Statistical Computing
Master Data Science
Winter Semester 2019/20

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 2 — Frequency tables and basic graphics

Introduction

The purpose of this short workshop is to put what was covered in the lecture into practice using R. It helps to learn R quickly if you use it every week. At the end there are a few simple exercises for you to attempt at home without using R.

Last Week you used R-Studio for the first time and learnt some basic principles of R. The most important aspects are:

`<-`	assigns the value (right hand side) to an object (left hand side)
`[...]`	square brackets are used to access an element or a range of elements in a vector object and
functions	are called using round brackets with arguments specified inside the brackets e.g. `log(3, base=10)`

Preliminaries

► Start R-Studio using the Icon on the desktop
► *Strg* + *Shift* + *N* opens a new R script.
► Type the comment

```
#Statistical Computing: Workshop 2
#Frequency tables and Introduction to Graphics
```

as a title to the script file.
► Save the file in your `H:\\StatComp` folder with the name `Workshop2.R` using *File > Save as..*
Throughout the workshop save your script file regularly using *Strg* + *s*. Sometimes the program crashes!
► Set your *working directory* to be `H:\\StatComp`. The code to do this is

```
> setwd(H://StatComp)
```

Work through this worksheet entering the commands and examples as you did last week. You should type in all your commands into the script file. In many of the code examples the output is not included, you should always see what the output is and make sure you understand it. Please add comments (in your own words) to your code so that it will make sense to you when you return several weeks (or years) later.

Please do not copy and paste the commands from this PDF file, type them in yourself. Copy/paste is the quickest way to forget what you have just done!

The Gotham City data set – Reading text data

In Lecture 2 the data for the students in a class at Gotham City University was presented. These data are in the file called `GCU.csv` (Moodle). Download the data into your `StatComp` directory and read in the data assigning it to an object called `Gotham`.

```
> Gotham<-read.csv(file="GCU.csv")
```

The function `read.csv()` was briefly mentioned in the *Further Reading* document in the section reading data from files.

Find out how many observations and how many variables there are in the data file.

```
> nrow(Gotham)
> ncol(Gotham)
> dim(Gotham)
```

Variable names and types

`Gotham` is stored in R as a data frame

```
class(Gotham)
```

You have already found out how many variables are in this data set. Use

```
names(Gotham)
```

to find the names of the variables in the data set.

Remember from last week there are several different ways to access a variable in a data frame, the easiest is to use the dollar sign

```
Gotham$Income
```

You can use class to find out the *variable type*

```
class(Gotham$NSiblings)
```

Find the data types for all the variables in the data set.

Frequencies for a nominal variable

Table of absolute frequencies

```
> table(Gotham$DegreeSubject)
```

Table of relative frequencies (proportions)

First a table is created and then the proportions are obtained

```
> prop.table(table(Gotham$DegreeSubject))
```

Seven decimal places is a bit overkill. Let's round the proportions to 3 decimal places and present them as percenatges

```
> round(prop.table(table(Gotham$DegreeSubject)),3)*100
```

We can present both the absolute and relative frequences in one table using the function `rbind()` ("row-bind").

```
> absolute<-table(Gotham$DegreeSubject)
> relative<-round(prop.table(table(Gotham$DegreeSubject)),3)*100
> rbind(absolute,relative)
```

What do you think the function `cbind()` does? Try it out.

Cumulative absolute frequencies

```
> cumsum(table(Gotham$DegreeSubject))
```

Notice that the cumulative frequencies do not make sense for this variable. Why not?

Graphics for frequency data

Presenting frequency data is usually done using a bar chart.

```
> barplot(table(Gotham$DegreeSubject))
```

Click on *zoom* to see all of the labelling. There are two purposes for obtaining graphics for our data

i) to investigate the data, "for your eyes only" or

ii) to communicate the properties of the data to others, "for the public's benefit"

For i) this plot is sufficient. For ii) we need to add arguments for the titles, colours or a legend. As an example

```
> barplot(table(Gotham$DegreeSubject),main="25 Gotham City Uni Students",
+ sub="Degree Subject",ylab="Frequency",
+ col=c("lightgreen","lightpink","lightblue","orange"))
```

Note that the '+' in the above code is a continuation prompt used because the command runs over multiple lines. You do not need to type the + in the same way that you do not need to type >. This notation is standard when presenting R code.

For continuous variables a bar chart is not a good idea. If you don't know why, try plotting a bar chart for the variable `Income`. Instead we use a histogram.

```
> hist(Gotham$Income,col="Grey")
```

You will learn the statistical details of a histogram in coming weeks, and a workshop will focus R graphics in detail.

In workshop exercise

(a) Obtain the four types frequency for the variable "number of siblings", and present them all in one table.

(b) Obtain a bar chart for the number of siblings.

Tidying up

Make sure your script file has sensible comments. At least one comment for each main section.

► Save the script file (source file) again: *Strg* + *s* or *File > Save as ....*

► Leave R-Studio by typing the command:

```
> q()
```

When R asks you *Save workspace image ...?*, click on **Don't save**!

► Feierabend!

Homework Exercises

Here are three easy exercises for you to work on with pen, papaer and maybe a calculator (i.e. not using R). These exercises are to reinforce the concepts covered in the lecture.

Aufgabe 1
Complete the gaps in the following text.

At the end of the 2017/18 school year, five-hundred students at a vocational college were asked about their decision to take a course at this college. There are in total 1251 students. The _____ consists of all 1251 students, whereas the students who took part in the survey is a _____. The _____ is $n = 500$.

Aufgabe 2 Frequency table

A "drop-in" tyre replacement garage asked 20 customers for feedback about their satisfaction with the work done. The responses are coded as s= satisfied, a= average, u=unsatisfied.

s, a, s, s, u, s, s, a, a, s, s, a, u, a, a, s, u, a, a, u

(a) Obtain the four types of frequencies covered in Lecture 2 and present them in a table.

(i) absolute frequency (h_i),

(ii) relative frequency (f_i),

(iii) absolute cumulative frequency (H_i) and

(iv) relative cumulative frequency (F_i).

(b) Plot the absolute frequencies in a bar chart.

Aufgabe 3

A comparison of PC-monitors considers the following variables:
Manufacturer, type of screen, diagonal dimension (inches), response time (miliseconds), connections (e.g. HDMI), number of USB ports, refresh rate (Hz), built in speakers (yes /no), features, colour, overall customer satisfaction (1 to 5 stars).

Assign the following *data types* to these variables: nominal, binary, ordinal, discrete or continuous.