Introduction to regression

Introduction to the randomisation approach

Throughout this module, we have conducted hypothesis tests in a similar way:

- Calculate test statistic (e.g., t, F, Chi-square)
- Compare our calculated statistic to the theoretical null distribution

Introduction to the randomisation approach

Throughout this module, we have conducted hypothesis tests in a similar way:

- Calculate test statistic (e.g., t, F, Chi-square)
- Compare our calculated statistic to the theoretical null distribution

P-value: Assuming the null hypothesis is true, what is the probability of getting the actual test statistic, or one more extreme than it, from this null distribution?

Introduction to the randomisation approach

Null distribution makes several assumptions about the data:

- Normal distribution
- Equality of variances

Introduction to the randomisation approach

Null distribution makes several assumptions about the data:

- Normal distribution
- Equality of variances

When these assumptions are violated, we then need to apply a transformation of some sort to the data, or to use a different non-parametric approach to testing our null hypothesis.

Introduction to the randomisation approach

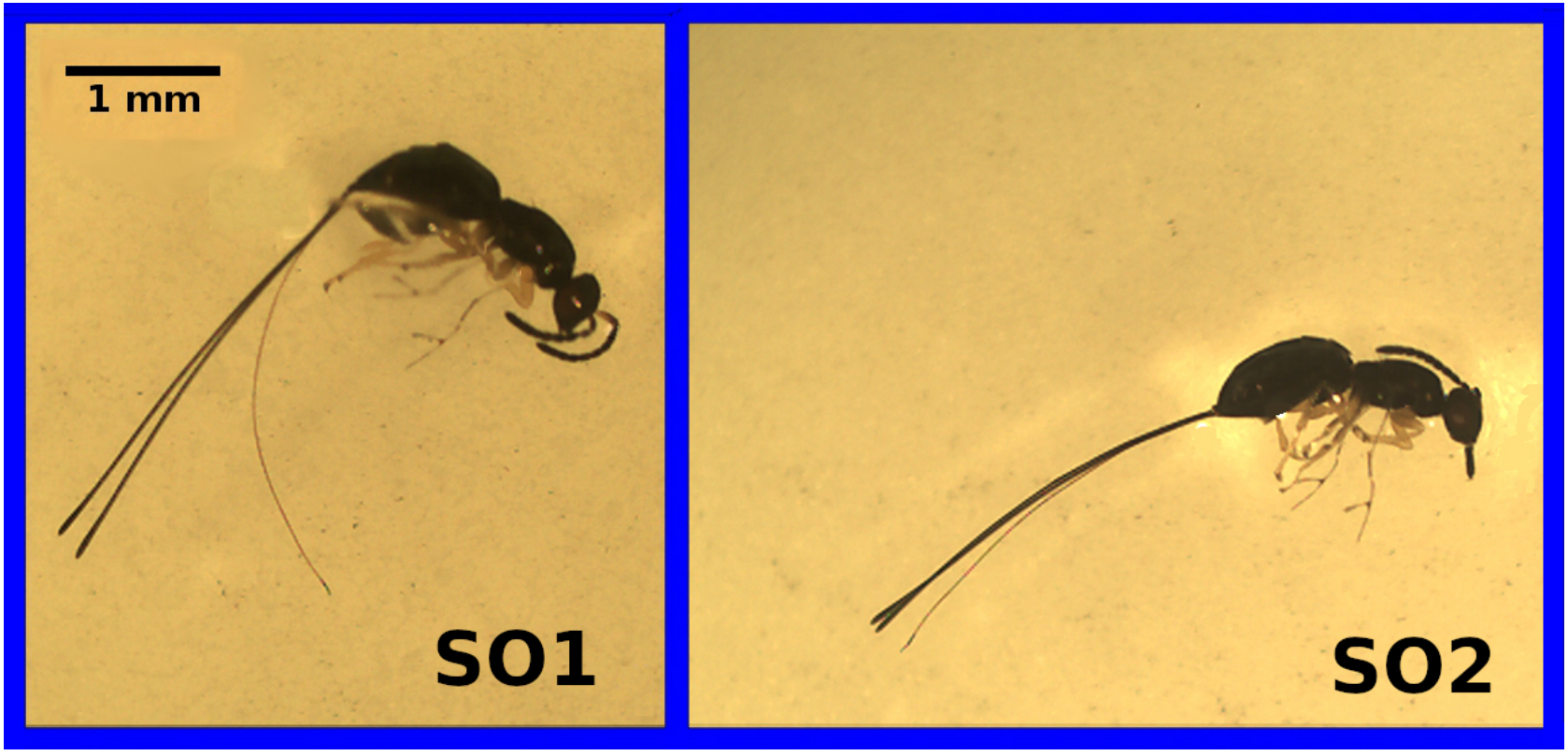
If the ordering of the data we collected was actually random, then what is the probability of getting a test statistic as or more extreme than the one that we actually did?

Introduction to the randomisation approach

If the ordering of the data we collected was actually random, then what is the probability of getting a test statistic as or more extreme than the one that we actually did?

- Build the null distribution by randomising our data in some useful way
- Conceptually, most students actually find randomisation methods easier to understand
- More challenging to implement in practice

An instructive example



Two fig wasps with (relatively) short ovipositors

An instructive example

Species	Ovipositor length (mm)
SO1	3.256
SO1	3.133
SO1	3.071
SO1	2.299
SO1	2.995
SO1	2.929
SO1	3.291
SO1	2.658
SO1	3.406

An instructive example

Standard approach would be to use a two sample t-test:

- **Null hypothesis:** the two means are the same
- **Alternative hypothesis:** the two means are not the same (two-sided)

An instructive example

Standard approach would be to use a two sample t-test:

- **Null hypothesis:** the two means are the same
- **Alternative hypothesis:** the two means are not the same (two-sided)

Check t-test assumptions:

- Data are normally distributed
- Both samples have similar variances

If data are not normally distributed, use a Mann Whitney test instead

An instructive example

Calculate the t-statistic:

$$t = \frac{mean(SO1) - mean(SO2)}{s_p}.$$

An instructive example

Calculate the t-statistic:

$$t = \frac{mean(SO1) - mean(SO2)}{s_p}.$$

The s_p is just being used as a short-hand to indicate the pooled standard deviation,

$$s_p = \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{n_1 + n_2}{n_1 n_2}\right)}.$$

Values of n1 and n2 are the sample sizes for SO1 (17) and SO2 (15), respectively, and s_1^2 and s_2^2 are the sample variances for SO1 and SO2, respectively.

An instructive example

Mean ovipositor lengths:

- SO1: 3.0301176 mm
- SO2: 2.8448667 mm

An instructive example

Mean ovipositor lengths:

- SO1: 3.0301176 mm
- SO2: 2.8448667 mm

Calculate our t-statistic, then get p-value with critical value table or SPSS:

- t-statistic (30 df): 2.4190497
- p-value: 0.0218352

An instructive example

Mean ovipositor lengths:

- SO1: 3.0301176 mm
- SO2: 2.8448667 mm

Calculate our t-statistic, then get p-value with critical value table or SPSS:

- t-statistic (30 df): 2.4190497
- p-value: 0.0218352

Hence, we would reject our null hypothesis and conclude that the difference between the group means is statistically significant.

An instructive example

Randomisation approach: If there really is no difference between group means, then we should be able to randomly shuffle group identities (species) and get a difference between means that is not far off the one we actually get from the data.

1. Shuffle the identities all of the species

An instructive example

Randomisation approach: If there really is no difference between group means, then we should be able to randomly shuffle group identities (species) and get a difference between means that is not far off the one we actually get from the data.

1. Shuffle the identities all of the species
2. Calculate the mean between shuffled groups

An instructive example

Randomisation approach: If there really is no difference between group means, then we should be able to randomly shuffle group identities (species) and get a difference between means that is not far off the one we actually get from the data.

1. Shuffle the identities all of the species
2. Calculate the mean between shuffled groups
3. Repeat steps 1 and 2 many times

An instructive example

Randomisation approach: If there really is no difference between group means, then we should be able to randomly shuffle group identities (species) and get a difference between means that is not far off the one we actually get from the data.

1. Shuffle the identities all of the species
2. Calculate the mean between shuffled groups
3. Repeat steps 1 and 2 many times
4. Compare the shuffled mean differences to the actual one

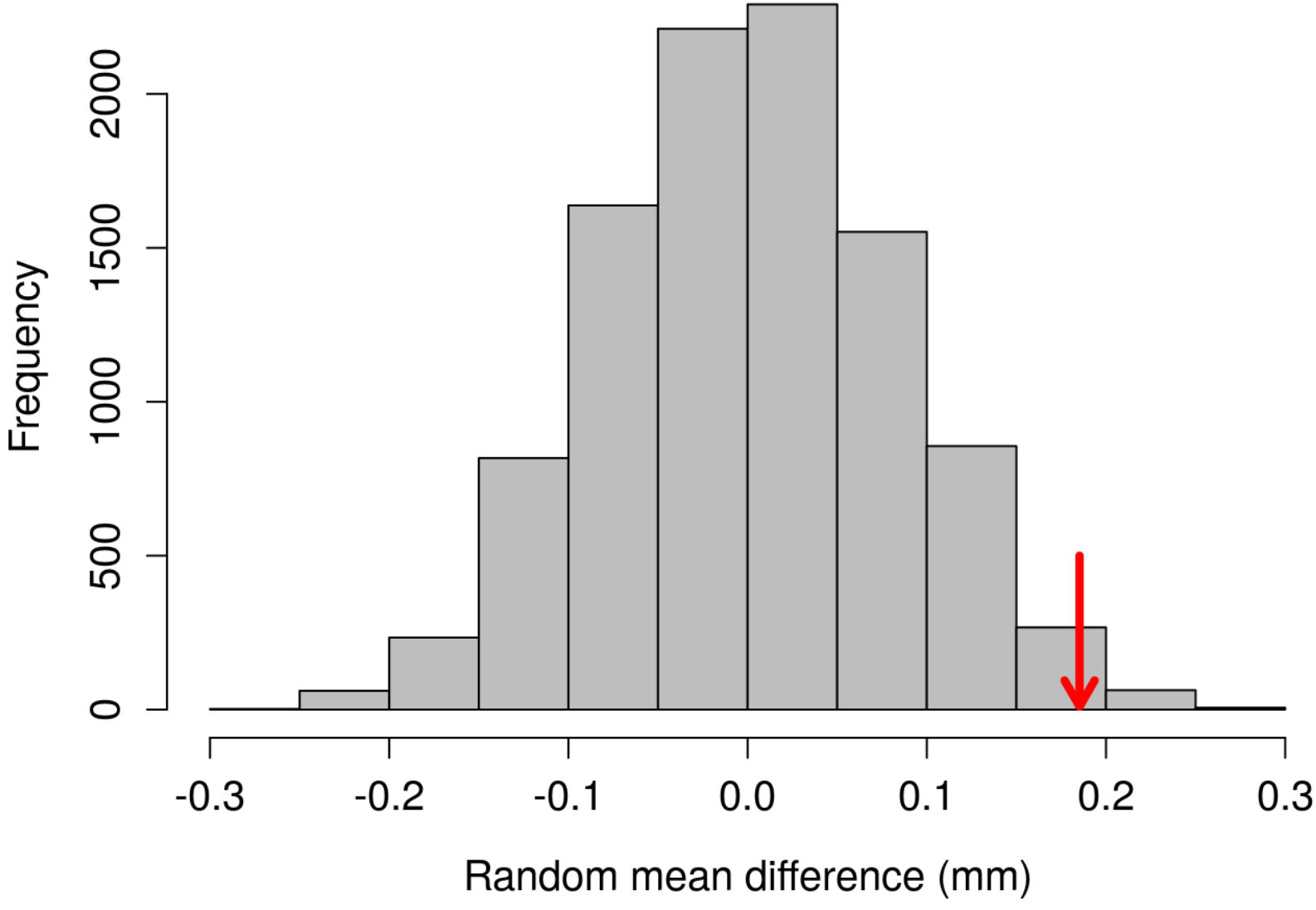
An instructive example

Randomisation approach: If there really is no difference between group means, then we should be able to randomly shuffle group identities (species) and get a difference between means that is not far off the one we actually get from the data.

1. Shuffle the identities all of the species
2. Calculate the mean between shuffled groups
3. Repeat steps 1 and 2 many times
4. Compare the shuffled mean differences to the actual one

This app illustrates how the process works.

An instructive example



An instructive example

Comparing the observed distance to the random mean differences:

- Observed difference was 0.185251
- 213 random differences were more extreme (above 0.185251 or below -0.185251)

An instructive example

Comparing the observed distance to the random mean differences:

- Observed difference was 0.185251
- 213 random differences were more extreme (above 0.185251 or below -0.185251)
- Probability of value as or more extreme

$$P = \frac{213 + 1}{9999 + 1}$$

An instructive example

Comparing the observed distance to the random mean differences:

- Observed difference was 0.185251
- 213 random differences were more extreme (above 0.185251 or below -0.185251)
- Probability of value as or more extreme

$$P = \frac{213 + 1}{9999 + 1}$$

Compare the randomisation and traditional approaches:

- Randomisation P value: 0.0214
- Traditional t-test p-value: 0.0218

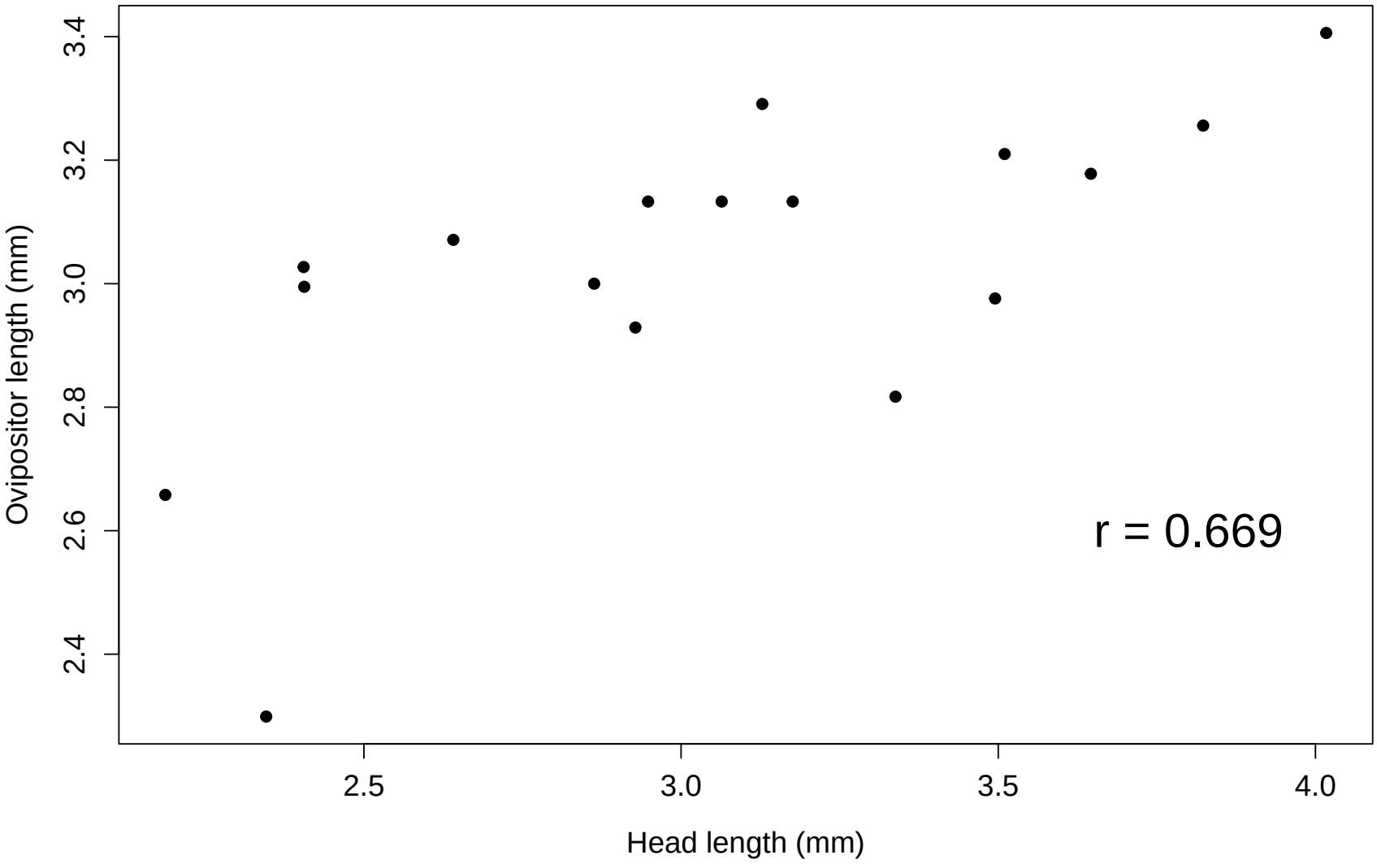
Randomisation is a general tool

Species	Ovipositor length (mm)
SO1	3.256
SO1	3.133
SO1	3.071
SO1	2.299
SO1	2.995
SO1	2.929
SO1	3.291
SO1	2.658
SO1	3.406

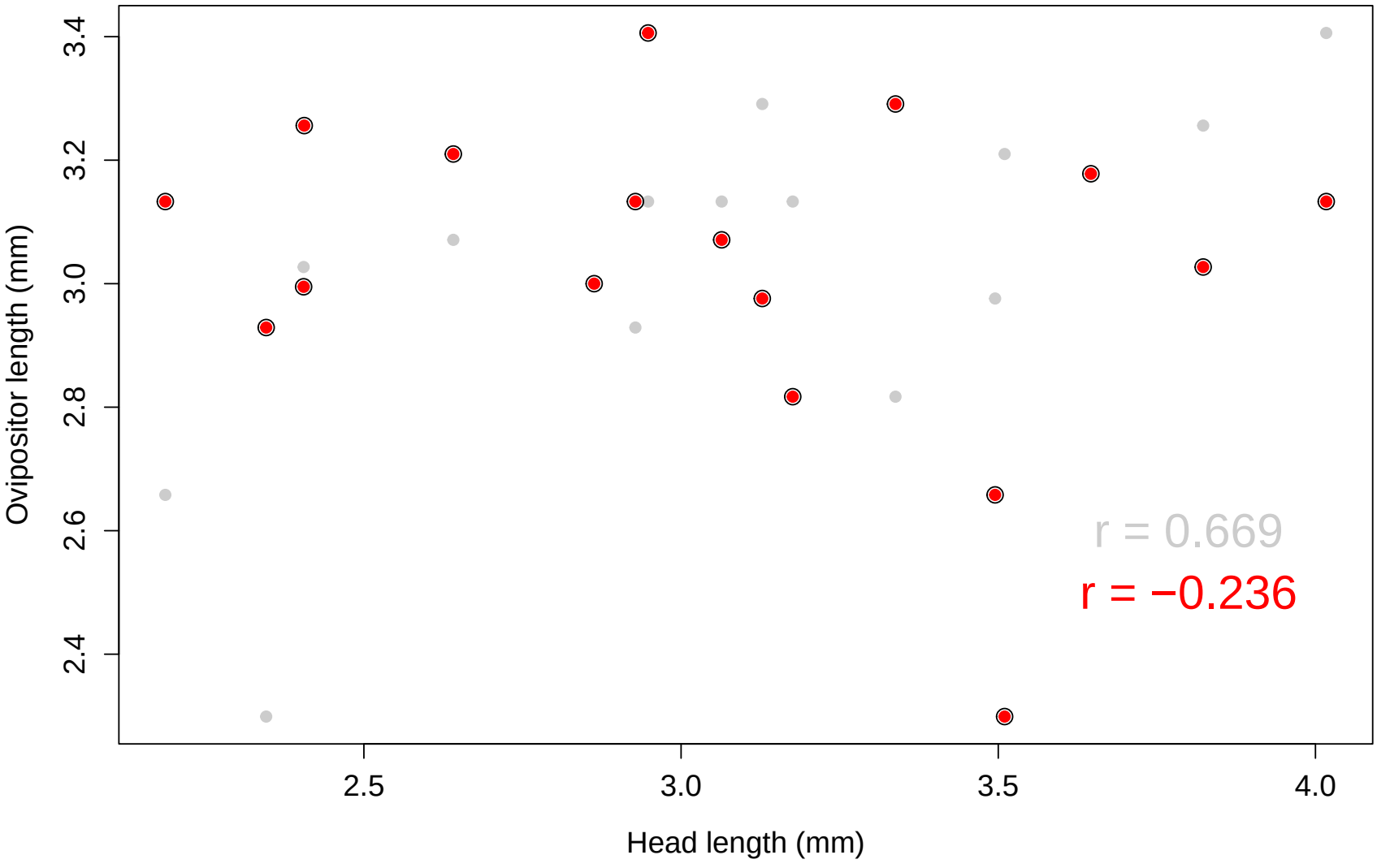
Randomisation is a general tool

Head length (mm)	Ovipositor length (mm)
3.823	3.256
2.948	3.133
2.641	3.071
2.346	2.299
2.406	2.995
2.928	2.929
3.128	3.291
2.187	2.658
4.017	3.406

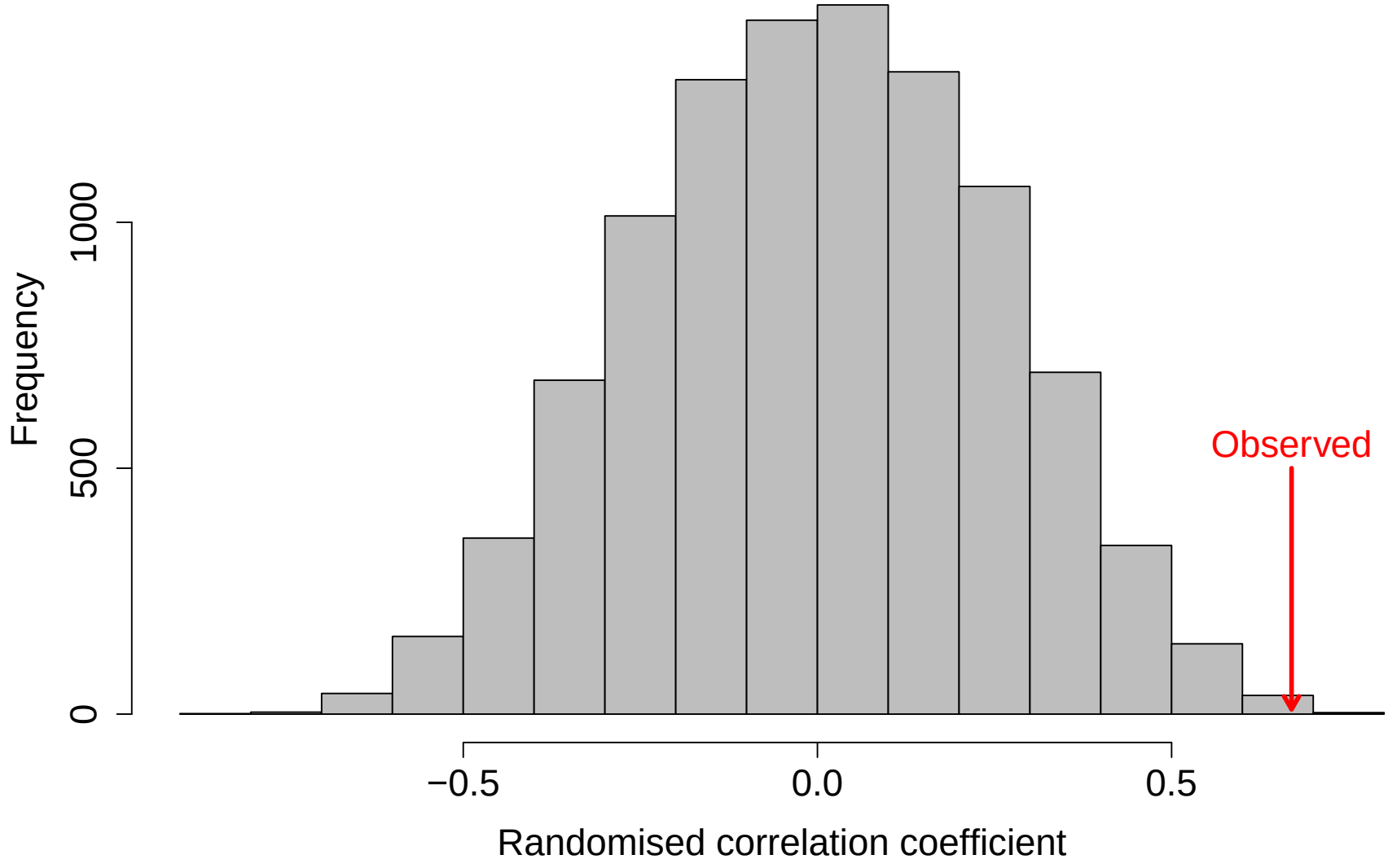
Randomisation is a general tool



Randomisation is a general tool



Randomisation is a general tool



Randomisation is a general tool

After randomising ovipositor length values 9999 times, 17 values more extreme than observed correlation (i.e., $>$ 0.669 or $<$ -0.669)

Randomisation is a general tool

After randomising ovipositor length values 9999 times, 17 values more extreme than observed correlation (i.e., $>$ 0.669 or $<$ -0.669)

$$P = \frac{17 + 1}{9999 + 1} = 0.0018.$$

Randomisation is a general tool

After randomising ovipositor length values 9999 times, 17 values more extreme than observed correlation (i.e., $>$ 0.669 or $<$ -0.669)

$$P = \frac{17 + 1}{9999 + 1} = 0.0018.$$

Traditional Pearson product moment p-value:

$$P = 0.003293.$$

Assumptions of randomisation tests

Assumptions of two sample t-test:

1. Data from indendent groups
2. The data are normally distributed
3. Observations independent of each other
4. Both samples have similar variances

Assumptions of randomisation tests

Assumptions of two sample t-test:

1. Data from indendent groups
2. The data are normally distributed
3. Observations independent of each other
4. Both samples have similar variances

With randomisation, we do not need to assume 1-3.

Assumptions of randomisation tests

The downsides of randomisation

- Statistical inferences are limited to the sample (not the population)
- This is because no formal assumption of random sample from populations

[1] Ernst, M D. 2004. *Stat. Sci.* 19:676-685.

[2] Ludbrook, J, & H Dudley. 1998. *Am. Stat.*, 52:127-132.

Assumptions of randomisation tests

The downsides of randomisation

- Statistical inferences are limited to the sample (not the population)
- This is because no formal assumption of random sample from populations
- Can still make the verbal argument generalising from sample to populations [1, 2]

---

[1] Ernst, M D. 2004. *Stat. Sci.* 19:676-685.

[2] Ludbrook, J, & H Dudley. 1998. *Am. Stat.*, 52:127-132.

Bootstrapping confidence intervals

We can use randomisation to calculate confidence intervals (CIs)

Recall the traditional approach to calculating CIs:

- Upper CI: mean + constant * standard error
- Lower CI: mean - constant * standard error

Bootstrapping confidence intervals

We can use randomisation to calculate confidence intervals (CIs)

Recall the traditional approach to calculating CIs:

- Upper CI: mean + constant * standard error
- Lower CI: mean - constant * standard error

Calculating standard error:

$$SE = \frac{s}{\sqrt{n}}.$$

Above, s is the standard deviation and n is the sample size.

Bootstrapping confidence intervals

We can use randomisation to calculate confidence intervals (CIs)

Recall the traditional approach to calculating CIs:

- Upper CI: mean + constant * standard error
- Lower CI: mean - constant * standard error

Finding the right constant

- Appropriate z score or t score
- Encompasses correct probability density

E.g., 95 per cent of standard normal distribution between z = -1.96 and 1.96

Bootstrapping confidence intervals

Values of SO1 ovipositor length:

3.256, 3.133, 3.071, 2.299, 2.995, 2.929, 3.291, 2.658, 3.406, 2.976, 2.817, 3.133, 3, 3.027, 3.178, 3.133, 3.21

Bootstrapping confidence intervals

Values of SO1 ovipositor length:

3.256, 3.133, 3.071, 2.299, 2.995, 2.929, 3.291, 2.658, 3.406, 2.976, 2.817, 3.133, 3, 3.027, 3.178, 3.133, 3.21

- Mean: 3.0301
- Standard error: 0.0631
- t-score for df $=$ 17 - 1: 2.120

Bootstrapping confidence intervals

Values of SO1 ovipositor length:

3.256, 3.133, 3.071, 2.299, 2.995, 2.929, 3.291, 2.658, 3.406, 2.976, 2.817, 3.133, 3, 3.027, 3.178, 3.133, 3.21

- Mean: 3.0301
- Standard error: 0.0631
- t-score for df = 17 - 1: 2.120

Upper CI: $3.0301 + 2.120(0.0631) = 3.164$

Lower CI: $3.0301 - 2.120(0.0631) = 2.896$

Bootstrapping confidence intervals

Randomisation approach:

- Resample the data *with replacement* many times

[1] Manly, B F J. 2007. *Randomization, Bootstrap and Monte Carlo Methods in Biology* (3rd ed.). Chapman & Hall/CRC.

Bootstrapping confidence intervals

Randomisation approach:

- Resample the data *with replacement* many times
- Approximate what would happen if we went out and collected more data from the population

[1] Manly, B F J. 2007. *Randomization, Bootstrap and Monte Carlo Methods in Biology* (3rd ed.). Chapman & Hall/CRC.

Bootstrapping confidence intervals

Randomisation approach:

- Resample the data *with replacement* many times
- Approximate what would happen if we went out and collected more data from the population
- Get the distribution of means from hypothetical resampling [1]

[1] Manly, B F J. 2007. *Randomization, Bootstrap and Monte Carlo Methods in Biology* (3rd ed.). Chapman & Hall/CRC.

Bootstrapping confidence intervals

Randomisation approach:

- Resample the data *with replacement* many times
- Approximate what would happen if we went out and collected more data from the population
- Get the distribution of means from hypothetical resampling [1]
- Rank random means, get lower 2.5% and upper 97.5% of ranked values

[1] Manly, B F J. 2007. *Randomization, Bootstrap and Monte Carlo Methods in Biology* (3rd ed.). Chapman & Hall/CRC.

Bootstrapping confidence intervals

Original SO1 ovipositor length values:

3.256, 3.133, 3.071, 2.299, 2.995, 2.929, 3.291, 2.658, 3.406, 2.976, 2.817, 3.133, 3, 3.027, 3.178, 3.133, 3.21

Bootstrapping confidence intervals

Original SO1 ovipositor length values:

3.256, 3.133, 3.071, 2.299, 2.995, 2.929, 3.291, 2.658, 3.406, 2.976, 2.817, 3.133, 3, 3.027, 3.178, 3.133, 3.21

SO1 ovipositor values resampled *with replacement*

3, **3.027**, 2.929, **3.291**, **3.291**, 2.976, **2.299**, **2.299**, **3.027**, **3.027**, **3.133**, **3.027**, **2.299**, **3.133**, **2.299**, 3.178, **3.133**

- Bold values appear multiple times in resampling
- Original values 3.256, 3.071, 2.995, 2.658, 3.406, 2.817, and 3.21 do not appear at all in resampling

Bootstrapping confidence intervals

Original SO1 ovipositor length values:

3.256, 3.133, 3.071, 2.299, 2.995, 2.929, 3.291, 2.658, 3.406, 2.976, 2.817, 3.133, 3, 3.027, 3.178, 3.133, 3.21

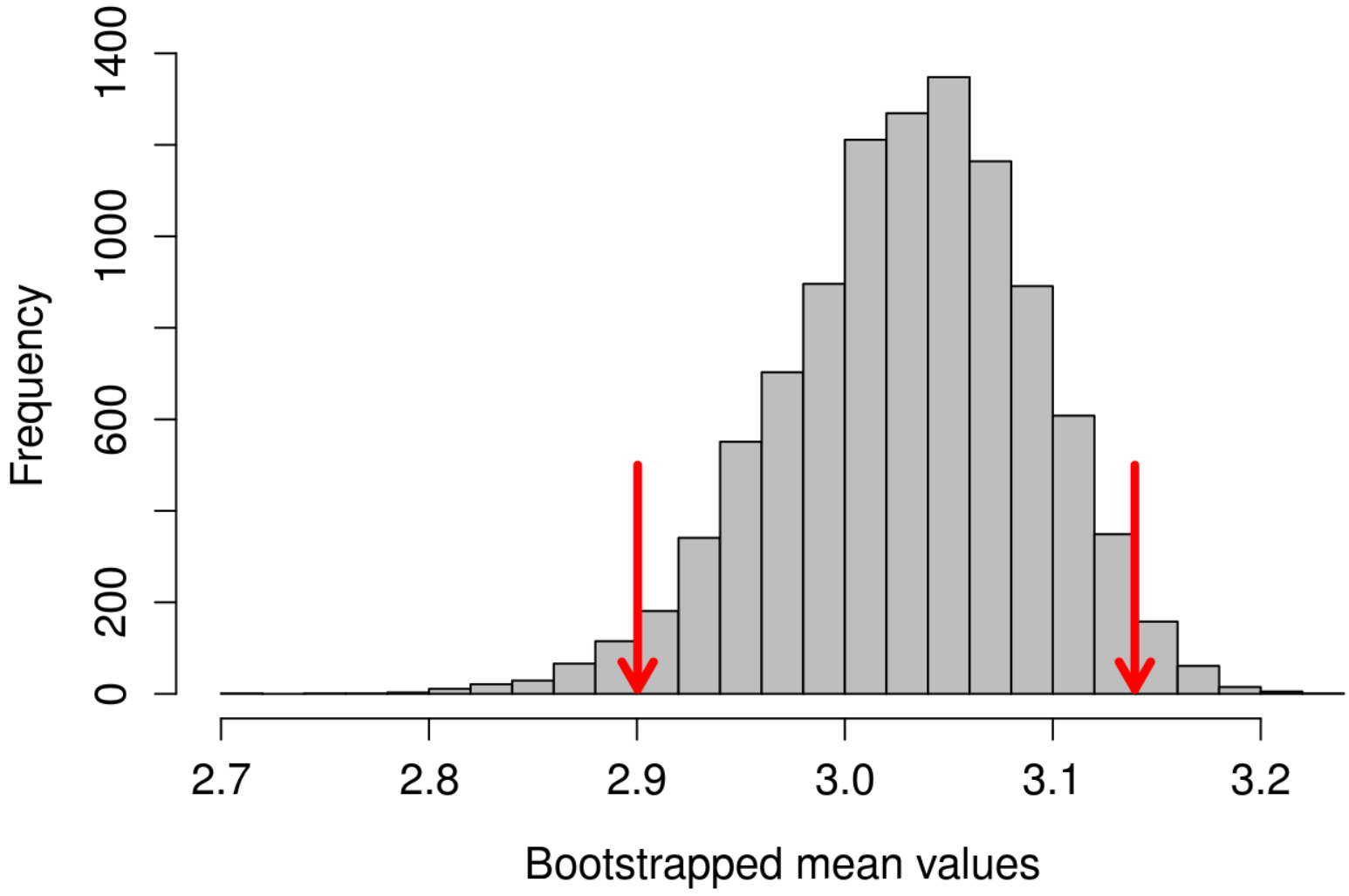
SO1 ovipositor values resampled *with replacement*

3, **3.027**, 2.929, **3.291**, **3.291**, 2.976, **2.299**, **2.299**, **3.027**, **3.027**, **3.133**, **3.027**, **2.299**, **3.133**, **2.299**, 3.178, **3.133**

- Bold values appear multiple times in resampling
- Original values 3.256, 3.071, 2.995, 2.658, 3.406, 2.817, and 3.21 do not appear at all in resampling

Now we calculate the mean of our values resampled with replacement 10000 times.

Bootstrapping confidence intervals



Monte Carlo on spatial data

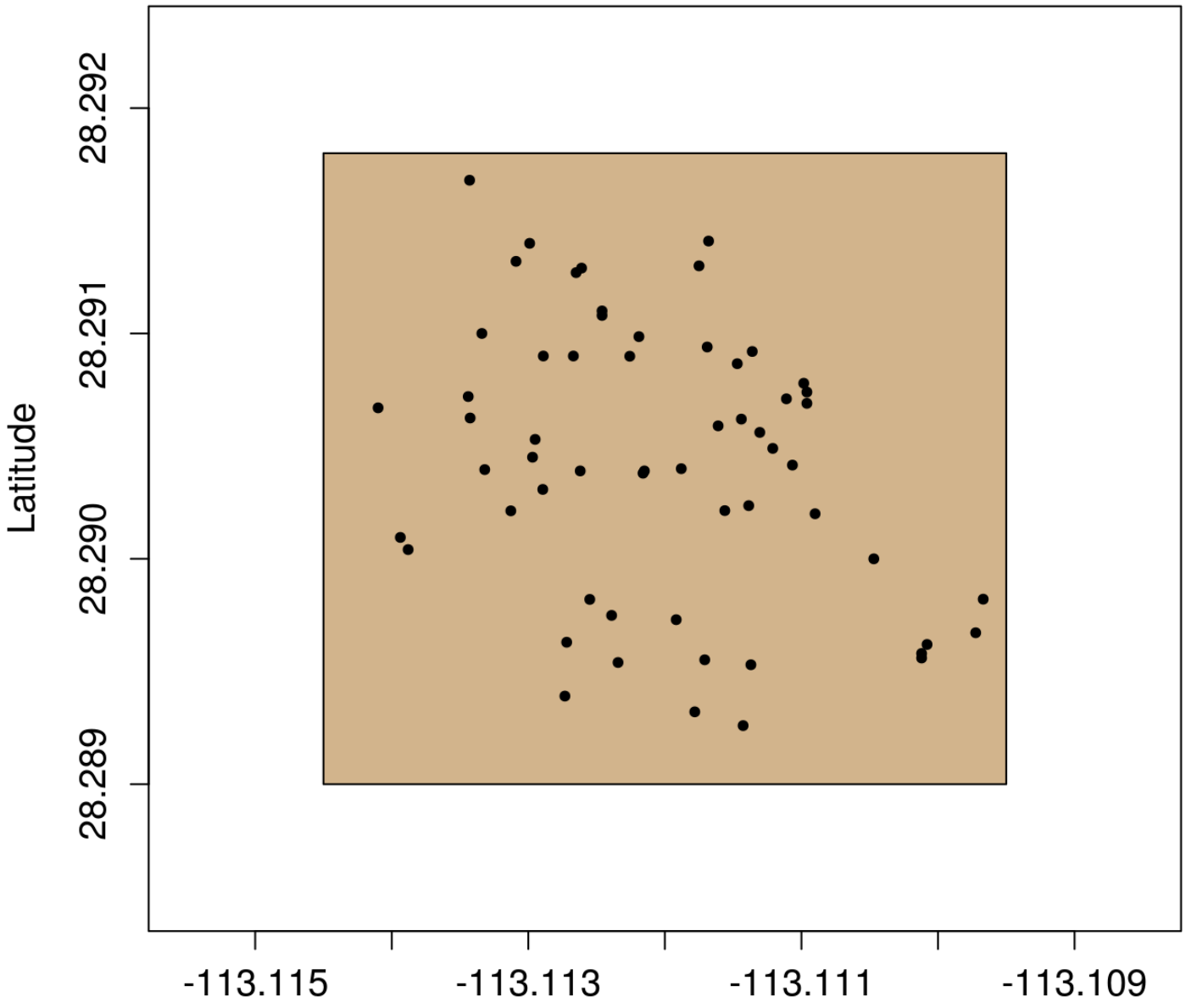
We can also use randomisation when a null distribution cannot be derived from the data.

Are fig trees are randomly distributed across a sampling area?

Monte Carlo on spatial data

Site	Tree	Latitude	Longitude	Elevation
S172	T34	28.29021	-113.1116	718.2859
S172	T01	28.29141	-113.1117	664.8726
S172	T02	28.29130	-113.1118	652.8560
S172	T03	28.29129	-113.1126	663.6709
S172	T04	28.29127	-113.1127	653.3367
S172	T05A	28.29110	-113.1125	676.8889

Monte Carlo on spatial data



Monte Carlo on spatial data

Are the trees randomly distributed in the brown box?

- Cannot randomise latitude and longitude coordinates
- Want to compare actual tree distribution with randomly placed trees
- Monte Carlo compares observed test statistic with one derived from theoretical null [1]
- Use mean distance to nearest neighbour as a test statistic

---

[1] Manly, B F J. 2007. *Randomization, Bootstrap and Monte Carlo Methods in Biology* (3rd ed.). Chapman & Hall/CRC.

Monte Carlo on spatial data

Null Hypothesis: Mean nearest neighbour distance for a focal tree in the observed data will not differ significantly from the mean nearest neighour distance obtained from the same number of trees randomly distributed within the sampling area.

1. Randomly place 59 trees in the sampling area
2. Calculate mean nearest neighbour distance of random trees

Monte Carlo on spatial data

Null Hypothesis: Mean nearest neighbour distance for a focal tree in the observed data will not differ significantly from the mean nearest neighour distance obtained from the same number of trees randomly distributed within the sampling area.

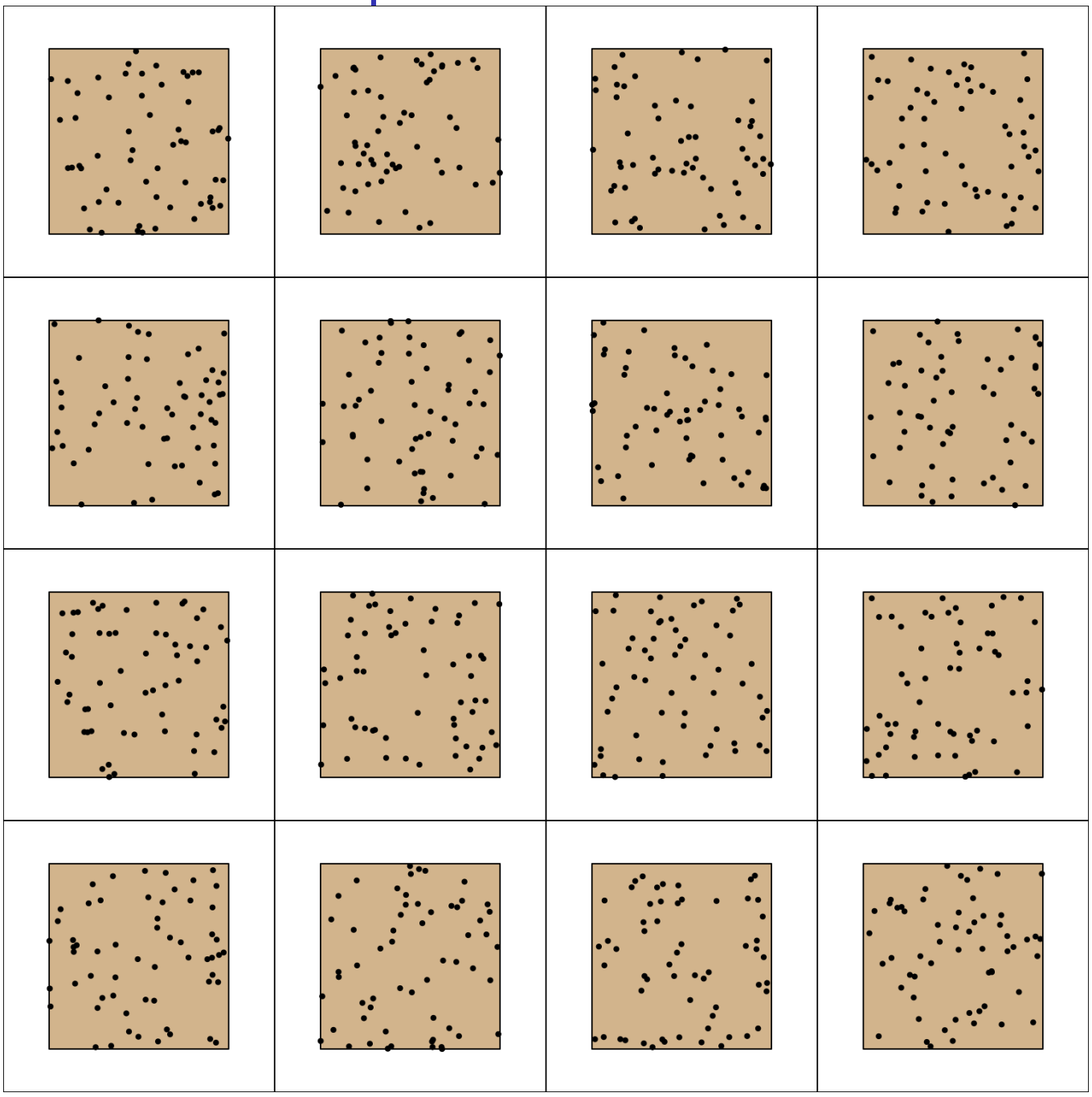
1. Randomly place 59 trees in the sampling area
2. Calculate mean nearest neighbour distance of random trees
3. Repeat steps 1 and 2 many times
4. Generate a null distribution of nearest neighbor distances

Monte Carlo on spatial data

Null Hypothesis: Mean nearest neighbour distance for a focal tree in the observed data will not differ significantly from the mean nearest neighour distance obtained from the same number of trees randomly distributed within the sampling area.

1. Randomly place 59 trees in the sampling area
2. Calculate mean nearest neighbour distance of random trees
3. Repeat steps 1 and 2 many times
4. Generate a null distribution of nearest neighbor distances
5. Compare actual mean nearest neighbour distance will null distribution

Monte Carlo on spatial data



Monte Carlo on spatial data