Introduction to Geostatistics, practical exercises

(c) Edzer J. Pebesma, Kristina Helle, Lydia Gerharz
Institute for Geoinformatics
University of Münster
edzer.pebesma@uni-muenster.de

July 1, 2009

Contents

1 Introduction to R

1.1 S, S-Plus and R

S is a language for data analysis and statistical computing. It has currently two implementations: one commercial, called S-Plus, and one free, open source implementation called R. Both have advantages and disadvantages.

S-Plus has the (initial) advantage that you can work with much of the functionality without having to know the S language: it has a large, extensively developed graphical user interface. R works by typing commands (although the R packages `Rcmdr` and `JGR` do provide graphical user interfaces). Compare typing commands with sending SMS messages: at first it seems clumsy, but when you get to speed it is fun.

Compared to working with a graphical program that requires large amounts of mouse clicks to get something done, working directly with a language has a couple of notable advantages:

- after a few commands, you can exactly (i) see and (ii) replicate what you did, and send the list of commands to someone else to replicate your analysis;
- when you have a large number of analyses that take much time, you can easily perform them as batches, possibly on a remote computer or on a cluster of computers;
- when you tend to repeat a set of commands often, you can very easily turn them into an S function, which then extends the language, is available forever, and can be redistributed. In the world of R, users soon become developers–vice versa, most developers are users in the first place.

Of course the disadvantage (as you may consider it right now) is that you need to learn a few new ideas.

Although S-Plus and R are both full-grown statistical computing engines, throughout this course we use R for the following reasons: (i) it is free (S-Plus may cost in the order of 10,000 USD), (ii) development and availability of novel statistical computation methods is much faster in R.

Recommended text books that deal with both R and S-Plus are e.g.:

1. W. Venables, B. Ripley: Modern applied statistics with S; Springer (emphasizes on statistical analysis)
2. W. Venables, B. Ripley: S Programming; Springer (emphasizes on programming)
3. J. Chambers: Programming with Data; Springer (fundamental; written by one of the designers of S)

Further useful books on R are:

1. P. Dalgaard, Introductory Statistics with R; Springer
2. R.S. Bivand, E.J. Pebesma, V. Gómez-Rubio: Applied Spatial Data Analysis in R. Springer (to appear in Jul 2008).

1.2 Downloading and installing R

To download R on a MS-Windows computer, do the following:

1. point your web browser to http://www.r-project.org
2. click on the link to CRAN (abbreviation of the Comprehensive R Archive Network)
3. choose a nearby mirror
4. click on the MS-Windows version of R
5. open the self-extracting executable, and follow the instructions

Instructions for installing R on a Macintosh computer are also found on CRAN. For installing R on a linux machine, you can download the R source code from CRAN, and follow the compilation and installation instructions; usually this involves the commands `./configure; make; make install`. There may be binary distributions for linux distributions such as RedHat FC, or Debian unstable.

1.3 Starting R; saving or resoring an R session

You can start an R session on a MS-Windows machine by double-clicking the R icon, or through the Start menu (Note that this is not true in the ifgi CIP pools: here you have to find the executable in the right direction on the C: drive). When you start R, you'll see the R console with the `>` prompt, on which you can give commands.

Another way of starting R is when you have a data file associated with R (meaning it ends on .RData). If you click on this file, R is started and reads at start-up the data in this file, along with the command history in the `.Rhistory` file in the same directory.

Try this using the `students.RData` file, that you will find at http://ifgi.uni-muenster.de/~epebe_01/Geostatistics/students.RData. Next, try the following commands:

```
> objects()
> summary(students)
> a = 1
> objects()
> print(a)
> a
```

(Note that you can copy and paste the commands *including the* `>` to the R prompt, only if you use "paste as commands")

You can see that your R workspace now has two objects in it. You can find out which class these objects belong by

```
> class(a)
> class(students)
> lapply(students, class)
```

If you want to remove an object, use

```
> rm(a)
```

If you want so safe a (modified) work space to a file, you can use the menu (windows) or:

```
> save.image(file = "students.RData")
```

2 Descriptive statistics

2.1 Frequencies and modes

First make sure that you finish the introductory material up to and including the chapter on *Reading data from files*

Import the students sample data, collected at the first lecture. Either do this by loading the R data file mentioned above, or importing the raw data in csv as explained on the course web site.

Exercise 2.1 Look at a summary of the data. Which of the variables are nominal?

Exercise 2.2 For each of the variables, find out what the modus is, by using function `table` on them.

One way to do this is to look up the maximum frequency manually. Automatic lookup is obtained as follows:

```
> x = table(students$Weight)
> x
> x[which(x == max(x))]
```

Distributions can be plotted by function `hist`. Execute the following two commands:

```
> hist(students$Weight)
> hist(students$Weight, plot = FALSE)
```

Exercise 2.3 [HAND IN] Try to explain how the results returned from the second command relate to the plot obtained by the first and how they relate to the data. Which of the following values can be read from the plot; give values if possible. How many times is Weight 100kg? How many times is Weight below or equal to100kg? Which Weight is the most frequent?

As the modus of continuous variables depends on the classification, we can use `hist` for this, e.g. by

```
> x = hist(students$Weight, plot = FALSE)
> data.frame(x[c("mids", "counts")])
> x = hist(students$Weight, 10, plot = FALSE)
> data.frame(x[c("mids", "counts")])
```

Exercise 2.4 Find out what the 10 refers to by reading the help for `hist`, and then compute the mode for Weight, when using 10, 20, 30, 40 and 50 classes.

2.2 Medians and quantiles, fractions

Read the help for functions `median` and `quantile`

Exercise 2.5 Compute the median for each of the variables in the `students` data set for which this makes sense

Compute the body-mass-index (bmi) as follows:

```
> students$bmi = students$Weight/(students$Length/100)^2
```

Exercise 2.6 Between which two values do the middel 90% of the values lie for variables `bmi` and `Length`, in the students data set?

Exercise 2.7 Above which value of the `Weight` variable do we have 75% of the observations?

Exercise 2.8 Which percentage of the `bmi` observations lie above 25 (overweight, according to Wikipedia)?

(For the last question, use the transformation to logical:

```
> students$bmi > 25
```

and function `sum`.)

2.3 Mean values

It seems there is an outlying value in the Weight data of `students`. Look at the plot given with

```
> library(lattice)
> xyplot(Weight ~ Length | Gender, students)
```

try to explain what you see, and point out which observation is (most) doubtful.

Compute the mean and median weight values:

```
> attach(students)
> split(Length, Gender)
> lapply(split(Length, Gender), mean)
> lapply(split(Length, Gender), median)
> detach()
```

As mean values depend on all observations, it is advisable to compute them from data that do not contain outliers. One approach to do this, is by removing the outlier, and continue, e.g. by

```
> students.clean = students[students$Weight < 200,
    ]
> attach(students.clean)
> lapply(split(Length, Gender), mean)
> lapply(split(Length, Gender), median)
> detach()
```

Another approach to get rid of outliers is by using trimmed means. Find out what is meant by this by reading carefully the documentation of the `trim` argument to function `mean`. Let us try:

```
> attach(students)
> lapply(split(Length, Gender), mean)
> lapply(split(Length, Gender), mean, trim = 0.1)
> lapply(split(Length, Gender), mean, trim = 0.2)
> lapply(split(Length, Gender), mean, trim = 0.3)
> lapply(split(Length, Gender), mean, trim = 0.4)
> lapply(split(Length, Gender), mean, trim = 0.5)
> lapply(split(Length, Gender), median)
> detach()
```

Exercise 2.9 [HAND IN] Explain why the last two commands give the same result.

Exercise 2.10 [HAND IN] What are the advantages and disadvantages of both approaches: (i) manual outlier removal, (ii) computing trimmed statistics, for obtaining mean values that are not influenced by outliers?

Exercise 2.11 What is the interquartile range for `Length` and `Weight`?

Exercise 2.12 What is the standard deviation for `Weight`?

Exercise 2.13 Between which two values do we find the middle 68% of the `Weight` values?

Exercise 2.14 How often does the standard deviation of `Weight` fit in this 68%-range?

3 Meeting 3; Descriptive plots

As a general hint: R under windows has the option `Record plots`; you will find it in the menu only when the plot window is activated (blue). Check this option and you can go to previous/next plots with PgUp/PgDown when in the plot window.

3.1 Histograms and cumulative distributions

Exercise 3.1 Using function `hist`, plot a histogram of variable `bmi` in the students data set.

The book Sachs and Hedderich suggest to use function `ecdf` for to compute empirical cumulative density functions, as in

```
> attach(students)
> plot(ecdf(bmi))
```

Exercise 3.2 [HAND IN] Estimate visually which percentage of the data has `bmi` values large than 25. As an aid, use `abline` to draw *vertical* lines.

There are different ways of drawing grouped histograms. Consider these two cases:

```
> library(lattice)
> histogram(~bmi | Gender, layout = c(2, 1))
> histogram(~bmi | Gender, layout = c(1, 2))
```

Exercise 3.3 [HAND IN] When the goal is to compare `bmi` as a function of gender, which of the two representations do you prefer? Briefly argue why.

3.2 Box-and-whisker plots

Consider the box-and-whisker plot obtained by

```
> boxplot(Length ~ Gender)
```

Exercise 3.4 Why are the three largest observations drawn as small circles for the female students, but not so for the male students?

For the following exercise, consider the following plots

```
> plot(Gender)
> hist(as.numeric(Gender), 2)
> plot(I.am.)
> hist(as.numeric(I.am.), 3)
```

Exercise 3.5 The help of function `hist` mentions, under section `See Also`, that "Typical plots with vertical bars are *not* histograms". (This may be taken as a comment towards a rather popular spreadsheet program.) What is the main difference between bar graphs and histograms, and for what kind of data is each of them used?

3.3 Time series

The famous sunspot data (H. Tong, 1996, Non-Linear Time Series. Clarendon Press, Oxford, p. 471.) contains a feature that is only visible when the aspect ratio of the plot is controlled well.

```
> library(lattice)
> plt <- xyplot(sunspot.year ~ 1700:1988, xlab = "",
     type = "l", scales = list(x = list(alternating = 2)),
     main = "Yearly Sunspots")
> print(plt, position = c(0, 0.3, 1, 0.9), more = TRUE)
> print(update(plt, aspect = "xy", main = "", xlab = "Year"),
     position = c(0, 0, 1, 0.3))
```

Exercise 3.6 [ADVANCED] What is this feature, i.e. what characteristic of these data are visible in the lower, but not in the upper plot?

3.4 Scatter plot matrix

The following exercise is about the scatter plot matrix, obtained e.g. for the `meuse` data set (top soil heavy metal contents; from: Burrough and McDonnell, Principles of GIS, 2nd ed, OUP) by

```
> library(sp)
> data(meuse)
> plot(meuse[c("cadmium", "lead", "copper", "zinc")])
```

Exercise 3.7 [HAND IN] Are the lower-left (column 1, row 4) and top-right (column 4, row 1) sub-plots in this matrix related, and if so how?

4 Probability I.

Some of the following exercises are easiest done by hand, or with a hand calculator. In absence of a hand calculator, expressions can be typed on the R prompt, try e.g.

```
> 3 * 50
> 4/3
```

Consider the following table, which tabulates gender agains the frequencies of passing a given test:

	Male	Female
passed	?	80
failed	10	40

Exercise 4.1 Given that passing the test is independent from gender, what is the expected frequency for the cell with a question mark?

Exercise 4.2 [HAND IN] Given that 10 male students passed the test in the incomplete table above, estimate the following probabilities:

a the probability that someone passes the test, condition to being male

b the probability that someone passes the test, condition to being female

c the probability that someone is female, conditional having passed the test

d the probability that someone is male, conditional having failed the test

e the probability that someone is male AND failed the test

e the probability that someone is female OR passed the test

In the answers, use the formal notations, such as Pr(male | pass), and the appropriate symbols for AND and OR.

Exercise 4.3 [HAND IN] Given there are 30 female and 60 male students, give cross tables with expected frequencies of passing a test, for the case where

a passing the test is independent from gender

b there is an extreme dependence between gender and passing the test (give both cases)

Now generate 6 vectors with random uniform numbers, of length 1, 10, 100, 1000, 10000 and 100000. Compute the fraction of values below 0.6 for each vector.

Exercise 4.4 [HAND IN] How do the values compare to the expected value (0.6), and how could you explain the differences?

Exercise 4.5 When would you expect obtaining a fraction of *exaclty* 0.6?

Exercise 4.6 [HAND IN] Suppose you take part in a quiz. The quiz master shows you three doors. Behind one of the three doors a price is hidden. Opening

the door with the price behind it gives you the price. You are asked to choose one door. Then, the quiz master opens one of the remaining doors, with no price behind it. Now you are allowed to switch your choice to the other remaining closed door. To maximize the probability of getting the price, should you (a) choose the other door, (b) stick with your original choice, or (c) is switching irrelevant for this probability? Explain.

5 Probability II: probability distributions

The following R commands give density, distribution function, quantile function and random value generation for the binomial distribution:

```
> dbinom(x = 5, size = 10, prob = 0.5)
```

is the probability to have 5 sucesses if 10 samples are drawn independently and each has a probability of sucess of 0.5 (`dbinom()` ist the density function).

```
> pbinom(q = 5, size = 10, prob = 0.5)
```

is the probability to have up to 5 (e.g. 0 or 1 ... or 5) sucesses in the same experiment (`pbinom()` is the cummulative distribution function). ■ = x = qbinom(p = 0.1,size = 10, prob = 0.5) is the 0.1-Quantile of the same distribution; i.e. (if the experiment is done very often) the lowest 10% of the results will have no more than x successes.

```
> rbinom(r = 5, size = 10, prob = 0.5)
```

generates randomly 5 possible outcomes of the experiment (for each draw 10 times with sucess of 0.5).

There exist similar functions for other distributions, try

```
> `?`(dpois)
> `?`(dunif)
> `?`(dnorm)
```

For all of them there are the four functions. The use of the first parameter is the same as above. The other parameters are needed to define the distribution, e.g. `pnorm(q = 1, mean = 100, sd = 5)` gives the values of a normal distribution with mean = 100 and standard deviation = 5.

In the following R commands, a Bernoulli sequence is generated

```
> as.numeric(runif(20) < 0.25)
```

Generate a similar sequence using the function `rbinom`.

Exercise 5.1 [HAND IN] Which generating probability of success (p) has to be used?

Exercise 5.2 Suppose we have a bowl with 25% red balls, and a single experiment involves the sampling of 10 balls with replacement, counting the number

of red balls in the sample. Generate the outcome of 100 such experiments.

Exercise 5.3 Draw a histogram of the result.

Exercise 5.4 Repeat this for 100000 experiments, and draw a histogram of the result. Why are there empty bars in this graph?

With the command

```
> plot(0:10, dbinom(0:10, 10, 0.25))
```

you plot the probabilities for all the outcomes.

Exercise 5.5 Create a similar plot for all the probabilities when when the experiment uses the same bowl, but draws 100 instead of 10 balls independently.

Exercise 5.6 Check that the latter probabilities sum to exactly one.

Suppose we are interested in the occurence pattern of the Green tiger beetle, in our study site of 2 km $\times$ 3 km. The hypothesis is that per square meter on average 0.05 of such beetles live. We have plots of 3 $\times$ 3 meter, and assume that we do catch all beetles in such a plot during an experiment.

For a normal distribution with mean 10 and standard deviation 1, the probability of having a value below 12 is found by

```
> pnorm(12, 10, 1)
```

Exercise 5.7 [HAND IN] Assuming that body length of the students data set is normally distributed, and assuming that the sample mean and standard deviation are the true values, What is the probability of having a student with length *larger* than 195 cm?

Exercise 5.8 Which of the distributions are discrete / continuous: binomial distribution, Poisson distribution, uniform distribution, normal distribution, Gaussian distribution?

Imagine a multiple choice test with 40 questions, each having 4 answers whereof 1 is correct. To pass the test you need 20 correct answers. For the

following exercises give the R command and the value(s). If you can not find the command, give the distribution and value (e.g. normal distribution, mean = 9, sd = 4, density in x = 2).

Exercise 5.9 You did not prepare and rely on guessing. How could you use R to give random answers?

Exercise 5.10 If you guess all answers, what is the probability that you fail? (What kind of experiment is it, how are the parameters?)

Exercise 5.11 The time students need is approximately normally distributed with mean 60 minutes and standard deviation 10 minutes. After how much time will 90% of the students have finished?

Exercise 5.12 [HAND IN] Create a normal probability plot for the variable `bmi` in the `students` data set. Add the line that indicates a reference normal distribution.

Can we consider `bmi` as approximately normally distributed? What are the mean and standard deviation (do not calculate but describe how to tell from the plot).

6 Sampling, standard error and confidence intervals

Load the http://www.ai-geostats.org/index.php?id=45SIC97 data, imported to an R file, from http://ifgi.uni-muenster.de/~epebe_01/Geostatistics/sic97.RData to your working directory. Start R by double-clicking the icon of the downloaded file.

Show the data, using

```
> library(sp)
> image(demstd, axes = T, col = terrain.colors(20))
```

(Note that `sp` needs to be loaded, otherwise the correct method for imaging spatial grids is not available.)

The area shown is a digital terrain model, projected in some lambert projection system (with undocumented offset). As an alternative, you might use

```
> spplot(demstd, col.regions = terrain.colors(20))
```

which takes longer and shows a scale bar on the side.

To plots generated by `image` it is however easier to add information incrementally:

```
> image(demstd, axes = T, col = terrain.colors(20))
> pts = spsample(demstd, n = 100, type = "random")
> points(pts, pch = 3)
```

We will consider this DTM to be the target population. Compute it's mean, variance and standard deviation, and write them down:

```
> mean(demstd[[1]])
> var(demstd[[1]])
> sqrt(var(demstd[[1]]))
```

Exercise 6.13 [HAND IN] How do we compute the standard error for the mean, when we take a simple random sample of 10 observations? Write down the equation used and the value found.

We can take a simple random sample of size 10:

```
> pts = spsample(demstd, 10, "random")
> summary(pts)
> image(demstd, axes = T, col = terrain.colors(20))
> points(pts)
> pts.val = overlay(demstd, pts)
> summary(pts.val)
```

Exercise 6.14 [HAND IN] compute the 95% confidence interval for the population mean, based on this sample of 10 observations, using the true (population) variance; give both the equation and the answer.

Look at the range and the distribution form of the altitude data:

```
> summary(demstd[[1]])
> hist(demstd[[1]])
```

Exercise 6.15 [HAND IN] Describe the range and form of the distribution, and explain what you see (this should be an easy question).

Now we are going to do draw 100 samples of size 3 each, and compute their mean

```
> m3.srs = replicate(100, mean(overlay(demstd, spsample(demstd,
     3, "random"))[[1]]))
```

```
> m3.srs
> mean(m3.srs)
> sqrt(var(m3.srs))
> hist(m3.srs)
```

Do the same thing for samples of size 10, and 100.

Exercise 6.16 [HAND IN] (i) How does the range of the distribution change when the sample size becomes larger? Explain this phenomenon, and give a general equation for a measure of spread for the simple random sample means of a given sample size. (ii) How does the distribution change when sample size increases?

Exercise 6.17 [HAND IN] Repeat the above procedure, instead of simple random sampling now use (i) regular sampling, (ii) stratified sampling and (iii) systematically unaligned sampling. How does the range and standard deviation of the means obtained change, compared to simple random sampling? Try to explain why this is the case.

7 Confidence intervals II.

Load the pm10 data file `pm10.txt` from the course web site. If needed, set path (`setwd`) to the place where you save the file, and read file (the header contains the names of the monitoring stations (VMUE, VMSS) and must therefore be read seperately, set: header=TRUE)

```
> pm10.file = "pm10.txt"
> pm10 = read.table(pm10.file, header = TRUE)
```

The data are the weights of respirable dust particles (below 10 μm) in mg/m^3 air from two monitoring stations in Münster (Steinfurter Straße: VMSS, Friesenring: VMUE) measured during winter 2006/07. The measured values are averages over consecutive periods of 30 min. Days with missing or negative data have been omitted.

7.1 Confidence intervals for different confidence levels, σ known

The whole dataset VMUE shall be the population. We know the standard deviation σ. Calculate it and call it `sVMUE`.

Exercise 7.1 Compute the standard error for the estimator of the mean for a sample size of 50 and call it `seVMUE50`.

Draw a random sample of size 50 from VMUE:

```
> VMUE50 = sample(pm10$VMUE, 50)
```

(Explain what VMUE50 is by looking at its construction, and repeating it.)

Exercise 7.2 [HAND IN] Calculate from this random sample

a the 90%-confidence interval for the mean using σ

b the 95%-confidence interval for the mean using σ

c the 99%-confidence interval for the mean using σ

Compare the intervals, explain the differences.

Exercise 7.3 Calculate with σ and the random sample the values

a above which the population mean is with probability 95%.

b below which the population mean is with probability 95%.

Write down the onesided confidence intervals and compare with the twosided confidence intervals.

7.2 Confidence intervals for the mean, σ not known

Draw a random sample of size 5 from VMSS, and call it `VMSS5`. Calculate from this random sample mean (=mVMSS5) and sample standard deviation (s=sVMSS5).

Exercise 7.4 [HAND IN] Calculate with the method described in the lecture:

a the 90%-confidence interval for the mean using s.

b the 95%-confidence interval for the mean using s.

c the 99%-confidence interval for the mean using s.

What is the difference from calculating confidence intervals using σ?

Hint: for the 90% confidence interval, use

```
> c(mVMSS5 + qt(0.05, 4) * sVMSS5/sqrt(5), mVMSS5 +
    qt(0.95, 4) * sVMSS5/sqrt(5))
```

Look for the required values in the results from

```
> t.test(VMSS5)
```

Exercise 7.5 [HAND IN] You want to know the population mean from a sample. The 95% confidence interval has to be smaller than 20 (i.e. you do not want to underestimate the mean by more than ten and not to overestimate it by more than ten, with probability 0.95). How big must the size of the sample be? (This can be calculated directly or by trying it out. Note you do not need the mean but only the population standard deviation `sVMSS` for the calculation.)

7.3 t-test

Draw random samples of size 100 from both columns (see 7.1), and save them to vectors with names `VMUE100` and `VMSS100`

Do a t-test for the difference of the means of VMUE and VMSS (mVMUE-mVMSS) based on those samples

```
> t.test(VMUE100, VMSS100, var.equal = TRUE)
```

Exercise 7.6 Can you be 95% confident that the population means mVMUE, mVMSS are different (short explanation)?

7.4 paired t-test

Draw samples

a)

```
> VMUE100 = sample(pm10$VMUE, 100)
> VMSS100 = sample(pm10$VMSS, 100)
```

b)

```
> r100 = sample(1:length(pm10$VMUE), 100)
> VMUE100 = pm10$VMUE[r100]
> VMSS100 = pm10$VMSS[r100]
```

Exercise 7.7 What is the difference?

Exercise 7.8 Give the 95% confidence interval for the difference of the means with the data from b).

Exercise 7.9 Do a paired t-test, give the 95% confidence interval for the mean of the differences. See hint below;

```
> t.test(VMUE100 - VMSS100, var.equal = TRUE)
> t.test(VMUE100, VMSS100, paired = TRUE, var.equal = TRUE)
```

Exercise 7.10 [HAND IN]Remember the data are measurements from two stations, each column belongs to a station, each row to a date. What happens in 7.8 compared to 7.9? Which method tells you more about the difference between the two stations at one time?

8 Classical tests: t-test

Load the student data set and try:

```
> attach(students)
> bmi[Gender == "male"]
```

Exercise 8.1 [HAND IN] Test, for the students data set, whether the bmi for students depends on gender, assuming equal variance of both groups. Write out the first 4 steps of the formal procedure (hypothesis testing), and use the p value to find the conclusion.

Exercise 8.2 [HAND IN] Explain in your own words what the p-value of the test result mean, and how it is related to the t value.

Exercise 8.3 [HAND IN] Test, for the students data set, whether the bmi for male students is equal to 25. Use the t-value and follow the 7 steps of formal testing. Compare your results to those obtained by using the function `t.test` for this.

Exercise 8.4 Test, for the students data set, whether the bmi for male students *exceeds* 25

Exercise 8.5 Test, for the students data set, whether the bmi for male students is *smaller than* 25

Exercise 8.6 Can you do a paired t-test to compare bmi for male and female students? If yes, show how. If no, explain.

Exercise 8.7 [HAND IN] Explain how in the first and third exercise of this section the t value is calculated; give also details how in both cases the standard error of mean in the denominator is calculated

9 Power, and ANOVA

9.1 Sample Size and Power

Work with the students data set, and try the following:

```
> attach(students)
> t.test(Length ~ Gender, var.equal = TRUE)
```

Exercise 9.1 [HAND IN] Assume that the groups have equal size, and that the variance of both groups is equal. How large a sample size would result in a confidence interval half as wide? (Hint: this is easiest done by hand)

Do the t computation of the example above step by step:

```
> m1 = mean(Length[Gender == "male"])
> m2 = mean(Length[Gender != "male"])
> var1 = var(Length[Gender == "male"])
> var2 = var(Length[Gender != "male"])
> n1 = length(Length[Gender == "male"])
> n2 = length(Length[Gender != "male"])
> ((n1 - 1) * var1 + (n2 - 1) * var2)/(n1 + n2 -
    2)
> v = ((n1 - 1) * var1 + (n2 - 1) * var2)/(n1 +
    n2 - 2)
> (m1 - m2)/sqrt(v * (1/n1 + 1/n2))
```

and check that the final value equals `t`.

Exercise 9.2 [HAND IN] Describe in words what `v` is.

The function `power.t.test` can compute the power, `delta` (difference of alternative means) or required sample size for a test if all other parameters are given:

```
> power.t.test(n = 10, delta = NULL, sd = 1, sig.level = 0.05,
    power = 0.95)
```

Computes the minimal distance of alternative hypotheses (`delta`) which can be destinguished by a test at significance level 0.05 and with power 0.95 if sample size is 10 and standard deviation of the data is 1. This can also be computed for one-sided tests with `alternative = "one.sided"` and for other test types (`type = "one.sample", type = "paired"`).

Exercise 9.3 [HAND IN] Compute the power of the test above: use the square-root of `v` as `sd` and `delta = m1 - m2`, set the correct sample size (see NOTE). You can use the default test type and alternative.

Exercise 9.4 Which sample size would have been required to obtain a power of *at least* 0.95, and of *at least* 0.9? (Note that sample size cannot be a broken number)

9.2 ANOVA

Exercise 9.5 [HAND IN] Does factor `Year` have a significant effect on `Weight`(please hand in the ANOVA output)?

```
> summary(aov(Length ~ Gender, students))
```

Exercise 9.6 [HAND IN] Explain how `F` got computed from the data, and what the Pr value means

A balanced sample from the students dataset is given by

```
> Length = c(165, 176, 158, 174, 180, 180, 163,
     187, 180, 189, 173, 185, 165, 159, 160, 159,
     183, 165, 163, 183, 182, 188, 185, 188)
> I.am. = rep(rep(c("small", "medium", "tall"),
     each = 4), 2)
> Year = rep(c(8, 9), each = 12)
> studentsSample = data.frame(Length, I.am., Year)
```

and compare the ANOVA models

```
> summary(aov(Length ~ I.am., studentsSample))
> summary(aov(Length ~ I.am. + Year, studentsSample))
```

Exercise 9.7 Why does significance of `I.am.` change between the models? (Explain how `F` got computed in each case.)

10 Simple linear regression

Consider the `meuse` data set.

```
> library(sp)
> data(meuse)
```

The heavy metal concentrations are thought to originate from sediment, deposited by the Meuse river, and the near-river soil samples clearly contain higher heavy metal concentrations. Also the concentrations of the different heavy metals are correlated.

10.1 Correlation

Plot the heavy metal concentrations against each other.

```
> plot(meuse[c("cadmium", "lead", "copper", "zinc")])
```

Exercise 10.8 [HAND IN] Try to tell which metals are strongest and which are weakest correlated. Compute these two correlations. Test, if the smaller one is still significant (see lecture for command; give the command and a concluding sentence).

10.2 Simple linear regression

Do a linear regression of `zinc` on `dist`:

```
> lr.out = lm(zinc ~ dist, meuse)
> lr.out
> summary(lr.out)
> sd(residuals(lr.out))
```

Exercise 10.9 [HAND IN] Describe in your own words what the residual standard error is, and how it is computed.

Plot the data and fitted line:

```
> plot(zinc ~ dist, meuse)
> abline(lm(zinc ~ dist, meuse))
```

Exercise 10.10 The plot shown contains the least-squares line of zinc as a linear function of dist, distance to river. Comment on whether this model is useful to predict zinc concentrations in the area.

Hint: also look at the predicted values (regression line values) for the data points, as in

```
> hist(predict(lm(zinc ~ dist, meuse)))
```

An alternative is to look at the same pair of variables, but to transform the dependent variable, e.g. by taking the logarithm:

```
> plot(log(zinc) ~ dist, meuse)
> summary(lm(log(zinc) ~ dist, meuse))
> abline(lm(log(zinc) ~ dist, meuse))
```

Exercise 10.11 [HAND IN] Is this model better, worse or of equal value compared to the first, when we consider how realistic the fitted, or predicted values are?

Exercise 10.12 [HAND IN] Write down the regression equation for the fitted values of log zinc, with the estimated values for the regression coefficients, as in

$$\log(\text{zinc}) = a + b \cdot \text{dist} + e$$

but then with a and b replaced with their estimated values.

Exercise 10.13 Use the regression equation from above to compute the modelled `log(zinc)` value for `zinc` where `dist` $= 0.3$. You may compare it with the results from `predict`

```
> new = data.frame(dist = c(0.3))
> predict(lm(log(zinc) ~ dist, meuse), new)
```

Exercise 10.14 What is the modelled value of `zinc` for `dist` $= 0.3$?

Exercise 10.15 [HAND IN] Compute a 95% confidence interval for the regression slope in this last regression model, either by hand or with R.

11 Linear regression II

11.1 Regression calculation by hand

This is a regression completely done by hand, you may use a calculator or R. Consider the following data (Wonnacott & Wonnacott p. 358),

X fertilizer (lb/acre)	Y yield (bu/acre)
100	40
200	50
300	50
400	70
500	65
600	65
700	80

For loading the data to R use

```
> X = c(100, 200, 300, 400, 500, 600, 700)
> Y = c(40, 50, 50, 70, 65, 65, 80)
```

Exercise 11.1 Plot the data by hand or by R. Copy the plot as a bitmap (.bmp) and add by hand the regression line, remember the residuals (vertical distance from points to line) should be minimal.

Exercise 11.2 To calculate the regression coefficients use the formulas (see lecture)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

You may do that completely by hand, filling the following table

	Data		Deviation		Products	
	X	Y	$x = X - X$	$y = Y - Y$	xy	x^2
	100	40	-300	-20	6000	90000
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\sum$	$\bar{X} = 400$	$\bar{Y} =$	$\sum x = 0$	$\sum y = 0$	$\sum xy =$	$\sum x^2 =$

Exercise 11.3 [HAND IN] Write the regression equation, using the regression coefficients calculated in the last exercises or by R.

Exercise 11.4 Plot the regression line (by hand or by R).

Exercise 11.5 [HAND IN] Which yield would you expect if using 250 lb fertilizer per acre?

11.2 Multiple Linear Regression with Dummy Variables

Load the `students` dataset into the path.

```
> attach(students)
```

Turn `Gender` into a numeric dummy variable (female: 0, male: 1)

```
> nGender = as.numeric(Gender) - 1
```

We want to do regression of Weight by Length and to include Gender if convenient. First have a look at the data

```
> plot(Weight ~ Length, col = (nGender + 1) * 2)
```

Exercise 11.6 Copy the plot as a bitmap and add a regression line (in black) by eye. Imagine you would do regression on males and females separately and add a coloured regression line for each.

Build three models.

```
> lm.L = lm(Weight ~ Length)
> lm.LaG = lm(Weight ~ Length + Gender)
> lm.LiG = lm(Weight ~ Length * Gender)
```

Exercise 11.7 [HAND IN] Complete the following table

Model	general Form	regression formula	R_a^2
lm.L	$Y = \beta_0 + \beta_1 X_1$	`Weight` = -120 + 1.1 * `Length`	0.4649
lm.LaG	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$		
lm.LiG	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 x_3$		

Save the coefficients

```
> b.L = coef(lm.L)
> b.LaG = coef(lm.LaG)
> b.LiG = coef(lm.LiG)
```

Exercise 11.8 [HAND IN] Consider the second model `lm.LaG` and the third model `lm.LiG`. Each of them can be split into a model for female and a model

for male (`Gender` is 0 or 1). Give for each of the models the equation for males and females.

Plot the models (male and female in colors)
(Hint: in R open one plot then set "History" to "record" then you can switch between plots by "page up" and "page down".)

```
> plot(Weight ~ Length, col = (nGender + 1) * 2,
     main = "Weight = -120 + 1.1 Length")
> abline(b.L[1], b.L[2])
> plot(Weight ~ Length, col = (nGender + 1) * 2,
     main = "Weight = -143 + 1.2 Length - 4.9 nGender",
     add = FALSE)
> abline(b.LaG[1] + b.LaG[3], b.LaG[2], col = "blue")
> abline(b.LaG[1], b.LaG[2], col = "red")
> plot(Weight ~ Length, col = (nGender + 1) * 2,
     main = "Weight = -204 + 1.6 Length - 155 nGender - 0.9 Length * nGender",
     add = FALSE)
> abline(b.LiG[1] + b.LiG[3], b.LiG[2] + b.LiG[4],
     col = "blue")
> abline(b.LiG[1], b.LiG[2], col = "red")
```

Exercise 11.9 [HAND IN] What do the models describe? (assign models to the descriptions below if they fit)

- The influence of only Length on Weight.
- The influence of Length on Weight together with indirect, hidden effects, e.g. that women are smaller.
- The influence of Length on Weight, included that women are smaller.
- The influence of Length on Weight, included that women are smaller and tall women compared to small ones are heavier than tall men compared to small men.
- The infuence of Length on Weight for women and the infuence of Length on Weight for men.

Exercise 11.10 What is the true influence of only Length on Weight? Which of the models is best?

Exercise 11.11 [HAND IN] Are all parameters in all models significant at the 5% level? If not, does it make sense to keep them anyway?

12 Regression extensions

12.1 Regression on non-linearly transformed variables

Carefully read the help of `predict.lm`, the predict method that applies to objects of class `lm`, e.g. by the command `?predict.lm`.

Load the data

```
> library(sp)
> data(meuse)
```

Compare the linear and quadratic models of log-zinc with distance to river in the meuse data set:

```
> plot(zinc ~ dist, meuse, ylim = c(-500, 2000))
> abline(0, 0)
> d = seq(-0.1, 1, by = 0.01)
> new = data.frame(dist = d)
> lm1 = lm(zinc ~ dist, meuse)
> pr1 = predict(lm1, new)
> lines(d, pr1, col = "red")
```

Exercise 12.1 What is given by `predict(lm(zinc ~ dist, meuse), new)`?

Add more regression lines

```
> lm2 = lm(zinc ~ dist + I(dist^2), meuse)
> pr2 = predict(lm2, new)
> lines(d, pr2, col = "green")
> lm3 = lm(zinc ~ dist + sqrt(dist), meuse)
> pr3 = predict(lm3, new)
> lines(d, pr3, col = "blue")
> legend(0.7, 2000, legend = c("lm1", "lm2", "lm3"),
     col = c("red", "green", "blue"), lty = c(1,
         1, 1))
```

(Set "History" to "record", you will need plots several times to add something.)

Exercise 12.2 Discuss the advantages and disadvantages of using a linear (lm1), a quadratic function in dist (lm2), or a quadratic function in square-root of distance (lm3). How realistic are the predictions?

Compare also the following two models for transformed data:

```
> plot(log(zinc) ~ dist, meuse)
> lm4 = lm(log(zinc) ~ dist, meuse)
> abline(lm4, col = "orange")
> lm5 = lm(log(zinc) ~ sqrt(dist), meuse)
> plot(log(zinc) ~ sqrt(dist), meuse)
> abline(lm5, col = "violet")
```

They can be plotted to the original scale, with a little bit work (go back to the plot of the above models or start again plotting the values by `plot(zinc ~ dist, meuse)`):

```
> pr4 = predict(lm4, new)
> lines(d, exp(pr4), col = "orange")
> pr5 = predict(lm5, new)
> lines(d, exp(pr5), col = "violet")
> legend(0.7, 1500, legend = c("lm4", "lm5"), col = c("orange",
    "violet"), lty = c(1, 1))
```

Exercise 12.3 [HAND IN] Which of the 5 models (lm1, ..., lm5) do you prefer, why?

Suppose that one of the observed values is zero, as in

```
> meuse$zinc - 113
> summary(meuse$zinc - 113)
```

we try to do regression on the logarithm of these values:

```
> lm(log(zinc - 113) ~ sqrt(dist), meuse)
```

Exercise 12.4 [HAND IN] Why does this fail?

12.2 Prediction

(Prediction is found at W&W p. 384-386)

Predicting values can be done under all circumstances, defined by the values of the X-variables, or predictors.

One particular case, common to geostatistics, is that where the X-variables come as maps. In the models defined in 12.1, `dist` and `zinc` vary in space

```
> data(meuse.grid)
> coordinates(meuse.grid) = c("x", "y")
> gridded(meuse.grid) = TRUE
> spplot(meuse.grid, col.regions = bpy.colors())
```

Therefore the models give a prediction for each point in space.

```
> meuse.grid$lzn.pred1 = predict(lm1, meuse.grid)
> meuse.grid$lzn.pred2 = predict(lm2, meuse.grid)
> meuse.grid$lzn.pred3 = predict(lm3, meuse.grid)
> meuse.grid$lzn.pred4 = exp(predict(lm4, meuse.grid))
> meuse.grid$lzn.pred5 = exp(predict(lm5, meuse.grid))
> spplot(meuse.grid[c("lzn.pred1", "lzn.pred2",
    "lzn.pred3", "lzn.pred4", "lzn.pred5")], col.regions = bpy.colors())
```

Exercise 12.5 [HAND IN] Why are there high values (yellow) far away from the river in the `lzn.pred2` plot? Why are the contours in the `lzn.pred5` plot closer for high values at the river than for low values far away?

12.3 Uncertainty, Confidence Intervals

Now we continue with only `lm5` on the transformed data.

```
> plot(log(zinc) ~ sqrt(dist), meuse)
> abline(lm5, col = "violet")
```

Exercise 12.6 [HAND IN] There is some uncertainty in estimating the parameters of the line. Copy the plot as a bitmap and add some other possible regression lines. For which X values we know best where the regression line should be?

The model can calculate uncertainties (here for the first point)

```
> predict(lm5, meuse[1, ], se.fit = TRUE, interval = "confidence")
```

The values are related

```
> upper = predict(lm5, meuse[1, ], se.fit = TRUE,
     interval = "confidence")$fit[1, 3]
> lower = predict(lm5, meuse[1, ], se.fit = TRUE,
     interval = "confidence")$fit[1, 2]
> sefit = predict(lm5, meuse[1, ], se.fit = TRUE,
     interval = "confidence")$se.fit
> (upper - lower)/(2 * sefit * qt(0.975, 153))
```

Exercise 12.7 Why is the result always = 1?

Add confidence intervals for the means to the plot:

```
> ciM = predict(lm5, interval = "confidence")
> points(sqrt(meuse$dist), ciM[, 2], pch = "-",
     col = "blue")
> points(sqrt(meuse$dist), ciM[, 3], pch = "-",
     col = "blue")
```

Exercise 12.8 [HAND IN] explain in words what the intervals between each pair of blue minus symbols represent.

Plot the standard errors as a map

```
> meuse.grid$lzn.se = predict(lm5, meuse.grid, se.fit = TRUE)$se.fit
> spplot(meuse.grid["lzn.se"], col.regions = bpy.colors())
```

Exercise 12.9 Why are the lowest standard error values not found near to or far from the river, but somewhere in between?

Exercise 12.10 Which value would you expect if you would really measure `zinc` at a location with `sqrt(dist) = 0.2`?

Add confidence intervals for the predictions to the plot (go back to the plot of the regression line):

```
> ciP = predict(lm5, interval = "prediction")
> points(sqrt(meuse$dist), ciP[, 2], pch = "-",
     col = "red")
> points(sqrt(meuse$dist), ciP[, 3], pch = "-",
     col = "red")
```

Exercise 12.11 [HAND IN] Explain in words what the intervals between each pair of red points represent, do the red minus symbols lie on straight lines?

12.4 Generalized linear models

Generalized linear models avoid the problem, not being able to transform observations, by defining the model in two steps:

$$\mathrm{E}(y_i) = \pi_i$$

$$f(\pi_i) = \beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p} + e_i$$

where $f()$ is called the link function. Typically, for count data this link function is the log-function, allowing zero as observed values. For binomial (0/1) data, the logistic link function is often used:

$$f(\pi_i) = \mathrm{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$$

and the resulting model is called *logistic regression*.

Exercise 12.12 What is the logistic transformed value for an observation of 0, and of 1?

A simple logistic model can be formed by regressing gender as a function of weight, in the student data set:

```
> student = students[students$Length < 210, ]
> glm1 = glm(Gender ~ Weight, student, family = "binomial")
> glm1
> summary(glm1)
> xrange = 50:150
> pr = predict(glm1, data.frame(Weight = xrange))
> plot(xrange, pr, type = "l")
> title("predictions on the logistic scale")
> plot(as.numeric(Gender) - 1 ~ Weight, student)
> pr = predict(glm1, data.frame(Weight = xrange),
    type = "response")
> lines(xrange, pr)
> title("predictions on the observation scale")
```

Exercise 12.13 Why are the observations not shown in the first plot?

Look at the predicted values for each of the Weight values:

```
> cbind(xrange, pr)
```

and consider the predicted values from the glm as probabilities of being one (i.e., having gender "male").

Exercise 12.14 [HAND IN] Above which weight value would you predict the gender of a person with unknown gender as being male?