Experimental Design and Data Analysis
Lecture 2

Eduard Belitser

VU Amsterdam

Lecture overview

1. bootstrap confidence intervals
2. bootstrap tests
3. one sample (two paired samples) tests for normal and not normal samples
   - 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 For a sample $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, we can estimate μ using the estimating statistic $\bar{X}$. The point estimate for μ is thus $\hat{\mu} = \bar{X}$.

Recall that 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) 95%-confidence interval for μ is the interval $[\bar{X} - m, \bar{X} + m]$, where $m = 1.96s/\sqrt{n}$.

The margin $m = 1.96s/\sqrt{n}$ 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 $n \to \infty$.

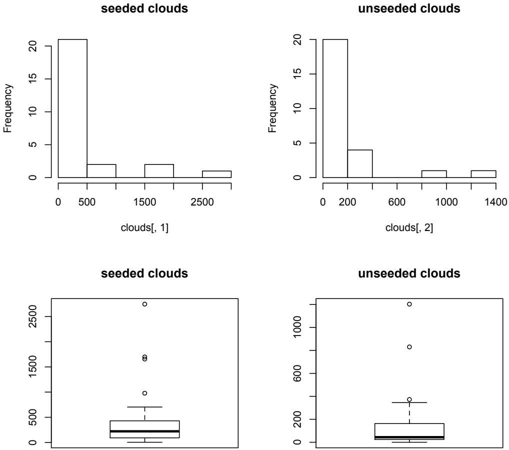
Confidence interval for nonnormal data

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

- Suppose we have a data sample $X = (X_1, \dots, X_n)$ and an estimating statistic $T = T(X_1, \dots, X_n)$ for a parameter, say, θ.
- We use simulation to find the distribution of the estimating statistic $T(X)$. The bootstrap CI is then found from this simulated distribution.
- The bootstrap method estimates the distribution of T by creating a sample of representative values $T_1^*, \dots, T_B^*$ with B large.
- The basic 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 $T_1^*, \dots, T_B^*$.
  In R: the sample β-quantile of $T^* = (T_1^*, \dots, T_B^*)$ is $T^*_{(\beta)} = \texttt{quantile}(T^*, \beta)$.
- The bootstrap estimate for the variance of statistics $T(X)$ is given by

  $$\widehat{\mathrm{Var}}(T) = S^2_{T^*} = \tfrac{1}{B-1} \sum_{b=1}^{B} \left(T_b^* - \overline{T^*}\right)^2. \quad \text{In R: } S^2_{T^*} = \texttt{var}(T^*).$$

This bootstrap CI is constructed in such a way that it uses T. A simplier version of bootstrap CI (called percentile bootstrap CI): $\left[T^*_{(\alpha/2)}, T^*_{(1-\alpha/2)}\right]$.

Heuristics for basic bootstrap CI

We interpret $T_1^*, \dots, T_B^*$ as realizations of some random variable T^*. Then

$$\begin{aligned} 1-\alpha &\approx \mathrm{P}\Big(T^*_{(\alpha/2)} \le T^* \le T^*_{(1-\alpha/2)}\Big) \ \text{(percentile bootstrap CI } [T^*_{(\alpha/2)}, T^*_{(1-\alpha/2)}]) \\ &= \mathrm{P}\Big(T^*_{(\alpha/2)} - T \le T^* - T \le T^*_{(1-\alpha/2)} - T\Big) \\ &\approx \mathrm{P}\Big(T^*_{(\alpha/2)} - T \le T - \theta \le T^*_{(1-\alpha/2)} - T\Big) \\ &= \mathrm{P}\Big(2T - T^*_{(1-\alpha/2)} \le \theta \le 2T - T^*_{(\alpha/2)}\Big), \end{aligned}$$

which gives us the basic bootstrap confidence interval for θ:

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

How to generate 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 empirical bootstrap.
How many different resamples are possible from a sample of size n? The number of ways to place n objects into n bins (some bins may be empty, i-th bin contains the copies of X_i). The method of stars and bars yields $\binom{2n-1}{n-1} = \binom{2n-1}{n}$.
If you want a reference and a rule of thumb for B, Wilcox (2010) writes "599 is recommended for general use."

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 $T_b^*, b = 1, \ldots B$
determine $T^*_{(\alpha/2)}$
determine $T^*_{(1-\alpha/2)}$

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 the CI, the better. The obtained two CI's are very large, because the estimating statistic $\bar{X}$ is not robust against outliers.
- A robust estimator for location is `median(X)`, estimating the population median. For the clouds data, the median is smaller than the mean.
- The 95% bootstrap CI for the median of seeded clouds is [139, 326] ([177, 669] for the mean); unseeded clouds: [-20, 62] ([42, 254] for the mean).
- For both data sets: the CI for the median is shorter and contains lower values. This confirms that the median is more robust than the mean.

General discussion on bootstrap confidence intervals

- Repeating the computation of a bootstrap confidence interval will always yield a different interval. Enlarging B will reduce the variation.
- The bootstrap interval still depends only on the sample $X_1, \ldots, X_n$.
- If the original data $X_1, \ldots, X_n$ caries little information about the parameter θ, the bootstrap interval will be off as well.

bootstrap tests

Idea

- Suppose we are given
  - a sample $X_1, \ldots, X_n$,
  - a null hypothesis H_0: 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 $X_1^*, \ldots, X_n^*$ (of the 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.

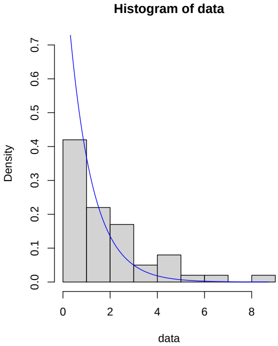
Bootstrap test: implementation in R (1)

We wish to test $H_0 : X_i \overset{iid}{\sim} \mathrm{Exp}(1)$, $i = 1 \ldots, n$, i.e. the data are a random sample from the standard exponential distribution.

First plot a histogram of the data and the density Exp(1) corresponding to H_0.

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

H_0 seems plausible.



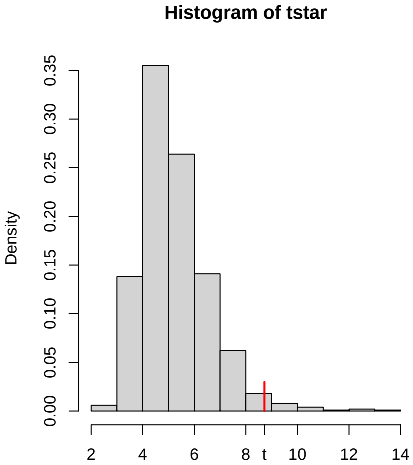
Bootstrap test: implementation in R (2)

As test statistic we use $T(X_1, \ldots, X_n) = \max(X_1, \ldots, X_n)$. The p-value is computed as the proportion of T^*-values exceeding the T-value.

```
> n=length(data); t=max(data); t
[1] 8.72055
> B=1000; tstar=numeric(B)
> for (i in 1:B){xstar=rexp(n,1)
  + tstar[i]=max(xstar)}
> pl=sum(tstar<t)/B;pr=sum(tstar>t)/B
> p=2*min(pl,pr); p
[1] 0.038
```

Since p-value$=0.038$, H_0 is rejected.
The R-code for the histogram

```
> hist(tstar,prob=T,
 + main="Histogram of tstar")
> lines(rep(t,2),c(0,0.03),
 + col="red",lwd=2)
> axis(1,t,expression(paste("t")))
```



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 that case B should be increased. 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 draw X_i^*'s from the original sample, whereas for tests you generate X_i^*'s according to H_0.

one sample (or two paired samples) from a normal distribution

t-test for one sample

Setting:
the data $(X_1, \ldots, X_n)$ is a result of 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 yearly amount of sun hours in diff. countries.

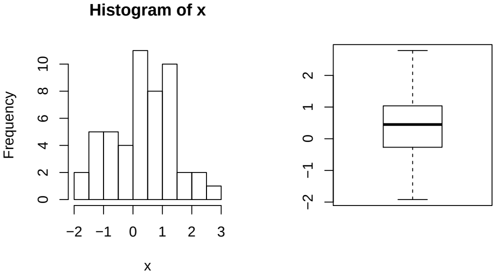
Analysis:

- t-test assumes that the data $(X_1, \ldots, X_n)$ stems from a normal distribution (or, at least, approximately normal).
- Test about the population mean μ: $H_0 : \mu \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} \mu_0$ vs. $H_1 : \mu \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} \mu_0$.
- The test statistic $T = \sqrt{n}(\bar{X} - \mu_0)/s$ has the t_{n-1}-distribution under H_0.

One sample t-test in R

Generate data:

```
> mu=0.2
> x=rnorm(50,mu,1)
> par(mfrow=c(1,2))
> hist(x)
> boxplot(x)
```



```
> t.test(x)  # by default H0: mu=0
        One Sample t-test
data:  x
t = 2.2701, df = 49, p-value = 0.02764.
   [ some output deleted ]
```

Conclusion: reject $H_0 : \mu = 0$.

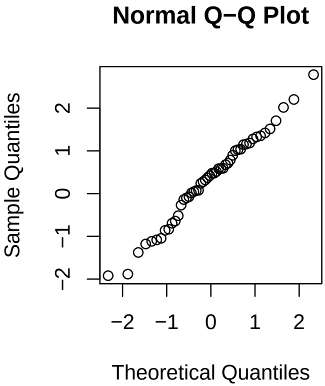
One sample t-test in R: diagnostics

- t-test is based on the (appr.) normality assumption, 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)
```

Besides `qqnorm`, one can also look at `hist`, `shapiro.test` and `boxplot`.

The main normality checks in this course are `histogram` and `qqnorm`. Sometimes, the Shapiro-Wilk normality test `shapiro.test` is also to be be reported (especially when it rejects normality).



Setting and design for two paired samples

Setting:

An experiment with two numerical outcomes per experimental unit. Interest is in a possible difference between the two outcomes.

EXAMPLE Comparing pain relief by a dedicated drug or by a placebo. Both treatments are applied to every individual (with recovery time in between).

EXAMPLE Comparing two car tire brands by putting both brands of tire on the same car and measuring the tires' wear.

Design:

- Take a random sample of experimental units from the relevant population.
- Measure the two outcomes on each unit (which are clearly related).
- The experiment should be set up so that any other type of "dependence" is eliminated and a difference in outcomes is due to the "treatment" only.

Remark. If subjects must perform two tasks, then they should be allowed sufficient time between the tasks to recover and forget. If a learning effect (the first measurement influences the second) is suspected, then, if possible, randomize the order of the two treatments within the units. The analysis must then follow the cross over design (studied later), not the paired samples design as discussed here.

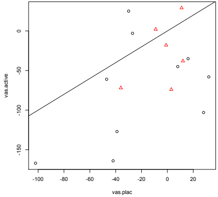
Paired t-test: analysis

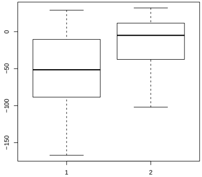
- Data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.
- In the paired t-test the differences $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$ are assumed to be (approx.) from a normal distribution $N(\mu, \sigma^2)$.
- Test about the mean difference $H_0 : \mu \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} 0$ versus $H_1 : \mu \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} 0$.
- Test statistic $T = \frac{\bar{Z}}{s_Z/\sqrt{n}}$, with $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$, $s_Z^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2$. Under H_0, T has the t_{n-1}-distribution.
- The analysis is simply a one sample analysis on the differences, and μ is the difference of the means of the X-population and the Y-population.

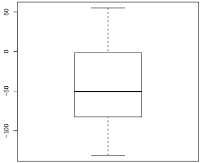
Paired t-test in R: graphics

The rows of the data set ashina.txt correspond to 16 subjects and give measures of pain (for chronic headache) when treated with an active drug or a placebo.

```
> ashina=read.table("ashina.txt",header=TRUE); ashina
   vas.active vas.plac grp
1        -167     -102   1
2        -127      -39   1
[ some output deleted ]
16        -72      -36   2
> plot(vas.active~vas.plac,pch=grp,col=grp,data=ashina); abline(0,1)
> boxplot(ashina[,1],ashina[,2]); boxplot(ashina[,1]-ashina[,2])
```







The third column of the data.frame ashina indicates the order of measurement (1=placebo first, 2=active first). This is used in the first plot (only) to determine the plotting character. A possible effect of the ordering of the measurements is ignored.

Paired t-test in R: estimation and testing

```
> t.test(ashina[,1],ashina[,2],paired=TRUE) # two sample paired t-test
        Paired t-test
data:  ashina[, 1] and ashina[, 2]
t = -3.2269, df = 15, p-value = 0.005644  # conclusion: H0 is rejected
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -71.1946 -14.5554
sample estimates:
mean of the differences
                -42.875
```

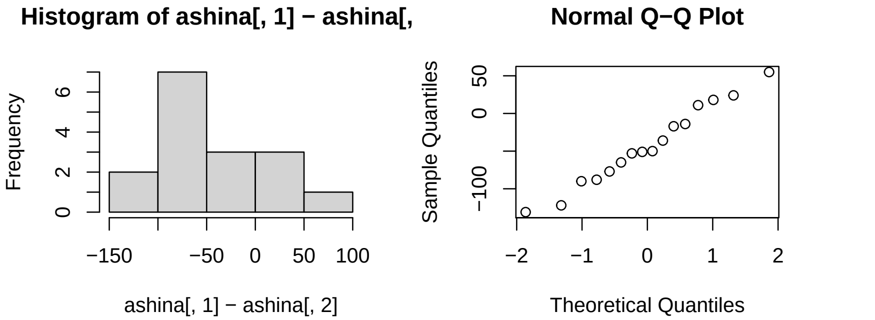
Without `paired=TRUE`, `t.test` with 2 arguments treats 2 samples as independent. With 1 argument `t.test` performs a one sample t-test. Applied to the differences this is equivalent to a paired two sample *t*-test.

```
> t.test(ashina[,1]-ashina[,2])  # one sample t-test for differences
        One Sample t-test
data:  ashina[, 1] - ashina[, 2]
t = -3.2269, df = 15, p-value = 0.005644  # conclusion: H0 is rejected
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -71.1946 -14.5554
 [ some output deleted ]
```

Paired t-test in R: diagnostics

Conclusion from the above analysis: H_0 is rejected, i.e., the mean of the differences is different from 0. Recall that we assumed (appr.) normality of the data. Check the normality assumption on the differences (not original samples):

```
> par(mfrow=c(1,2));hist(ashina[,1]-ashina[,2]);qqnorm(ashina[,1]-ashina[,2])
> shapiro.test(ashina[,1]-ashina[,2])  ## gives $p$-value 0.9377
```

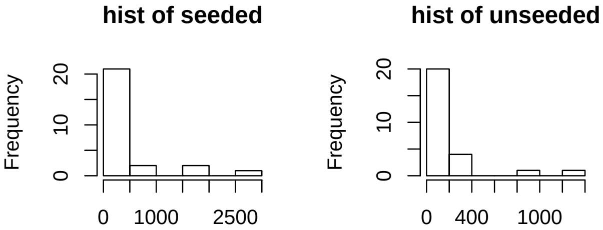


Here no reason to suspect that the differences are not taken from a normal population.

Example of non-normal sample

Not all data can be assumed to come from a (appr.) 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.

one sample (or two paired samples) from a nonnormal distribution

One sample (or two paired samples): setting and design

Setting:

- An experiment with one numerical outcome per experimental unit. Interest is in the location (e.g., median) of the population distribution.
- An experiment with two numerical outcomes per experimental unit. Interest is in a possible difference between the locations of the two outcomes. This setting is called two paired samples (or, matched pairs).

Design:

- Take a random sample of experimental units from the relevant population.
- Measure the outcome on each unit, or measure the two outcomes on each unit (will be clearly related as the they are measured on the same unit).

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

EXAMPLE The exam grades for a certain course.

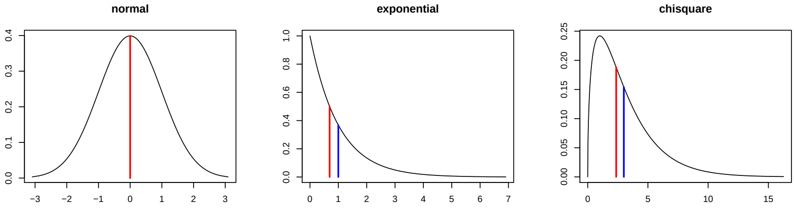
EXAMPLE The blood pressure of a person before and after a drug treatment.

The median: recap

The median of a population is the middle value in the sorted populat. values. Formally: m is the median of a (contin.) random variable X if $P(X \leq m) = \frac{1}{2}$.

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

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



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

Sign test for one sample or matched pairs

Setting:

- A sample $X_1, \dots, X_n$ from some population. We want to test about the population median m.
- A sample $(Z_1, Y_1) \dots, (Z_n, Y_n)$ of matched pairs from some population. We want to test about the median m of the differences $X_i = Z_i - Y_i$.

Hypotheses: we test $H_0 : m \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} m_0$ versus $H_1 : m \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} m_0$.

Test statistic: $T = \#(i : X_i < m_0)$, where "#" means "the number of".

Distribution of T under H_0: exactly $\text{Bin}(n, \frac{1}{2})$ (a norm. approx. is possible). Depending on H_1 the test is one-sided or two-sided.

In R: `binom.test(t,n,p=0.5,alt=...)` (for example, `alt="g"` if $H_1 : m > m_0$)

If $m = m_0$, about $\frac{n}{2}$ values are expected to be bigger/smaller than m_0. Large deviations from this indicate that H_0 may not be true. In case of matched pairs $\#(i : X_i < m_0) = \#(i : Z_i < Y_i)$.

Sign test in R: example

We want to test whether the median exam grade is 6. Because of the small sample size, we are not sure about normality. (Grades are not always 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
 [ some output is deleted ]
   p-value = 0.2668
```

Conclusion from the above output of `binom.test`: H_0 is not rejected.

To test the claim of interest correctly, one should reduce to the right version of the binomial test: the relevant one-sided or two sided version. For example, to test whether the exam is not too difficult, we can set $H_1 : m > 6$ leading to test `binom.test(9,13,p=0.5,alt="g")`. One can also work with other choices of statistics T, e.g., $T = \#(i : X_i > m_0)$.

Wilcoxon signed rank test for one sample or matched pairs

Setting:

- A sample $X_1, \dots, X_n$ from a symmetric population (a stronger assumption than for the sign test!). Want to test about the population median m.
- A sample $(Z_1, Y_1) \dots, (Z_n, Y_n)$ of matched pairs from some population. Test about the median m of the (symm.) differences $X_i = Z_i - Y_i$.

Hypotheses: $H_0 : m \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} m_0$ vs. $H_1 : m \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} m_0$.

Test statistic: $T = \sum_{i: X_i > m_0} R_i$ of the ranks of $|X_i - m_0|$ over such i for which $X_i > m_0$. Large values of T indicate that $m > m_0$, small T that $m < m_0$.

Distribution of T under H_0: known in R (normal approximation for large n).

In R: `wilcox.test(data,mu=m0,alt=...)` Dep. on H_1, one- or two-sided test.

Rank of an observation is the order number assigned to it if the observations are ordered from smallest to largest. For example, the ranks of observations $X_1 = 3$, $X_2 = 5$, $X_3 = 2$, $X_4 = 7$ are $R_1 = 2, R_2 = 3, R_3 = 1, R_4 = 4$ resp. In R, the ranks of the sample `x` are computed by `rank(x)`. Norm. approx.: $\frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$.

Wilcoxon signed rank test in R: example

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

```
> sum(rank(abs(examresults-6))[examresults-6>0]) # value test statistics
[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 (two paired samples) tests for normal and not normal samples
    - t-test
    - sign test
    - Wilcoxon signed rank test

Next time: two sample tests.