107.330 Statistische Simulation und computerintensive Methoden

Bootstrap Methods I

Alexandra Posekany

WS 2020

Monte Carlo methods and bootstrap

The Monte Carlo methods discussed so far are useful to get an idea about the performance of statistical methods under various models and assumptions or to verify theoretic properties.

They are however not that often useful in practical data analysis.

A special class of Monte Carlo methods are bootstrap methods which we will consider in the following.

The main idea of bootstrap is to use simulation techniques **without** having to specify the distribution of the data.

Bootstrap

Bootstrapping as suggested by Efron 1979 revolutionized computational statistics.

He named it after "The Surprising Adventures of Baron Munchhausen" who told the tall tale of pulling himself and his horse out of a swamp by his hair. In English "to pull oneself up by one's bootstraps" is proverbial based on this story.

Notion of Bootstrapping

Since then bootstrapping became an important part of the statistical toolbox.

The notion of bootstrapping is that inference about a population from sample data can be modelled by resampling the sample data and performing inference about a sample from resampled data.

The bootstrap is related to many methods like the permutation methods, jackknife, cross validation, ...

The bootstrap makes many "difficult" things "easy", it is does however not always work.

We will look at the bootstrap in the context of estimation and hypothesis testing for iid data, linear regression and time series analysis.

Prelimenaries

Consider in the following having an iid sample $x_1, \ldots, x_n$ coming from a distribution with cdf F.

For now consider we the problem to make inference about a population characteristic θ. Let then T denote a statistic whose value when computed for the sample is t.

In the context of bootstrapping the interest is on the probability distribution of T.

Like what are its - bias - standard errors - quantiles - likes values in the context of hypothesis testing

And then, how can we get for example based on this knowledge confidence intervals for θ?

Parametric models

In parametric models F is fully specified by a finite set of (adjustable) parameters Ψ.

In this case θ is either a component of Ψ or a function of Ψ.

Parametric models have then many different well-established methods to estimate Ψ/θ .

E. g. for a Gaussian model, $F = \mathcal{N}(\mu, \sigma^2)$ with parameters $\Psi = (\mu, \sigma^2)$. Then, θ can be the mean μ or any function of the parameter such as the t-test statistic $t = \frac{\bar{x}_n - \mu_0}{\sigma}$.

Sampling methods

The representation of $\theta = T(F)$ defines the parameter and estimator in way that no assumptions on F are made other than θ needs to exist.

For example the trimmed mean defines the "mean"önly for symmetric data. Hence usually characteristics of the underlying distributions need to be made to ensure that the quantity of interest is well defined.

Due to this broad specification it is then often challenging to study the theoretical properties of T.

In the following sampling methods will be investigated for this purpose.

Parametric simulation

It is assumed that the original data set is a realization of a random sample from the parametric distribution F with a finite set of (adjustable) parameters Ψ. In this case a parametric model is fitted by a parameter θ, which is either a component of Ψ or a function of Ψ, often the maximum likelihood estimator. Then, samples of random numbers are drawn from this fitted model with the parameter estimate obtained from the original sample.

Parametric sampling is then to use the fitted model to draw m samples of sizes n and compute for each sample t and then study the distribution of the m estimates $\hat{t}_1, \ldots, \hat{t}_m$.

Example of parametric approach

- assumption: data follow (approximately) a normal distribution
- we can then estimate $\overline{x}$ and s_n^2 as approximations of mean μ and variance σ^2
- based on this we then simulate artificial samples from the $\mathcal{N}(\overline{x}, s_n^2)$ distribution and calculate estimators based on this

non-paramteric models

When no parametric model is used the statistical analysis is **non-parametric**.

Then usually just general assumptions are made like:

- iid sample
- symmetry
- unimodality
- and additional ones depending on the question and scenario

Again, there are many non-paramteric methods usually to infer about Ψ / θ.

Often also for the same data results of a parametric analysis are compared to the non-paramteric results to assess the robustness of the conclusion.

ECDF

An important role in non-paramteric statistics plays the empirical distribution which puts the equal probability of n^{-1} on each sample value x_i.

The empirical distribution (ecdf) $\hat{F}$ is the natural estimator of F and is defined as:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} H(x - x_i),$$

where $H(u)$ is the Heaviside function, a step function with jumps from 0 to 1 when $u = 0$. Hence the values of the ecdf are fixed as $(0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n})$.

non-paramteric simulation

In a non-paramteric context there is no real "fitted model" and the ecdf has to play its role.

Sampling and from the empirical distribution is quite easy as it is puts equal weight on all observed data points and in this case then using the inverse method to get random samples equals resampling the data points with replacement.

Then again m samples of sizes n are drawn and for each sample t is computed and then the distribution of the m estimates $\hat{t}_1, \ldots, \hat{t}_m$ studied.

This is often called non-paramteric bootstrap.

Bootstrap sampling

Given a data set x of size n, the general strategy for nonparametric bootstrapping is:

1. Identify the parameters of interest θ.
2. Compute the parameters for the given data set, denoted $\hat{\theta}$.
3. Generate m "data sets' ' x^i of length n by resampling with replacement from x. x^i is often called the bootstrap sample.
4. Compute for each data set x^i the corresponding bootstrap estimate $\hat{\theta}^i$, $i = 1, \ldots, m$.

The choice of m depends then mainly of what needs to done using the bootstrap estimates.

Difference parametric - nonparametric sampling

- nonparametric sampling has an anavoidable discreteness. Hence the bootstrap statistics have a discrete distribution (however approximating a continuous distribution).
- nonparametric sampling has problems when the data has outliers - as these might appear even more often in a bootstrap sample.

When to use bootstrapping?

- **When the theoretical distribution of a statistic of interest is complicated or unknown**.

  Non-parametric bootstrapping is distribution-independent and allows assessing the properties of the distribution underlying the sample and estimators derived from this distribution.

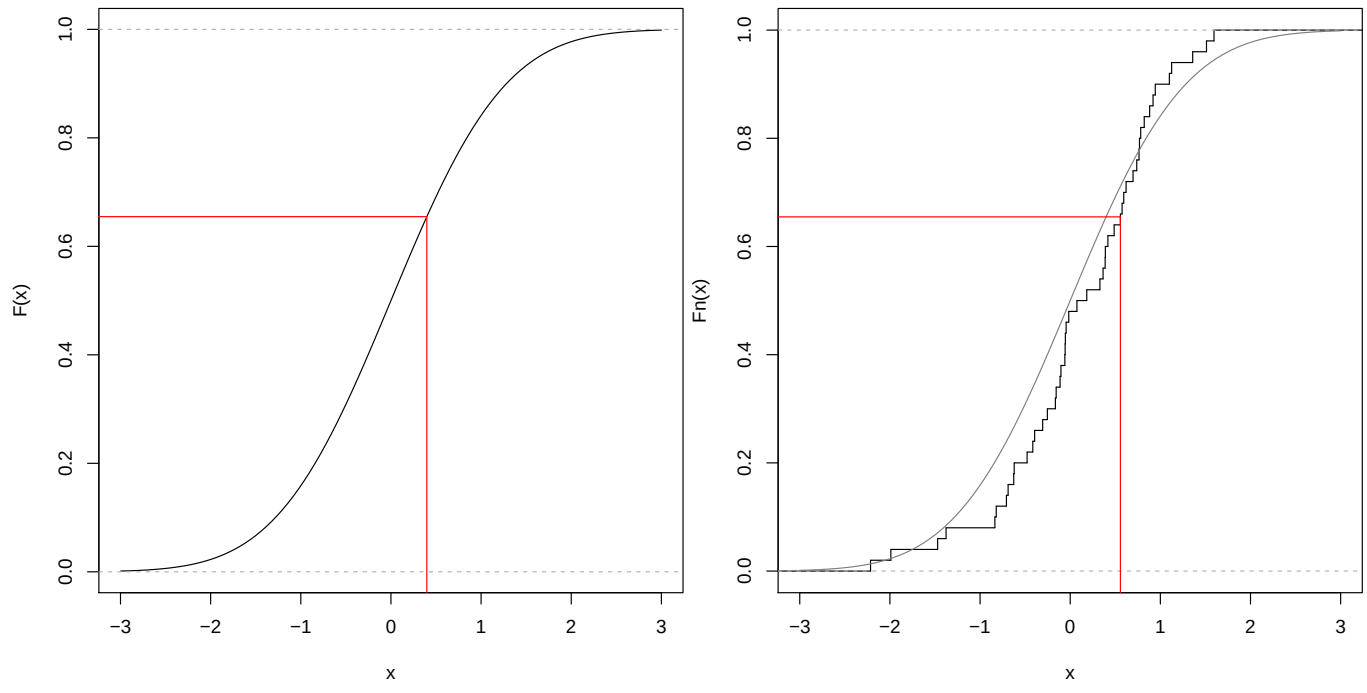
- **When the sample size is insufficient for straightforward statistical inference**.

  If the underlying distribution is known, parametric bootstrapping accounts for sample distortions.

- **When power calculations have to be performed and a small pilot sample is available**.

  To estimate the variation of the statistic using a small pilot sample perform bootstrapping and estimate BS variance.

Connection inverse method and sampling with replacement



Bootstrapping standard error estimation

To estimate the standard error of (a univariate) $\hat{\theta}$ using nonparametric bootstrap the standard deviation of the bootstrap estimates is used:

$$\hat{se}(\hat{\theta}) = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}^i - \bar{\hat{\theta}}\right)^2},$$

where $\bar{\hat{\theta}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}^i$.

In univariate problems is often stated that $m = 50$ to $m = 200$ would be sufficient.

Bootstrapping standard errors example

```
> set.seed(123456)
> n <- 50; m <- 200
> x <- rnorm(n)
> mean(x); sd(x)/sqrt(n)
> B.MEANS <- replicate(m, mean(sample(x, replace=TRUE)))
> mean(B.MEANS); sd(B.MEANS)
```

```
a sample of size 50  with mean 0.117362532834952
and standard error 0.147288076217546 sampled from N(0,1)
which would imply a theoretical mean of 0
and standard error of 0.14142135623731
bootstrap simulation of size 200
with bootstrap mean 0.101763019654491
and standard error 0.14612950105207
```

Bootstrapping confidence intervals

There are several ways to get bootstrap confidence intervals, among others:

1. **Normal bootstrap CI**: Makes many assumptions and basically uses the bootstrapped standard error to construct intervals of the form $\hat{\theta} \pm z_{\alpha/2} se(\hat{\theta})$, where $z_{\alpha/2}$ is the corresponding quantile from the the normal distribution (or some times also the corresponding quantile from the t distribution). This approach is based on the notion of the central limit theorem applying for the estimate $\hat{\theta}$.
2. **Percentile bootstrap CI**: Use the sample percentiles of the bootstrap estimates $\hat{\theta}^i$, $i = 1, \ldots, m$.

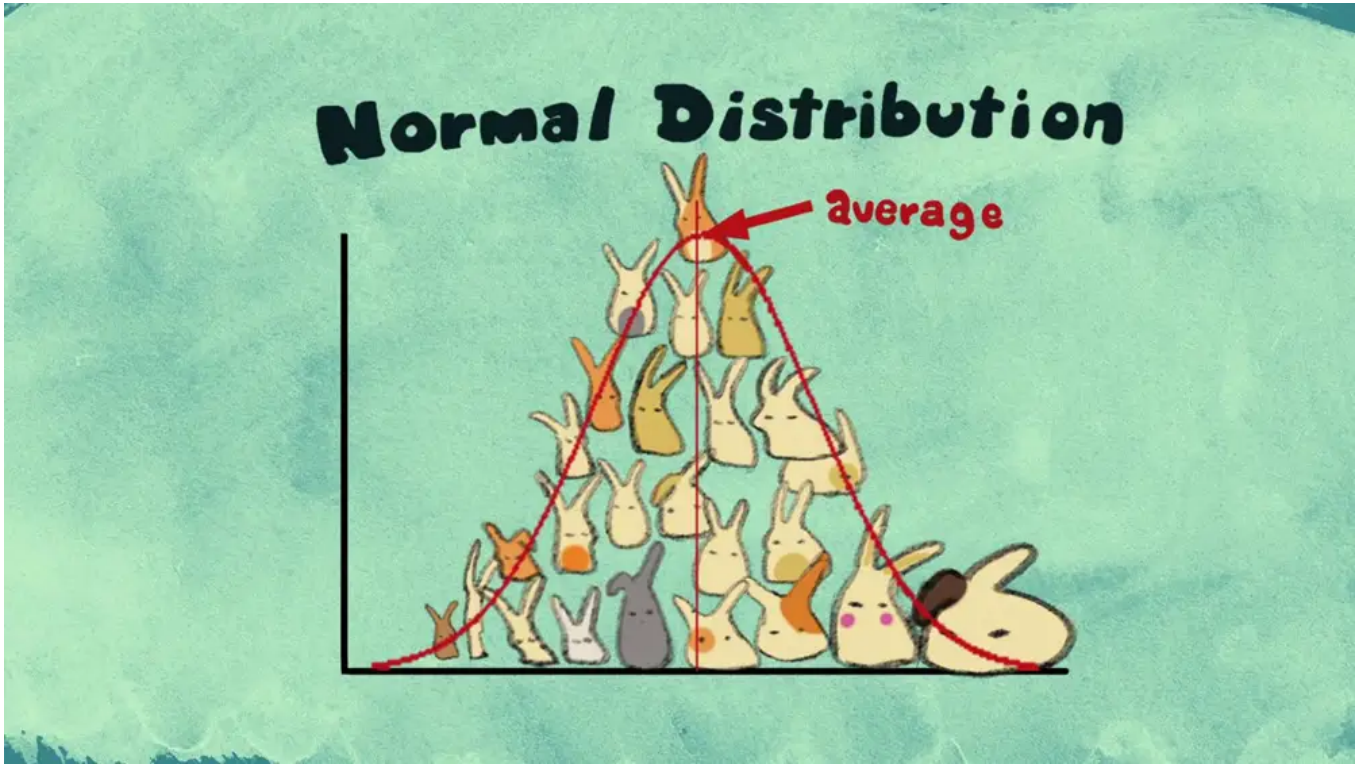
Central Limit Theorem

Central Limit Theorem Let $X1, X2, \ldots$ be a sequence of i. i. d. random variables with expectation $\mathbb{E}[X_i] = \mu$ and $Var[Xi] = \sigma^2 < \infty$. Then as n grows towards infinity, the random variables $\sqrt{n}(\overline{X} - \mu)$ converge in distribution to a normal distribution $N(0, \sigma^2)$.

$$\overline{X} \sim^P N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow$$

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim^P N(0, 1) \tag{1}$$

Central Limit Theorem for Bunnies and Dragons



Percentile bootstrapping CI example

```
> set.seed(123456)
> n <- 30
> m <- 1000
> alpha <- 0.05
> quant <- qchisq(alpha, df=n-1)
> sigma2 <- 1
> x <- rnorm(n)
> B.s2 <- replicate(m, var(sample(x, replace=TRUE)))
> quantile(B.s2, 0.95, type=1)
     95%
1.125701
> (n-1)*var(x)/quant
[1] 1.399839
```

Heart attack risk data

NYT published 27.01.1987 an article on its front page entitled

Heart attack risk found to be cut by taking Aspirin

The double-blinded trial yielded the following data

	Heart attacks	No Heart attacks	Number of subjects
Aspirin	104	10933	11037
Placebo	189	10845	11034

How to decide now if the claim is true?

Odds ratio

The odds ratio (OR) is a popular measure of association between an exposure (here the drug) and an outcome (here heart attack yes/no). The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

Interpretation:

- OR=1: exposure does not affect odds of outcome
- OR>1: exposure associated with higher odds of outcome
- OR<1: exposure associated with lower odds of outcome

Calculating the OR

When comparing 2 categorical variables we construct contigency tables. The visualisation of a contingency table is the mosaic plot.

Consider the 2x2 table

	Outcome: YES	Outcome: NO
Exposure: YES	a	b
Exposure: NO	c	d

$$OR = \frac{\#exposed\ cases / \#non-exposed\ cases}{\#exposed\ non-cases / \#non-exposed\ non-cases}$$

hence

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

CI for OR

A popular formula for the 95% intercal for the OR is

$$[exp(log(OR) - 1.96s), exp(log(OR) + 1.96s)],$$

with $s = \sqrt{1/a + 1/b + 1/c + 1/d}$.

Contingency tables and relative frequencies

To compare the probability of Aspirin and the Placebo to cause heart attacks, we obtain row-wise relative frequencies.

	Heart attacks	No Heart attacks	Number of subjects
Aspirin	104	10933	11037
Placebo	189	10845	11034

$$\mathbb{P}(\text{heart attack}|\text{Aspirin}) = \frac{104}{11037}$$

and

$$\mathbb{P}(\text{heart attack}|\text{Placebo}) = \frac{189}{11034}$$

Odds and Odds ratios

These relative frequencies can be tranformed into odds which are well-known in betting and gambling.

$$\text{Odds}(\text{heart attack}|\text{Aspirin}) = \frac{P(\text{heart attack}|\text{Aspirin})}{1 - P(\text{heart attack}|\text{Aspirin})}$$

To compare two odds, we calculate the *Odds Ratio*, which is

$$OR = \frac{\text{Odds}(\text{heart attack}|\textit{Aspirin})}{\text{Odds}(\text{heart attack}|\textit{Placebo})} = \frac{\frac{P(\text{heart attack}|\text{Aspirin})}{1-P(\text{heart attack}|\text{Aspirin})}}{\frac{P(\text{heart attack}|\text{Placebo})}{1-P(\text{heart attack}|\text{Placebo})}}$$

Interpreting Odds

For Aspirin the odds of causing a heart attack are $0.009512485/0.9904875 = 0.009603842$, approximately 1:100.

For the Placebo the odds of causing a heart attack are $0.01712887/0.9828711 = 0.01742738$, approximately 1:58.

To relate both odds, we caluclate the *Odds Ratio*. The odds ratio $0.009603842/0.01742738 = 0.5458355$ tells us that the odds of Aspirin to cause a heart attack are about half of the odds of a placebo for causing a heart attack. The 95% confidence interval for this OR is [0.4050499 ; 0.6581108]

Analysis in R

```
> DATA <- matrix(c(104,10933,189,10845),
+                2,2, byrow=TRUE)
> DATA
     [,1]  [,2]
[1,]  104 10933
[2,]  189 10845
>
> OR <- DATA[1,1]*DATA[2,2]/(DATA[1,2]*DATA[2,1])
> OR
[1] 0.5458355
>
> s <- sqrt(sum(1/DATA))
> CI.OR <- exp(c(log(OR) - 1.96*s, log(OR) + 1.96*s))
> CI.OR
[1] 0.4290391 0.6944271
```

A resampling idea

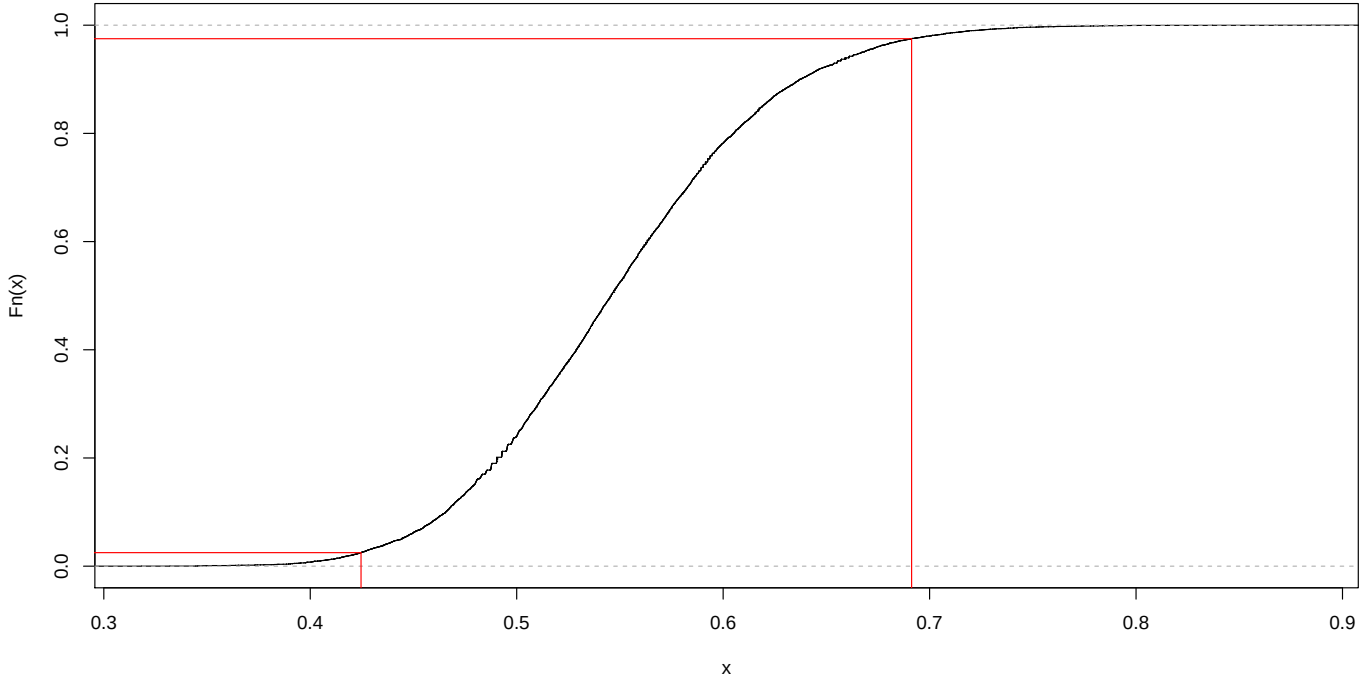
```
> set.seed(1234)
> g1 <- rep(c(TRUE, FALSE),
+     DATA[1, ])
> g2 <- rep(c(TRUE, FALSE),
+     DATA[2, ])
>
> g1new <- sample(g1, replace = TRUE)
> g2new <- sample(g2, replace = TRUE)
>
> ORnew <- mean(g1new)/(1 -
+     mean(g1new))/(mean(g2new)/(1 -
+     mean(g2new)))
> ORnew
[1] 0.5445964
```

Repeating the resampling idea

```
> ORnew <- function(g1, g2) {
+     g1new <- sample(g1, replace = TRUE)
+     g2new <- sample(g2, replace = TRUE)
+
+     ORnew <- mean(g1new)/(1 -
+         mean(g1new))/(mean(g2new)/(1 -
+         mean(g2new)))
+     ORnew
+ }
>
> ORnews <- replicate(10000,
+     ORnew(g1, g2))
> summary(ORnews)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3183  0.5012  0.5458  0.5490  0.5925  0.8850
```

Resampling quantiles

```
> quantile(ORnews, c(0.025,c(0.975)))
     2.5%     97.5%
0.4245977 0.6912860
> # compare to
> CI.OR
[1] 0.4290391 0.6944271
```



Odds ratio and relative risk

Relative risk is frequently used in medical terminology. It described the ratio of probabilities

$$RR = \frac{\pi_{Aspirin}}{\pi_{Placebo}} = \frac{0.009512485}{0.01712887} = 0.5553481$$

Aspirin are therefore 0.56 times more likely to cause a heart attack than the Placebo.

Constructing Confidence intervals for the relative risk

```
> RRnew <- function(g1,g2)
+ {
+ g1new <- sample(g1, replace=TRUE)
+ g2new <- sample(g2, replace=TRUE)
+
+ RRnew <- mean(g1new)/mean(g2new)
+ RRnew
+ }
>
> RRnews <- replicate(10000,RRnew(g1,g2))
> summary(RRnews)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3426  0.5072  0.5499  0.5535  0.5987  0.8755
> quantile(RRnews, c(0.025,c(0.975)))
     2.5%     97.5%
0.4299336 0.6958906
```