Project 1: Heating in Sønderborg

Formalities, structure and expectations for the first mandatory project

The assignment consists of two parts. The first part focuses on descriptive analysis of the data. The second part is primarily about confidence intervals and hypothesis tests.

The assignment is formulated in such a way that it can be solved in small "easy" steps. In practice, the assignment must be solved using the statistical software R. Some R code is provided in order to make it easy to get started with the project. However, the code is not complete, and you are encouraged to explore new features in R while working on the project. For example, you could add suitable titles to the plots, or use R's built-in functions for computing confidence intervals and testing hypotheses.

The results of the analysis must be documented in the report using tables, figures, mathematical notation, and explanatory text. Relevant figures and tables must be included within the text, not in the appendix. Present the results of your analysis as you would when explaining them to one of your peers.

Divide the report into subsections, one for each of the questions to be answered.

The report must be handed in as a pdf file. R code should not be included in the report itself but must be handed in as an appendix (a .R-file). The report and appendix must be handed in under Assignments on Learn at: `Projekt 1: Varmeforbrug i Sønderborg`

The report text should not exceed 6 pages (excluding figures, tables, and the appendix). A normal page contains 2400 characters.

Figures and tables cannot stand alone - it is important that you describe and explain the R output in words.

Figures and tables are not included in the assessment of the length of the report. However, it is not in itself an advantage to include many figures, if they are not relevant!

You may work together in groups, but the report must be written individually. Questions about the project can be addressed to the teaching assistants, see the guidelines on the `Projects` page of the course website.

Introduction

This project focuses on the daily heat consumption for four houses in Sønderborg during the period from October 2008 to June 2011. The relation between the surrounding climate and the heat consumption can give insight into the actual level of insulation of the houses. E.g. the relation between wind speed and heat consumption is an indicator of how airtight/exposed the houses are.

Ideally, this kind of data can be used to make an empirical energy signature for houses. As long as the residents' behaviour doesn't change too much during the period of observation, the analysis will be independent of the indoor temperature. The houses are detached single-family houses.

Reading the data into R

Make a folder for the project on your computer. Download the project material from Learn and unzip it to the folder that you just made.

Then, open the data file `soenderborg1_data.csv` (e.g., in RStudio, File → Open File) in order to see the contents of the file. Note that the first row (referred to as a *header*) contains variable names, and that the subsequent rows contain the actual observations. Variable names and observations of the individual variables are separated by a ';' (therefore `.csv`: "comma separated values", though here it is a semi-colon).

The data consists of daily observations of the heat consumption in four houses together with daily averages of centrally observed climatic variables. The file contains the following columns/variables:

Variable	**Explanation**
t	Date
Ta	Ambient air temperature (°C)
G	Global radiation (W/m^2)
Ws	Wind speed (m/s)
Q1	Heat consumption in House 1 (kW/day)
Q2	Heat consumption in House 2 (kW/day)
Q3	Heat consumption in House 3 (kW/day)
Q4	Heat consumption in House 4 (kW/day)

Open the file soenderborg1_english.R, which contains some R code that can be used for the analysis. First, the "working directory" must be set to the directory on the computer, which contains the files for the project:

```
## In RStudio the working directory is easily set via the menu
## "Session -> Set Working Directory -> To Source File Location"
## Note: In R only "/" is used for separating in paths
## (i.e. no backslash).
setwd("Replace with path to directory containing project files.")
```

Now the data may be read into R using the following code:

```
## Read data from soenderborg1_data.csv
D <- read.table("soenderborg1_data.csv", header=TRUE, sep=";",
                as.is=TRUE)
```

D becomes a "data.frame" (a kind of table), which contains the data that was read into R (see the introduction to R in Section 1.5 of the book).

Descriptive analysis

The purpose of the first part of the project is to carry out a descriptive analysis of the data. In a report it is important to present the data and describe it to the reader. For example, this can be done using summary statistics and suitable figures.

Start by running the following commands to get a simple overview of the data:

```
## Dimensions of D (number of rows and columns)
dim(D)
##  Column/variable names
```

```
names(D)
## The first rows/observations
head(D)
## The last rows/observations
tail(D)
## Selected summary statistics
summary(D)
## Another type of summary of the dataset
str(D)
```

a) Write a short description of the data. Which variables are included in the dataset? Are the variables *quantitative* and/or *categorized* (or date variables)? (Categorized variables are only introduced in Chapter 8, but they are simply variables which divide the observations into categories/groups - e.g. three categories: low, medium, and high.) How many observations are there? Which time period is covered by the observations (date of first and last observations)? Are there any missing values?

The following code may be used to generate a "density histogram" describing the empirical density of the heat consumption of House 1 (see Section 1.6.1):

```
## Histogram describing the empirical density of the daily heat
## consumptions of House 1 (histogram of daily consumptions normalized
## to have an area of 1)
hist(D$Q1, xlab="Heat consumption (House 1)", prob=TRUE)
```

b) Make a density histogram of the daily heat consumption of House 1. Use this histogram to describe the empirical distribution of the daily heat consumption. Is the empirical density symmetrical or skewed? Can the heat consumption be negative? Is there much variation to be seen in the observations?

Note: In a *skewed* distribution, the probability mass is not symmetrically distributed around the median. In a left-skewed distribution, the left tail is longer than the right tail (and, typically, the mean will lie to the left of the median). Similarly, in a right-skewed distribution, the right tail is the longer of the two (usually, with the mean to the right of the median).

When observations are recorded regularly over time, the data is often referred to as a *time series*. Thus, the daily heat consumptions of each house constitute a times series. For time series, it is often relevant to make figures illustrating the data over time. Here, it is first necessary to tell R that the variable t should be treated as a date variable. This can be done using the following code:

```
## Converts the variable 't' to a date variable in R
D$t <- as.Date(x=D$t, format="%Y-%m-%d")
## Checks the result
summary(D$t)
```

A plot illustrating the daily heat consumption over time for each house for the period 2 October 2008 to 1 October 2010 (coloured according to house) can now be made using the following R code:

```
## Plot of heat consumption over time
plot(D$t, D$Q1, type="l", xlim=as.Date(c("2008-10-02","2010-10-01")),
     ylim=c(0,9), xlab="Date", ylab="Heat consumption", col=2)
lines(D$t, D$Q2, col=3)
lines(D$t, D$Q3, col=4)
lines(D$t, D$Q4, col=5)
## Add a legend
legend("topright", legend=paste0("Q", c(1,2,3,4)), lty=1, col=2:5)
```

Note that the data has missing values – shown as gaps in the time series plot.

c) Make a plot illustrating the daily heat consumption over time for the period 2 October 2008 to 1 October 2010 (coloured according to house). Describe the development of the heat consumption over time in words. Is it similar across the four houses? Is the daily heat consumption stable? Is is possible to identify the heating season? Are there any time periods with unexpected levels of heat consumption?

When doing data analysis, it is often useful to be able to take subsets of the data. This can be done in R using, e.g., the subset function. See the remark on p. 11 as well.

In the further analysis, only data from the period January-February 2010 is to be used. Use the R code below to make a subset of the data which only includes the observations from these two months:

```
## Subset of the data: only Jan-Feb 2010
Dsel <- subset(D, "2010-01-01" <= t & t < "2010-3-01")
```

The following R code makes a box plot of the daily heat consumption by house for the first two months of 2010:

```
## Box plot of daily heat consumption by house
boxplot(Dsel[ ,c("Q1","Q2","Q3","Q4")],
        xlab="House", ylab="Heat consumption")
```

d) Make a box plot of the daily heat consumption in January-February 2010 by house. Use this plot to describe the empirical distribution of the daily heat consumption of the four houses. Are the distributions symmetrical or skewed? Does there seem to be a difference between the distributions (if so, describe the difference)? Are there extreme observations/outliers?

The empirical distribution of the daily heat consumptions during January-February 2010 for each of the four houses may also be quantified using summary statistics as in the following table:

House	**Number of obs.**	**Sample mean**	**Sample variance**	**Sample Std. dev.**	**Lower quartile**	**Median**	**Upper quartile**
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
House 1							
House 2							
House 3							
House 4							

R code like the following may be used to fill in the empty cells in the table (see also the remark on p. 12):

```
## Total number of observations for House 1 during Jan-Feb 2010
## (doesn't include missing values if there are any)
sum(!is.na(Dsel$Q1))
## Sample mean of daily heat consumption for House 1, Jan-Feb 2010
mean(Dsel$Q1, na.rm=TRUE)
## Sample variance of daily heat consumption for House 1, Jan-Feb 2010
var(Dsel$Q1, na.rm=TRUE)
## etc.
##
## The argument 'na.rm=TRUE' ensures that the statistic is
## computed even in cases where there are missing values.
```

e) Fill in the empty cells in the table above by computing the relevant summary statistics for the daily heat consumption of each of the four houses during the first two months of 2010. Which additional information may be gained from the table, compared to the box plot?

Statistical analysis

The purpose of the second part of the project is to perform a simple statistical analysis of the daily heat consumption of the houses. This includes specifying statistical models for heat consumption, estimating the parameters of these models, performing hypothesis tests, and computing confidence intervals.

Confidence intervals and hypothesis tests

The following R code may be used to make a qq-plot. This plot can be used to investigate whether the daily heat consumptions of House 1 may be assumed to be normal distributed:

```r
## qq-plot of daily heat consumption (House 1)
qqnorm(Dsel$Q1)
qqline(Dsel$Q1)
```

f) Specify separate statistical models describing the daily heat consumption of each of the four houses (see Remark 3.2). Estimate the parameters of the four models (mean and standard deviation). Carry out model validation (see Chapter 3 and Section 3.1.8). Since, in this case, confidence intervals and hypothesis tests involve the distribution of an average, it might also be useful to include the central limit theorem (Theorem 3.14) in the discussion.

In practice, situations will arise where it is *not* appropriate to assume that the assumptions of a model are satisfied. In these cases, one often considers whether a transformation of the data might improve the situation. (See Chapter 3.1.9.) Note that after a transformation, the interpretation of the results on the original scale changes. In this specific project, however, the intention is *not* for you to transform the data.

g) State the formula for a 95% confidence interval (CI) for the mean daily heat consumption of House 1 during January - February 2010 (see Section 3.1.2 of the book). Insert values and calculate the interval. Compute corresponding intervals for the three other houses and fill in the table below.

	Lower bound of CI	**Upper bound of CI**
House 1		
House 2		
House 3		
House 4		

Compare the CI for House 1 computed above with the result of the following R code:

```
## CI for the mean daily heat consumption of House 1
t.test(Dsel$Q1, conf.level=0.95)$conf.int
```

It was estimated that the average daily heat consumption of House 1 over a full year is approximately 2.38 kW/day.

h) Carry out a hypothesis test in order to assess whether the mean daily heat consumption of House 1 during January-February 2010 is significantly different from 2.38 kW/day. This can be done by testing the following hypothesis:

$$H_0 : \mu_{\text{House1}} = 2.38,$$
$$H_1 : \mu_{\text{House1}} \neq 2.38.$$

Specify the significance level α, the formula for the test statistic, as well as the distribution of the test statistic (remember to include the degrees of freedom). Insert relevant values and compute the test statistic and p-value. Write a conclusion in words. If there is a significant difference: Is the mean daily heat consumption in January-February more or less than 2.38 kW/day? Furthermore, comment on whether it was necessary to perform the hypothesis test or whether the same conclusion could have been reached using the confidence interval for House 1 computed above.

Compare the results of the test with the results of the following R code:

```
##  Testing hypothesis mu=2.38 for daily heat consumption
## (House 1, Jan-Feb 2010)
t.test(Dsel$Q1, mu=2.38)
```

We would also like to investigate whether the mean daily heat consumption in January-February 2010 differs between House 1 and House 2.

i) Carry out a hypothesis test in order to investigate whether the mean daily heat consumption during the first two months of 2010 differs between House 1 and 2. Specify the hypothesis as well as the significance level α, the formula for the test statistic, and the distribution of the test statistic (remember the degrees of freedom). Insert relevant values and compute the test statistic and p-value. Write a conclusion in words. Is the daily heat consumption of House 1 significantly different than that of House 2? If so, is it higher or lower?

Compare the results from the hypothesis test with the results of the following R code:

```r
## Comparing the heat consumption of House 1 and 2
t.test(Dsel$Q1, Dsel$Q2)
```

j) Comment on whether it was necessary to carry out the statistical test in the previous question, or if the same conclusion could have been drawn from the confidence intervals alone? (See Remark 3.59).

Correlation

In relation to the subsequent development of models for the description of the level of heat consumption, e.g. in Project 2, we will focus on the correlation between two variables of interest.

k) *In this question, data from the whole period of observation, October 2008 through June 2011, should be used (not only the data from January-February 2010).* State the formula for computing the correlation between the daily heat consumption of House 1 ($Q1$) and the global radiation (G). Insert values and determine the correlation (note, insert only numbers in the correlation formula, i.e. three numbers) . Make a scatter plot of one variable against the other. Assess whether the relation between the plot and the correlation is as you would expect.

Compare the correlation computed above to the result of the following R code:

```r
## Correlation between heat consumption and global radiation
cor(D[, c("Q1","G")], use="pairwise.complete.obs")
```

Remark 2.1 Extra R tips

This is an optional extra remark about different ways to take subsets in R (useful but not necessary for solving the project):

```r
## Optional extra remark about taking subsets in R
##
## A logical vector with TRUE or FALSE for each row in D, e.g.:
## Finding days with frost
D$Ta < 0
## Can be used to extract heat consumptions for House 1 on
## days with frost
D$Q1[D$Ta < 0]
## The 'subset' function can be used as well
subset(D, Ta < 0)
## More complex logical expressions can be made, e.g.:
## Observations from days with frost before 2010
subset(D, t < "2010-01-01" & Ta < 0)
```

Remark 2.2 Extra R tips

Optional remark with some extra R tips. The table can also be generated more effectively using a ′for′-loop:

```
## Use a 'for'-loop to calculate the summary statistics for each house
## and assign the result to a new data.frame
Tbl <- data.frame()
## Find the relevant columns
selected <- c("Q1","Q2","Q3","Q4")
## Calculate the summary statistics for each house
for(i in selected){
  ## Take the relevant column
  x <- Dsel[, i]
  ## Compute the sample mean
  Tbl[i, "mean"] <- mean(x, na.rm=TRUE)
  ## Compute the sample variance
  Tbl[i, "var"] <- var(x, na.rm=TRUE) }
## View the content of Tbl
Tbl

## In R there are even more condensed ways to do such
## calculations, e.g.:
apply(Dsel[, selected], 2, mean, na.rm=TRUE)
## or several calculations in one go
apply(Dsel[, selected], 2, function(x){
  c(mean=mean(x, na.rm=TRUE),
    var=var(x, na.rm=TRUE)) })
## See more useful functions with: ?apply, ?aggregate and ?lapply
## For extremely efficient data handling see, e.g., the packages:
## dplyr, tidyr, reshape2 and ggplot2

## LaTeX tips:
## The R package "xtable" can generate LaTeX tables written to a file
## and thereby they can automatically be included in a .tex document.
## The R package "knitr" can be used very elegantly to generate .tex
## documents with R code written directly in the document. This
## document and the book were generated using knitr.
```