Biological Statistics III – Lecture 12

Computer based methods: randomization, bootstrap

January 13, 2020

Frank Miller, Department of Statistics, SU

frank.miller@stat.su.se Office: B736

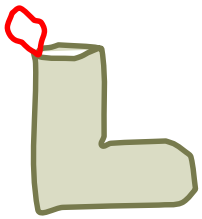
Stockholm University

Outline

- Bootstrap
  - Idea, method and examples
  - Bootstrap in the context of regression models
- Permutation tests
  - Idea, method and examples
  - The Mantel test in ecology

Why bootstrap?

- You have tested hypotheses or constructed confidence intervals (CIs) for parameters of interest (e.g. mean) assuming normal distribution with help of formulae
- Sometimes not reasonable to assume normal distribution
  - Using formulae for normal distribution might then mislead
  - With resampling methods discussed here, we can **obtain CIs without any distributional assumption**
- The available sample is our best information about the population – we take the **available sample as assumption for distribution of population**
- We pull ourselves up by our own capabilities – like "pulling us up from the mud by our own **bootstraps**"



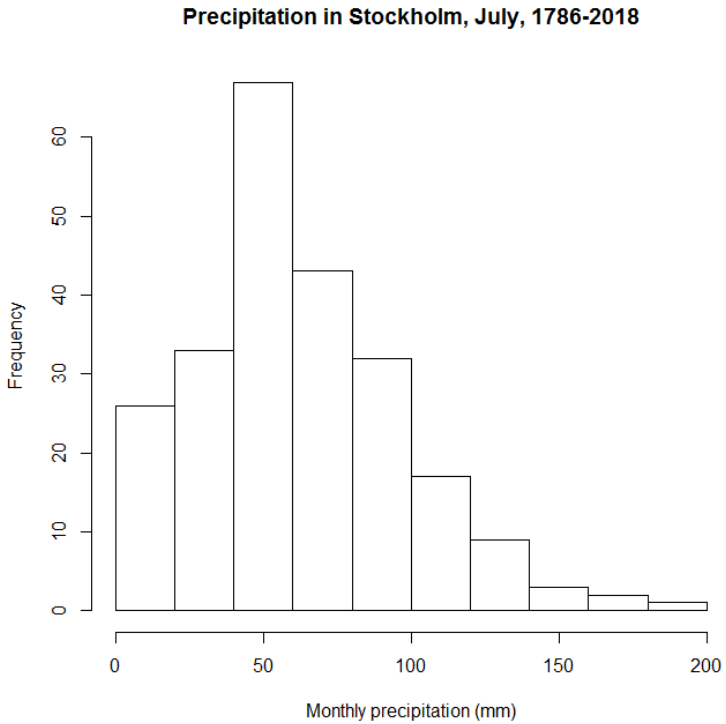
Analysis of precipitation data

- Rainfall data from July in 233 years in Stockholm
- What is the mean and a 95%-CI for the mean?
- With methods from earlier in course, we can calculate CI:

  $\bar{x} = 62.6mm, s = 35.0, n = 233,$

  $s_{\bar{x}} = s/\sqrt{n} = 2.29,$

  $t_{0.05,233} = 1.970$
- 95%-CI: $(58.1, 67.1)$



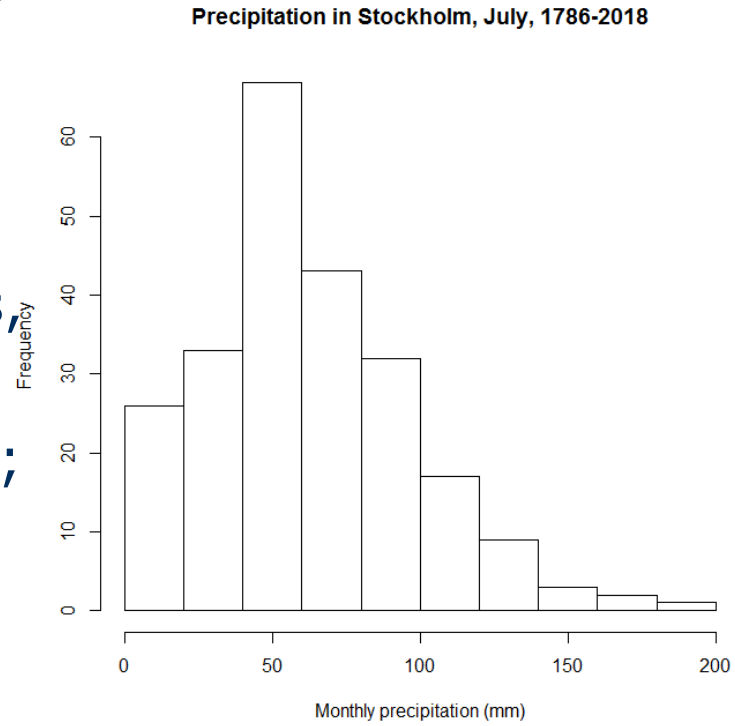
Data source: SMHI

- But: normal distribution assumed

Stockholm University

Analysis of precipitation data

- This data is not normally distributed
- We want to make a more realistic assumption: Actual sample distribution is best information about distribution
- Idea: Given the 233 observations, **sample from them with replacement** until you have 233; calculate mean;
  repeat this B=1000 times;
  we have now 1000 means: the ”middle 950” give a 95%-CI



Stockholm University

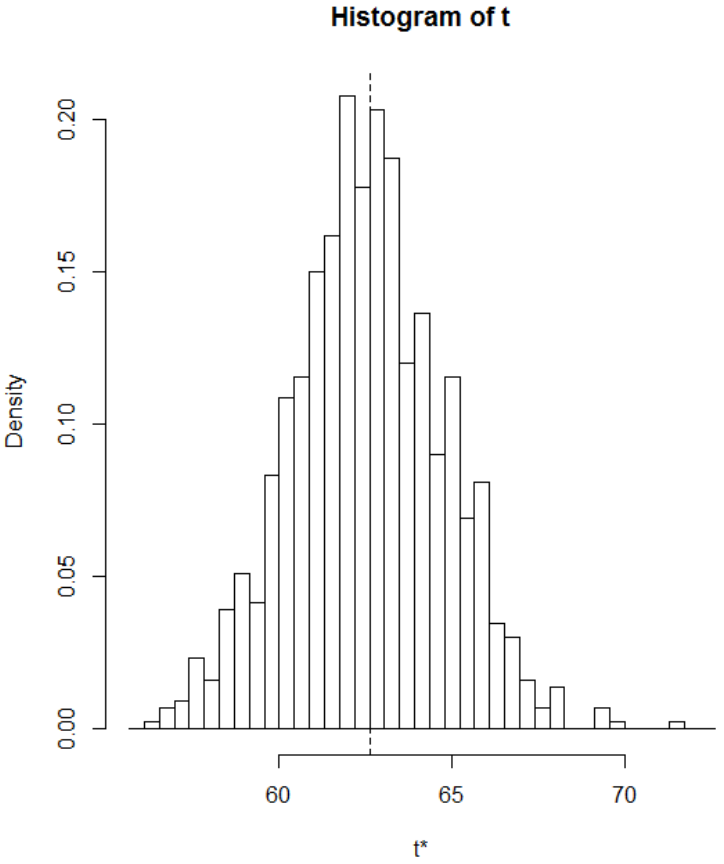
Analysis of precipitation data

- We illustrate the bootstrap using only the last 6 years: `42.3, 44.1, 91.9, 47.6, 14.6,  5.9`
- First resample: `5.9, 42.3,  5.9, 47.6, 91.9, 91.9`
- Second resample: `42.3, 44.1, 42.3, 91.9, 42.3, 14.6`
- Third resample: `47.6, 44.1, 42.3, 14.6, 91.9, 14.6`
- …
- 1000th resample: `47.6, 42.3, 91.9, 91.9,  5.9, 42.3`

- The mean of each resample: 47.6, 46.3, 42.5, …, 53.7

Analysis of precipitation data

- From the complete data, we made 1000 resamples; the 1000 means of those are in the histogram
- The mean of the means: 62.6 mm
  (bootstrap estimate is here the same as the usual estimate of the mean $\bar{x}$)
- The middle 95% of the means are from 58.2 to 66.7 – this is our 95%-bootstrap-CI for the mean



Stockholm University

Bootstrap idea

- Original data of size n is given, a certain property (e.g. mean, variance, …) is of interest and it's uncertainty should be quantified (e.g. CI)
- Draw B resamples of size n of the original data with replacement

  B=500 or 1000 has been used historically; B=10000 is nowadays no problem
- Usually, there are repetitions in a resample
- Calculate the property of interest for each resample – you have now B of them; the distribution of these B values ("bootstrap distribution") can be used e.g. to compute a CI
- Advantage: no assumption for distribution of original data

Bootstrap in R

- In R, package `boot` with it's functions `boot` and `boot.ci` can be used

  ```
  library(boot)
  ```

- Define first function of interest, e.g. the mean:

  ```
  bootmean <- function(x, i) mean(x[i])
  ```

- Generate B bootstrap resamples with function `boot`:

  ```
  bss <- boot(mrain, bootmean, R=1000)
  ```

- You can plot a histogram of the bootstrap distribution:

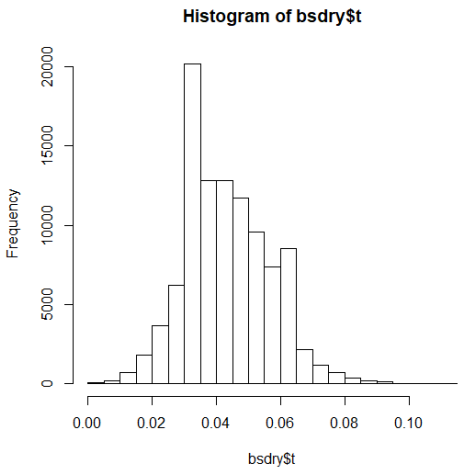
  ```
  hist(bss$t)
  ```

- A 95%-CI is between the 2.5%- and 97.5%-percentile of the bootstrap distribution:

  ```
  boot.ci(bss, type="perc")
  ```

Analysis of precipitation data

- What is an estimated probability for "less than 10mm rain in next July"? How good is our estimation? (➔ CI)
- Reasonable to calculate proportion of years with July-rain < 10mm. Here: in 10 of 233 years = 0.043
- To calculate a 95%-CI, we generate a bootstrap distribution
  (We resample B times and compute for each resample the proportion of years with July-rain < 10 mm)
- We use it's 2.5%- and 97.5%-percentile:
  `(0.0172, 0.0687)`
- Conclusion: The probability for < 10mm rain in July is between 1.7% and 6.9%; estimate is 4.3%
- (With normal assumption an estimate would be 6.6%. But a probability for < 0 mm rain would be 3.7%... To use bootstrap gives here much better estimates than with normal assumption! You get easily a confidence interval as well.)



Stockholm University

Bootstrap in R

- Define function of interest, here proportion below 10mm:

```
bootdry <- function(x,i) mean((x[i]<10))
```

- Generate B=100000 bootstrap resamples:

```
bsdry <- boot(mrain, bootdry, R=100000)
```

- Plot a histogram of bootstrap distribution:

```
hist(bsdry$t)
```

- Estimate proportion:

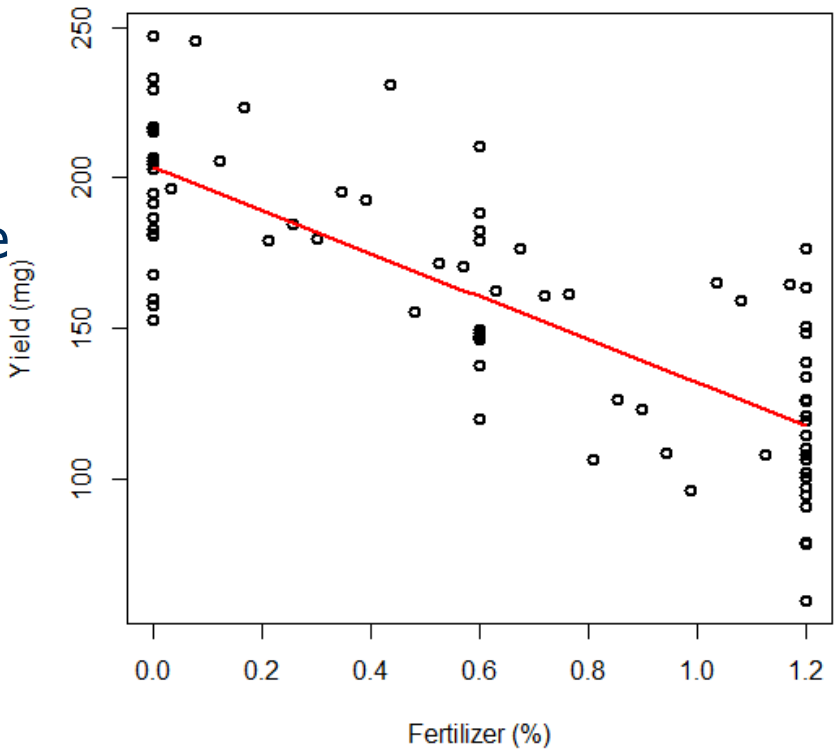
```
bootdry(mrain)
```

- A 95%-CI is between the 2.5%- and 97.5%-percentile of the bootstrap distribution:

```
boot.ci(bsdry, type="perc")
```

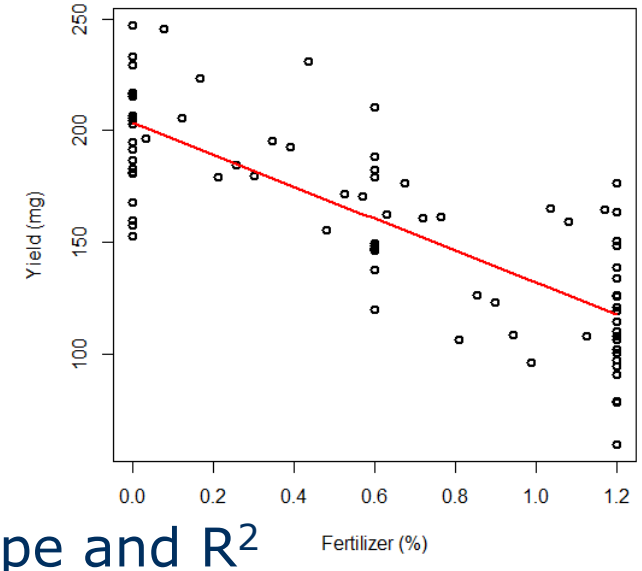
Bootstrap for regression models

- We can use the bootstrap method very flexibly, e.g. **in linear regression** if we want a **CI for the slope or for R^2**
- Example: The (toxic) influence of a fertilizer on growth of garden cress was investigated in an experiment (yield vs. amount of fertilizer, n=81)
- Estimated linear regression: $yield = 203.3 - 71.3 \cdot fertilizer$ with R^2=0.63
- CI for slope? CI for R^2?



Stockholm University

Bootstrap for regression models



- The dataset has n=81 pairs of fertilizer-yield-values
- The bootstrap resamples **n pairs** with replacement, computes regression-slope and R^2
- This is done B times; R-code:

```
cressdat <- data.frame(fertilizer,yield)
cmslope  <- function(dat, i)
{
  cm  <- lm(yield~fertilizer, subset=i, data=dat)
  coef(cm)[2]
}
cb <- boot(cressdat, cmslope, R=10000)
boot.ci(cb, type="perc")
```

- Result for CI-limits: -83.5, -58.7

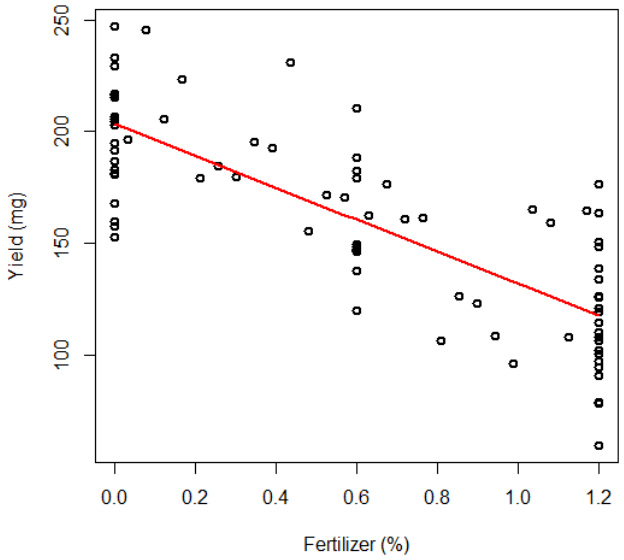
Stockholm University

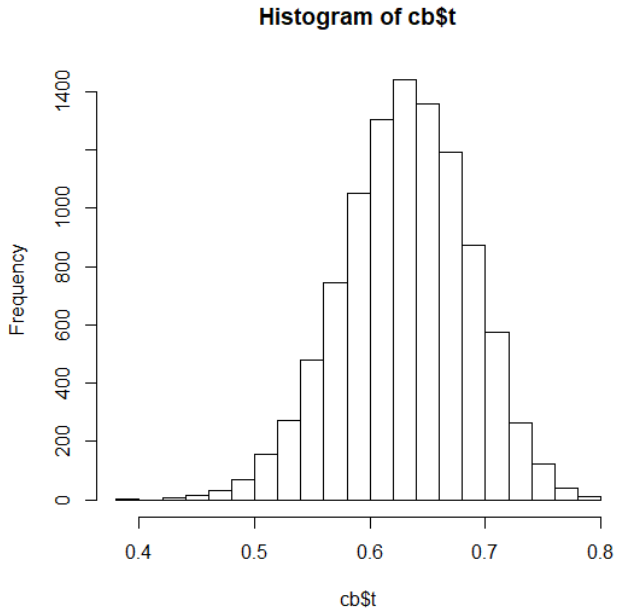
Bootstrap for regression models

- A function for analysis of regression-R^2 instead of slope is:

```
cmr2  <- function(dat, i)
{
  cm  <- lm(yield~fertilizer, subset=i, data=dat)
  summary(cm)$r.squared
}
```

- Result for CI-limits: 0.518, 0.733

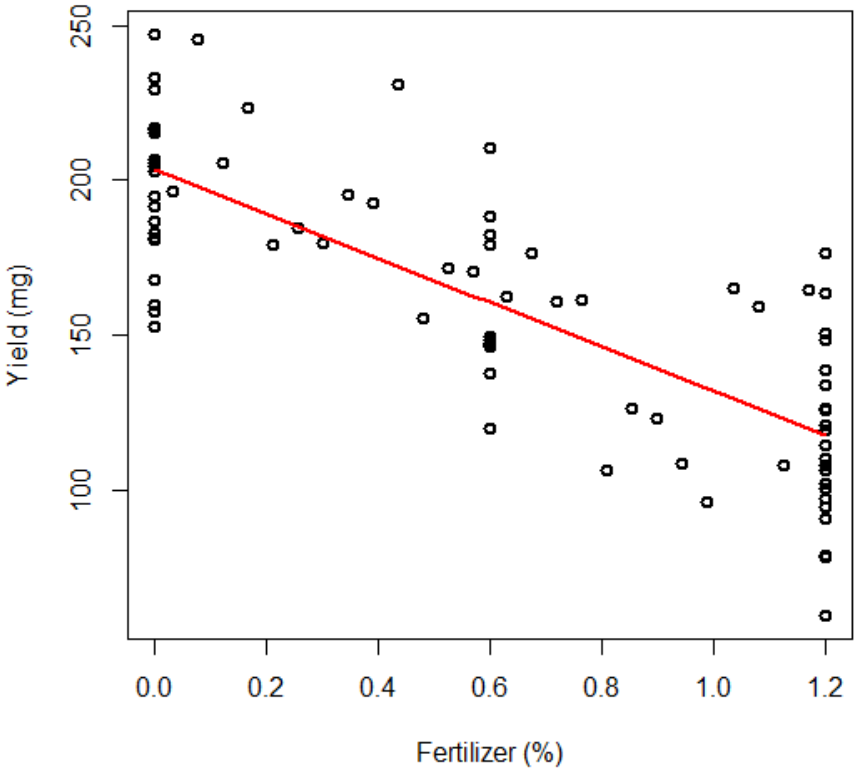




Stockholm University

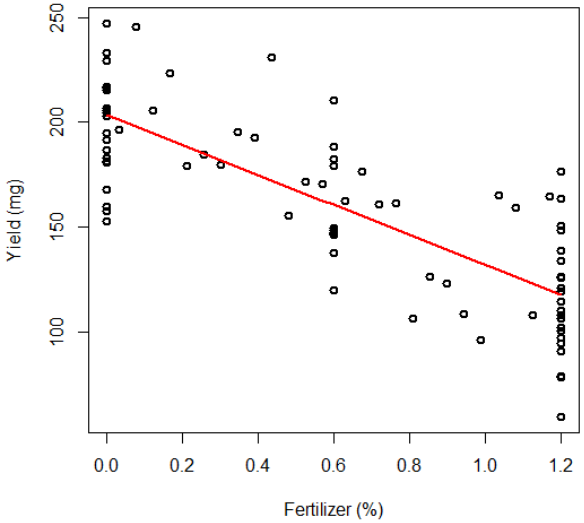
Permutation tests

- Want to **test if there is association between two variables**
- Example: Is there an association between amount of fertilizer and cress yield?
  - Is the *slope* significantly different from 0?
  - Alternative: Is the *correlation* significantly different from 0?
- We could perform t-test from linear regression but we want to avoid the assumptions (here avoid normality assumption, in other examples independence)

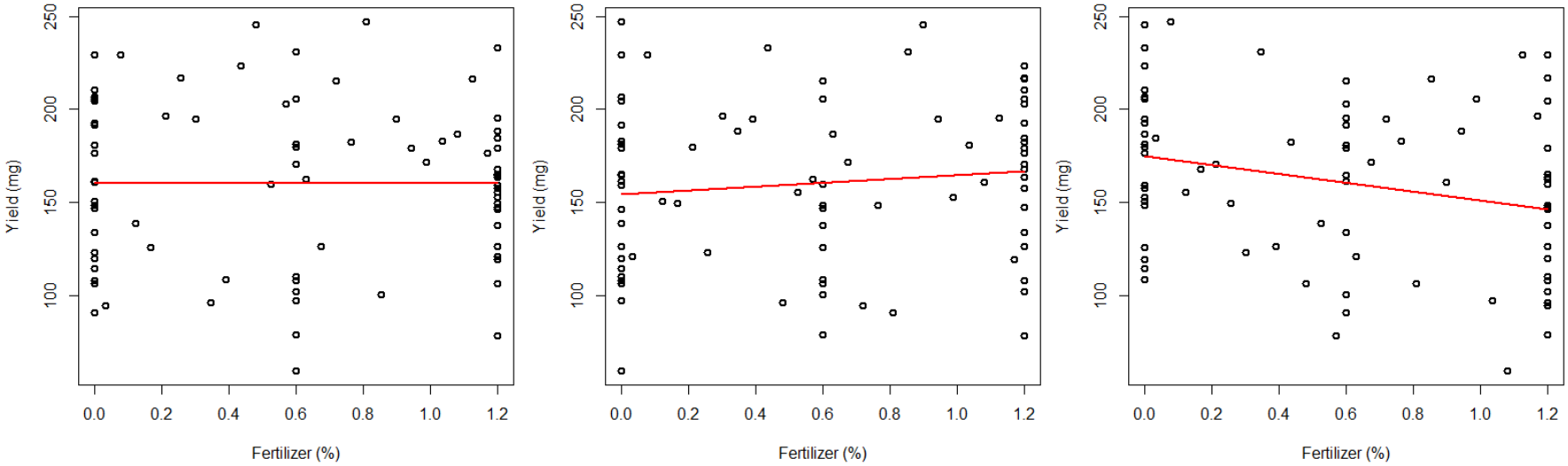


Stockholm University

Permutation tests

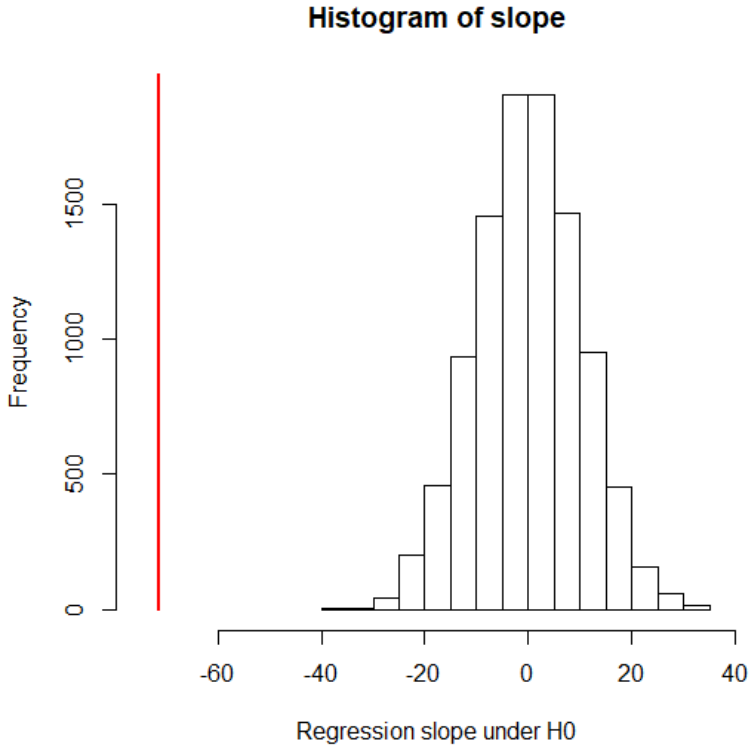


- Idea: If we permute yield-results (assign them randomly to fertilizer-values), we have no association, but we compute a slope (=chance-slope)
- We do this repeated times (e.g. 10000) and obtain a distribution for chance-slopes; three of the chance-slopes:

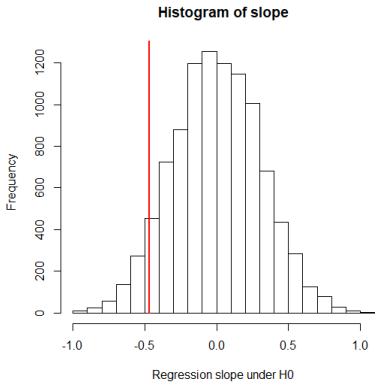


Permutation tests

- If the observed slope is different from the chance-slopes, we conclude that there is a real association
- Here: It is evident that the real slope (-71.3) is not by chance



- In general: We calculate proportion of resamples which give a more extreme slope than the real one
- Proportion is the p-value: conclude that there is an association if $p<0.05$



Stockholm University

Permutation tests – analysis with R

- The function `independence_test` in the package `coin` can be used:

```
library(coin)
independence_test(yield ~ fertilizer,
alternative="less")
```

- Result:

```
        Asymptotic General Independence Test

data:  yield by fertilizer
Z = -7.0994, p-value = 6.265e-13
alternative hypothesis: less
```

- Conclusion: Since p-value<0.05, we conclude that there is an association between fertilizer and yield

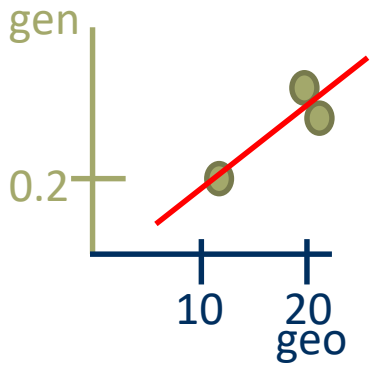
Permutation tests: The Mantel test

- Association between genetic and geographic distances?
- Genetic distances:

	A	B	C
A		0.2	0.35
B	0.2		0.4
C	0.35	0.4	

Geographic distances:

	A	B	C
A		12	20
B	12		19
C	20	19	



- Compute correlation between genetic and geographic dist.
- Permute one of the distance matrices e.g. 10000 times; compute correlation for all of them (=chance-correlations)
- Is real correlation different from chance-correlations?
- This specific permutation test is the Mantel test

Stockholm University

Permutation tests: The Mantel test

- Why do we not just compute the correlation for the real data and test by "usual" test if it is 0 or not?
  - Assumption of independence is violated
  - Usually also normality questionable
- The Mantel test avoids these problems
- To read more about the Mantel test:
  - p.597-608 in Legendre P, Legendre L (2012). Numerical Ecology. 3rd edition.
  - Urban DL (2003). Spatial analysis in ecology – Mantel's test. URL: https://www.nceas.ucsb.edu/files/scicomp/doc/SpatialEcologyMantelTest.pdf

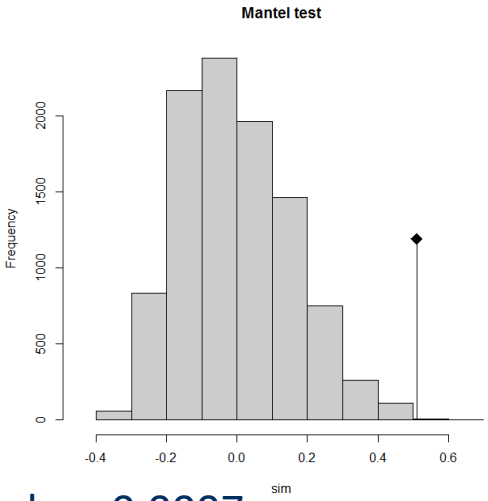
The Mantel test – analysis with R

- Function `mantel.randtest` in package `ade4` can be used:

```
gen <- as.dist(geneticdist)
geo <- as.dist(geographicdist)
gg  <- mantel.randtest(geo, gen, nrepet = 10000)
gg
plot(gg, main = "Mantel test")
```

- Example: distance matrices for 19 villages of Yanomama Indians in `ade4`; can be analysed by running first following:

```
data(yanomama)
geneticdist    <- yanomama$gen
geographicdist <- yanomama$geo
```



p-value=0.0007

Stockholm University