Computational statistics, lecture 5

Frank Miller, Department of Computer and Information Science, Linköping University

February 18, 2025

LINKÖPING UNIVERSITY

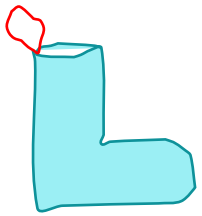
Today

(Literature: Givens and Hoeting, 9.1-9.3, 9.8; Gentle, 12, 13)

- Bootstrap
  - Nonparametric bootstrap
  - Jackknife
  - Parametric bootstrap
- Hypothesis tests
- Permutation test

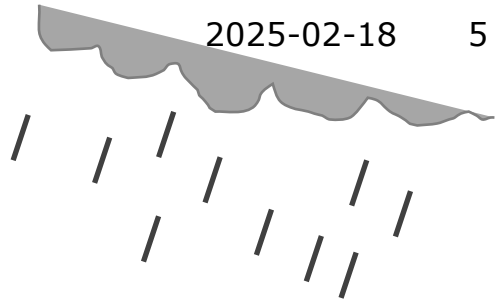
Why bootstrap?

- Assume you have independent samples of some population
- In statistics, we have methods to construct confidence intervals (CIs) for a parameter θ of interest (e.g., mean) based on distributional assumptions; e.g., explicit formulas exist in case of normal distribution
- Sometimes not reasonable to make distributional assumptions
- Aim here: **obtain CIs without these distributional assumption**
- We take the **available sample as assumption for distribution of population** and **resample** from it
- We pull ourselves up by our own capabilities –
  like “pulling us up from the mud by our own **bootstraps**”



Bootstrap method

- Observed data: $D = (X_1, \dots, X_n)$
- Of interest: An estimator $T(D) = \hat{\theta}$ for some parameter θ and its uncertainty (e.g., CI for θ)
- Draw B resamples $D_i^* = (X_1^*, \dots, X_n^*)$ **of size n** from original data D **with replacement**
  - $B = 500$ or 1000 has been used historically; $B = 10000$ is nowadays often no problem
  - Usually, there are repetitions in a resample
- Calculate the property of interest for each resample: $\hat{\theta}_i = T(D_i^*), i = 1, \dots, B$
- The distribution of these B values (”bootstrap distribution”) gives information about distribution of $T(D)$
  - E.g., a CI for θ can be computed

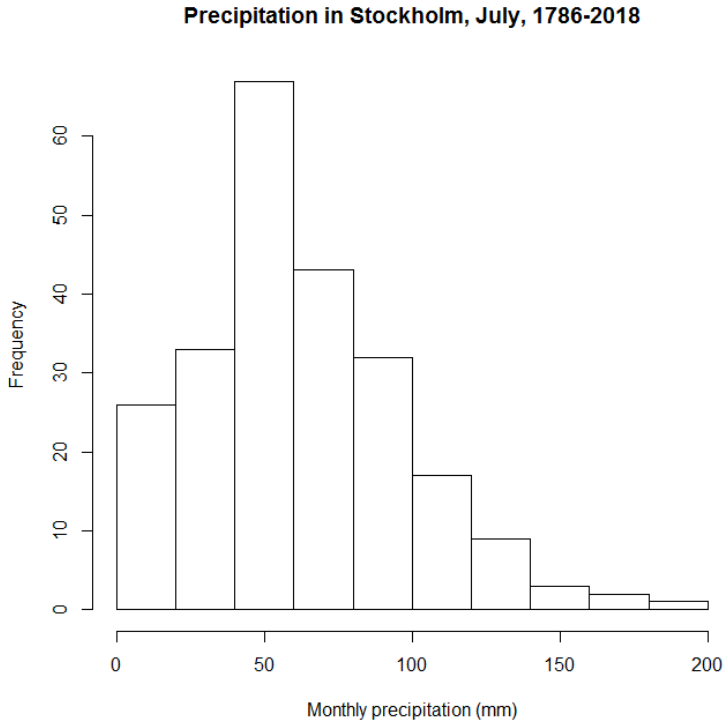


Example: precipitation data

- Rainfall data from July in 233 years in Stockholm
- What is the mean and a 95%-CI for the mean?
- A standard formulae for the CI assumes that data is normally distributed and uses therefore the t-distribution:

$$\bar{x} = 62.6mm, s = 35.0, n = 233,$$
$$s_{\bar{x}} = s/\sqrt{n} = 2.29,$$
$$t_{0.025,233} = 1.970$$

- 95%-CI-bounds: $\bar{x} \pm s_{\bar{x}} \cdot t_{0.025,233}$; here: $(58.1, 67.1)$
- But data here is not normally distributed
- Now, we construct a CI using the bootstrap method



Data source: SMHI

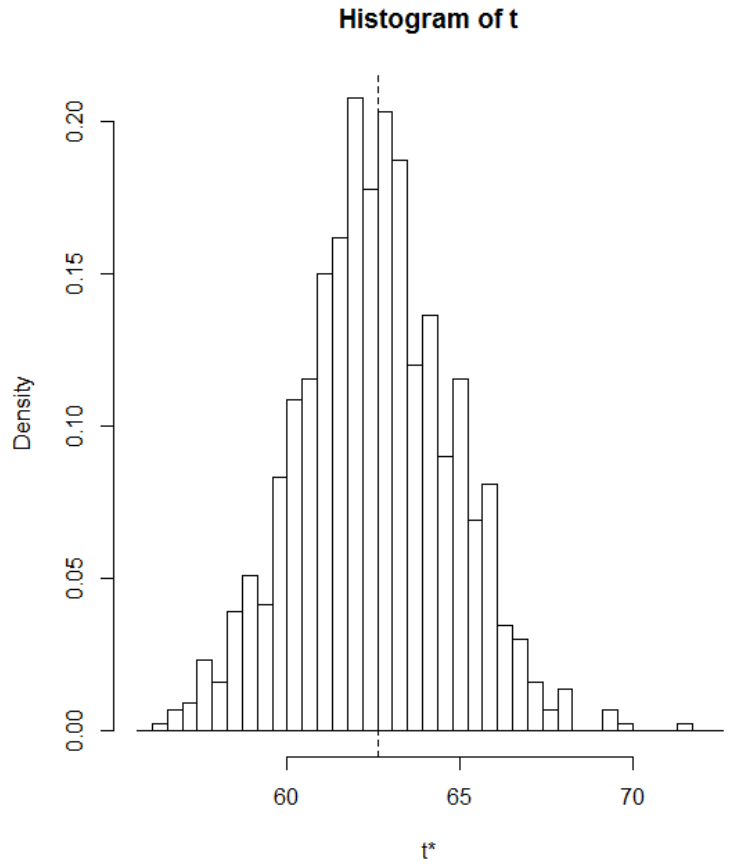
Example: precipitation data

- 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`
- ...
- B-th 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

Example: precipitation data

- From the complete data, we made $B = 1000$ 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
  This is: limits are the 2.5% and 97.5% percentiles
- This way to define the CI is called **percentile method**



Bootstrap in R

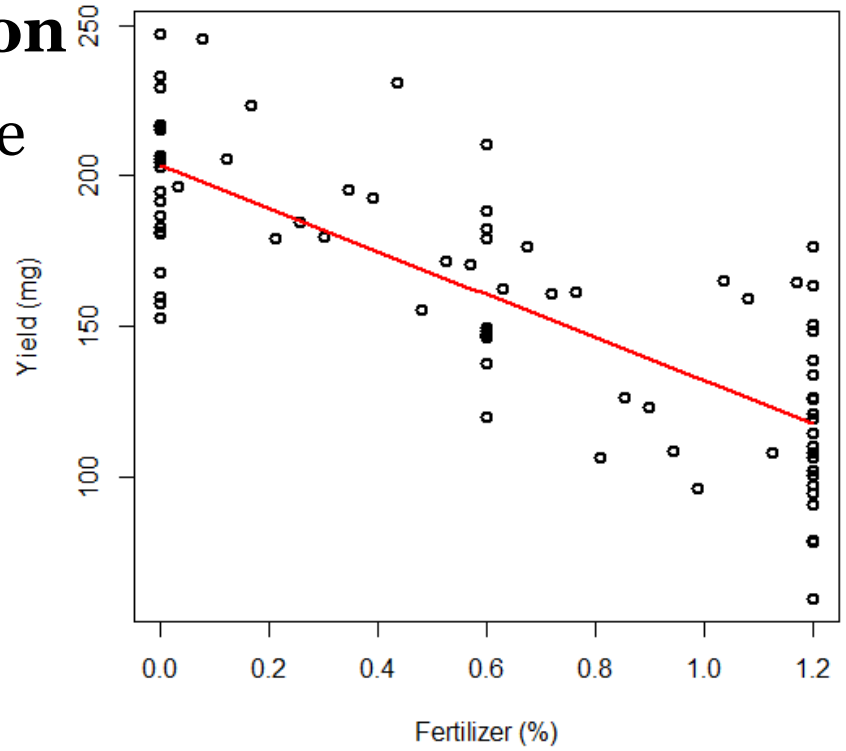
- R code using a loop for bootstrap replicates:

```
bo <- 1000   # bootstrap replicates
bs <- c()    # to save the results for the means
for (l in 1:bo){
  x  <- sample(mrain, size=length(mrain), replace=TRUE)
  bs <- c(bs, mean(x))
}
hist(bs)
bss  <- sort(bs)
ci95 <- c(bss[round(bo*0.025)], bss[round(bo*0.975)])
ci95
```

- Running this code gave the 95% bootstrap confidence interval (58.2, 66.7)
- Alternatively, the package `boot` with functions `boot` and `boot.ci` can be used (see R-code on homepage)

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 the residual standarddeviation**
- Example: Experiment about the (toxic) influence of a fertilizer on the growth of garden cress (yield vs. amount of fertilizer, $n = 81$)
- Estimated linear regression:
$$yield = 203.3 - 71.3 \cdot fertilizer$$
with residual standarddeviation $\hat{\sigma} = 26.7$
- CI for slope? CI for $\hat{\sigma}$?

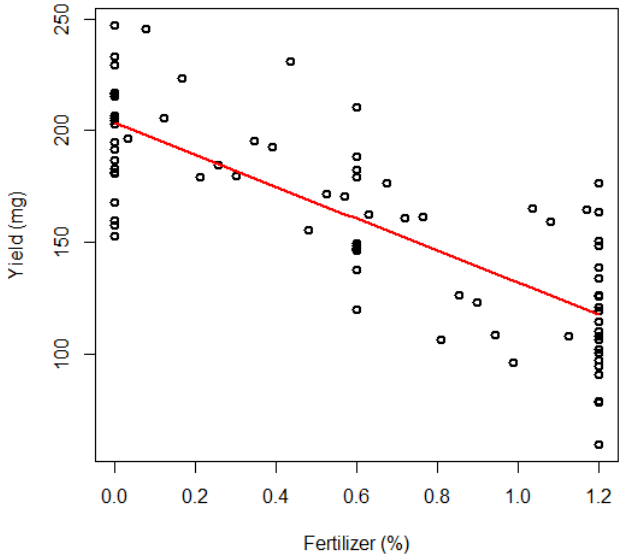


Bootstrap for regression models

- The dataset has $n = 81$ pairs of fertilizer-yield-values
- The bootstrap resamples ***n* pairs** with replacement, computes regression-slope and $\hat{\sigma}$
- 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")
```

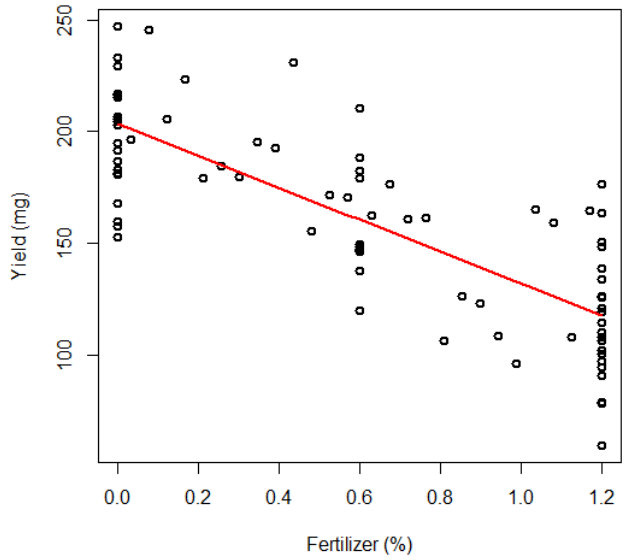
- Result for CI-limits: -83.8, -59.1



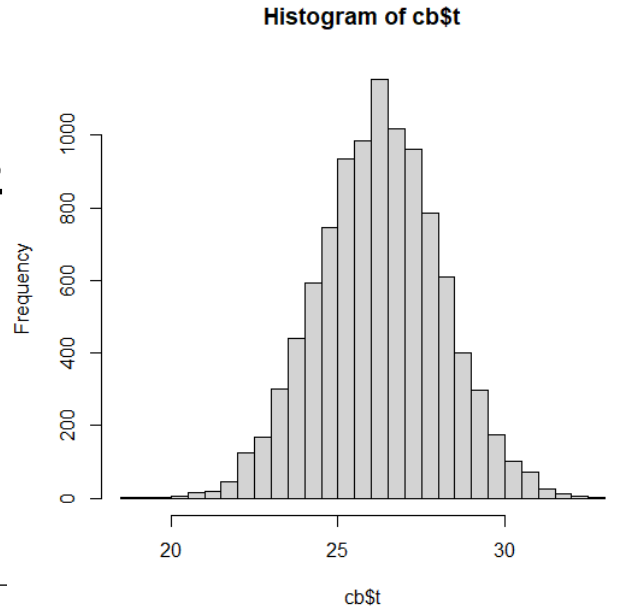
Bootstrap for regression models

- A function for analysis of the residual $\hat{\sigma}$ is:

```
cmressd <- function(dat, i){
  cm  <- lm(yield~fertilizer, subset=i, data=dat)
  summary(cm)$sigma
}
```



- Result for CI-limits: 22.62, 29.89 (percentile method)
- Median (50% percentile) of bootstrap distribution: 26.32
- Residual $\hat{\sigma}$ of data: 26.72
- Percentile CI is constructed around 26.32 while it should be constructed around 26.72 ➔ the CI is biased



Percentile method for CIs and alternatives

- The percentile method can have drawbacks
  - Bias: Estimate $\hat{\theta}$ might be very different from median of bootstrap distribution, median($\hat{\theta}_i$), but we would like a CI constructed around $\hat{\theta}$
  - The bootstrap distribution might be skewed implying that the se($\hat{\theta}$) changes with the true θ
- The BC_a method (bias correction – accelerated) improves the percentile method by
  - correcting for bias and
  - adjusting the boundary alpha-levels to handle dependence of se($\hat{\theta}$) on θ
- If bootstrap distribution has not these issues, BC_a = percentile
- For other methods (and BC_a) see Givens and Hoeting (2013), Chapter 9.3.

Jackknife

- Observed data: $D = (X_1, \dots, X_n)$
- Of interest: An estimator $T(D)$ for some parameter
- n resamples defined as $D_i^* = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ (leave-one-out sample)
- $T(D_1^*), \dots, T(D_n^*)$ give information about distribution of $T(D)$

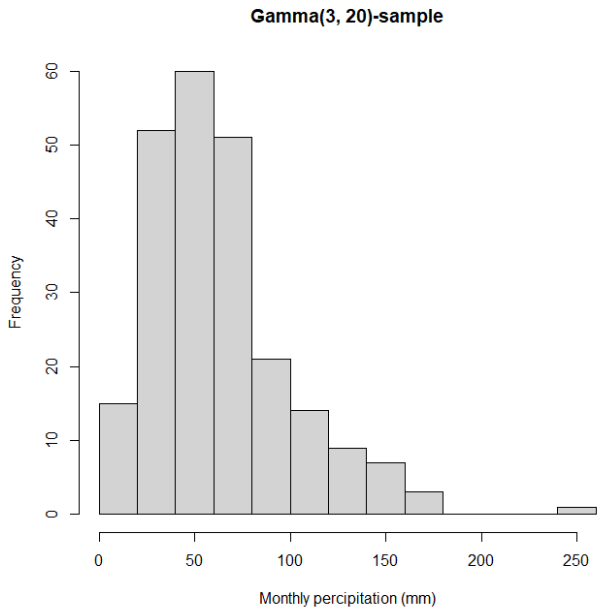
- Jackknife variance estimation for $T(D)$:

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} (T(D_i^*) - J)^2, \text{ where } J = \frac{1}{n} \sum_{i=1}^{n} T(D_i^*)$$

- Important application both for Jackknife and bootstrap is variance estimation
- Jackknife is resampling method like bootstrap, but it is deterministic

Parametric bootstrap

- When a parametric model for the data is known or believed to represent the reality well, we can do parametric bootstrap and sample according to the assumed model
- Example: We assume that monthly precipitation in July follows a Gamma(3, 20)-distribution
- We sample 233 datapoints from Gamma(3, 20) and calculate parameter of interest
- Do this B times and derive e.g. a confidence interval



Recap: Hypothesis testing

- Given n observations $X_1, \dots, X_n$ with mean μ
- Test H_0: $\mu = \mu_0$ versus H_1: $\mu < \mu_0$
- (Here: one sample problem)

	H_o is false	H_o is true
Reject H_o	√	Type I error
Accept H_o	Type II error	√

Power = 1 - type II error
Type I error should be limited, $\leq \alpha$

- Example: bakery is baking breads supposed to have 750 g, each; n breads measured $(X_1, \dots, X_n)$; assumption $X_i \sim N(\mu, \sigma^2)$; question H_0: $\mu = 750$ or H_1: $\mu < 750$?

known

- Test statistic, e.g., $T(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Reject H_0 if $T(X) < c_\alpha$ (i.e., if $T(X)$ unlikely under H_0)

Ex.: properties of test assuming data distribution

- Given n independent and identically distributed observations $X_1, \ldots, X_n$ with mean μ, one can test H_0: $\mu = 0$ versus H_1: $\mu > 0$ with the one-sample t-test

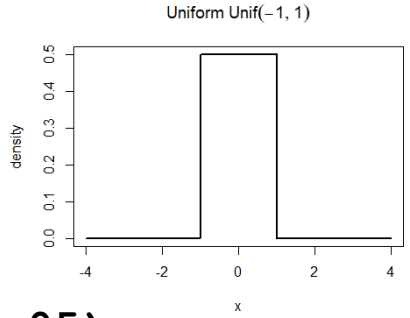
$$\text{reject } H_0 \text{ if and only if } \frac{\sqrt{n}\bar{x}}{s_x} > t_{n-1;1-\alpha}$$

- Assumption for test: normal distribution of observations with unknown variance
- How sensitive is t-test if observations not normal?
- We focus on H_0 first: Can type I error be larger than α (such that it matters) for certain distributions?
- Idea:
  - Choose some distributions with mean=0, simulate n repetitions, perform t-test, and record if rejected
  - Repeat this s times and check rejection rate

Ex.: properties of test assuming data distribution

- For $n = 10$, simulate rejection rate for Unif[-1,1]

```
s      <- 100000
n      <- 10
count <- 0
for (sim in 1:s){
  x      <- runif(n, min = -1, max = 1)
  reject <- (t.test(x, alternative = "greater")$p.value < 0.05)
  count  <- count + reject
}
#Rejection rate estimate:
rre    <- count/s
```



This is 1 if the condition in (...) is true, otherwise it is 0

- Note that there are possibilities to make simulation more efficient, e.g., avoiding the loop
- Precision of result?

Ex.: properties of test assuming data distribution

```
s     <- 100000
n     <- 10
count <- 0
for (sim in 1:s){
  x      <- runif(n, min = -1, max = 1)
  reject <- (t.test(x, alternative = "greater")$p.value < 0.05)
  count  <- count + reject
}
rre   <- count/s
```

- Precision of result?

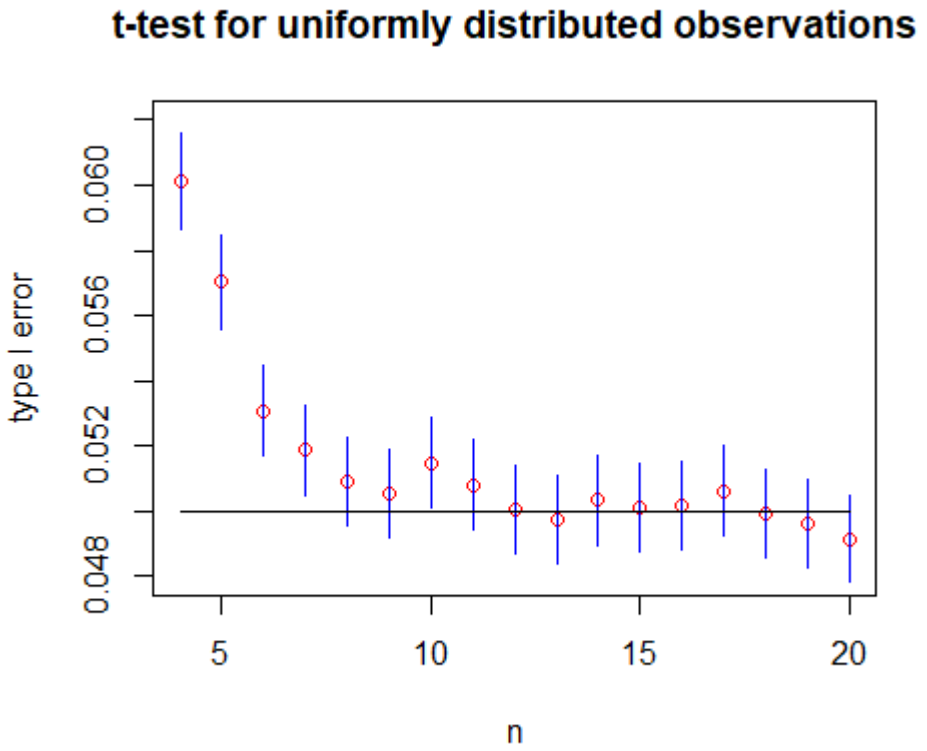
p = true rejection rate; reject$\sim Bin(1,p)$, count$\sim Bin(s = 100000, p)$

$$\text{Var}(\text{count}) = p(1-p)s, \text{Var}\left(\frac{\text{count}}{s}\right) = \frac{p(1-p)}{s}, \text{sd}(\text{rre}) = \sqrt{\frac{p(1-p)}{s}}$$

≈ 0.0007 for $p = 0.05$.

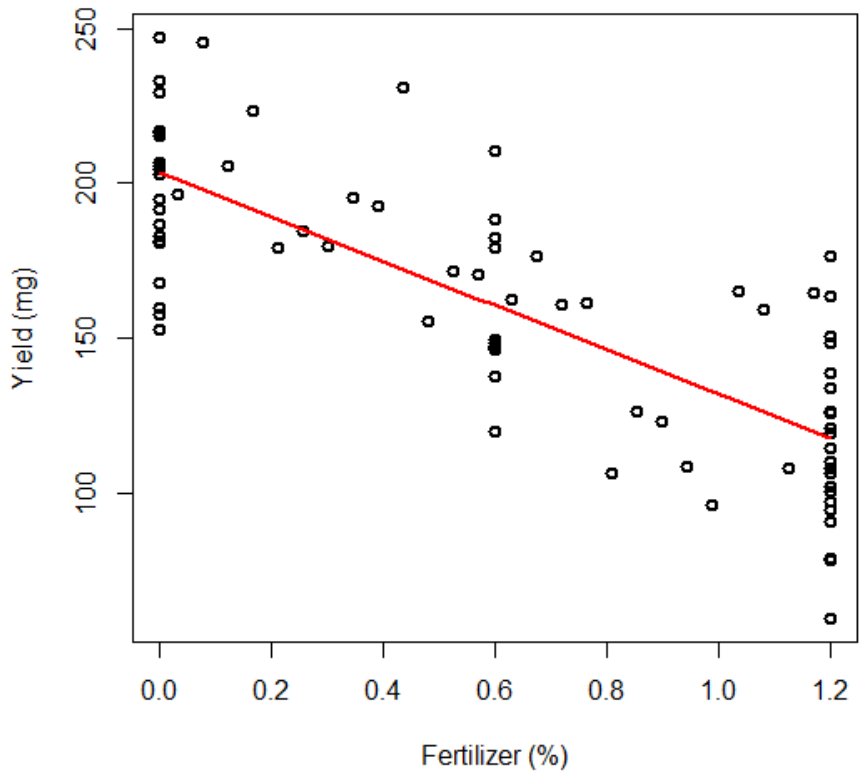
Ex.: properties of test assuming data distribution

- Simulated rejection rate for Unif[-1,1] for $n = 4, 5, \ldots, 20$ with 95%-simulation-error-CIs based on 100 000 simulations for each n
- One more loop for n used
- Took ~1 min to simulate

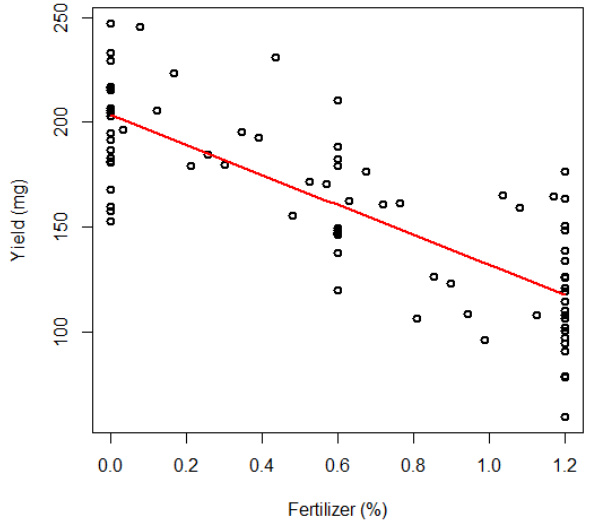


Permutation tests

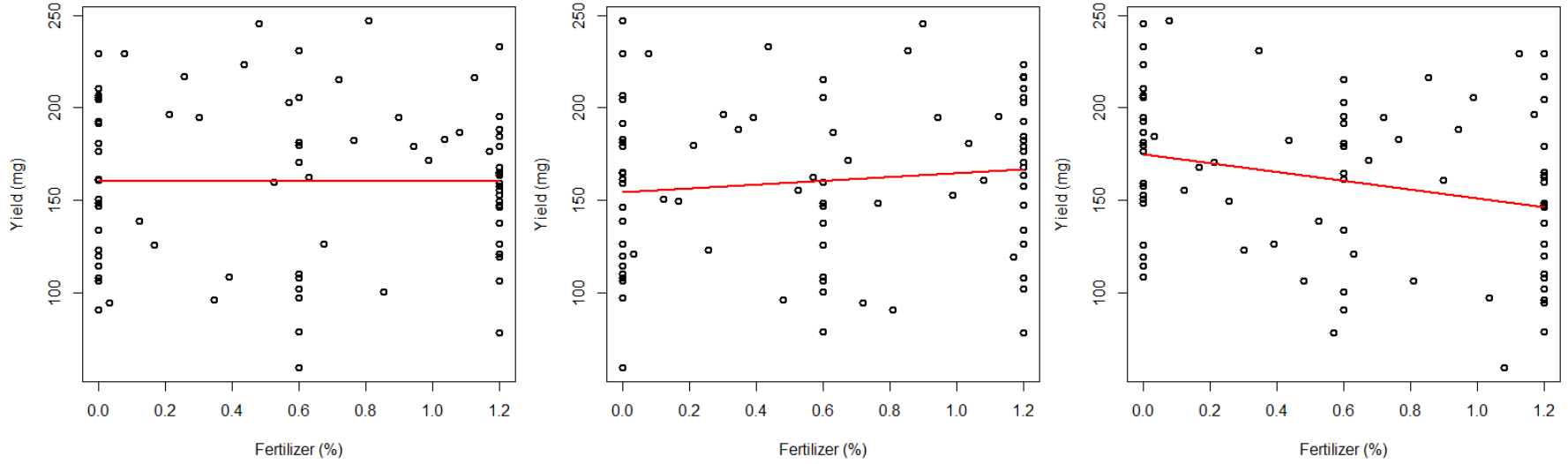
- Want to **test if there is association between two variables**
- Example: Is there an association between amount of fertilizer and cress yield?
  - Is the *slope* in the regression model significantly different from 0?
- We could perform t-test from linear regression, but we want to **avoid the assumptions**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



Permutation tests

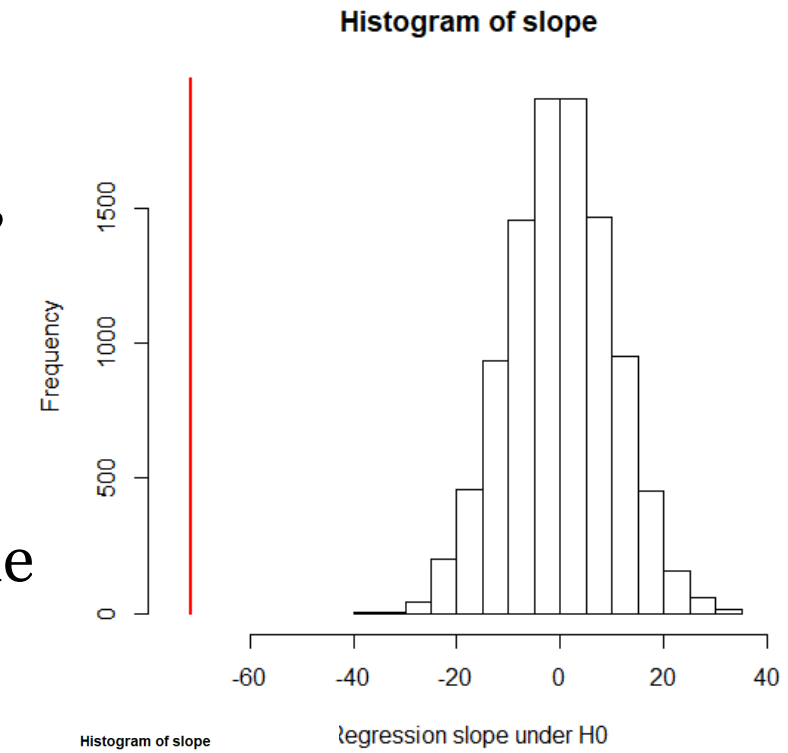


- Idea: If we permute yield-results (assign them randomly to fertilizer-values), we have no association, but we compute a slope (=chance-slope)
- We do this repeated times (e.g. 10000) and obtain a distribution for chance-slopes; three of the chance-slopes:



Permutation tests

- If observed slope different from chance-slopes, conclude that association is real
- Here: Evident that real slope (-71.3) not by chance
- In general: We calculate proportion of resample more extreme slope than the real one



- Proportion is the p-value: conclude that there is an association if $p < 0.05$