Experimental Design and Data Analysis
Lecture 2

Eduard Belitser

VU Amsterdam

Lecture overview

1. bootstrap confidence intervals
2. bootstrap tests
3. one sample tests (normal and not normal sample)
    - *t*-test
    - sign test
    - Wilcoxon signed rank test

bootstrap confidence intervals

Confidence interval for normal data

A point estimate for an unknown parameter μ is some function of the data.

EXAMPLE Suppose we have a sample $X_1, \dots, X_n$ from a normal population with unknown population mean μ. We can estimate μ using the estimating statistic $\bar{X}$. The point estimate for μ is $\hat{\mu} = \bar{X}$.

A confidence interval for an unknown parameter μ is a random interval around the point estimate, containing μ with, e.g., 95% confidence.

EXAMPLE (continued) An (asymptotic) confidence interval for μ with 95% confidence level is the interval $[\bar{X} - m, \bar{X} + m]$, where $m = 1.96 s/\sqrt{n}$.

The margin $m = 1.96 s/\sqrt{n}$ is based on the asymptotic normality of $\bar{X}$ and the fact that s is a good estimator of σ. If in the CI we use the upper t-quantile $t_{0.025,n-1}$ instead of $z_{0.025} \approx 1.96$, the CI will be bigger (i.e., more "conservative") because always $t_{\alpha,n-1} > z_\alpha$, but $t_{\alpha,n-1} \to z_\alpha$ as $n \to \infty$.

Confidence interval for nonnormal data

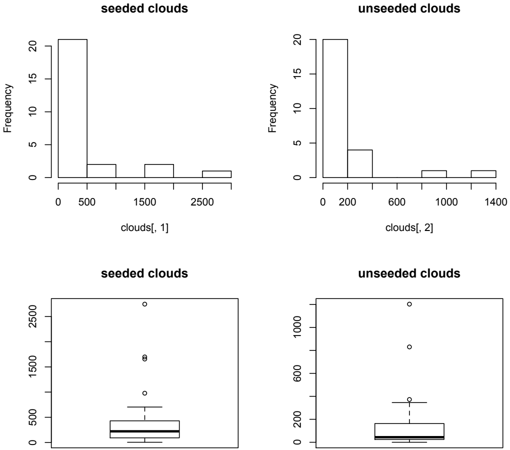
If we have a (small) sample from an unknown distribution and the distribution of $\bar{X}$ is not close to normal, we cannot rely on the above (asympt.) normal CI.

EXAMPLE

Estimate the rainfall means of the two clouds data sets: seeded (with a chemical, silver nitrate, to cause a rainfall) and unseeded

```
> c1=clouds[,1] #   seeded
> c2=clouds[,2] # unseeded
> T1=mean(c1); T2=mean(c2)
> T1
[1] 441.9846
> T2
[1] 164.5619
```

How to determine confidence intervals?



Bootstrap confidence interval

- A bootstrap confidence interval uses simulation to find the distribution of the estimating statistic. The left and right margins for the confidence interval are found from this simulated distribution.
- Denote the data sample as $X_1, \dots, X_N$ and the estimating statistic as $T = T(X_1, \dots, X_N)$. The bootstrap method estimates the distribution of T by using a sample of representative values $T_1^*, \dots, T_B^*$ with B large.
- The formula for the bootstrap confidence interval of level $1 - \alpha$ is

$$[2T - T^*_{(1-\alpha/2)}, 2T - T^*_{(\alpha/2)}],$$

where $T^*_{(\beta)}$ is the T^*-value such that $\beta \times 100\%$ of the T^*-values are lower than $T^*_{(\beta)}$. $T^*_{(\beta)}$ is called the sample β-quantile of the sample $T_1^*, \dots, T_B^*$.

For $T^* = (T_1^*, \dots, T_B^*)$, compute sample β-quantile in R: $T^*_{(\beta)} = \texttt{quantile}(T^*, \beta)$.

T^*-values

The generation of T^* values is as follows.

Repeat B times ($i = 1, \dots, B$):

- generate a surrogate data set $X_1^*, \dots, X_N^*$ by sampling N values from the original data set $X_1, \dots, X_N$ with replacement,
- compute $T_i^* = T(X_1^*, \dots, X_N^*)$ for the surrogate sample.

This procedure yields $T_1^*, \dots, T_B^*$.

Notice that we sample from the data that we have. Some data points X_i may be chosen more than once amongst the X^*-values, whereas other data points X_i may not be chosen at all. We do not introduce any new X-values, we only determine new T^*-values. This bootstrap procedure is called the empirical bootstrap.

Bootstrap CI in R: example with cloud sets

EXAMPLE (continued) Determine this interval for the seeded clouds (c1):

```
> B=1000
> Tstar=numeric(B)
> for(i in 1:B) {
+  Xstar=sample(c1,replace=TRUE)
+  Tstar[i]=mean(Xstar) }
> Tstar25=quantile(Tstar,0.025)
> Tstar975=quantile(Tstar,0.975)
> sum(Tstar<Tstar25)
[1] 25
> c(2*T1-Tstar975,2*T1-Tstar25)
176.8857 668.9462
```

generate $X_1^*, \ldots, X_N^*$
compute T_i^*
determine $T_{(\alpha)}^*$
determine $T_{(1-\alpha)}^*$

The 95% bootstrap confidence interval for the population mean of seeded clouds is [177, 669] around its mean T1=442.

For unseeded clouds the interval is [42, 254] around its mean T2=165.

Example with cloud sets: discussion

- The smaller a confidence interval (with fixed confidence), the more accurate our estimation is. The obtained two intervals are very large, because the estimating statistic $\bar{X}$ is not robust against outliers.
- A robust estimator for location is `median(X)`, the estimating statistic for the population median. For the clouds data, the median is smaller than the mean.
- The 95% bootstrap confidence interval for the population median of seeded clouds is [139, 326] (cf. [177, 669] for population mean). For unseeded clouds, we find the interval [-20, 62] (cf. [42, 254] for population mean).
- For both data sets: the confidence interval for the median is shorter and contains lower values. This confirms that the median is more robust than the mean.

Bootstrap confidence intervals — discussion

- Repeating the computation of a bootstrap confidence interval will always yield a different interval. Enlarging B will reduce the variation.
- Whereas the bootstrap interval is for a population parameter, this interval still depends only on the sample $X_1, \ldots, X_N$.
- In case these values are somewhat extreme, then the bootstrap interval will be off as well. We cannot correct for this, our only information is the sample.

bootstrap tests

Idea

- Suppose we are given
  - a sample $X_1, \ldots, X_N$,
  - a null hypothesis H_0 stating some claim about the population distribution,
  - a (sensible) test statistic $T = T(X_1, \ldots, X_N)$,

  but we lack
  - the distribution of T under H_0.
- Then we cannot perform the test, because we do not have a critical value for T, that acts as border between rejecting and not rejecting H_0.
- But if we somehow can simulate “pseudo-observations” characterizing H_0, we can use a bootstrap test.
- It uses simulations to "mimic" the distribution of T under H_0.

For a bootstrap test, no standard *R*-command — we have to program it ourselves.

Set up of a bootstrap test

Given our sample $X_1, \ldots, X_N$, we can compute the test statistic $T = T(X_1, \ldots, X_N)$ based on our sample.

Simulating the distribution of T under H_0 in the bootstrap fashion means generate a bunch of surrogate T-values $(T_1^*, \ldots, T_B^*)$ that are representative values for T under H_0.

The simulation set up is

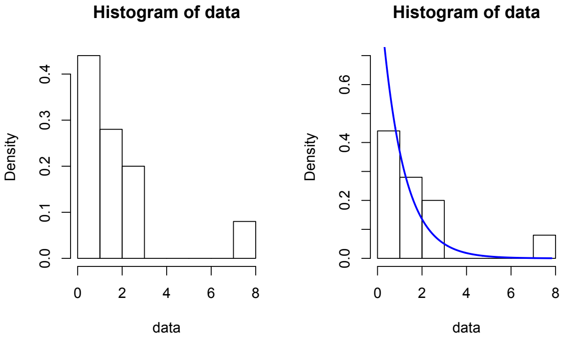
- repeat B times ($i = 1, \ldots, B$):
  1. generate a surrogate data sample $X_1^*, \ldots, X_N^*$ (same sample size as original data set) according to H_0,
  2. Compute the test statistic $T_i^* = T(X_1^*, \ldots, X_N^*)$ for the surrogate sample.
- compare the T-value of the original data to the surrogate T^*-values and determine a p-value.

(By simulating the unknown distribution we make an estimation error. This error can be made arbitrarily small by choosing B large enough.)

Bootstrap test — implementation in R (1)

We wish to test $H_0 : X_i \sim \exp(1)$, i.i.d. $i = 1 \ldots, N$, i.e. the data are a random sample from the standard exponential distribution.

```
> hist(data,prob=T)
> hist(data,prob=T,ylim=c(0,0.7))
> x=seq(0,max(data),length=1000)
> lines(x,dexp(x),type="l",col="blue",lwd=2)
```

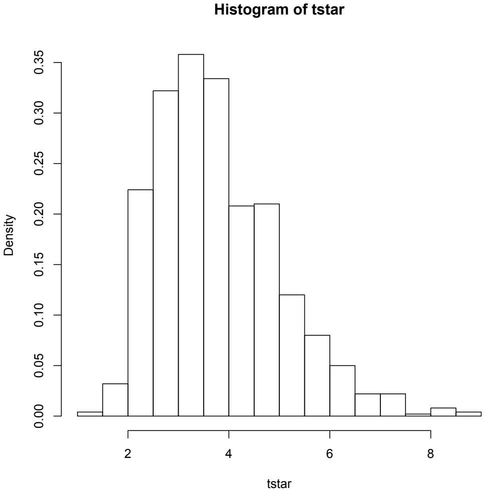


Bootstrap test — implementation in R (2)

We use as test statistic the maximum of the sample: $T(X_1, \dots, X_N) = max(X_1, \dots, X_N)$.

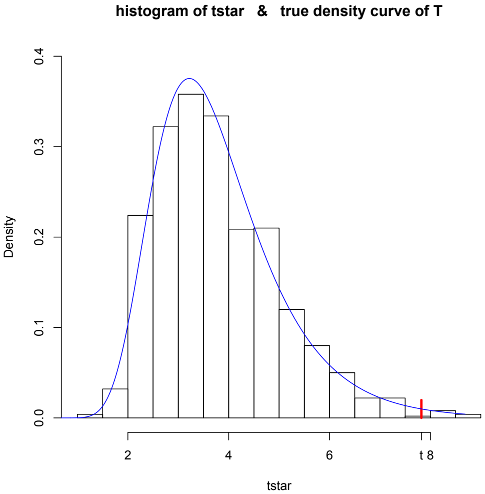
```
> t=max(data)
> t
[1] 7.821847

> B=1000
> tstar=numeric(B)
> n=length(data)
> for (i in 1:B){
+  xstar=rexp(n,1)
+  tstar[i]=max(xstar)}
> hist(tstar,prob=T)
```



Bootstrap test — p-value in R (1)

The *p*-value is found by considering the proportion of T^*-values exceeding the T-value of the data.



Bootstrap test — p-value in R (2)

The R-code for the *p*-value:

```
> pl=sum(tstar<t)/B; pr=sum(tstar>t)/B; p=2*min(pl,pr)
> pl;pr;p
[1] 0.994
[1] 0.006
[1] 0.012
```

The *p*-value is 0.012 and H_0 is rejected.
The R-code for the histogram in the previous slide:

```
> hist(tstar,prob=T,ylim=c(0,0.4),
+ main="histogram of tstar & true density curve of T")
> densmaxexp=function(x,n) n*exp(-x)*(1-exp(-x))^(n-1)
> lines(rep(t,2),seq(0,2*densmaxexp(t,n),length=2),
+ type="l", col="red", lwd=3)
> axis(1,t,expression(paste("t") ) )
> u=seq(0,max(tstar),length=1000)
> lines(u,densmaxexp(u,n),type="l",col="blue")
```

Bootstrap test — discussion

- The resulting p-value depends on the realised T^*-values. It is recommended to repeat a bootstrap test a few times to see whether the p-value is stable.
- When B is too small, there is a lot of variation in the p-value. In most cases $B = 1000$ is adequate.
- A bootstrap test can be performed with any test statistic. E.g., in the example taking `min` as a test statistic yields a bootstrap p-value of about 0.19 (check this yourselves!) and does not lead to rejecting H_0.
- The difference between the simulation of T^*-values for bootstrap confidence intervals and bootstrap tests is in the way the $X_1^*, \ldots, X_N^*$ are generated. For confidence intervals you choose X_i^* from your sample, whereas for tests you generate X_i^* according to H_0.

one sample from a normal distribution

Setting and design

Setting:

an experiment with one numerical outcome per experimental unit. Interest is in the location of the population distribution.

Design:

- Take a random sample of experimental units from the relevant population
- Measure the outcome on each unit

EXAMPLE Measurement of the height of 4 years old children.

EXAMPLE Measurement of the time it takes to find a certain document in a web design for different users.

EXAMPLE Measurement of the yearly amount of sun hours in different countries.

Analysis

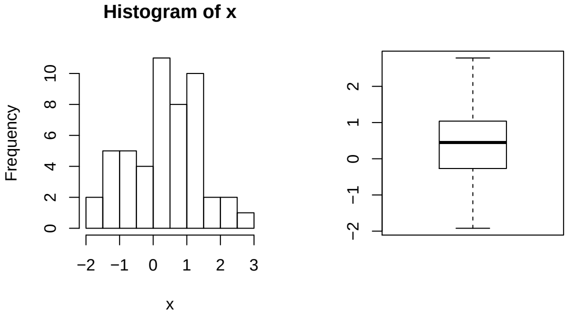
- Data: $(X_1, \ldots, X_N)$.
- The *t*-test assumes that the data $X_1, \ldots, X_N$ are a random sample from a normal population.
- We test the null hypothesis $H_0 : \mu = \mu_0$ that the mean of this population is μ_0, e.g. $\mu_0 = 0$.
- The test statistic is

$$T = \frac{\bar{X}_N - \mu_0}{S_N},$$

which has the t_{N-1}-distribution under H_0.

Analysis in R — data input and graphics

```
> mu=0.2
> x=rnorm(50,mu,1);   # creating artificial data
> par(mfrow=c(1,2))   # two plots next to each other
> hist(x)
> boxplot(x)
```



Analysis in R — estimation and testing

```
> t.test(x)

        One Sample t-test

data:  x
t = 2.2701, df = 49, p-value = 0.02764
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.03804252 0.62504011
sample estimates:
mean of x
0.3315413
```

Conclusion?

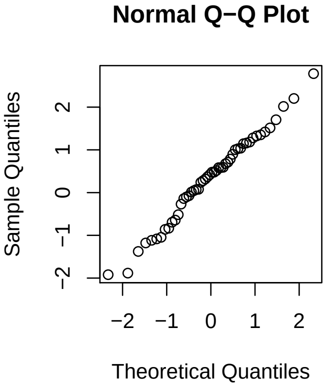
(By default `t.test` tests $H_0 : \mu = 0$.)

Analysis in R — diagnostics

- The t-test is based on the normality assumption, we need to check this.
- The assumption of normality is crucial. If the data do not follow a normal distribution, the *p*-value from the *t*-test cannot be trusted.

```
> qqnorm(x)
```

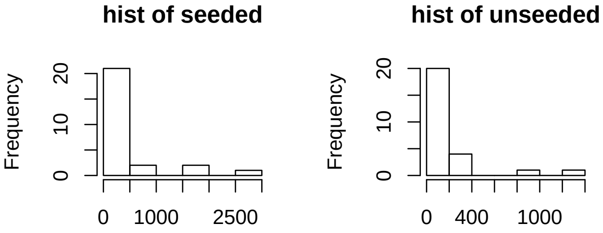
One can also look at the boxplot, histogram, and the Shapiro-Wilk normality test (`shapiro.test(x)`).



Discussion (1)

Not all data can be assumed to come from a normal distribution. Histograms and QQ-plots can be used to check the normality assumption.

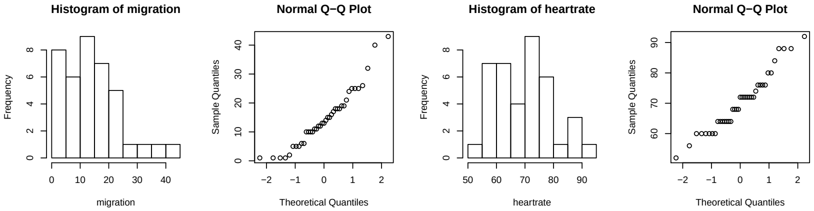
EXAMPLE Cloud seeding is a technique used to change the amount and type of precipitation, by dispersing substances into clouds. Precipitation values of seeded and unseeded clouds were measured.



Assuming normality here is clearly wrong.

Discussion (2)

EXAMPLE From a sample of 39 Peruvian men that had moved from a native culture to a modern society, the following variables were measured (amongst others): years since migration, systolic and diastolic blood pressure, heart rate, weight, length.



Normality is doubtful for both `migration` (seems not symmetric) and `heartrate`.

one sample from a nonnormal distribution

Setting and design

Setting:

an experiment with one numerical outcome per experimental unit. Interest is in the location of the population distribution.

Design:

- Take a random sample of experimental units from the relevant population
- Measure the outcome on each unit

EXAMPLE The number of infected people by a certain disease in different countries.

EXAMPLE The times between eruptions of a volcano.

EXAMPLE The exam grades for a certain course.

Analysis A — sign test

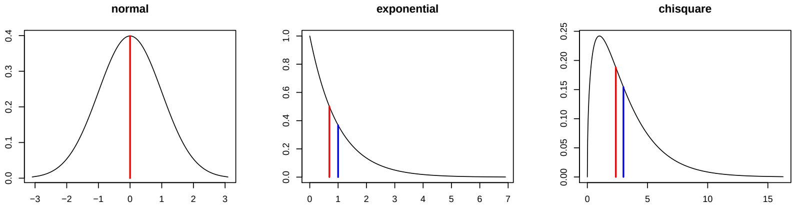
- Data $(X_1, \dots, X_N)$.
- The sign test assumes that the data $X_1, \dots, X_N$ are a random sample from a population with a certain median m.
- We test the null hypothesis $H_0 : m = m_0$ that the median of this population is m_0, e.g. $m_0 = 0$.
- The test statistic is $T = \#(X_i > m_0)$, which has the $\mathrm{Bin}(N, 0.5)$-distribution under H_0.

The median — recap

The median of a population is the middle value in the sorted population values.

For a given population median m, we have that $P(X < m) = P(X > m) = \frac{1}{2}$ for a random value X from the population. Being bigger or smaller than the median is like tossing a fair coin.

For skewed distributions (e.g., clouds) the mean is highly influenced by the high/low values. In such cases it is better to test location in terms of the median, instead of the mean.



The more skewed, the bigger the distance between median and mean.

Analysis A in R — data input and sign test

We want to test whether an exam is of adequate level, that is whether the median is equal to 6. Because of the small sample size, we are not sure about normality. (Grades are never really normally distributed!) Data are the exam grades of 13 randomly selected students.

```
> examresults=c(3.7,5.2,6.9,7.2,6.4,9.3,4.3,8.4,6.5,8.1,7.3,6.1,5.8)
> sum(examresults>6)
[1] 9
> binom.test(9,13,p=0.5)
        Exact binomial test
data:  9 and 13
number of successes = 9, number of trials = 13, p-value = 0.2668
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3857383 0.9090796
sample estimates:
probability of success
             0.6923077
```

The sign test computes the number of values bigger than m_0. If $m = m_0$ then we expect about $N/2$ values bigger/smaller than m_0. Conclusion from the above output of `binom.test`: H_0 is not rejected.

Analysis B — Wilcoxon test

- Data $(X_1, \ldots, X_N)$.
- The Wilcoxon signed rank test assumes that the data $X_1, \ldots, X_N$ are a random sample from a symmetric population with a certain median m. This is a stronger assumption than the one for the sign test!
- We test the null hypothesis $H_0 : m = m_0$ that the median of this population is m_0, e.g., $m_0 = 0$.
- The test statistic T is based on the ranks R_i of the absolute differences $|X_i - m_0|$.

  $$T = \sum_{i : X_i > m_0} R_i.$$

  The distribution of T under H_0 is known, and can be approximated by a normal distribution if N is large.
- Large values of T indicate that $m > m_0$, whereas small values of T indicate that $m < m_0$.

Analysis B in R — Wilcoxon test

The signed rank test takes into account the ranks of the deviations from the proposed median m_0. If the data are symmetric around m_0, the ranks at both sides should be approximately equal.

```
> examresults-6
 [1] -2.3 -0.8 0.9 1.2 0.4 3.3 -1.7 2.4 0.5 2.1 1.3 0.1 -0.2
> rank(abs(examresults-6))
 [1] 11 5 6 7 3 13 9 12 4 10 8 1 2
> rank(abs(examresults-6))[examresults-6>0]
[1] 6 7 3 13 12 4 10 8 1
> sum(rank(abs(examresults-6))[examresults-6>0])
[1] 64
> wilcox.test(examresults,mu=6)

        Wilcoxon signed rank test

data:  examresults
V = 64, p-value = 0.2163
alternative hypothesis: true location is not equal to 6
```

Conclusion: H_0 is not rejected.

To finish

Today we discussed:

1. bootstrap confidence intervals
2. bootstrap tests
3. one sample tests (normal and not normal samples)
    - *t*-test
    - sign test
    - Wilcoxon signed rank test

Next time: two sample tests.