Permutation tests for Clinical Trials

Livio Finos

Outline

Introduction

- Well established nonparametric approach to **inference**: Fisher, 1935; Pitman, 1937; Pitman, 1938.

- (In general) it requires less assumptions about the data generating process than the parametric counterpart.

- Very good inferential properties, typically:
  - exactness (i.e. exact control of the type I error)
  - asymptotically optimality and convergence to the parametric counterpart when it does exist.

Introduction

○ Fisher exact test is a prototypical example, but

○ the general approach has restricted applicability without the support of a computer.

Renewed interest toward permutation testing

○ A milestone: Westfall and Young (1993). Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment. Wiley.

○ Many actives areas of research adopt these methods in their daily statistical analysis (e.g. genetics and neuroscience: Nichols and Holmes (2002); Pantazis et al. (2009); Winkler et al. (2014)).

○ Permutation approach:
  ○ Ideal for **randomized experimental design**
  ○ deals with very complex models, without formal definition of the data generating process.

Outline

A Naive approach to Permutation Testing

Comparison of Two Samples (i.e. one factor with two levels)

- **Control**: 3 observations,
- Treated: 3 observations:

$$1.025, 1.949, 3.477, 2.391, 3.676, 4.816$$

Hypothesis testing

- H_0: Two groups are equal
- H_1: Treated is greater than Control (on average)

p-value: probability to get the observed evidence against H_0 if the two groups were equal (i.e. H_0 were true)

Test: if $p \leq \alpha$ (e.g. $\alpha = .05$): we decide for H_1,
    otherwise: we stay with H_0

Parametric approach

Assumptions on $y_1, y_2, \ldots, y_6$

- independent

- identically distributed

    - normally distributed
      OR
    - have finite mean and variance (but inference is only asymptotically valid in this case!)

Parametric approach

Assumptions on $y_1, y_2, \ldots, y_6$

- independent

- identically distributed

    - normally distributed
      OR
    - have finite mean and variance (but inference is only asymptotically valid in this case!)

We can perform a t-test:

$$T = \frac{\bar{y}(Treated) - \bar{y}(Control)}{\widehat{sd(\bar{y}(Treated) - \bar{y}(Control))}} \sim t_4$$

(i.e. T test statistic follow a t distribution with n-2=4 d.f.)

Parametric approach

With toy data:

```
t = -1.4545,
df = 4,
p-value = 0.1098
```

Remark

The hypotheses tested are:

- $H_0 : \mu_{Treated} = \mu_{Control}$
- $H_1 : \mu_{Treated} > \mu_{Control}$ (only a difference in mean is allowed)

Outline

A Naive approach to Permutation Testing

The p-value is computed **under** H_0:
Controls and Treated have the **same distribution**.

Collection of equally likely outcomes:

$$f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) =$$

A Naive approach to Permutation Testing

The p-value is computed **under** H_0:
Controls and Treated have the **same distribution**.

Collection of equally likely outcomes:

$$f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) =$$
$$= f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) =$$

A Naive approach to Permutation Testing

The p-value is computed **under** H_0:
Controls and Treated have the **same distribution**.

Collection of equally likely outcomes:

$$
\begin{aligned}
& f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) = \\
= \; & f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) = \\
= \; & f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) =
\end{aligned}
$$

A Naive approach to Permutation Testing

The p-value is computed **under** H_0:
Controls and Treated have the **same distribution**.

Collection of equally likely outcomes:

$$f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) =$$
$$= \ f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) =$$
$$= \ f(1.025, 1.949, 3.477, 2.391, 3.676, 4.816) =$$
$$= \ \ldots$$

There are $\binom{6}{3} = \frac{6!}{3!3!} = 20$ equally likely outcomes

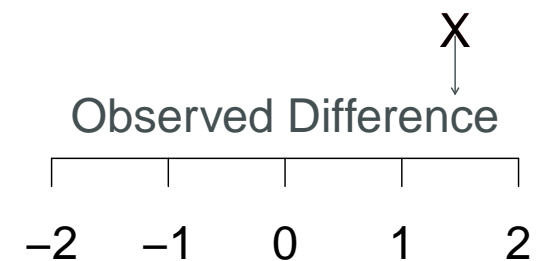# A Naive approach to Permutation Testing
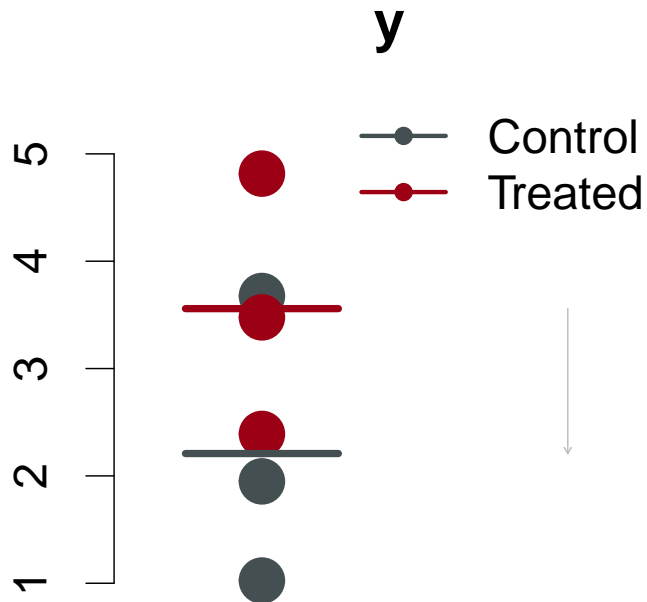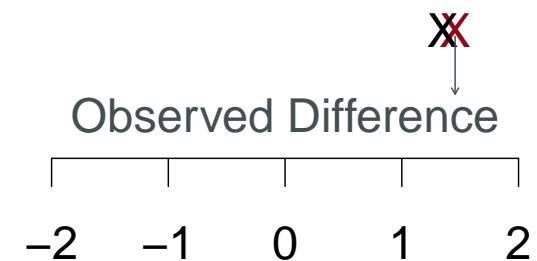
Compute the difference in mean of the two samples
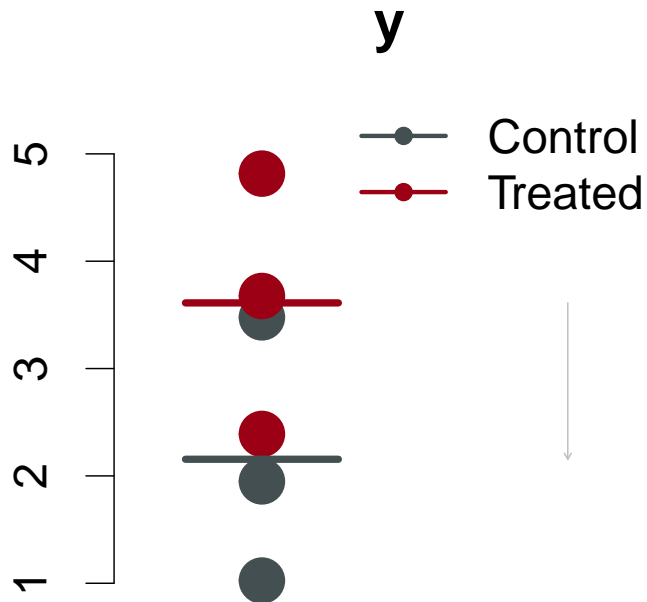
A Naive approach to Permutation Testing

Compute the same difference on another hypothetical experiment

y

Control
Treated

$T = \bar{y}(\text{Treated}) - \bar{y}(\text{Control})$

$T = 1.35$

X X

Observed Difference

−2   −1   0   1   2

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

y

$T = \bar{y}(\text{Treated}) - \bar{y}(\text{Control})$

$T = 0.6$

X   XX

Observed Difference

−2   −1   0   1   2

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

y

Control
Treated

$T=\bar{y}(\text{Treated})-\bar{y}(\text{Control})$

$T=2.18$

X X X X

Observed Difference
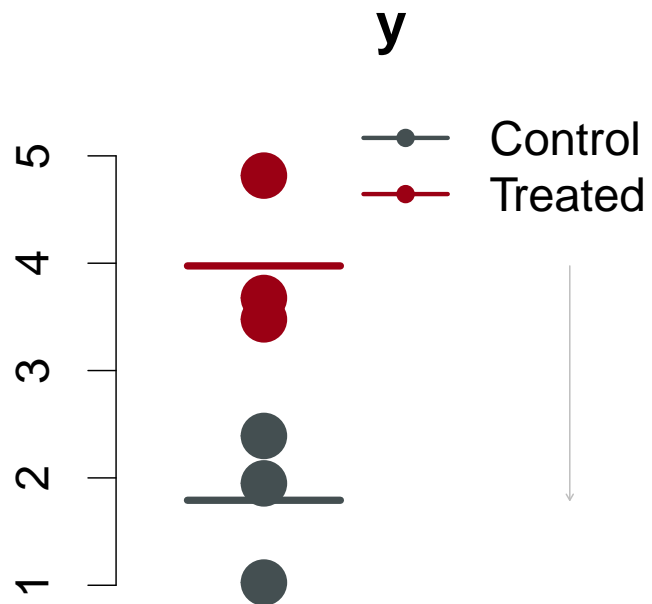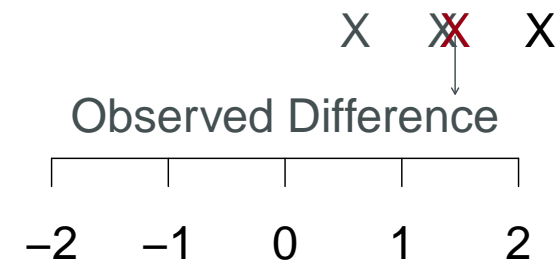
−2   −1   0   1   2

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

A Naive approach to Permutation Testing

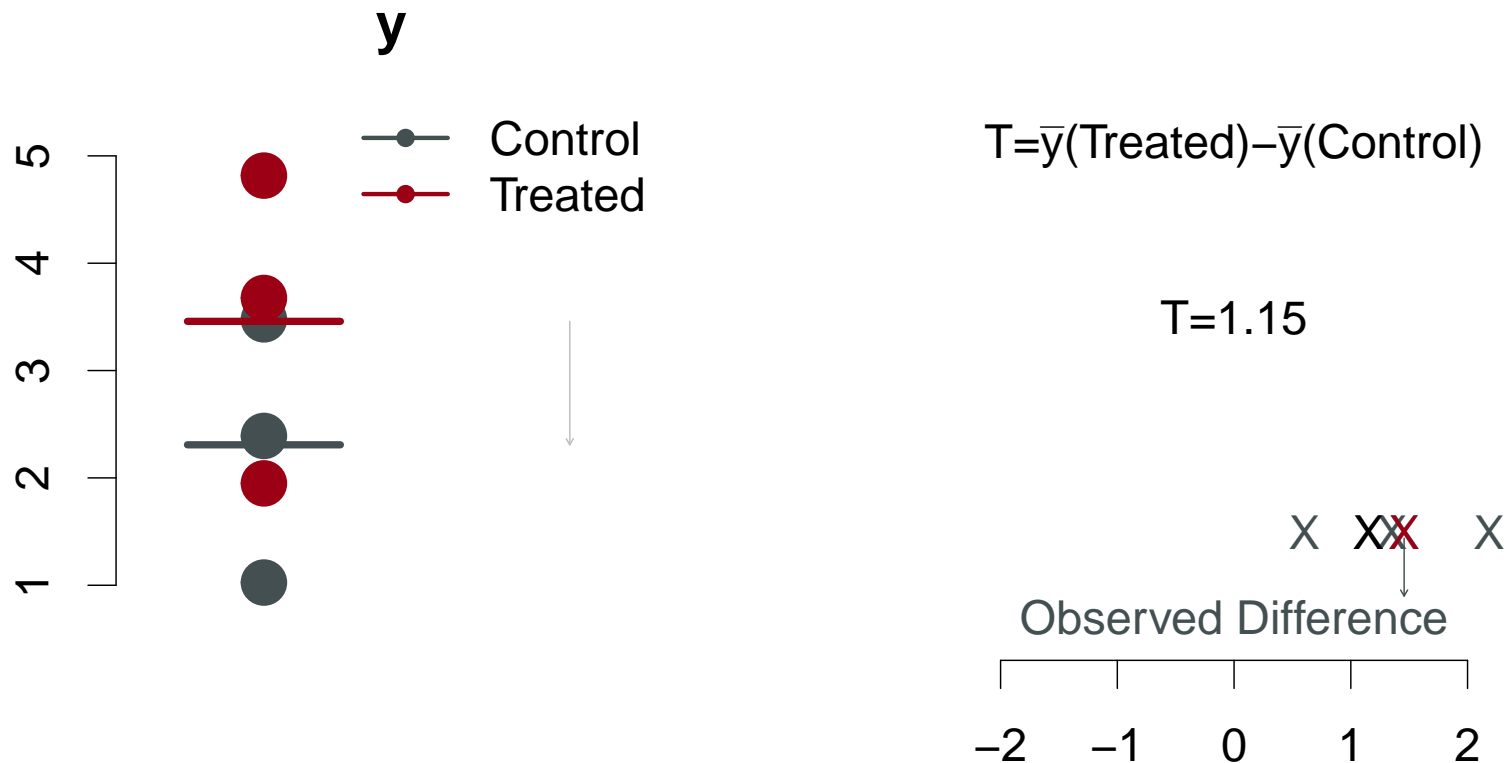. . . and go on with all hypothetical experiments. . .

y

Control
Treated

$T=\bar{y}(\text{Treated})-\bar{y}(\text{Control})$

$T=0.32$

XX XX X X

Observed Difference

−2   −1   0   1   2

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

y

Control
Treated

$T=\bar{y}(\text{Treated})-\bar{y}(\text{Control})$

$T=-0.43$

X   XX   XXX   X
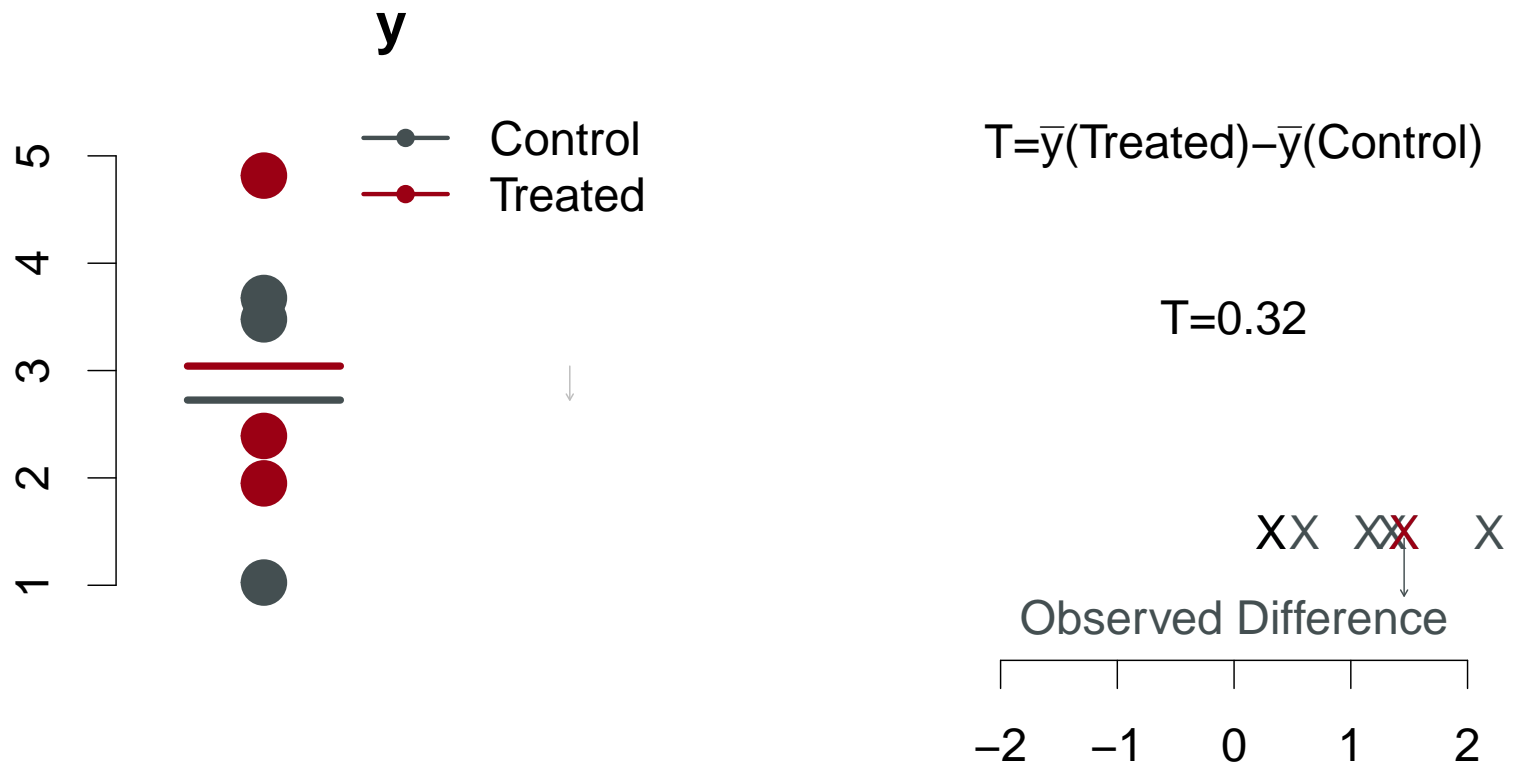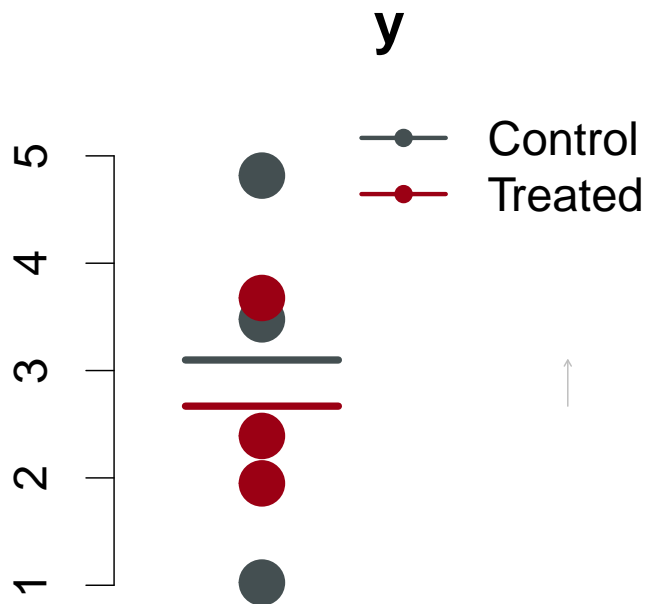
Observed Difference

−2   −1   0   1   2

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

A Naive approach to Permutation Testing

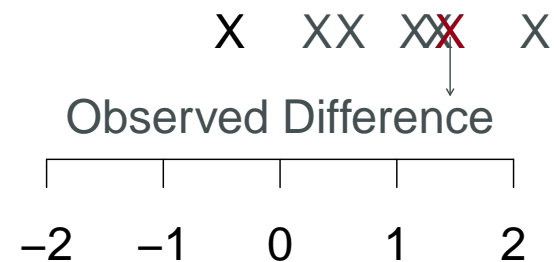. . . and go on with all hypothetical experiments. . .

y

Control
Treated

$T = \bar{y}(\text{Treated}) - \bar{y}(\text{Control})$

$T = 0.3$

X **X** X XXX X

Observed Difference

−2   −1   0   1   2

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

A Naive approach to Permutation Testing

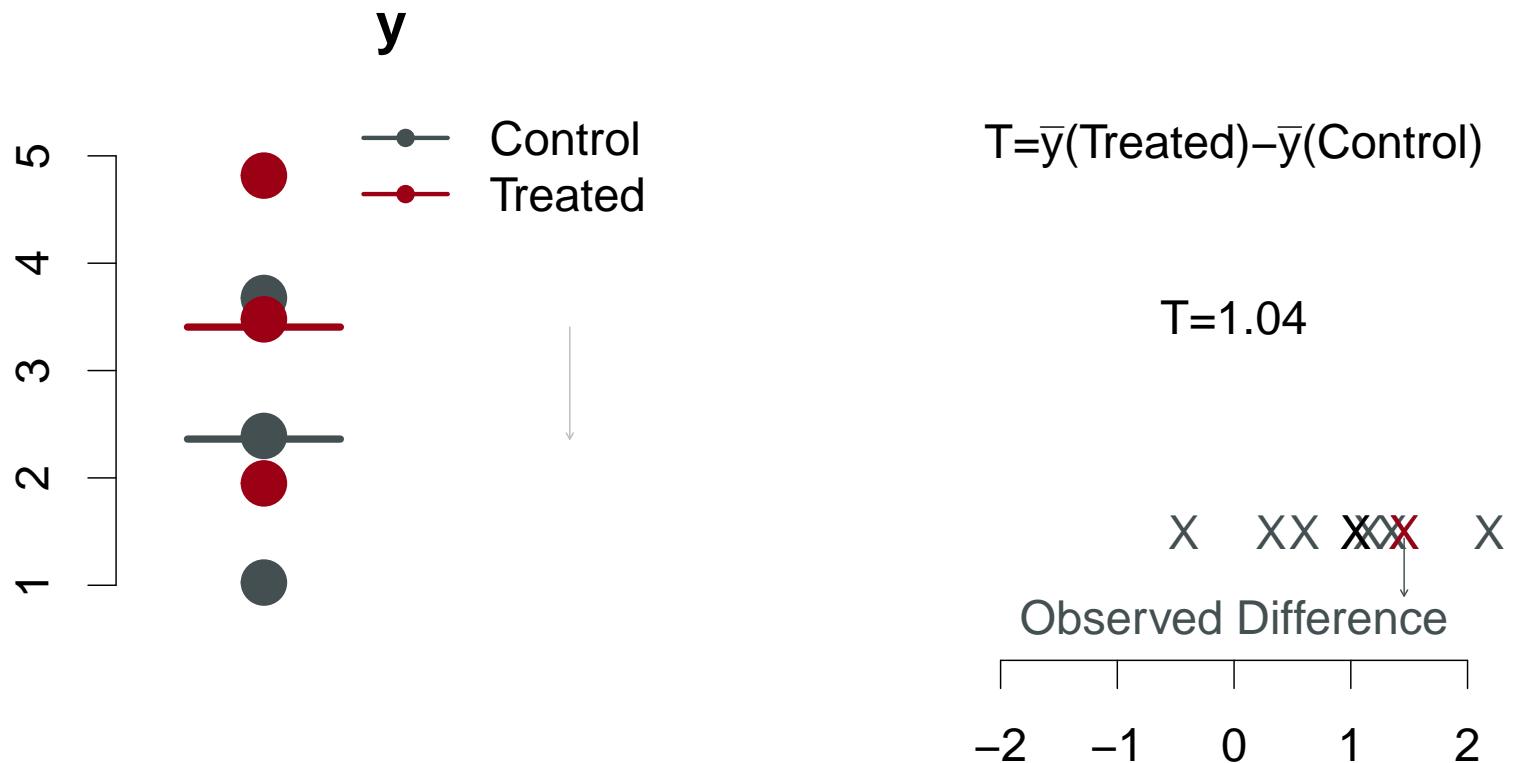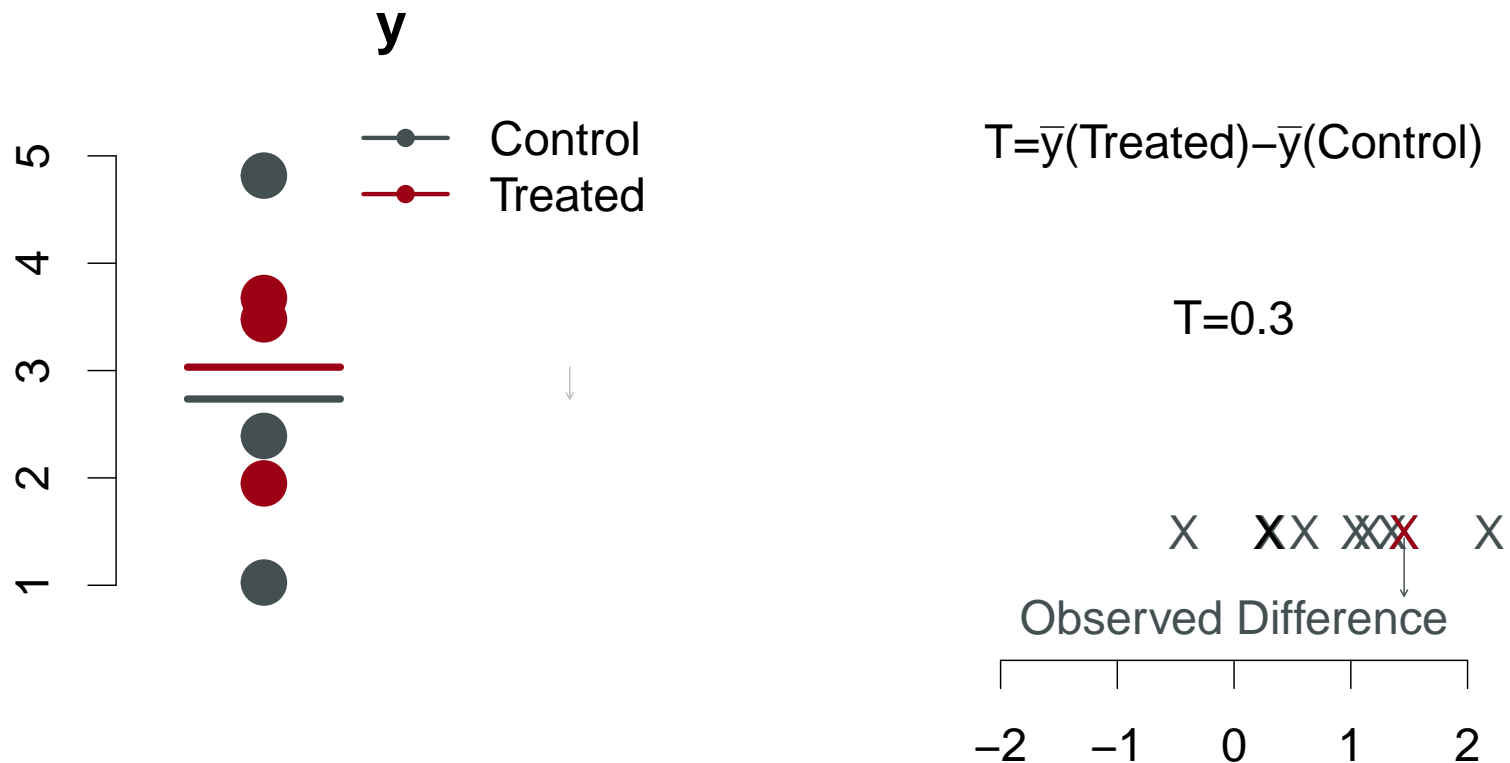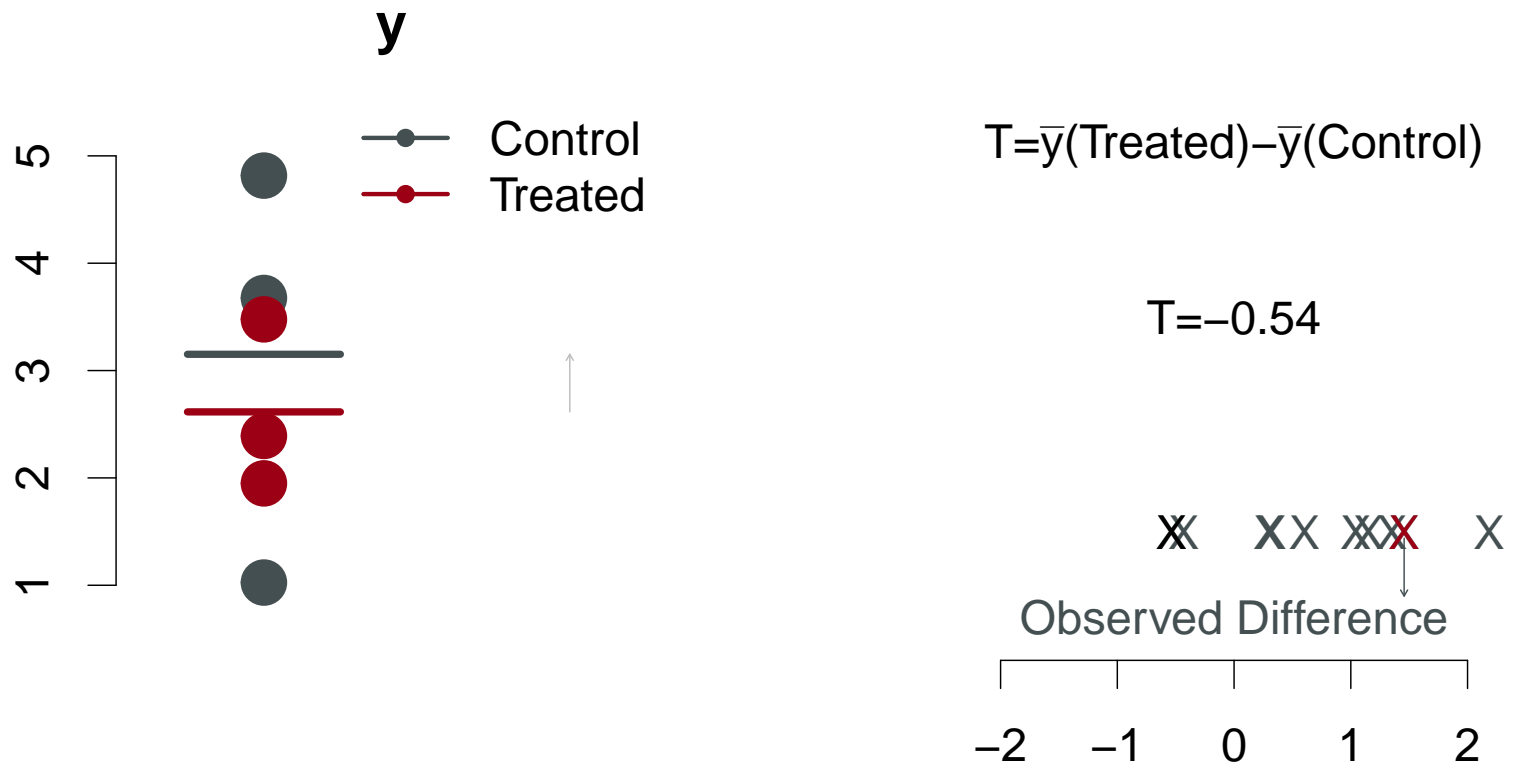. . . and go on with all hypothetical experiments. . .

y

Control
Treated

$T = \bar{y}(\text{Treated}) - \bar{y}(\text{Control})$

T=0.54

Observed Difference

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

y

$T=\bar{y}(\text{Treated})-\bar{y}(\text{Control})$

$T=-0.3$

Observed Difference

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

y

Control
Treated

$T=\bar{y}(\text{Treated})-\bar{y}(\text{Control})$

$T=-1.04$

X XXX XX XXX X X

Observed Difference

−2   −1   0   1   2

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

y

$T = \bar{y}(\text{Treated}) - \bar{y}(\text{Control})$

$T = 0.43$

Observed Difference

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

y

$T=\bar{y}(Treated)-\bar{y}(Control)$

$T=-0.32$

Observed Difference

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

y

Control
Treated

$T=\bar{y}(\text{Treated})-\bar{y}(\text{Control})$
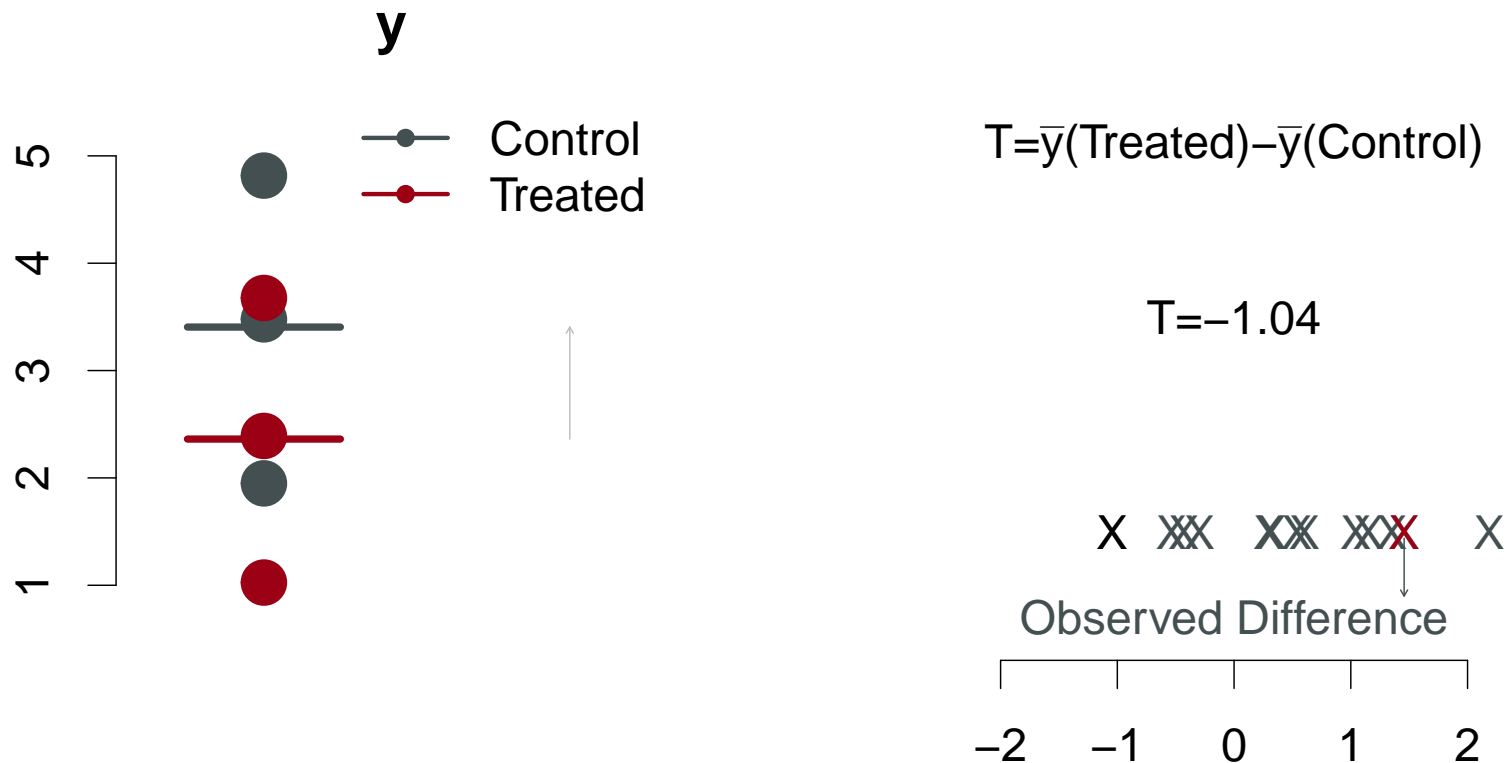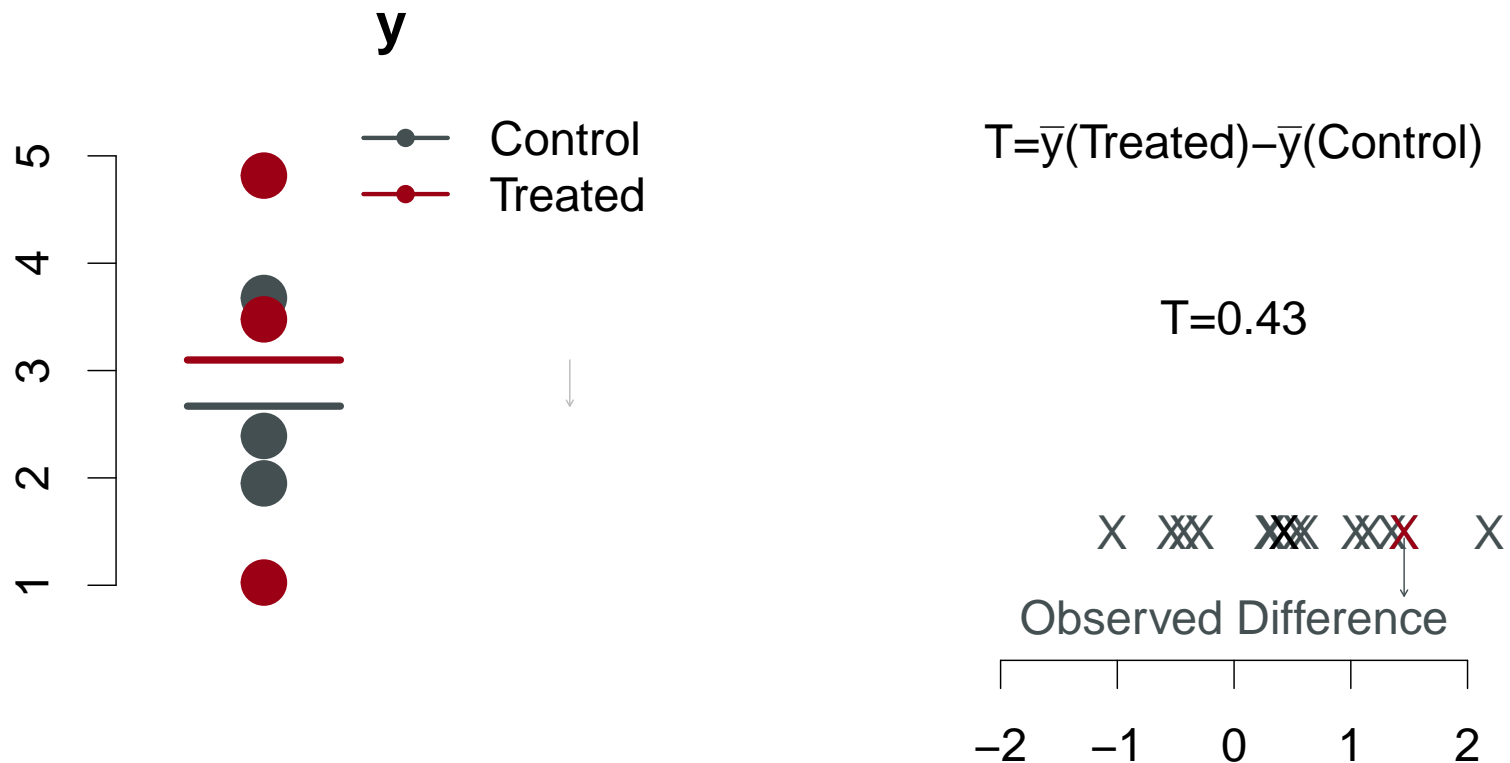
$T=-1.15$

Observed Difference
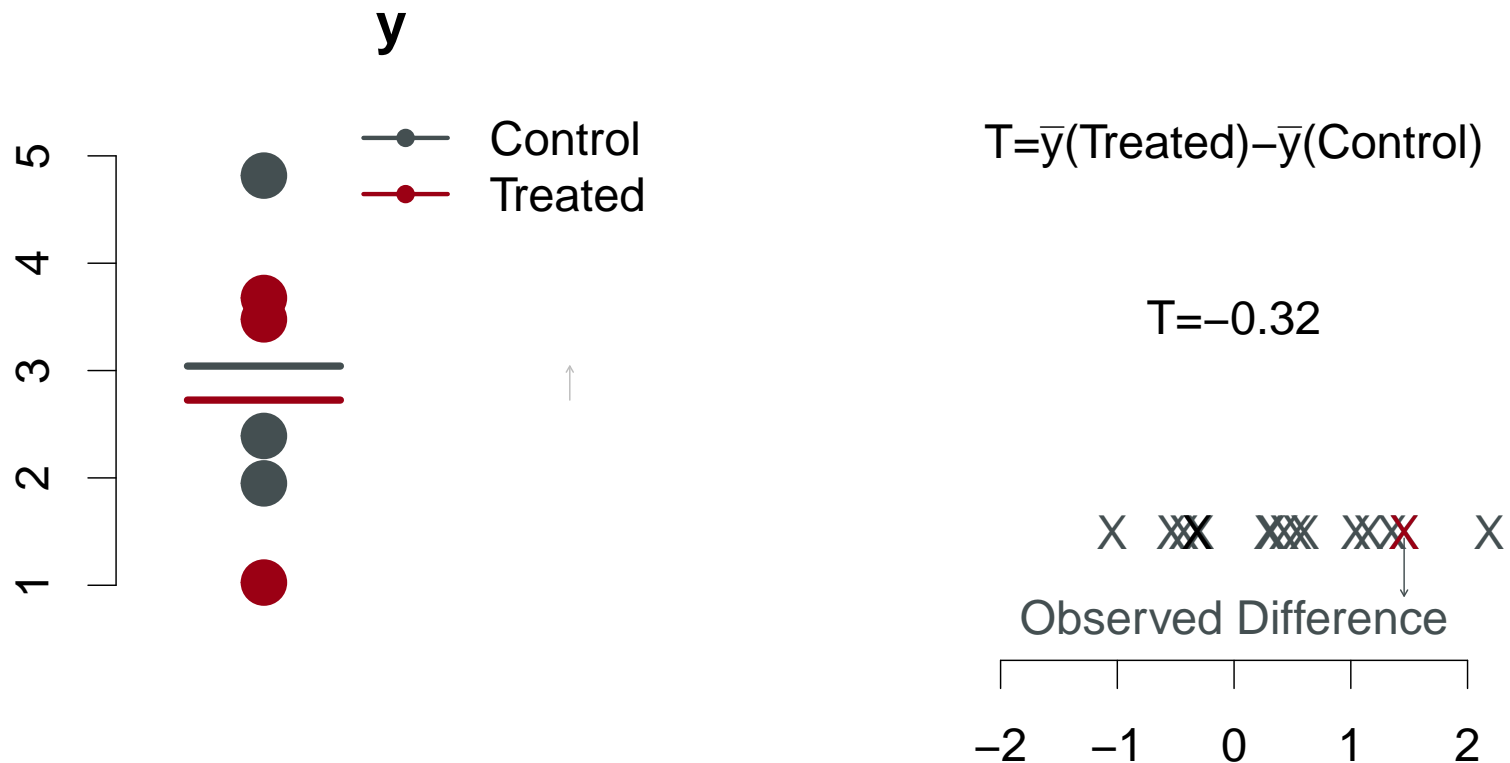
−2   −1   0   1   2

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .

A Naive approach to Permutation Testing
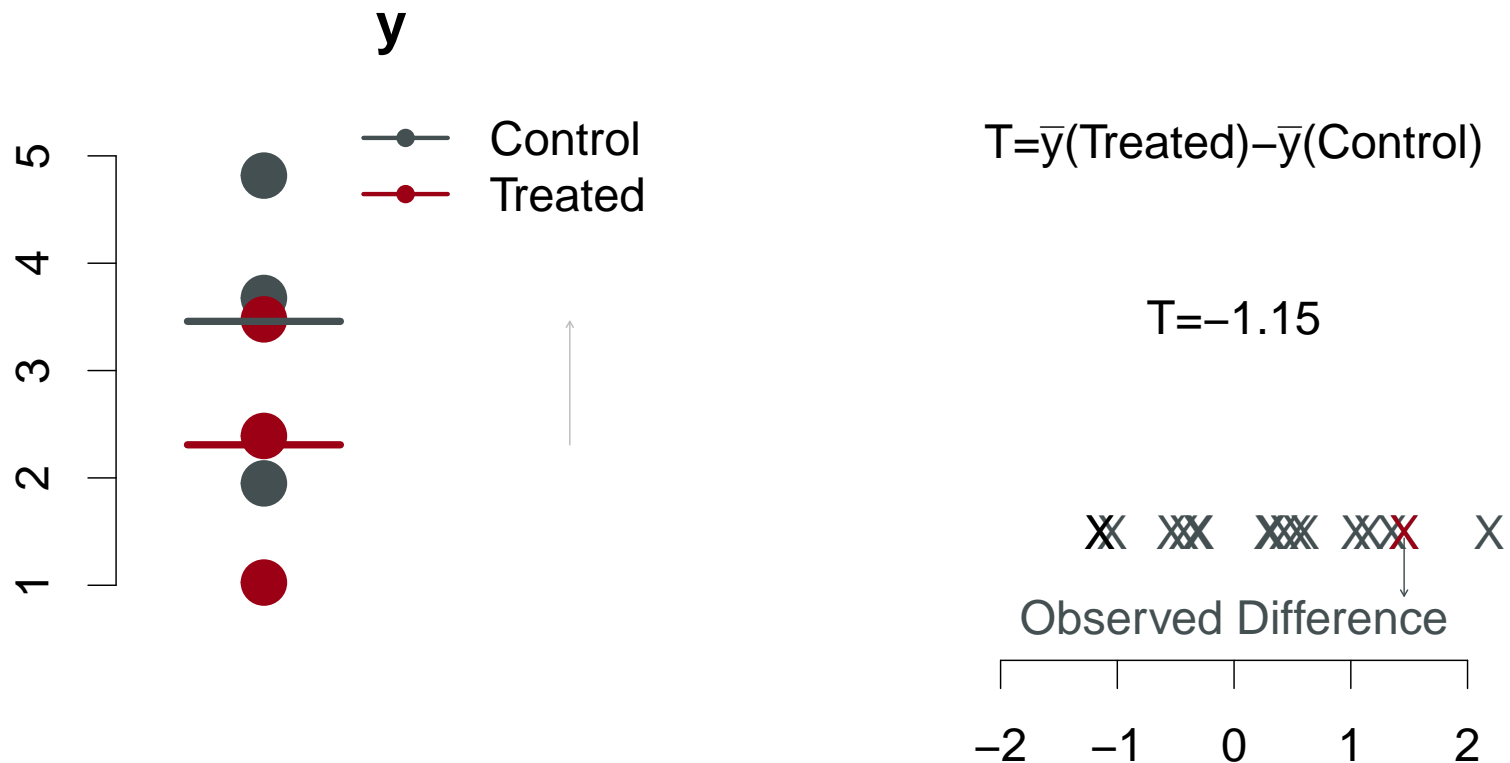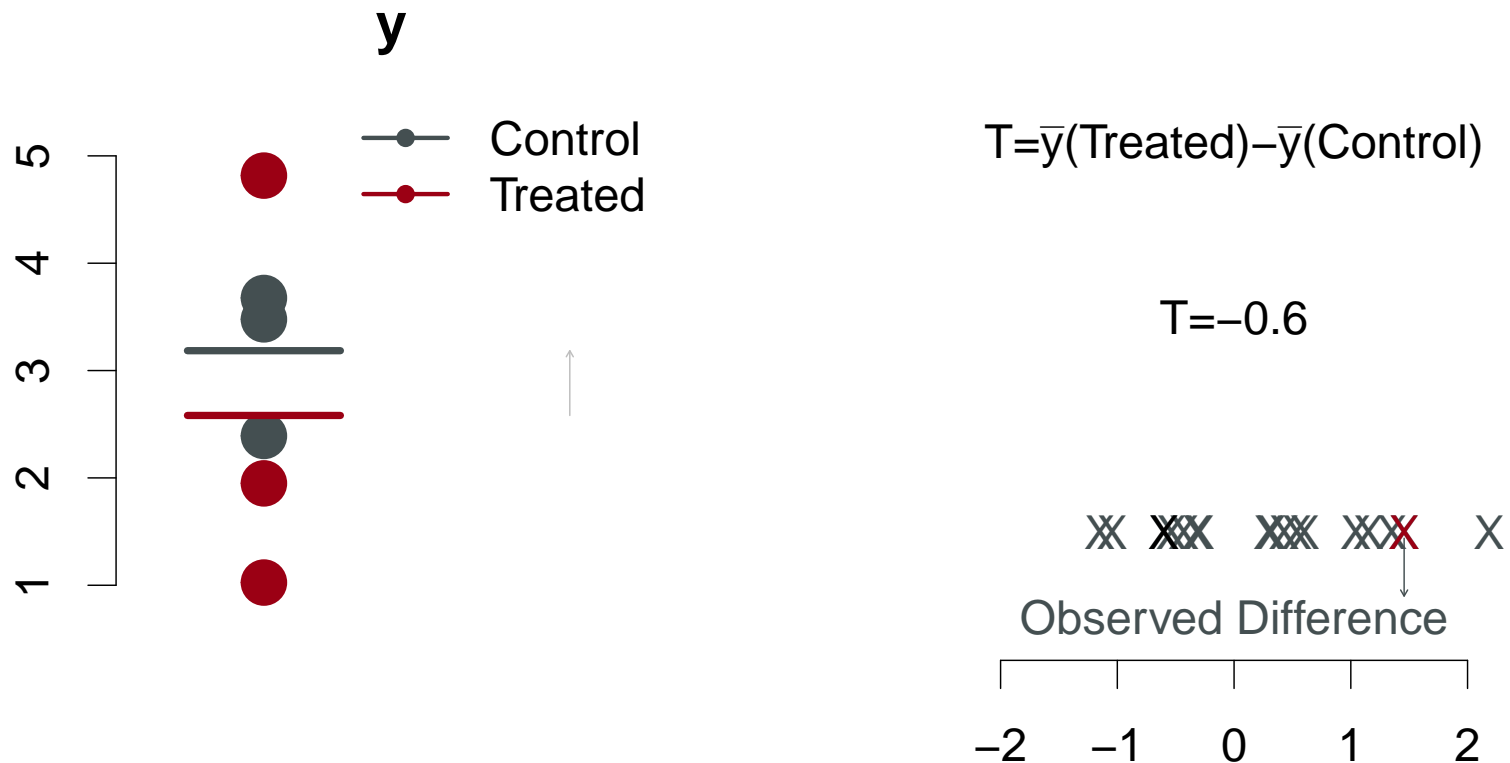
. . . and go on with all hypothetical experiments. . .

$T = \bar{y}(\text{Treated}) - \bar{y}(\text{Control})$

$T = -1.35$

Observed Difference

A Naive approach to Permutation Testing

. . . and go on with all hypothetical experiments. . .
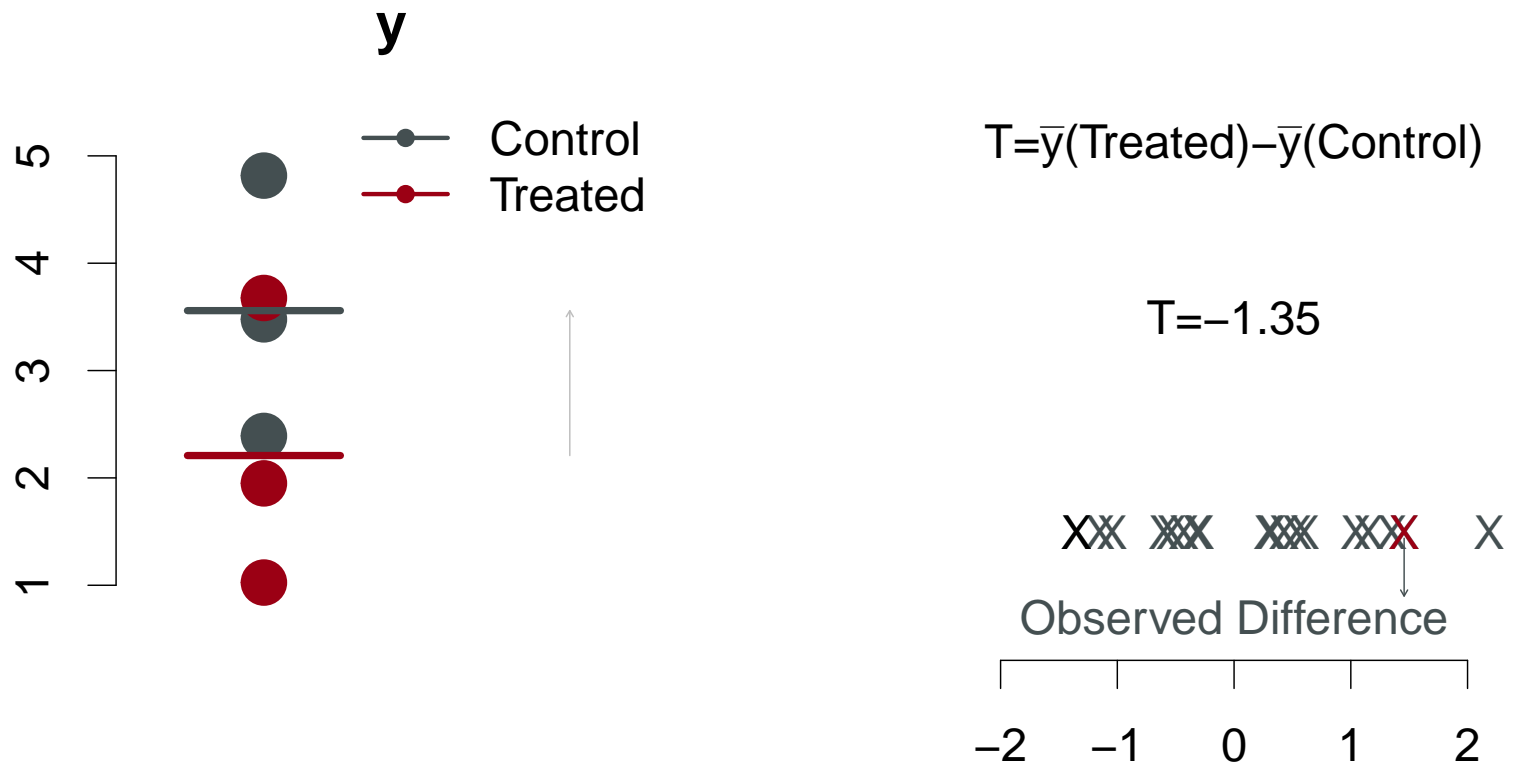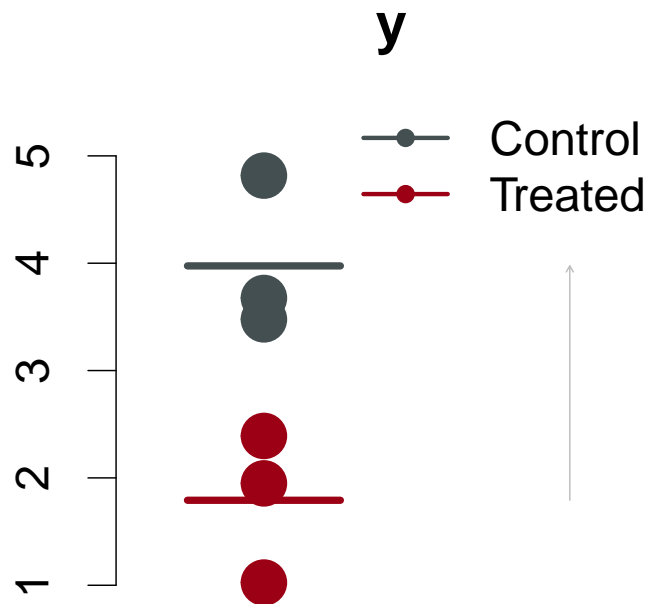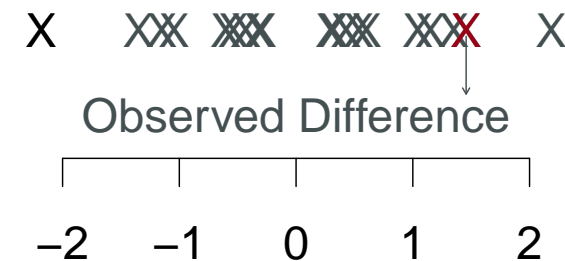
A Naive approach to Permutation Testing

. . . and compute the p-value!

y
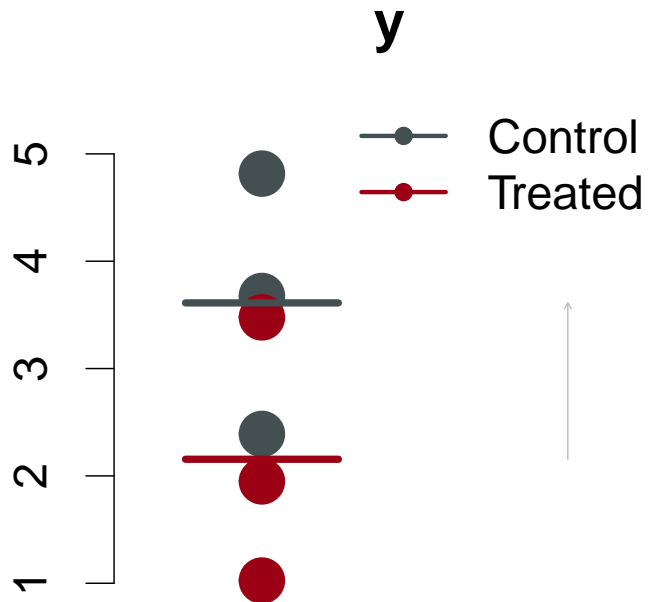
Control
Treated

$T = \bar{y}(\text{Treated}) - \bar{y}(\text{Control})$

p−value = 2/20 = 0.10

T=−1.46

Observed Difference

−2   −1   0   1   2

Summary

The Permutation Test:

- Conditioned to observed data (i.e. the distribution of the test statistic depends on the data).

- Under H_0 cases and controls have the same distribution (eg. they have the same probability to get hight values),

- explore all possible experiments that we can observe with the data (ie. exchanging cases and controls),

- compute the p-value as the proportion of experiments providing equal or more evidence against H_0 with respect to observed data.

Outline

A (slight) more formal approach

(see also Pesarin, 2001)

$\mathbf{y} = (y_1, y_2, \ldots, y_n)$ the vector of observed data

Orbit: the set of all samples having the same likelihood under H_0.

$$\mathcal{O} = \{\mathbf{y}^* : \ f(\mathbf{y}^*) = f(\mathbf{y})\}$$

(and $|\mathcal{O}|$ number of elements of $\mathcal{O}$)

If we assume exchangeability of observations, then:

$$\mathcal{O} = \{\text{all permutations of the observed data } \mathbf{y}\} = \{\mathbf{y}^* : \pi^* \circ \mathbf{y}\}$$

$(\pi^* \in \Pi, \ \Pi \text{ set of all possible permutations})$

A (slight) more formal approach

Remark: exchangeable observations: $f(y_1, y_2) = f(y_2, y_1)$.
It implies observations:

- are **identically distributed**
  t-test and linear models assume normality, only asymptotic control of the tye I error

- have the **same dependence**
  t-test and linear models assume independence, which is just a special case, i.e. more stringend assumptions

A (slight) more formal approach

p-value: proportion of experiments providing equal or more evidence against H_0 with respect to observed data.

To compute it, we need the Orbit $\mathcal{O}$ and a

A (slight) more formal approach

p-value: proportion of experiments providing equal or more
evidence against H_0 with respect to observed data.

To compute it, we need the Orbit $\mathcal{O}$ and a
Test statistic (T: $\mathbb{R}^n \to \mathbb{R}$) quantifies the evidence against H_0

- higher values provide more evidence against H_0

- compute a test statistic for each element of the Orbit $\mathcal{O}$,
  this induces an ordering on $\mathcal{O}$.

In our example: $T = \bar{y}(Treated) - \bar{y}(Control)$ is the difference in
mean, higher the difference, higher the evidence for H_1.

A (slight) more formal approach

$$f(\mathbf{y}^*|\mathcal{O}) = \frac{f(\mathbf{y}^* \cap \mathcal{O})}{f(\mathcal{O})} = \frac{f(\mathbf{y}^*)}{f(\mathcal{O})} = \frac{f(\mathbf{y}^*)}{f(\cup_{y \in \mathcal{O}} y)} = \frac{1}{|\mathcal{O}|} \; \forall \; \mathbf{y}^* \in \mathcal{O}$$

i.e. each permutation is equally likely in the Orbit $\mathcal{O}$.

The p-value:

$$P(T(\mathbf{y}^*) \geq T(\mathbf{y})|\mathbf{y}^* \in \mathcal{O}, H_0) =$$

$$= \int_{T(\mathbf{y})}^{+\infty} f(T(\mathbf{y}^*))dT(\mathbf{y}^*) =$$

$$= \sum_{\mathbf{y}^* \in \mathcal{O}} I(T(\mathbf{y}^*) \geq T(\mathbf{y}))/|\mathcal{O}| \quad \forall \mathcal{O}$$

The package flip

It is on CRAN and on github (https://github.com/livioivil/flip)

To install the github version type (in R):

```
library(devtools)
install_github('livioivil/flip')
```

Seeds data (Pesarin, 2001)

Standard fertilizer ($grp = 0$) vs New fertilizer ($grp = 1$)
Total weight of the plant and average leaves length is recorded.
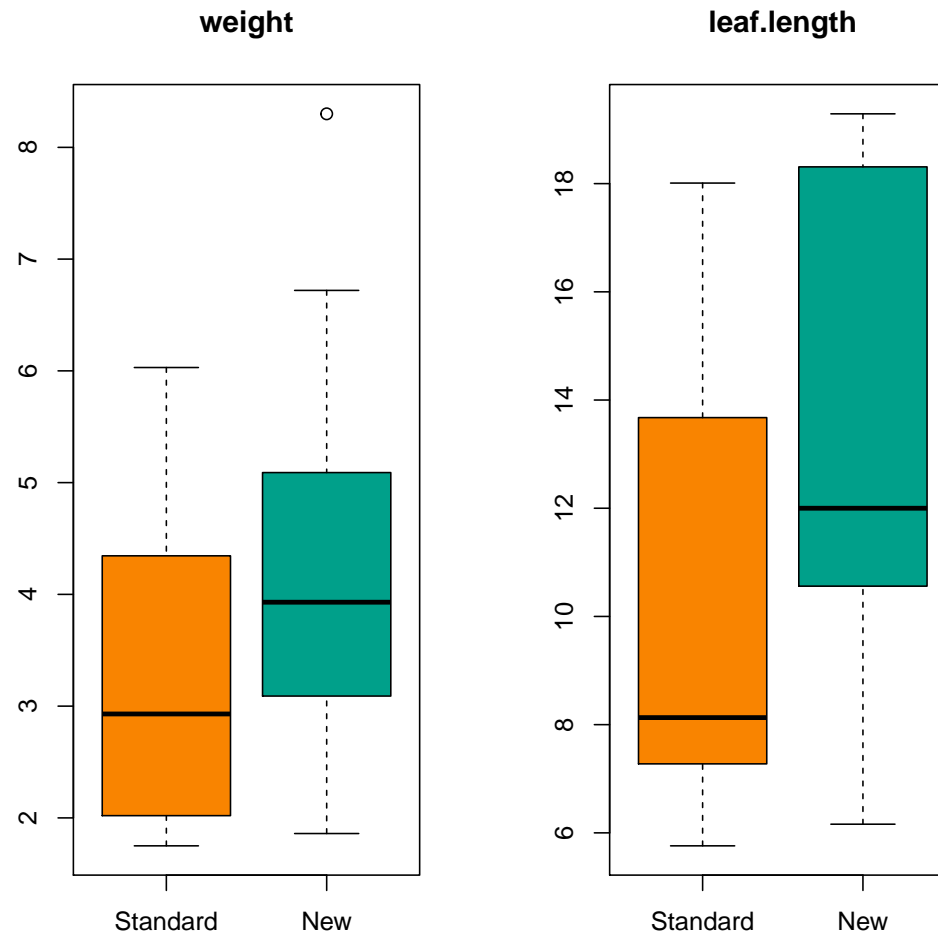
Hypothesis testing

About weight:

- $H_0(weight) : F(weight|grp = 0) = F(weight|grp = 1)$

  vs

- $H_1(weight) : F(weight|grp = 0) > F(weight|grp = 1)$

And about leaf length:

- $H_0(leaf.len) : F(leaf.len|grp = 0) = F(leaf.len|grp = 1)$

  vs

- $H_1(leaf.len) : F(leaf.len|grp = 0) > F(leaf.len|grp = 1)$

Hypothesis testing

```
res=flip(.~ grp, data=seeds, tail=1)
```

	Test	Stat	tail	p-value
weight	t	1.320	>	0.098
leaf.length	t	2.061	>	0.030

together with some visualization `hist(res)`

weight

Two-tailed tests

About weight:

- $H_0(weight) : F(weight|grp = 0) = F(weight|grp = 1)$

  vs

- $H_1(weight) : F(weight|grp = 0) \neq F(weight|grp = 1)$

And about leaf length:

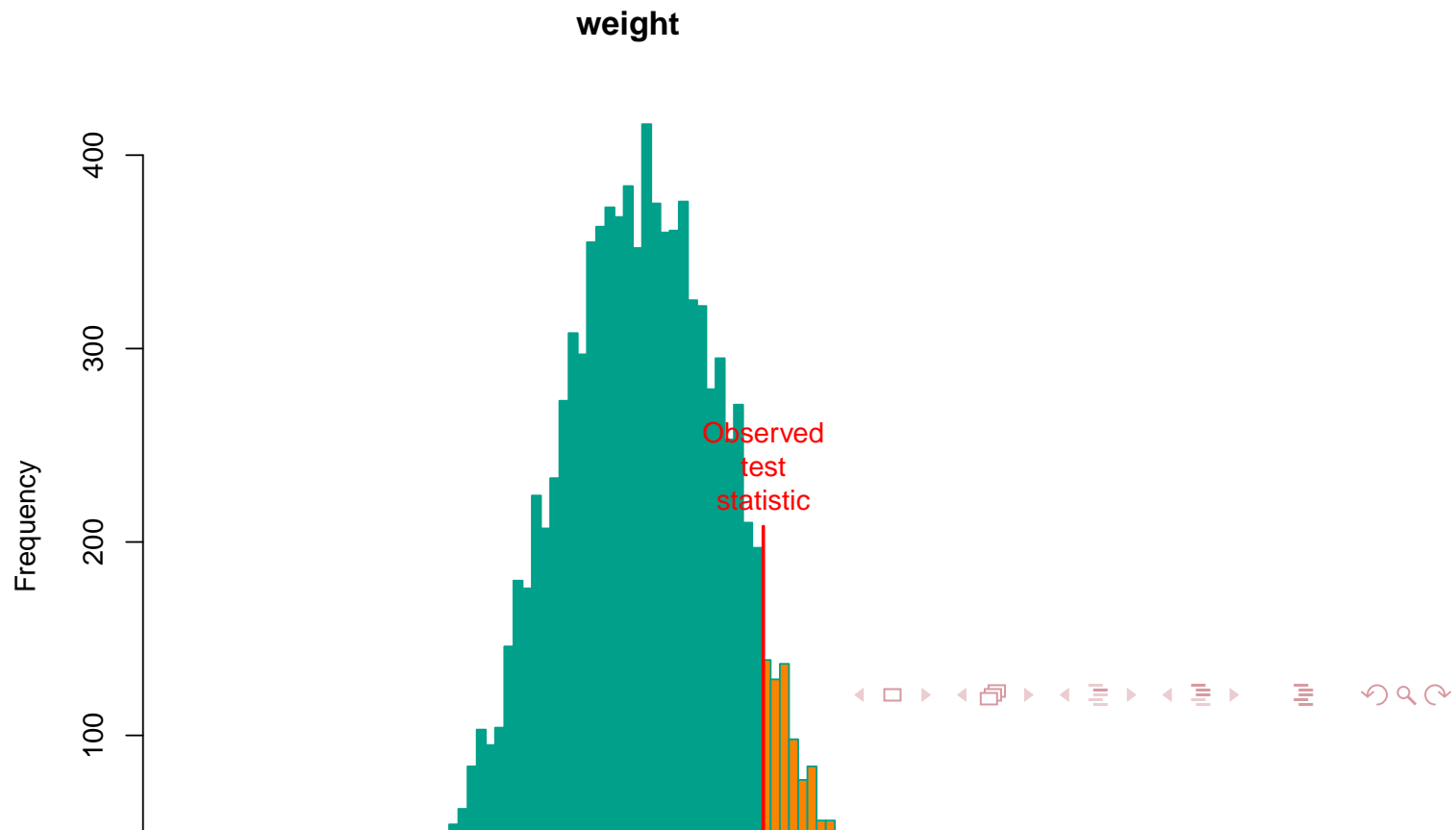- $H_0(leaf.len) : F(leaf.len|grp = 0) = F(leaf.len|grp = 1)$

  vs
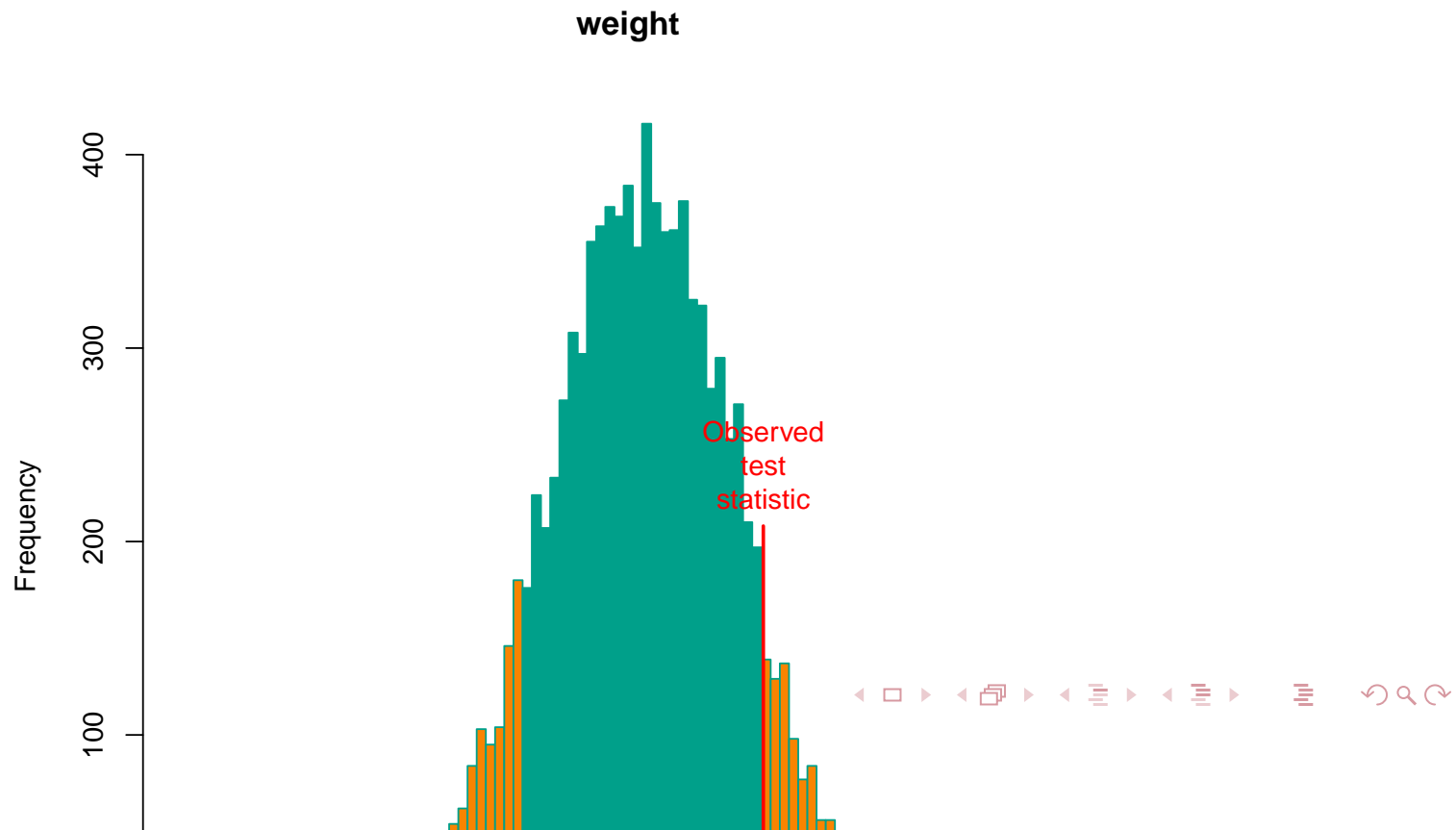
- $H_1(leaf.len) : F(leaf.len|grp = 0) \neq F(leaf.len|grp = 1)$

Two-tailed tests

```
res=flip(.~ grp, data=seeds, tail=0)
```

	Test	Stat	tail	p-value
weight	t	1.320	><	0.202
leaf.length	t	2.061	><	0.049

Also very negative values provide evidence against H_0

weight

Properties

- Exact control of the Type I Error:
  $P(p \leq \alpha | H_0) < \alpha \ \forall \ (\text{attainable}) \alpha$

- Consistency: $P(p \leq \alpha | H_1) \to 1$ when $n \to \infty$

- Converges to parametric counterpart (i.e. asymptotic optimality if the parametric test is optimal)

Remark

The number of possible permutations (size of the Orbit $|\mathcal{O}|$) is often huge, we can not compute the test statistic for all elements. Common use to sample from the Orbit (i.e. randomly permute the labels B times).
The properties remain the same.

Simulation: normal distribution

- Comparison of Two groups (labels A, B) of size 5
- $y_i \sim N(0, 1)$
    - $H_0 : f(y|grp = A) = f(y|grp = B)$
    - $H_0 : f(y|grp = A) \neq f(y|grp = B)$
      (i.e. two-sided alternatives)
- 10000 replications
- 1000 random permutations for each test
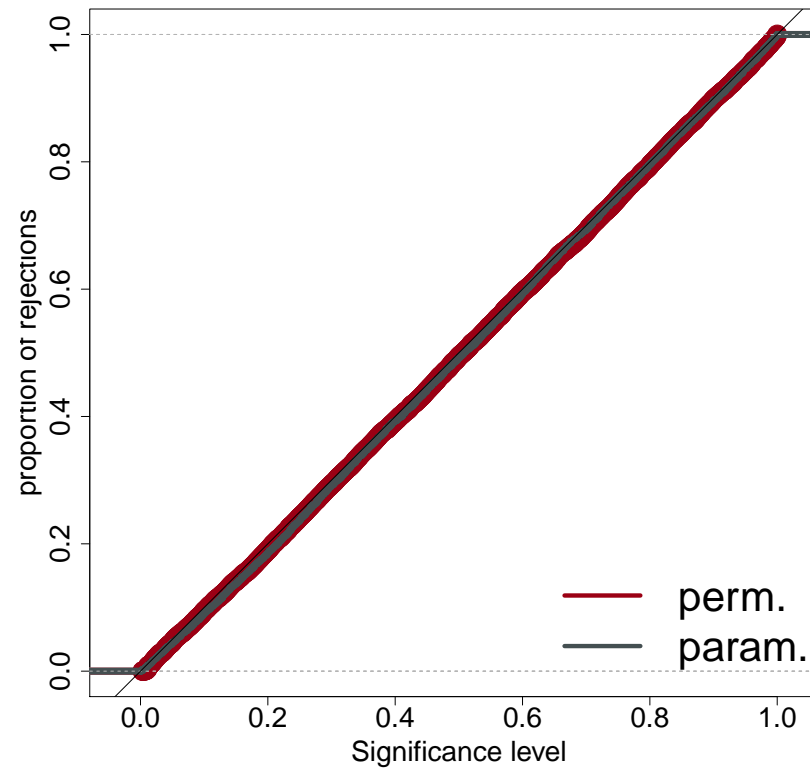
Simulation: H_0

Empirical Type I error	≤ 0.01	≤ 0.05	≤ 0.1	≤ 0.5	≤ 0.75
Permutation	0.00	0.05	0.09	0.49	0.74
Paramatric t.test	0.01	0.04	0.09	0.49	0.75

Simulation: H_1

now $(y|grp = A) \sim N(0,1)$, $(y|grp = B) \sim N(2,1)$

Empirical Power	≤ 0.01	≤ 0.05	≤ 0.1	≤ 0.5	≤ 0.75
Permutation	0.40	0.78	0.89	0.99	1.00
Paramatric t.test	0.41	0.77	0.88	0.99	1.00

Simulation: Cauchy distribution

- Comparison of Two groups (labels A, B)

- $y_i \sim Cauchy$
  - $H_0 : f(y|grp = A) = f(y|grp = B)$
  - $H_0 : f(y|grp = A) \neq f(y|grp = B)$
    (i.e. two-sided alternatives)

- 10000 replications
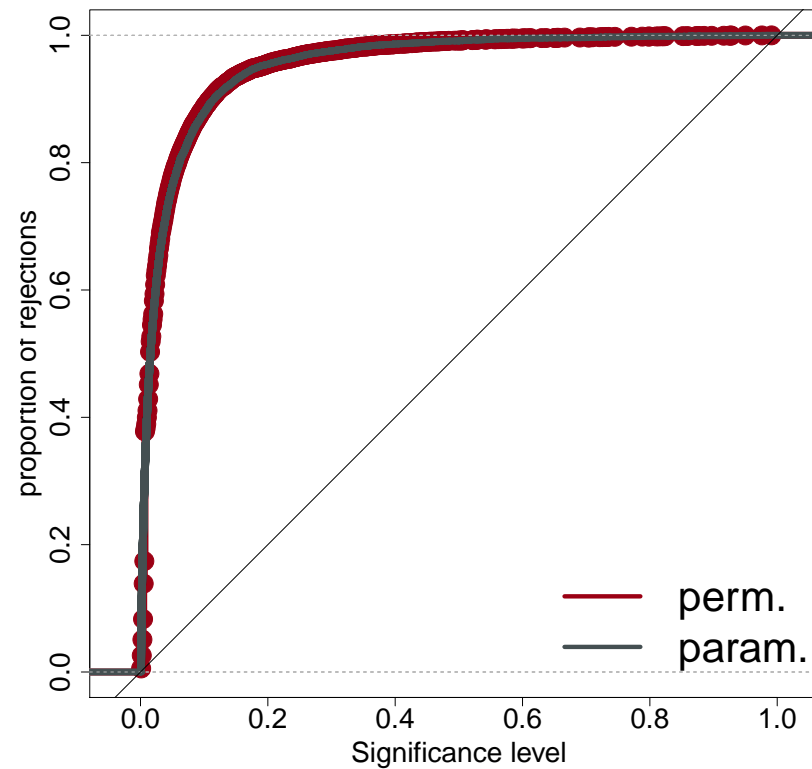
- 1000 random permutations for each test
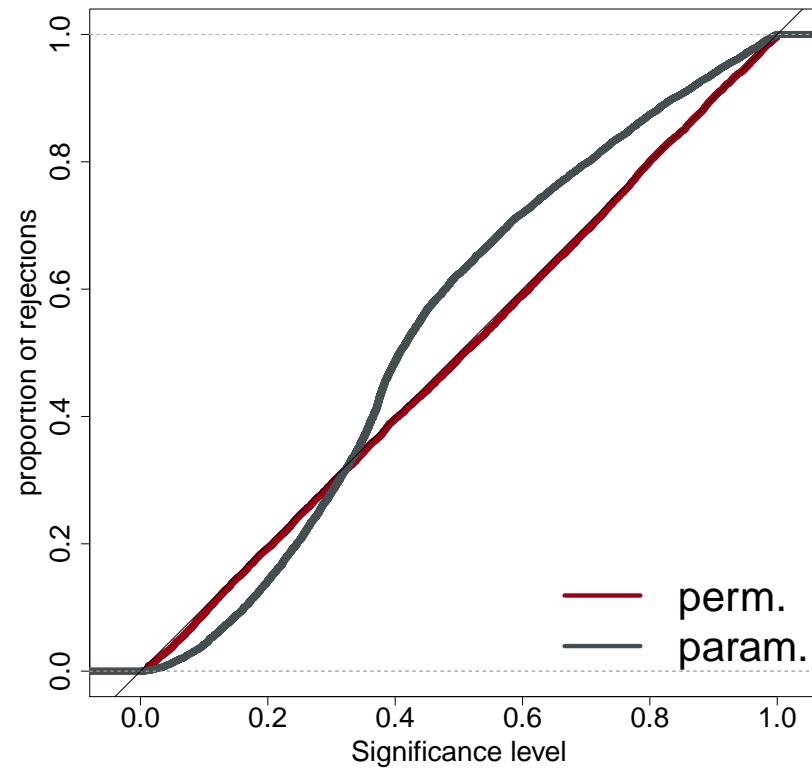
Simulation: H_0

Empirical Type I error	≤ 0.01	≤ 0.05	≤ 0.1	≤ 0.5	≤ 0.75
Permutation	0.00	0.04	0.09	0.49	0.74
Paramatric t.test	0.00	0.01	0.04	0.62	0.84

Simulation: H_1

now $(y|grp = A) \sim Cauchy(0)$, $(y|grp = B) \sim Cauchy(10)$
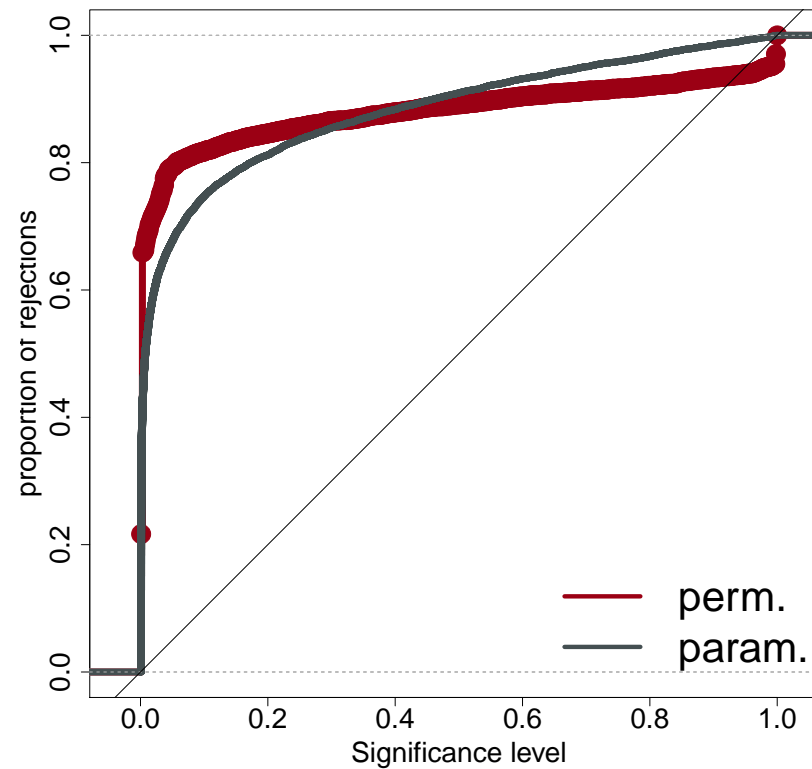
Empirical Power	≤ 0.01	≤ 0.05	≤ 0.1	≤ 0.5	≤ 0.75
Permutation	0.66	0.78	0.81	0.89	0.92
Paramatric t.test	0.53	0.68	0.75	0.91	0.96

Outline

A very general approach

This approach (Orbit $\mathcal{O}$ + Test statistic T) is very general.
It includes:

- ChiSquare test
- Fisher exact test
- McNemar test
- rank tests
- ANOVA tests
- linear models
- other models difficult to deal within the parametric framework
- . . .

The case of contingency table

$x \setminus y$	0	1	2
A	4	3	0
B	1	2	4

```
> chisq.test(x,y)
Pearson's Chi-squared test
data:  x and y
X-squared = 6, df = 2, p-value = 0.04979
Warning message:
In chisq.test(x, y) :  Chi-squared approximation may
be incorrect
```

The case of contingency table

Use: simulate.p.value = TRUE

```
> chisq.test(x,y,simulate.p.value = TRUE)
Pearson's Chi-squared test (based on 2000 replicates)
data:  x and y
X-squared = 6, df = NA, p-value = 0.09345
```

Same schema:

- Orbit $\mathcal{O}$: all possible permutation of $\mathbf{y}$ (under H_0 $\mathbf{x}$ and $\mathbf{y}$ are independent)

- test statistic $T(\mathbf{y}^*)$: the χ^2 statistic computed (higher is better)

- p-value: proportion of T^* greater than the one computed on observed data $\mathbf{y}$: $p = \sum_{\mathbf{y}^* \in \mathcal{O}} I(T(\mathbf{y}^*) \geq T(\mathbf{y}))/|\mathcal{O}|$.

The case of experimental design with blocks (or within-subject)

For example: n lots/subjects, each with two treatments (A vs B) we can assume a specific effect for each lot/subject:

$$y_{ij} \sim (\nu_i + \mu_j, \sigma_i)\ i = 1, \ldots, n,\ j = A, B$$

$$H_0 : \mu_A = \mu_B \Leftrightarrow \mu_A - \mu_B = 0$$

Parametric approach:

- define: $z_i = y_{iB} - y_{iA} \sim (\mu_B - \mu_A, 2 \cdot \sigma_i)\ i = 1, \ldots, n$
- assume $\sigma_i = \sigma,\ \forall i = 1, \ldots, n$
- perform a 1-sample t-test (i.e. t-test for 2 paired samples)
- test is exact only it z_i is normal, it is aproximated otherwise.

The case of experimental design with blocks (or within-subject)

What about permutation approach? how to define the Orbit $\mathcal{O}$?
$$H_0 \Rightarrow \mu_A = \mu_B = \mu \Rightarrow y_{ij} \sim (\nu_i + \mu, \sigma_i) \ i = 1, \ldots, n, \ j = A, B$$

$$f(y_{1A}, y_{1B}, y_{2A}, y_{2B}, \ldots, y_{nA}, y_{nB})(\text{observed}) =$$
$$= \ f(y_{1A}, y_{1B}, y_{2A}, y_{2B}, \ldots, y_{nA}, y_{nB}) =$$
$$= \ f(y_{1A}, y_{1B}, y_{2A}, y_{2B}, \ldots, y_{nA}, y_{nB}) \ldots$$

- We exchange observations only within the same lot/subject!
- There are 2^n possible configurations: $|\mathcal{O}| = 2^n$
- we don't need to assume: $\sigma_i = \sigma, \ \forall i = 1, \ldots, n$
- (even we may allow non addictive effect: $\nu_{ij} \neq \nu_i + \mu_j$)
- A part from $\mathcal{O}$, the procedure is the same.

Outline

Multivarite hypotheses

Testing $H_0(weight)$ + Testing $H_0(leaf.length)$
is different from testing

- $H_0 : \ H_0(weight) \ \cap \ H_0(leaf.length)$
  (i.e. simultaneously true)

  vs

- $H_1 : \ H_1(weight) \ \cup \ H_1(leaf.length)$
  (i.e. at least one null hypo is false)

Here test H_0: the New is equal to the Standard in both variables

For 1-tailed alternative:

$p_{weight} = .098$ and $p_{leaf.length} = .030$.

Shall we reject the multivariate H_0?

Multivariate hypotheses

Common approaches:

- MANOVA test

  is OK but only for linear models (2 or more samples).
  assumes multivariate normality.
  Does not allow for one-sided alternatives.
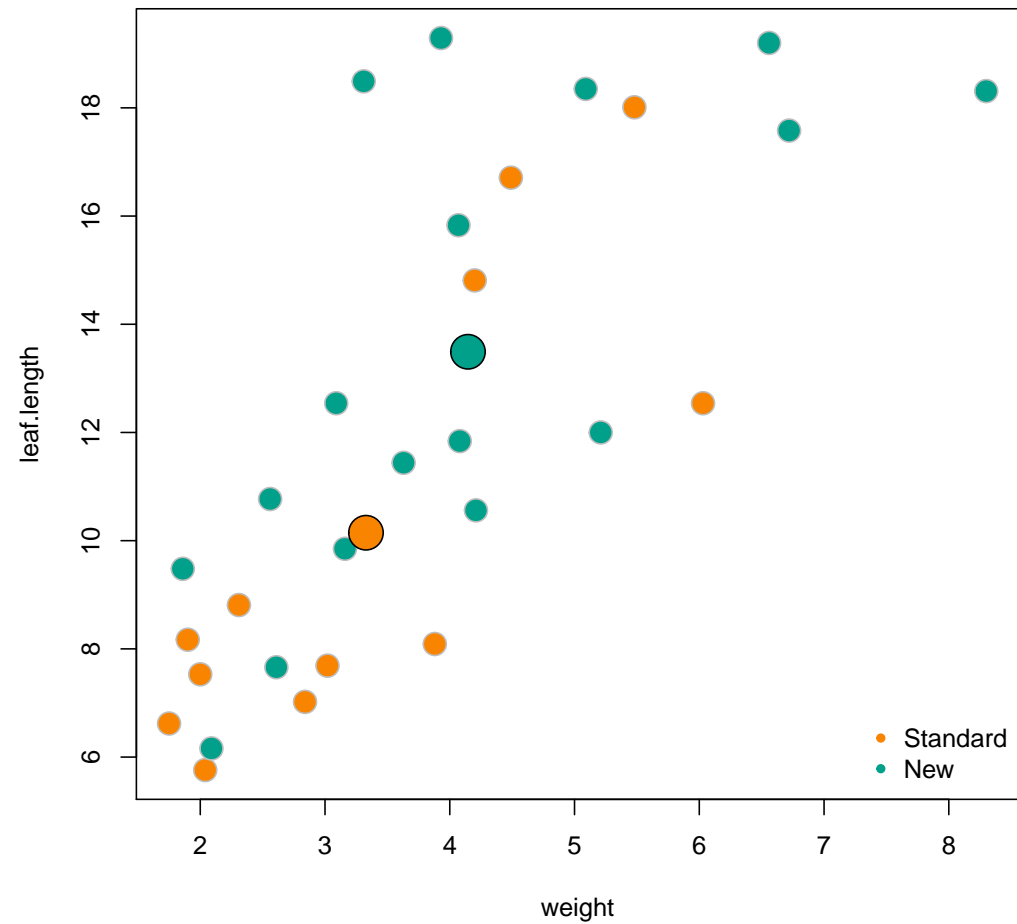
- Bonferroni correction $(p = min(p_1, p_2) * 2)$

  very simple, always valid
  does not take in account dependences among data
  (i.e. may be very conservative, i.e. high final p-value)
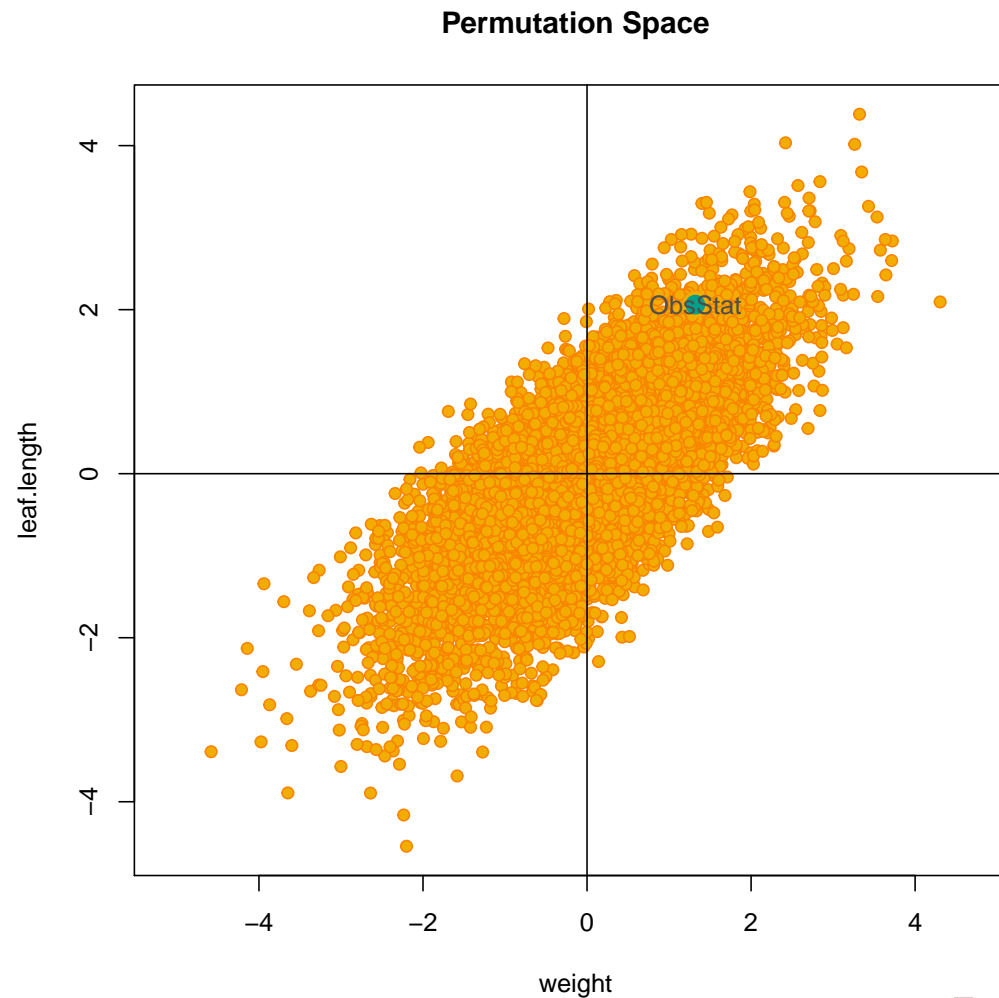
Joint distributions of the data

The two variables are dependent:

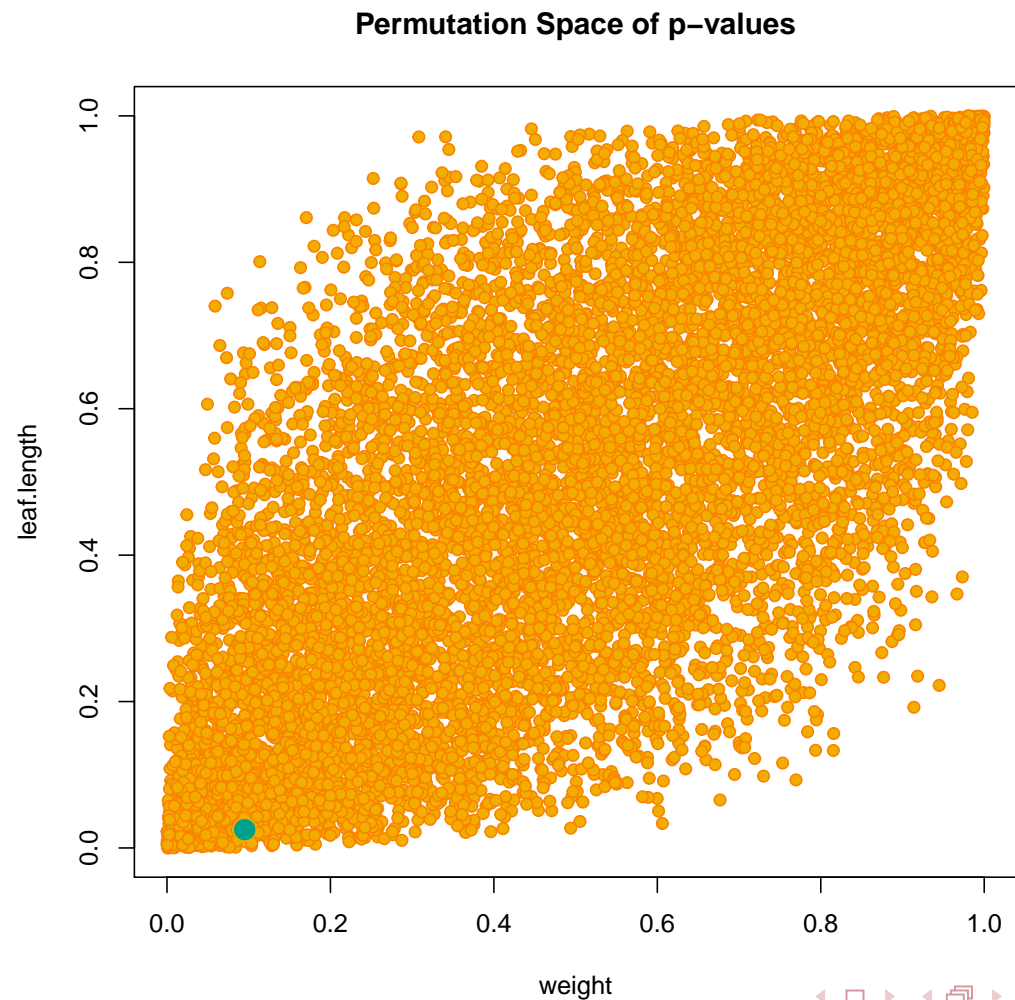# Joint distributions of test statistics

This dependence induce a dependence into the joint distribution of the test statistic

Permutation Space

Joint distributions of the p-values

… and into the p-values joint distributions (i.e. compute the p-values for observed samples and all elements $\mathbf{y}^* \in \mathcal{O}$)

Permutation Space of p−values

Nonparametric Combination methodology (Pesarin, 2001)

The Orbit is now defined in a multivariate framework:

$$\mathcal{O} = \{(\mathbf{y}_1, \mathbf{y}_2)^* : f((\mathbf{y}_1, \mathbf{y}_2)^*) = f((\mathbf{y}_1, \mathbf{y}_2))\}$$

(in practice: when you permute one observation in a variable, do the same in the other variables)

Also compute the p_1^*, p_2^* associated to each $\mathbf{y}_1^*, \mathbf{y}_2^*$.

Nonparametric Combination methodology (Pesarin, 2001)

The Orbit is now defined in a multivariate framework:
$$\mathcal{O} = \{(\mathbf{y}_1, \mathbf{y}_2)^* : f((\mathbf{y}_1, \mathbf{y}_2)^*) = f((\mathbf{y}_1, \mathbf{y}_2))\}$$
(in practice: when you permute one observation in a variable, do the same in the other variables)

Also compute the p_1^*, p_2^* associated to each $\mathbf{y}_1^*, \mathbf{y}_2^*$.

Define a Combining Function $\psi(p_1, \ldots, p_m)$ having the following properties:
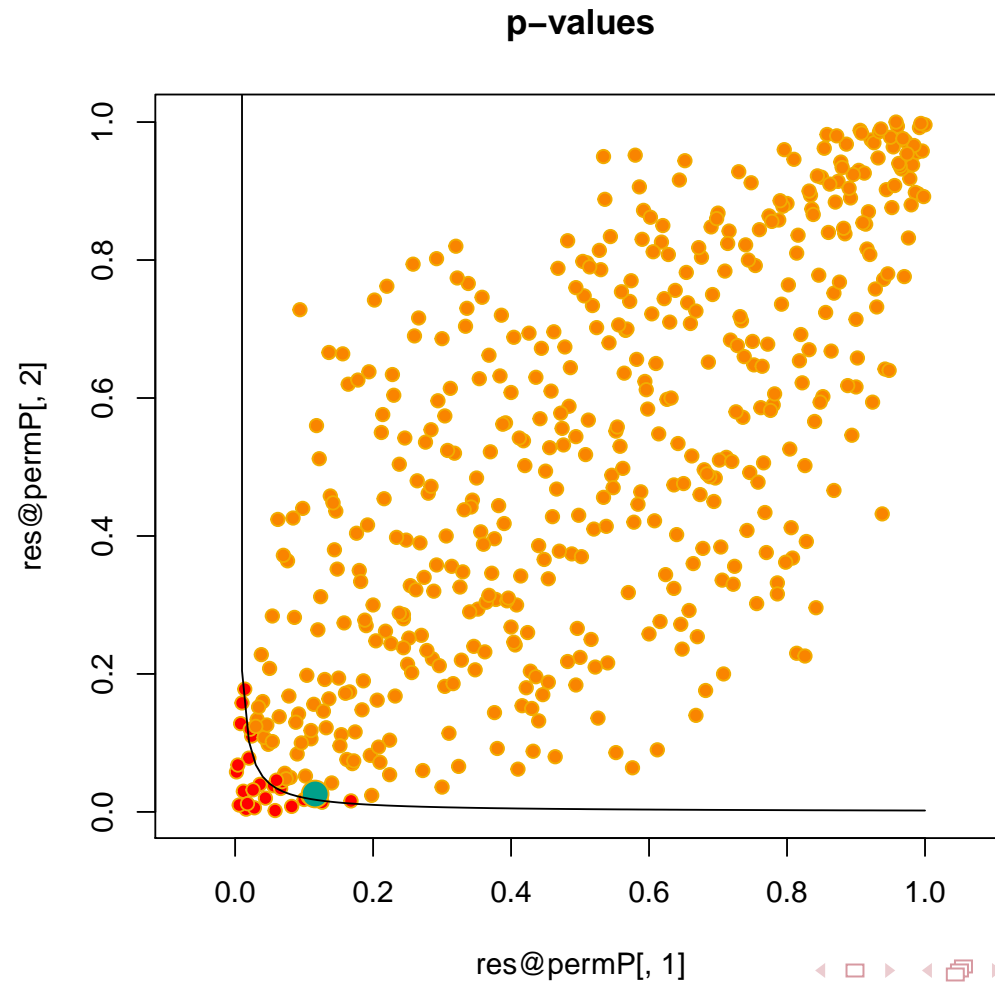
   i is non-increasing in each argument: $p_k < p_k^{'}$ implies
$$\psi(\ldots, p_k, \ldots) \geq \psi(\ldots, p_k^{'}, \ldots);$$

  ii attains its supremum ψ° if at least one argument attains 0;

 iii $\alpha > 0$ implies the critical value is such that $T_{\psi\alpha} < \psi^\circ$, i.e. no concentration of points at ψ° under H_0.

Apply $T((p_1, p_2)^*) = \psi((p_1, p_2)^*)$ and compute the p_{global}.

Fisher combining function

$$\psi = -2 \cdot (log(p_1) + log(p_2)) = 6.029$$
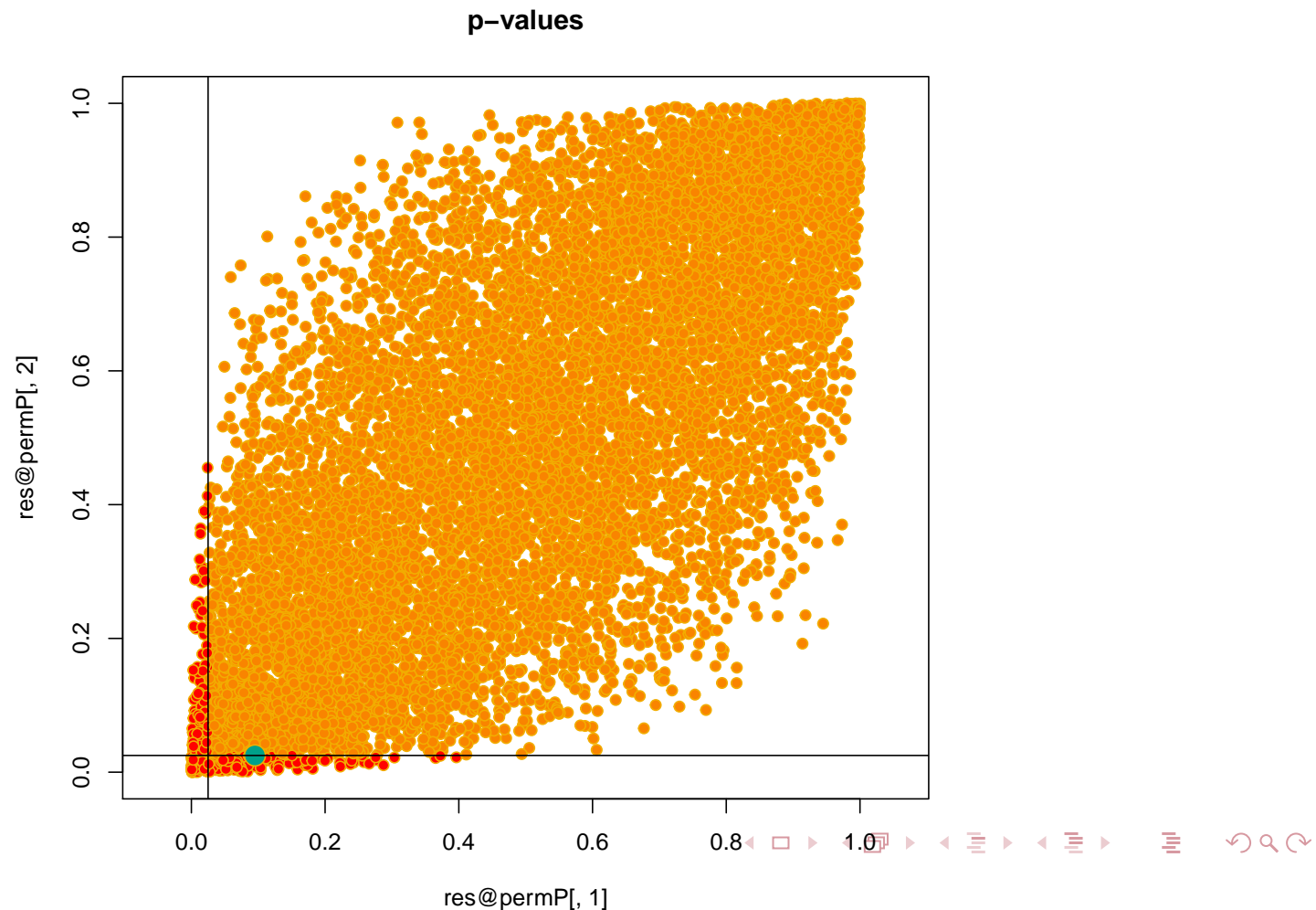$$p_{global} = 0.0400$$

p−values

Tippett (min-p) combining function

$\psi = 1 - min(p_1, p_2) = 1 - 0.0280$

$p_{global} = 0.0460 \ (\leq 2 \cdot 0.0280 = 0.0560$ similar to Bonferroni, but more powerfull)

p–values

The need for Multiple Testing

○ Here we only have global p-values (in analogy to Manova, F-tests etc).

The need for Multiple Testing

- Here we only have global p-values (in analogy to Manova, F-tests etc).

- If the test is significant at least one among $H_0(weight)$ or $H_0(leaf.len)$ is false.

- We don't know which one! (weak control of the FamilyWise Error)

The need for Multiple Testing

- Here we only have global p-values (in analogy to Manova, F-tests etc).

- If the test is significant at least one among $H_0(weight)$ or $H_0(leaf.len)$ is false.

- We don't know which one! (weak control of the FamilyWise Error)

- How to cast them into multiple testing procedures?

The need for Multiple Testing

- Here we only have global p-values (in analogy to Manova, F-tests etc).

- If the test is significant at least one among $H_0(weight)$ or $H_0(leaf.len)$ is false.

- We don't know which one! (weak control of the FamilyWise Error)

- How to cast them into multiple testing procedures?

- (strong) control of the FWER is easy!

Take Home Message

Permutation approach:

- very general uni/multi-variate approach
- few assumptions on the data-generating process
- natural approach in **randomized experimental design**
- good inferential properties
  (in most of the cases: exact control of the type I error, consistency, asymptotic optimality)
- very convenient for multiplicity control methods, since it deals easily with dependent tests.

Warnings:

- more complex experimental design can be dealt, but with caution (Pesarin, 2001)
- multiple (generalized) linear models need some care (Solari, Finos & Goeman, 2014; Hemerik, Goeman & Finos, 2018)

Software 'R': libraries 'coin','flip', 'permuco'