutations

Example: permutation test on 1D functional data: barcodes in TDA or cumulative distribution functions (CDF)



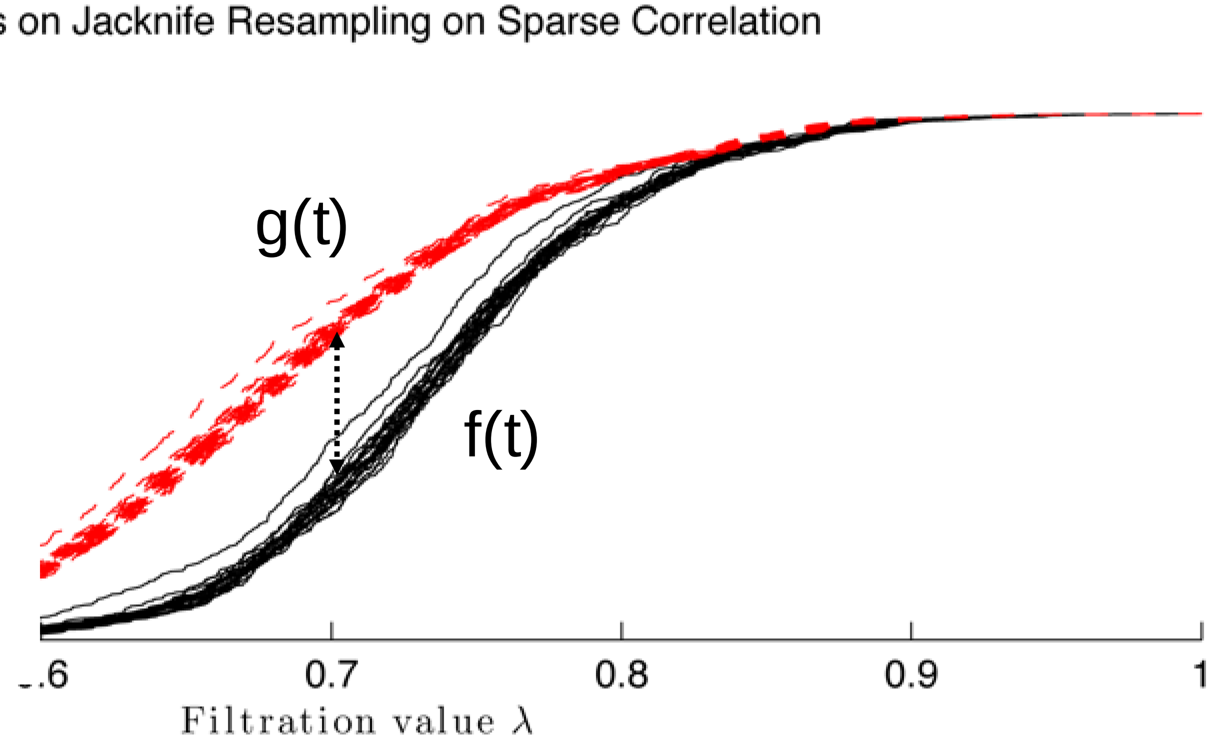


Average barcodes

Chung et al. 2015 TMI

Barcodes for all the subjects

$$d(f_i, g_j) = \sup_t |f_i(t) - g_j(t)|$$



Test statistic

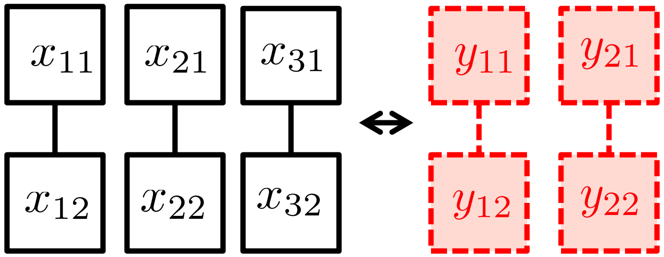
$$\sup_t |\bar{f}(t) - \bar{g}(t)|$$

Permutation test on a collection of barcodes and determine the statistical significance of curve shape differences.

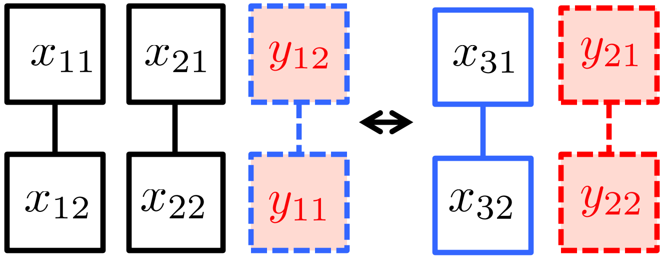
Matlab demo

Example: Permutation of paired data –twins

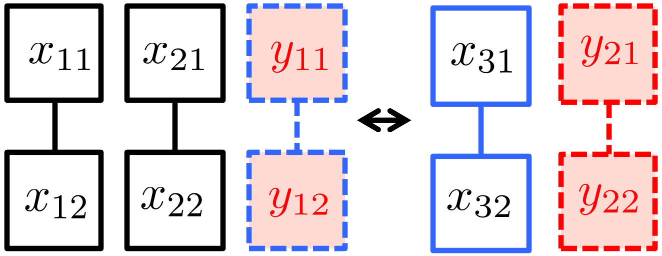
MZ-twins DZ-twins



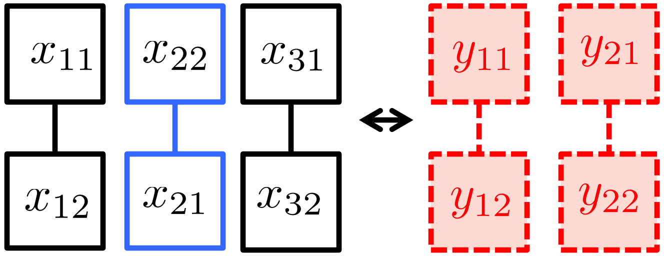
Permutation is the composition of two types of permutations



Permutation between groups



Permutation within twins



Used masterfully by a previous student
http://arxiv.org/pdf/2012.00675.pdf

Permutation of paired data

The sample space is generated by the permutation within and between groups.

# of permutations = $2^{m+n-2}\binom{m+n}{n}$

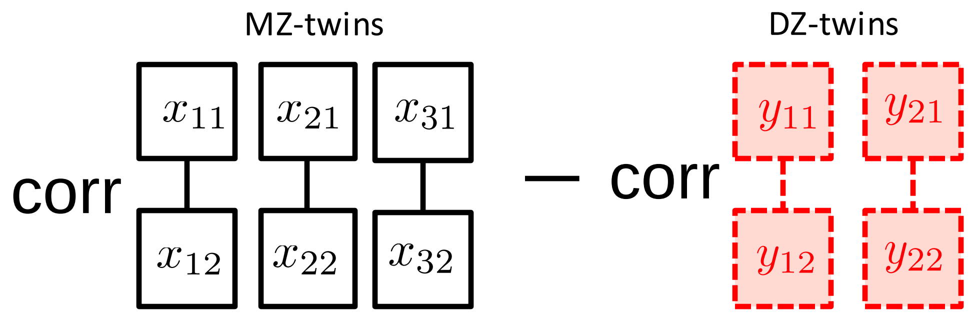
Exercise. Determine if it is the correct number of permutations.

The sample space is so large, you need extremely fast algorithm

→ online algorithms for resampling -transpositions

Permutation of paired data

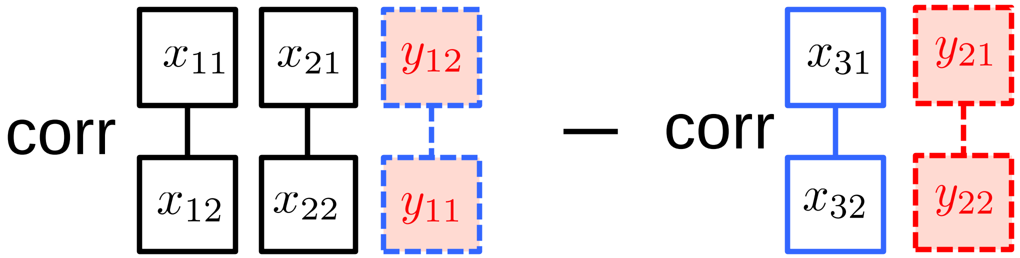
How to design a test statistic?



Permutation within group

Permutation within group

Observed test statistic



Permutation within and between groups

Permutation within and between groups

test statistic under the permutation within and between groups

Research problem

Have you ever seen the permutation test on three groups?

$$\mathbf{x} = (x_1, x_2, \cdots, x_m)$$

$$\mathbf{y} = (y_1, y_2, \cdots, y_n)$$

$$\mathbf{z} = (z_1, z_2, \cdots, z_l)$$

Other Resampling Techniques

Permutation test

Resampling in two samples

Jackknife

Resampling without replacement

Bootstrap

Resampling with replacement

History: Jackknife

Mosteller and Tukey (1977, p. 133) described a resampling method, the jackknife, in the following way:

> *“The name ‘jackknife’ is intended to suggest the broad usefulness of a technique as a substitute for specialized tools that may not be available, just as the Boy Scout’s trustworthy tool serves so variedly…”*

Resampling without replacement

1   2 3 4

leave-one-out
(2 3 4), (1 3 4), (1 2 4), (1 2 3)

leave-two-out
(3 4), (2 4), (2 3), (1 4), (1 3), (1 2)

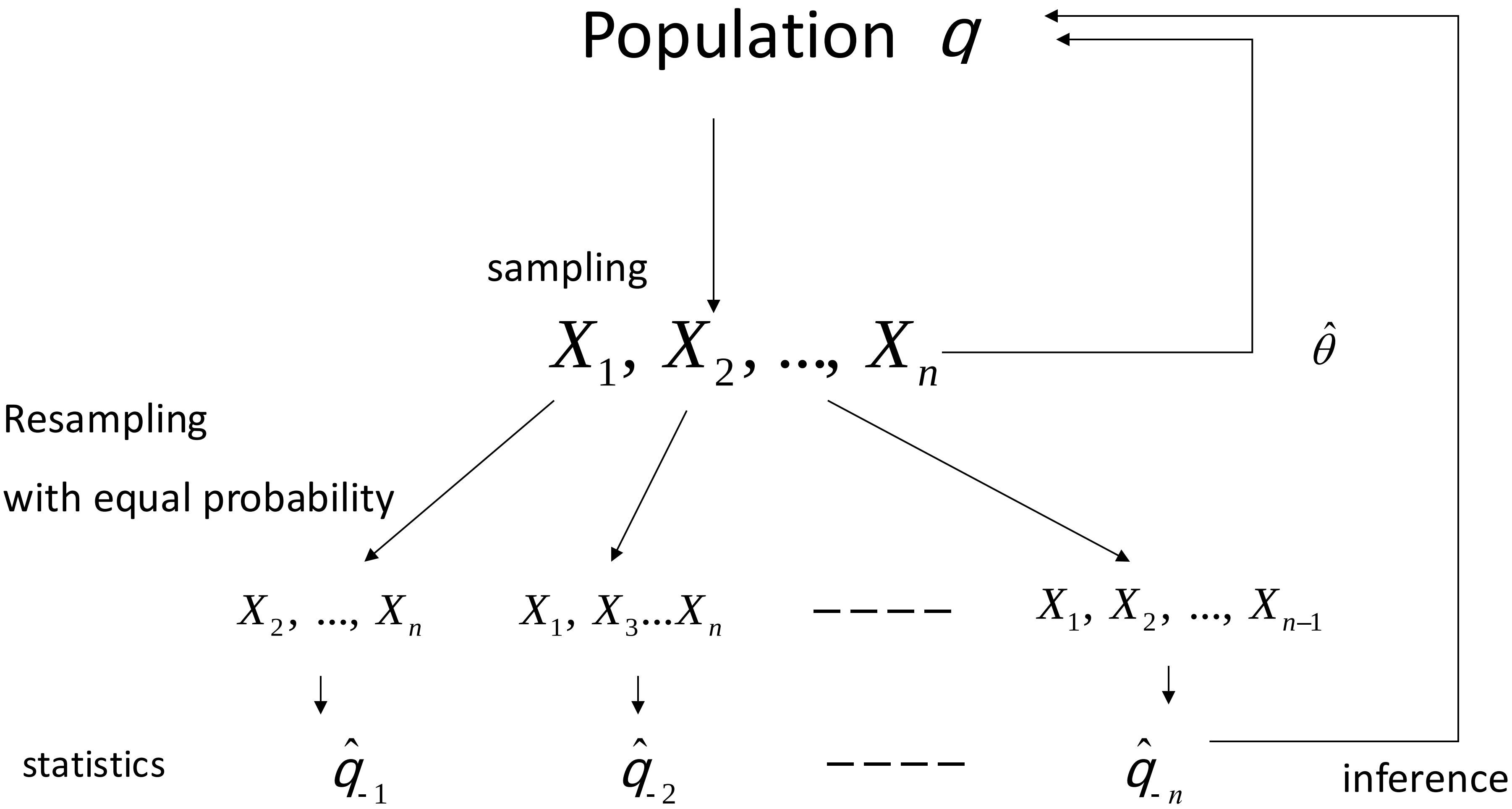
For sample size n,
leave-k-out $\binom{n}{k} = \frac{n!}{(n-k)!k!}$

Jackknife resampling method

For a group with *n* subjects, one subject is removed and the remaining *n*-1 subjects are used in computing statistics (leave-one-out scheme).

This process is repeated for each subject to produce *n* statistics.

Schematic of Jackknife



Population q

sampling

$X_1, X_2, \ldots, X_n$

$\hat{\theta}$

Resampling

with equal probability

$X_2, \ldots, X_n$ $X_1, X_3 \ldots X_n$ $----$ $X_1, X_2, \ldots, X_{n-1}$

statistics $\hat{q}_{-1}$ $\hat{q}_{-2}$ $----$ $\hat{q}_{-n}$

inference

Jackknife

population
mean

$$X_1,\ \ldots,\ \mathrm{X}_\mathrm{n} \overset{iid}{\sim} F(X;q)$$

$$\cancel{X}_1,\ X_2 \ldots,\ \mathrm{X}_\mathrm{n} \quad \text{Þ} \qquad \hat{q}_{-1} = \frac{X_2 + \ldots + X_n}{n - 1}$$

$$X_1,\ \cancel{X}_2 \ldots,\ \mathrm{X}_\mathrm{n} \quad \text{Þ} \qquad \hat{q}_{-2} = \frac{X_1 + X_3 + \ldots + X_n}{n - 1}$$

$$\vdots \qquad\qquad \vdots$$

$$X_1,\ X_2 \ldots,\ \cancel{\mathrm{X}}_\mathrm{n} \quad \text{Þ} \qquad \hat{q}_{-n} = \frac{X_1 + \ldots + X_{n-1}}{n - 1}$$

Jackknife estimation

Leave-one-out mean

$$\mu_{(i)} = \frac{1}{n-1}\sum_{j\neq i} X_j = \frac{n\bar{X} - X_i}{n-1}$$

Jackknife estimate of mean

$$m_{(.)} = \frac{1}{n}\overset{n}{\underset{i=1}{\mathring{a}}} m_{(i)} = \frac{n}{n-1}\bar{X} - \frac{1}{n-1}\bar{X} = \bar{X}$$

<u>Jackknife estimate of mean is unbiased</u>.

Example: Correlation between gill weight and body weight in 12 crabs

Sample correlation

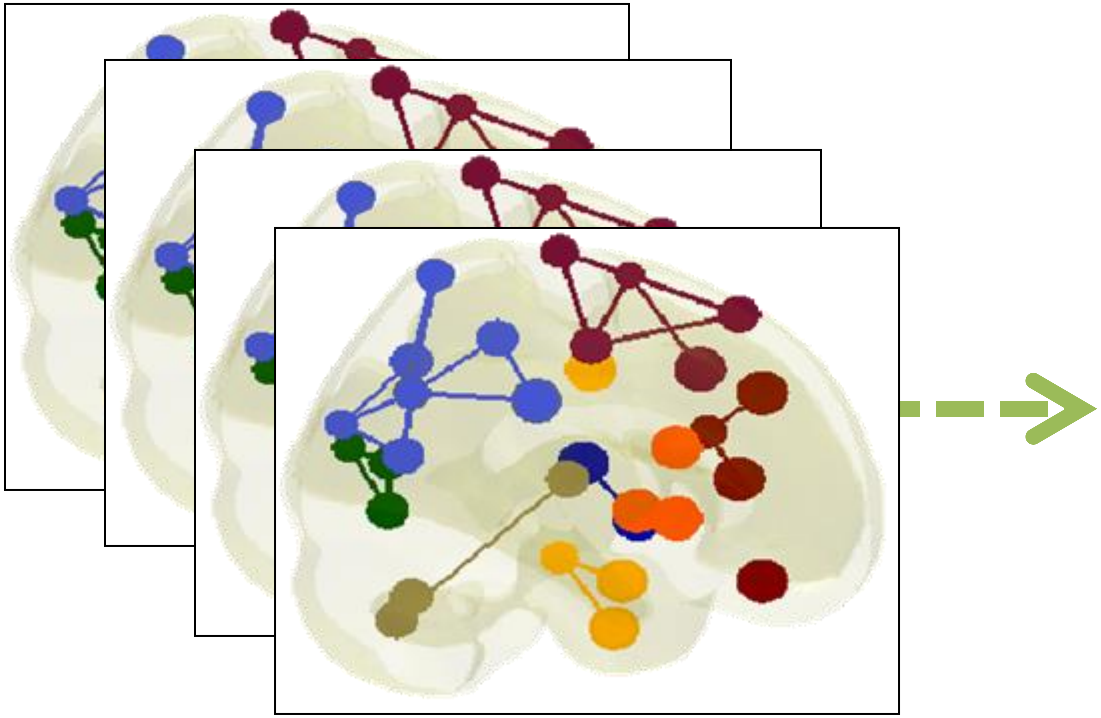
$r = 0.865$

Jackknife

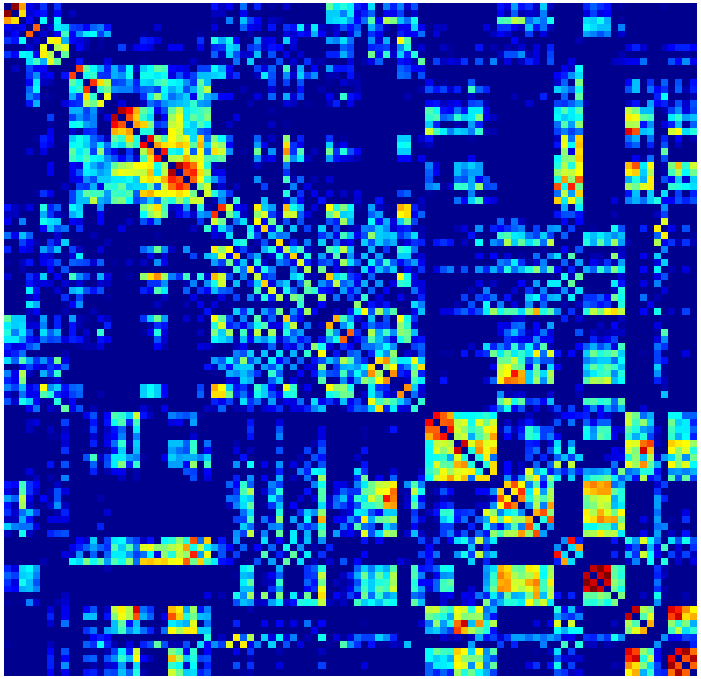
$r = 0.878$

Gill(mg)	Body(g)	r_{-i}	φ_i
1590	14.40	0.888	0.607
1790	15.20	0.884	0.656
10	11.30	0.892	0.570
450	2.50	0.830	1.249
3840	22.70	0.811	1.452
23	14.90	0.863	0.879
10	1.41	0.875	0.751
32	15.81	0.872	0.779
8	4.19	0.845	1.078
22	15.39	0.867	0.843
32	17.25	0.858	0.940
21	9.52	0.877	0.725

Example: Jackknife on correlation matrix



Multiple PET images on 90 voxels



90 x 90 correlation map per group

Lee, H. et al. 2012 IEEE Trans. Medical Imaging 31:2267-2277

MATLAB function

`y = randsample(n,k)` returns a k-by-1 vector y of values sampled uniformly at random, without replacement, from the integers 1 to n.

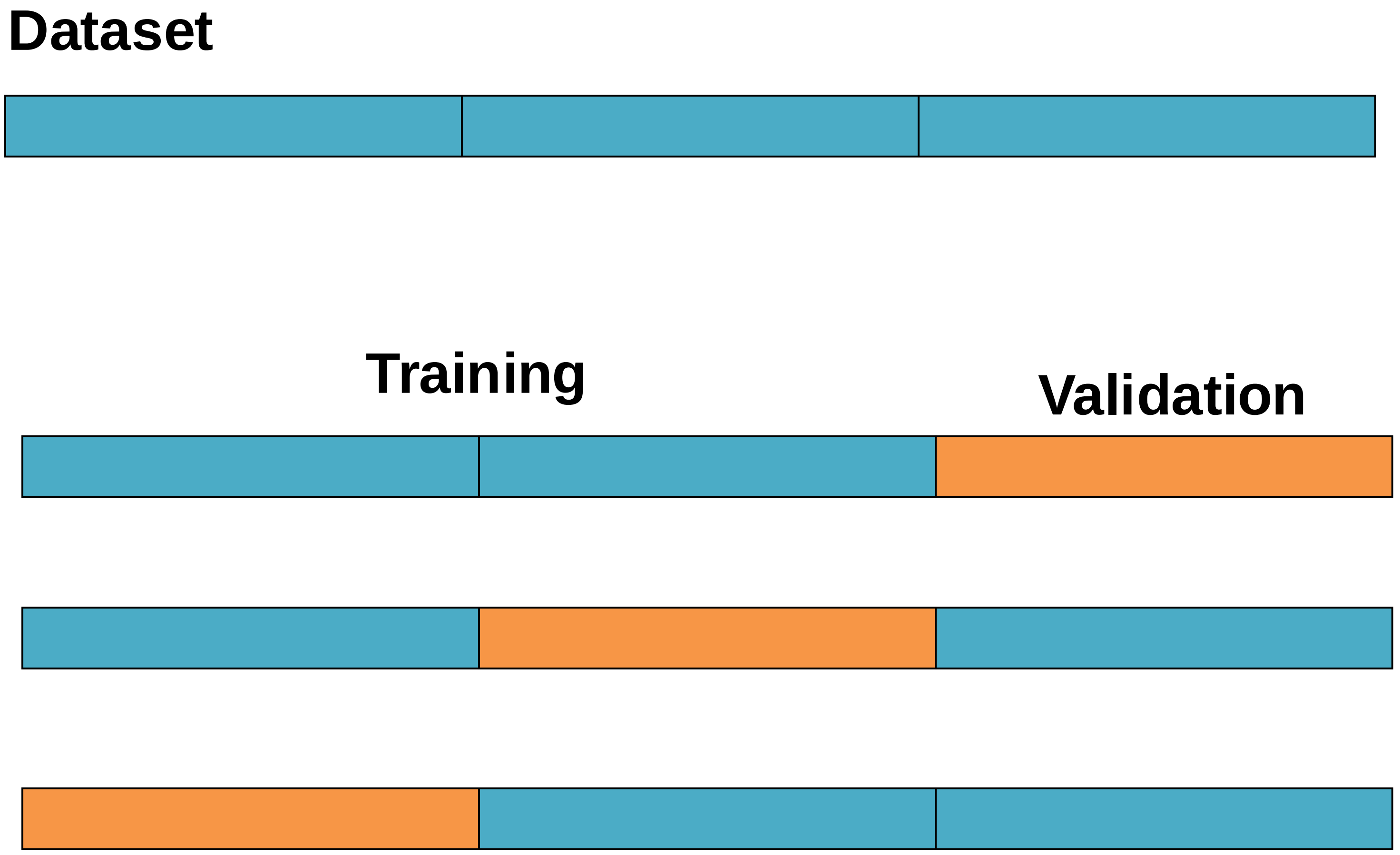
`jackstat = jackknife(jackfun,X)` draws jackknife data samples from the n-by-p data array X, computes statistics on each sample using the function jackfun, and returns the results in the matrix jackstat. jackknife regards each row of X as one data sample, so there are n data samples.

Cross-validation

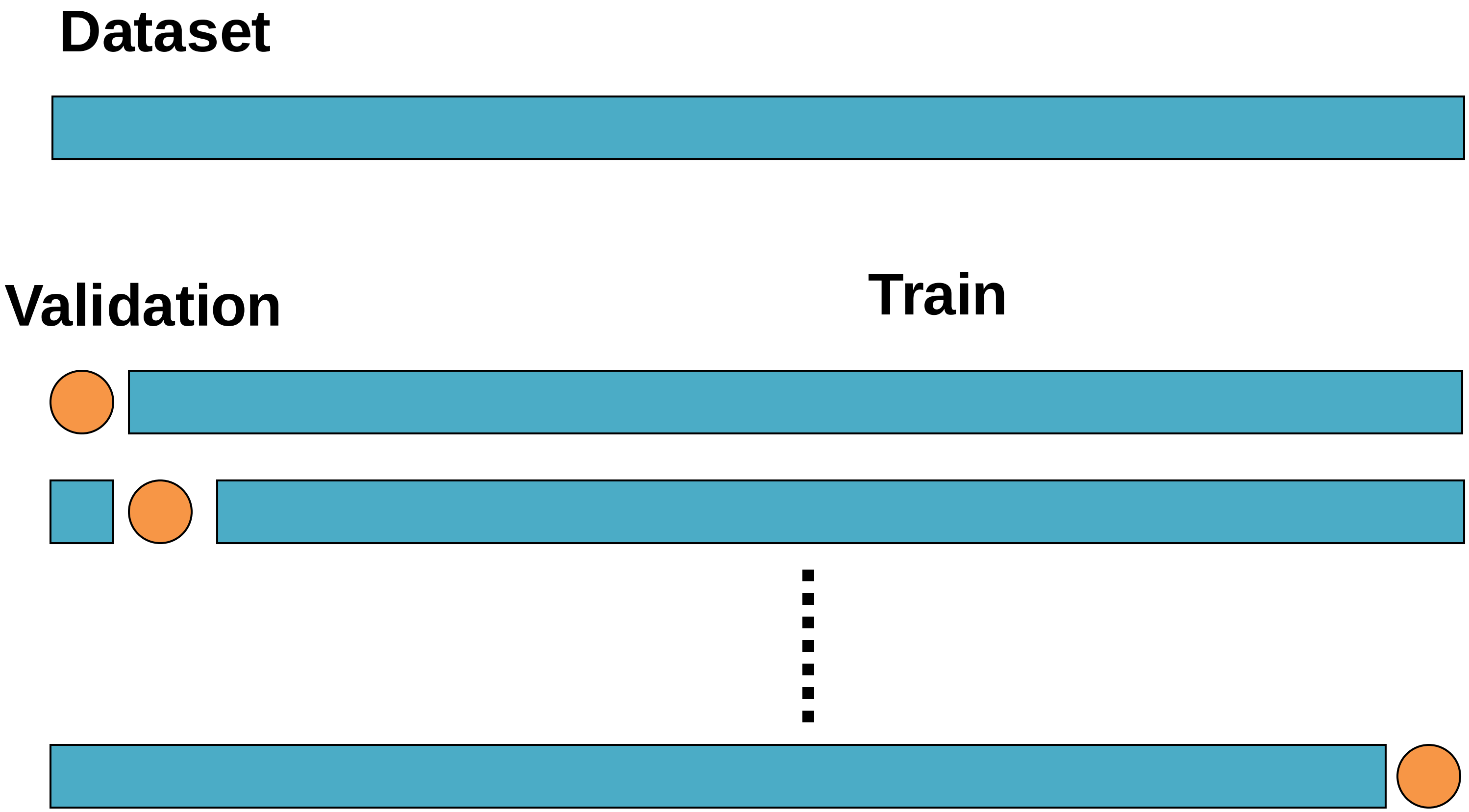
Cross-validation is a model validation technique for estimating the performance of a (predictive) model.

Invented by statistician A.K. Kurtz in 1948
(A research test of Rorchach test, Personal Psychology 1:41-53). Extended further by Mosier (1951) and Krus and Fuller (1982)

3-fold cross-validation: splitting strategy



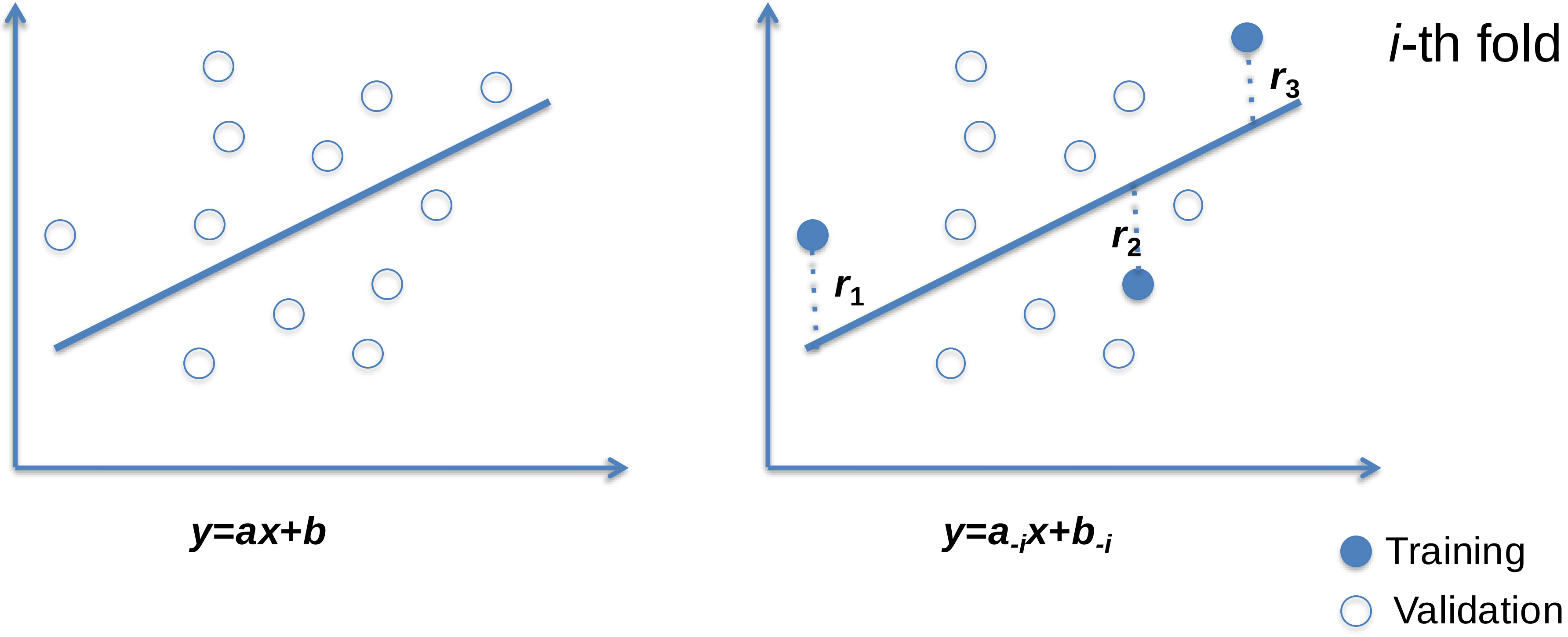
Leave-one-out cross-validation



Question: what is the optimal strategy for split?

Survey cross-validation strategy → project

Model Accuracy with k-fold cross-validation



Prediction error: the model accuracy is often measured in terms of the sum of squared residuals.

Bootstrap

In 1979, the statistician Brad Efron made an ingenious suggestion.

Most of what we know about the true probability distribution comes from the data. So let’s treat the data as a *proxy* for the true distribution.

We draw multiple samples from this proxy. This is called *resampling*. And compute the statistic of interest on each of the resulting pseudo-datasets.

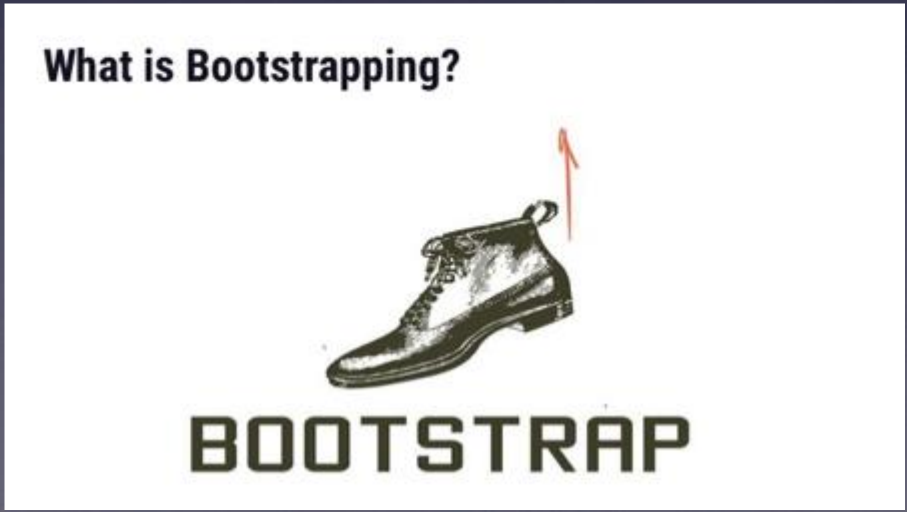
The start of modern statistics

computation vs. theory

“Bootstrapping has requires very little in the way of modeling, assumptions, or analysis, and can be applied in an automatic way to any situation, no matter how complicated”. *Efron and Gong 1983*

In ‘*Singular Travels, Campaigns and Adventures of Baron Munchausen*’ by R. E. Raspe (1786), the main character, finding himself in a deep hole, extracts himself using only the straps of his boots.

Pull yourself by pulling the bootstrap



Bootstrap resampling

- Introduced by Efron in 1979.
- Motivated by Jackknife.
- A "bootstrap" data set is one created by randomly selecting *n* points from the training set *D*, with replacement.
- In bootstrap estimation, this selection process is independently repeated *B* times to yield *B* bootstrap data sets.

Resampling with replacement

1   2 3 4

(1 1 1 1), (1 1 1 2), (1 1 1 3), ….., (4 4 4 3), (4 4 4 4)

(1 2 2 4) = (4 2 2 1)

$\binom{2n-1}{n} = \frac{(2n-1)!}{n!(n-1)!}$ possible resamples

Can be exhaustively large → Monte Carlo simulation

Problem: Prove the above combinations.

Proof: Resampling with replacement

1 2 3 4

(y_1, y_2, y_3, y_4)

(1 1 1 1) → (4, 0, 0, 0)

(1 2 2 4) → (1, 2, 0, 1)

(4 2 2 1) → (1, 2, 0, 1)

Equivalent representation

Method 1) The number of nonnegative solutions in $y_1+y_{2,}+y_3+y_4 = 4$

# of grid points on the hyperplane.

Method 2) Replace # with bar and , with +:

| | | | + + +

| + | | + + |

| + | | + + |

4 vertical lines + 3 plus signs

$$\binom{4+3}{4}$$

possible solutions

The Bootstrap

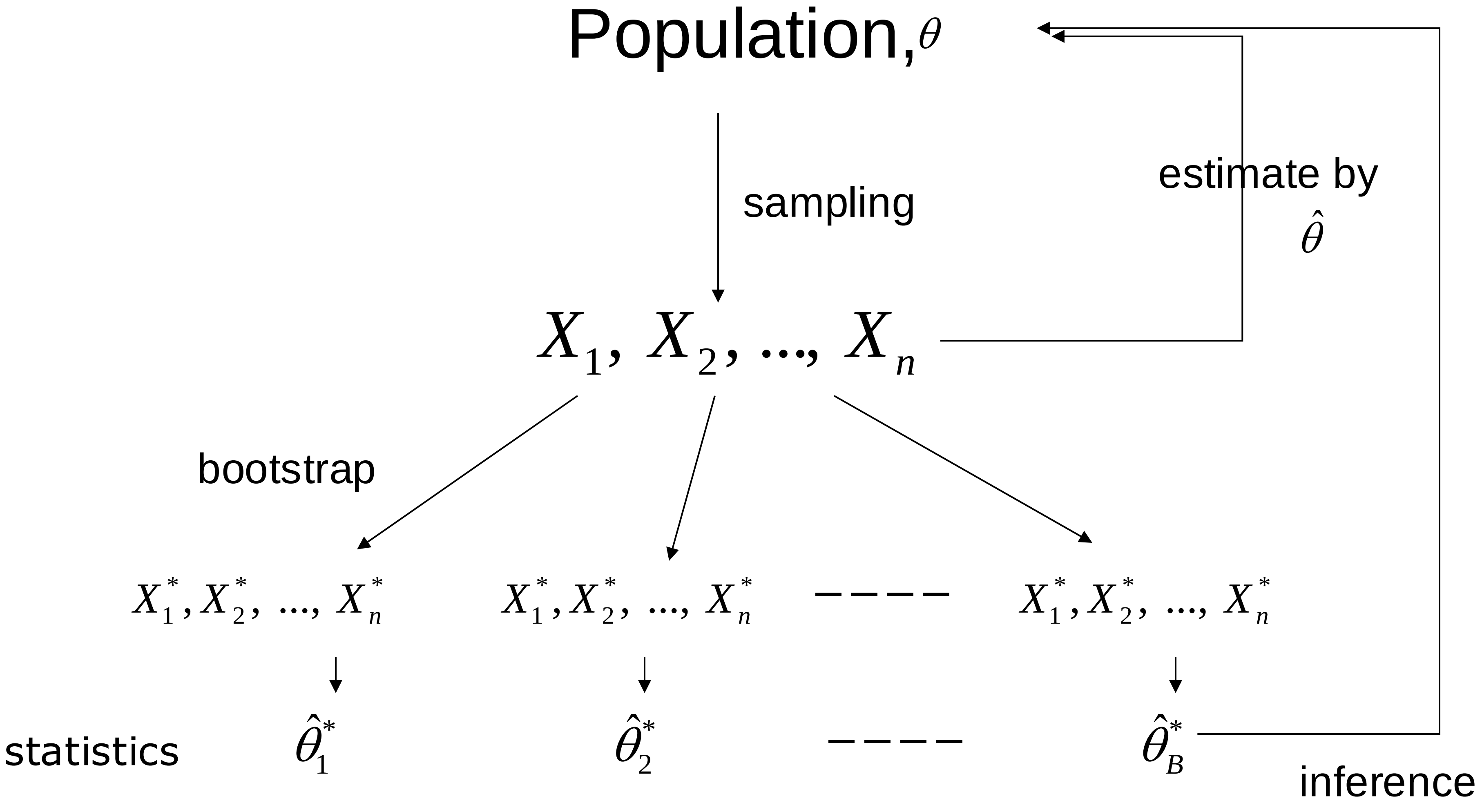
- Data **D** = X_1, X_2, X_3, ... ,X_n → statistic *s*

- Bootstrap replicate:
  - **D***1 = $X^*{}_1$, $X^*{}_2$, $X^*{}_3$, ... ,$X^*{}_n$ → statistic *s**1
  - **D***2 = $X^*{}_1$, $X^*{}_2$, $X^*{}_3$, ... ,$X^*{}_n$ → statistic *s**2
  - …

- $X^*{}_1$, $X^*{}_2$, $X^*{}_3$, ... ,$X^*{}_n$ are randomly selected with replacement, from X_1, X_2, X_3, ... ,X_n

- <u>Usually</u> use 1000-10,000 bootstrap replicates and obtain the empirical distribution

Schematic of Bootstrap



Population, θ

sampling

estimate by $\hat{\theta}$

$X_1, X_2, \ldots, X_n$

bootstrap

$X_1^*, X_2^*, \ldots, X_n^*$ $X_1^*, X_2^*, \ldots, X_n^*$ $----$ $X_1^*, X_2^*, \ldots, X_n^*$

statistics $\hat{\theta}_1^*$ $\hat{\theta}_2^*$ $----$ $\hat{\theta}_B^*$

inference

Example: *Median Gill Weight in Crabs*

Gill weights (in mg):

`159 179 100  45 384 230 100 320  80 220 320 210`

Median = 195mg

													Median
Real	**159**	**179**	**100**	**45**	**384**	**230**	**100**	**320**	**80**	**220**	**320**	**210**	**195**

Bootstrap replicates:

B1	**320**	**159**	**45**	**320**	**100**	**320**	**100**	**320**	**100**	**230**	**100**	**210**	**185**
B2	**384**	**384**	**45**	**384**	**45**	**384**	**100**	**80**	**45**	**179**	**230**	**230**	**205**
B3	**159**	**320**	**80**	**45**	**45**	**80**	**220**	**210**	**230**	**320**	**230**	**220**	**215**
B4	**220**	**179**	**384**	**100**	**80**	**100**	**230**	**230**	**179**	**230**	**384**	**45**	**200**
B5	**320**	**220**	**210**	**100**	**159**	**320**	**220**	**210**	**100**	**80**	**100**	**210**	**210**
B6	**80**	**100**	**230**	**100**	**210**	**384**	**159**	**220**	**320**	**45**	**45**	**210**	**185**
B7	**179**	**210**	**80**	**320**	**100**	**230**	**159**	**320**	**100**	**45**	**384**	**320**	**195**
B8	**384**	**159**	**100**	**159**	**100**	**179**	**100**	**179**	**220**	**384**	**220**	**159**	**169**
B9	**320**	**210**	**45**	**320**	**179**	**159**	**100**	**210**	**159**	**45**	**210**	**100**	**169**

. . .

Empirical distribution

Bootstrap properties

The bootstrap estimate of mean

$$\hat{\theta}^*_{(\bullet)} = \frac{\hat{\theta}^*_1 + \hat{\theta}^*_2 + \ldots + \hat{\theta}^*_B}{B},$$

$$\widehat{bias} = \hat{\theta}^*_{(\bullet)} - \hat{\theta} = \hat{\theta}^*_{(\bullet)} - \bar{X} \xrightarrow[B\to\infty]{} 0$$

Problem: Prove asymptotic unbiasness.

MATLAB function

```
boot=inline('x(unidrnd(length(x),m,length(x)))','x',
'm')

X=[5 8 3 2]

>>bs=boot(X,3)

bs =

     2     5     2     3
     5     2     2     3
     2     3     5     8
```

3 replicates

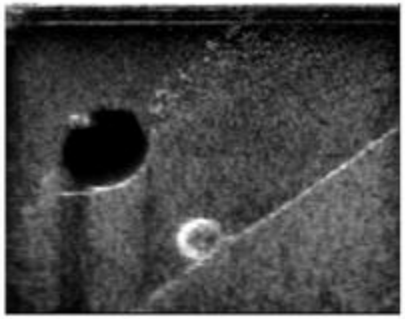
bootstat = bootstrp(nboot,bootfun,d1,...)
draws nboot bootstrap data samples, computes statistics on each sample using bootfun, and returns the results in the matrix bootstat.nboot

Limitation

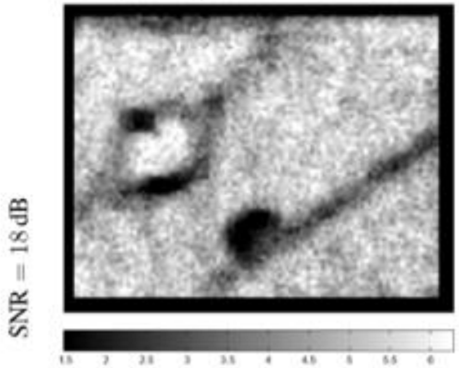
Bootstrap resampling assumes i.i.d. data. If observations are NOT independent and identically distributed, you need a different sampling strategy.

This is why bootstrap method is *rarely* used in images.

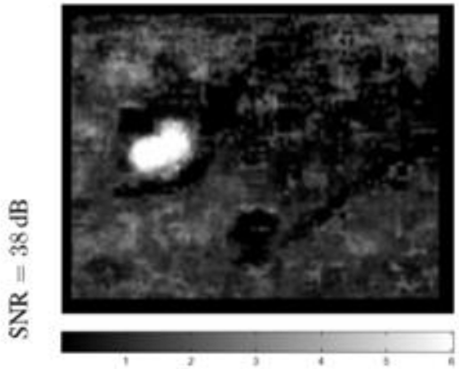
Block bootstrap on images



(a) input image



(d) bootstrap



(g) bootstrap

Can be used to estimate the performance of image registration when there is no ground truth. The residual of the model fit is block bootrapped.

Kybic, J. 2010 IEEE Trans. Image Processing 19:64-73

Question: This this a good validation method or not?

→Topology invariant resampling
Resampling is done in the spectral domain

Wang et al. 2018 Annals of Applied Stat.

References on bootstrap

Davison and Hinkley, Bootstrap Methods and their Application

Efron and Gong (1983) A Leisurely Look at the Bootstrap, American Statistician

Efron and Tibshirani (1986) Bootstrap Methods for Standard Errors, Statistical Science

Derrig, Ostaszewski, Rempala (2000) Applications of Resampling Methods in Actuarial Practice, PCAS

B. Efron (1979) Computers and the theory of statistics: thinking the unthinkable, SIAM Review, 21, 460-480.

B. Efron and R. J. Tibshirani (1993) An Introduction to the Bootstrap. Chapman & Hall.

J. I. De la Rosa and G. A. Fleury (2006) Bootstrap methods for a measurement estimation problem. IEEE Transactions on Instrumentation and Measurement, 55, 3, 820–827.