ESMA 3101: Introduction to Statistics I

Dr. Wolfgang Rolke

August 12, 2018

Contents

1 Table of Contents

Resma3.RData (Ver 3.0)

All these web pages are also available as a single pdf here.

For a nice introduction to Statistics watch the PBS-NOVA episode Prediction by the Numbers

1.0.6 6 Statistical Inference

6.1. Population - Sample

6.2. Estimation - Confidence Interval

6.3. Hypothesis Testing

6.4. The Lady Tasting Tea

6.5. Inference for the Mean μ

6.6. Inference for a Proportion π

6.7. Bayesian Inference

6.8. Exercise Problems

1.0.7 7 Correlation and Regression

7.1. Correlation

7.2. Regression

7.3. Exercise Problems

1.1 Syllabus

Professor: Dr. Wolfgang Rolke

The web address is http://academic.uprm.edu/wrolke/esma3101.

The official prontuario for the course is available from the usual site or directly from here. If there is any difference between the prontuario and the information on the webpage use the webpage.

Time and Place:

Section 1: Tuesday, Thursday 7:30 - 8:45am SH005
Section 2: Tuesday, Thursday 9:00- 10:15am SH005

Textbook: Statistics, Informed Decisions using Data, Michael Sullivan (**highly recommended but not required**)

Office hours:
Tuesday, Thursday 12:00-12:30pm OF407
Tuesday, Thursday 3:15-5:15pm OF407 (by appointment)
Wednesday 1:30-3:00pm via email

email: wolfgang[dot]rolke[at]upr[dot]edu

when you send me an email ALWAYS start the subject line with ESMA3101

Grading:

1. Quizzes: 35%
2. Partial Exams 35%
3. Final 30%

All quizzes and exams will be done using moodle. To get to the quizzes go to https://ecourses.uprm.edu/, log on with your UPR ID and password.

The first time use the enrollment key:

- 7:30am Section: Esma 3101 - 017
- 9:00am Section: Esma 3101 - 026

1.2 Warning

If you want to pass, here is what you should expect to do:

- **Take it serious!**
- Come to class, (almost) always
- Pay attention while in class. If you don't understand something **ASK**
- After each class, spend **at least one hour** to go over what was discussed. If there is anything you don't understand, **ASK** about it in the next class. You can not expect to understand this material just by sitting in class and listening, you have to work through it again on your own later.
- There will be a lot of material to memorize. Without doing so you have no chance to pass this course.

2 Introduction

This page discusses some general concepts of Statistics.

2.1 Motivation for learning Statistics

Quote:

If I had only one hour to live, I would choose to live it in statistics class because it would seem to last forever

(A student's complaint)

Why does Statistics appear to be so boring?

Consider the following questions:

- Does aspirin lower the risk of heart attacks? (Medical research)
- Are there fewer Manatees in Puerto Rico today than 10 years ago? (Biology)
- Does a certain brand of gasoline really clean the engine? (Chemistry and Mechanical Engineering)
- Do tax cuts work? (Economics and Government Policy)
- Does anger management training work? (Psychology and Law)
- Do frequent flier programs increase airline ticket sales? (Business)
- Do the salaries of men and women differ? (Social Sciences and Law) Amazingly enough, a person investigating any of these questions might well end up using the same statistical method to answer them! (It is called the one-sample t test and we will study it at some time during this semester) The power, strength (and beauty) of statistics lies in its universal applicability!

2.2 What is Statistics?

Answer 1: Statistics is the Science of **data** (or information)

- How to collect data
- How to analyze data
- How to present data

Answer 2: Statistics is the Science of **Uncertainty**

- where does it come from
- what types are there
- how to deal with it

2.3 Why everybody should know a little bit about Statistics - Misuse of Statistics

Statistics can be used in many ways to make things appear to be something that it is not (lying!)

Another quote:

There are Lies, Damn Lies and Statistics

(maybe Benjamin Disraeli, probably not)

2.4 What can we do with Statistics?

2.4.1 Case Study: WRInc

WR Inc. is a large (ficticous) company. It recently did a survey of all its employees, asking them to fill out a questionnaire with questions regarding their gender, income etc.
In addition they randomly selected 500 employees and asked them some additional questions.

Let's start by checking what's in the data set:

```
head(wrinccensus)
```

```
##   Id.Number Gender Income Job.Level Years Satisfaction
## 1     10001 Female  22800         1     6            4
## 2     10003 Female  18600         1     1            2
## 3     10004 Female  23900         1     4            2
## 4     10008   Male  37200         1    13            4
## 5     10010   Male  29800         1     9            1
## 6     10014   Male  53700         5    16            1
```

shows us the names of the variables and the first six rows of data.

Note: when something appears in a box like this it means you can type (or copy-paste) this into R and get the same answer.

```
dim(wrinccensus)
```

```
## [1] 23791     6
```

shows us that there are 23791 **observations**, one for each employee, and with 6 pieces of information (variables) for a total of 23791*6 = 142746.

Trying to look at so much information is very difficult, so organizing it in some fashion is very useful.

Often just making a little table is a good idea:

```
attach(wrinccensus)
table(Gender)
```

```
## Gender
## Female   Male
##   9510  14281
```

Sometimes it is better to consider percentages:

```
length(Gender)
```

```
## [1] 23791
```

```
table(Gender)/length(Gender)*100
```

```
## Gender
##  Female    Male
## 39.9731 60.0269
```

```
round(table(Gender)/length(Gender)*100, 2)
```

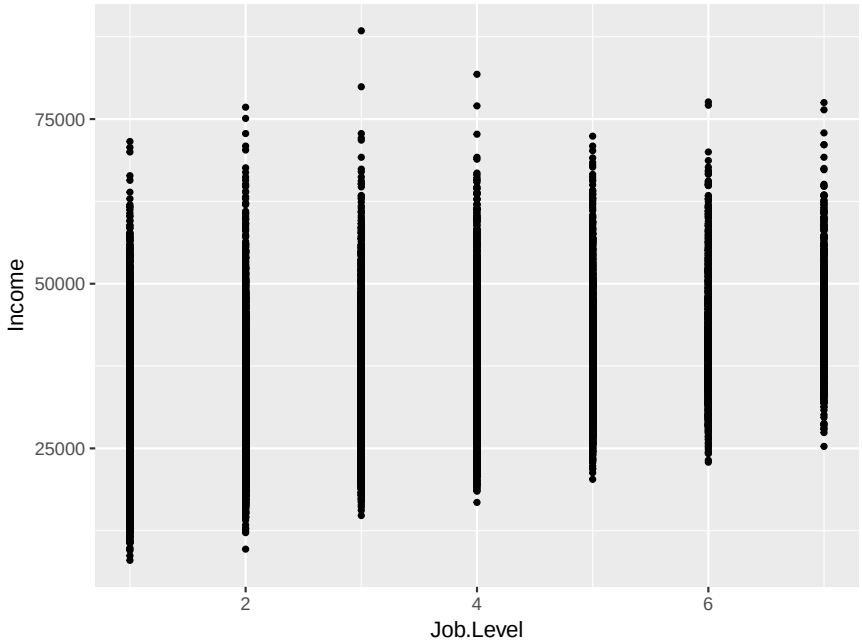
```
## Gender
## Female   Male
##  39.97  60.03
```

or maybe even both:

Gender	N	Percentage
Female	9510	39.97%
Male	14281	60.03%

Another good way to study a dataset is via **graphs**. For example, it seems reasonable that there should be a connection between job level and income, after all usually people with a better job make more money. Is this true for our company? For this we can use

```
splot(Income, Job.Level)
```



As we will see soon, often we have different ways to do the same thing in Statistics. For example, instead of the scatterplot above we could draw something called a boxplot, which we will discuss at some point in this class. And in fact, this would be a very good idea because there is a serious problem with the scatterplot. Can you see what it is?

An important question would be whether there is **job discrimination** in this company, that is whether men are paid more than women. How can we find out? Let's compute the average income of the men and the average income of the women. But before we do we need to understand that

- the two will **not be the exactly the same**!
- so one will be higher than the other, just by random chance.

In fact even if there is no job discrimination there is a 50-50 chance that the average income of the men is (a little bit) higher than the average income of the women. Of course if there is job discrimination we would expect the average income of the men to be substantially higher than the average income of the women. What we need to find out is whether the men's income is statistically significantly higher!

Something is statistically significant if it cannot be expained by random chance alone.

Example 4 heads in 4 flips of a fair coin has a probability of 1 in 16 or 6.25%, so this would not be considered unusual

Example 10 heads in 10 flips of a fair coin has a probability of 1 in 1028 or 0.01%, so this would be considered very unusual. In fact one would now conclude that this coin is not a fair coin.

Note What is and what is not statistically significant is a question of probability.

back to WRInc:

```
stat.table(Income , Gender)
```

```
##        Sample Size    Mean Standard Deviation
## Female        9510 33150.9              9373.1
## Male         14281 33521.1              9455.8
```

We find that the average income is

Female: $33150.9
Male: $33521.1

so the difference is $33521.1 - 33150.9 = \$370.2$.

- Is this a "substantial" or a "little" difference?
- Is this a "statistically significant" difference?
- Does it "prove" discrimination?

"prove" here has just about the same meaning as it does in a criminal trial: beyond any reasonable doubt.

To answer the question we would need to do a hypothesis test.

One way to answer that question is to use a method discussesd in ESMA 3102 called the two-sample t test. You might be surprised to learn that indeed the difference is to large to be due to random chance!

Careful, though: although the difference is statistically significant, it still does not mean that there is discrimination, because the difference in salaries might be caused by other factors such as the fact that there are more men at the higher job levels, that men tend to have more years at the company etc. This is an example of one of the major issues in Statistics:

Correlation does not imply Causation or here: there is some connection between gender and income (specifically men are paid more than women) but it is not clear yet why that is. One

possible reason is discrimination, but there could be others.

2.5 Some Basic Terminology of Statistics

Population : all of the entities (people, events, things etc.) that are the focus of a study

Example 1 Say we are interested in the average age of the undergraduate students at the Colegio.

Example 2 A company is considering to sell a new product in Puerto Rico, but before they do they want to know how many people in Puerto Rico might be interested in buying it.

Example 3 All possible hurricanes, past and **future**. Clearly this last one is much more complicated population than the undergraduate students at the Colegio. In order to properly describe it we will need **probability**.

Census : If all the entities of a population are included in the study.

Example 1 if we ask the Registrars Office they might give as the ages of all these students, and if we then ask all the students how old they are we we would have done a census.

Example 2 impossible for practical reasons (we cannot ask every person in Puerto Rico)

Example 3 impossible for theoretical reasons (future?) **Sample** : any subset of the population

Example 1 let's take the students in the room as a sample.

Example 2 ask all our friends and relatives

Example 3 all the hurricanes during the last 10 years. **Random sample** : a sample found through some randomization (flip of a coin, random numbers on computer etc.)

Example 1: Are you a random sample?

Example 2 do a telephone survey

Example 3 yes

Simple Random Sample (SRS) : each "entity" in the population has an equal chance of being chosen for the sample.

Example 1 Are you a simple random sample?

Example 2 depends on exactly how the telephone numbers are chose, but generally ok

Example 3 all the hurricanes during the last 10 years: yes

Data : the collection of many pieces of information

Example 1 a table with the ages of the students in our sample

Example 2 list of answers (yes - I would buy the product, no - I would not)

Example 3 all the data available about a hurricane: track, windspeed, air pressure etc.

Parameter : any numerical quantity associated with a population

Example 1 If we had the ages of **all** 10000 or so undergraduate students we could calculate the average, and it would be a parameter

Example 2 the percentage of all the people in PR who would buy the product - impossible to find exactly because we cannot do a census.

Example 3 the average top windspeed of the strongest hurricane in any one year. This is a number that nobody knows or can know, even theoretically.

Example 4 The mean income of the employees of WRInc. Because we have the income data for **all** the employees this is a parameter.

Statistic : any numerical quantity associated with a sample

Example 1 let's calculate **your** average age. You are a sample, so this is a statistic.

Example 2 the percentage of people in our sample who say they would buy the product

Example 3 Take the last 10 years as a sample, and calculate the average of the top windspeeds of the strongest hurricane in each year.

Example 4 The mean IQ of the employees of WRInc. Because we have the IQ data for only 500 of the employees this is a statistic. Note there is **one** value of the population parameter (at a fixed moment in time) but there are many different values of the statistic, depending on the sample that was selected.

Statistical Error the uncertainty in the value of the statistic due to the fact we only used a sample.

Example 1 If the average age of the students in the classroom is 21.25, what does this tell us about the average age of **all** the undergraduate students? Is it possible that this might be much higher, maybe even over 22 years?

Example 2 If 30% of the people in the sample say they would buy the product, what might be the number for the whole island?

Example 3 If the strongest hurricane in the last 10 years had indspeeds up to 130 miles per hour, does this mean we will not have one with 150 miles per hour this year?

Bias Any systematic difference between the population and the sample with respect to a variable.

Example 1 Are you (the class) a biased sample?

Example 2 depends on how the selection of telephone numbers is done.

Example 3 Are the last 10 years a biased sample? Avoiding bias is the main reason for using a Simple Random Sample.

What is better?

Large Variance - Small Bias

or

Small Variance - Large Bias

2.5.0.1 App

for an illustration of the bias vs variance issue run

```
run.app(bias-variance)
```

[Australia] ("http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language")

2.6 Categorical vs. Quantitative Variables

Example Let's consider again **WRInc**

a) Variable Gender: most obvious thing to do: count how many males and females are in the company

b) Variable Income: find average income of employees.

Why the difference?

In real live we always use a computer for all the calculations. That leaves two tasks for the human being doing Statistics:

Decide

- what is the best method for analysing a specific dataset?
- what is the result of the analysis telling you about the experiment?

Most important here are

- **the computer will not do these steps for you**
- (Almost always) the computer will do the analysis you ask it to do, even if this analysis is **complete nonsense**

In order to know what method to use it is important to understand some basic features of your data. One is its **data type**:

We categorize variables as follows:

A. **Quantitative** data is numeric, and arithmetic makes sense (adding, multiplying etc.)

Example

1) Yearly income of a family in Puerto Rico
2) Temperature in Mayaguez at 12 Noon
3) Amount paid for the phone bill

B. **Categorical**

everything else

Example

1) A students major
2) in an experiment to grow wheat three different fertilizers were labeled 1,2 and 3

3) Your student id number

Note Often whether a variable is categorical or quantitative depends on how (and how precisely) it is measured.

Example Our variable is "amount of rain fall"

- Is it raining at all? "Yes" or "No" → categorical
- We put a cup outside. The cup has marks for each cubic inch of rain. Our data is the number of cubic inches. Values will be 0, 1, maybe 2. → quantitative

Categorical data comes in one of two versions - **ordered or unordered**:

Examples

1. grades in a course: A, B, C, D, W - ordered
2. gender: Male, Female - unordered
3. Treatments in a clinical trial: A, B, C - unordered
4. Treatments in a clinical trial: 1, 2, 3 - unordered
5. blood pressure: low medium high - ordered
6. directions: north east south west - unordered

One consequence of having an ordering is that it should be used in graphs, tables etc.

2.6.1 Case Study: WRInc

Let's look at the variables in the survey of **wrinccensus**

```
head(wrinccensus)
```

```
##   Id.Number Gender Income Job.Level Years Satisfaction
## 1     10001 Female  22800         1     6            4
## 2     10003 Female  18600         1     1            2
## 3     10004 Female  23900         1     4            2
## 4     10008   Male  37200         1    13            4
## 5     10010   Male  29800         1     9            1
## 6     10014   Male  53700         5    16            1
```

3 Data Collection - Design of Experiments

Things to decide before collecting data:

- define goal of study
- define population

- define sampling frame
- define variables
- define data collection methods
- define measuring methods
- define analysis methods
  Collection Methods: scientific experiment, survey, ..

Example Say we want to do a survey to see whether people prefer Coke over Pepsi - how do we do it?

Example risk of smoking

Methodology

- experiment on lab animals (rats, mice, monkeys,..)
- survey of smokers / nonsmokers
  Variables
- what measures effect of smoking?
- lung cancer deaths
- other diseases
- bad teeth

3.0.1 Population - Sampling Frame

A sampling frame is a "listing" of all the elements of a population **Example** 1948 presidential elections Harry S. Truman vs. Thomas E. Dewey

Example Phone surveys (random dialing, bias, recalls etc.)

3.1 Sampling Methods

We already mentioned the Simple Random Sample, the most basic sampling method. There are others, though:

Stratified Sampling: First divide population into subgroups, then do a SRS in each subgroup.

Example 1 Gender (Male-Female), Year (Freshman - Sophomore - Junior - Senior), Departments (English - Math - ..)

Example 2 in the telephone survey use a list of prefixes (832,. . .) to get people from different parts of the Island

Example 3 hurricanes by category 1-5

Systematic Sampling: choose sample according to some deterministic rule.

Example 1 pick randomly a number between 100 and 999, then pick every student whose student id ends with those three digits

Example 2 pick randomly a number between 0 and 99, call all phone numbers ending in those digits.

Example 3 pick randomly a number n between 1 and 5, pick every n^{th} hurricane.

Cluster Sampling: First divide population into subgroups, first do an SRS of these subgroups, then do a SRS in each of those subgroups.

Example 1 randomly pick 10 buildings on campus, then do SRS in each building.

Example 2 randomly pick 10 municipalities, then do SRS in each municipality.

Example 3 divide by week, randomly select 10 weeks, do SRS for hurricanes that formed that week.

3.1.0.1 App

for an illustration of several sampling strategies run the **sampling** app

3.2 Designed Experiments

There are two basic methods for designing experiments with people.

An observational study observes individuals and measures variables of interest but does not attempt to influence the responses

A designed experiment deliberately imposes some treatment on individuals in order to observe their responses.

Example say we want to test two different diets for their effectiveness in losing weight. We advertise in the newspaper and on a certain day people who want to participate come to our office.

There

A) we have two tables. On each there is an explanation of one of the diets and a sign-up sheet. Our participants can read the material and eventually sign up for one of the two diets → observational study.

B) when a participant comes in we have him/her flip a coin. If it comes up headsthe person does diet 1, otherwise diet 2 → designed experiment.

Which of the two methods is better, and why?

The individuals on which an experiment is done are called experimental units. When the units are humans they are called subjects. A specific experimental condition applied to a unit is called a treatment.

Example let's consider again the diet experiment above, specifically A). Let's say that after two month we find that those subjects on diet 1 have lost on average 2.3 pounds more than those on diet 2, and that such a difference of 2.3 pounds is statistically significant. Can we conclude that diet 1 is better than diet 2? The problem is this: it was the the subjects themselves who picked their diet (we say they were self-selected). But why did subject #237 pick diet 1 and not 2? If she did it essentially randomly, there is no problem. But maybe there was a tendency for heavier subjects to pick diet 1 (a selection bias) and of course heavier people generally lose more weight, and that is why there is a difference between the groups. There is another variable in play, let's call it diet preference, and maybe it is this variable that is the reason for the difference in weight loss.

This type of problem is called confounding.

The big advantage of doing a designed experiment is that such a confounding variable is certain to not exist.

3.2.1 Case Study: Physicians Health Study

Does Aspirin lower the risk of heart disease? By 1980 there was a lot of anectotal evidence to suggest this was true, so in 1982 a large scale clinical trial was conducted.

anectotal evidence - non- scientific evidence **clinical trial** - designed experiment in medicine

A randomized, double-blind, placebo-controlled clinical trial designed to test the effects of low-dose aspirin and beta-carotene in the primary prevention of cardio-vascular disease and cancer among 22,071 US male physicians, aged 40 to 84 at baseline in 1982. Baseline blood specimens were collected and frozen for later analyses from 14,916 participants. The trial was started in 1982 and was designed to run until 1995.

The trial used a 2x2 factorial design: 325 mg of aspirin (Bufferin, supplied by Bristol-Myers Products on alternate days)

50 mg of beta-carotene (Lurotin, supplied by BASF AG on alternate days)

placebo - something exactly like the treatment.

randomized - each subject was assigned randomly to one of four groups (Aspirin-Beta-carotine, Aspirin-Placebo, Beta-Carotine-Placebo, Placebo-Placebo).

double-blind - neither the subjects knew which group they were in (blind) nor the medical personnel working with them (double-blind).

2x2 factorial design - there were two factors (aspirin and beta-carotine) each with two levels (yes-no) and all 2x2=4 possible combinations were included in trial.

Why were the subjects chosen to be doctors?

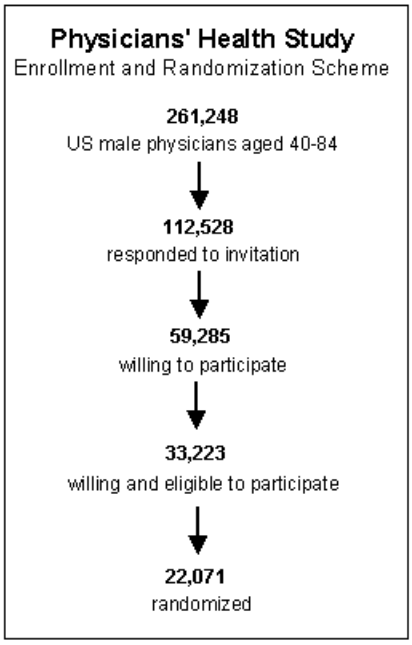


Figure 1:

Figure 2:

- Ability to give true informed consent
- Knowledge of possible side effects
- Accuracy and completeness of information
- Possibly higher rate of compliance

Primary endpoints (outcomes of interest)

- Cardiovascular disease
- Total cancer
- Prostate cancer
- Eye disease

By 1988 of the 11037 doctors who received aspirin 139 had developed heart disease, but of the 11034 doctors in the control group 239 had done so. This was statistically significant evidence that aspirin works to prevent heart disease and this part of the trial was stopped. The beta-carotine part continued until 1995 and did not show any effect.

For more on this study see http://phs.bwh.harvard.edu/phs1.htm

output: html_document: default pdf_document: fig_caption: no —

4 Using the Computer and R

This page contains some basic information on how to use the computer and the R program.

To log on to computers in Ch115:

Username: .\esma (important: do not forget to include ". " before the word esma)
Password: Mate1234 (important: uppercase letter"M")

To log on to computers in SH005:

Username: Estudiante
Password: salon005

The class webpages are at http://academic.uprm.edu/wrolke/esmaXXXX (3015, 3101, 3102, 6661 etc)

At the end of each session log off

4.0.1 General Info

You can get a free version of R for your computer from a number of sources. The download is about 70MB and setup is fully automatic. Here are some links:

Windows

MacOS

After the installation is finished close R (if it is open). From now on ALWAYS open R by clicking on the link to to the RESMA3 file on top of the homepage. You can also download and save that file to your own computer and start R from there. The first time you do this the program will download a number of additional stuff, just let it. Also a window might pop up and ask whether to save something, if so click on yes.

Note

- You might be asked at several times whether you want to do something (allow access, run a program, save a library, . . .), always just say yes!
- You will need to connect to a reasonably fast internet for these steps.
- This will take a few minutes, just wait until the > sign appears.

FOR MAC OS USERS ONLY

There are a few things that are different from MacOS and Windows. Here is one thing you should do:

Download XQuartz - XQuartz-2.7.11.dmg
Open XQuartz
Type the letter R (to make XQuartz run R)
Hit enter Open R Run the command .First()
Then, every command should work correctly.

4.0.2 RStudio

there is a program called RStudio that a lot of people like to use to run R. You can download it at RStudio. Before you can use RStudio with Resma3 you need to run Resma3 JUST ONCE from R itself.

So do this

1) follow **ALL** the instructions above
2) only if everything is running correctly install RStudio.

For the purpose of the class R itself is enough, we don't need RStudio.

4.0.3 Troubleshooting

if you try to run a command and get an error

could not find function "ggplot"

(or grid or shiny)

first try this: run the command

```
ls()
```

You should see a listing of many things (over 200). If you do not Resma3 did not load correctly. Close R and restart it by clicking on the link to Resma3 on the homepage.

If you do see the listing, type

```
one.time.setup()
```

A number of things should be happening, just wait until you see the > again and see whether that fixes the problem.

If this does not work turn off R and restart it with a new version of Resma3 from the top of the class homepage.

If this also does not work send me an email with the explanation of the problem. The best thing to do is to include a screenshot. Here is how:

Windows

MacOS

You can also just use your cell phone to take a picture of the screen, but make sure it is is readable!

I often get an email saying that something is not working, and my answer is simply:

RGDM

this means: **R**ead the **G**od-**D**amn **M**anual!

that is the answer to your problem is somewhere on these pages, and you should have found it there before sending an email!

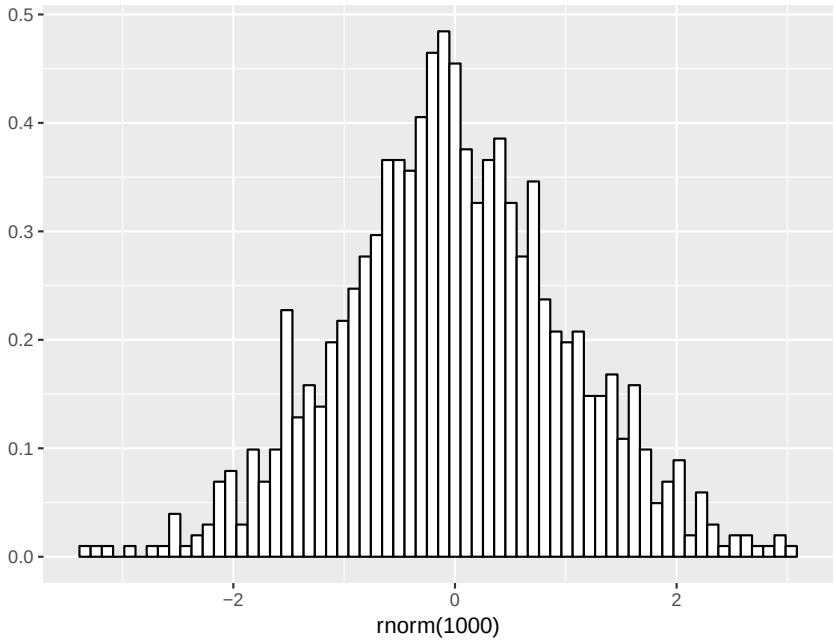
Throughout this class when you see something like this:

```
text
```

it means commands you should type (or copy-paste) into R.

To see whether everything is installed correctly copy-paste the following line into R and hit enter:

```
hplot(rnorm(1000))
```



You should see a graph like this (called a histogram)

For a much more extensive introduction to R go here

Once you have started a session the first thing you see is some text, and then the > sign. This is the **R prompt**, it means R is waiting for you to do something. Sometimes the prompt changes to a different symbol, as we will see.

Let's start with

```
ls()
```

shows you a "listing"" of the files (data, routines etc.)

If you have worked for a while you might have things you need to save, do that by clicking on

File > Save Workspace

If you quit the program without saving your stuff everything you did will be lost. R has a somewhat unusual file system, everything belonging to the same project (data, routines, graphs etc.) are stored in just one file, with the extension .RData.

To quit R, type

```
q()
```

or click the x in the upper right corner.

R has a nice recall feature, using the up and down arrow keys. Also, typing

```
history()
```

shows you the most recent things entered.

R is case-sensitive, so a and A are two different things.

Often during a session you create objects that you need only for a short time. When you no longer need them use **rm** to get rid of them:

```
x <- 10
x^2
```

```
## [1] 100
```

```
rm(x)
```

the **<-** is the *assignment* character in R, it assigns what is on the right to the symbol on the left.

4.1 Data Entry

4.1.1 With the keyboard

For a few numbers the easiest thing is to just type them in:

```
x <-  c(10, 2, 6, 9)
x
```

```
## [1] 10  2  6  9
```

c() is a function that takes the objects inside the () and **c**ombines them into one single object (a vector).

4.1.2 idataio

This section can be left out unless data i/o is discussed in class

idataio won't run in CH115 until computers have been updated

We have data on the age and the position of people. So there were 10 old people in the first position, and so on:

Age	First	Second	Third
Old	10	16	21
Young	15	12	26

To get this into R use the the routine **idataio**.

CAREFUL: **idataio** currently does not work in CH115 because of old version of R!

It can be used to enter the values directly from the keyboard, a table that was copied to the clipboard or read it from a file like an excel worksheet.

Say we want to get the table above into R. Here are three ways to do this using idataio:

```
x <- idataio()
```

this will bring up the browser with a spreadsheet and you can just enter the values. Change Number of Cases to 2 and Number of Variables to 4. Type the column names (Age First Second Third) in the box on the right and enter the values in the spreadsheet. Click on the button Close App to return to R.

2) use the mouse to highlight the whole table, switch to R and run

```
x <- idataio()
```

select the Copy from Clipboard option. Change Number of Variables to 4. Highlight the table in the browser and right-click Copy. Hit Go! and see whether the table appears correctly. If not maybe you need to play around a bit with the Number ofr cases etc. When it is ok hit the Close App button on top.

copying from an Excel worksheet works exactly the same way.

NOTE : the current version does not allow for empty cells. If there are any enter NA first. Also any names can not include spaces.

3) Open Microsoft Excel and enter the info as usual. Save the file as an excel spreadsheet (with the xlsx extension). Now run idataio and choose the Read data from file option.

4.1.3 Getting Data from Moodle Quizzes

Most moodle quizzes will require you to transfer data from the quiz to R. This is done with the command *get.moodle.data()*. There are two steps:

- in moodle use the mouse to highlight the data. If it is a table with several columns ALWAYS include the column headers (names of variables).
- switch to R and run

```
get.moodle.data()
```

Now the data should be in R. It is called x. You can always check by typing x and ENTER.

```
x
```

```
## [1] 10  2  6  9
```

Here are some examples:

a) single set of numbers:

101.6 115.0 100.9 103.8 77.6 102.6 99.6 108.5 100.8 92.5 101.8 81.6 103.7 94.9 103.3 86.7 101.6 106.6 101.5 96.9

highlight the data with the mouse, copy it, go to R and type

```
get.moodle.data()
x
```

```
##  [1] 101.6 115.0 100.9 103.8  77.6 102.6  99.6 108.5 100.8  92.5 101.8
```

```
## [12]  81.6 103.7  94.9 103.3  86.7 101.6 106.6 101.5  96.9
```

this also works if the data is not numbers:

Old Old Young Old Young Young

```
get.moodle.data()
```

```
## [1] "Old"   "Old"   "Young" "Old"   "Young" "Young"
```

sometimes parts of the data are separated by some symbol, for example a comma. In that case you can use the *sep* argument:

1.5, 2.3, 5.3, 2.4, 7.9, 8.1, 2.7, 4.2

```
get.moodle.data(sep = ",")
```

```
## [1] 1.5 2.3 5.3 2.4 7.9 8.1 2.7 4.2
```

b) data is in the form of a table with several columns:

Age	Gender
23	Male
23	Male
20	Female
24	Female
25	Male
21	Male
20	Male
22	Female
21	Female
25	Male

```
get.moodle.data()
```

```
##    Age Gender
## 1   23   Male
## 2   23   Male
## 3   20 Female
## 4   24 Female
## 5   25   Male
## 6   21   Male
## 7   20   Male
## 8   22 Female
## 9   21 Female
## 10  25   Male
```

Note if the data is a single vector it is given the name x, and you can now do things like

```
mean(x)
```

if the data is a table it is immediately attached and you can use the column names, for example

```
mean(Age)
```

Note on rare occasions the routine can fail if the data is a table but everyting is text. In that case use the argument is.table=TRUE.

Note sometimes you might get a warning from R, as long as the data is transfered correctly you can ignore that.

4.1.4 Data Types in R

the most basic type of data in R is a **vector**, simply a list of values.

Say we want the numbers 1.5, 3.6, 5.1 and 4.0 in an R vector called x, then we can type

```
x <- c(1.5, 3.6, 5.1, 4.0)
x
```

```
## [1] 1.5 3.6 5.1 4.0
```

Often the numbers have a structure one can make use of:

```
1:10
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
10:1
```

```
##  [1] 10  9  8  7  6  5  4  3  2  1
```

```
1:20*2
```

```
##  [1]  2  4  6  8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
```

```
c(1:10, 1:10*2)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10  2  4  6  8 10 12 14 16 18 20
```

Sometimes you need parentheses:

```
n <- 10
1:n-1
```

```
##  [1] 0 1 2 3 4 5 6 7 8 9
```

```
1:(n-1)
```

```
## [1] 1 2 3 4 5 6 7 8 9
```

The *rep* ("repeat") command is very useful:

```
rep(1, 10)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1
```

```
rep(1:3, 10)
```

```
##  [1] 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
```

```
rep(1:3, each=3)
```

```
## [1] 1 1 1 2 2 2 3 3 3
```

```
rep(c("A", "B", "C"), c(4,7,3))
```

```
##  [1] "A" "A" "A" "A" "B" "B" "B" "B" "B" "B" "B" "C" "C" "C"
```

what does this do?

```
rep(1:10, 1:10)
```

4.1.4.1 Commands for Vectors

To find out how many elements a vector has use the *length* command:

```
x <- c(1.4, 5.1, 2.0, 6.8, 3.5, 2.1, 5.6, 3.3, 6.9, 1.1)
length(x)
```

```
## [1] 10
```

The elements of a vector are accessed with the bracket [ ] notation:

```
x[3]
```

```
## [1] 2
```

```
x[1:3]
```

```
## [1] 1.4 5.1 2.0
```

```
x[c(1, 3, 8)]
```

```
## [1] 1.4 2.0 3.3
```

```
x[-3]
```

```
## [1] 1.4 5.1 6.8 3.5 2.1 5.6 3.3 6.9 1.1
```

```
x[-c(1, 2, 5)]
```

```
## [1] 2.0 6.8 2.1 5.6 3.3 6.9 1.1
```

Instead of numbers a vector can also consist of characters (letters, numbers, symbols etc.) These are identified by quotes:

```
c("A", "B", 7, "%")
```

```
## [1] "A" "B" "7" "%"
```

A vector is either numeric or character, but never both (see how the 7 was changed to "7").

You can turn one into the other (if possible) as follows:

```
x <- 1:10
x
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
as.character(x)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
```

```
x <- c("1", "5", "10", "-3")
x
```

```
## [1] "1"  "5"  "10" "-3"
```

```
as.numeric(x)
```

```
## [1]  1  5 10 -3
```

A third type of data is logical, with values either TRUE or FALSE.

```
x <- 1:10
x
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
x > 4
```

```
##  [1] FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

these are often used as conditions:

```
x[x>4]
```

```
## [1]  5  6  7  8  9 10
```

This, as we will see shortly, is EXTREMELY useful!

4.1.5 Data Frames

data frames are the basic format for data in R. They are essentially vectors put together as columns.

The main thing you need to know about working with data frames are the following commands:

4.1.5.1 Case Study: UPR Admissions

consider the **upr** data set . This is the application data for all the students who applied and were accepted to UPR-Mayaguez between 2003 and 2013.

```
dim(upr)
```

```
## [1] 23666    16
```

tells us that there were 23666 applications and that for each student there are 16 pieces of information.

```
colnames(upr)
```

```
##  [1] "ID.Code"        "Year"           "Gender"         "Program.Code"
##  [5] "Highschool.GPA" "Aptitud.Verbal" "Aptitud.Matem"  "Aprov.Ingles"
##  [9] "Aprov.Matem"    "Aprov.Espanol"  "IGS"            "Freshmen.GPA"
## [13] "Graduated"      "Year.Grad."     "Grad..GPA"      "Class.Facultad"
```

shows us the variables

```
head(upr, 3)
```

```
##      ID.Code Year Gender Program.Code Highschool.GPA Aptitud.Verbal
## 1 00C2B4EF77 2005      M          502           3.97            647
## 2 00D66CF1BF 2003      M          502           3.80            597
## 3 00AB6118EB 2004      M         1203           4.00            567
##   Aptitud.Matem Aprov.Ingles Aprov.Matem Aprov.Espanol IGS Freshmen.GPA
## 1           621          626         672           551 342         3.67
## 2           726          618         718           575 343         2.75
## 3           691          424         616           609 342         3.62
##   Graduated Year.Grad. Grad..GPA Class.Facultad
## 1        Si       2012      3.33           INGE
## 2        No         NA        NA           INGE
## 3        No         NA        NA       CIENCIAS
```

shows us the first three cases.

Let's say we want to find the number of males and females. We can use the table command for that:

```
table(Gender)
```

```
## Error: object 'Gender' not found
```

What happened? Right now R does not know what Gender is because it is "hidden" inside the upr data set. We need to make it visible to R first:

```
attach(upr)
table(Gender)
```

```
## Gender
##     F     M
## 11487 12179
```

there is also a detach command to undo an attach, but this is not usually needed because the attach goes away when you close R.

Note: you need to attach a data frame only once in each session working with R.

Note: Say you are working first with a data set "students 2016" which has a column called Gender, and you attached it. Later (but in the same R session) you start working with a data set "students 2017" which also has a column called Gender, and you are attaching this one as well. If you use Gender now it will be from "students 2017".

4.2 Subsetting of Data Frames

Consider the following data frame (not a real data set):

```
students
```

```
##    Age GPA Gender
## 1   22 3.1   Male
## 2   23 3.2   Male
## 3   20 2.1   Male
## 4   22 2.1   Male
## 5   21 2.3 Female
## 6   21 2.9   Male
## 7   18 2.3 Female
## 8   22 3.9   Male
## 9   21 2.6 Female
## 10  18 3.2 Female
```

Here each single piece of data is identified by its row number and its column number. So for example in row 2, column 2 we have "3.2", in row 6, column 3 we have "Male".

As with the vectors before we can use the [ ] notation to access pieces of a data frame, but now we need to give it both the row and the column number, separated by a ,:

```
students[6, 3]
```

```
## [1] "Male"
```

As before we can pick more than one piece:

```
students[1:5, 3]
```

```
## [1] "Male"   "Male"   "Male"   "Male"   "Female"
```

```
students[1:5, 1:2]
```

```
##   Age GPA
## 1  22 3.1
## 2  23 3.2
## 3  20 2.1
## 4  22 2.1
## 5  21 2.3
```

```
students[-c(1:5), 3]
```

```
## [1] "Male"   "Female" "Male"   "Female" "Female"
```

```
students[1, ]
```

```
##   Age GPA Gender
## 1  22 3.1   Male
```

```
students[, 2]
```

```
##  [1] 3.1 3.2 2.1 2.1 2.3 2.9 2.3 3.9 2.6 3.2
```

```
students[, -3]
```

```
##    Age GPA
## 1   22 3.1
## 2   23 3.2
## 3   20 2.1
## 4   22 2.1
## 5   21 2.3
## 6   21 2.9
## 7   18 2.3
## 8   22 3.9
## 9   21 2.6
## 10  18 3.2
```

4.3 Vector Arithmetic

R allows us to apply any mathematical functions to a whole vector:

```
x <- 1:10
2*x
```

```
##  [1]  2  4  6  8 10 12 14 16 18 20
```

```
x^2
```

```
##  [1]   1   4   9  16  25  36  49  64  81 100
```

```
log(x)
```

```
##  [1] 0.0000000 0.6931472 1.0986123 1.3862944 1.6094379 1.7917595 1.9459101
##  [8] 2.0794415 2.1972246 2.3025851
```

```
sum(x)
```

```
## [1] 55
```

```
y <- 21:30
```

```
x+y
```

```
##  [1] 22 24 26 28 30 32 34 36 38 40
```

```
x^2+y^2
```

```
##  [1]  442  488  538  592  650  712  778  848  922 1000
```

```
mean(x+y)
```

```
## [1] 31
```

Let's try something strange:

```r
c(1, 2, 3) + c(1, 2, 3, 4)
```

```
## [1] 2 4 6 5
```

so R notices that we are trying to add a vector of length 3 to a vector of length 4. This should not work, but it actually does!

When it runs out of values in the first vector, R simply starts all over again.

In general this is more likely a mistake by you, check that this is what you really wanted to do!

4.4 Subsetting

One of the most common tasks in Statistic is to select a part of a data set for further analysis. There is even a name for this: **data wrangling**.

4.4.1 Case Study: New York Air Quality Measurements

Description: Daily measurements of air quality in New York, May to September 1973.

A data frame with 154 observations on 6 variables.

Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island

Solar.R: Solar radiation in Langleys in the frequency band 4000-7700 Angstroms from 0800 to 1200 hours at Central Park

Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport

Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

Source: The data were obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data).

```r
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Let's say that instead of looking at the whole data set we want to consider only the months of August and September. Those have Month = 8, 9 and we can select this part of the data set with the [,] notation we discussed earlier:

```r
attach(airquality)
airAugSept <- airquality[Month>=8, ]
head(airAugSept)
```

```
##    Ozone Solar.R Wind Temp Month Day
## 93    39      83  6.9   81     8   1
## 94     9      24 13.8   81     8   2
## 95    16      77  7.4   82     8   3
## 96    78      NA  6.9   86     8   4
## 97    35      NA  7.4   85     8   5
## 98    66      NA  4.6   87     8   6
```

This task of data wrangling is so important, there are quite a lot of routines that are helping with it. One of them is **isubset**.

Here is what you do:

```
airAugSept<- isubset(airquality)
```

The app lets you use up to three conditions, we just have one (Month ≥ 8), so we can leave that alone. Now choose the condition and then hit "Click when ready to run"

Here is a screenshot:

now hit Close App and return to R.

In this example we used a very simple condition: Month ≥ 8. These conditions can be much more complicated using & (AND), | (OR) and !(NOT).

Let's say what we want only those days in August and September with a Temperature less than 80:

```
airAugSeptTemp80 <- isubset(airquality)
```

Finally let's say we want only either those days in August and September with a Temperature less than 80, or days with Wind>10:

Let's get back to the days in August and September. What we want to do with those days is to find the mean Ozone level:

```
airAugSept <- isubset(airquality)
mean(Ozone)
```

```
## [1] NA
```

Oh! Something went wrong! The problem is that the column Ozone has *missing values*, which R codes as NA. These are just what it says, for some days the Ozone level was not measured and so is missing. One way to go is to tell R to ignore the missing values:

```
mean(Ozone, na.rm=TRUE)
```

```
## [1] 42.12931
```

or we could use:

```
stat.table(Ozone)
```

```
## Warning:  37  missing values were removed!
```

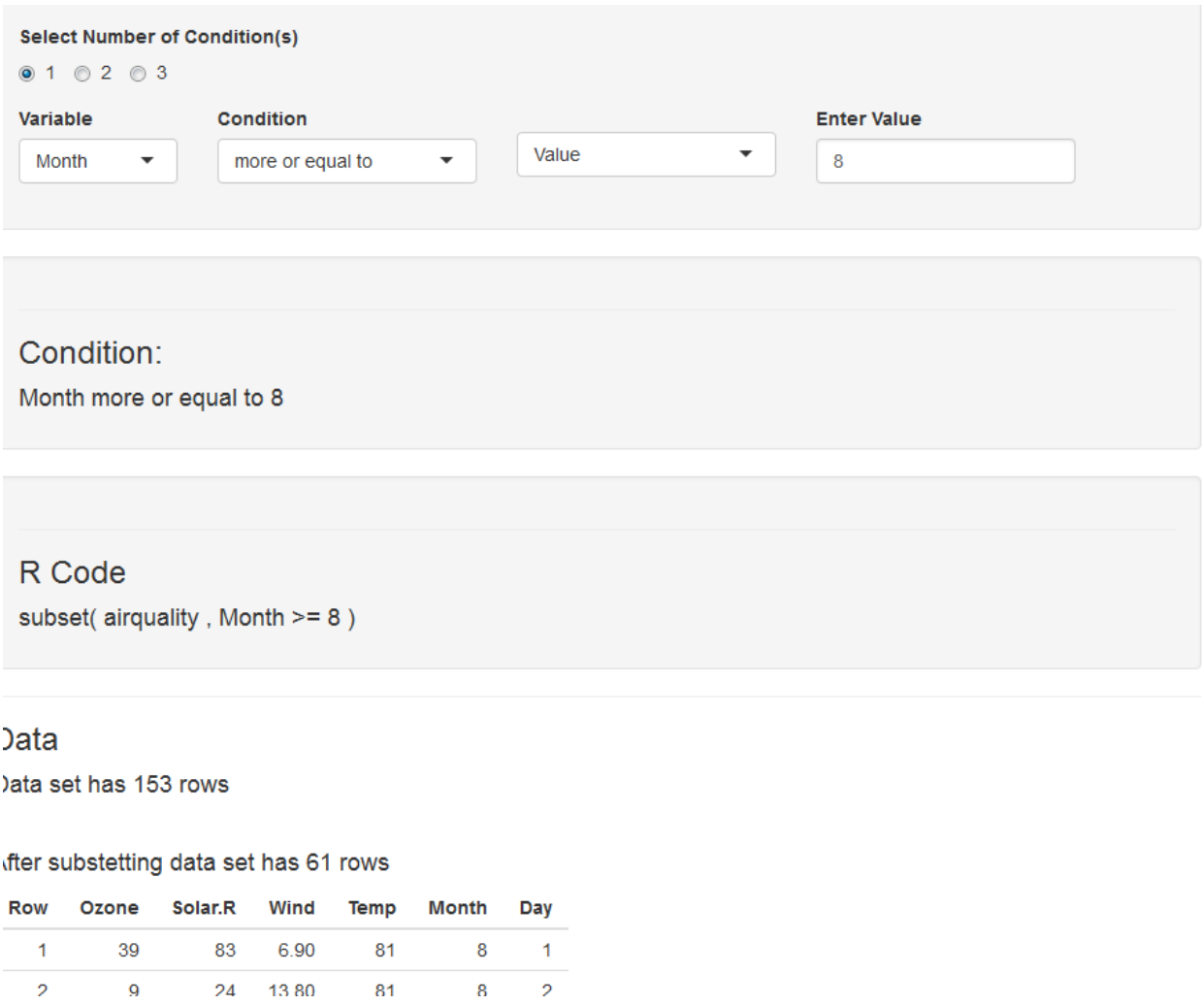

Select Number of Condition(s)
1 2 3
Variable
Month
Condition
more or equal to
Value
Enter Value
8
Condition:
Month more or equal to 8
R Code
subset(airquality , Month >= 8)
Data
Data set has 153 rows
After substetting data set has 61 rows
Row Ozone Solar.R Wind Temp Month Day
1 39 83 6.90 81 8 1
2 9 24 13.80 81 8 2

Figure 3:

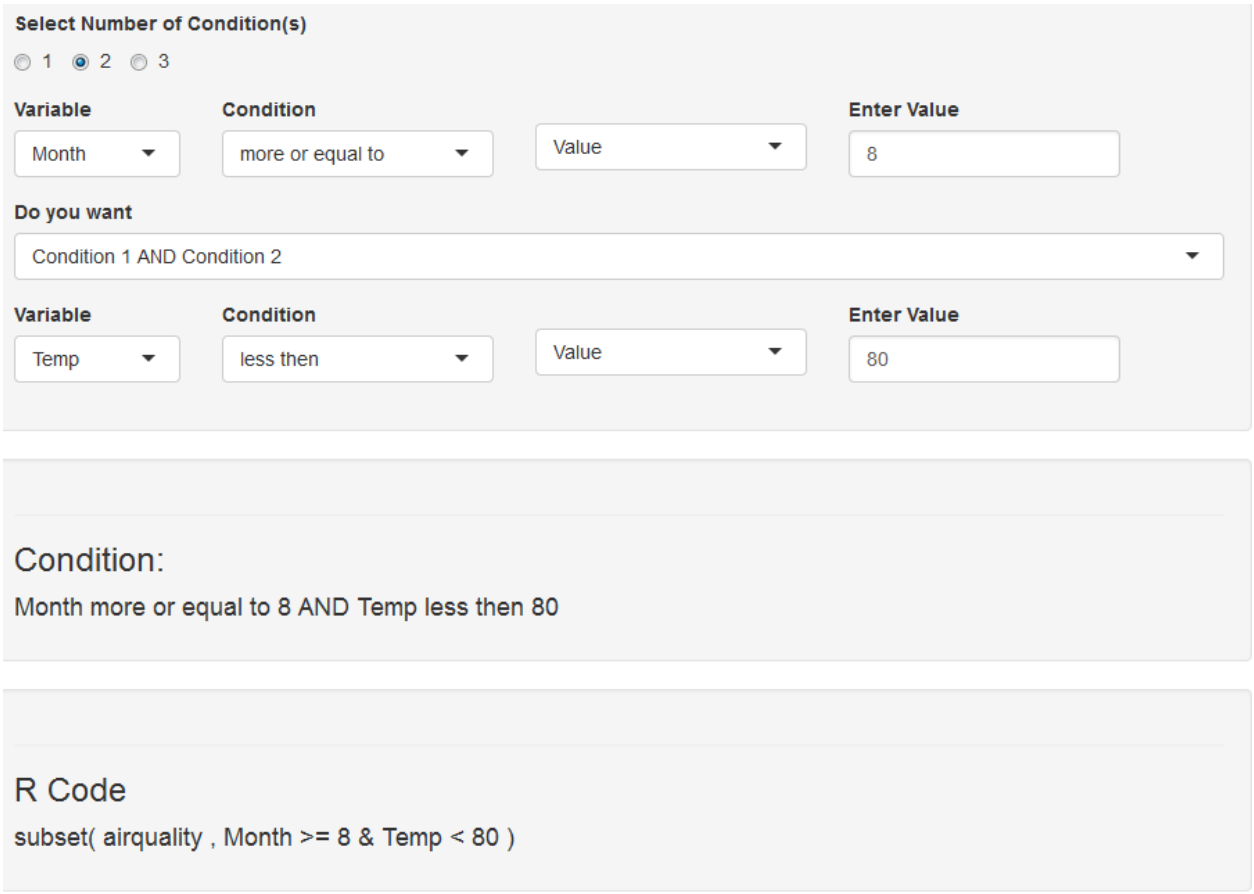

Select Number of Condition(s)
1
2
3
Variable
Month
Condition
more or equal to
Value
Enter Value
8
Do you want
Condition 1 AND Condition 2
Variable
Temp
Condition
less then
Value
Enter Value
80
Condition:
Month more or equal to 8 AND Temp less then 80
R Code
subset(airquality , Month >= 8 & Temp < 80)

Figure 4:

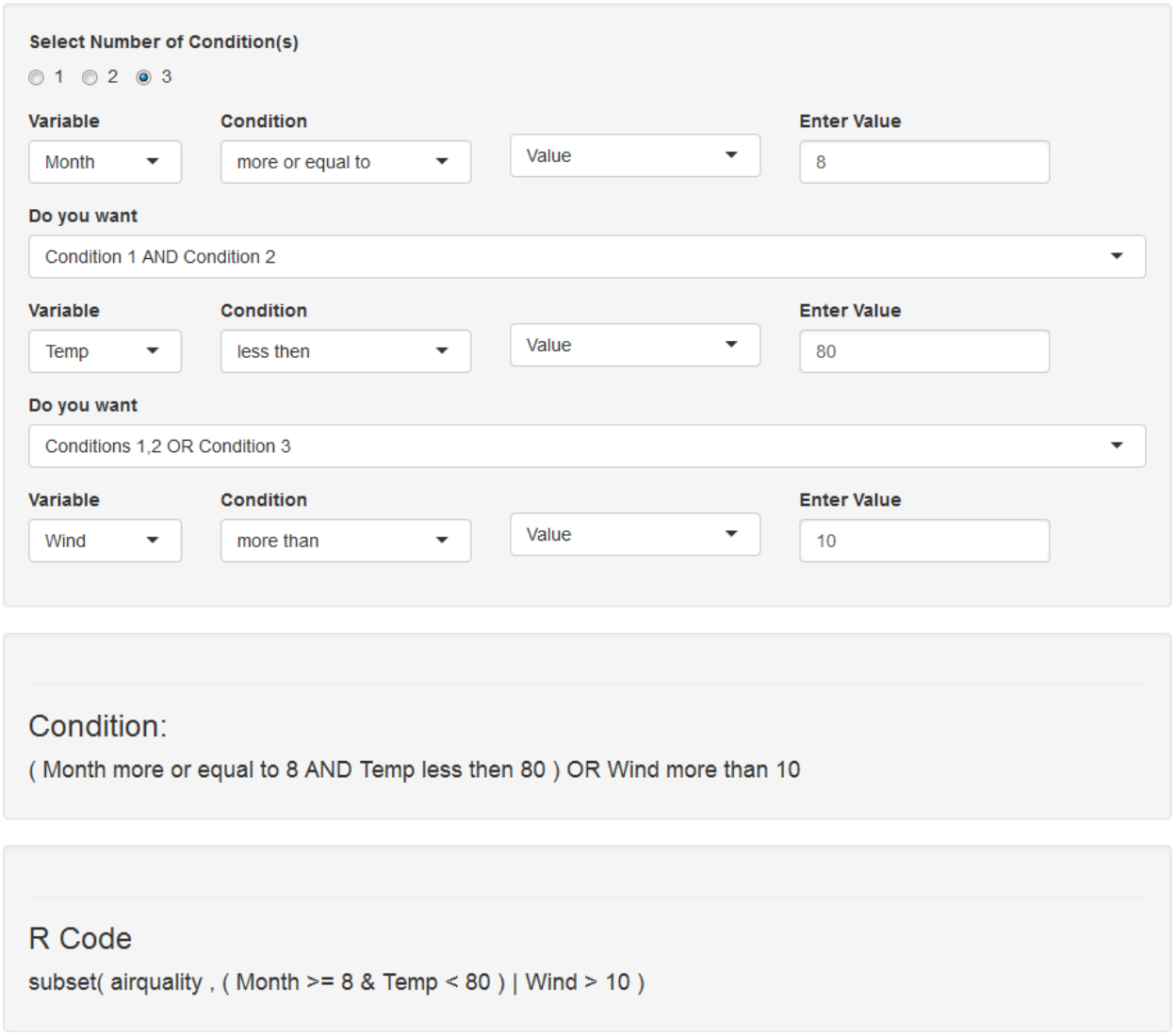

Select Number of Condition(s)
1 2 3
Variable
Month
Condition
more or equal to
Value
Enter Value
8
Do you want
Condition 1 AND Condition 2
Variable
Temp
Condition
less then
Value
Enter Value
80
Do you want
Conditions 1,2 OR Condition 3
Variable
Wind
Condition
more than
Value
Enter Value
10
Condition:
(Month more or equal to 8 AND Temp less then 80) OR Wind more than 10
R Code
subset(airquality , (Month >= 8 & Temp < 80) | Wind > 10)

Figure 5:

```
##       Sample Size Mean Standard Deviation
## Ozone         116 42.1                 33
```

OK!

But wait a minute: we are told there are 37 missing values and 116 "good" ones, for a total of 37+116=153. But there are supposed to be only 61 rows (or observations) in airAugSept. Let's check:

```
length(Ozone)
```

```
## [1] 153
```

```
nrow(airAugSept)
```

```
## [1] 61
```

What's wrong?

The problem is that Ozone still comes from the original airquality data set, but our Ozone is still hidden inside airAugSept. One solution would be to

```
attach(airAugSept)
```

but as R is warning us, now there are two Ozones, and it can get quite confusing. To be sure we work with the correct data we can do this:

```
detach(airquality)
stat.table(Ozone)
```

```
## Warning:  6  missing values were removed!
```

```
##       Sample Size Mean Standard Deviation
## Ozone          55 44.9               35.2
```

4.4.2 Case Study: Age and Gender in Puerto Rico in 2000

Breakdown of the population of USA and Puerto Rico by age and gender, according to the 2000 Census

```
head(agesex)
```

```
##           Age  Male Female
## 1 Less than 1 29601  28442
## 2           1 29543  28130
## 3           2 30252  28881
## 4           3 30643  28867
## 5           4 31248  29799
## 6           5 31621  29696
```

```
tail(agesex)
```

```
##           Age Male Female
## 98         97  282    418
```

```
## 99          98  189    296
## 100         99  123    196
## 101 100 - 104  258    448
## 102 105 - 109   47     59
## 103  Over 110   17     27
```

shows us that the data set consists of three vectors: the ages, the number of males and the number of females. The first one is a character vector ("less than 1") and the other two are numeric.

Let's answer a few questions about the age and gender in PR in 2000:

- What was the number of men and women in PR in 2000?

```
attach(agesex)
sum(Male)
```

```
## [1] 1833577
```

```
sum(Female)
```

```
## [1] 1975033
```

- How many people where there in PR?

Simple:

```
sum(Male)+sum(Female)
```

```
## [1] 3808610
```

we will need the column with the Male and Female counts a few more times, so maybe we should do it this way:

```
People <- Male + Female
head(People)
```

```
## [1] 58043 57673 59133 59510 61047 61317
```

```
sum(People)
```

```
## [1] 3808610
```

Note

we now have another variable called People among the data sets, as we can see with

```
ls()
```

It will stay there until we close R. If we want to keep it for the next time we use R we need to save everything with File > Save Workspace. If we want to save the workspace but not this variable we first have to

```
rm(People)
```

- How many newborns were there?

```
People[1]
```

```
## [1] 58043
```

- How many teenagers were there?

teenagers (Age from 13 to 19) are in rows 14 - 20, so

```
sum(People[14:20])
```

```
## [1] 433764
```

- What percentage of the population was male, rounded to 1 digit behind the decimal point?

```
sum(Male)/sum(People)*100
```

```
## [1] 48.14294
```

```
round(sum(Male)/sum(People)*100, 1)
```

```
## [1] 48.1
```

- In how many age groups were there more males than females?

Let's start with

```
Male > Female
```

```
##   [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [12]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
##  [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE FALSE FALSE FALSE
```

and now we can find

```
sum(Male > Female)
```

```
## [1] 21
```

- What age group had the largest population?

```
max(People)
```

```
## [1] 64795
```

```
People==max(People)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
##  [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE FALSE FALSE FALSE
```

```
Age[People==max(People)]
```

```
## [1] " 10"
```

Note == is the symbol for "is equal to". The others are

- < "is less than"
- <= "is less or equal to"

- > "is greater than"
- >= "is greater or equal to"

So the age group of 10 year olds is the largest. Why is this answer a bit strange?

Here is another way to do this:

```
order(People, decreasing = TRUE)
```

```
##   [1]  11  21  19  18  20  10   6   8  17   5  22  23  16   7  13  12  15
##  [18]  14   9   4   3  24   1   2  25  26  30  35  36  29  31  37  28  38
##  [35]  27  41  40  34  39  33  32  43  44  46  42  45  51  53  47  48  54
##  [52]  50  49  52  55  56  57  58  59  61  60  62  63  64  66  65  68  67
##  [69]  69  70  72  71  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86]  86  87  88  89  90  91  92  93  94  95  96  97 101  98  99 100 102
## [103] 103
```

```
head( agesex[ order(People, decreasing = TRUE), ])
```

```
##    Age  Male Female
## 11  10 33188  31607
## 21  20 32441  32154
## 19  18 32216  31705
## 18  17 32735  31070
## 20  19 32038  31744
## 10   9 31798  30101
```

another useful command is **sort**, which we can use to order one variable, by default from smallest to largest:

```
sort(People)
```

```
##   [1]    44   106   319   485   700   706   847  1122  1332  1728  2285
##  [12]  2694  3640  4466  5261  6278  7279  8414  8726  9132 10436 11659
##  [23] 13449 14211 15293 16657 17514 19403 19673 20588 21421 21865 23123
##  [34] 24982 25596 26222 26929 30387 30552 30690 32035 32737 34118 34715
##  [45] 36268 38544 39146 40807 44265 45004 45280 45875 45926 46155 46311
##  [56] 46579 48142 48987 49262 49499 50003 50009 50828 50951 51259 52213
##  [67] 52395 52553 52795 52807 53293 53573 53709 54352 54815 55124 55313
##  [78] 55754 56337 57673 58043 58725 59133 59510 60020 60112 60216 60221
##  [89] 60456 60695 60707 60748 60786 61047 61221 61231 61317 61899 63782
## [100] 63805 63921 64595 64795
```

- What was the mean age of the population?

Because the data is grouped the mean is found as follows:

$$\frac{(0 \times \text{newborns} + 1 \times \text{one year olds} + 2 \times \text{two year olds} + ...+)}{\text{total population}}$$

Age is a character variable but we need a quantitative one to do arithmetic, so let's make one as close to Age as possible:

```
Ages  <- c(0:99, 102, 107, 112)
Ages
```

```
##   [1]   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
##  [18]  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33
##  [35]  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50
##  [52]  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67
##  [69]  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84
##  [86]  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 102 107
## [103] 112
```

```
round(sum(Ages*People)/sum(People), 1)
```

```
## [1] 34
```

5 Short List of Important R Commands

- **head**: show the first k elements of a dataset

```
head(agesex, 3)
```

```
##           Age  Male Female
## 1 Less than 1 29601  28442
## 2           1 29543  28130
## 3           2 30252  28881
```

- **ls**: list of all elements of the RData file

```
head(ls())
```

```
## [1] "acorn"      "Ages"       "agesex"     "agesexUS"   "aids"
## [6] "airAugSept"
```

- **attach**: make column names of a data frame usable

```
table(Gender)
```

```
## [1] "Error: object Gender not found"
```

```
attach(upr)
head(Gender)
```

```
## [1] "Male"   "Male"   "Male"   "Male"   "Female" "Male"
```

- **args**: show arguments of a routine

```
args(stat.table)
```

```
## function (y, x, Mean = TRUE, Sort = FALSE, ndigit = 1)
## NULL
```

this does not always work:

```
args(mean)
```

```
## function (x, ...)
## NULL
```

in that case use

- **?**: show details of routine

```
?mean
```

will open a help file in a browser.

- **length**: number of elements of a vector
- **dim**: number of row and columns of a data frame
- **colnames**: names of columns of a data frame
- **rownames**: names of rows of a data frame
- **sum, mean, sd**
- **table**: count the number of occurances

```
table(Gender)
```

```
## Gender
##     F     M
## 11487 12179
```

```
table(Gender, Year)
```

```
##        Year
## Gender 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
##      F 1102 1040 1162 1137 1208 1219 1180  958  853  769  859
##      M 1151 1118 1138 1098 1256 1219 1237 1073  919  979  991
```

6 Routines in Resma3

6.0.1 Interactive Apps

idataio input and output of data into R.

isummary - graphs and numerical summaries, with or without groups.

ihist - histogram

isplot - scatterplot, with or without groups

isubset - data subsetting

6.0.2 Routines

barchart - Barcharts, one or two Variables

bplot - Boxplot

change.order - Change Ordering of Categorical Variable

chi.gof.test - Chisquare Goodnes-of-fit Test

chi.ind.test - Chisquare Test for Independence

ci.mean.sim - Simulation of Confidence Intervals for one Mean

dlr - Least Squares Regression with one Dummy Variable

dlr.predict Prediction for SLR with Dummy Variable

fivenumber - Five Number Summary

flplot - Fitted Line Graph

get.moodle.data - read data from moodle quizzes

hplot - Histogram

iplot - Interaction Plot

kruskalwallis - Kruskal-Wallis test

mallows - Best Subset Regression

mlr Multiple Regression

mlr.predict - Prediction for Multiple Regression

mplot - Marginal Plot

multiple.graphs - Combine Several Graphs into one

nplot - Normal Probability Plot

one.sample.t - Infrerence for one Mean

one.sample.prop - Inference for one Proportion

one.sample.wilcoxon - Wilcoxon Rank Sum Test, non parametric alternative to one.sample.t

oneway - One-way ANOVA

pearson.cor - Test and Interval for Correlation

prop.ps - Power and Sample Size for one Proportion

slr - Regression for One Predictor

slr.predict Prediction for Regression with one Predictor

splot Scatterplot, also with groups

stat.table - Summary Statistics

t.ps - Power and Sample Size for one Mean

test.mean.sim - Simulation of Hypothesis esting for one Mean

tukey - Tukey Multiple Comparison, one or two Factors

twoway - Two-way ANOVA

6.1 Interactive Apps

These are apps that open a new window and then allow the user to do all the work using (mostly) point and click.

Most of these apps are called with data sets as arguments. They will accept any number of arguments, which can be either vectors, matrices or data frames. If any of the the later arguments do not match the first one in length they are ignored. Some apps also return a data set.

Most of the apps also show the commands that could be used in R directly to produce the same results, either with the Resma3 commands or without them.

6.1.1 idataio

Routine to read data into R and export data to a file. It allows for

- data entered from the keyboard into a spreadsheet
- data read from a file
- data downloaded from the internet
- data copied from another program such as a browser or an Excel spreadsheet

Almost all standard file formats are supported, such as csv, excel, html, etc. For a complete list see

Examples:

```
dta <- idataio()
```

6.1.2 isummary

graphical and numerical summaries of one numerical vector, optionally rouped by a categorical variable

Examples

```
attach(mtcars)
isummary(mtcars)
isummary(mpg)
isummary(mpg, gears)
```

6.1.3 ihist

draws histograms

Examples

```
ihist(mtcars)
```

6.1.4 isplot

scatterplots

Examples

```
isplot(mtcars)
isplot(mpg, disp, gear, cyl)
```

6.1.5 isubset

subsetting a data frame or vector

Examples:

```
new.mtcars <- isubset(mtcars)
```

7 General Comments on Resma3 Routines

The routines I wrote for this course all use the following standard (where it makes sense)

first argument y is a numeric vector ("Response")

second argument x is either a numeric or categorical vector or matrix ("Predictor" or "Factor")

Sometimes there is a third argument z, always a categorical vector ("Group")

Obvious exceptions: routines for categorical data analysis (barchart, chi.ind.test, chi.gof.test)

Many of the routines have the following arguments:

return.result=FALSE (Optional): if TRUE returns results as vector for further use. This allows storing the results, for example to do simulation.

You can get all the routines and data sets by downloading and opening Resma3.RData

---

sometimes you might make a mistake entering the data, or you want to change a few values. In that case use

```
students <- edit(students)
```

This brings up the spreadsheet and you can do the changes there!

7.1 Standard R Routines

7.1.1 attach

Arguments
x: a data frame
makes column names "visible" to R

Examples:

```
attach(mothers)
mean(Length)
```

Note: you need to do this only once in any R session, it will stay until you close R.

7.1.2 mean, median, sd, IQR, quantile, cor

Summary statistics for quantitative data

Arguments
x: a numeric vector
na.rm = FALSE

Examples:

```
mean(Length)
```

```
median(Length)
```

```
sd(Length)
```

```
 IQR(Length)
```

```
quantile(Length, c(0.25,0.75))
```

Note: all these routines have an argument na.rm = FALSE, so if the data set has missing values (NA) the result is NA. Simply use na.rm = TRUE

7.1.3 table

Tables and cross-tabulation for categorical data

Arguments:
x: either a categorical vector or a data frame with two categorical columns
y: a second categorical vector (if x is a vector as well)

Examples:

```
head(rogaine,3)
```

```
table(rogaine)
```

7.1.4 cor

Pearson's correlation coefficient **Arguments:**
x: either a numeric vector or a data frame with two or more numeric columns
y: a second numeric vector (if x is a vector as well)
use = "everything", set to use="complete.obs" if NA's in the data

Examples:

```
x <- rnorm(50)
y <- rnorm(50)
cor(x, y)
```

```
cor(cbind(x,y))
```

7.1.5 subset

find a subset of a data set based on some condition(s)

Arguments:

x: a data frame

cond: some logical condition

select (Optional): which columns should be returned, default is all of them

drop=FALSE, if just one column is selected as output use drop=TRUE

Examples:

```
head(subset(wrinccensus, Satisfaction>=4, select=Income),3)
head(subset(wrinccensus, Satisfaction>=4 & Gender=="Male"),3)
head(subset(wrinccensus, Satisfaction>=4 & Gender=="Male", select=c(Income,Job.Level)),3
head(subset(wrinccensus, Satisfaction>=4 & Gender=="Male", select=Income),3)
```

Note that the last one results in a data frame with one column. You might want it as a numeric vector:

```
head(subset(wrinccensus, Satisfaction>=4 & Gender=="Male", select=Income, drop=TRUE),3)
```

NOTE: see also interactive app **isubset**

8 Resma3 routines

8.0.1 get.moodle.data

read data from moodle quizzes

highlight the data, use mouse to copy, switch to R and run

```
get.moodle.data()
```

8.1 Routines for Summary Statistics

8.1.1 stat.table

tables of summary statistics, with or without groups **Arguments** y: numeric vector (Required) x: categorical variable (Optional) Mean=TRUE: if set to FALSE table finds medians and IQRs **Examples:**

```
stat.table(Length)
stat.table(Length,Status)
stat.table(Length,Status,Mean=FALSE)
```

8.2 Routines for One Variable

8.2.1 fivenumber

five number summary and IQR, with or without groups

Arguments:

y: quantitative vector x: (optional) categorical vector

Example:

```
fivenumber(Length)
```

8.2.2 one.sample.t

Confidence interval or hypothesis test for one mean

Arguments:

y: either a vector with numbers or the sample mean of the data shat, n: standard deviation and sample size (only needed if y is sample mean)

mu.null: mean in null hypothesis (if missing confidence interval is found)

alternative = "equal": alternative hypothesis

conf.level = 95

ndigit = 1 (number of digits for rounding)

Examples:

```
one.sample.t(Length, conf.level=90)
one.sample.t(49.55, 3.38, 94, conf.level=90, ndigit=2)
one.sample.t(Length, mu.null=50, alternative="less")
```

8.2.3 t.ps

power and sample size calculations for one mean

Arguments:

n: sample size diff: difference in means
sigma: standard deviation
power: power of test
E (optional): error of confidence interval (for sample size calculation only)
conf.level=90: confidence level of confidence interval (for sample size calculation only)
alpha = 0.05: type I error probability
alternative = "equal": alternative hypothesis
routine finds whatever argument is left out (n, diff or power)

Examples:

```
t.ps(n=100, diff=1.23, sigma=5, alpha=0.1, alternative="greater")
t.ps(power=90, d=1, sigma=13, alpha=0.1, alternative="greater")
t.ps(sigma= 0.5, E=0.125, conf.level=99)
```

8.2.4 wilcoxon

Wilcoxon rank sum test for one quantitative variable - non parametric alternative to one.sample.t

Arguments:

y: quantitative vector
mu.null: mean in null hypothesis (if missing confidence interval is found)
alternative = "equal": alternative hypothesis
conf.level = 95

Examples:

```
wilcoxon(Length, conf.level=90)
wilcoxon(Length, mu.null=50, alternative="greater")
```

8.2.5 one.sample.prop

Confidence interval or hypothesis test for one proportion (percentage, probability)

Arguments:

x: number of successes
n: number of trials
pi.null: proportion in null hypothesis (if missing confidence interval is found)
alternative = "equal": alternative hypothesis
conf.level = 95

Examples:

```
one.sample.prop(40, 100, conf.level=90)
one.sample.prop(40, 100, pi.null=0.5, alternative=less)
```

8.2.6 prop.ps

Power and sample size calculations for one proportion

Arguments:

n: sample size phat: alternative proportion
pi.null: proportion under null hypothesis
power: power of test
E (optional): error of confidence interval (for sample size calculation only)
conf.level=90: confidence level of confidence interval (for sample size calculation only)
alpha = 0.05: type I error probability
alternative = "two.sided": alternative hypothesis
routine finds whatever argument is left out (n, phat or power)

Examples:

```
prop.ps(n=100, phat=0.65, pi.null=0.5)
prop.ps(power=90, phat=0.65, pi.null=0.5)
```

8.2.7 chi.gof.test

Chisquare test for multinomial proportions

Arguments:

x: observed counts p: hypothesized proportions

Example

```
chi.gof.test(c(12, 17, 20, 15, 10, 26), rep(1,6)/6)
```

8.3 Routines for Two Variables

8.3.1 pearson.cor

Confidence interval and hypothesis test for Pearson's correlation coefficient

Arguments:

y: quantitative vector
x: quantitative vector
rho.null (if missing confidence interval is found, only rho.null = 0 accepted)

conf.level = 95 confidence level of interval

Note: when the routine is run R sometimes gives a

Warning message:
Continuous x aesthetic – did you forget aes(group=...)?

just ignore this

Example:

```
pearson.cor(Draft.Number, Day.of.Year, rho.null = 0)
```

8.4 Routines for Simulations

8.4.1 ci.mean.sim

does a simulation for coverage of the t test confidence intervals

Arguments:

n : sample size mu: mean sigma: standard deviationconf.level: nominal coverage

Example:

```
ci.mean.sim(n=500,mu=75,sigma=30,conf.level=99)
```

8.4.2 test.mean.sim

does a simulation of the p value of the t test. If mu.null=mu it finds the true type I error α, otherwise the power of the test. In either case it draws the histogram of p values.

Arguments:

n : sample size
mu: mean
mu.null=mu: value of mean under null hypothesis
sigma: standard deviation
alpha: nominal alpha

Examples:

```
test.mean.sim(n=20, mu=5, sigma=1, alpha=0.1)
test.mean.sim(n=20, mu=5, mu.null=5.5, sigma=1, alpha=0.1)
```

8.5 Routines for Graphs

8.5.1 barchart

bar charts

Arguments:

y: a table (often from a call to the table routine)

Percent: if missing graph uses counts. Other values are "Grand", "Row" or "Column" for respective percentages

new.order: for changing the order of the bars

Polygon = FALSE if TRUE adds polygon

Examples:

```
attach(rogaine)
barchart(table(Growth))
barchart(table(Growth), Percent="Grand")
barchart(table(Growth), Percent="Grand",Polygon=TRUE)
barchart(table(rogaine))
barchart(table(rogaine),Percent="Row")
```

8.5.2 hplot

Histogram, if desired with fitted density

Arguments:

x: numerical data
f: name of distribution (Optional)
par: parameters of distribution(Optional)
n: number of bins (Optional) label_x, main_title: x axis label and graph title (Optional)

Examples:

```
hplot(Length)
hplot(Length, label_x = "Length of Babies (cm)", main_title = "Mothers, Babies and Cocai
hplot(Length, f = "norm", par = c(mean(Length), sd(Length)))
```

8.5.3 bplot

Boxplot / do.violinplot

Arguments:

y: numeric vector or matrix or data frame
x: catagorical vector (Optional)
do.violin = FALSE: if TRUE does violin plot
orientation="vertical", if orientation="horizontal" boxplot is drawn horizontally
new_order: change the order of the boxes. Either a vector of position numbers or "Sort", then sorted from smallest mean to largest.
label_x, label_y, main_title: axes labels and graph title (Optional)
Examples:

```
bplot(Length)
bplot(Length, Status)
bplot(Length, Status, label_y  = "Length of Babies (cm)",
      label_x = "Drug Status",
      main_title = "Mothers, Babies and Cocain Use")
```

8.5.4 splot

Scatterplot, possibly with groups and fits

Arguments:

y: numeric vector , y axis
x: numeric vector, x axis
z: catagorical variable (Optional)
w: second catagorical variable (Optional)
plot.points=TRUE: if FALSE dots are not plotted add.line = 0: adds lines, if add.line=1 least squares regression line, if add.line=2 LOESS, if add.line=3 it does the line graph

jitter = FALSE: if true jitters dots
use.facets = FALSE: if TRUE usess facets instead of colors for z
errorbars = FALSE: if TRUE adds error band to fit

label_x, label_y, label_z, main_title: axes labels and graph title (Optional)
add.text, add.text_x, add.text_y: add text to graph (Optional)
plotting.size = 1: size of plotting symbols
plotting.symbols: change plotting symbols. can use either symbols added on keyboard or numbers corresponding to R symbols key(Optional)

plotting.colors: change colors, can use either numbers corresponding to R color key or explicit text : pcolor="red" (Optional)

ref_x, ref_y: add reference lines (Optional)

log_x = FALSE, log_y = FALSE: change to log scale

no.legends = FALSE: rmove all alegends

Examples:

```
attach(salaries)
splot(Salary,Years)
splot(Salary,Years, add.line=1)
splot(Salary,Years, Level, add.line=1)
splot(Salary,Years, add.line=3)
```

```
attach(upr)
splot(y = Freshmen.GPA, x = IGS, z = Gender, use.facets = TRUE, add.line = 1, label_y =
```

NOTE: see also ineractive app **isplot**

8.5.5 mplot

Marginal plot with scatterplot and boxplots

Arguments:

y: numeric vector , y axis

x: numeric vector, x axis

z: catagorical variable (Optional)

add.line = 0: adds lines, if add.line=1 least squares regression line, if add.line=2 LOESS, if add.line=3 it does the line graph

Examples:

```
mplot(Salary, Years)
```

Note: when the routine is run R sometimes gives a Warning message: Continuous x aesthetic – did you forget aes(group=...)? Just ignore that

8.5.6 flplot

Fitted line plot, allows for log transforms or polynomial fitting

Arguments:

y: numeric vector , y axis

x: numeric vector, x axis

z: catagorical variable (Optional)

additive = FALSE: if true fits parallel lines

logx = FALSE, logy = FALSE: if true applies log transforms

polydeg = 1: degree of polynomial to be fit

jitter = FALSE: if true jitters dots

Examples:

```
attach(longjump)
flplot(LongJump, Year)
flplot(LongJump, Year, polydeg=2)
attach(elusage)
flplot(elusage[,3], elusage[,4], logx=TRUE, logy=TRUE)
```

8.5.7 nplot

Normal probability plot

Arguments:

y: numerical vector

x: categorical vector (Optional)

Examples:

```
nplot(euros[,1])
```

8.5.8 iplot

Interaction plot

Arguments:

y: numerical vector

x and z: categorical vectors

Examples:

```
attach(fermentation)
iplot(Ethanol, Sugar, Oxygen)
```

8.5.9 multiple.graphs

combine (up to four graphs) in one

8.5.9.1 Arguments:

ggplt objects, likely generated using other graph functions with the argument returnGraph=TRUE

titles (Optional) titles for each graph

Examples:

```
attach(gasoline)
plt1 <- bplot(MPG, Gasoline, returnGraph=TRUE)
plt2 <- bplot(MPG, Automobile, returnGraph=TRUE)
multiple.graphs(plt1,plt2)
```

```
x<-rnorm(1000)
multiple.graphs(
  hplot(x, n=10, returnGraph=TRUE),
  hplot(x, n=25, returnGraph=TRUE),
  hplot(x, n=50, returnGraph=TRUE),
  hplot(x, n=100, returnGraph=TRUE),
  titles =  paste(c(10, 25, 50, 100), "bins")
  )
```

8.6 Routines for Testing with two or more Variables

8.6.1 chi.ind.test

Chisquare test of independence

Arguments:

x: a table of counts

Examples:

```
chi.ind.test(table(rogaine))
```

8.6.2 oneway

ANOVA with one factor

Arguments:

y: numeric vector

x: categorical vector

ndigit = 1: rounding answer to 1 digit

var.equal = TRUE: assume equal variance

conf.level = 95: in the case of a categorical variable with 2 levels finds a 95% confidence interval for the difference in means

Examples:

```
oneway(Length, Status)
```

8.6.3 kruskalwallis

Non-parametric ANOVA

Arguments:

y: numeric vector

x: categorical vector

Examples:

```
kruskalwallis(Length, Status)
```

8.6.4 twoway

ANOVA with two factors

Arguments:

y: numeric vector

x, z: categorical vectors

with.interaction = TRUE: assume interaction is present (defaults to FALSE if there are no repeated measurements)

Examples:

```
attach(gasoline)
twoway(MPG, Gasoline, Automobile)
twoway(MPG, Gasoline, Automobile, with.interaction="FALSE")
```

8.6.5 tukey

Tukey's Multiple Comparison in ANOVA

Arguments:

y: numeric vector

x : categorical vector

z : second categorical vector (Optional)

with.interaction = TRUE: assume interaction is present (defaults to FALSE if there are no repeated measurements)

which="first": do comparison for first categorical variable (x), or change to which="second" or which="interaction"

Examples:

```
tukey(mothers[,2], mothers[,1])
tukey(MPG, Gasoline, Automobile, which="first")
tukey(MPG, Gasoline, Automobile, which="interaction")
```

8.6.6 slr

Linear Regression with one predictor, including polynomial regression

Arguments:

y, x: numerical vectors

no.intercept = FALSE: fit intercept?

polydeg = 1: fit polynomial of higher degree?

show.tests=FALSE: if TRUE t tests for coefficients are shown

Examples:

```
slr(wine[,3],wine[,2])
slr(wine[,3],wine[,2],polydeg=2)
slr(log(wine[,3]),wine[,2],polydeg=2)
```

8.6.7 slr.predict

Prediction for simple linear regression

Arguments:

same as slr. In addition:

newx = x: predict for values for x (can be vector). If missing predict for values in data set.

interval: either "PI" for prediction intervals or "CI" for confidence intervals

conf.level = 95

Examples:

```
slr.predict(wine[,3], wine[,2],newx=c(2,2.5,3), interval="PI", conf.level=90)
```

8.6.8 mlr

Linear Regression with more than one predictor

Arguments:

y: numerical vector

x: numeric matrix with predictors in columns

show.tests=FALSE: if TRUE t tests for coefficients are shown

returnModel=FALSE, if TRUE fit object is returned (and can be used in other routines)

Examples:

```
mlr(houseprice[,1], houseprice[, -1])
```

8.6.9 mlr.predict

Prediction for regression with more than one predictor

Arguments:

same as slr.predict but here x and newx are matrices

Examples:

```
newx <- cbind(c(2000, 2100, 2200), rep(1, 3), rep(2, 3), rep(2, 3))
mlr.predict(houseprice[,1], houseprice[, -1], newx=newx, interval="PI", conf.level = 99)
```

8.6.10 mallows

Best subset regression with Mallow's Cp

Arguments:

same as mlr

Examples:

```
mallows(houseprice[,1], houseprice[, -1] )
```

8.6.11 dlr

Linear regression with one dummy variable

Arguments:

y: numerical vector

x: numeric vectorz: categorical vector

additive = FALSE: if parallel lines set to TRUE

show.tests=FALSE: if TRUE t tests for coefficients are shown

Examples:

```
dlr(salaries[,1], salaries[,2], salaries[,3])
dlr(salaries[,1], salaries[,2], salaries[,3], additive=T)
```

8.6.12 dlr.predict

Prediction for regression with a dummy variable

Arguments:

same as slr.predict but also needs newz: values of categorical variable for prediction

Examples:

```
dlr.predict(salaries[, 1], salaries[, 2], salaries[, 3],
  newx=5, newz="Low", interval="PI")
```

8.7 Miscellaneous Routines

8.7.1 change.order

Change the order of a categorical variable

Arguments:

z: categorical variable

NewOrder: can be a numeric vector specifying a certain order or a categorical vector with ordered values of z

Examples:

```
bplot(Length, Status)
bplot(Length, change.order(Status,c(2,1,3)))
bplot(Length, change.order(Status,c("Throughout","First Trimester","Drug Free")))
```

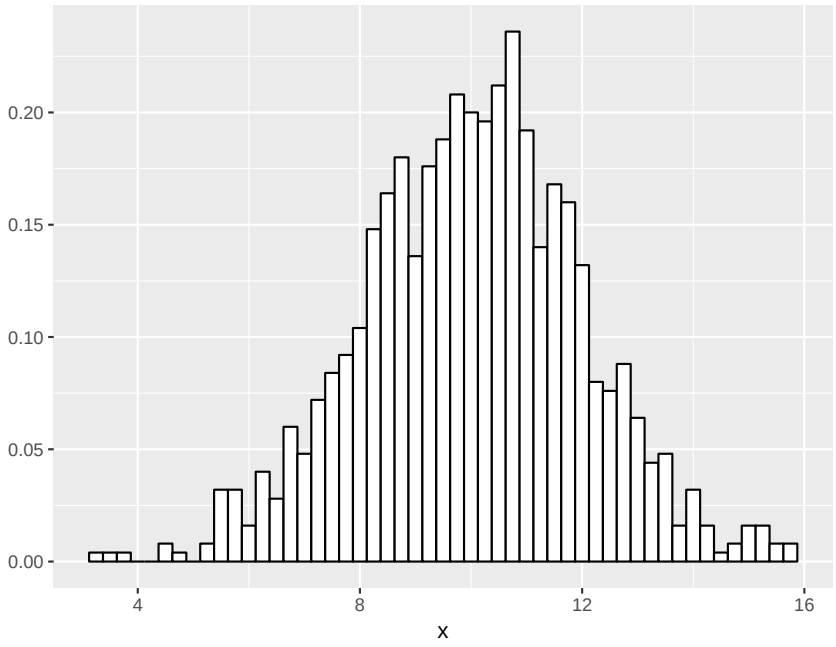
9 Resma3 vs Basic R

In this section we will see how some of our problems could be done with base R.

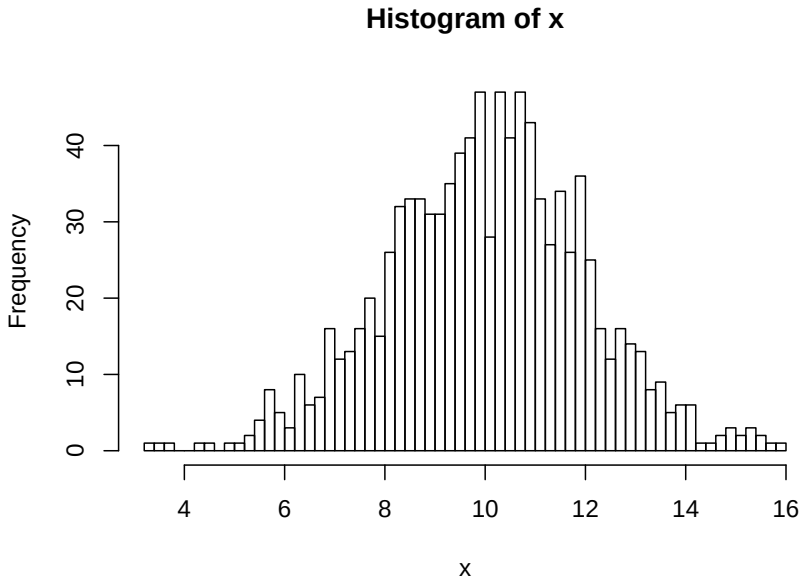
9.1 Graphs

9.1.1 Histogram

```
x <- rnorm(1000, 10, 2)
hplot(x, n=50)
```

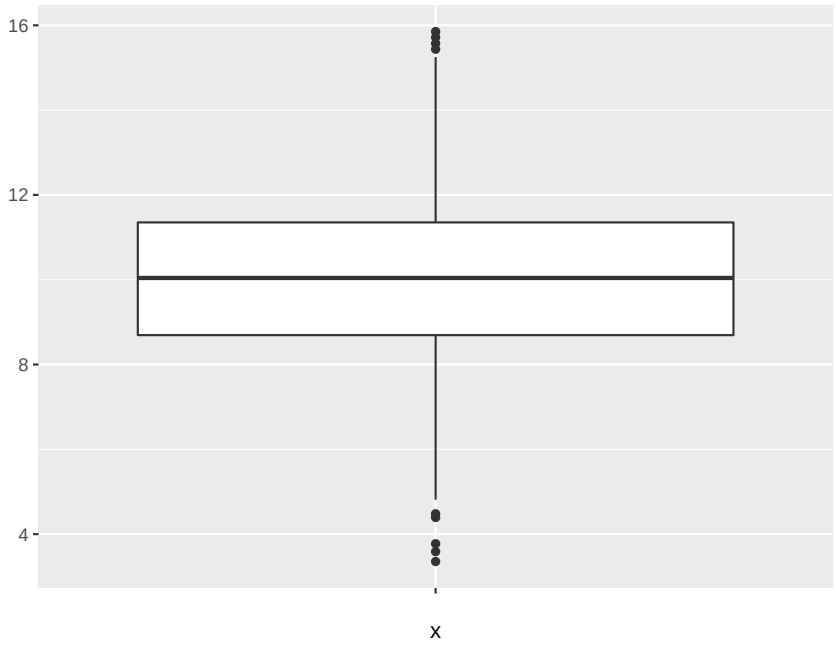



```
hist(x, 50)
```

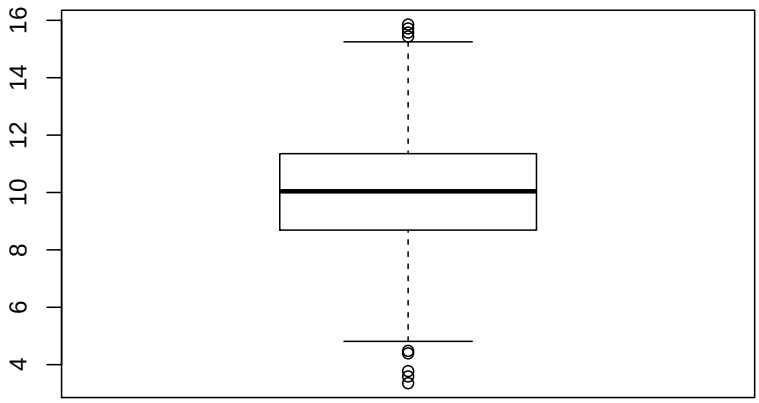


9.1.2 Boxplot

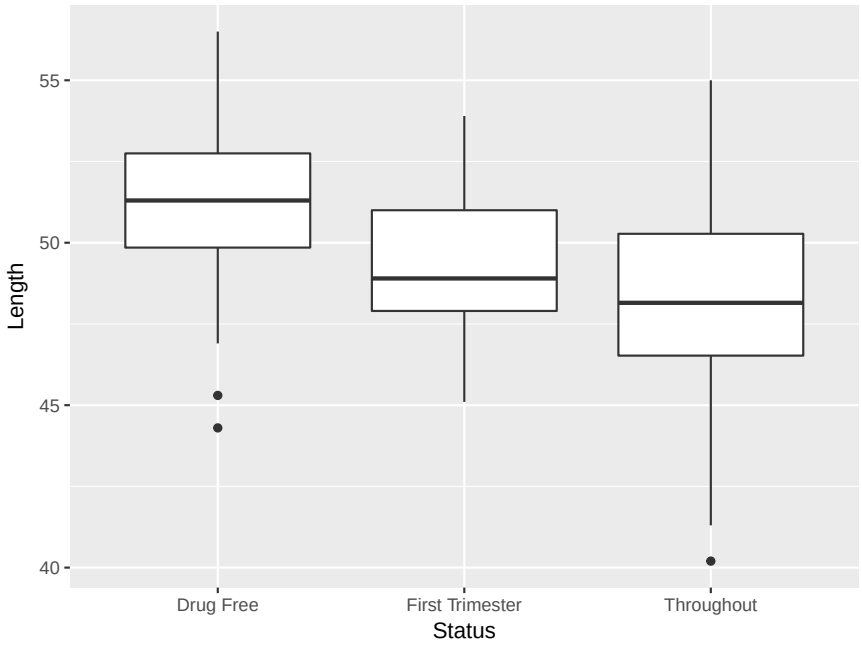
```
bplot(x)
```



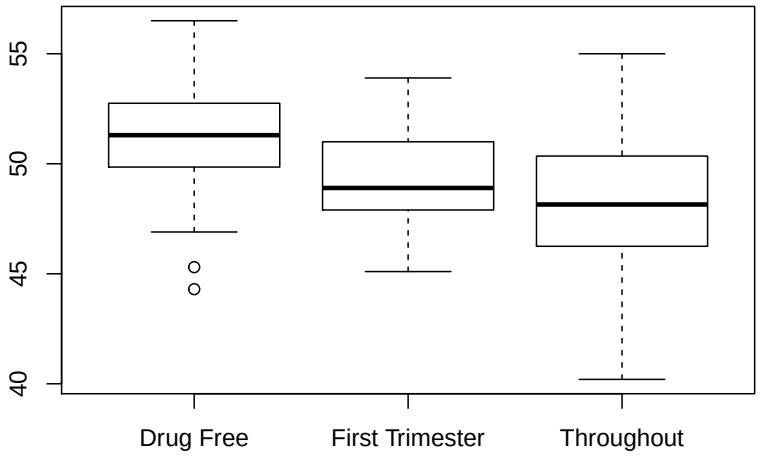

```
boxplot(x)
```




```
attach(mothers)
bplot(Length, Status)
```

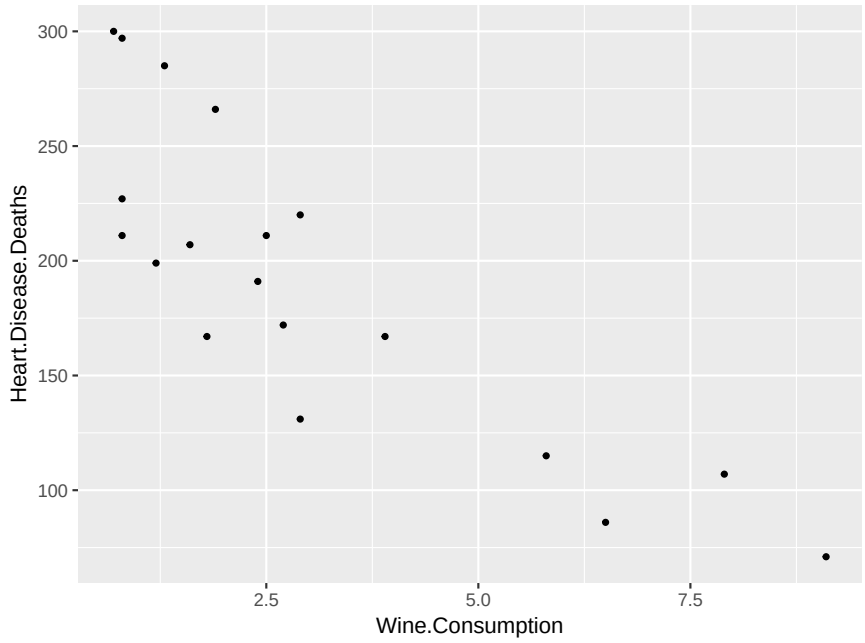



```
boxplot(Length~Status)
```

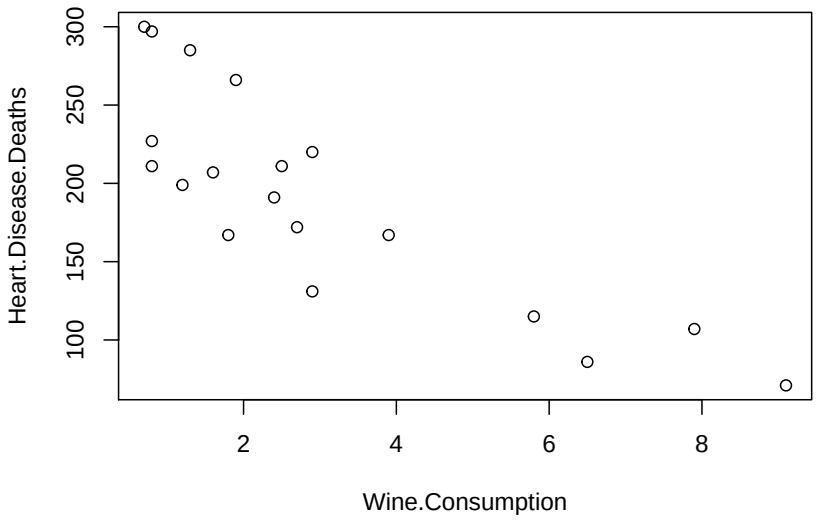


9.1.3 Scatterplot

```
attach(wine)
splot(Heart.Disease.Deaths, Wine.Consumption)
```




```
plot(Wine.Consumption, Heart.Disease.Deaths)
```



9.2 Summary Statistics

```
fivenumber(x, ndigit = 2)
```

```
##  Minimum   Q1 Median    Q3 Maximum
##     3.35 8.69  10.04 11.35   15.85
## IQR =  2.66
```

```
round(c(min(x), quantile(x, 0.25), median(x), quantile(x, 0.75), max(x)), 2)
```

```
##          25%          75%
##  3.35  8.69 10.04 11.35 15.85
```

```
stat.table(x, ndigit = 2)
```

```
##   Sample Size  Mean Standard Deviation
## x        1000 10.02                1.98
```

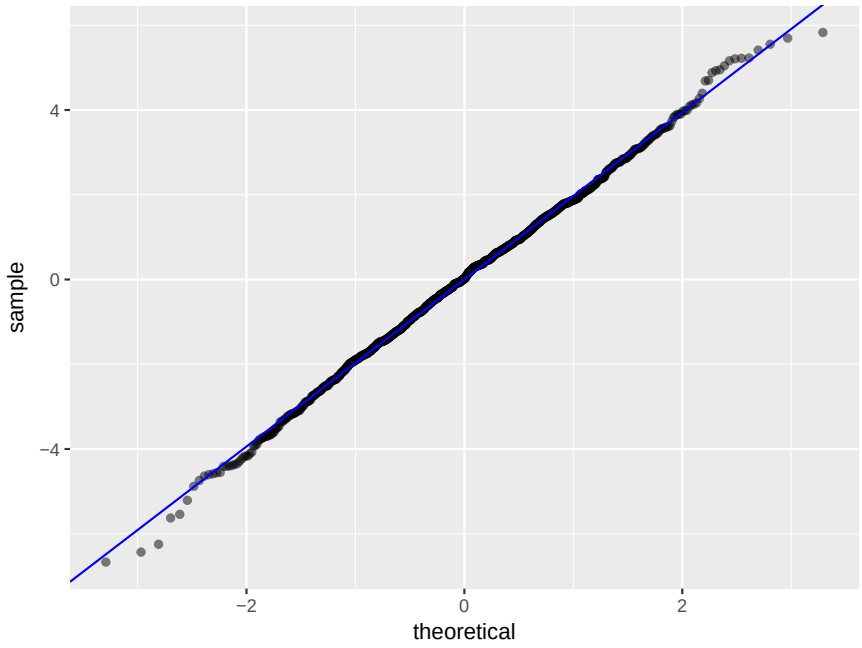
```
round(c(length(x), mean(x), sd(x)), 2)
```

```
## [1] 1000.00   10.02    1.98
```

9.3 Confidence Intervals/Hypothesis Tests

9.3.1 Mean

```
one.sample.t(x, conf.level = 90,  ndigit = 3)
```



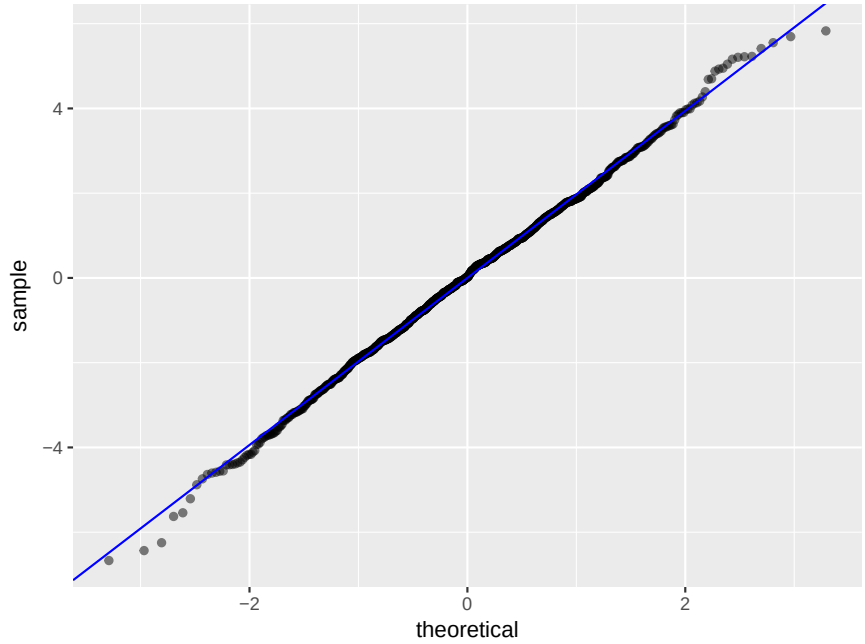

```
## A 90% confidence interval for the population mean is (9.919, 10.125)
```

```
t.test(x, conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 160.32, df = 999, p-value < 0.00000000000000022
## alternative hypothesis: true mean is not equal to 0
```

```
## 90 percent confidence interval:
##   9.919363 10.125213
## sample estimates:
## mean of x
##  10.02229
```

```
one.sample.t(x, mu.null = 10,
             alternative = "greater", ndigit = 3)
```




```
## p value of test H0: mu=10 vs. Ha: mu > 10:  0.3608
```

```
t.test(x, mu=10, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  x
## t = 0.35652, df = 999, p-value = 0.3608
## alternative hypothesis: true mean is greater than 10
## 95 percent confidence interval:
##  9.919363      Inf
## sample estimates:
## mean of x
##  10.02229
```

the *t.ps* command does not exist in base R.

9.3.2 Proportion

```
one.sample.prop(60, 100, conf.level = 90,  ndigit = 3)
```

```
## A 90% confidence interval for the population proportion is (0.513, 0.682)
```

```
prop.test(60, 100, conf.level = 0.9)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  60 out of 100, null probability 0.5
## X-squared = 3.61, df = 1, p-value = 0.05743
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.5127842 0.6816248
## sample estimates:
##   p
## 0.6
```

```
one.sample.prop(60, 100, pi.null = 0.5,
             alternative = "greater", ndigit = 3)
```

```
## p value of test H0: pi=0.5 vs. Ha: pi > 0.5:  0.0287
```

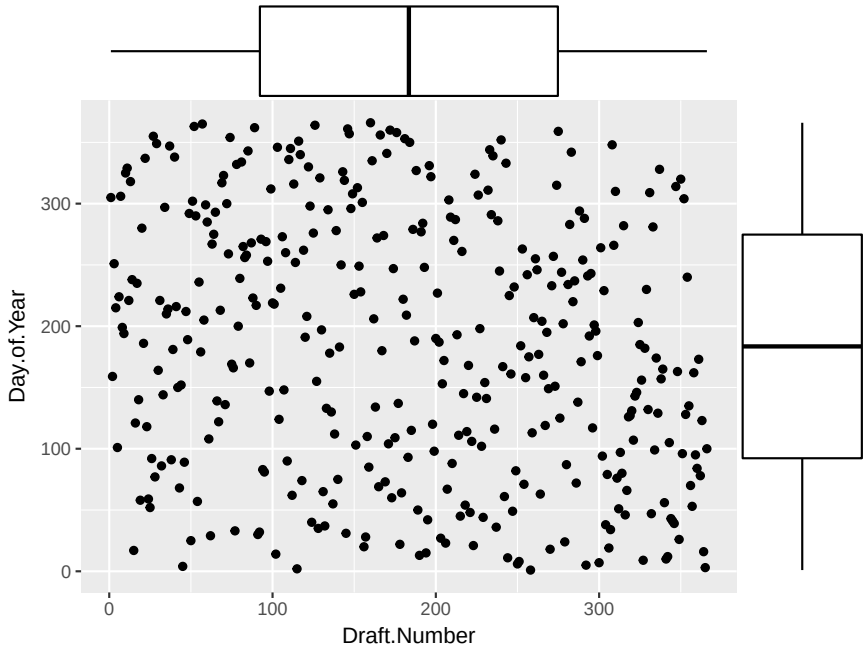
```
prop.test(60, 100, p=0.5, alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  60 out of 100, null probability 0.5
## X-squared = 3.61, df = 1, p-value = 0.02872
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5127842 1.0000000
## sample estimates:
##   p
## 0.6
```

the *prop.ps* command does not exist in base R.

9.3.3 Correlation

```
attach(draft)
pearson.cor(Draft.Number, Day.of.Year, conf.level = 90)
```




```
## A 90% confidence interval for the
## population correlation coefficient is ( -0.306, -0.143 )
```

```
cor.test(Draft.Number, Day.of.Year, conf.level = 0.9)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Draft.Number and Day.of.Year
## t = -4.4272, df = 364, p-value = 0.00001264
## alternative hypothesis: true correlation is not equal to 0
## 90 percent confidence interval:
##  -0.3061994 -0.1427007
## sample estimates:
##        cor
## -0.2260414
```

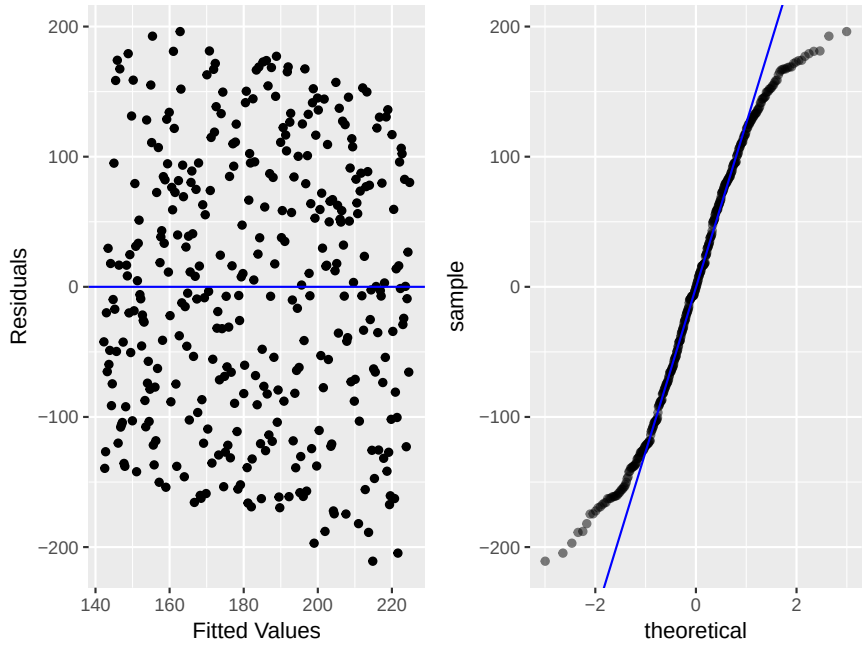
```
pearson.cor(Draft.Number, Day.of.Year, rho.null = 0)
```

```
## p value of test H0: rho=0 vs. Ha: rho <> 0:  0.000
```

9.3.4 Regression

- Simple Regression

```
slr(Draft.Number, Day.of.Year)
```



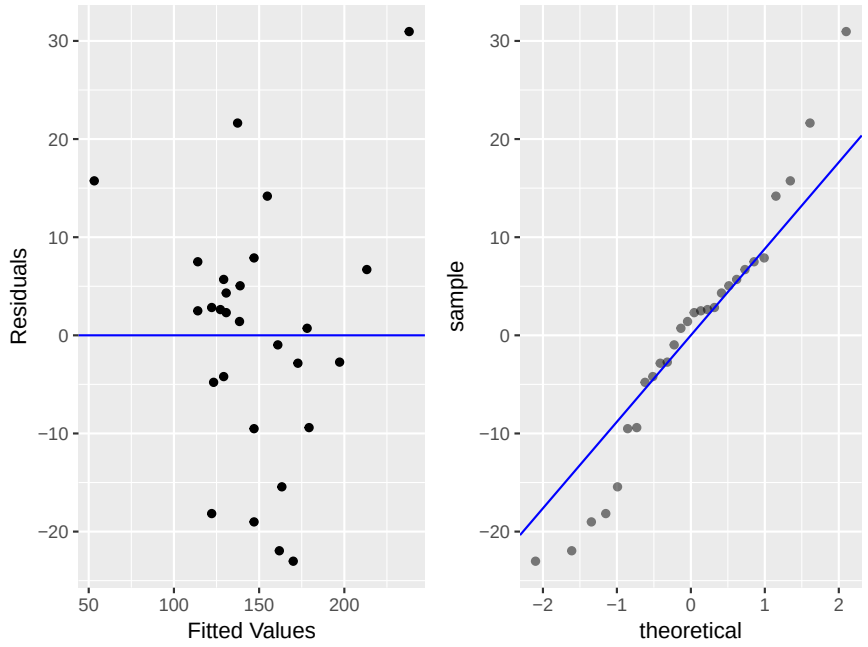

```
## The least squares regression equation is:
##  Draft.Number  = 225.009 - 0.226 Day.of.Year
## R^2 = 5.11%
```

```
summary(lm(Draft.Number~Day.of.Year))
```

```
##
## Call:
## lm(formula = Draft.Number ~ Day.of.Year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -210.837  -85.629   -0.519   84.612  196.157
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 225.00922   10.81197  20.811 < 0.0000000000000002
## Day.of.Year  -0.22606    0.05106  -4.427            0.0000126
##
## Residual standard error: 103.2 on 364 degrees of freedom
## Multiple R-squared:  0.05109,    Adjusted R-squared:  0.04849
## F-statistic:  19.6 on 1 and 364 DF,  p-value: 0.00001264
```

- Multiple Regression

```
attach(houseprice)
mlr(Price, houseprice[, -1])
```




```
## The least squares regression equation is:
##  Price  =  -67.62 + 0.086 Sqfeet - 26.493 Floors - 9.286 Bedrooms + 37.381 Baths
## R^2 = 88.6%
```

```
summary(lm(Price ~ ., data=houseprice))
```

```
##
## Call:
## lm(formula = Price ~ ., data = houseprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.018  -5.943   1.860   5.947  30.955
##
## Coefficients:
##              Estimate Std. Error t value     Pr(>|t|)
## (Intercept) -67.61984   17.70818  -3.819     0.000882
## Sqfeet        0.08571    0.01076   7.966 0.0000000462
## Floors      -26.49306    9.48952  -2.792     0.010363
## Bedrooms     -9.28622    6.82985  -1.360     0.187121
## Baths        37.38067   12.26436   3.048     0.005709
##
## Residual standard error: 13.71 on 23 degrees of freedom
## Multiple R-squared:  0.8862, Adjusted R-squared:  0.8665
## F-statistic:  44.8 on 4 and 23 DF,  p-value: 0.0000000001558
```

- Best Subset Regression

```
library(leaps)
mallows(Price, houseprice[, -1])
```

```
##  Number of Variables Cp   Sqfeet Floors Bedrooms Baths
##  1                   8.83 X
##  2                   8.81 X                     X
##  3                   4.85 X      X              X
##  4                   5    X      X      X       X
```

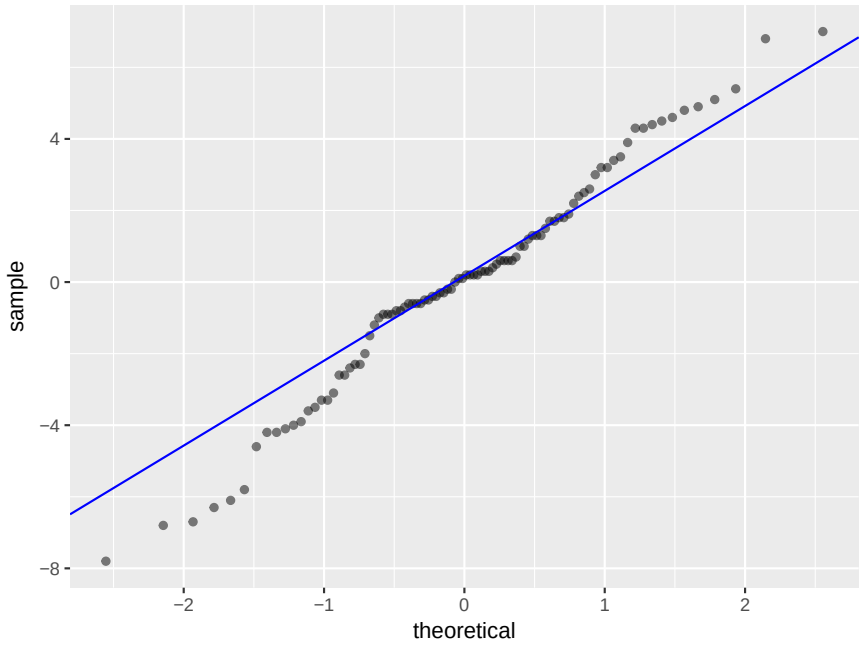
```
leaps(houseprice[, -1], Price)
```

```
## $which
##       1     2     3     4
## 1  TRUE FALSE FALSE FALSE
## 1 FALSE FALSE FALSE  TRUE
## 1 FALSE FALSE  TRUE FALSE
## 1 FALSE  TRUE FALSE FALSE
## 2  TRUE FALSE FALSE  TRUE
## 2  TRUE  TRUE FALSE FALSE
## 2  TRUE FALSE  TRUE FALSE
## 2 FALSE FALSE  TRUE  TRUE
## 2 FALSE  TRUE FALSE  TRUE
## 2 FALSE  TRUE  TRUE FALSE
## 3  TRUE  TRUE FALSE  TRUE
## 3  TRUE FALSE  TRUE  TRUE
## 3  TRUE  TRUE  TRUE FALSE
## 3 FALSE  TRUE  TRUE  TRUE
## 4  TRUE  TRUE  TRUE  TRUE
##
## $label
## [1] "(Intercept)" "1"           "2"           "3"           "4"
##
## $size
##  [1] 2 2 2 2 3 3 3 3 3 4 4 4 4 5
##
## $Cp
##  [1]   8.834171  92.088525 104.303380 161.057329   8.812489  10.306028
##  [7]  10.812154  66.886236  77.214388  87.881962   4.848657  10.794275
## [13]  12.289752  66.450032   5.000000
```

9.3.5 ANOVA

- oneway

```
oneway(Length, Status)
```




```
## p value of test of equal means: p = 0.000
## Smallest sd:  2.5    Largest sd : 3.6
```

```
summary(aov(Length~Status))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Status       2  181.4   90.69   9.319 0.000208
## Residuals   91  885.6    9.73
```

- twoway

```
attach(gasoline)
twoway(MPG, Gasoline, Automobile)
```

```
##             Df Sum Sq Mean Sq F value             Pr(>F)
## x            3 25.405   8.468  90.464 0.000000000000321
## z            2  0.527   0.263   2.813             0.0799
## x:z          6  0.909   0.151   1.618             0.1854
## Residuals   24  2.247   0.094
##                   [,1]
## Gasoline  p =   0.0000
## Automobile  p = 0.0799
## Interaction p = 0.1854
```

```
G <- as.factor(Gasoline)
A <- as.factor(Automobile)
summary(aov(MPG ~ G * A))
```

```
##             Df Sum Sq Mean Sq F value             Pr(>F)
## G            3 25.405   8.468  90.464 0.000000000000321
## A            2  0.527   0.263   2.813             0.0799
```

```
## G:A          6  0.909   0.151   1.618            0.1854
## Residuals   24  2.247   0.094
```

```
twoway(MPG, Gasoline, Automobile, with.interaction = FALSE)
```

```
##             Df Sum Sq Mean Sq F value             Pr(>F)
## x            3 25.405   8.468  80.510 0.0000000000000189
## z            2  0.527   0.263   2.504             0.0987
## Residuals   30  3.156   0.105
##                   [,1]
## Gasoline  p =   0.0000
## Automobile  p = 0.0987
```

```
summary(aov(MPG ~ G + A))
```

```
##             Df Sum Sq Mean Sq F value             Pr(>F)
## G            3 25.405   8.468  80.510 0.0000000000000189
## A            2  0.527   0.263   2.504             0.0987
## Residuals   30  3.156   0.105
```

10 Categorical Data

10.0.1 Case Study: wrinccensus

Consider the variable Gender. Clearly this is categorical data. Usually the first thing one would do is simply count how many of each type there are:

```
attach(wrinccensus)
table(Gender)
```

```
## Gender
## Female   Male
##   9510  14281
```

10.0.2 Case Study: Race and Education

According to a table from the US Department of Education there were 19,980,000 students in US colleges in the fall of 2010. Their breakdown by race was as follows:

```
##
## American Indian           Asian           Black        Hispanic
##          196000         1282000         3039000         2741000
##           White
##        12722000
```

If a table is used for presentation purposes it should usually include a little more information and maybe a better ordering, for example by size. Also, big numbers are often expressed in bigger units:

	Number (in 1000)	Percentage
White	12722	63.7
Black	3039	15.2
Hispanic	2741	13.7
Asian	1282	6.4
American Indian	196	1.0

In order to compute the percentages we need to divide by the total and multiply by 100. The total is found using the sum command:

```
x <- c(12722, 3039, 2741, 1282, 196)
round(x/sum(x)*100,1)
```

```
## [1] 63.7 15.2 13.7  6.4  1.0
```

As we said before, some categorical variables have a built-in (natural) ordering, for example t-shirt size (small, medium, large, x-large) or grades (A,B, . . .). Such an ordering can also be used.

10.0.3 Graphs for Categorical Data

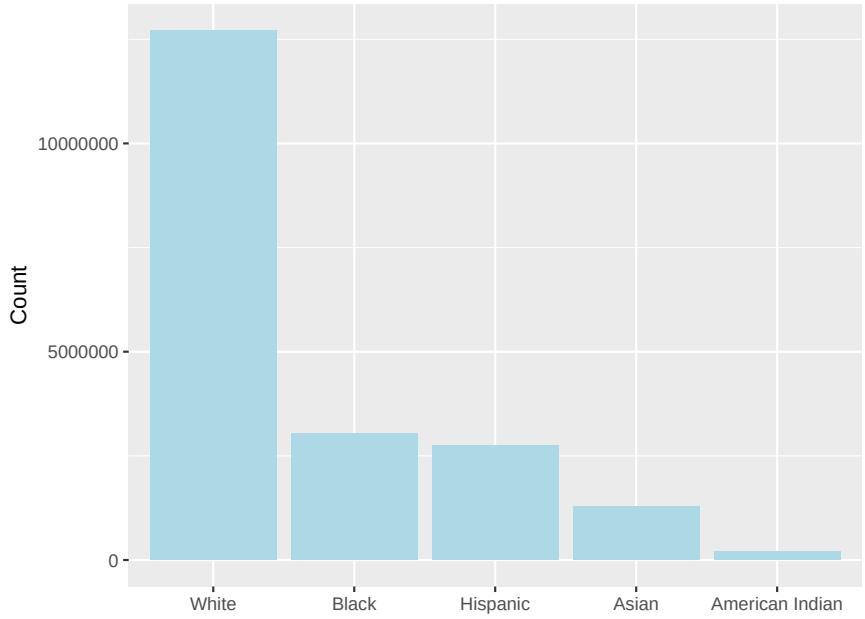
A very popular Choice: Pie Charts

but:

[Death to Pie Charts] (http://www.storytellingwithdata.com/blog/2011/07/death-to-pie-charts)

Much better: Bar charts

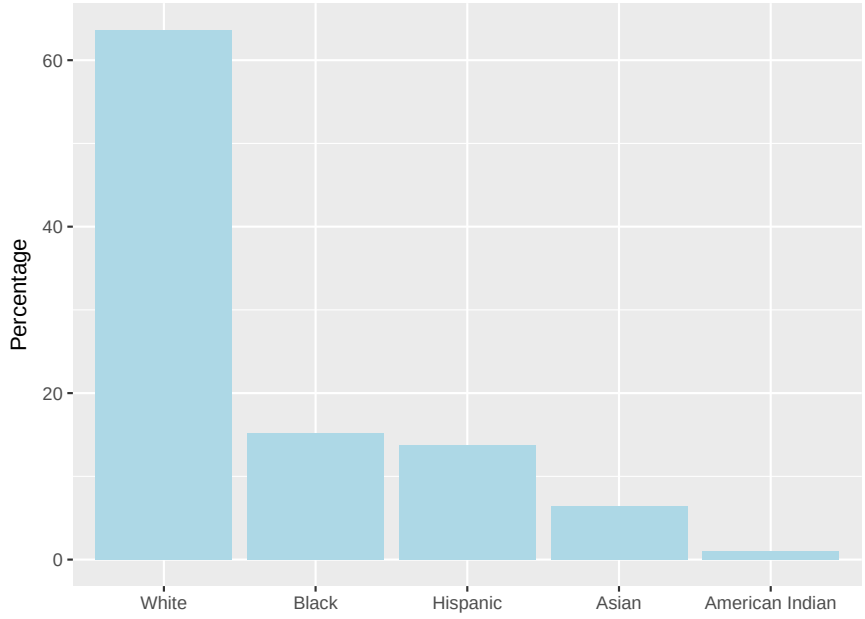
```
barchart(race.table, new.order=c(5, 3, 4, 2, 1))
```



Note: without the new.order=c(5, 3, 4, 2, 1)) argument alphabetic ordering is used, which is not very nice!

Note: to show the graph based on percentage use the argument Percent="Grand":

```
barchart(race.table, new.order=c(5, 3, 4, 2, 1), Percent="Grand")
```



Note The argument to *barchart* has to be a table, usually from a call to the R table function:

```
attach(wrinccensus)
barchart(table(Job.Level))
```

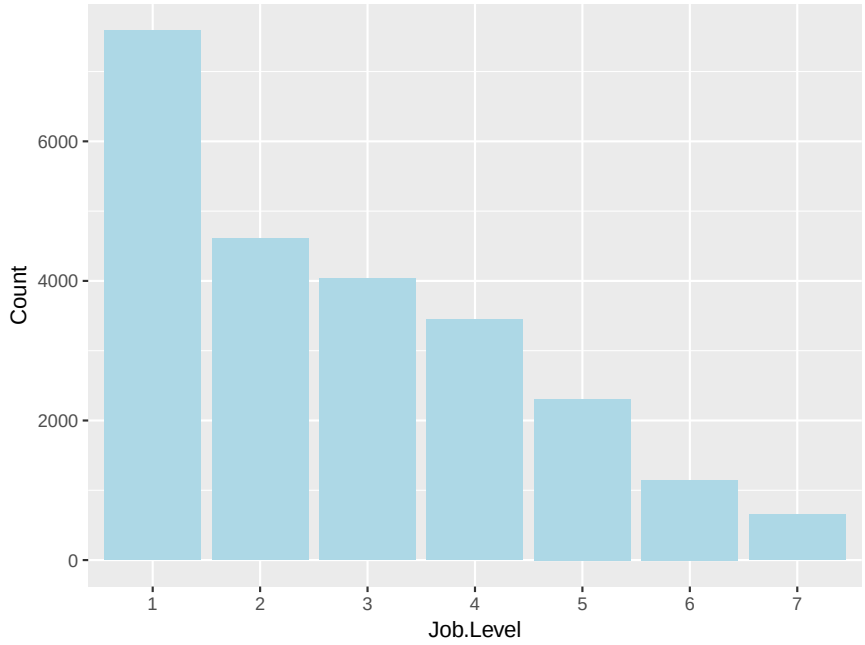


TABLE. Number and percentage of cigarette smoking-attributable conditions* among current and former smokers†, by condition — United States, 2000§

Condition	Current smokers No.	(%)	Former smokers No.	(%)	Overall No.	(%)
Chronic bronchitis	2,633,000	(49)	1,872,000	(26)	4,505,000	(35)
Emphysema	1,273,000	(24)	1,743,000	(24)	3,016,000	(24)
Heart attack	719,000	(13)	1,755,000	(24)	2,474,000	(19)
All cancer except lung cancer	358,000	(7)	1,154,000	(16)	1,512,000	(12)
Stroke	384,000	(7)	637,000	(9)	1,021,000	(8)
Lung cancer	46,000	(1)	138,000	(2)	184,000	(1)
Total¶	5,412,000	(100)	7,299,000	(100)	12,711,000	(100)

*Cigarette smoking-attributable conditions considered are stroke, heart attack, emphysema, chronic bronchitis, and cancer of the lung, bladder, mouth/pharynx, esophagus, cervix, kidney, larynx, and pancreas.
†Current smokers were defined as persons who reported smoking ≥100 cigarettes during their lifetime and who now smoke some days or every day. Former smokers were defined as persons who reported having smoked ≥100 cigarettes during their lifetime but did not smoke at the time of interview.
§Results are adjusted for age, race, sex, and state/area of residence and rounded to the nearest 1,000.
¶Numbers might not add to total because of rounding.

Figure 6:

Example This is a nice professional table from the website of the CDC (Centers for Disease Control) about the dangers of smoking:

10.1 Totals (Frequencies) vs. Percentages

Decide based on the background of the data which number is more relevant/important/interesting.

Some of the things to consider are:

If the data is a random sample from a larger population percentages are often better:

Example of 150 randomly selected people in a phone survey 85 said they would vote for candidate AA in the next election → use 57% instead.

Example in a company with 150 employees 85 said they like their job → use these numbers

For small numbers use frequencies, for large numbers use percentages When using percentages it has to be clear what the totals were.

Example an advertisement in the newspaper reads: "Almost 70% of the participants in a scientific study said they prefer Coke over Pepsi".

Now if this study had 1000 participants and about 700 of those said they like Coke better than Pepsi, that is quite impressive. On the other hand, if it had 3 participants, two of whom liked Coke (2 out of 3 = 67%, "almost" 70%) than this may not be so interesting! When comparing groups of unequal sizes, percentages are almost always necessary:

Example in a survey of the employees in a company they were asked whether they liked there current position:

	Yes	No
Male	123	88
Female	85	61

At first glance it seems that men are happier with their position than women (123 vs 88) but notice that there are more men than women in total (208 vs 149) so even if they are equally

happy we would expect more men who said yes then women. Changing to percentage gives

	Happy
Male	59.1
Female	59.0

Notice another advantage of the table with percentages: because there are only the two options yes and no, we need only the perentage of one, the other is simply 100-..

	Unappy
Male	40.9
Female	41.0

and in general the smaller a table, the better (as long as it has all the information).

These are just guidelines, there can always be exceptions if there is a good reason.

10.2 Rounding

When doing a calculation the rules are:

- if the number is used again in a later calculation, use three digits more than the data
- if the number is the final result, round to 1 digit more than the data.

Example Data is the weights of patients on two visits:

A	B	C	D
102.5	156.3	139.7	188.2
101.2	149.8	141.0	185.9

We want to find the average percentage change from visit 1 to visit 2:

Intermediate calculation: visit 1/ visit 2 *100

101.2845849802 104.3391188251 99.0780141844 101.2372243141

Data has 4 digits, so this should have 7 digits

101.2846 104.3391 99.0780 101.2372

Find mean: (101.2846+104.3391+99.0780+101.2372)/4 = 101.484725

This is final answer, round to 1 more digit than data: 101.48

By default R always uses 7 digits:

```
x
```

```
##  [1]  4.7  6.6  7.2  7.8  7.8  9.1 10.5 11.4 12.0 13.9 17.2
```

```
mean(x)
```

```
## [1] 9.836364
```

but many of the routines we use do some rounding already

```
stat.table(x)
```

```
##   Sample Size Mean Standard Deviation
## x          11  9.8                3.6
```

and you can use the ndigit argument to change how much:

```
stat.table(x, ndigit=2)
```

```
##   Sample Size Mean Standard Deviation
## x          11 9.84               3.62
```

10.3 Contingency Tables

10.3.1 Case Study: Treatment of Drug Addiction

Cocaine addiction is hard to break. Addicts need cocaine to feel any pleasure, so perhaps giving them an antidepressant drug will help. A 3 year study with 72 chronic cocaine users compared an antidepressant called desipramine with standard treatment for cocaine addiction (lithium) and a placebo. One third of the subjects chosen at random received each drug. After 3 years for each addict it was determined whether he/she was drug free or relapsed.

The data, from D.M. Barnes, "Breaking the Cycle of Addiction", Science, 241 1988).

```
head(drugaddiction)
```

```
##          Drug Relapse
## 1 Desipramine     Yes
## 2 Desipramine     Yes
## 3 Desipramine     Yes
## 4 Desipramine     Yes
## 5 Desipramine     Yes
## 6 Desipramine     Yes
```

So here for each subject we have two variables, "Drug" with values "Desipramine", "Lithium" and "Placebo", and "Relapsed" with values "Yes" and "No". Both variables are categorical.

Usually the first thing to do with this type of data is to just count each combination of values and write them up in a **contingency table**:

```
attach(drugaddiction)
table(Drug, Relapse)
```

```
##              Relapse
## Drug          No Yes
##   Desipramine 14  10
##   Lithium      6  18
##   Placebo      4  20
```

If the table is for publication you probably want to add some row and column totals:

	No	Yes	Totals
Desipramine	14	10	24
Lithium	6	18	24
Placebo	4	20	24
Totals	24	48	72

Often instead of the totals (frequencies) these tables might be based on percentages. Here, though, there are three types of percentages:

Percentages based on Grand Total:

	No	Yes	Totals
Desipramine	19.4	13.9	33.3
Lithium	8.3	25.0	33.3
Placebo	5.6	27.8	33.3
Totals	33.3	66.7	100.0

Percentages based on Row Totals:

	No	Yes	Totals
Desipramine	58.3	41.7	100
Lithium	25.0	75.0	100
Placebo	16.7	83.3	100
Totals	33.3	66.7	100

Percentages based on Column Totals:

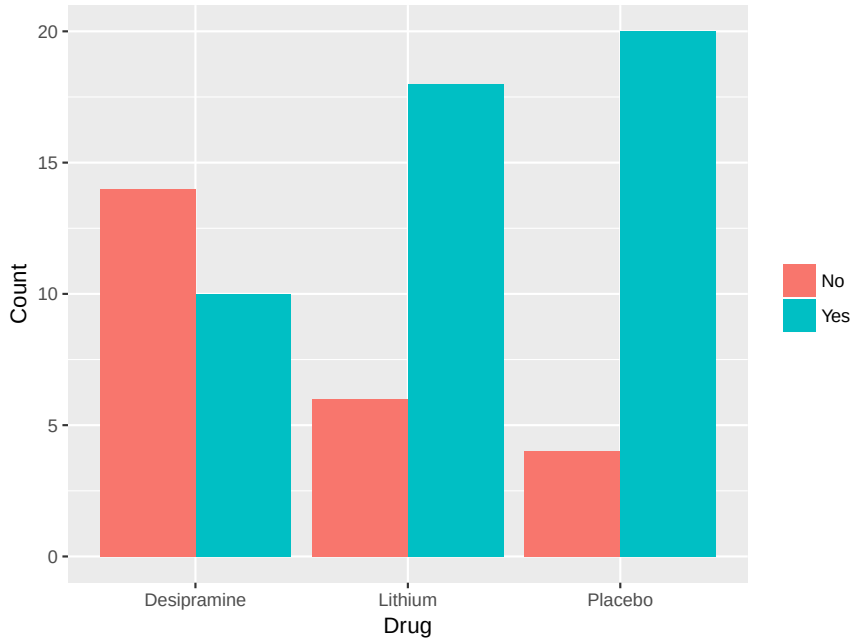
	No	Yes	Totals
Desipramine	58.3	20.8	33.3
Lithium	25.0	37.5	33.3
Placebo	16.7	41.7	33.3
Totals	100.0	100.0	100.0

Which of these 4 tables is the most interesting? It depends on the story behind the data and the result you wish to highlight. Here it is probably the third table which shows clearly that the "relapse rate" for desipramine is much smaller (41.7%) than for either Lithium (75%) or the Placebo (83.3%)

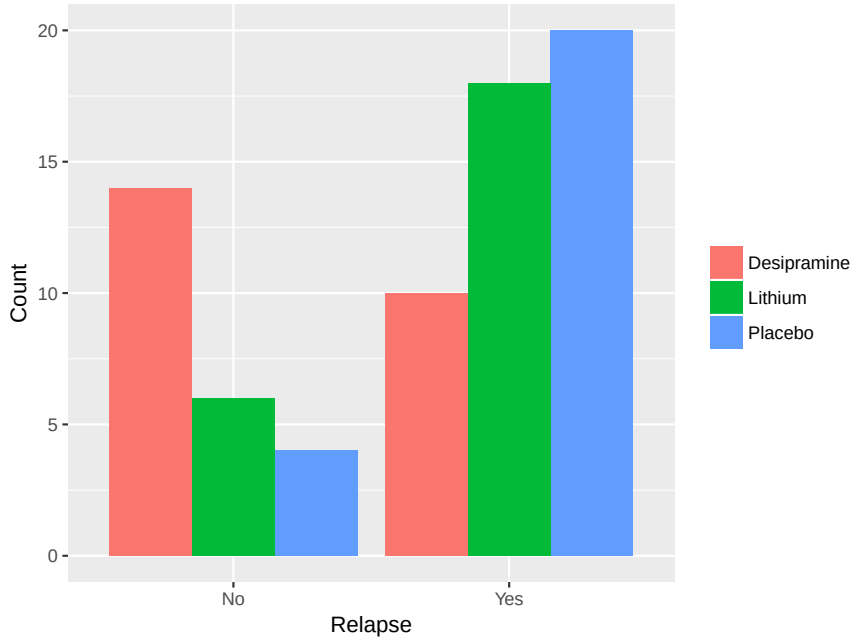
The standard graph for this data is a **multiple bar chart**. It is done with the same command as before.

There are always two depending on which way the bars are grouped together, see

```
barchart(table(Drug, Relapse))
```

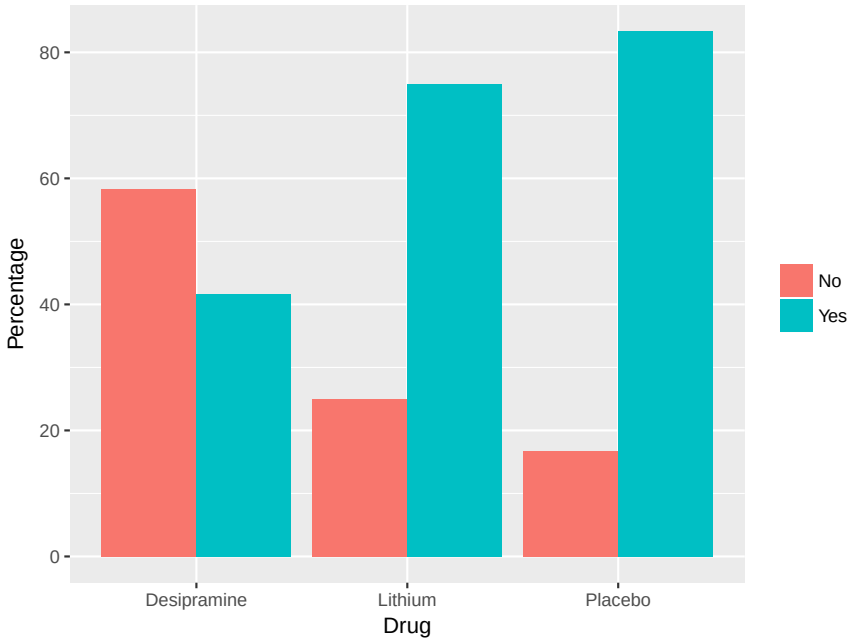



```
barchart(table(Relapse, Drug))
```



or based on percentages:

```
barchart(table(Drug, Relapse), Percent="Row" )
```



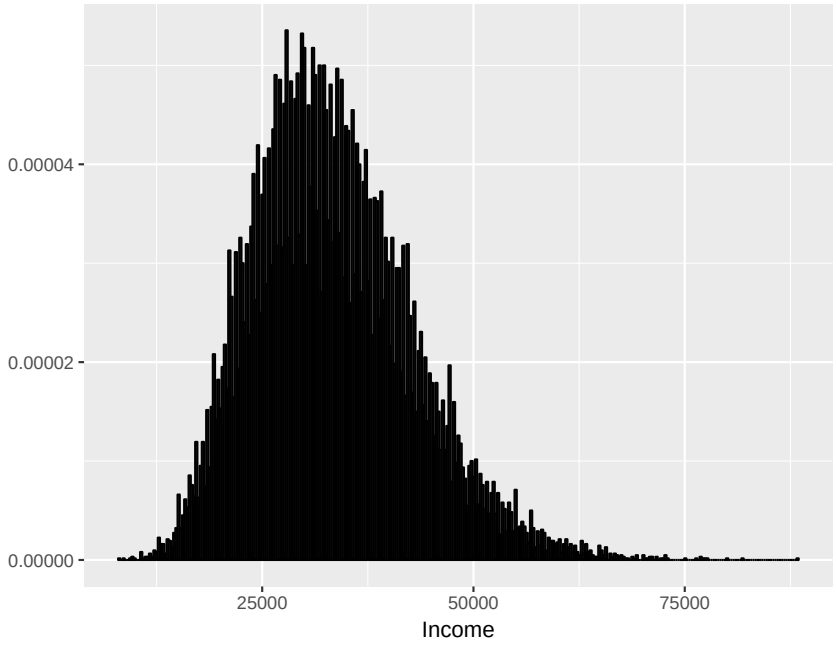
As with the tables the graphs can also show each of the three types of percentages.

#Quantitative Variables

10.4 Histograms

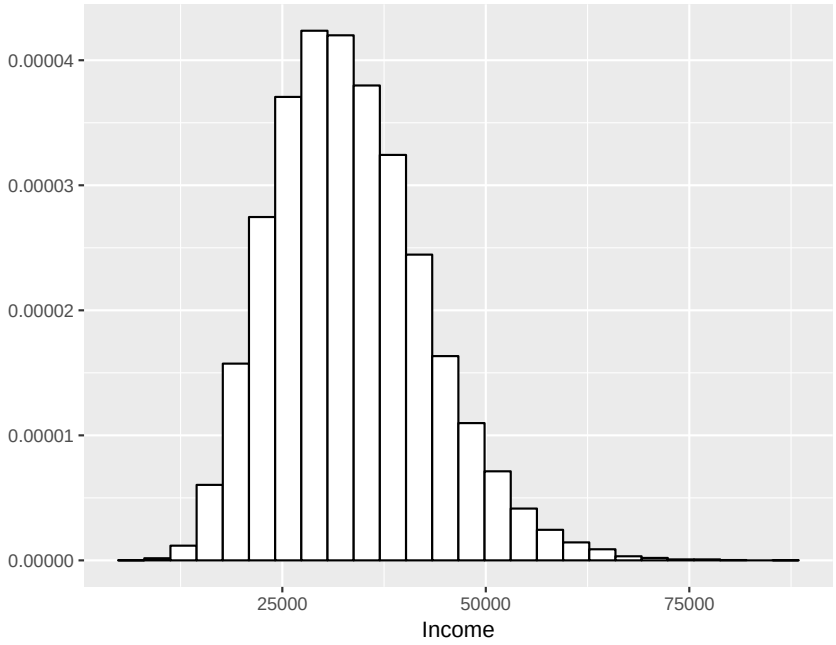
The standard graph for one quantitative variable is the histogram:

```
attach(wrinccensus)
hplot(Income)
```



It can be useful to draw a couple of histograms, with different numbers of bins:

```
hplot(Income, n=25)
```



Now that we have numbers we can do arithmetic:

10.5 Measures of Central Tendency

10.5.1 Case Study: Population Sizes of States and Puerto Rico

According to the 2010 US Census the population of Puerto Rico was 3725789. How does this compare to the rest of the US? here is the data:

```
us.population.2010
```

```
##                Alabama                 Alaska                Arizona
##                4779736                 710231                6392017
##               Arkansas             California               Colorado
##                2915918               37253956                5029196
##            Connecticut               Delaware  District of Columbia
##                3574097                 897934                 601723
##                Florida                Georgia                 Hawaii
##               18801310                9687653                1360301
##                  Idaho               Illinois                Indiana
##                1567582               12830632                6483802
##                   Iowa                 Kansas               Kentucky
##                3046355                2853118                4339367
##              Louisiana                  Maine               Maryland
##                4533372                1328361                5773552
```

```
##         Massachusetts              Michigan             Minnesota
##               6547629               9883640               5303925
##           Mississippi              Missouri               Montana
##               2967297               5988927                989415
##              Nebraska                Nevada         New Hampshire
##               1826341               2700551               1316470
##            New Jersey            New Mexico              New York
##               8791894               2059179              19378102
##        North Carolina          North Dakota                  Ohio
##               9535483                672591              11536504
##              Oklahoma                Oregon          Pennsylvania
##               3751351               3831074              12702379
##          Rhode Island        South Carolina          South Dakota
##               1052567               4625364                814180
##             Tennessee                 Texas                  Utah
##               6346105              25145561               2763885
##               Vermont              Virginia            Washington
##                625741               8001024               6724540
##         West Virginia             Wisconsin               Wyoming
##               1852994               5686986                563626
```

So how does Puerto Rico compare? One way to answer this question is to find the **average** population size:

We want just **one** number to describe **all** the numbers in the dataset.

How do we calculate an "average"?

Usual answer: **mean**

Example Three of your friends are 19, 20 and 23 years old. What is their average age?

Answer: (19+20+23)/3 = 62/3 = 20.7

Formula:

$$\overline{X} = \frac{1}{n} \sum x$$

Note $\overline{X}$ (spoken: X bar) is the standard symbol in Statistics for the sample mean

10.5.2 Case Study: Population Sizes of States and Puerto Rico

We find

$$\frac{4779736 + 710231 + ... + 563626}{51} =$$

$$\frac{308745538}{51} = 6053834$$

PR had a population of 3725789, so ours is lower than average.

Now in R we have the command *mean*:

Figure 7:

```
mean(us.population.2010)
```

```
## [1] 6053834
```

Note the mean command does not do any rounding. According to our rules we should round to one digit behind the decimal. Except that often for large numbers we actually round the other way, so here I might end up using 6,054,000!

10.5.3 Case Study: Babe Ruth's Homeruns

Many still consider Babe Ruth the greatest baseball player of all time. In 1919 he moved to the New York Yankees, where he played until 1934. Here are the number of homeruns he hit in those years

```
babe
```

```
##    Year Homeruns
## 1  1920       54
## 2  1921       59
## 3  1922       35
## 4  1923       41
## 5  1924       46
## 6  1925       25
## 7  1926       47
## 8  1927       60
## 9  1928       54
## 10 1929       46
## 11 1930       49
## 12 1931       46
## 13 1932       41
## 14 1933       34
## 15 1934       22
```

what was his homerun average while with the Yankees?

$\overline{X} = (54 + +59 + .. + 22)/15 = 659/15 = 43.9$

of course we can use R:

```
attach(babe)
mean(Homeruns)
```

```
## [1] 43.93333
```

again, you should round the answer, here 43.9.

Advice

The most important thing you can do in this class (and, more importantly, in life!) after you did some calculation is to ask yourself:

Does my answer make sense?

If you find that the average age of your three friends in the example above is 507.9, you have to know that this answer **is wrong**.

Example Which of the following are obviously **not** correct for the mean of Babe Ruths homeruns, and why?

a. 43.2
b. 17.9
c. -45.6
d. 49.5
e. 59.0
f. 35.4

There are other methods for computing an "average", though. For example:

Median: the observation "in the middle" of the **ordered** data set:

22, 25, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60

What if the Babe had left the Yankees a year earlier?

25, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60

Median = (46+46)/2 = 46

Using R:

```
median(Homeruns)
```

```
## [1] 46
```

Let's find the mean and median of the salaries of the WRInc employees:

```
mean(Income)
```

```
## [1] 33373.13
```

```
median(Income)
```

```
## [1] 32400
```

Here there is a difference of almost $1000 between the mean and the median. So which one is the right "average"?

10.6 Mean vs. Median

10.6.1 Case Study: Weights of Mammals

Weights of the bodies of 62 mammals (in kg)

```
brainsize
```

```
##                       Animal body.wt.kg brain.wt.g
## 1            African elephant   6654.000    5712.00
## 2  African giant pouched rat      1.000       6.60
## 3                  Arctic Fox      3.385      44.50
## 4     Arctic ground squirrel      0.920       5.70
## 5              Asian elephant   2547.000    4603.00
## 6                      Baboon     10.550     179.50
## 7               Big brown bat      0.023       0.30
## 8             Brazilian tapir    160.000     169.00
## 9                         Cat      3.300      25.60
## 10                Chimpanzee     52.160     440.00
## 11                Chinchilla      0.425       6.40
## 12                       Cow    465.000     423.00
## 13           Desert hedgehog      0.550       2.40
## 14                    Donkey    187.100     419.00
## 15     Eastern American mole      0.075       1.20
## 16                   Echidna      3.000      25.00
## 17         European hedgehog      0.785       3.50
## 18                    Galago      0.200       5.00
## 19                     Genet      1.410      17.50
## 20           Giant armadillo     60.000      81.00
## 21                   Giraffe    529.000     680.00
## 22                      Goat     27.660     115.00
## 23            Golden hamster      0.120       1.00
## 24                   Gorilla    207.000     406.00
## 25                 Gray seal     85.000     325.00
## 26                 Gray wolf     36.330     119.50
## 27           Ground squirrel      0.101       4.00
## 28                Guinea pig      1.040       5.50
## 29                     Horse    521.000     655.00
## 30                    Jaguar    100.000     157.00
## 31                  Kangaroo     35.000      56.00
## 32 Lesser short-tailed shrew      0.005       0.14
## 33          Little brown bat      0.010       0.25
## 34                       Man     62.000    1320.00
## 35                  Mole rat      0.122       3.00
## 36           Mountain beaver      1.350       8.10
## 37                     Mouse      0.023       0.40
## 38                Musk shrew      0.048       0.33
```

```
## 39       N. American opossum      1.700       6.30
## 40     Nine-banded armadillo      3.500      10.80
## 41                     Okapi    250.000     490.00
## 42                Owl monkey      0.480      15.50
## 43              Patas monkey     10.000     115.00
## 44                Phanlanger      1.620      11.40
## 45                       Pig    192.000     180.00
## 46                    Rabbit      2.500      12.10
## 47                   Raccoon      4.288      39.20
## 48                       Rat      0.280       1.90
## 49                   Red fox      4.235      50.40
## 50             Rhesus monkey      6.800     179.00
## 51    Rock hyrax (Hetero. b)      0.750      12.30
## 52 Rock hyrax (Procavia hab)      3.600      21.00
## 53                  Roe deer     83.000      98.20
## 54                     Sheep     55.500     175.00
## 55                Slow loris      1.400      12.50
## 56           Star nosed mole      0.060       1.00
## 57                    Tenrec      0.900       2.60
## 58                Tree hyrax      2.000      12.30
## 59                Tree shrew      0.104       2.50
## 60                    Vervet      4.190      58.00
## 61             Water opossum      3.500       3.90
## 62     Yellow-bellied marmot      4.050      17.00
```

```
attach(brainsize)
mean(body.wt.kg)
```

```
## [1] 199.8895
```

```
median(body.wt.kg)
```

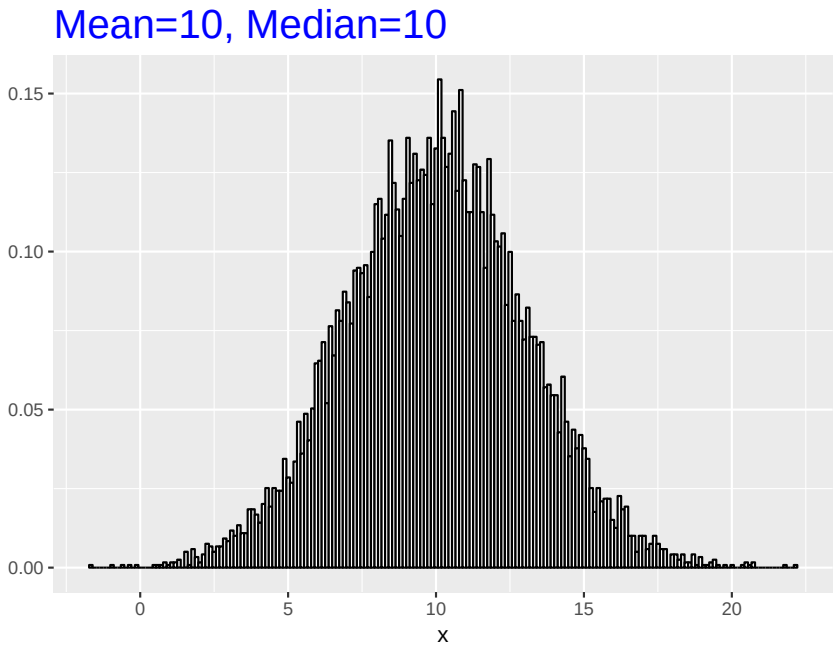
```
## [1] 3.3425
```

here we find Mean=199.9 and Median=3.3!!!

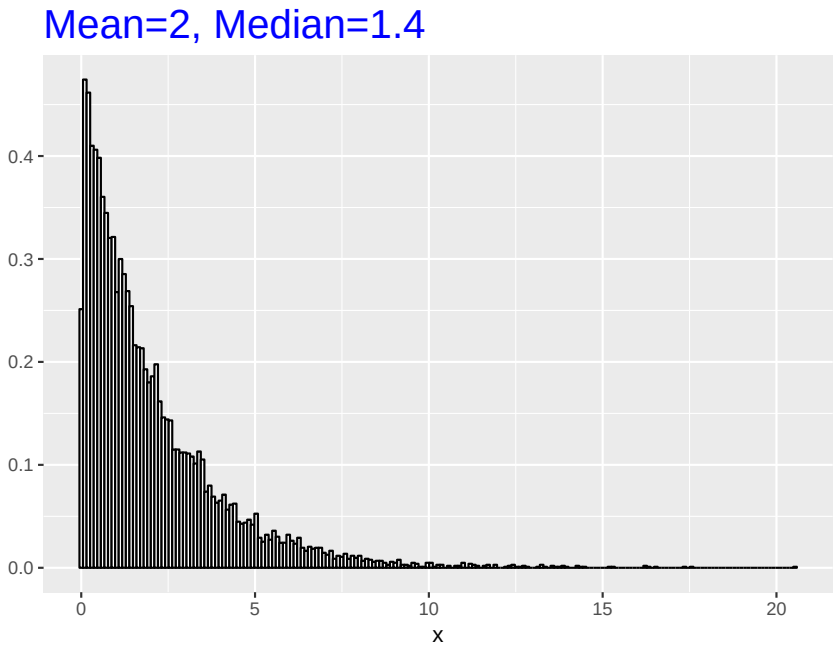
So what is the AVERAGE???

The reason for this huge difference is obvious: there are two mammals that are much larger than the rest, the African and the Asian elephants. Observations like these that are "unusual" are often called **outliers**.

Often the mean and the median are very similar if a histogram of the data is **symmetric**, that is it looks the same from right to left as from left to right:



compared to for example the following, which is called **skewed to the right**:



Whether the mean or the median is a better measure of "average" is NOT a simple question. It often depends on the question asked:

Example 1: what is the weight of a "typical" mammal? Median = 3.34kg

Example 2: say we randomly choose 50 mammals. These are to be transported by ship. How large a ship do we need (what carrying capacity?)

Now if we use the median we find $50 \times 3.3 = 165$ kg, but if one of the 50 animals is an elephant we are sunk (literally!) So we should use

estimated total weight = 50 × mean weight = 50 × 199.9 = 9995.

Example The government has just released the data for a study of Puerto Rican households. One of the variables was **household income**

- you read in El Nuevo Dia that the mean income in PR is $23100
- you hear on the local news that the median income in PR is $20400

Which of these number is better?

Without any explanation what the number will be used for this question has no answer, both the mean and the median are perfectly good ways to calculate an "average"

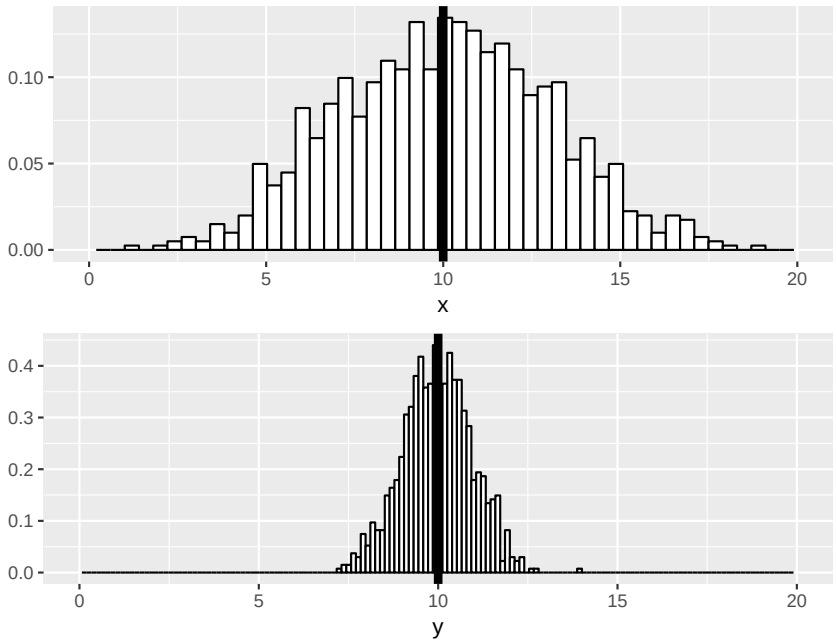
Misuse of Statistics: Mean vs. Median

Say the owner of a McDonalds wants to compute the "average" hourly wage for the people working there. Do you think she will use the mean or the median? What if it is the Union that wants to find the "average"?

10.7 Measures of Variability

A statistician is standing with one foot in an icebucket and the other foot in a burning fire. He says: on average I feel fine.

A "measure of central tendency" is a good start for describing a set of numbers, but it does not tell the whole story. Consider the the two examples in the next graph:



Here we have two datasets, both have a mean of 10 but they are clearly very different, with different "spreads". We would like to have some way to measure this "spread-out-ness".

Range: the first is the range of the observations, defined as Largest-Smallest observation.

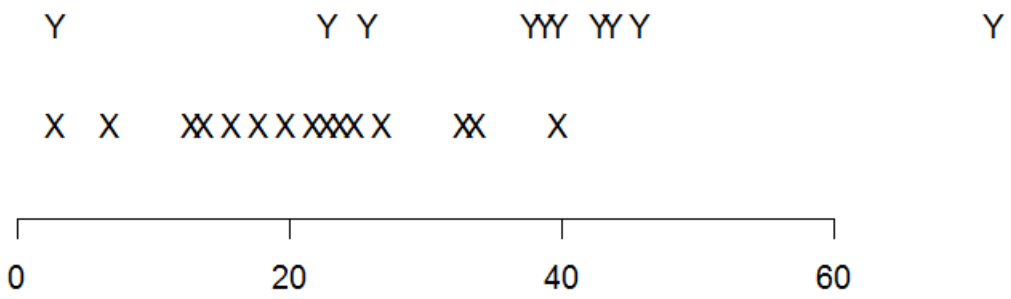


Figure 8:

Example in the graph above the x data seems to go from about 0 to about 19, so the range is 19-0=19. The y data seems to go from about 7 to about 13, so the range is 13-7=6.

Example For Babe Ruth Homeruns we find range = 60-22 = 38.

Note Some textbooks and/or computer programs define the range as the pair of numbers (smallest, largest).

10.7.1 Standard Deviation

This is the most important measure of variation, so it is very important that you learn what it is and what it is telling you.

Consider the following example. Say we have done a survey. We went to a number of locations, and among other things we asked people their age. We found:

Mall: 3 7 13 14 16 18 20 22 23 24 25 27 33 34 40

Plaza: 3 23 26 38 39 40 43 44 46 72

Let's look at the data with a graph:

Now it seems the variation of the Y's is a bit larger than the variation of the X's. But also the mean of Y's and X's are different. If we want to concentrate on the variation we can eliminate the differences of the means by subtracting them from each observation:

$\overline{X} = (3 + 7 + .. + 40)/15 = 319/15 = 21.27$

$\overline{Y} = (3 + 23 + .. + 72)/10 = 374/10 = 37.40$

and with this we get:

$x - \overline{X}$: -18.27 -14.27 -8.27 -7.27 -5.27 -3.27 -1.27 0.73 1.73 2.73 3.73 5.73 11.73 12.73 18.73

$y - \overline{Y}$: -34.4 -14.4 -11.4 0.6 1.6 2.6 5.6 6.6 8.6 34.6

Let's look at these numbers with a graph again:

and it is now more obvious that the variation of the Y's is little bit larger than those of the X's.

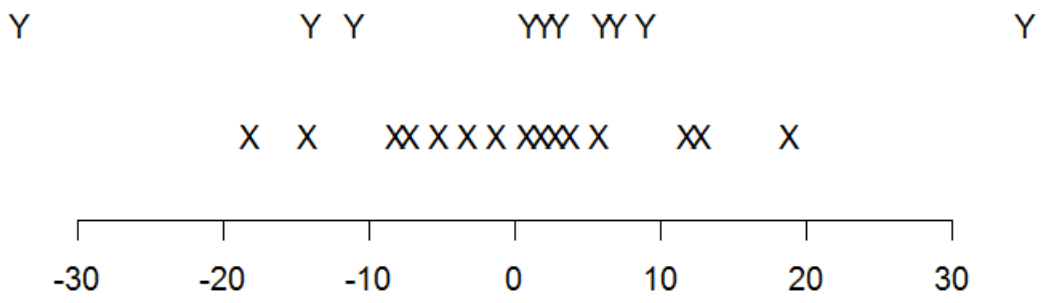


Figure 9:

Notice that the mean of the x-$\overline{X}$ numbers (and of course also the y-$\overline{Y}$ numbers) is now 0.

Because these new numbers are centered at 0, a larger variation means "farther away from 0", so how about as a measure of variation the "mean of x-$\overline{X}$", that is

$$\frac{1}{n}\sum\left(x-\overline{X}\right)$$

But no, that won't work because

$$\frac{1}{n}\sum\left(x-\overline{X}\right)=0$$

always! (Not obvious? Try it out!)

The problem is that some (actually about half) of the x-$\overline{X}$ are negative, the other are positive, so in the sum they just cancel out.

So somehow we need to get rid of the - signs. One way to do that would be to use absolute values: |x-$\overline{X}$|. It turns out, though, that for some mathematical reasons it is better to use squares:

$$\frac{1}{n}\sum\left(x-\overline{X}\right)^2$$

Another change from the "obvious" is that we should devide this by n-1 instead of n, and with this we have the famous formula for the Variance:

$$s^2=\frac{1}{n-1}\sum\left(x-\overline{X}\right)^2$$

So in essence the variance is the mean distance from the sample mean (squared).

One problem with having squared everything is that now the units are in "squares". For example, if our data is the height of people, the variance is height^2. Usually we want everything in the same units, and this is easy to do by taking square roots, and so we finally have the formula for the Standard Deviation:

$$s = \sqrt{\frac{1}{n-1}\sum\left(x - \overline{X}\right)^2}$$

The R command to find a standard deviation is *sd*.

10.7.2 Case Study: Babe Ruth Homeruns

what was the standard deviation of his homeruns?

```
sd(Homeruns)
```

```
## [1] 11.24701
```

and again we should round the answer to 11.2.

Now we have two ways to measure the "spread-out-ness", range and standard deviation. Unfortunately the two don't quite work together. For example we have found range=38 and s=11.2 for Babe Ruths Homeruns. As a rule of thumb we often have

s is close to range/4

Example Babe Ruth's Homeruns: range/4 = 38/4 = 9.5, s = 11.2.

10.7.3 Case Study: Weights of Mammals

Weights of the bodies of 62 mammals (in kg)

We saw before that a few outliers can have a HUGE effect on the mean. The same is true (actually even worse!) for the standard deviation:

```
sd(body.wt.kg)
```

```
## [1] 898.971
```

```
sd(body.wt.kg[body.wt.kg<1000])
```

```
## [1] 119.4329
```

If we want to ignore the outliers in the calculation of an average, we can use the median. What can we do if we want to find a measure of variation?

We will see in a little bit!

10.8 Population vs Sample

As we discussed earlier, real life research questions are about populations. If we can do a census and get all the information we can find the number (a **Parameter**) we are looking for. In real life, though, that is very rarely possible. So the next best thing we can do is get a sample, and find the corresponding number (a **Statistic**).

	Population	Sample
Mean	$\mu = \frac{1}{N}\sum x$	$\overline{X} = \frac{1}{n}\sum x$
Std. Dev.	$\sigma = \sqrt{\frac{1}{N}\sum(x-\mu)^2}$	$s = \sqrt{\frac{1}{n-1}\sum(x-\overline{X})^2}$

Figure 10:

So the mean and standard deviation come in two forms, as a parameter and as a statistic. Sometimes the formulas for the two are the same, sometimes there is a slight difference. Also, when talking about parameters we usually use greek letters.

Note that the formula for the population mean is exactly the same as for the sample mean, only we use N (population size) instead of n (sample size) and μ instead of $\overline{X}$ (parameter instead of statistic). The formula for the population standard deviation is a little different, we devide by N instead of N-1. The reason the sample standard deviation uses n-1 is that if we used n the answers would come out a little to small.

10.9 z score

in the discussion of the standard deviation we saw that if we want to compare two sets of numbers, subtracting the mean is a good idea because then the datsets are both centered at 0. Now we go a step further and also devide by the standard deviation, getting to the **z scores**:

$$z = \frac{x - \overline{X}}{s}$$

the idea here is that no matter what scale the original data is on, the z scores are of the same "size" and can therefore be compared directly.

Note z scores are usually rounded to three digits behind the decimal.

Example say you have taken two exams. In exam 1 you got 13 out of 20 points and in exam 2 you had a 58 out of 100 points. In which exam did you do better?

At first glance you might say exam 1, because if we want to rescale exam 1 to also have a total of 100 we need to multiply by 5 (20*5 = 100), so your "equivalent" score is

13*5 = 65 > 58

But "doing better" often means doing better with respect to how everyone else did. So let's say

$\overline{X}_1 = 10.1$, $s_1 = 4.5$

$\overline{X}_2 = 45.7$, $s_2 = 16.5$

Let's find your respective z scores:

$$z_1 = \frac{13 - 10.1}{4.5} = 0.64 z_2 = \frac{58 - 45.7}{16.5} = 0.745$$

and because your z score in exam 2 was higher, that is the one you did better.

Clearly if x is close to the mean, the z score will be 0. It turns out that often z is somewhere between -2 and +2. Both of your z scores are a bit larger than 0 but not much, so they probably are B's!

10.9.1 Case Study: Population Sizes of States and Puerto Rico

What is the z score of PR's population of 3725789?

We have

```
mean(us.population.2010)
```

```
## [1] 6053834
```

```
sd(us.population.2010)
```

```
## [1] 6823984
```

so

$$\overline{X} = 6053834$$
$$s = 6823984$$
$$z = \frac{3725789 - 6053834}{6823984} = -0.341$$

and so PR's z score is -0.341.

Of course we can use R as well:

```
(3725789 - mean(us.population.2010))/sd(us.population.2010)
```

```
## [1] -0.3411563
```

which we should round to $z = -0.341$

10.10 Empirical Rule

Above we learned the following:

- If we have the dataset, how do we calculate the mean and the standard deviation?

Now we will look at the following question:

- If we know **only** the mean and the standard deviation, what do they tell us about the dataset, or more precisely, what do they tell us about an individual observation in the dataset?

Example You read in the newspaper about a study on the age when a criminal committed his first crime. They found that the mean age was 18.3 with a standard deviation of 2.6 years. What is this telling you?

The information "mean age was 18.3", or with our notation $\overline{X} = 18.3$, is pretty easy to understand - somewhere around age 18 people start to commit crimes. But what about "with a standard deviation of 2.6 years"?

For this we can use the **empirical rule**

if a data set has a bell-shaped histogram, then 95% of the observations fall into the interval

$$\left(\overline{X} - 2s,\, \overline{X} + 2s\right)$$

Notice the connection to the z scores. We previously said the z score is usually between -2 and 2, so z=2 would indicate a score almost at the maximum. But then

$$\begin{aligned} 2 = z &= \frac{x - \overline{X}}{s} \\ 2s &= x - \overline{X} \\ x &= \overline{X} + 2s \end{aligned}$$

Example Back to the example. We have $\overline{X}$=18.3 and s=2.6, so

```
18.3 - 2*2.6
```

```
## [1] 13.1
```

```
18.3 + 2*2.6
```

```
## [1] 23.5
```

so 95% of the criminals are between 13.1 and 23.5 years old when they first commit a crime.

Knowing the mean and the standard deviation and using the empirical rule makes it possible to make a guess about the size of the actual observations.

Above we said that s is often close to range/4. The reason for this is explained by the empirical rule: $(\overline{X} - 2s, \overline{X} + 2s)$ contains 95% of the data, so $\overline{X} - 2s$ should be close to the smallest observation and $\overline{X} + 2s$ should be close to the largest observation. So

range = largest-smallest is close to

$$(\overline{X} + 2s) - (\overline{X} - 2s) = 4s$$

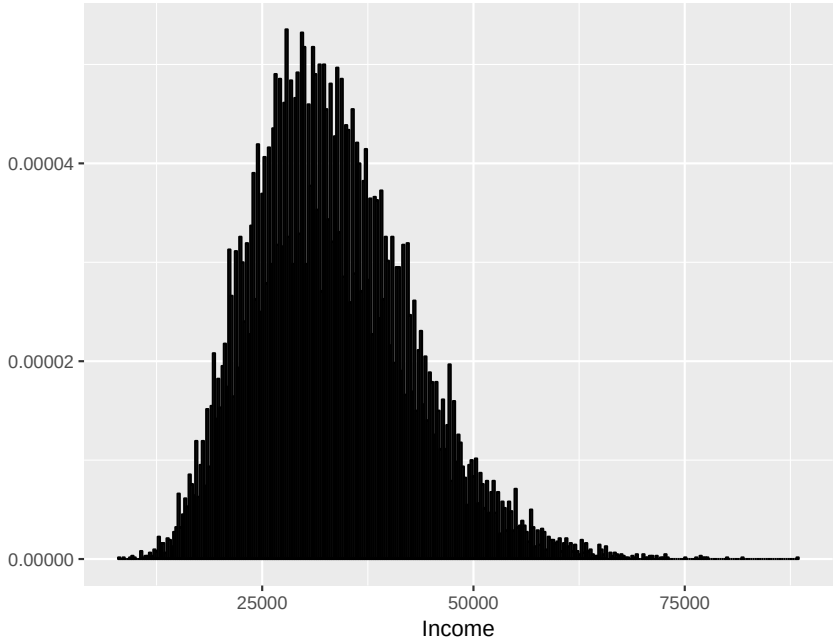
or

s is close to range/4

Example Again back to the example with the criminals. For the empirical rule to work the data should have a bell shaped histogram. Do you think this is true for this example?

Example let's check whether the empirical rule holds for the income data of the WRInc dataset.

```
hplot(Income)
```



histogram is reasonably bell-shaped

What does the empirical rule say?

```
mean(Income) - 2*sd(Income)
```

```
## [1] 14524.37
```

```
mean(Income) + 2*sd(Income)
```

```
## [1] 52221.9
```

so we should have about 95% of the incomes between $14524 and $52221.

Let's check:

```
sum(Income>14524 & Income<52221)/length(Income)*100
```

```
## [1] 95.92283
```

looks about right!

11 Percentiles and Boxplots

11.1 Percentiles (Measures of Location)

11.1.1 Case Study: Population Sizes of States and Puerto Rico

According to the 2010 US Census the population of Puerto Rico was 3725789.

```
us.population.2010
```

```
##                Alabama                 Alaska                Arizona
##                4779736                 710231                6392017
##               Arkansas             California               Colorado
##                2915918               37253956                5029196
##            Connecticut               Delaware  District of Columbia
##                3574097                 897934                 601723
##                Florida                Georgia                 Hawaii
##               18801310                9687653                1360301
##                  Idaho               Illinois                Indiana
##                1567582               12830632                6483802
##                   Iowa                 Kansas               Kentucky
##                3046355                2853118                4339367
##              Louisiana                  Maine               Maryland
##                4533372                1328361                5773552
##          Massachusetts               Michigan              Minnesota
##                6547629                9883640                5303925
##            Mississippi               Missouri                Montana
##                2967297                5988927                 989415
##               Nebraska                 Nevada          New Hampshire
##                1826341                2700551                1316470
##             New Jersey             New Mexico               New York
##                8791894                2059179               19378102
##         North Carolina           North Dakota                   Ohio
##                9535483                 672591               11536504
##               Oklahoma                 Oregon           Pennsylvania
##                3751351                3831074               12702379
##           Rhode Island         South Carolina           South Dakota
##                1052567                4625364                 814180
##              Tennessee                  Texas                   Utah
##                6346105               25145561                2763885
##                Vermont                Virginia             Washington
##                 625741                8001024                6724540
##           West Virginia              Wisconsin                Wyoming
##                1852994                5686986                 563626
```

Previously we found the mean population size for the states to be 6053834, so that PR's population was lower than average. Here is a different way to compare PR to the states: If we order them from smallest to largest and add PR we find:

563626 601723 625741 672591 710231 814180 897934 989415 1052567 1316470 1328361 1360301 1567582 1826341 1852994 2059179 2700551 2763885 2853118 2915918 2967297 3046355 3574097 3725789 3751351 3831074 4339367 4533372 4625364 4779736 5029196 5303925 5686986 5773552 5988927 6346105 6392017 6483802 6547629 6724540 8001024 8791894 9535483 9687653 9883640 11536504 12702379 12830632 18801310 19378102 25145561 37253956

so PR's population is the 24th. So of the 52 numbers 23 are smaller than PR's, 23 out of 52 is 23/52*100% = 44.2%. We say that

PR is at the 44.2nd percentile.

Definition:

The p^{th} percentile of a data set is the value that has at most p% of the data below it and at most $(100 - p)$% above it.

Example consider the first employee in our WRInc dataset. She has an income of $22800.

```
attach(wrinccensus)
sum(Income<22800)
```

```
## [1] 2880
```

shows that there are 2880 employees with a lower income. 2880 out of 23791 means she is at the $2880/23791 \times 100 = 12.1^{st}$ percentile.

So 12.1% have an income less than her and $(100 - 12.1)\% = 87.9\%$ have an income higher than her.

The R command to find a percentile is *quantile*.

11.1.2 Case Study: Babe Ruth's Homeruns

Find the 67^{th} percentile of the data

```
attach(babe)
quantile(Homeruns, 0.67)
```

```
##   67%
## 47.76
```

11.1.3 Case Study: WRinc

Find the 10^{th} and the 90^{th} percentile of the WRinc incomes:

```
quantile(Income, c(0.1, 0.9))
```

```
##   10%   90%
## 22000 45900
```

11.2 Quartiles, Five-Number Summary and IQR

The quartiles of a data set are defined as

1st quartile Q_1 = 25th percentile

3rd quartile Q_3 = 75th percentile

Using these we can also find the **Interquartile Range**

IQR = Q_3 - Q_1

and the **five number summary**:

Minimum | Q_1 | Median | Q_3 | Maximum

11.2.1 Case Study: Babe Ruth's Homeruns

```
fivenumber(Homeruns)
```

```
##  Minimum Q1 Median   Q3 Maximum
##       22 38     46 51.5      60
## IQR =  13.5
```

Example Find the 5-number-summary and the IQR of the incomes of WRInc:

```
fivenumber(Income)
```

```
##  Minimum    Q1 Median    Q3 Maximum
##     8000 26600  32400 39200   88400
## IQR =  12600
```

What is the meaning of these percentiles?

- $Q_1 = P_{25}$ = \$26600, so 25% (or 1 in 4) of the employees make **less** than \$26600.
- Median = \$32400, so half of the employees make **less** than \$32400, half make **more**.
- $Q_3 = P_{75}$ = \$39200, so 25% (or 1 in 4) of the employees make **more** than \$39200.

What is the meaning of IQR? Actually it is a 3rd way to calculate a measure of variation, after the range and the standard deviation.

Example The standard deviation of the incomes is s = 9424. Now IQR = 12600.

Now we have several formulas (methods) for finding an "average" (mean, median) and a variation (range/4, s, IQR). How do you decide which to use?

- Use the range only if you can't find either of the other two, for example if you only know the smallest and the largest observation, or if you have to do a quick calculation in your head.
- decide whether to use mean or median as we discussed before.
- If you use the mean, also use the standard deviation. If you use the median, use IQR.

11.2.2 Case Study: Weights of Mammals

Weights of the bodies of 62 mammals (in kg)

We saw before that a few outliers can have a HUGE effect on the standard deviation:

```
attach(brainsize)
sd(body.wt.kg)
```

```
## [1] 898.971
```

```
sd(body.wt.kg[body.wt.kg<1000])
```

```
## [1] 119.4329
```

If we want to ignore the outliers we can use the median. But then we should also ignore the outliers in the calculation of a measure of variation, which happens if we use the IQR:

```
IQR(body.wt.kg)
```

```
## [1] 54.065
```

```
IQR(body.wt.kg[body.wt.kg<1000])
```

```
## [1] 39.755
```

11.3 Boxplot

From the five number summary we can construct another graph for quantitative data, the boxplot:

11.3.1 Case Study: Babe Ruth's Homeruns

```
bplot(Homeruns)
```

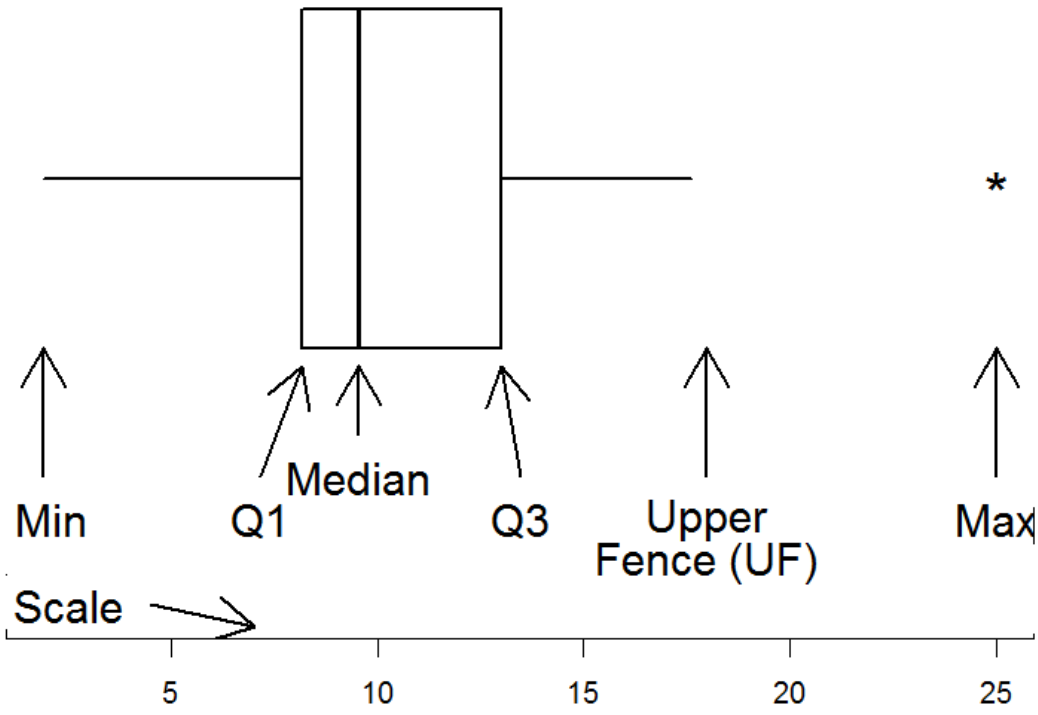
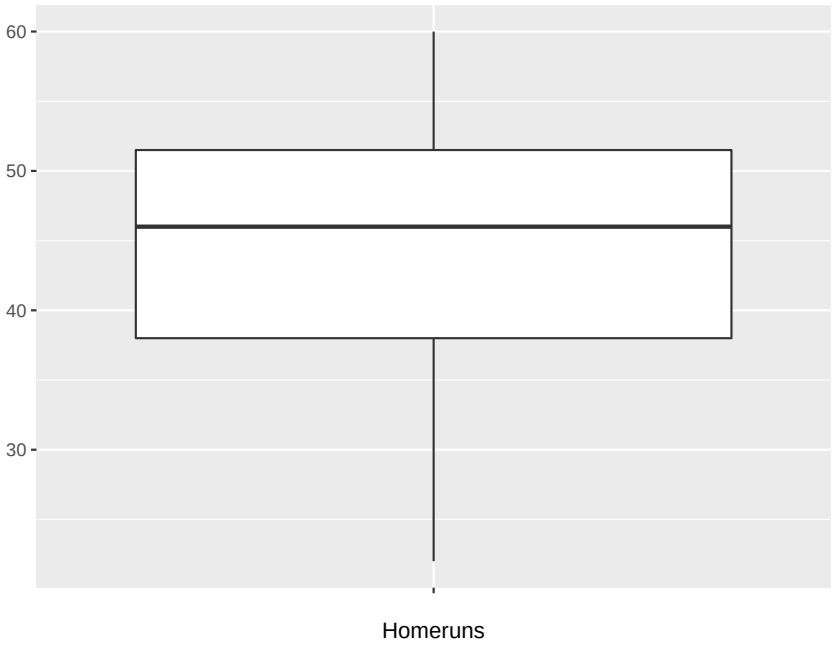


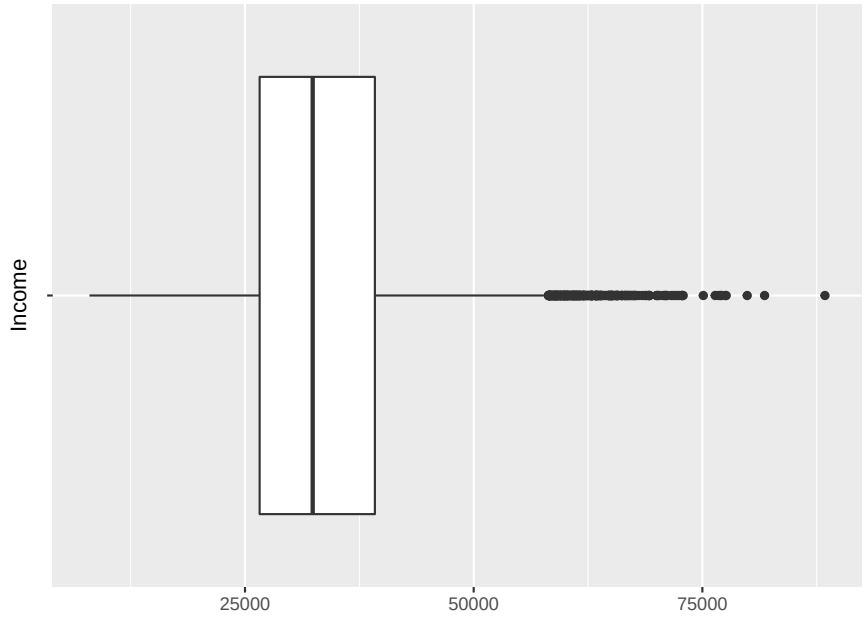
Figure 11:



Note that by its definition the box contains 50% of the data

11.3.2 Case Study: Wrinccensus

```
bplot(Income, orientation="Horizontal")
```



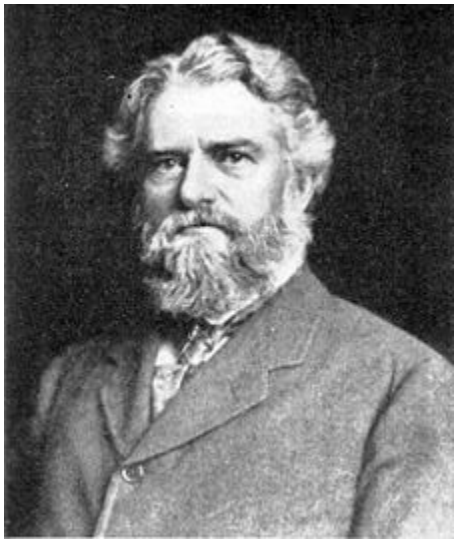
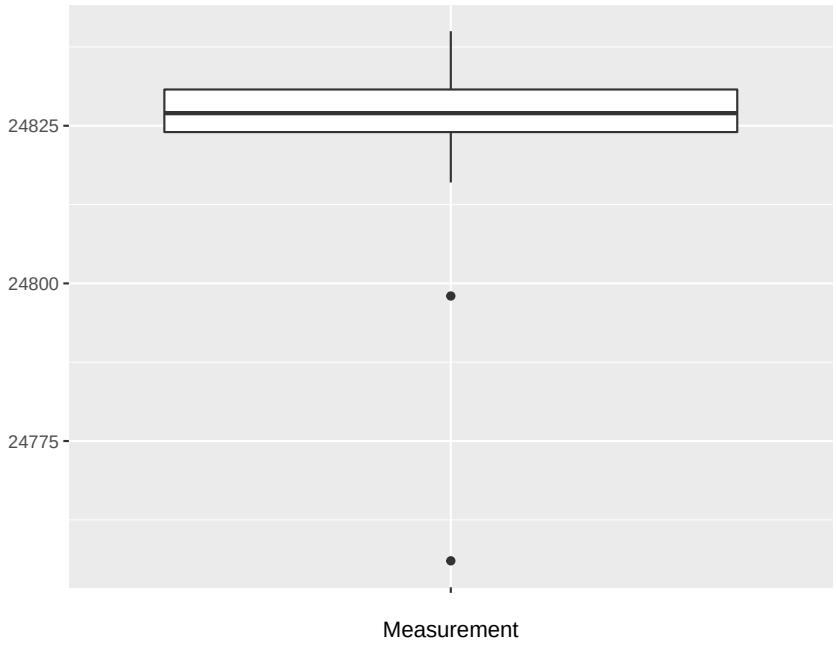


Figure 12:

11.3.3 Case Study: Simon Newcomb's Measurements of the Speed of Light

Simon Newcomb made a series of measurements of the speed of light between July and September 1880. He measured the time in seconds that a light signal took to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of 7400m. His first measurement was 0.000024828 seconds, or 24,828 nanoseconds (10^9 nanoseconds = 1 second).

```
attach(newcomb)
bplot(Measurement)
```



Observations marked with a dot are (possible) outliers, that is "unusual" observations.

One of the effects of outliers is that we then often get a difference between the mean and the median:

```
mean(Measurement)
```

```
## [1] 24826.21
```

```
median(Measurement)
```

```
## [1] 24827
```

How did Newcomb handle this problem? After careful consideration he dropped the 24756 and found the mean of the other 65 observations (24827.3), an answer much closer to the median than the original mean. Eliminating data from the analysis is something that should be done with great care! At very least one needs to be honest about this and discuss the issue, just like Newcomb.

Notice also that the effect of outliers is even greater on the standard deviation, but not so much on the IQR

```
sd(Measurement)
```

```
## [1] 10.74532
```

```
sd(Measurement[Measurement>24756])
```

```
## [1] 6.249308
```

```
IQR(Measurement)
```

```
## [1] 6.75
```

```
IQR(Measurement[Measurement>24756])
```

```
## [1] 7
```

The handling of outliers is one of the more difficult and dangerous jobs in Statistics:

11.3.4 Case Study: Ozone Hole over South Pole

In 1985 British scientists reported a hole in the ozone layer of the earth's atmosphere over the South Pole.

This news is disturbing, because ozone protects us from cancer-causing ultraviolet radiation. The British report was at first disregarded, because it was based on ground instruments looking up.

More comprehensive observations from satellite instruments looking down had shown nothing unusual.

Then, examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software used to analyze the data had automatically set these values aside as suspicious outliers.

Readings dating back to 1979 were reanalyzed and showed a large and growing hole in the ozone layer that is unexplained and considered dangerous.

Computers analyzing large volumes of data are often programmed to suppress outliers as protection against errors in the data. As the example of the hole in the ozone illustrates, suppressing an outlier without investigating it can conceal valuable information.

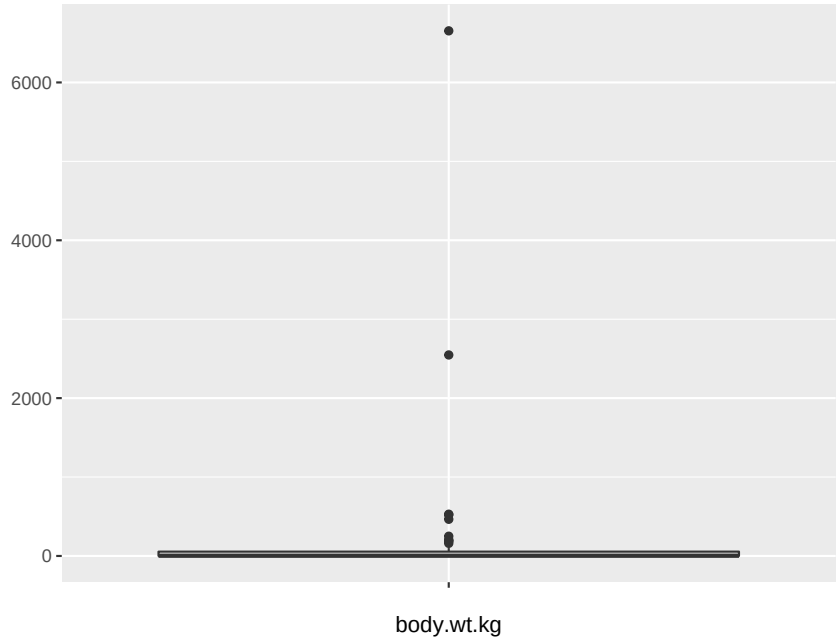
(From More and McCabe)

Sometimes it is the outliers that are the most interesting feature of a dataset!

11.3.5 Case Study: Weights of Mammals

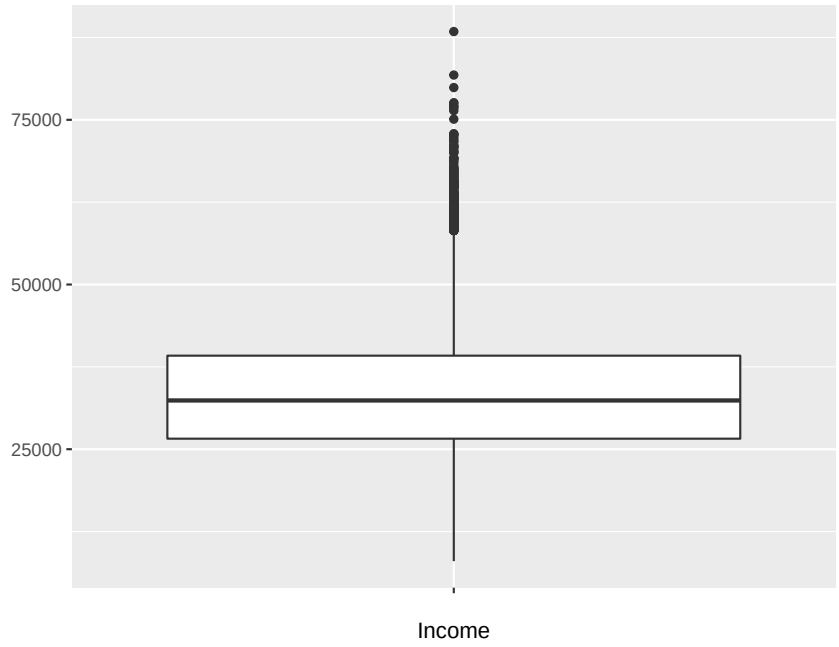
Weights of the bodies of 62 mammals (in kg)

```
bplot(body.wt.kg)
```

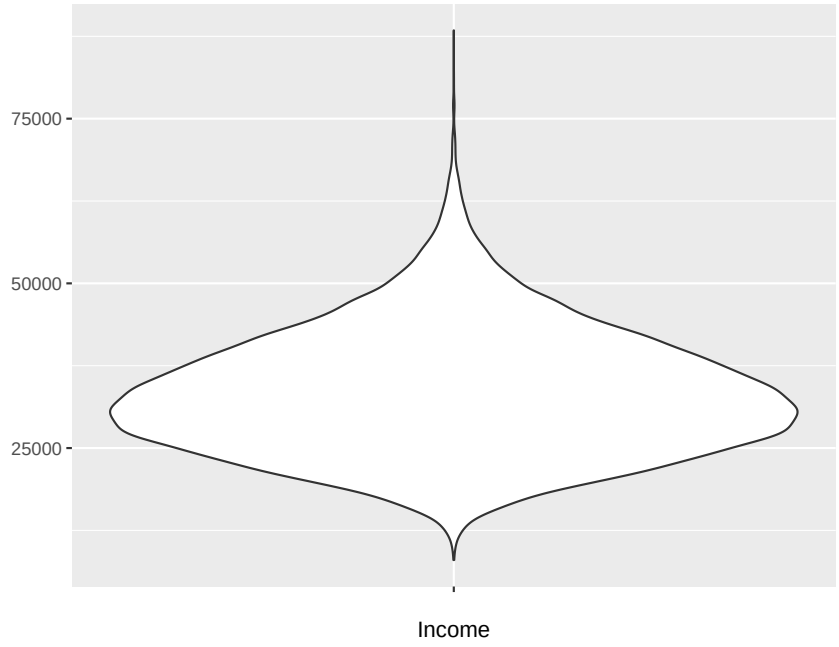


There is a nice alternative to the boxplot called the violinplot:

```
bplot(Income)
```




```
bplot(Income, do.violin = TRUE)
```



In addition to the box this also gives us some information on how many observations we have at various levels.

11.3.6 Case Study: Drug Use of Mothers and the Health of the Newborn

Chasnoff and others obtained several measures and responses for newborn babies whose mothers were classified by degree of cocain use. The study was conducted in the Perinatal Center

for Chemical Dependence at Northwestern University Medical School. The measurement given here is the length of the newborn.

Source: Cocaine abuse during pregnancy: correlation between prenatal care and perinatal outcome Authors: SN MacGregor, LG Keith, JA Bachicha, and IJ Chasnoff

Obstetrics and Gynecology 1989;74:882-885

Here we have two variables, Length (quantitative) and Status (categrorical). Another way (and in many ways a more natural way) to look at this data is as quantitative measurements from different groups. For this type of data we might first compute the summary statistics for each group separately:

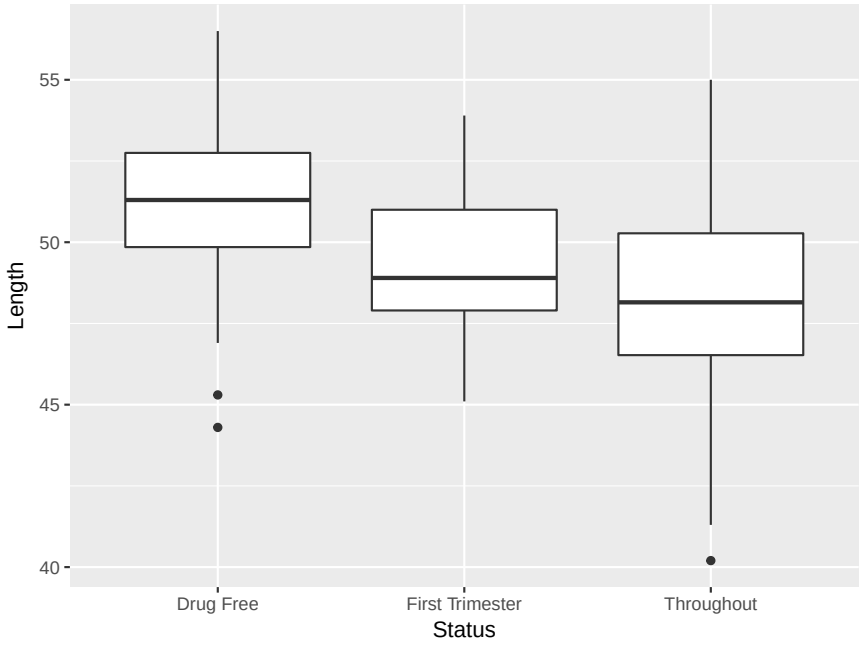
```
attach(mothers)
stat.table(Length,Status)
```

```
##                  Sample Size Mean Standard Deviation
## Drug Free                 39 51.1                 2.9
## First Trimester           19 49.3                 2.5
## Throughout                36 48.0                 3.6
```

Note that the discussion on Mean vs. Median still holds: If there are outliers it might be better to use the median and IQR here.

The standard graph for this data is a multiple boxplot. Note that all the boxes are on the same scale!

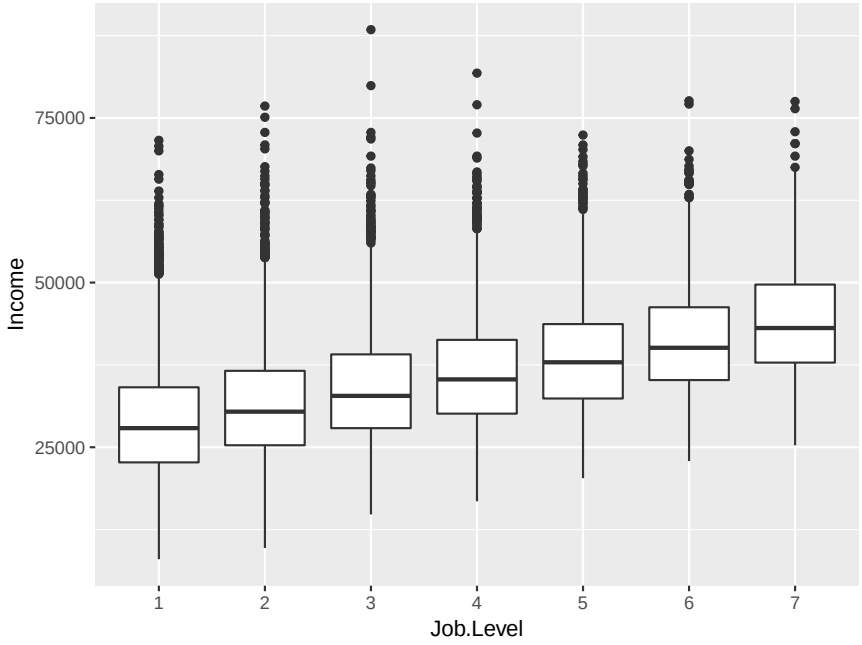
```
bplot(Length,Status)
```



11.3.7 Case Study:WRInc

Recall that when we looked at the relationship of job level and income we did a scatterplot but then noticed that to be a bad graph. So here is a better one:

```
bplot(Income, Job.Level)
```



and this shows much clearer the increase in income.

12 Two Quantitative Variables - Correlation

Generally if there are more than two variables we are interested in their relationships. We want to investigate the two questions:

1. Is there a relationship?
2. If there is a relationship, can we describe it?

If both variables are quantitative, for the first question we can find the correlation and for the second we can do a regression.

12.0.1 Case Study: Olympic Men's Long Jump

Data on the gold medal winning performances in the men's long jump for the modern Olympic games.

```
longjump
```

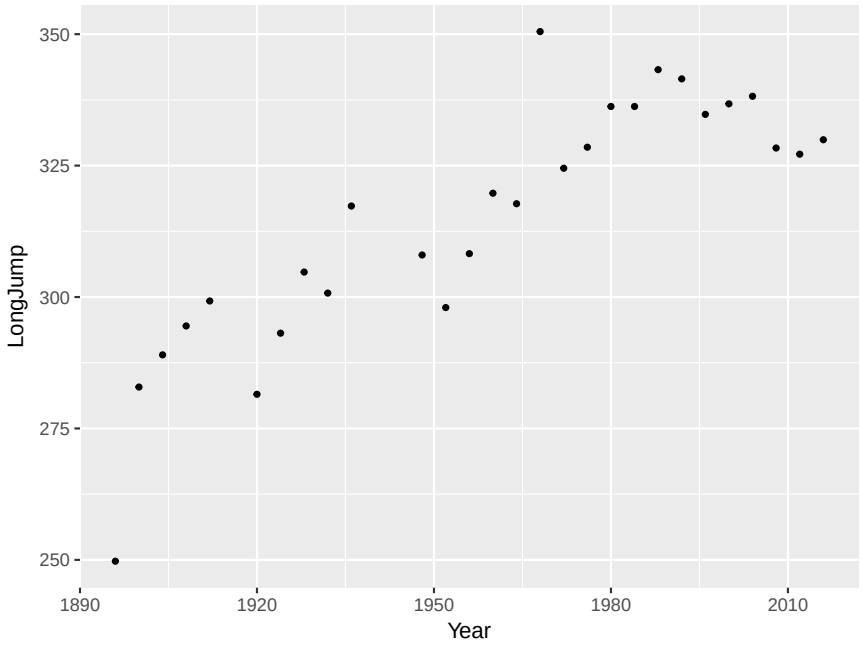
```
##              City Year LongJump
## 1          Athens 1896 249.7500
```

```
## 2           Paris 1900 282.8750
## 3        St.Louis 1904 289.0000
## 4          London 1908 294.5000
## 5       Stockholm 1912 299.2500
## 6         Antwerp 1920 281.5000
## 7           Paris 1924 293.1250
## 8       Amsterdam 1928 304.7500
## 9      LosAngeles 1932 300.7500
## 10         Berlin 1936 317.3125
## 11         London 1948 308.0000
## 12       Helsinki 1952 298.0000
## 13      Melbourne 1956 308.2500
## 14           Rome 1960 319.7500
## 15          Tokyo 1964 317.7500
## 16     MexicoCity 1968 350.5000
## 17         Munich 1972 324.5000
## 18       Montreal 1976 328.5000
## 19         Moscow 1980 336.2500
## 20     LosAngeles 1984 336.2500
## 21          Seoul 1988 343.2500
## 22      Barcelona 1992 341.5000
## 23        Atlanta 1996 334.7500
## 24         Sydney 2000 336.7500
## 25         Athens 2004 338.1880
## 26         Bejing 2008 328.3460
## 27         London 2012 327.1719
## 28 Rio de Janeiro 2016 329.9213
```

Both Year and LongJump are quantitative. Neither is very interesting by itself, the real interest is in their **relationship.** What does the year tell us about the length of the jump?

Here we usually start by drawing a **scatterplot**.

```
attach(longjump)
splot(LongJump, Year)
```



12.0.2 Case Study: The 1970's Military Draft

In 1970, Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The first date drawn from the drum received draft number one and eligible men born on that date were drafted first. In a truly random lottery there should be no relationship between the date and the draft number.

Question: **was the draft was really "random"?**

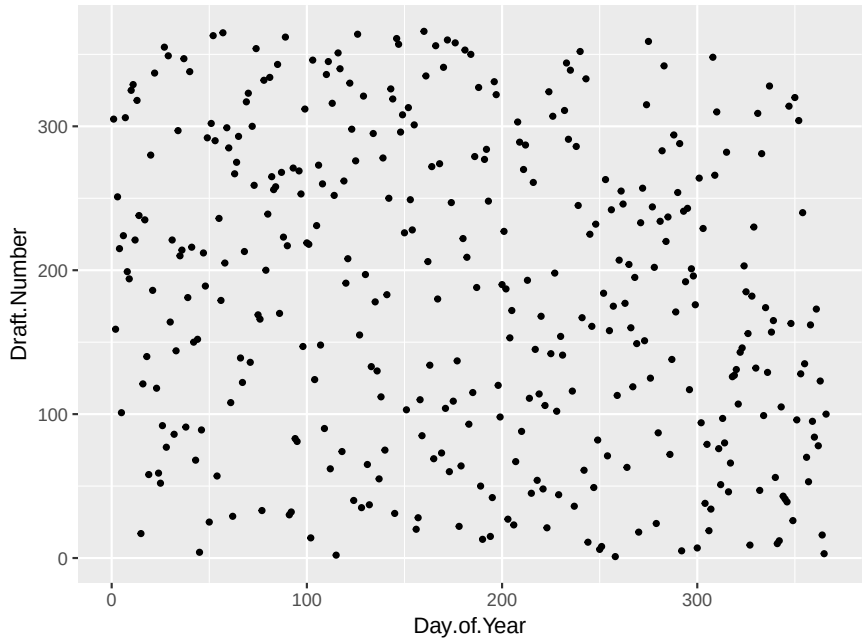
```
head(draft[, 4:5])
```

```
##   Day.of.Year Draft.Number
## 1           1          305
## 2           2          159
## 3           3          251
## 4           4          215
## 5           5          101
## 6           6          224
```

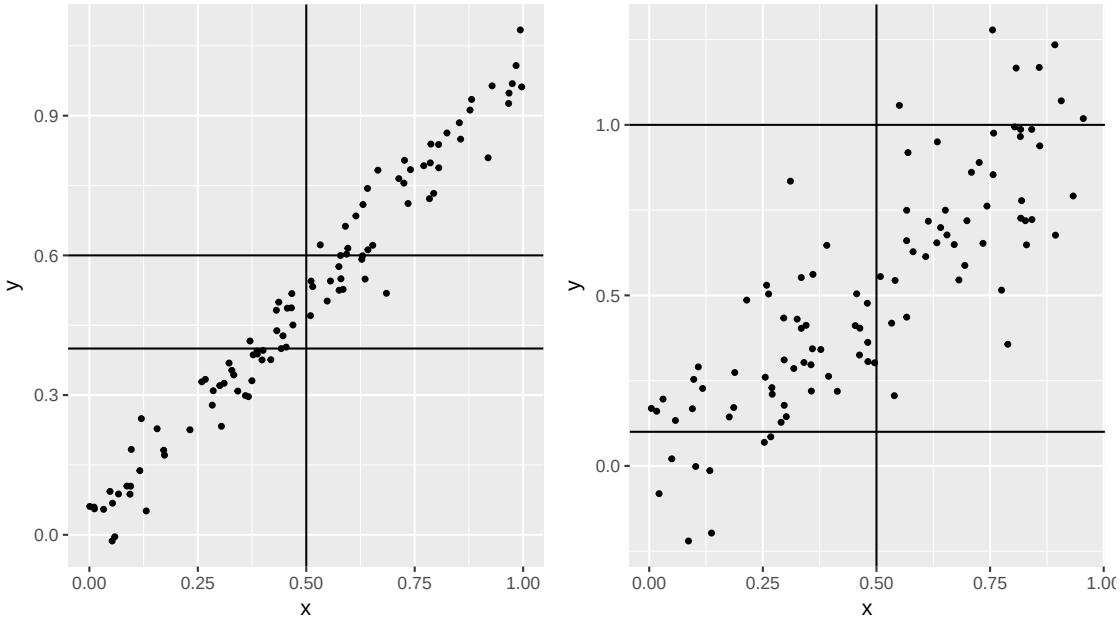
Let's have a look at the scatterplot of "Day.of.Year" and "Draft.Number":

```
attach(draft)
splot(Draft.Number, Day.of.Year)
```

Figure 13:



It certainly does not appear that there is a relationship between "Day of the Year" and "Draft Number", but is this really true? What we want is a number that can tell us if there is a relationship between two quantitative variables, and if so how strong it is. Consider the following two examples:



Clearly in the case on the left we have a much stronger relationship than on the right. For example, if I knew x = 0.5, then on the left I could reasonably guess that y is between 0.4 and 0.6, whereas on the right I could only guess 0.1 to 1.0.

The most popular choice for such a number is **Pearson's correlation coefficient r**, which we can find with the R command **cor**:

```r
cor(Draft.Number, Day.of.Year)
```

```
## [1] -0.2260414
```

correlations are usually rounded to three digits, so we have a correlation between Draft.Number and Day.of.Year of r = -0.226.

The correlation coefficient is like the mean, median, standard deviation, Q_1 etc.: it comes in two versions:

- it is a statistic when it is found from a sample
- it is a parameter when it belongs to a population

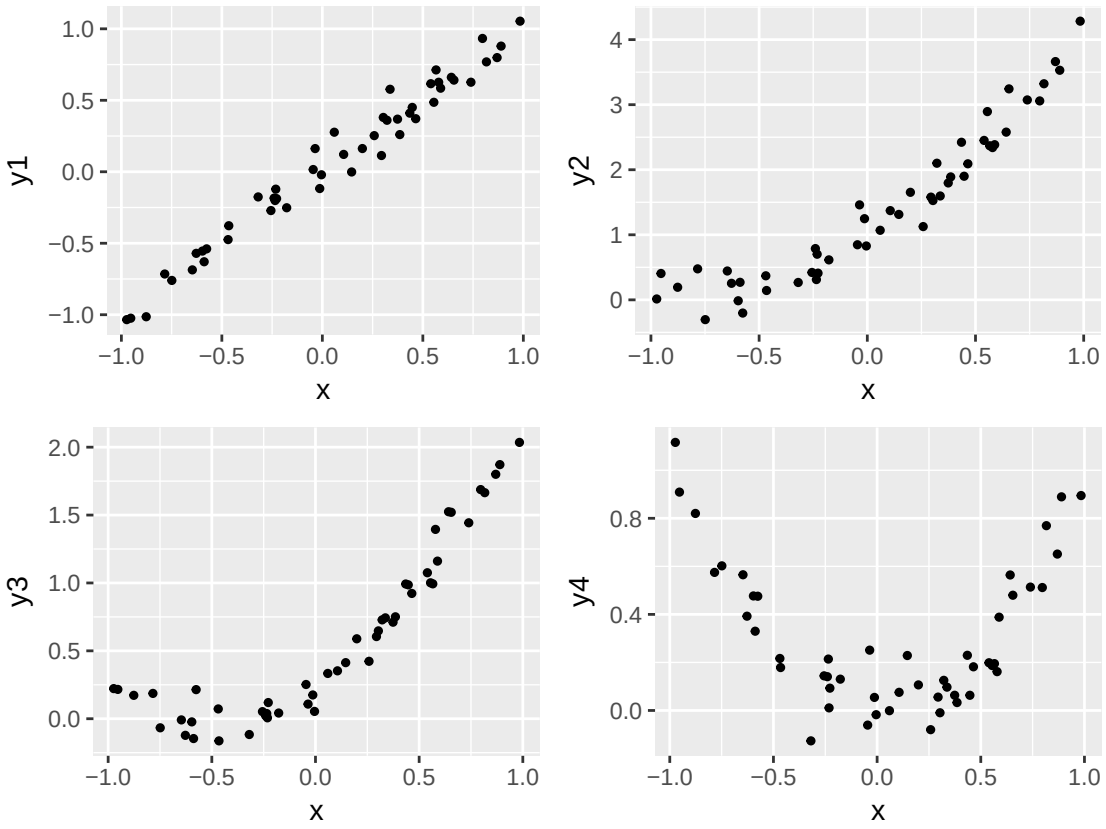
In the first case we use the symbol **r**, in the second case we use ρ.

Properties of the Correlation Coefficient:

- always -1 < r < 1
- r close to 0 means very small or even no correlation (relationship)
- r close to ± 1 means a very strong correlation
- r = -1 or r = 1 means a perfect linear correlation (that is in the scatterplot the dots form a straight line)
- r < 0 means a negative relationship (as x gets bigger y gets smaller)
- r > 0 means a positive relationship (as x gets bigger y gets bigger)
- r treats x and y symmetricaly, that is cor(x,y) = cor(y,x)

Peason's correlation coefficient only measures **linear** relationships, it does not work if a relationship is nonlinear.

Here is an example:



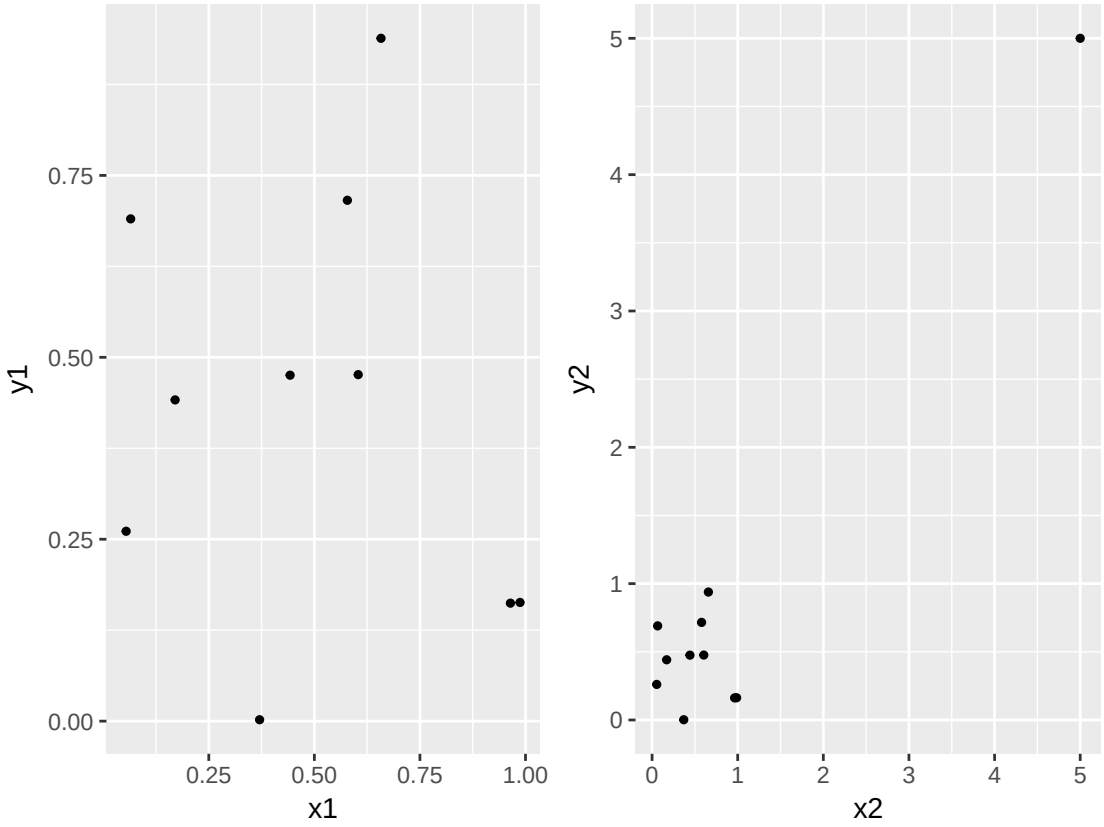
All for of these have about the same "strength" of a relationship.

BUT

```
## cor(x, y1) = 0.986
## cor(x, y2) = 0.937
## cor(x, y3) = 0.88
## cor(x, y4) = -0.111
```

Peason's correlation coefficient is only useful for the first case.

Another situation where Pearson's correlation coefficient does not work is if there are outliers in the dataset. Even just one outlier can determine the correlation coefficient:




```
## Correlation between x1 and y1:  -0.154
```

```
## Correlation between x2 and y2:  0.948
```

Weak vs. no Correlation

It is important to keep two things separate: a situation with two variables which are **uncorrelated** ($\rho = 0$) and two variables with a weak correlation ($\rho \neq 0$ but small).

In either case we would find an r close to 0 (but never = 0 !) Finding out which case it is might be impossible, especially for small datasets.

App - correlation

```
run.app(correlation)
```

this app illustrates the correlation coeffcient.

Move slider around to see different cases of the scatterplot of correlated variables

include a few outliers and see how that effects that "look" of the scatterplot and the sample correlation coefficient

On the Histogram tab we can study the effect of changing ρ and/or n on the sample correlation r.

Back to the draft

So, how about the draft? Well, we found r = -0.226. But of course the question is whether -0.226 is close to 0, close enough to conclude that all went well. Actually, the question really is whether the corresponding parameter $\rho = 0$! Let's do a **simulation:**

12.1 Simulation for the 1970's Military Draft

Doing a simulation means teaching the computer to repeat the **essential** part of an experiment many times. Here the experiment is the draft. What are the important features of this experiment?

- there are the numbers 1-366 in the order from 1 to 366 (in "Day.of.Year")
- there are the numbers 1-366 in some random order (in "Draft.Number")

In R we can do this as follows:

- get the numbers in Day.of.Year in random order with the *sample* command:

```
sample(Day.of.Year)
```

```
##   [1] 169 349 260 145  43  87 311 157 179 248 271  56 301 335 208 177  67
##  [18]   1 306  47   8 325 101  57 137 359 148 176 287  19 187 231 220 221
##  [35] 341 321 133 280 249 175 285 152   3 236 353  61 207 173 107 203 263
##  [52] 224 110  41 122 289 250 235 295 348 186   9 103  85  36  14 112 344
##  [69]  52 185 118 282 193  97  58  34 360 314 162 211 310 164   4 257 218
##  [86] 108  27 354 229 230 181  37  94 200 247 189  66 174  76 256 362 111
## [103] 350  25 217 138 269 106 308 190  90 242 165 305 151 105  20 132 239
## [120] 339 342  60 283 159 201 332 116 119  91 318 294 219  35 327 227 226
## [137]  81 238 262 316 212  78 120 121 243 267 291  75 246  13 184  46 240
## [154]  30  69 328  54 330 102 195 351  77  65 366  92  64 286 228 258 338
## [171] 178  59 276  39 153 266   6 163 129 206 126 156 264  72 104  18 234
## [188] 259 134 154 270 140  11 281  62 299 297 334  63 113 128 361 136  86
## [205]  82  68 292 273  48  44 223 170  23  79 278 205 160 255   2 326  12
## [222]  74 357 356 331 192 265  24 196 204 127 312 171 333 143  10 355 322
## [239] 155 345 202 320  15 293 210  53 284 194  50 209 365  70 182  80 213
## [256]  95 275 149 347 244 225 150 141 237  73  31  29  83 253   5 317 298
## [273] 251  26 337 180 191 309 364 125 117 252 319  88 290  16 147 144  89
## [290] 167 245 146 198 277 199 358  40  51 268 158  32 135 340 168 115 329
## [307] 109  45 183 343 188 172  55  93 254  98 272 124  22 324 130  84  42
## [324]  49 216 166 131  21 261 274  38 161  96 232 215 323  99  28 363 214
## [341] 100 313 296 336 139  71 307 114 123 288 304 241 300 279  17 142 233
## [358] 315  33 352 222   7 197 346 302 303
```

- and then calculate the correlation with

```
cor(Day.of.Year, sample(Day.of.Year))
```

```
## [1] 0.01648229
```

Now of course we should do this many times:

```
cor(Day.of.Year, sample(Day.of.Year))
```

```
## [1] -0.08308161
```

```
cor(Day.of.Year, sample(Day.of.Year))
```

```
## [1] -0.06702214
```

```
cor(Day.of.Year, sample(Day.of.Year))
```

```
## [1] 0.01222493
```

```
cor(Day.of.Year, sample(Day.of.Year))
```

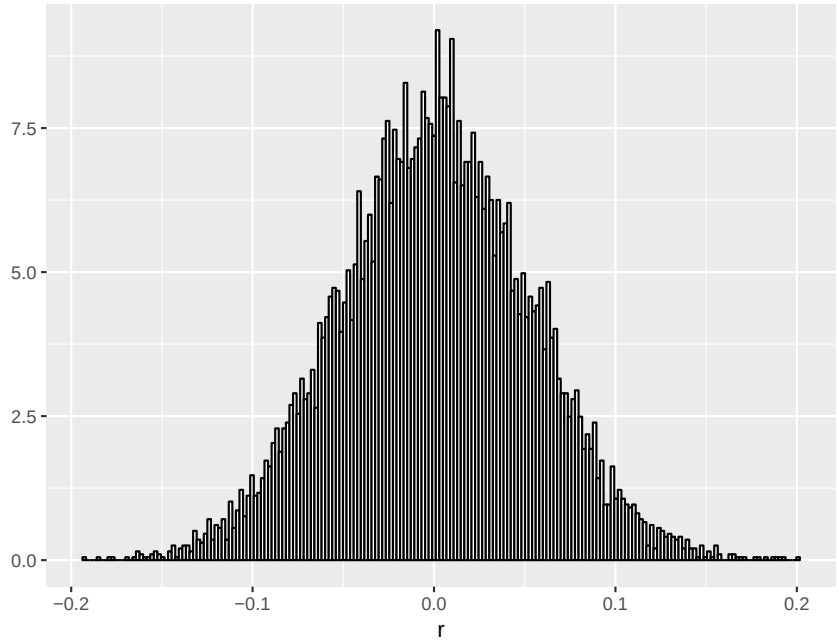
```
## [1] 0.1354924
```

```
cor(Day.of.Year, sample(Day.of.Year))
```

```
## [1] -0.01182866
```

And so we have not yet seen a correlation as far from 0 as 0.226. But maybe we need to do it many more times than that. Here is how:

```
r <- rep(0, 10000)
for(i in 1:10000)
  r[i] <- cor(Day.of.Year, sample(Day.of.Year))
hplot(r)
```




```
length(r[abs(r)>0.226])
```

```
## [1] 0
```

As you can see, none of the 10000 simulations had a sample correlation as far from 0 as

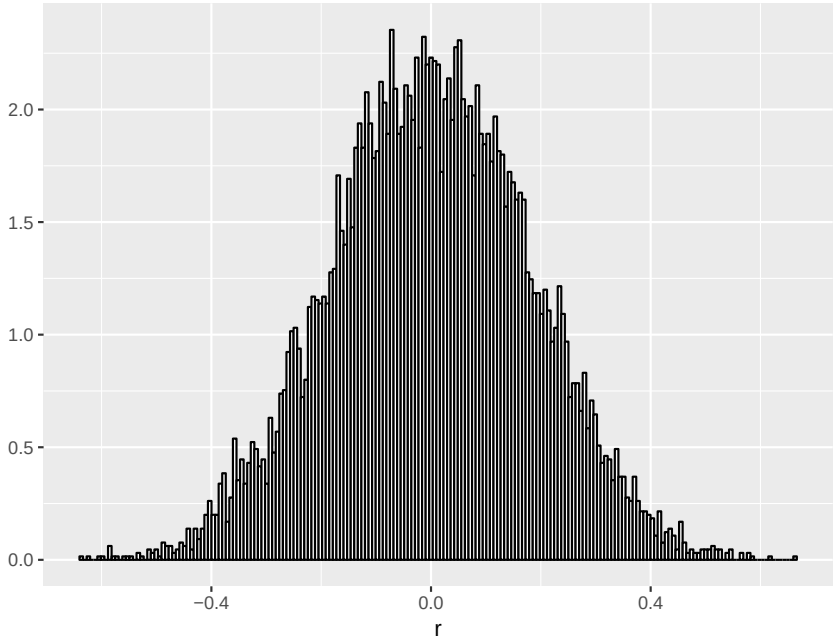
-0.226! So either

- The draft went fine, but something extremely unlikely happened (something with a probability less than 1 in 10000)
- Something went wrong in the draft.

A probability of less than 1 in 10000 is generally considered to unlikely, so we will conclude that something did go wrong.

So the next time you see a sample correlation coefficient r=-0.226, can you again conclude that the corresponding population correlation coefficient $\rho \neq 0$? Unfortunately no! For example, say that instead of using the day of the year the military had used the day of the month (1-31). Now we do the simulation

```
r <- rep(0, 10000)
for(i in 1:10000)
  r[i] <- cor(1:31, sample(1:31))
hplot(r)
```




```
length(r[abs(r)>0.226])
```

```
## [1] 2142
```

As you can see, now quite a few of the simulations had a sample correlation as far from 0 as -0.226 (about 22%), so this would not be unusual.

In a while we learn more about doing simulations as well as how to decide whether something is or is not statistically significant.

12.2 Correlation vs. Causation

Say we have found correlation between variables "x" and "y". How can we understand and interpret that relationship?

One possible explanation is a **Cause-Effect** relationship. This implies that if we can "change" x we expect a change in y.

Example

x = "hours you study for an exam"

y = "score on exam"

Example

Say we have the following data: for one year in some city we have data on fires that happenend during the year. Specifically we recorded

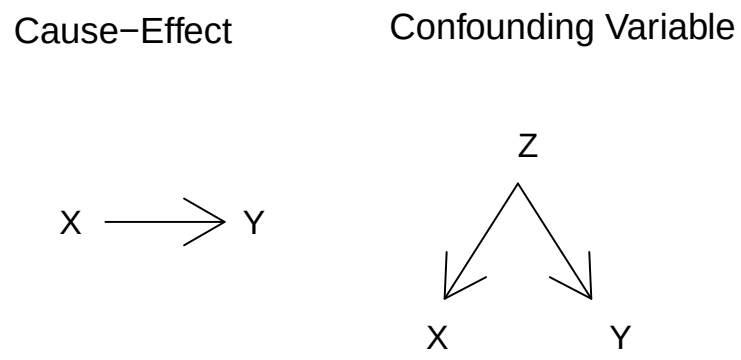
x = "Number of fireman responding to a fire"

y = "damages done by the fire"

say there is a positive correlation between x and y (and in real live there will be!).

Now if the correlation is due to a Cause-Effect, than changing x means changing y. Clearly we want a small y (little or no damamges), and because the correlation is positive we can get that by making sure x is small. So never call the firebrigade!

If this does not make sense, there has to be another explanation for the positive correlation:



Under the latent variable explanation we find (if all correlations are positive):

small z leads to small x and small y, so we get pairs of small (x,y)

large z leads to large x and large y, so we get pairs of large (x,y)

Finally cor(x,y) is positive!

[Online Resource: Bizzare Correlations] (http://www.buzzfeed.com/kjh2110/the-10-most-bizarre-correlation

---

Please note saying x causes y is not the same as x by itself determines y.

There are usually many other factors besides x that influence y, maybe even some more important than x.

Example

x = "hours you study for an exam

y = "score on exam"

but there are also many other factors that determine your score in an exam such as

- general ability
- previous experience

- being healthy on the day of the exam
- exam anxiety
- having a hang-over
- etc.

12.2.1 Case Study: Smoking and Lung Cancer

There have been hundreds of studies all over the world that have shown a correlation between smoking rates and lung cancer deaths, usually with correlations of about 0.5 to 0.7. And yet, none of these studies has shown that smoking causes lung cancer because all of the were observational studies, not clinical trial.

The only perfectly satisfactory way to establish a causation is to find a random sample, for example to do a **clinical trial**. An **observational study** is always somewhat suspect because we never know about hidden biases. Nevertheless, even only using observational studies the evidence for cause-effect can be quite strong:

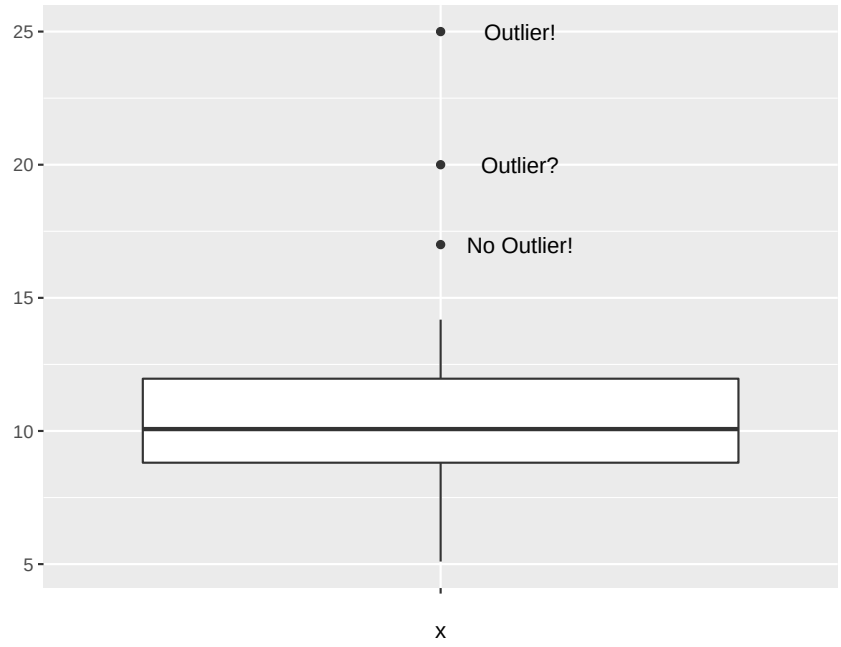
Things to look for when trying to establish a causation:

- correlation is strong - the correlation between smoking and lung cancer is very strong
- correlation is consistent over many experiments - many studies of different kinds of people in different countries over a long time period all have shown this correlation
- higher doses are associated with stronger responses - people who smoke more have a higher chance of lung cancer
- the cause comes before the response in time - lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women started to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer deaths among women.
- the cause is plausible - lab experiments on animals show that nicotin causes cancer.

13 Outliers - Detection and Treatment

Many of the methods discussed in this class don't work well if the dataset has **outliers**. An outlier is any observation that is in some way unusual/strange/weird.

We have already seen that an observation that is unusual with respect to one variable appears as a separate dot in an R boxplot:



Unfortunately there are no hard rules exactly when an observation becomes an outlier. To a large part that depends on the method of analysis we want to use, some methods are **sensitive** to outliers, others are more **robust**.

In addition to the case discussed above, there are other ways in which an observation can be an outlier:

13.0.1 Case Study: Alcohol vs. Tobacco Expenditure

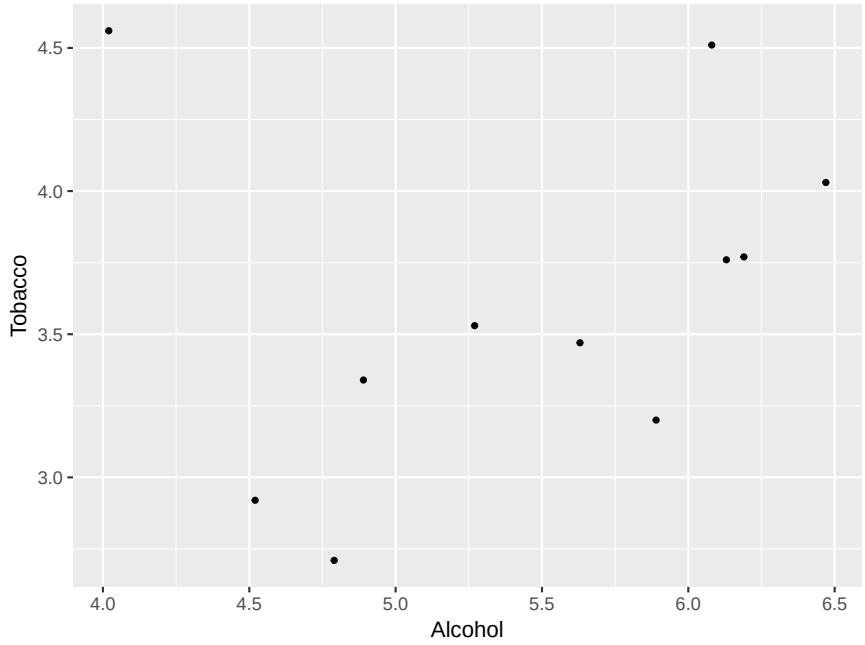
Data from a British government survey of household spending may be used to examine the relationship between household spending on tobacco products and alcoholic beverages. The numbers are the average expenditure for each of the 11 regions of England.

```
alcohol
```

```
##              Region Alcohol Tobacco
## 1             North    6.47    4.03
## 2         Yorkshire    6.13    3.76
## 3         Northeast    6.19    3.77
## 4     East_Midlands    4.89    3.34
## 5     West_Midlands    5.63    3.47
## 6       East_Anglia    4.52    2.92
## 7         Southeast    5.89    3.20
## 8         Southwest    4.79    2.71
## 9             Wales    5.27    3.53
## 10         Scotland    6.08    4.51
## 11 Northern_Ireland    4.02    4.56
```

Here we have two quantitative variables, so the obvious thing to do is draw the scatterplot:

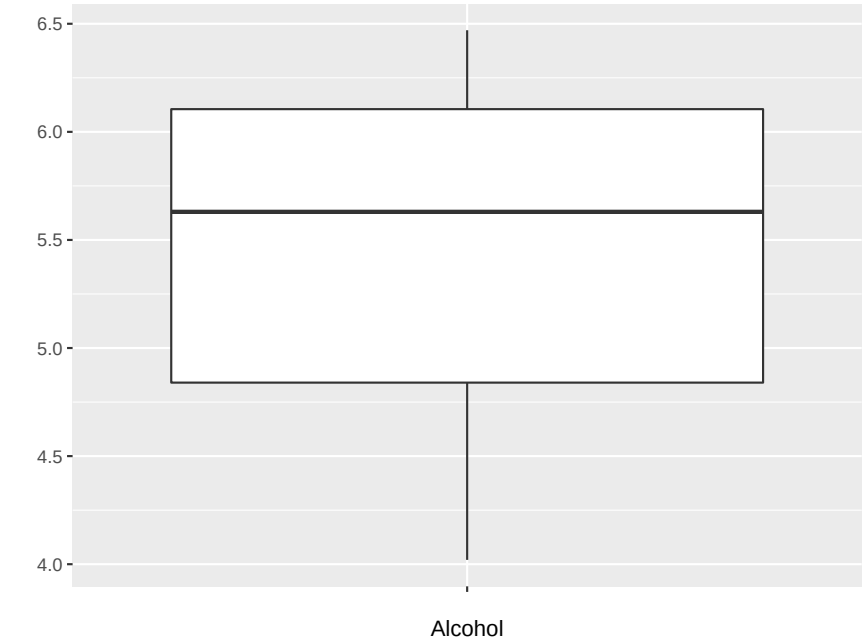
```
attach(alcohol)
splot(Tobacco , Alcohol)
```



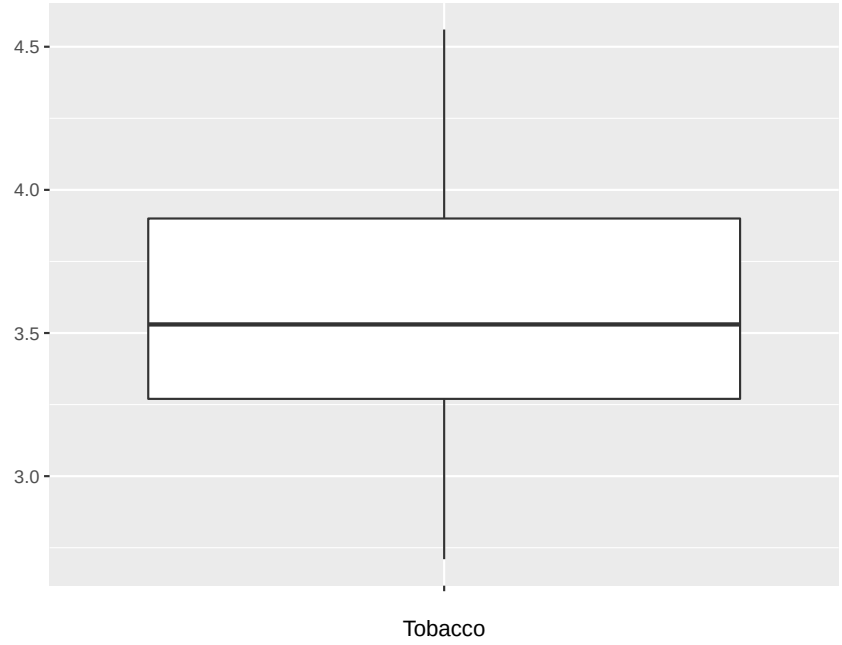
There seems to be generally a positive relationship, but also one case that does not fit. It seems it has the smallest value for Alcohol, which we can see in the data is for Northern Ireland, where there is a fairly high expenditure on Tobacco but not on Alcohol (???)

Note that neither Alcohol nor Tobacco have any outliers by themselves:

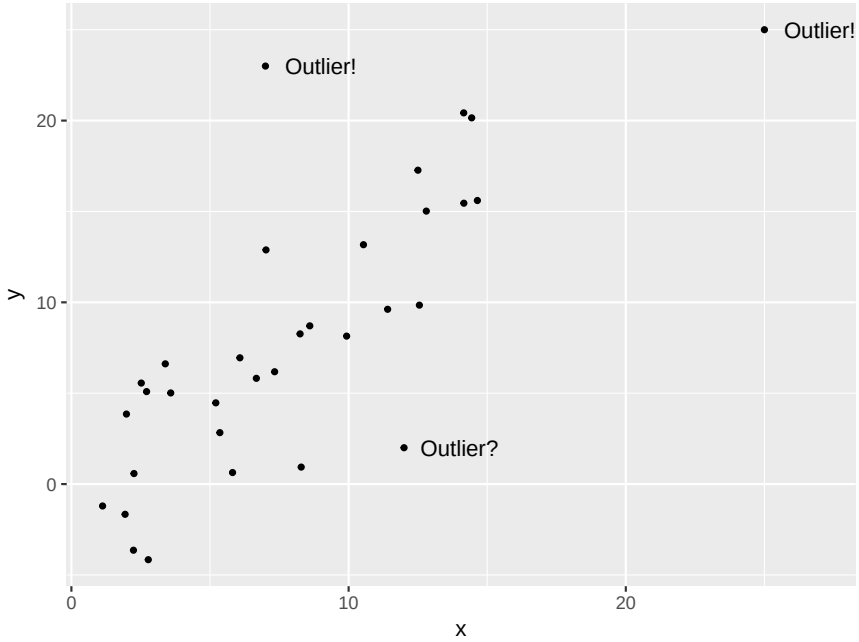
```
bplot(Alcohol)
```



```
bplot(Tobacco)
```



Again, it is not always obvious when an observation becomes an outlier:

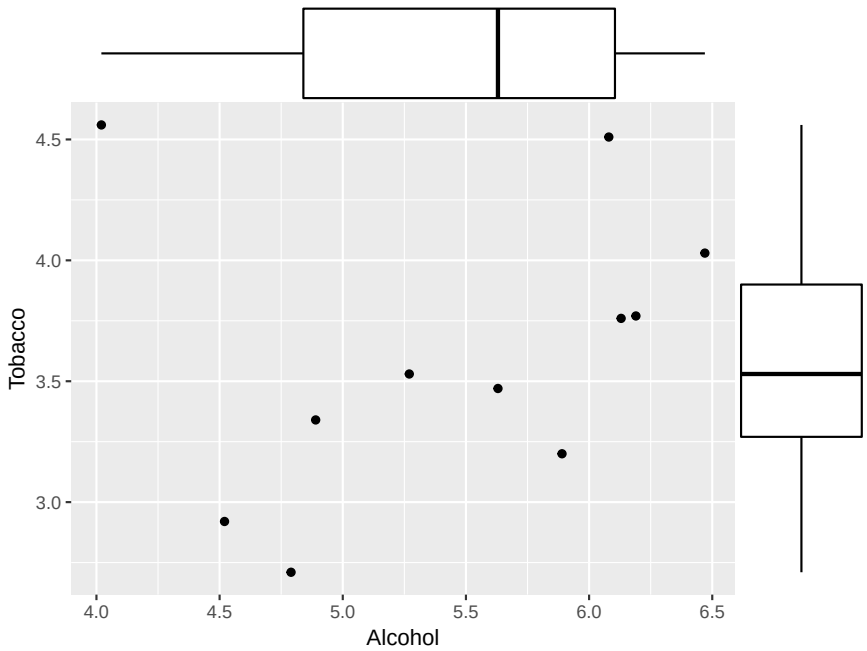


If we have two quantitative variables an outlier can happen in one of three ways:

- in the x variable, which we can check in the boxplot of x
- in the y variable, which we can check in the boxplot of y
- in the relationship between the xand the y variable, which we can check in the scatterplot of x and y

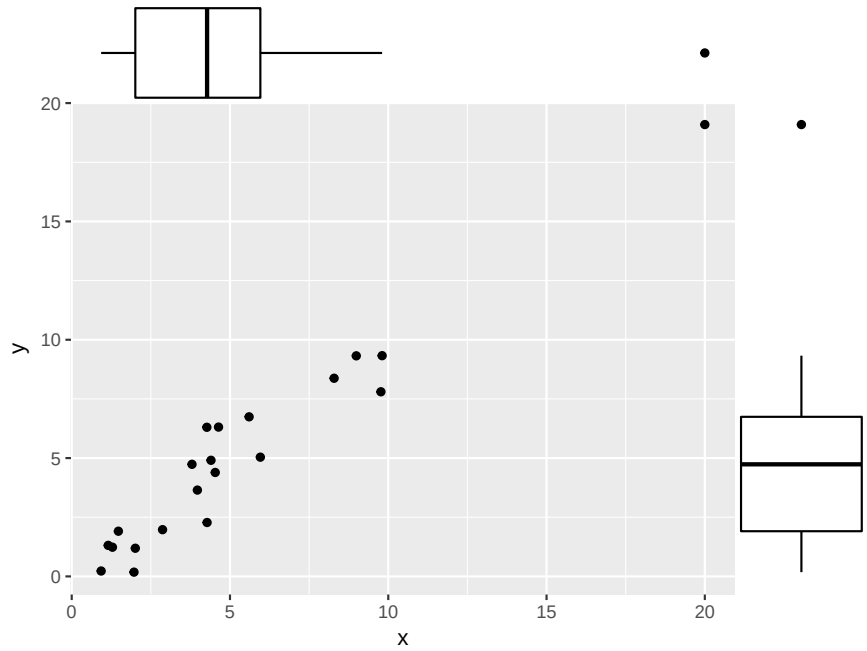
In fact we can do all three in one step:

```
mplot(Tobacco, Alcohol)
```



Consider the following case:

```
x <- c(runif(20, 0, 10), 20)
y <- x+rnorm(21)
mplot(y, x)
```



How many outliers does this data set have? Actually just one, but it appears in each of three graphs.

13.1 Treatment of Outliers

If we have an outlier in a dataset, what do we do then? First and foremost, **don't ignore them!** Most statistical methods are very sensitive to outliers, often they simply don't work.

Example Is there a relationship between Alcohol and Tobacco expenditures in England? Because we have two quantitative variables we might use Pearson's correlation coefficient to answer this question:

```
cor(Tobacco, Alcohol)
```

```
## [1] 0.2235721
```

```
cor(Tobacco[-11], Alcohol[-11])
```

```
## [1] 0.7842873
```

So with Northern Ireland we find a weak positive correlation, but without Northern Ireland it is a fairly strong positive correlation.

Which one is right? Clearly the first one is wrong because of the outlier!

So, if there are outliers, what do we do?

1. Learn as much as you can about the "story" behind the data and understand why there is an outlier. Is it an error? Is it something we should expect to see in this kind of data? etc.
2. Find a method that is not sensitive to outliers. For example, alternatives to Pearson's correlation coefficient include Spearman's rank correlation coefficient and Kendall's coefficient of concordance , although neither of them works any better here.
3. Try and "adjust" the outliers. We know what "caused" the Alcohol number for Northern Ireland to be off, so maybe we can adjust it.
4. If all else fails, eliminate the outlier(s)

14 Exercises - Descriptive Statistics - Data Summaries

Problem 1

For each of the following variables decide whether the data is categorical or quantitative

Daily low temperature in New York

Brand of cereal in supermarket

Telephone number

License plates of cars

Weight lost in a weight loss program

Time spent on studying for the class during last week

14.0.1 Problem 2

0.5 , 1 , 1.5 , 1.9 , 2.1 , 2.2 , 2.7 , 2.8 , 3.3 , 3.6 , 3.9 , 3.9 , 3.9 , 4 , 4 , 4.3 , 4.3 , 4.5 , 4.5 , 5 , 5 , 5.1 , 5.1 , 5.5 , 5.6 , 5.9 , 6.2 , 6.3 , 7.1 , 27.1

a. Find the mean, median, range and standard deviation.

b. Find the 20^{th} and the 64^{th} percentile of this dataset.

c. draw the boxplot for this dataset

Problem 3

Using the data from problem 2, find the z score of x = 3.6. What x would have a z score = 1?

Problem 4

Consider the data set for Friday the 13^{th}. This data set has several comparisons of a Friday the 13th and the previous Friday the 6th, for example the number of cars passing through a junction (traffic), shoppers for a supermarket (shopping), or admissions due to transport accidents (accident)

```
head(friday13)
```

```
##   Dataset    six thirteen
## 1 traffic 139246   138548
## 2 traffic 134012   132908
## 3 traffic 137055   136018
## 4 traffic 133732   131843
## 5 traffic 123552   121641
## 6 traffic 121139   118723
```

Use R to compute the mean and the standard deviation for the two Fridays and the three data sets separately (so there will be the mean and st. dev. for the number of accidents on Friday the 6^{th}, the mean and st. dev. for the number of accidents on Friday the 13^{th}, the mean and st. dev. for the number of shoppers on Friday the 6^{th} and so on).

Does any of these numbers support the idea that Friday the 13^{th} is special?

Problem 5 Consider the data in AIDS in Americas in 1995.

```
head(aids)
```

```
##           Country  AIDS
## 1        Anguilla   0.0
## 2 Antigua Barbuda   7.3
## 3       Argentina   5.6
## 4         Bahamas 131.4
## 5        Barbados  44.1
## 6          Belize   4.5
```

a. Find the 20^{th} and the 80^{th} percentiles of the AIDS rates.

b. Find the 5 number summary for the aids rates.

c. According to the boxplot, which countries are outliers in this data set?

d. Let's say the WHO wants to use the "average" rate of AIDS infection (together with the number of people living in the Americas) to estimate the number of AIDS infected people in the Americas. Should they use the mean or the median to find the "average"?

Problem 6

In this exercise we study the dataset

```
head(headache)
```

```
##   Time Dose Sex BP.Quan
## 1   35    2   0    0.25
## 2   43    2   0    0.50
## 3   55    2   0    0.75
## 4   47    2   1    0.25
## 5   43    2   1    0.50
## 6   57    2   1    0.75
```

a. What is the type of data of the variables?

b. Find the mean and standard deviation of Time.

c. Find the 5-number summary and draw the boxplot of Time

Problem 7

Company XYZ has a contract with a supplier for metal rods. The contract says that all the rods have to be between 15.26cm and 15.47cm long. XYZ just received a shipment of 50000 rods. They randomly select 100 of them and measure the length of each. They find $\overline{X} = 15.344$ and $s = 0.041$. Should they except this shipment?

Problem 8

Consider the **rogaine** dataset. Draw a good graph for this dataset.

Problem 9 Consider the following data set:

```
x <- 1:10
y <- c(3,    0,  0,  10,     8,  4,  5,  8,  14,     9)
kable(data.frame(x=x,y=y))
```

x	y
1	3
2	0
3	0
4	10
5	8
6	4
7	5
8	8
9	14
10	9

Now we find

```
cor(x, y)
```

```
## [1] 0.7041777
```

Find another observation (a,b) such that

```
round(cor( c(x,a), c(y,b)),2)
```

14.1 Solutions

Problem 1 For each of the following variables decide whether the data is categorical or quantitative

Daily low temperature in New York - quantitative

Brand of cereal in supermarket - categorical

Telephone number - categorical

License plates of cars - categorical

Weight lost in a weight loss program - quantitative

Time spent on studying for the class during last week - quantitative

Problem 2

0.5 , 1 , 1.5 , 1.9 , 2.1 , 2.2 , 2.7 , 2.8 , 3.3 , 3.6 , 3.9 , 3.9 , 3.9 , 4 , 4 , 4.3 , 4.3 , 4.5 , 4.5 , 5 , 5 , 5.1 , 5.1 , 5.5 , 5.6 , 5.9 , 6.2 , 6.3 , 7.1 , 27.1

the data is comma delimeted, so after copying it in R type

```
x <- getx(sep=",")
```

```
mean(x)
```

```
## [1] 4.76
```

```
median(x)
```

```
## [1] 4.15
```

```
sd(x)
```

```
## [1] 4.51859
```

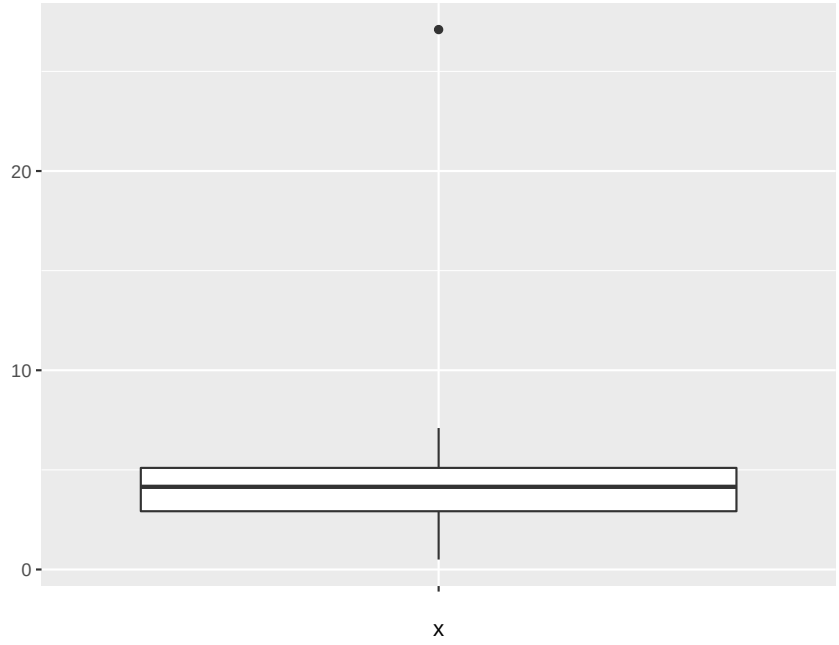
b. Find the 20^{th} and the 64^{th} percentile of this dataset.

```
quantile(x, c(0.2, 0.64))
```

```
##  20%  64%
## 2.60 4.78
```

c) draw the boxplot for this dataset

```
bplot(x)
```



Problem 3 Using the data from problem 2, find the z score of x = 3.6. What x would have a z score = 1?

We found $\overline{X} = 4.76$ and $s = 4.519$, so the z score of x=3.6 is
$z = (x - \overline{X})/s = (3.6 - 4.76)/4.519 = -0.2567$

We want z score =1, so

$1 = (x - \overline{X})/s$

or

$s = x - \overline{X}$

or

$x = s + \overline{X} = 4.519 + 4.76 = 9.279$

Problem 4

Consider the data set for Friday the 13^{th}. This data set has several comparisons of a Friday the 13th and the previous Friday the 6th, for example the number of cars passing through a junction (traffic), shoppers for a supermarket (shopping), or admissions due to transport accidents (accident)

Use R to compute the mean and the standard deviation for the two Fridays and the three data sets separately (so there will be the mean and st. dev. for the number of accidents on Friday the 6^{th}, the mean and st. dev. for the number of accidents on Friday the 13^{th}, the mean and st. dev. for the number of shoppers on Friday the 6^{th} and so on). Does any of these numbers support the idea that Friday the 13^{th} is special?

a.

```r
attach(friday13)
mean(friday13[Dataset=="accident",3])
```

```
## [1] 10.83333
```

```r
mean(friday13[Dataset=="accident",3])
```

```
## [1] 10.83333
```

```r
mean(friday13[Dataset == "traffic", 2])
```

```
## [1] 128385.3
```

```r
sd(friday13[Dataset == "traffic", 2])
```

```
## [1] 7259.223
```

```r
mean(friday13[Dataset == "traffic", 3])
```

```
## [1] 126549.5
```

```r
sd(friday13[Dataset == "traffic", 3])
```

```
## [1] 7664.282
```

```r
mean(friday13[Dataset == "shopping", 2])
```

```
## [1] 4970.511
```

```r
sd(friday13[Dataset == "shopping", 2])
```

```
## [1] 1165.615
```

```r
mean(friday13[Dataset == "accident", 2])
```

```
## [1] 7.5
```

```r
sd(friday13[Dataset == "accident", 2])
```

```
## [1] 3.331666
```

```r
mean(friday13[Dataset == "accident", 3])
```

```
## [1] 10.83333
```

```r
sd(friday13[Dataset == "accident", 3])
```

```
## [1] 3.600926
```

There does not appear to be anything special about Friday the 13^{th}

Problem 5

Consider the data in AIDS in Americas in 1995.

a. Find the 20^{th} and the 80^{th} percentiles of the AIDS rates.

```
attach(aids)
quantile(AIDS, c(0.2,0.8))
```

```
##   20%   80%
##  0.96 13.82
```

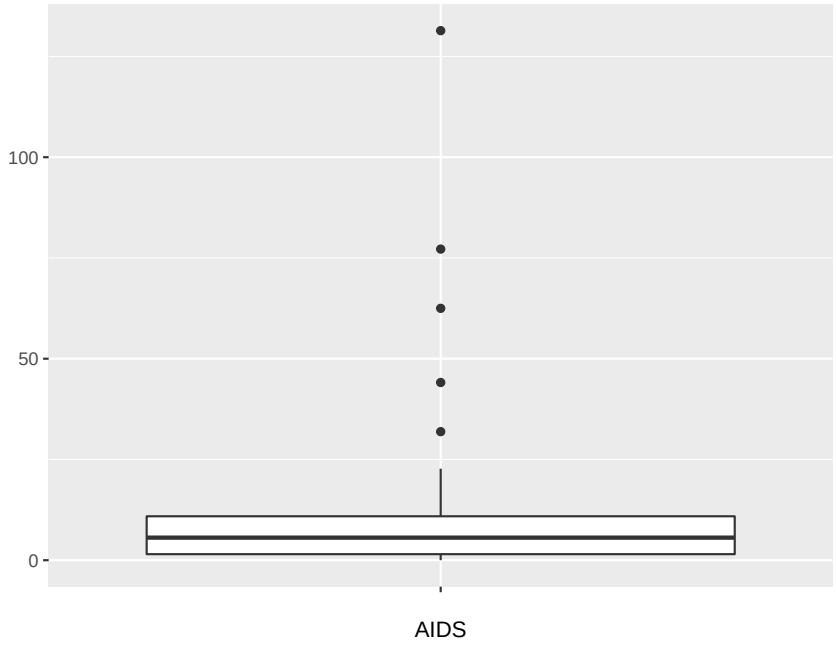
b. Find the 5 number summary for the aids rates.

```
fivenumber(AIDS)
```

```
##  Minimum  Q1 Median   Q3 Maximum
##        0 1.5    5.6 10.9   131.4
## IQR =  9.4
```

c. According to the boxplot, which countries are outliers in this data set?

```
bplot(AIDS)
```



the 5 countries with rates over 30are outliers, so

```
aids[AIDS > 30,]
```

```
##          Country  AIDS
## 4        Bahamas 131.4
## 5       Barbados  44.1
## 7        Bermuda  77.2
## 21 French Guiana  62.5
## 23    Guadaloupe  31.9
```

d. Let's say the WHO wants to use the "average" rate of AIDS infection (together with the number of people living in the Americas) to estimate the number of AIDS infected people in the Americas.

Should they use the mean or the median to find the "average"?

Mean, because the countries with the highest AIDS rates have to influence our "average", and they don't if we use the median.

Problem 6

In this exercise we study the dataset **headache**

a. What is the type of data of the variables?

Time: quantitative
Dose: quantitative
Sex: categorical
BP Quan: categorical.

b. Find the mean and standard deviation of Time.

```
attach(headache)
mean(Time)
```

```
## [1] 26.33333
```

```
sd(Time)
```

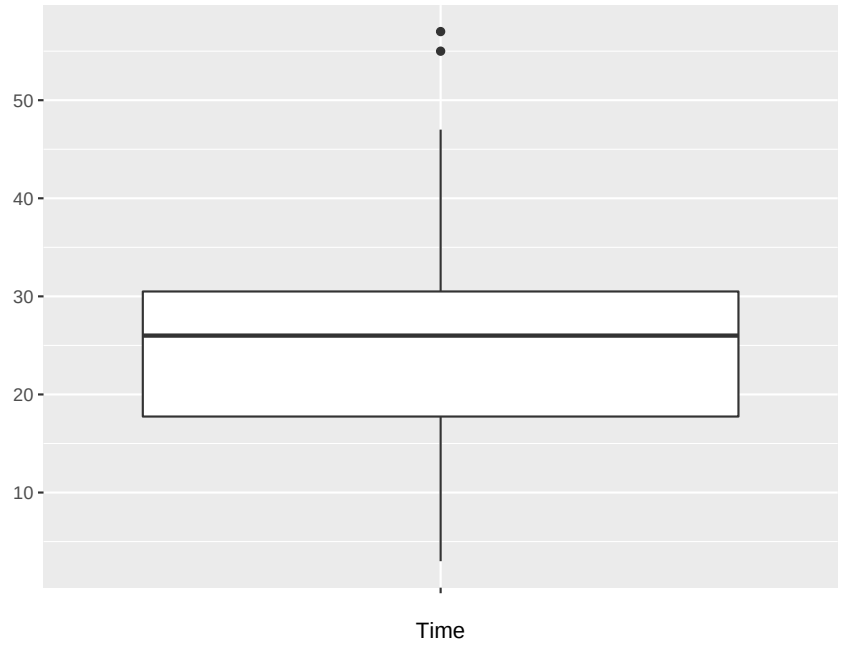
```
## [1] 14.56818
```

c. Find the 5-number summary and draw the boxplot of Time

```
fivenumber(Time)
```

```
##  Minimum   Q1 Median   Q3 Maximum
##        3 17.8     26 30.5      57
## IQR =  12.7
```

```
 bplot(y=Time)
```



Problem 7

Company XYZ has a contract with a supplier for metal rods. The contract says that all the rods have to be between 15.26cm and 15.47cm long. XYZ just received a shipment of 50000 rods. They randomly select 100 of them and measure the length of each. They find $\overline{X} = 15.344$ and $s = 0.041$. Should they except this shipment?

We can use the empirical rule to decide. This requires that the lengths of the rods have a bell-shaped histogram, which of course should be checked. Then

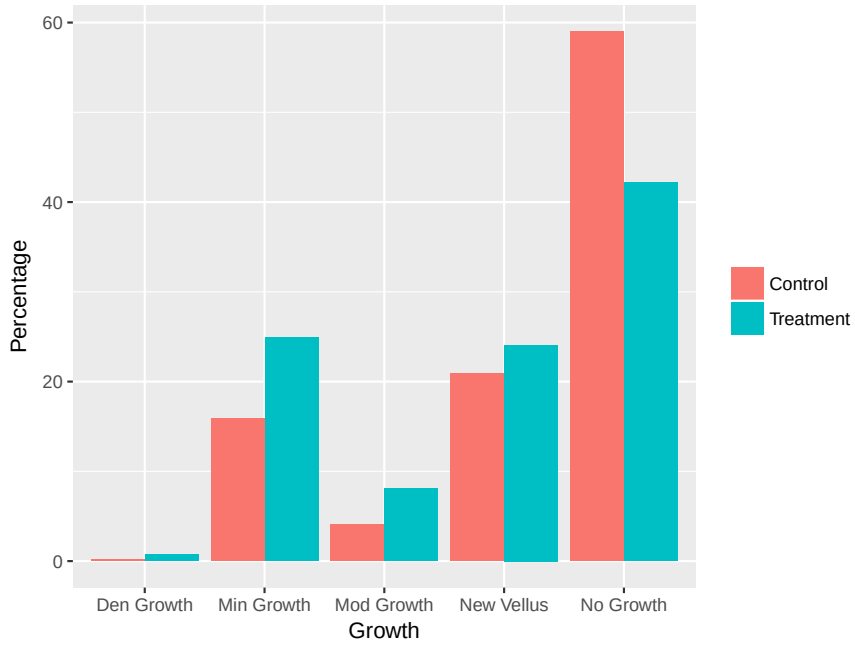
$\overline{X} \pm 2s = 15.344 \pm 2 \times 0.041 = (15.262, 15.426)$

The interval is supposed to include 95% the observations (or lengths of the rods), so we can conclude that 95% the rods have a length at least 15.262cm and at most 15.462cm, which is in accordance with the contract. So XYZ should accept the shipment.

Problem 8

Consider the **rogaine** dataset. Draw a good graph for this dataset. Here we have two categorical variables: Growth and Group. The standard graph for two categorical variables is the multiple barchart. There are 6 possible graphs, depending whether we show totals, percentages based on grand total, percentages based on row total or percentages based on column total, and on the grouping. The most useful of these is probably the bar chart based on percentages within Growth:

```
attach(rogaine)
barchart(y=table(Growth, Group), Percent="Column")
```



Problem 9 Consider the following data set:

```
x <- 1:10
y <- c(3,    0,  0,  10,     8,  4,  5,  8,  14,     9)
kable(data.frame(x=x,y=y))
```

x	y
1	3
2	0
3	0
4	10
5	8
6	4
7	5
8	8
9	14
10	9

Now we find

```
cor(x,y)
```

Find another observation (a,b) such that

```
round(cor( c(x,a), c(y,b)),2)
```

There are many solutions, here is one of them:

```
round(cor(c(x,20),c(y,-0.5)),2)
```

```
## [1] 0
```

15 Why Probability?

15.0.1 Case Study: Is this Coin Fair?

Let's say you have a specific coin and you want to see whether it is actually a **fair** coin, or whether it comes up heads more than tails. So you sit down and flip the coin 1000 times and get N heads.

What do you think N should be for this to NOT be a fair coin? Because (for some reason) we only care about the case were we get more heads than tails obviously we need $N > 500$ before we would conclude that this is an unfair coin. But also quite clearly (say) $N = 510$ would not be convincing enough to say this is an unfair coin.

But why not?

Simply because a fair coin can easily give 510 heads in 1000 flips, or the probability of 510 heads in 1000 flips of a **fair*** coin is not so small.

Actually it is very small (0.0207) but there are many possible outcomes, and so the probability of any one of them has to be small, for example 500 heads in 1000 flips has a probability of 0.0252, even though it is the most likely outcome. And relative to 0.0252 0.0207 is large.

Because of this what we will calculate is the probability of **N or more heads** in 1000 flips of a fair coin. Now we find

N or more	Probability
500	0.513
505	0.388
510	0.274
515	0.180
520	0.109
525	0.061
530	0.031
535	0.015
540	0.006
545	0.002
550	0.001

and so somewhere around N=530 these probabilities get very small, it is then more reasonable to think that the coin is actually not fair.

The topics we have discussed so far in this class (making graphs, computing things like the mean or the median) go under the heading of **descriptive statistics**. What we want to do now is make a guess what the true "state of nature" is based on the available information, namely decide whether this is a fair coin or not. This type of problem goes under the heading of **inferential statistics**. As we just saw, this generally means calculating some probabilities.

By the way eventually we will call the probabilities in the table **p-values**.

15.1 Statistically Significant

In Statistics you often hear statements like:

"The new drug is statistically significantly better than the old one"

What does this mean?

The statement **"The new drug is statistically significantly better than the old one"** means that an experiment was carried out, and the new drug performed better than the old one. More than that, if the same experiment were repeated again (with different subjects) we expect the new drug to perform better again **with a high probability** (90%? 99%?)

15.2 Probability Distributions

A **Probability Distribution** is a model for a real live experiment

Example We roll a fair die:

x	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Example We roll a fair die until the first six appears. How many rolls are needed?

P(six on first roll) = 1/6

P(first six on second roll) = P(no six on first roll, a six on second roll) = $5/6 \times 1/6$

P(first six on third roll) = P(no six on first two rolls, a six on third roll) = $5/6 \times 5/6 \times 1/6$

It's easy to guess the general case:

P(first six on k^{th} roll) = $(5/6)^{k-1}1/6$, k=1,2,..

So for example

P(first six on the 5^{th} roll) = P(X=5) = $(5/6)^{5-1}1/6 = 0.0804$

15.2.1 App

To illustrate this experiment run the app

```
run.app(geometric)
```

Probability distributions describe populations. The distribution in the first example tells us everything we might want to know about the population of all possible outcomes of the experiment "roll a fair die". There are formulas for all sorts of things. For example, say we

Figure 14:

roll the die a million times and keep track of the rolls. Then we find the mean. What would it be? Theory tells us it is 3.5. How about the standard deviation? it would be 1.7. How do we know this? Because we can do some math and calculated them from the distribution.

There is a fairly small list of "basic" distributions that cover a wide variety of everyday experiments and random phenomena. Here is one:

The most basic experiment possible is one that has only two possible outcomes:

- flip a coin - heads or tails
- roll a die - get a six or don't
- take a class - pass or fail
- person smokes - yes or no
- person has open heart surgery - person survives or dies

any such experiment is called a **Bernoulli trial**, named after the Swiss mathematician Jakob Bernoulli (1655-1705)

Note often one of the two outcomes is called a "success" and the other one a "failure". But for the calculations it makes no difference what is what. This sometimes leads to a bit of nastiness:

- person has open heart surgery - person survives (=failure) or dies (=success)

Usualy we "code" one outcome as 0 (=failure) and the other as 1 (=success) . Then the distribution is given by

x	P(x)
0	1-p
1	p

Note that this does not describe the experiment completely but only up to the p. This lets us "tailor" the distribution to specific experiments:

Example: Flip a fair coin, Heads = success = 1: p = 0.5

Example: Roll a fair die, "get a six" = success = 1: p=1/6

Example: Choose employee from WRInc, employee is female = success = 1, p=9510/23791=0.3997

Example choose people from some population, person has a genetic condition = success = 1: p=0.015

The number p then is a number that belongs to the population described by the distribution: it is a **parameter**!

Once we have a distribution we can find formulas for the **population** mean and standard deviation. For the Bernoulli distribution we have

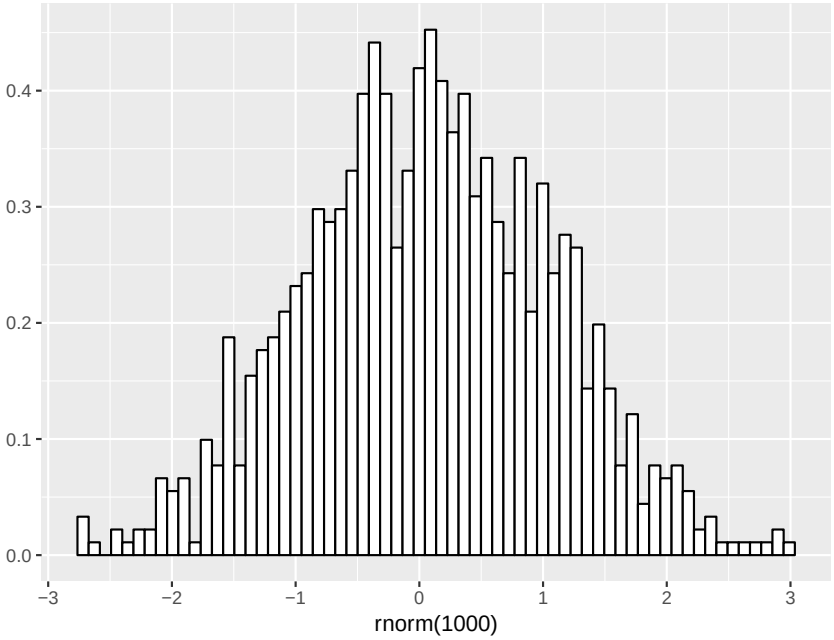
Figure 15:

- **Population Mean**: $\mu = p$
- **Population Standard Deviation**: $\sigma = \sqrt{p(1-p)}$

16 Normal (Gaussian) Distribution

Named after Karl Friedrich Gauss

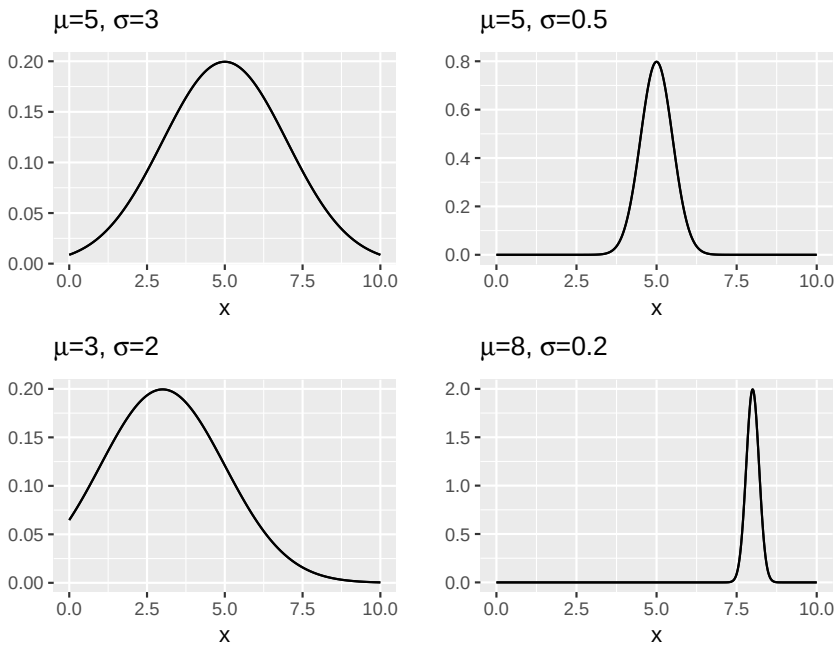
This (for good reasons we will see shortly) is the most important distribution of them all! First, it is already familiar to you because it results in data with bell-shaped histograms:



A normal random distribution has two parameters, denoted by μ and σ.

What is the meaning (interpretation) of the parameters? It is of course that μ is the **population mean** and σ is the **population standard deviation**.

Example: In the next picture we have 4 examples of normal densities with different means and standard deviations, drawn on the same scale:



16.0.1 App

```
run.app(normal)
```

this app draws the histogram of data from a normal distribution with different means and standard deviations.

16.1 Central Limit Theorem

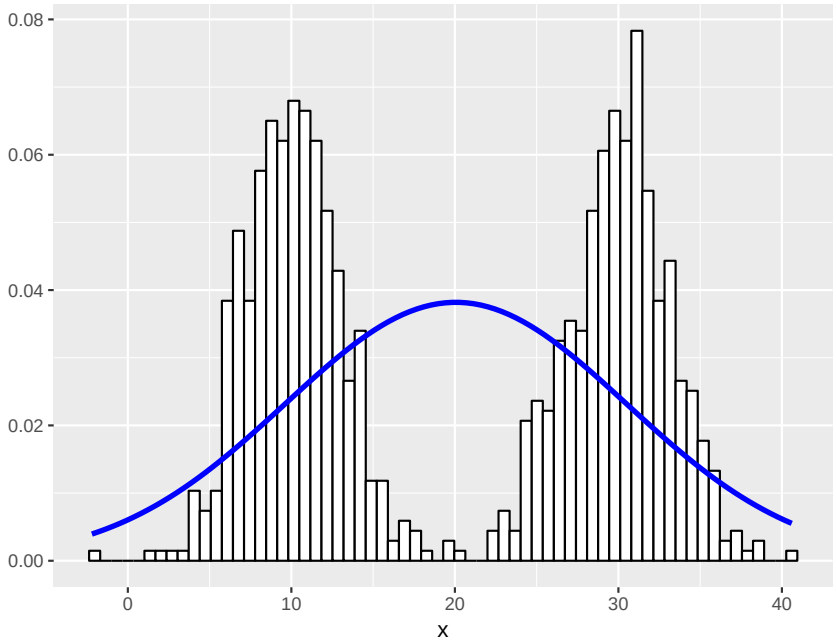
Why is the normal distribution so important? The reason is the **Central Limit Theorem**, which states that under some very general conditions the **sample mean** has (approximately) a normal distribution, no matter what the distribution of the observations.

Example As an illustration let's do the following. We start by getting some data that is very much NOT normally distributed. I have a routine to that called **clt_illustration**. To see what the data looks like run

```
clt.illustration(1)
```

```
## x
## [1] 25.8
## [1] 32.1
## [1] 24.6
## [1] 26
## [1] 23.3
## ...
## [1] 31.4
## [1] 7.1
```

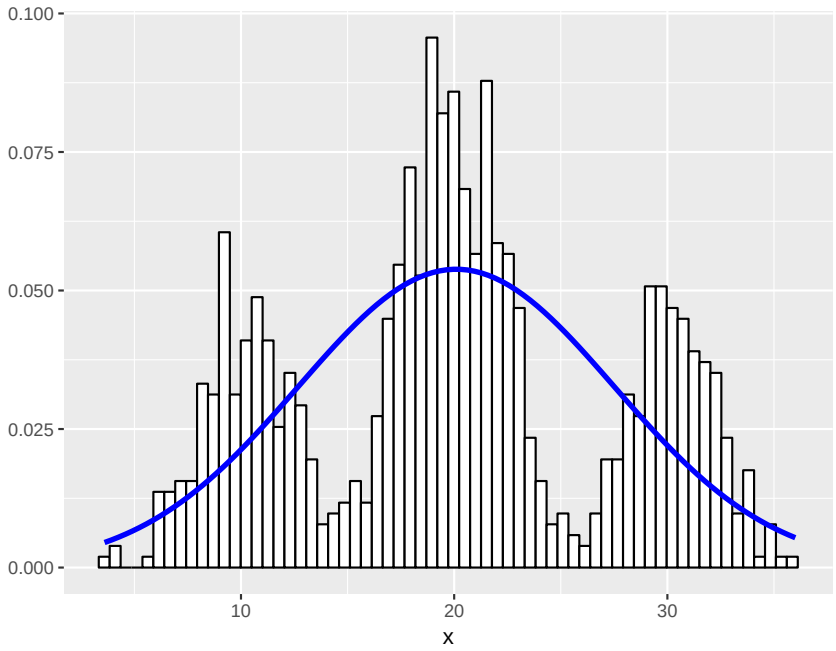
```
## [1] 9.6
## [1] 33.4
## [1] 6.2
```



Now this clearly is not a bell-shaped histogram! Now let's do the following: generate pairs of numbers x1, x2 and find their mean with (x1+x2)/2. We can do this with

```
clt.illustration(2)
```

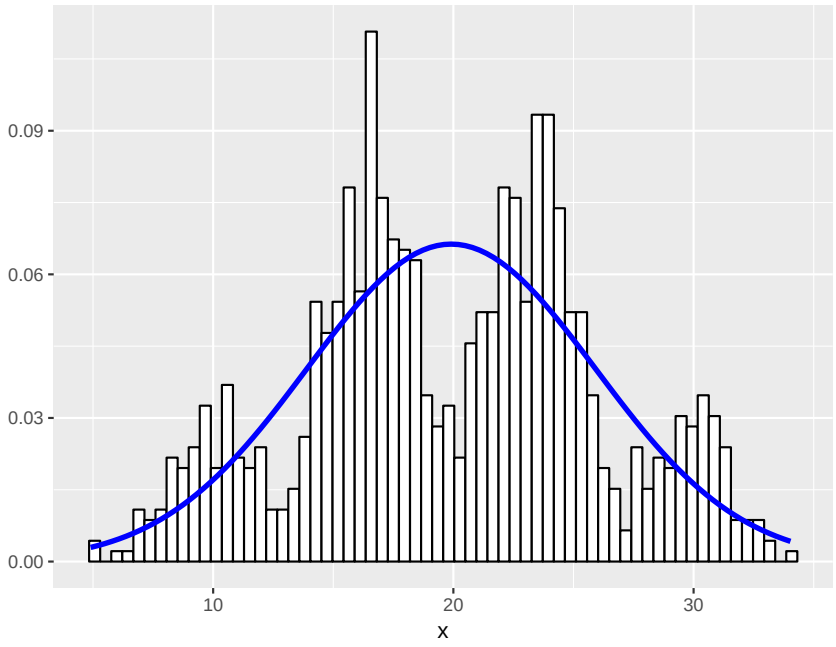
```
## (x1 + x2)/2 = xbar
## (12.7  + 36.2 )/ 2  =  24.45
## (14.7  + 6.7 )/ 2  =  10.7
## (28.4  + 27.7 )/ 2  =  28.05
## (34.2  + 12.4 )/ 2  =  23.3
## (31.1  + 28.5 )/ 2  =  29.8
## ...
## (35.1  + 29.1 )/ 2  =  32.1
## (32.3  + 30.6 )/ 2  =  31.45
## (9.9  + 8.1 )/ 2  =  9
## (14.7  + 28.7 )/ 2  =  21.7
## (13.9  + 10.8 )/ 2  =  12.35
```



Still not much of a bell-shaped histogram. But if we keep numbers to the mean we quickly get there, here is what it looks like for 10:

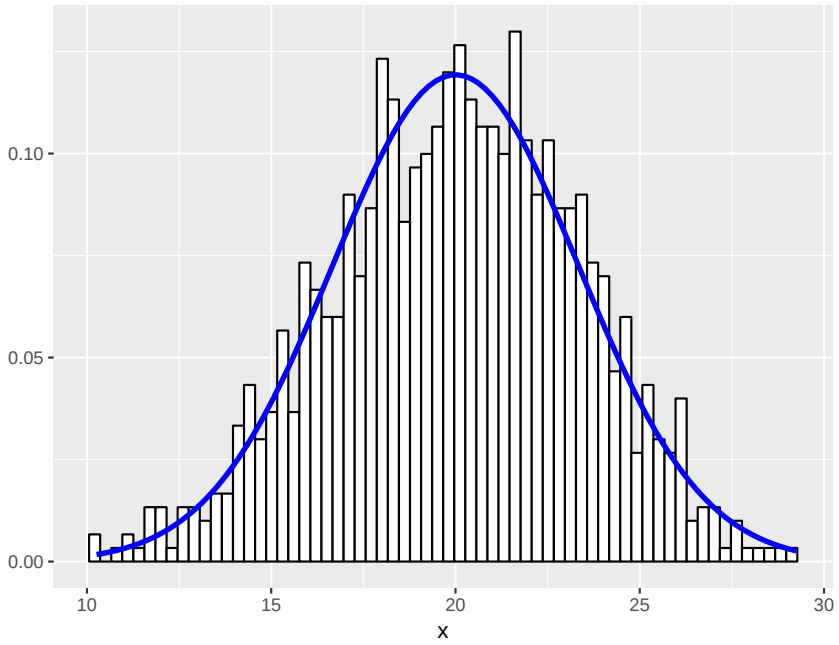
```
clt.illustration(3)
```

```
## (x1 + x2 + x3)/3 = xbar
## (30.8  + 35.4  + 25.9 )/ 3  =  30.7
## (10  + 11.6  + 14.3 )/ 3  =  11.96667
## (31.2  + 30.5  + 8.8 )/ 3  =  23.5
## (29.3  + 10.9  + 37.1 )/ 3  =  25.76667
## (27.6  + 34.7  + 28.9 )/ 3  =  30.4
## ...
## (8.5  + 28.9  + 5.3 )/ 3  =  14.23333
## (14.1  + 7.3  + 25.4 )/ 3  =  15.6
## (4.1  + 9.5  + 14.5 )/ 3  =  9.366667
## (11.9  + 10.5  + 27.7 )/ 3  =  16.7
## (7.7  + 9.3  + 31.4 )/ 3  =  16.13333
```




```
clt.illustration(10)
```

```
## (x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10)/10 = xbar
## (29.4  + 8.4  + 32.6  + 28.7  + 27.5  + 8.1  + 33.7  + 32.1  + 10.9  + 24.6 )/ 10  =
## (25.4  + 27.5  + 30.6  + 27.3  + 32.3  + 12  + 30.7  + 9.6  + 9.6  + 8.6 )/ 10  =  21
## (11  + 31.5  + 13.5  + 32.7  + 8.5  + 27  + 12.3  + 7.1  + 34.6  + 5.3 )/ 10  =  18.3
## (12.1  + 7.2  + 28.1  + 9.3  + 8.6  + 28.7  + 11.4  + 7.9  + 8.7  + 14.5 )/ 10  =  13
## (32.5  + 9.9  + 12.3  + 33.3  + 6.4  + 8.4  + 11.3  + 12.5  + 33.2  + 31 )/ 10  =  19
## ...
## (8.5  + 31.4  + 10.4  + 28.7  + 35.7  + 7.7  + 34.9  + 8.8  + 12  + 22.6 )/ 10  =  20
## (11.4  + 30  + 28.3  + 37.1  + 11.7  + 29.7  + 32.9  + 28.8  + 29.5  + 10.9 )/ 10  =
## (8.9  + 10.8  + 14.6  + 13.4  + 7.1  + 10  + 9.3  + 11  + 33.1  + 8.1 )/ 10  =  12.63
## (10.4  + 9  + 33  + 8  + 30  + 37.2  + 29.2  + 9.2  + 30.6  + 30.6 )/ 10  =  22.72
## (5.1  + 13.3  + 9.3  + 7.7  + 30.1  + 9.6  + 13.1  + 7.9  + 28.5  + 27.2 )/ 10  =  15
```



There is very nice way to illustrate the workings of the central limit theorem called a Galton Board Video

16.1.1 App

this app does various illustrations of the central limit theorem

```
run.app(clt)
```

16.2 Theory of Errors

In real life almost any measuring device makes some errors. Some instruments are lousy and make big ones, other instruments are excellent and make small ones.

Example You want to measure the length a certain streetlight is red. You ask 10 friends to go with you and everyone makes a guess.

Example You want to measure the length a certain streetlight is red. You ask 10 friends to go with you. You have a stopwatch that you give to each friend.

Clearly in the second case we expect to get much smaller errors.

Around 1800 Gauss was thinking about what one could say in great generality about such measurment errors. He came up with the following rules that (almost) all measurement errors should follow, no matter what the instrument:

1. Small errors are more likely than large errors.
2. an error of ϵ is just as likely as an error of $-\epsilon$

3. In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average.

Now it is quite astonishing that JUST FROM THESE THREE rules he was able to derive the normal curve.

For the math people, the mathematical function is

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

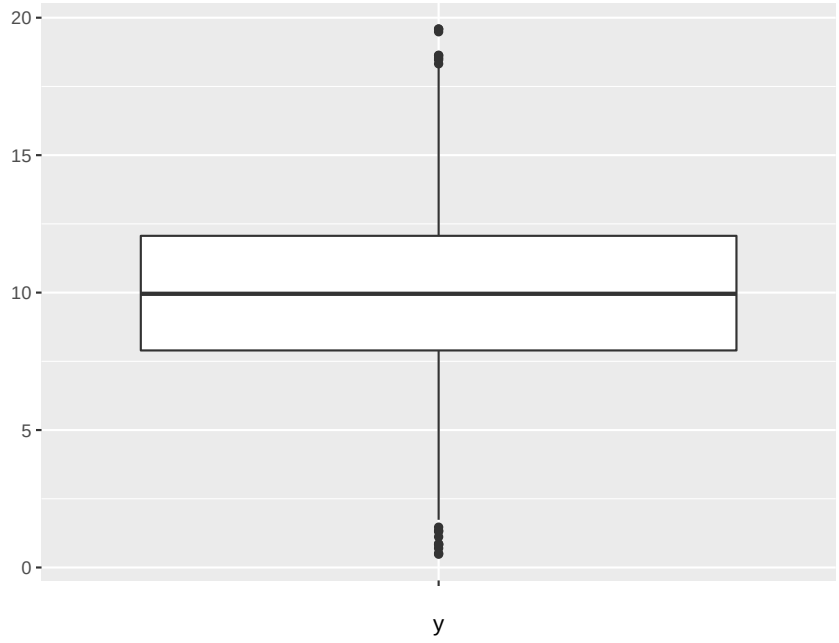
notice it has two very famous math numbers in it, π and e For more on the central limit theorem see page 407 of the textbook.

17 Checking for Normality

As you will see shortly, many of the methods for statistical inference we discuss here (and that are widely used in practise) require the data to come from a normal distribution. How do we check that?

17.1 Boxplot

We will check the assumption of normality via two graphs. The first of these we already talked about previously, namely the boxplot. Here are some boxplots for data from a normal distribution:

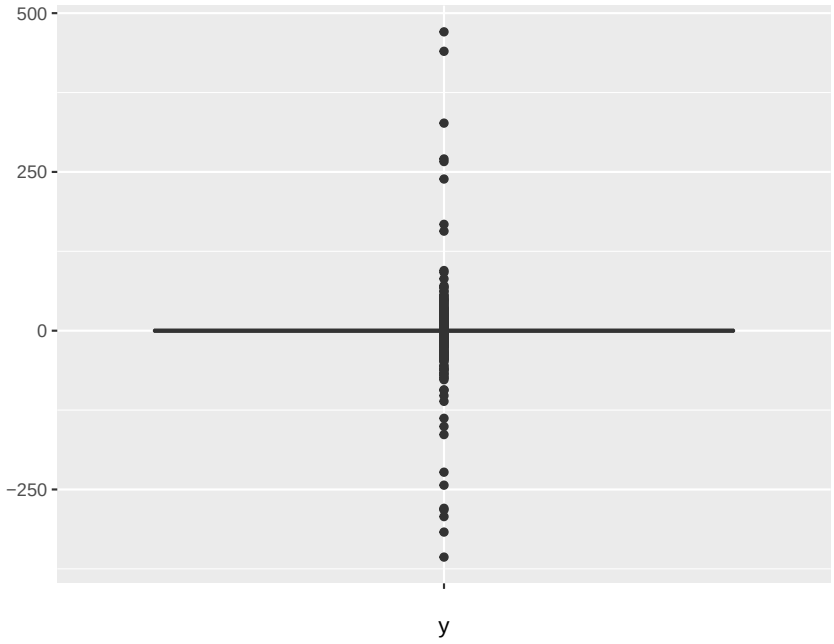


Here are some features of boxplots for normal data:

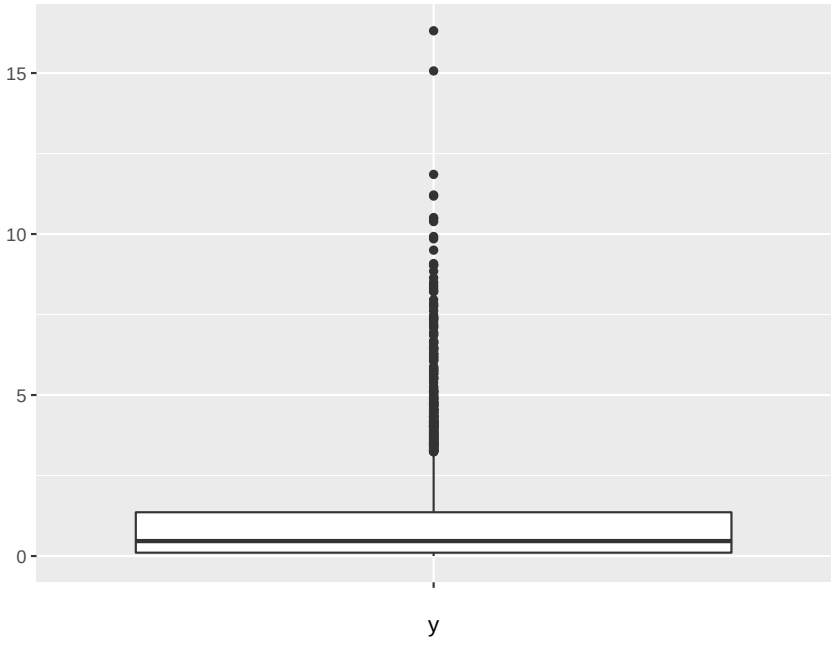
- There are very few "outliers", and those are close to the boxplot

- The lower fence, the box and the upper fence are all about the same size.

Here are some examples of non - normal data:



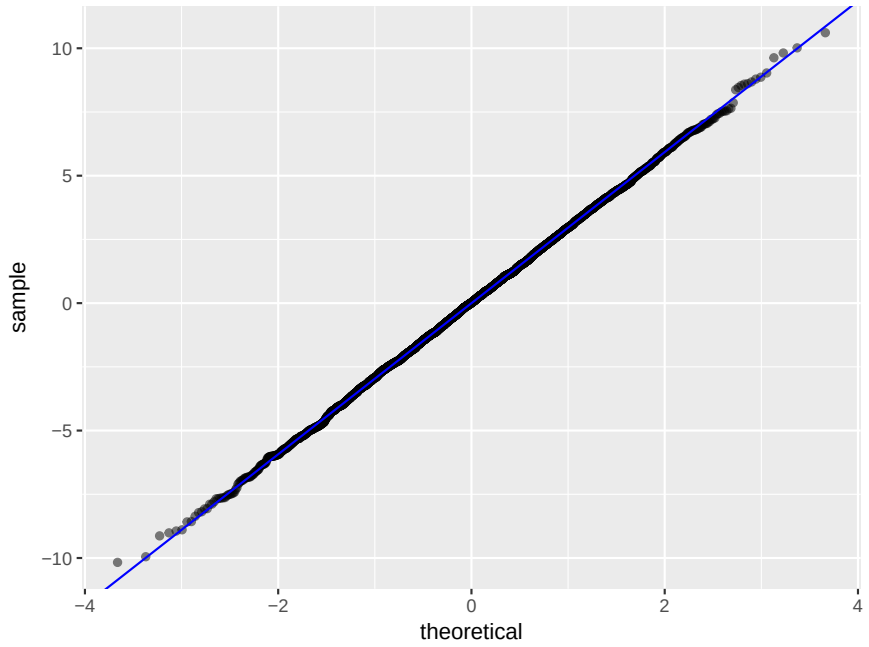
or this one



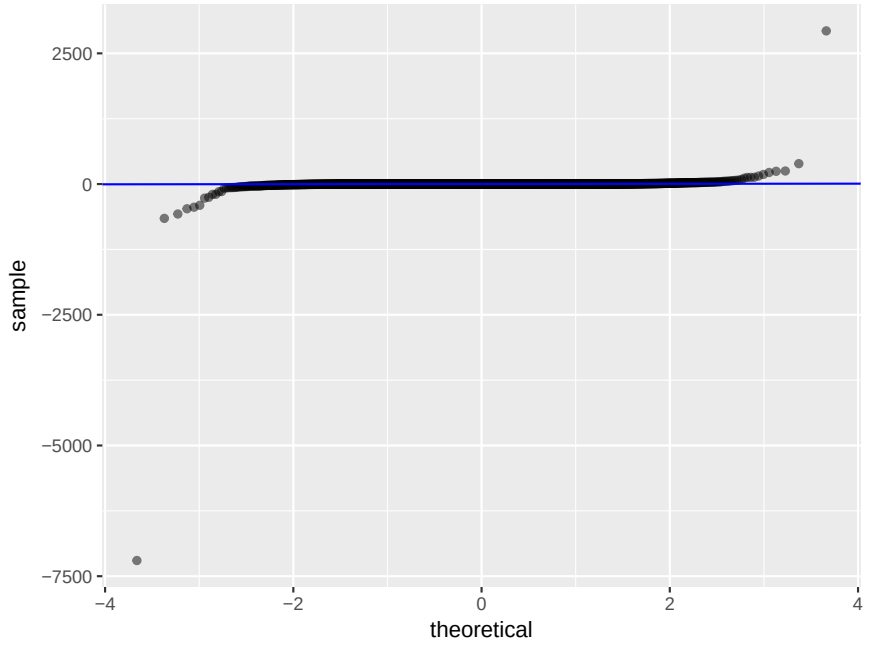
17.2 Normal Probability Plot

This is a graph specifically designed to check for normality. If the data comes from a normal distribution the points should form a line. Again, let's start with some examples of normal

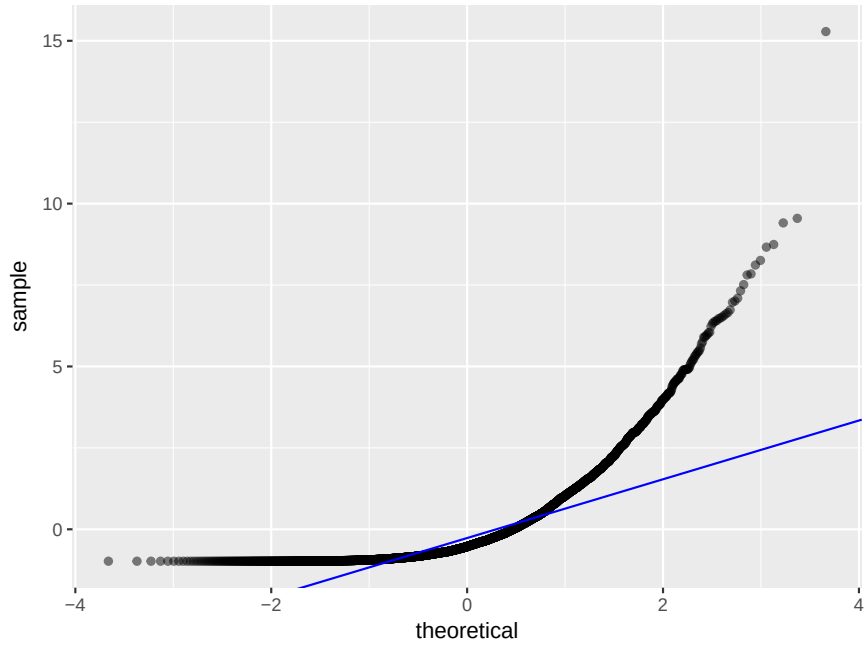
data:



and some examples of non-normal data:



or this one



17.2.1 Case study: Euro coins

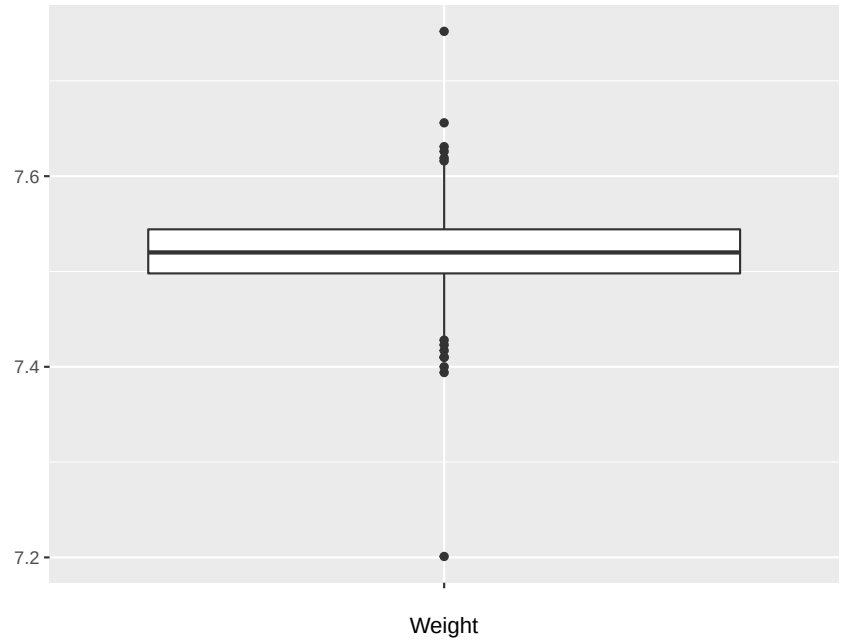
The data is the weight of 2000 1-euro coins, 250 each in eight "rolls". The data were collected by Herman Callaert at Hasselt University in Belgium. The euro coins were "borrowed" at a local bank. Two assistants, Sofie Bogaerts and Saskia Litiere weighted the coins one by one, in laboratory conditions on a weighing scale of the type Sartorius BP 310s.

```
head(euros)
```

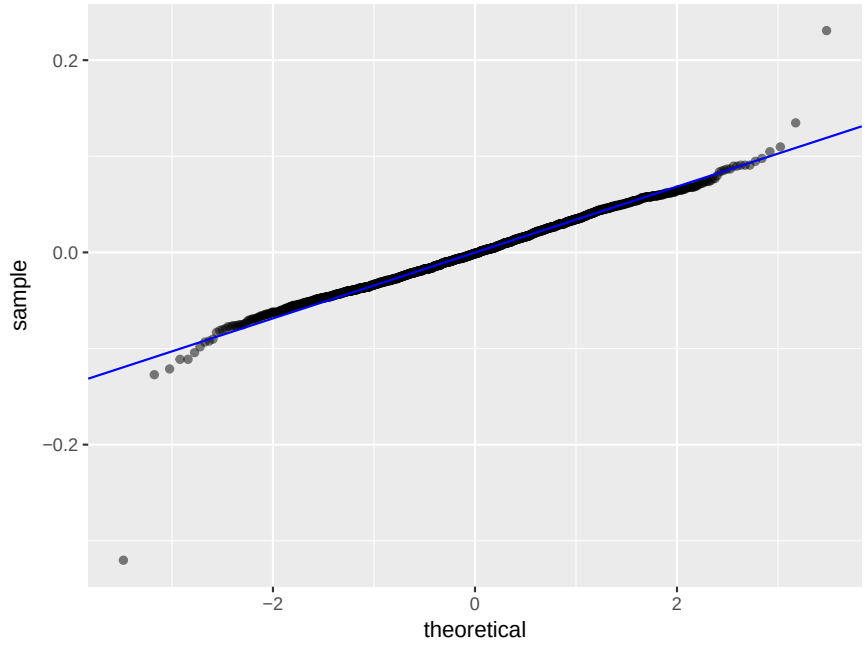
```
##   Weight Roll
## 1  7.512    1
## 2  7.502    1
## 3  7.461    1
## 4  7.562    1
## 5  7.528    1
## 6  7.459    1
```

The manufacturing process of the coins might suggest that the weights have a normal distribution. Is this true?

```
attach(euros)
bplot(Weight)
```




```
nplot(Weight)
```



both the boxplot and the normal probability plot indicate that the data does **not** come from a normal distribution but from some symmetric distribution with **heavier tails**, that is some outliers on both sides.

18 Simulation

A simulation is a way to do experiments on a computer. This can be useful to do various calculations as well as to test the performance of different methods.

A simulation generally consists of the following parts:

- generate random data
- calculate something for this data

- repeat the above many times, keep track of results
- analyze those results

Example: say we are interested in the following: if a fair die is rolled, what is the probability of a six?

Now we already know the answer (1/6) but let's do a simulation anyway.

- generate random data
  in the real live experiment the data are the numbers 1-6, each with probability 1/6. We can generate data like this as follows:

```
x <- sample(1:6, size = 20, replace = TRUE)
x
```

```
##  [1] 3 4 6 4 1 4 2 3 5 2 4 1 3 2 5 2 5 3 6 3
```

- calculate something for this data

we want to know if x is 6 or not. This we can do with

```
y <- (x==6)
rbind(x, y)
```

```
##   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## x    3    4    6    4    1    4    2    3    5     2     4     1     3
## y    0    0    1    0    0    0    0    0    0     0     0     0     0
##   [,14] [,15] [,16] [,17] [,18] [,19] [,20]
## x     2     5     2     5     3     6     3
## y     0     0     0     0     0     1     0
```

- repeat the above many times, keep track of results

in the calculations above we already did the simulation 20 times. In general one wants to do this 10000 times (or so). Of course then we can't print all the stuff to the screen.

```
B <- 10000
x <- sample(1:6, size = B, replace = TRUE)
```

- analyze those results

we want to know the proportion of 6's, so

```
sum(x==6)/B
```

```
## [1] 0.1667
```

Note that each time we run a simulation the computer generates new data, so the answer will always be a little bit different:

```
sum(sample(1:6, size = B, replace = TRUE)==6)/B
```

```
## [1] 0.1699
```

In the example above we could generate all the data (the 10000 x's) in one step. In many examples though this has to be done one at a time. In that case we can use a **for** loop:

```
B <- 10000
y <- rep(0, B)
for(i in 1:B) y[i] <- (sample(1:6, size = 1) == 6)
sum(y)/B
```

```
## [1] 0.1631
```

Sometimes the calculation done inside the for loop needs several steps. If so use { } to keep them together:

```
for(i in 1:B) {
 x <- sample(1:6, size = 1)
 y[i] <- (x == 6)
}
sum(y)/B
```

```
## [1] 0.1687
```

Note When you type this sequence of commands into R after the { the cursor will change from > to +

18.0.1 Generating Data

R has a great many standard probability distributions built in. The general format is **r** followed by the name (sort of). So for example to generate 20 observations from a normal distribution with mean 10 and standard deviation 3 use

```
rnorm(20, 10, 3)
```

```
##  [1] 17.491968 11.409922  9.354232  9.540684  7.207115  8.351231  8.937012
##  [8] 12.725255  8.739160 10.634908 10.957552 16.082267 11.253672  8.953572
## [15]  8.864783 11.907892 11.064801  7.152860  9.279088 11.858517
```

The sample command above generally works for distributions that take finite discrete values with different probabilities. You can also generate many values in one step.

Example: say we roll a fair die 5 times. What is the probability of no sixes?

```
B <- 10000
y <- rep(0, B)
for(i in 1:B) {
```

```
   x <- sample(1:6, size=5, replace = TRUE)
   z <- (x == 6)
   y[i] <- (sum(z) == 0)
}
sum(y)/B
```

```
## [1] 0.4079
```

Example: say we flip a fair coin 20 times. What is the probability of at most 7 heads?

a coin comes up "heads" or "tails", so flipping it once we could do with

```
sample(c("heads","tails"), size=1)
```

```
## [1] "tails"
```

Now of course we want to flip the coin 20 times , so we need:

```
x <- sample(c("heads","tails"), size = 20, replace=TRUE)
x
```

```
##  [1] "tails" "heads" "tails" "tails" "heads" "tails" "heads" "heads"
##  [9] "heads" "heads" "heads" "tails" "heads" "tails" "tails" "tails"
## [17] "tails" "tails" "tails" "tails"
```

Next we need to figure out how many "heads" we have. Here are two ideas:

```
sum(x == "heads")
```

```
## [1] 8
```

```
table(x)["heads"]
```

```
## heads
##     8
```

Now we can do the whole simulation:

```
B <- 10000
y <- rep(0,B)
for(i in 1:B) {
  x <- sample(c("heads","tails"), size = 20, replace = TRUE)
  if(sum(x == "heads") <= 7) y[i] <- 1
}
sum(y)/B
```

```
## [1] 0.1286
```

Example: say it is known that people from a certain population are 40% white, 25% black, 20% hispanic and 15% others. If we randomly select 10 of them, what is the probability that there is at least one of each group?

First, we are selecting 10 people who are either white, black, hispanic or other, we again can do that with the sample command:

```
races <- c("white", "black", "hispanic","other")
sample(races , size = 10, replace = TRUE)
```

```
##  [1] "hispanic" "black"    "black"    "other"    "hispanic" "other"
##  [7] "white"    "white"    "black"    "other"
```

but they are in different proportions in the population, so we need to use

```
p <- c(40, 25, 20, 15)/100
x <- sample(races, size = 10, replace = TRUE, prob = p)
x
```

```
##  [1] "hispanic" "hispanic" "white"    "other"    "black"    "other"
##  [7] "hispanic" "black"    "white"    "hispanic"
```

Next we need some way to check whether we have one of each group in x. (Clearly the case here but we need to do this automatically, not by checking ourselves). So is there a white person among our sample?

```
x == "white"
```

```
##  [1] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
```

if there is we have at least one TRUE, if not all of them will be FALSE. Remember that we can add these up, with FALSE = 0 and TRUE = 1:

```
sum(x == "white")
```

```
## [1] 2
```

so if sum(x== "white") > 1 there was at least one white person, if sum(x == "white") = 0 there is none. Say we set y = 0 and then run

```
y <- 0
if(sum(x == "white") == 0) y <- 1
if(sum(x == "black") == 0 ) y <- 1
if(sum(x == "hispanic") == 0) y <- 1
if(sum(x == "other") == 0) y <- 1
y
```

```
## [1] 0
```

what happens? If there are no whites, on the first line y is set to 1, if there are no blacks, on the second line y is set to 1 and so on.

If there is at least one of each, y is never set to 1, it stays at 0!

So now we put it all together:

```
B <- 10000
y <- rep(0,B)
for(i in 1:B) {
  x<-sample(races, size = 10, replace = TRUE, prob=p)
 if(sum(x == "white") == 0) y[i] <- 1
```

```
 if(sum(x == "black") == 0) y[i] <- 1
 if(sum(x == "hispanic") == 0) y[i] <- 1
 if(sum(x == "other") == 0) y[i] <- 1
}
sum(y)/B
```

```
## [1] 0.3443
```

Note In the above simulation we used the exact words from the real live problem. This is not actually necessary, we can easily "code" those details:

```
B <- 10000
y <- rep(0,B)
for(i in 1:B) {
 x <- sample( 1:4, size = 10, replace = TRUE, prob =p)
 for(j in 1:4)
   if(sum(x == j) == 0) y[i] <- 1
}
sum(y)/B
```

```
## [1] 0.3458
```

also, we often have different ways to do things. In the current ploblem notice the following:

```
x1 <- sample(races, size = 10, replace = TRUE, prob =p)
x1
```

```
##  [1] "white"    "hispanic" "white"    "white"    "black"    "black"
##  [7] "white"    "white"    "black"    "black"
```

```
table(x1)
```

```
## x1
##    black hispanic    white
##        4        1        5
```

```
x2 <- sample(races, size = 10, replace = TRUE, prob =p)
x2
```

```
##  [1] "black"    "black"    "black"    "black"    "other"    "hispanic"
##  [7] "white"    "hispanic" "other"    "white"
```

```
table(x2)
```

```
## x2
##    black hispanic    other    white
##        4        2        2        2
```

In the first case one of the 4 races ("other") is missing whereas in the second all are there. We can distinguish between them with

```
length(table(x1))
```

```
## [1] 3
```

```
length(table(x2))
```

```
## [1] 4
```

So if length(table(x))<4 we do not have one of each race, and so we can rewrite the simulation again:

```
B <- 10000
y <- rep(0,B)
for(i in 1:B) {
 x <- sample( 1:4, size = 10, replace = TRUE, prob =p)
 y[i] <- (length(table(x)) < 4)
}
sum(y)/B
```

```
## [1] 0.3495
```

Notice that this is quite a bit less to type. On the other hand it is not as clear what the program does. At least in the beginning I recommend to stay as close to the real live problem as possible!

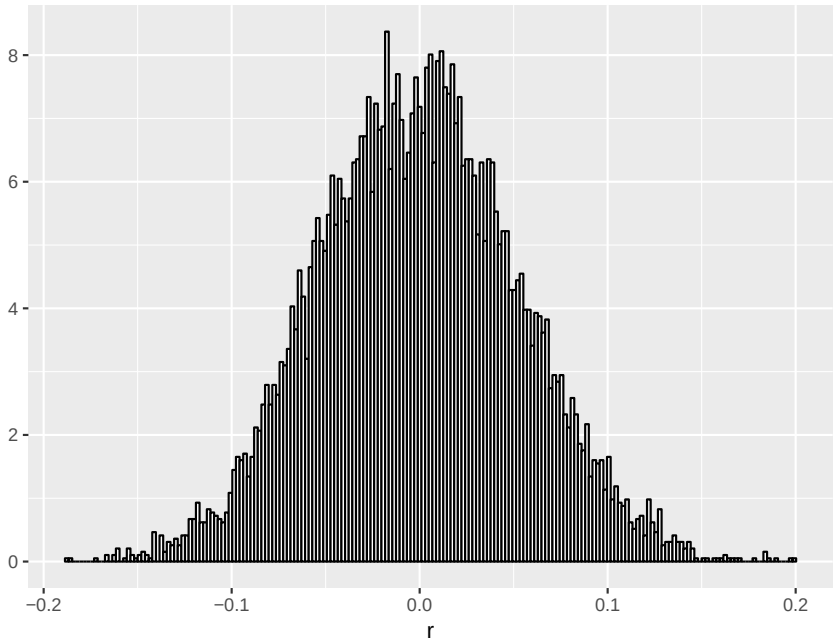
Example: say we have data from a normal distribution with mean 100 and standard deviation 25. What is the probability that an observation is less than 90?

```
B <- 10000
x <- rnorm(B, 100, 25)
sum(x<90)/B
```

```
## [1] 0.3424
```

Example use simulation to generate the histogram of correlation coefficients from the Draft data.

```
B <- 10000
r <- rep(0, B)
for(i in 1:B) {
  x <- sample(1:366, size=366)
  r[i] <- cor(1:366, x)
}
hplot(r)
```



Example: say we roll a fair die 10 times. What is the mean and the standard deviation of the number of sixes?

The number of "successes" in a sequence of Bernoulli trials is one of those special distributions. It is called the *Binomial* distribution and we can generate data from it with **rbinom(B, n, p)** where n is the number of trial and p is the success probability. So

```
B <- 10000
x <- rbinom(B, 10, 1/6)
mean(x)
```

```
## [1] 1.665
```

```
sd(x)
```

```
## [1] 1.168294
```

Example: In the beginning of the discussion on probability I had a table of probabilities. Let's recreate it!

Say we flip a fair coin 1000 times. What is the probability of 525 or more heads?

```
B <- 10000
x <- rbinom(B, 1000, 1/2)
sum(x>=525)/B
```

```
## [1] 0.058
```

Now we had that table for 500, 505, .., 550, so:

```
B <- 10000
a <- 0:10*5+500
y <- rep(0, 11)
x <- rbinom(B, 1000, 1/2)
```

```
for(i in 1:11)
  y[i] <- sum(x>=a[i])/B
cbind(a, y)
```

```
##          a      y
##  [1,] 500 0.5201
##  [2,] 505 0.3923
##  [3,] 510 0.2782
##  [4,] 515 0.1820
##  [5,] 520 0.1095
##  [6,] 525 0.0633
##  [7,] 530 0.0315
##  [8,] 535 0.0142
##  [9,] 540 0.0066
## [10,] 545 0.0029
## [11,] 550 0.0008
```

Example: At the end of the discussion on probability I had told you the formulas for the mean and the standard deviation of a Bernoulli trial: $\mu = p$ and $\sigma = \sqrt{p(1-p)}$. Let's test that!

```
p <- 0.5
B <- 10000
x <- sample(0:1, size=B, replace=TRUE, prob=c(1-p, p))
round(c(p, mean(x)), 4)
```

```
## [1] 0.5000 0.5067
```

```
round(c(sqrt(p*(1-p)), sd(x)), 4)
```

```
## [1] 0.5 0.5
```

```
p <- 0.15
B <- 10000
x <- sample(0:1, size=B, replace=TRUE, prob=c(1-p, p))
round(c(p, mean(x)), 4)
```

```
## [1] 0.1500 0.1471
```

```
round(c(sqrt(p*(1-p)), sd(x)), 4)
```

```
## [1] 0.3571 0.3542
```

```
p <- 0.75
B <- 10000
x <- sample(0:1, size=B, replace=TRUE, prob=c(1-p, p))
round(c(p, mean(x)), 4)
```

```
## [1] 0.7500 0.7552
```

```r
round(c(sqrt(p*(1-p)), sd(x)), 4)
```

```
## [1] 0.433 0.430
```

19 Exercise 2: Probability and Simulation

19.0.1 Problem 1

a. The routine **normal.ex(k)** generates data from different probability distributions. Decide for which k (from 1 to 10) the data comes from a normal distribution.

b. for those that do not come from a normal distribution, how large a sample size is needed so that the mean has a normal distribution?

19.0.2 Problem 2

Say a huge box contains 5000 red balls, 3000 blue balls and 1000 green balls. Write a simulation that finds the smallest number of balls we need to pick out of the box so that the probability we get at least one of each color is 90%.

19.0.3 Problem 3

Say we flip a fair coin n times. Write a simulation that finds the smallest n so that the probability we get at least one of each heads is 99%.

19.0.4 Problem 4

Say we roll a fair die in n times. Write a simulation that finds the smallest n so that the probability we get at least two sixes is 90%.

19.0.5 Problem 5

Write a simulation that checks the empirical rule when generating data from

a. rnorm(50)

b. rt(50,1)

19.0.6 Problem 6

A standard deck as used to play Poker consists of 52 cards. Each card has a denomination (numbers 2 to 10, Jack, Queen, King and Ace) and a suit (Hearts, Clubs, Diamonds and Spades. In many games a player first gets 5 cards (called a hand). Use simulation to find the probability that his hand is a

a. “Three of a kind” (3 cards of the same denomination)

b. “Flush” (all 5 cards of the same suit)

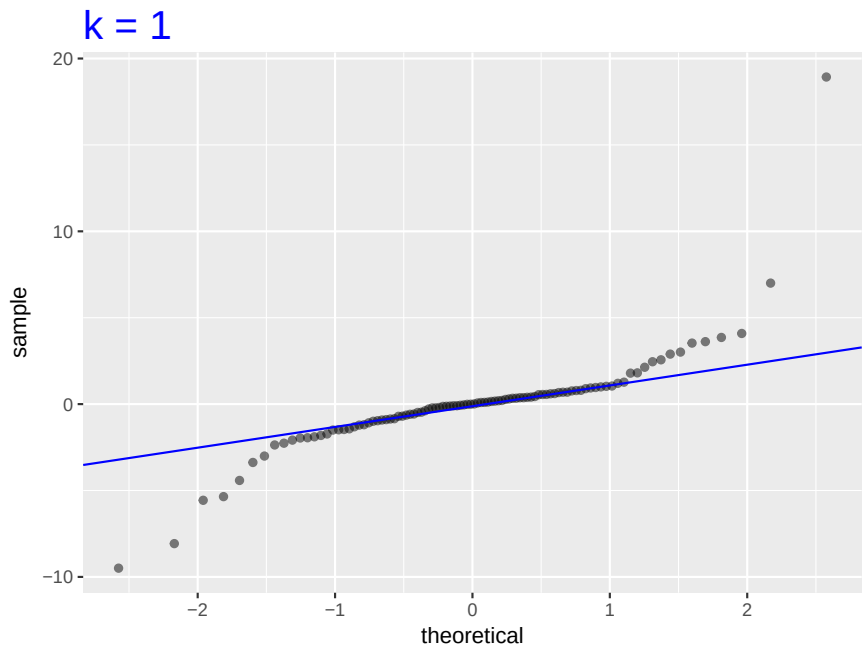
19.1 Solutions

19.1.1 Problem 1

The routine **normal.ex(k)** generates data from different probability distributions. Decide for which k (from 1 to 10) the data comes from a normal distribution.

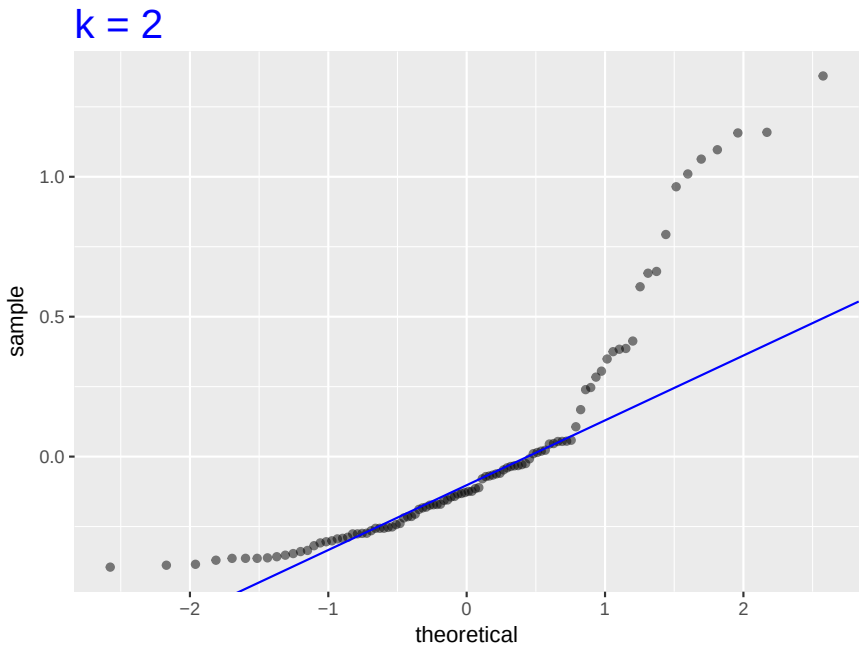
We will use the normal plot to check. So for the first case we can run

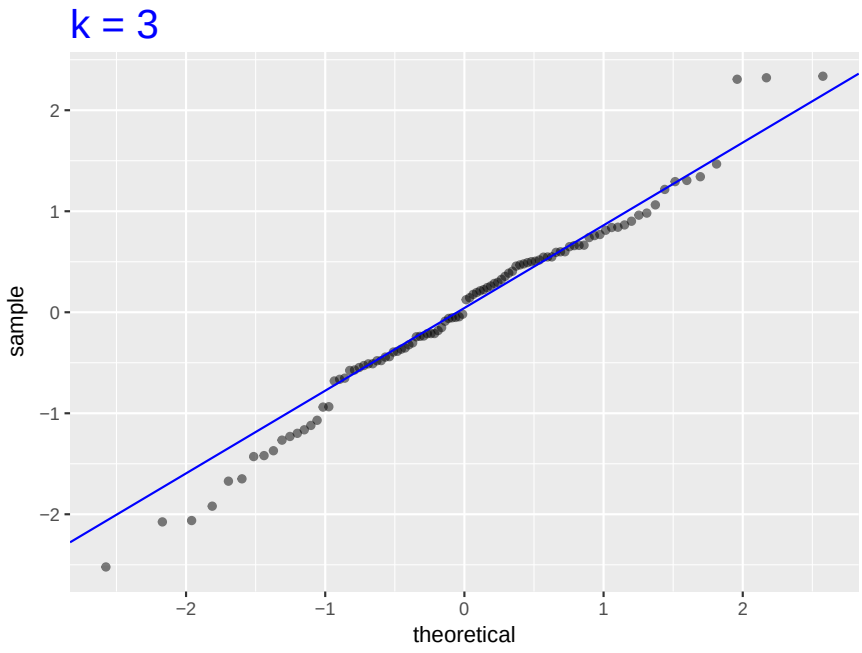
```
nplot(normal.ex(1), main_title = "k = 1")
```

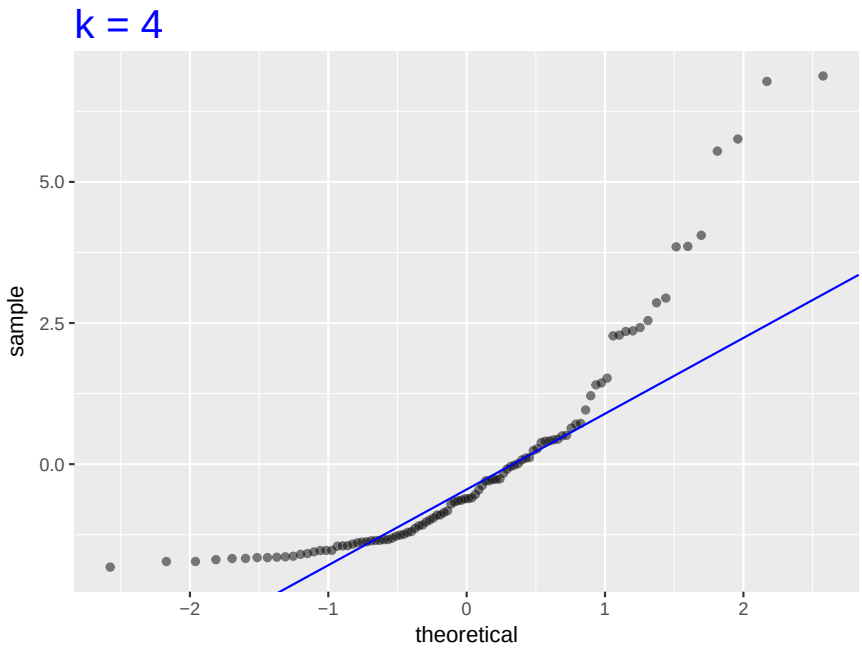


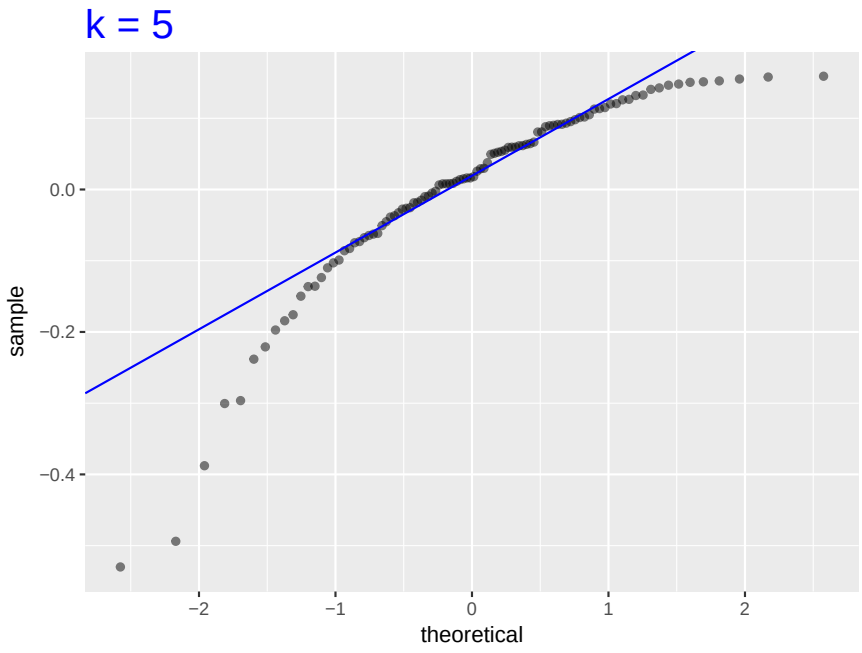
Here are all the graphs:

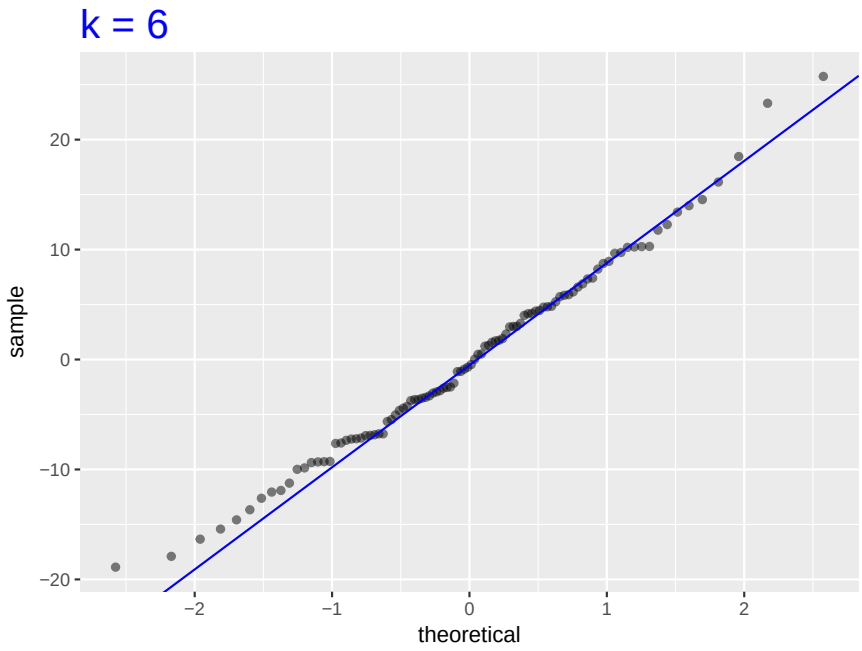
```
for(i in 2:10) nplot(normal.ex(i), main_title = paste("k =", i))
```

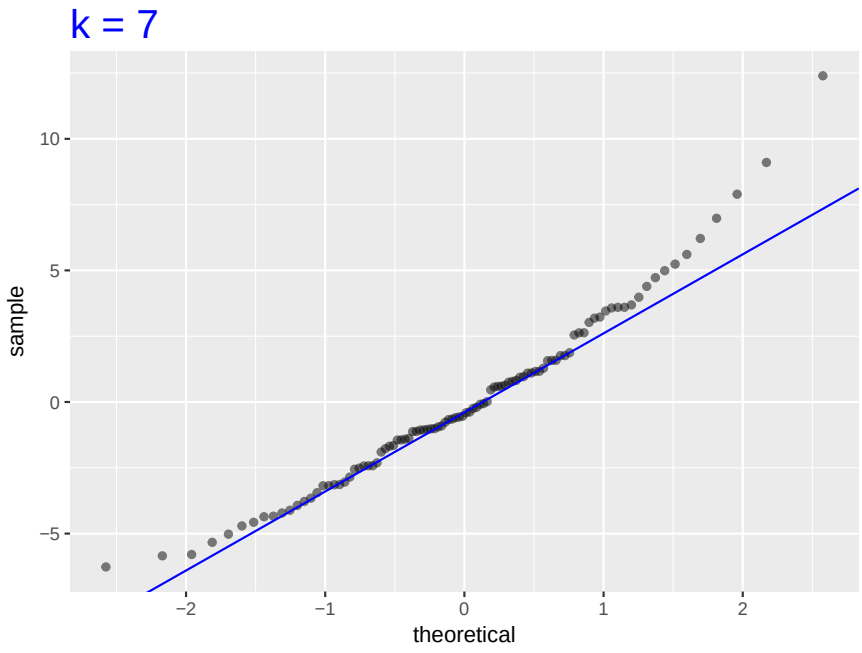

k = 2
sample
theoretical
1.0
0.5
0.0
−2
−1
0
1
2

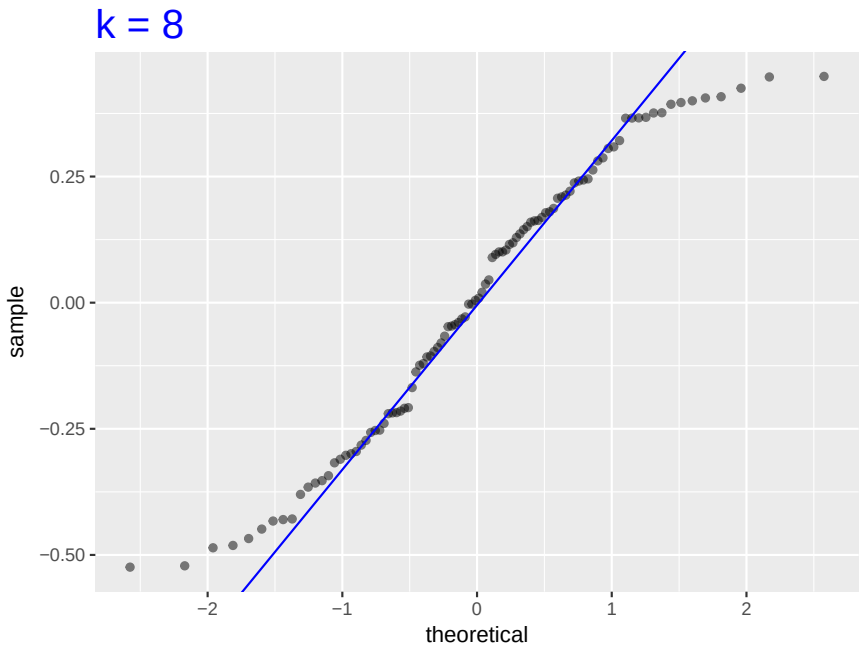

k = 3
sample
theoretical
2
1
0
−1
−2
−2
−1
0
1
2

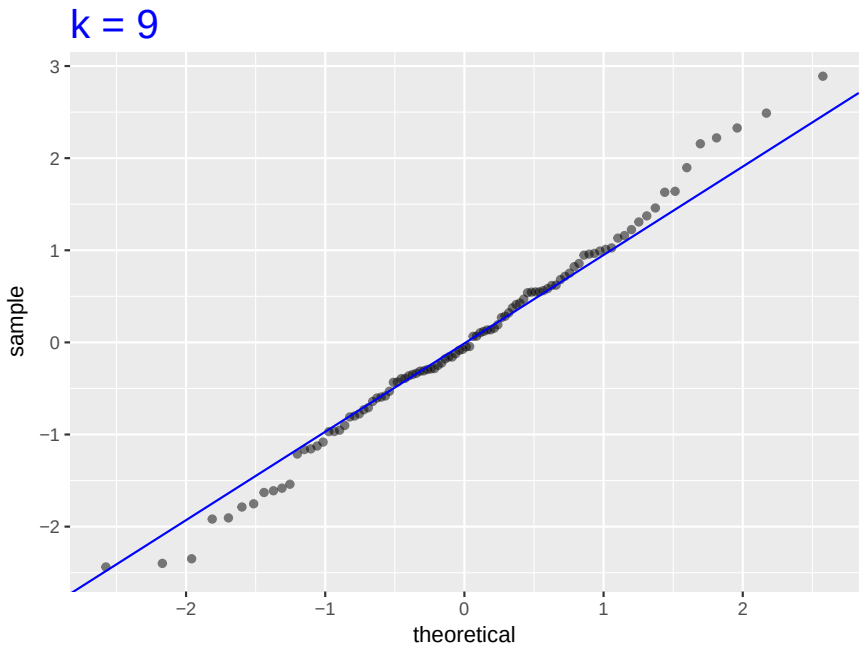

k = 4
5.0
2.5
0.0
sample
−2
−1
0
1
2
theoretical

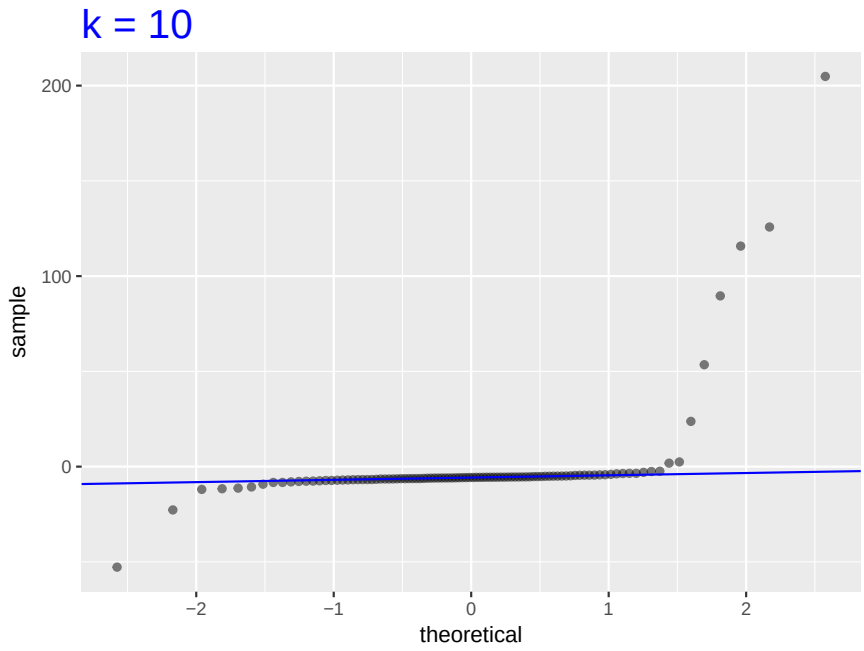

k = 5
0.0
−0.2
−0.4
sample
−2
−1
0
1
2
theoretical


k = 6
sample
theoretical
20
10
0
−10
−20
−2
−1
0
1
2


k = 7
sample
theoretical
10
5
0
−5
−2
−1
0
1
2


k = 8
0.25
0.00
−0.25
−0.50
sample
−2
−1
0
1
2
theoretical


k = 9
3
2
1
0
−1
−2
sample
−2
−1
0
1
2
theoretical

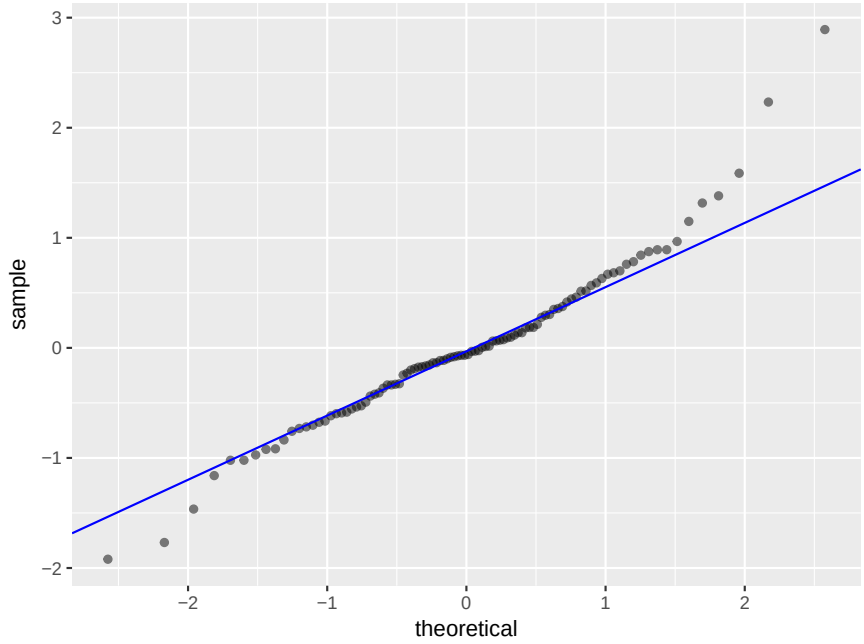


It seems k = 3, 6 and 9 have a normal distribution

b. for those that do not come from a normal distribution, how large a sample size is needed so that the mean has a normal distribution?

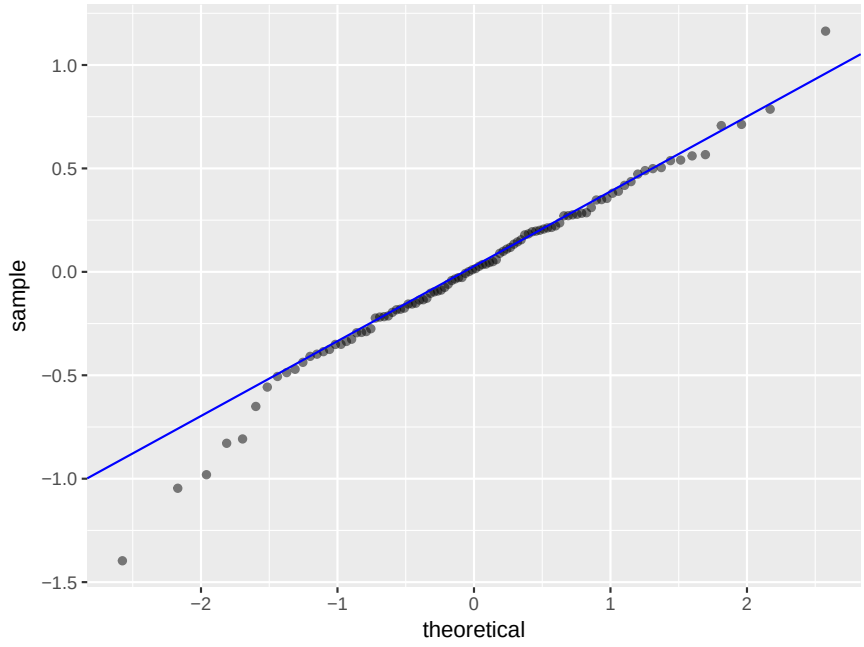
for this we keep generating data, add them to the vector and check whether the result has anormal distribution. For example

```
xbar <- normal.ex(1)
for(i in 1:10) xbar <- xbar + normal.ex(1)
nplot(xbar/11)
```



well not quite yet. Maybe 50?

```
xbar <- normal.ex(1)
for(i in 1:50) xbar <- xbar + normal.ex(1)
nplot(xbar/51)
```



and that's not so bad.

For the others:

k=2: about 30
k=4 about 10
k=5 about 10
k=7 about 10
k=8 about 10
k=10 never!

19.1.2 Problem 2

Say a huge box contains 5000 red balls, 3000 blue balls and 1000 green balls. Write a simulation that finds the smallest number of balls we need to pick out of the box so that the probability we get at least one of each color is 90%.

say we want to find the probability of at least one ball of each color if we pick 10 balls:

```
B <- 10000
y <- rep(0, B)
for(i in 1:B) {
 x <- sample( c("red", "blue", "green"),
    size = 10, replace = TRUE, prob = c(5, 3, 1))
 if(sum(x == "red") == 0) y[i] <- 1
```

```
 if(sum(x == "blue") == 0) y[i] <- 1
 if(sum(x == "green") == 0) y[i] <- 1
}
sum(y)/B
```

```
## [1] 0.3289
```

so the probability of at least one of each color is 1-0.329 = 0.671, a little to small. Next let's try 20 balls:

```
B <- 10000
y <- rep(0,B)
for(i in 1:B) {
 x <- sample( c("red", "blue", "green"), size = 20,
    replace = TRUE, prob=c(5, 3, 1))
 if(sum(x == "red") == 0) y[i] <- 1
 if(sum(x == "blue") == 0) y[i] <- 1
 if(sum(x == "green") == 0) y[i] <- 1
}
sum(y)/B
```

```
## [1] 0.0926
```

and now the probability of at least one of each color is 1-0.093 = 0.907, almost there.

Would 19 have been enough? Let's see:

```
B <- 10000
y <- rep(0,B)
for(i in 1:B) {
 x <- sample( c("red", "blue", "green"), size = 19,
    replace = TRUE, prob=c(5, 3, 1))
 if(sum(x == "red") == 0) y[i] <- 1
 if(sum(x == "blue") == 0) y[i] <- 1
 if(sum(x == "green") == 0) y[i] <- 1
}
1 - sum(y)/B
```

```
## [1] 0.8918
```

and the answer is no!

19.1.3 Problem 3

Say we flip a fair coin n times. Write a simulation that finds the smallest n so that the probability we get at least one of each heads is 99%.

let's try 10 heads:

```r
B <- 10000
y <- rep(0,B)
for(i in 1:B) {
 x <- sample(0:1, size = 10, replace = TRUE)
 if(sum(x) == 0) y[i] <- 1
}
1 - sum(y)/B
```

```
## [1] 0.9995
```

a bit to large. Now trying different n's we find

n = 9: 0.9980
n = 8: 0.9961
n = 7: 0.9922
n = 6: 0.9844

so n = 7 is just large enough.

19.1.4 Problem 4

Say we roll a fair die in n times. Write a simulation that finds the smallest n so that the probability we get at least two sixes is 90%.

the probability of a six is 1/6, so the probability of two sixes is 1/6*1/6 = 1/36, so we would expect two or more sixes every 36 rolls or so. Let's try that:

```r
B <- 10000
y <- rep(0, B)
for(i in 1:B) {
 x <- sample(0:1, size = 36, replace = TRUE,
    prob = c(5,1)/6)
 if(sum(x) > 1) y[i] <- 1
}
sum(y)/B
```

```
## [1] 0.9876
```

a bit to large, so let's try some smaller n's:

- n=30: p=0.97
- n=25: p=0.94

- n=20: p=0.87

- n=22: p=0.9
- n=21: p=0.89 and so we find n=22!

19.1.5 Problem 5

Write a simulation that checks the empirical rule when generating data from

a. rnorm(1000)

```
B <- 10000
y <- rep(0, B)
for(i in 1:B) {
   x <- rnorm(1000)
   low <- mean(x) - 2*sd(x)
   high <- mean(x) + 2*sd(x)
   z <- x[x > low]
   z <- z[z < high]
   y[i] <- length(z)
}
mean(y)
```

```
## [1] 954.6478
```

the interval should include 95% of the data, 95% of 1000 is 950, it includes 954.6, so that's good.

b. rt(1000, 1)

```
B <- 10000
y <- rep(0, B)
for(i in 1:B) {
   x <- rt(1000, 1)
   low <- mean(x) - 2*sd(x)
   high <- mean(x) + 2*sd(x)
   z <- x[x > low]
   z <- z[z < high]
   y[i] <- length(z)
}
mean(y)
```

```
## [1] 989.9439
```

the interval should include 95% of the data, 95% of 1000 is 950, it includes 989.9, so that's not good

---

19.1.6 Problem 6

A standard deck as used to play Poker consists of 52 cards. Each card has a denomination (numbers 2 to 10, Jack, Queen, King and Ace) and a suit (Hearts, Clubs, Diamonds and Spades. In many games a player first gets 5 cards (called a hand). Use simulation to find the probability that his hand is a

a. “Three of a kind” (3 cards of the same denomination)

b. “Flush” (all 5 cards of the same suit)

Let’s denote the denominations by numbers 1-13. Each of them appears 4 times in the deck, so we can use

```
den <- rep(1:13, 4)
```

We draw 5 cards without repitition, so we can do that with

```
x <- sample(den,5)
```

In a “Three of a kind” we then need 3 equal numbers, for example (1, 5, 5, 5, 7) or (4, 6, 10, 10, 10). We can use the table command to get how many of each we have

```
tbl.x <- table(x)
tbl.x
```

```
## x
##  6  7  8 12
##  2  1  1  1
```

Now if there is a three, it must be the largest number (because there are only 5 cards), so we can check with

```
if(max(tbl.x) == 3)
```

and we can do all of it in one step

```
if(max(table(sample(rep(1:13, 4), 5))) == 3)
   y[i] <- 1
```

How about the “Flush”? Now we are interested in the denominations so we pick them with sample(rep(1:4, 13), 5). In a “Flush” all of them are the same number, for example (3, 3, 3, 3, 3) and we can check this with if(length(table()) == 1). With this we have

```
if(length(table(sample(rep(1:4, 13), 5))) == 1)
   y[i] <- 1
```

Now:

```
B <- 10000
y <- rep(0,B)
y1 <- rep(0, B)
for(i in 1:B) {
   if(max(table(sample(rep(1:13, 4), 5))) == 3)
      y[i] <- 1
   if(length(table(sample(rep(1:4, 13), 5))) == 1)
      y1[i] <- 1
}
sum(y)/B
```

```
## [1] 0.023
```

```
sum(y1)/B
```

```
## [1] 0.0018
```

20 Population - Sample

Now we can go back to Statistics. To begin with, recall the following:

Population: all of the entities (people, events, things etc.) that are the focus of a study

Sample: any subset of the population

Parameter: any numerical quantity associated with a population

Statistic: any numerical quantity associated with a sample After our discussion of probability we can now be a little bit more precise

Example Say we roll a fair die until the first time we get a six. We always are in one of two situations:

20.1 Theoretical

We have not actually rolled any die, maybe we don't even have a die, we are just studying this exercise theoretically. That is we are studying the **population** of all possible outcomes of the experiment "Number of rolls of a fair die until a six". We now have a **theoretical description** of this experiment., namely its distribution.

For the first couple of values of k we find

k	Probability
1	0.167
2	0.139
3	0.116
4	0.096
5	0.080
6	0.067

There are formulas for all sorts of numbers for various distributions. For ours we have

- $\mu = 6.0$
- $\sigma = 5.48$
- third quartile Q_3=8
- 95^{th} Percentile P_{95}=16

and so on.

But because they are computed for the whole population they are **parameters.**

20.2 Practical

On the other hand we can study this exercise by actually rolling a fair die many times and observing what happens. Actually (a lot faster and less work!) we can do a simulation. The distribution discribing this experiment is called a **geometric** and we can generate data with **rgeom(B, 1/6)+1**

```
B <- 10000
x <- rgeom(B,1/6)+1
```

Now we can use this dataset to find probabilities:

```
round(table(x)/B, 3)
```

```
## x
##     1     2     3     4     5     6     7     8     9    10    11    12
## 0.167 0.137 0.120 0.091 0.080 0.066 0.057 0.053 0.037 0.035 0.023 0.020
##    13    14    15    16    17    18    19    20    21    22    23    24
## 0.020 0.016 0.012 0.012 0.008 0.007 0.006 0.006 0.005 0.004 0.004 0.002
##    25    26    27    28    29    30    31    32    33    34    35    36
## 0.002 0.002 0.001 0.001 0.001 0.001 0.001 0.001 0.001 0.000 0.000 0.000
##    37    40    41    43    44    47    52
## 0.000 0.000 0.000 0.000 0.000 0.000 0.000
```

and of course we can find the summary statistics:

- mean

```
round(mean(x), 2)
```

```
## [1] 6.01
```

- standard deviation s

```
round(sd(x), 2)
```

```
## [1] 5.47
```

- third quartile Q_3 and 95^{th} Percentile P_{95}

```
round(quantile(x, c(0.75, 0.95)), 2)
```

```
## 75% 95%
##   8  17
```

Because these are computed from a sample they are **statistics**

20.2.1 Population - Sample

The real powerful idea here is to **combine** these two approaches: Say we have a die that we suspect is **not** a fair die and we wish to test this. So we roll the die and then compare the practical results with the theoretical ones. For our die we find (for one run of the simulation)

k	Theory	Sample
1	0.167	0.167
2	0.139	0.137
3	0.116	0.120
4	0.096	0.091
5	0.080	0.080
6	0.067	0.066

or we can do this by looking at some summaries:

- Mean: $\mu = 6$, $\overline{X} = 6.01$
- Standard Deviation: $\sigma = 5.48$, $s = 5.47$
- Third Quartile: Q_3(Theory) $= 8$, Q_3(Sample)$= 8$
- 95^{th} Percentile: P_{95}(Theory) $= 18$, P_{95}(Sample)$= 17$

It seems our die is pretty much a fair die.

The most important feature of the **scientific method** is that any scientific theory has to be **falsifyable**, that is it has to be possible to carry out experiments and compare the results of these experiments to predictions made by the theory. If they agree, the theory looks good, if not we need to change the theory or even find a new one. But how do we decide whether or not they "agree"? That is one place where Statistics comes into play.

- Theory: our die is fair
- predictions made using this theory: P(X=1)=0.167, $\mu = 6.0$, ...
- carry out an experiment (6, 1, 7, 4, 1, ...)
- compare predictions with results of the experiment P(X=1)=0.168 (theory), P(X=1)=0.182 (experiment) $\mu = 6$, $\overline{X} = 6.01$
- do they agree or is the theory bad?

Note: most "theories" we look at are not **big** scientific theories but simple things like "Our new drug works better than the currently available one". Well, if this a new drug for cancer maybe it is a pretty big theory afterall!

21 Estimation

21.1 Point Estimation

Here is the type of problem we often see in Statistics:

Example We want to know the percentage π of students at the Colegio who like the food in the Cafeteria. π is the percentage for **all** students, that is for the whole population, so π is a **parameter**.

For each student there are two possibilities: he/she likes the food or he/she does not like the food. So if we randomly select a student he/she is a **Bernoulli trial**. If we randomly select n students we have a sample of n Bernoulli trials.

We know that for a Bernoulli trial with parameter π the **population mean** $\mu = \pi$.

If we did a good job when selecting the students for the sample, then we would hope that the sample mean $\overline{X}$ is close to the population mean μ, that is

$\overline{X} \sim \mu = \pi$ (roughly)

and so we can **estimate** the parameter π with $\overline{X}$

Example Confused? For once I actually on purpose made something simple look complicated. All I just told you is this: if you ask 500 people whether they like Coca-Cola and 300 say yes, your best guess for the proportion of people who like Coca-Cola is 300/500! Here $\overline{X} = 300/500$.

The point is that this simple idea is useful in much more complicated cases as well.

Often in Statistics we use greek letters for parameters and regular letters with a hat or a bar for estimators.

What we have done is the following:

- we decided that the population can be described by a certain distribution, **but without knowing the parameter**
- formulas let us compute the parameter ($\mu = \pi$)
- Statistics lets us find the corresponding statistics ($\overline{X}$)
- combine the two to estimate the parameter ($\pi \sim \overline{X}$)

Each population parameter has a corresponding statistic and vice versa.

Sample mean - population mean
Sample percentage - population percentage
Sample median - population median
Sample standard deviation - population standard deviation
Sample 1^{st} quartile - population 1^{st} quartile
etc ....

Each of these sample numbers is called a **point estimate**.

Example: for data from a geometric distribution we have

$$\mu = \frac{1}{\pi}$$
$$\sigma = \frac{\sqrt{\pi(1-\pi)}}{\pi}$$

```
B <- 10000
p <- 0.5
x <- rgeom(B, p) + 1
round(c(mean(x), 1/p, sd(x), sqrt(1-p)/p), 3)
```

```
## [1] 1.998 2.000 1.414 1.414
```

```
B <- 10000
p <- 0.1
x <- rgeom(B, p) +1
round(c(mean(x), 1/p, sd(x), sqrt(1-p)/p), 3)
```

```
## [1] 10.081 10.000  9.602  9.487
```

```
B <- 10000
p <- 0.8
x <- rgeom(B, p) + 1
round(c(mean(x), 1/p, sd(x), sqrt(1-p)/p), 3)
```

```
## [1] 1.236 1.250 0.545 0.559
```

21.2 Interval Estimation

In real life a point estimate is rarely enough, usually we also need some estimate of the **error** in our estimate.

Example Again we did a survey of the undergraduate students at the Colegio. For that we interviewed 150 randomly selected students, and found a (sample) mean GPA of 2.53. We really want to know the (population) mean GPA of all the undergraduates at the Colegio.

Now the "2.53" is the mean GPA specifically for the sample we collected, if we repeated the whole process and found a different sample, we would also get a different sample mean. Let's say the (population) mean GPA for all the undergraduates is μ_{GPA}. It is is pretty clear that $\mu_{GPA} \neq 2.53$, but hopefully μ_{GPA} is close to 2.53. Only, how close?

One way to answer such questions is to find an **interval estimate** rather than a point estimate. Specifically we will consider a type of interval estimate called a **confidence interval**.

We will learn about confidence intervals using the mean μ as an example. Here the formal definition is

A 100(1-α)% confidence interval for the population mean μ is given by

$$\overline{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

First notice that the interval is given in the form **point estimate** $\pm$ **error**, which is quite often true in Statistics, although not always.

We already know all the ingredients of this formula with the exception of $t_{n-1,\alpha/2}$. You can think of this as a mathematical adjustment factor. At any rate, R will do all of this for us:

Example Say in our survey we found a sample mean GPA of 2.53 with a standard deviation of 0.65. Find a 90% confidence interval for the mean GPA.

This is done with

```
one.sample.t(2.53, shat=0.65, n=150, conf.level=90, ndigit=3)
```

```
## A 90% confidence interval for the population mean is (2.442, 2.618)
```

and so our 90% confidence interval is

$(2.442, 2.618)$

Note the 90% is called the *confidence level*.

What does that mean: a 90% confidence interval for the mean is $(2.442, 2.618)$? The interpretation is this:

suppose that over the next year statisticians (and other people using statistics) all over the world compute 100,000 90% confidence intervals, many for the mean, others maybe for medians or standard deviations or ..., than about 90% or about 90,000 of those intervals will actually contain the parameter that is supposed to be estimated, the other 10,000 or so will not.

Let's do a **simulation** to illustrate all this. We are going to generate data with 50 observations from a normal distribution with mean 100 and standard deviation 25. Then we will find 90% confidence intervals.

Let's just do it slowly at first:

```
x <- rnorm(50, 100, 25) #generate data
xbar <- mean(x)
shat <- sd(x)
one.sample.t(xbar, shat=shat, n=50, conf.level=90)
```

```
## A 90% confidence interval for the population mean is (98, 112)
```

now $98 < 100 < 112$, so the population mean is in the interval, and it is a good one!

Now again, but with the argument *return.result=TRUE*. We will need this in a minute, because we will need to store the result for later:

```
x <- rnorm(50, 100, 25) #generate data
xbar <- mean(x)
shat <- sd(x)
one.sample.t(xbar, shat=shat, n=50,
      conf.level=90, return.result = TRUE)
```

```
##  Low High
## 86.3 97.9
```

So now the interval is $(86.3, 97.9)$, and that is a bad one because 100 is not in it!

Now for the simulation:

```
B <- 10000
good.interval <- 0
for(i in 1:B) {
  x <- rnorm(50, 100, 25) #generate data
```

```
  limits <- one.sample.t(mean(x), shat=sd(x), n=50,
                conf.level=90, return.result = TRUE)
  if(limits[1]<100 & 100<limits[2])
      good.interval <- good.interval+1
}
round(good.interval/B*100, 1) #find percentage of good intervals
```

```
## [1] 90
```

There is already an R routine that does this whole simulation and also let's us change the mean, the standard deviation, the sample size and the confidence level:

```
ci.mean.sim(n=500, mu=75, sigma=30, conf.level = 99)
```

```
## Nominal coverage: 99%
## True coverage: 99%
```

21.2.1 App

```
run.app(confint)
```

this app does exactly the same as above - generate 50 observations from a N(100,25) and find the 90% confidence interval. By clicking on the run button we get a few examples.

Then we can switch to the many examples tab, where we get the interval for 100 examples. For each we see the interval and whether it is good or bad. Finally we see the percentage of good intervals, which should match (somewhat) the chosen confidence level.

But why would we be willing to accept a 10% chance of being "wrong", that is of getting an interval that does not contain the true parameter? Well, we don't have to, after all we chose to compute a 90% confidence interval. Instead we could have found a 99% confidence interval and only leave a 1% chance being "wrong".

```
limits90 <- one.sample.t(2.53, shat=0.65, n=150,
     conf.level=90, ndigit=3, return.result = TRUE)
limits99 <- one.sample.t(2.53, shat=0.65, n=150,
     conf.level=99, ndigit=3, return.result = TRUE)
limits90
```

```
##   Low  High
## 2.442 2.618
```

```
limits99
```

```
##   Low  High
## 2.392 2.668
```

Notice the lengths of these intervals:

```
limits90[2]-limits90[1]
```

```
## [1] 0.176
```

```
limits99[2]-limits99[1]
```

```
## [1] 0.276
```

So the 99% interval is larger!

Here a five of them:

```
## Length of  68 % confidence interval:  0.106
## Length of  90 % confidence interval:  0.176
## Length of  95 % confidence interval:  0.210
## Length of  99 % confidence interval:  0.276
## Length of  99.5 % confidence interval:  0.302
```

So finding confidence intervals involves a trade-off: if we make the probability of being wrong smaller we (almost always) make the interval larger.

The only way to make an interval smaller without changing the confidence level is to get a larger data set!

For more on confidence intervals see section 9.1 and 9.2 of the textbook.

22 Hypothesis Testing

Introduction

Formalism

p-value- Level of Significance

H_0 and H_a

Type I and Type II errors

Type II error β and Power

Importance of Sample Size

Statistical vs. Practical Significance

Warning

22.0.1 Case study: Coin Tossing

During World War II the South African mathematician Jon Kerrich was in a German prisoner of war camp. He tossed a coin 10000 times. He got 5067 heads and 4933 tails.

Question: Was his coin fair?

This type of problem is answered by a **hypothesis test.**

22.1 Introduction

A hypothesis test is a statistical method that answers a yes-no question.

Example Is the average GPA of undergraduates at the Colegio less than 2.8?

Example Is the average income of men in Puerto Rico higher than the average income of women?

Example Was Jon Kerrichs coin fair?

Analogy: Criminal Trial

You can think of a hypothesis test as a trial in a criminal court: is the accused guilty or innocent? Sometimes there is overwhelming proof of guilt - maybe a video showing the crime. Similarly sometimes the data is so obvious no statistics is needed. Usually, though, there is only circumstancial evidence - partial fingerprints, a motive, no alibi, and then the jury has to make a decision.

A hypothesis test is usually phrased in the form of **two statements** rather than a question. These statements are called the **null hypothesis** (H_0) and the **alternative** or research hypothesis (H_1 or H_a)

Example

H_0: The average GPA of undergraduates at the Colegio is 2.8 (or maybe even higher).

H_a: The average GPA of undergraduates at the Colegio is less than 2.8.

Example

H_0: The average income of men and women in Puerto Rico is the same.

H_a: The average income of men in Puerto Rico is higher than the average income of women.

Example H_0: Jon Kerrichs coin was fair

H_a: Jon Kerrichs coin was not fair

Now instead of deciding whether we should answer the question with yes or no we are going to decide which statement we believe is true, but of course this is (almost) the same thing.

Analogy: Criminal Trial

What is the "null hypothesis" in a criminal trial? In the US we start with an "assumption of innocence" (Innocent until proven guilty), so

H_0: accused is innocent
H_a: accused is guilty

Often we will make our decision based on a parameter, and the value of the corresponding statistic. If so we can also express the hypotheses in terms of population parameters. If the hypothesis is written in terms of parameters this (almost always) means that the

null hypothesis has the = sign

Example Is the average GPA of undergraduates at the Colegio less than 2.8? Here we are looking at an "average". In Statistics we have several ways to compute an "average", such as the mean or the median. Which of these is better depends on many considerations. Let's say we use the mean. Now the standard symbol for a population mean is μ, and so we can write the hypotheses as follows:

$$H_0 : \mu = 2.8$$
$$H_a : \mu < 2.8$$

Example Is the average income of men in Puerto Rico higher than the average income of women? Again we are interested in "averages", and let's say here we decide to use the median. The population median is sometimes denoted by λ. But there are two medians: the median income of men and the median income of women. Let's denote them by λ_M and λ_W, respectively. Then the hypotheses are:

$$H_0 : \lambda_M = \lambda_W$$
$$H_a : \lambda_M > \lambda_W$$

Example What does it mean: "a coin is fair"? It means that is has the same chance of coming up "heads" or "tails". That is the probability of heads is 0.5. If we denote by $\pi =$ P("heads) then

$$H_0 : \pi = 0.5$$
$$H_a : \pi \neq 0.5$$

22.2 Hypothesis Testing: Formalism and Notation

A complete hypothesis test has to have all of the following parts:

1. Parameter of interest
2. Method of analysis
3. Assumptions of Method
4. Type I error probability α
5. Null hypothesis H_0
   (in terms of parameter and in plain language)
6. Alternative hypothesis H_a
   (in terms of parameter and in plain language)
7. p value (from R)

8. decision on test

9. Conclusion (in plain language)

Warning

In any homework or exam if any of these parts are missing you will loose points!

22.2.1 p-value

In step 8 we have to make a decision on the test - reject the null hypothesis or not. This is done by comparing the p value from step 8 with the α from step 4:

$p < \alpha \rightarrow$ reject H_0

$p \geq \alpha \rightarrow$ fail to reject H_0

Example Over the last five years the average score in the final exam of a course was 73 points. This semester a class with 27 students used a new textbook, and the mean score in the final was 78.1 points with a standard deviation of 7.1.

Question: did the class using the new text book do (statistically significantly) better?

The R command that calculates the p value is called **one.sample.t**, actually the same command we used before to get a confidence interval. To do a test we need to add the argument muNULL for the null hypothesis and possibly alternative = "greater" (or "less").

1. Parameter: mean μ

2. Method: 1-sample t test

3. Assumptions: data comes from normal distribution, or n large. Checked normal plot

4. $\alpha = 0.05$

5. $H_0 : \mu = 73$ (mean score is still 73)

6. $H_a : \mu > 73$ (mean score is higher than 73)

7. p = 0.000

```
one.sample.t(78.1, shat = 7.1, n = 27, mu.null = 73,
    alternative = "greater")
```

```
## p value of test H0: mu=73 vs. Ha: mu > 73:  0.000
```

8. $p = 0.000 < \alpha = 0.05$, so we reject the null hypothesis

9. The mean score in the final is statistically significantly higher than before.

What is the p value?

Let's assume for a moment that the null hypothesis is true, $\mu = 73$, the mean score is still 73. Then in our experiment we saw something unlikely, the class did much better than they should have.

Now let's say we repeat the same experiment again next year. Chances are the same unusual thing is not going to happen again. How unlikely is it that it is going to happen again? That is the p-value:

$$p = P(\overline{X} > 78.1 \text{ if actually } \mu = 73)$$

One nice feature of the p-value approach is that in addition to the decision on whether or not to reject the null hypothesis it also gives us some idea on how close a decision it was. Here with $p = 0.000$ we would have rejected the null hypothesis even if we had chosen $\alpha = 0.01$, so it was not a close thing at all.

Example: say we flip a coin die 100 times and get 62 heads. Is this an indication that the coin is not fair?

So we want to test

$H_a : \pi = 0.5$ (coin is fair) $H_a : \pi \neq 0.5$ (coin is not fair)

Now the p-value is the following:

What is the probability to flip a **fair** coin and get 62 or more heads? It turns out to be p = 0.012, so we would reject the null hypothesis of a fair coin if we use $\alpha = 0.05$.

Here are a couple of cases:

Number of Heads	p value
50	0.920
51	0.764
52	0.617
53	0.484
54	0.368
55	0.271
56	0.193
57	0.133
58	0.089
59	0.057
60	0.035
61	0.021
62	0.012
63	0.007
64	0.004
65	0.002

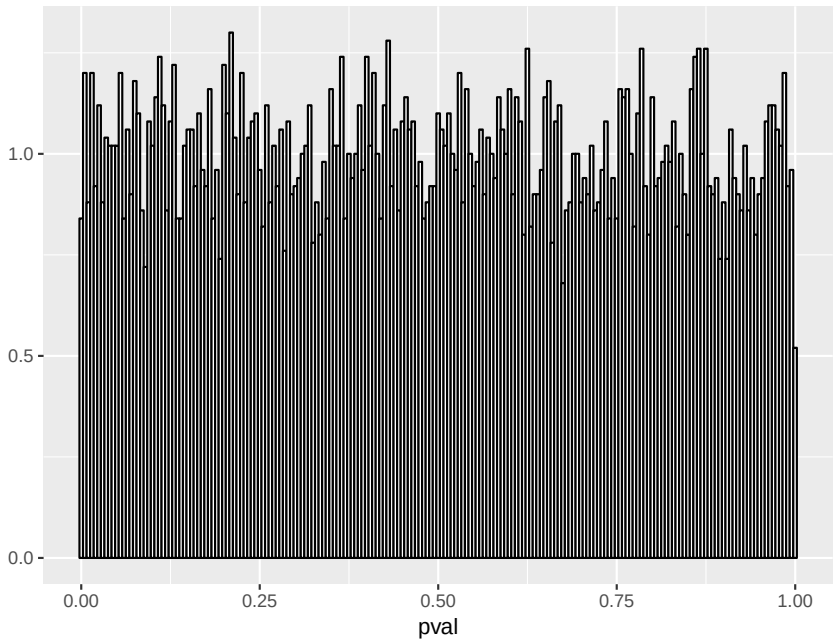
Let's look at another example of coin tossing and testing whether it is a fair coin:

Tosses	Heads	Percentage	p value	Reject H0?
10	6	60%	0.344	No
20	12	60%	0.263	No
30	18	60%	0.200	No
40	24	60%	0.154	No
50	30	60%	0.119	No
60	36	60%	0.092	No
70	42	60%	0.072	No
80	48	60%	0.057	No
90	54	60%	0.045	Yes
100	60	60%	0.035	Yes

So although in all cases we have 60% of the tosses result in Heads, but at a sample size of 80 or less that is not enough to reject the null hypothesis of a fair coin. Whether or not something is statistically significant is always also a question of the sample size. With a small sample size it is difficult to find anything statistically significant!

Let's do a little simulation to see how the p value works. For this we will generate 100 observations from a normal distribution with mean 50 and standard deviation 10. The we do the test $H_0 : \mu = 50$ vs $H_a : \mu \neq 50$, and record the p value.We repeat this 10000 times and look at the histogram of p values:

```
B <- 10000
pval <- rep(0, B)
for(i in 1:B) pval[i ] <- one.sample.t(rnorm(100, 50, 10),
          mu.null = 50, return.result = TRUE)
hplot(pval)
```



Now we reject H_0 if $p < \alpha$. Say we do the test at the 5% level, so $\alpha = 0.05$. How many of the 10000 simulation runs did we (FALSELY!) reject H_0?

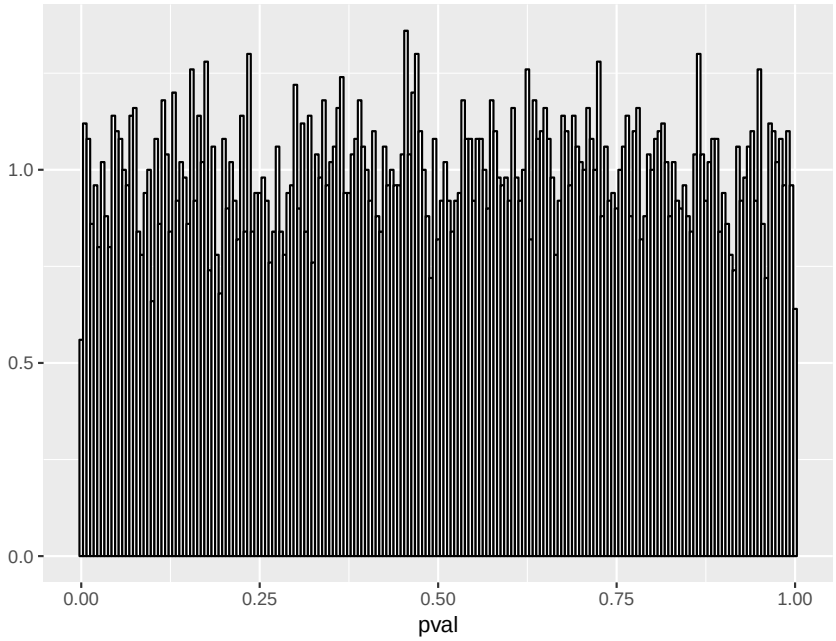
```
sum(pval<0.05)/B
```

```
## [1] 0.0535
```

and so the actual level of significance (0.0535) is just about the nominal one ($\alpha = 0.05$).

There is a routine that does this simulation for us called test.mean.sim. It also let's us choose different values for the mean, standard deviation and sample size:

```
test.mean.sim(n=20, mu=5, sigma=1, alpha=0.1)
```




```
## Nominal alpha:  0.1
## True alpha:  0.0972
```

Caution

Above we stated that we reject H_0 if $p < \alpha$ and fail to reject otherwise. In this class we will adhere to this rule, but in real live things are a little bit more complicated.

p is calculated from the data, so it is itself a **statistic**, so it also has an error. Say we do some experiment, then carry out a hypothesis test and find a p-value of 0.041. If we use $\alpha = 0.05$, then we find $0.041 < 0.05$ and we reject H_0.

But let's say that then we repeat the exact same experiment, and again run the same test on the new data. Just as we would likely not find the same sample mean (say) we would also not find the (exactly) same p-value. But if we find a p-value just a little large, say 0.053, we might then fail to reject the null hypothesis.

In these "borderline" cases it might be better not to either reject or not reject the null hypothesis but to simply

"reserve judgement".

22.3 H_0 and H_a

Example Say a pharmaceutical company has developed a new drug, and they want to show that it is better than the currently available ones.

They carry out a clinical trial with a treatment and a control group. For each patient they record the days until the disease is cured.

Let μ_T be the mean number of days for the treatment group, and μ_C be the mean number of days for the control group. Eventually they will carry out a hypothesis test to see whether the new drug is better. Here they would use the hypotheses

$H_0 : \mu_T = \mu_C$ (the new drug does **not** work better)
$H_a : \mu_T < \mu_C$ (the new drug does work better)

At first it seems a little strange to students that we would choose "new drug is not better than the old one" as H_0, but there are good reasons for this approach as we will see later. In practise it is very easy for us:

H_0 always has the = sign!

There is another reason why the null hypothesis has to be "new drug is not better than the old one", and that has to do with the Philosophie of Science. In general in Science we have that

it is in principle impossible to prove that a scientific theory is correct but it must always be possible to prove that the theory is false (a theory can be falsified)

Example: Newton's theory of gravity has been tested numerous time since it was invented in 1687, in fact everytime someone drops something it is tested again, and (as far as I now) so far the object has always fallen down. Yet strictly speaking the theory of gravity has not been proven to be correct, and it never will be!

(Of course I do think it is a very good theory and I do trust it quite a bit, and so I am careful when I hold something valuable and fragile!)

For our discussion this has the following implication: we have to choose the null hypothesis so it is in principle possible to prove the null hypothesis wrong. So we can't choose

$\mu_T < \mu_C$ (the new drug does work better)

as the null because if this is wrong then the two treatments are exactly the same. But how could we possibly prove that two treatments take exactly the same amount of time to cure a patient? There is not even 10 seconds difference? Impossible to do!

Warning: In the 9 parts of a hypothesis test, the first 6 (at least in theory) should be done **before** looking at the data. The following is not allowed:
say we did a study of students at the Colegio. Among other things we asked them to rate the food at the cafeteria as either "good" or "bad". When looking at the data we find that 54.2% of the students chose "good". Based on this we carry out a hypothesis test with

$H_0 : \pi = 0.5$ vs $H_a : \pi > 0.5$

(maybe we have been hired by the company that runs the cafeteria and they want to argue that "most students find the food good")

The problem here is that this hypothesis test was suggested to us by the data, but hypothesis tests only work as advertised if the hypotheses are formulated without consideration of the data.

Here is a different story: say that 10 years ago a study like this showed that "over half the students say the food is good". We want to to know whether this is still true, and so we do the survey and then the test above. Now, though, we can write down the hypotheses **before** looking at the data, and everything is ok.

Related to this problem is the following issue: as we said we will always use H_0 with $=$ (for example $\mu = 0$). On the other hand there are three commonly used alternative hypotheses:

- $H_a : \mu > 0$
- $H_a : \mu < 0$
- $H_a : \mu \neq 0$

Go back to our example of the new textbook. Here we have the following:

Correct: we pick $H_a : \mu > 73$ because we want to proof that the new textbook works better than the old one.

Wrong: we pick $H_a : \mu > 73$ because the sample mean score was 78.1, so if anything the new scores are higher than the old ones.

Warning

A null hypothesis looks something like this

$H_0 : \mu = 14.5$

not like this:

$H_0 = 14.5$ (What is the parameter?)

or like this:

$\mu = 14.5$ (Is this H_0 or H_a?)

Warning

getting the correct null hypothesis is very important - if you do everything else right but picked the wrong statement as your null hypothesis you will **always** get the wrong answer!

22.3.1 Practice Example

A Biologist reads in a journal article that in the population of a certain animal historically more than 10% of the newborns carry a special gene defect. He takes samples from 250 newborns and tests them. He finds 30 of them have this defect.

Write down the null hypothesis and alternative both in terms of a parameter and in words.

	State of Nature	
Decision we make	H_o is true	H_o is false
Accept H_o	ok	Type II error probability β
Reject H_o	Type I error probability α	ok

Figure 16:

First of remember that the null hypothsis and alternative **can not depend on anything from the sample**. Therefore the information

"He takes samples from 250 newborns and tests them. He finds 30 of them have this defect."

is irrelevant for us.

That leaves only the info

"...historically more than 10% of the newborns ..."

"10%" tells us our parameter is a percentage / proportion, so the symbol we need is π. Always the **null hypothsis has the = sign**, so we find

$H_0 : \pi = 0.1$

he is especially interested in knowing whether the percentage is more than 10% , so the alternative is

$H_a : \pi > 0.1$

Finally we can add the words:

$H_0 : \pi = 0.1$ (Percentage in his population is 10%)

$H_a : \pi > 0.1$ (Percentage in his population is higher than 10%)

22.4 Type I and Type II errors

When we carry out a hypothesis test in the end we always face one of the following situations:

So there are two possible mistakes (type I and type II) and the probabilities for making them (α and β).

These two mistakes, though, are treated completely differently in statistics: when we do a hypothesis test we decide ahead of time what we are willing to accept as a type I error

probability α, and then accept whatever the type II error probability β is. Well, not quite, but wait and see.

Analogy: Criminal Trial

The famous line "innocent until proven guilty" shows that here as well the two possible mistakes are not taken to be equal!

Example: if we flip a coin 100 times and get 60 heads we might conclude that it is not a fair coin. Most of the time this would be the correct conclusion, but on occasion even a fair coin might come up heads 60 time, and then we would commit the type I error.

On the other hand if the coin comes up heads 52 time we would conclude that it is a fair coin, but if the actual probability of heads is 52% we would have committed the type II error.

We have already talked about confidence intervals. At first a confidence interval and a hypothesis test seem to be very different but they are actually closely related.

So finding a 90% confidence interval is related to carrying out a hypothesis test with $\alpha = 0.1$ because $100(1-\alpha)\% = 90\%$ leads to $\alpha = 0.1$.

As we saw before if you find a 95% CI instead of a 90% CI you make the interval wider. Similarly if you make α smaller, thereby reducing the probability of falsely rejecting the null hypothesis you (almost always) make β larger, that is you increase the probability of falsely accepting a wrong null hypothesis. We always have

if $\alpha \downarrow$ then $\beta \uparrow$

The only way to make both α and β smaller is by increasing the sample size n.

How do you choose α? This in practice is a very difficult question. What you need to consider is the

consequences of the type I and the type II errors.

Example In our example above with the new textbook, what does it mean to "commit the type I error"? If we do, what are the consequences? What does it mean to "commit the type II error" and what are its consequences?

Type I error: Reject H_0 although H_0 is true

$H_0 : \mu = 73$ (mean score on final is the same as before, new textbook is **not** better than old one)

This is the truth, but we don't know this, based on our experiment and the hypothesis test we reject H_0, that is now we think the textbook is better

Consequences?

- We will change textbooks for everybody
- New students will not be able to buy used books, previous students will not be able to sell their books
- Professors have to rewrite their material, prontuarios etc.

- Professors will not consider other new textbooks that might really be better
- but scores will **not** go up, all of this is for nothing

Type II error: Fail to reject H_0 although H_0 is false

$H_0 : \mu = 73$ (mean score on final is the same as before, new textbook is **not** better than old one)

This is **false**, but we don't know this, based on our experiment and the hypothesis test we **fail to reject** H_0, that is now we think the textbook is **not** better

Consequences?

- We will **not** change to the new textbook
- scores will **not** go up, but they would have if we had changed
- more students would have passed the course, got an A etc., but now they won't
- Professors will consider other new textbooks, but those might really be worse than the one we just rejected.

Note that in this example we will probably find out that we committed the type I error because we will observe that over the next few years the scores are **not** going up. On the other hand if we comit the type II error we are not likely to ever find out!

Many fields such as psychology, biology etc. have developed standards over the years. The most common one is $\alpha = 0.05$ or 5%. It has mainly historical reasons:

Here is an excerpt from this book:

. . . , therefore, we know the standard deviation of a population, we can calculate the standard deviation of [p. 102] the mean of a random sample of any size, and so test whether or not it differs significantly from any fixed value.

If the difference is many times greater than the standard error, it is certainly significant, and it is a convenient convention to take twice the standard error as the limit of significance ; this is roughly equivalent to the corresponding limit $\alpha = 0.05$ or 1 in 20, . . .

Recent research in psychology has shown that $\alpha = 0.05$ is a fairly good standard, and we will use this if nothing else is said.

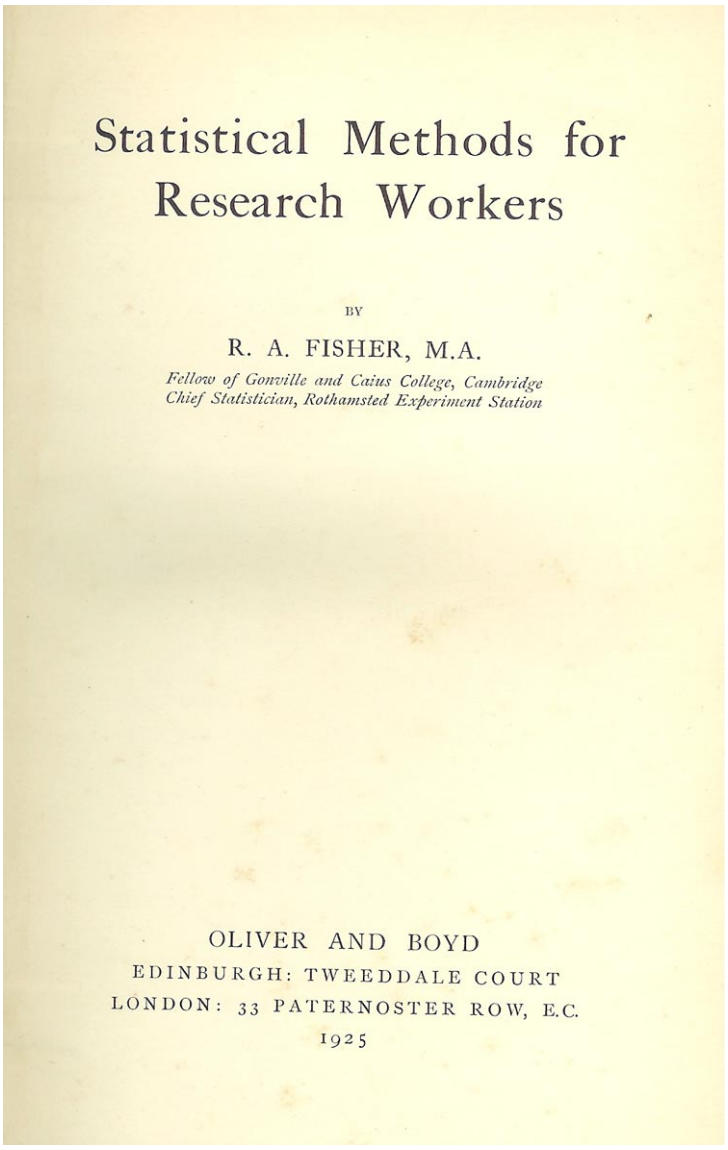
It is not the only one, though. For example for the recent discover of the Higgs boson at the Large Hadron Collider in Geneva, Switzerland we used

$$\alpha = 2.9 \cdot 10^{-7} = 0.000000029$$

22.4.1 Example

A pharmaceutical company has developed a new treatment for terminal cancer. They do a clinical trial and at the end do the following hypothesis test:

H_0: New treatment is the same as old one
H_a: New treatment is better than the old one



Statistical Methods for Research Workers

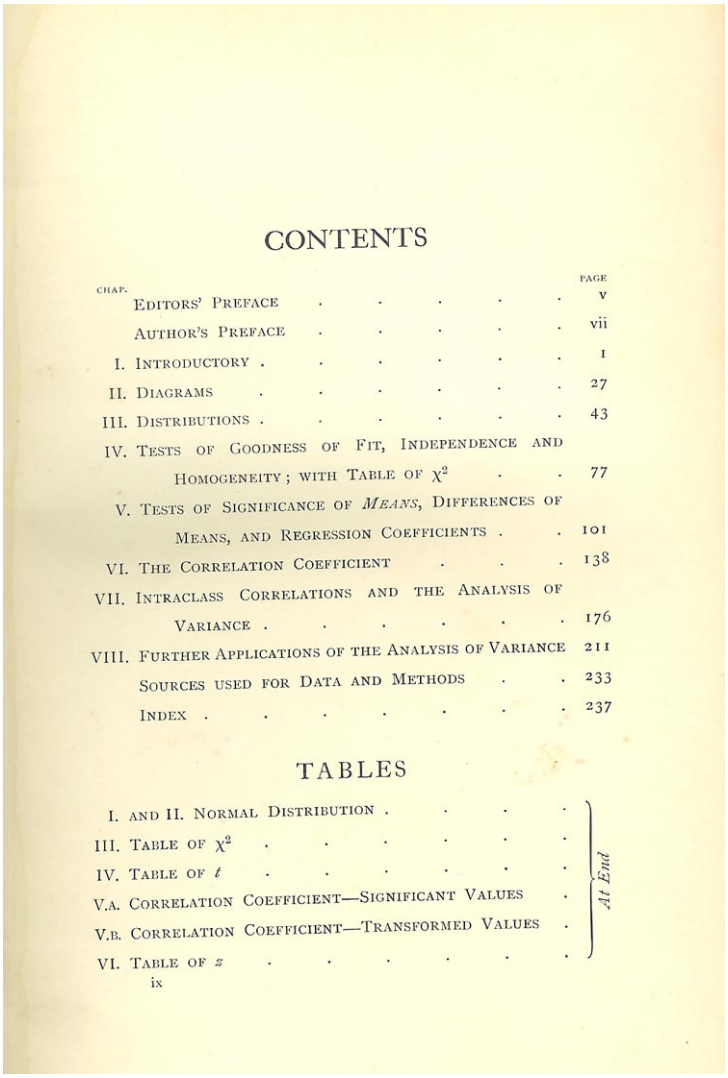
BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

Figure 17:



CONTENTS

TABLES

Figure 18:

What are the type I and type II errors here? Find one consequence each of the type I and type II errors. What should they use as α? You can assume that the new treatment is more expensive than the old one.

Type I error: reject H_0 although it is true

"reject H_0" means because of statistical fluctuation the new treatment did very well in the clinical trial (better than it should have) and therefore we now believe that the treatment is better. "although it is true" means that in reality the new treatment is really the same as the old one.

If they think it is better patients with this type of cancer will start using it

They will expect to live longer but they won't. They (or their insurance) will pay more money for the treatment without any benefit.

Type II error: fail to reject H_0 although it is false

"fail to reject H_0" means because of statistical fluctuation the new treatment did not as good in the clinical trial as it should have and therefore we now believe that the treatment is not better.

"although it is false" means the new treatment is really better than the old one.

If they think it is the same as the old treatment patients with this type of cancer will not be using it, especially if it is more expensive.

They could have lived longer but they won't.

Clearly the worst thing here is for people to die quicker than they might have. This is a consequence of the type II error, so if we want to make that less likely we need β to be smaller, which we can get by allowing α to be larger, say 10% instead of 5%.

22.5 Type II error β and Power

In hypothesis testing we choose α but we don't have any influence on β. One thing we can do is study its behaviour:

Example Let's illustrate the issue with a little simulation. For this we will generate some data and carry out a hypothesis test as follows:

```
x <- rnorm(100, 50, 10)
```

Now we carry out the following hypothesis test:

1. Parameter: mean μ

2. Method: 1-sample t test

3. Assumptions: data comes from normal distribution (true because this is how data is generated)

4. $\alpha = 0.05$

5. $H_0 : \mu = 50.0$

6. $H_a : \mu \neq 50.0$

But we generated the data, so we **know** that $\mu = 50.0$. Therefore we know that the null hypothesis is true, and so if we commit an error it will be the type I error.

But what if we generated the data with

```
x <- rnorm(100, 52.6, 10)
```

Now $\mu = 52.6$ but we test $H_0 : \mu = 50.0$ vs. $H_a : \mu \neq 50.0$, so H_0 is **false**. If we do not reject it we commit the type II error. What is the probability that this happens? Let's do a simulation to find out:

```
set.seed(1111)
B <- 10000
pvals <- rep(0, B)
for(i in 1:B) pvals[i] <- one.sample.t(rnorm(100, 52.6, 10),
          mu.null = 50, return.result = TRUE)
sum(pvals > 0.05)/B
```

```
## [1] 0.2649
```

so we find $\beta = 0.2649$.

So far we talked about the type II error β. In real live one usually calculates the **power** of a test, which is simply

Power = 1 - β =
P(correctly reject H_0 when in fact the H_0 is wrong)

so in the above example for $\mu = 52.6$ we found a power of $100 - 26.49 = 73.51\%$.

We previously used the routine test.mean.sim to study the p value of the test for a mean. We can use the same routine to study the p values in the case when the null hypothesis is false:

```
test.mean.sim(n=100, mu=52.6, mu.null=50,
            sigma=10, alpha=0.05)
```

```
## Power of Test: 73.31%
```

so now we get many more small ($< \alpha$!) p values, which is good because now $H_0 : \mu = 50.0$ is FALSE and should be rejected!

What would have happened if the true mean was 53? Let's see:

```
test.mean.sim(n = 100, mu = 53, mu.null = 50,
              sigma = 10, alpha = 0.05)
```

```
## Power of Test: 84.26%
```

In the case of the one sample t test there are actually exact formulas for the power. We can use the routine **t.ps**:

```
t.ps(n = 100, diff = 52.6-50, sigma = 10)
```

```
## Power of Test = 73%
```

```
t.ps(n = 100, diff = 53-50, sigma = 10)
```

```
## Power of Test = 84.4%
```

Example: for testing $H_0 : \mu = 50.0$ vs. $H_a : \mu \neq 50.0$ with n = 100 and sigma = 10.0, how large would the true mean have to be to have $\alpha = \beta = 0.05$?

$\beta = 0.05$ means power = 1 - β = 1 - 0.05 = 0.95 or 95%.

We can do a little trial and error:

```
t.ps(n = 100, diff = 54-50, sigma = 10)
```

```
## Power of Test = 97.7%
```

```
t.ps(n = 100, diff = 53.5-50, sigma = 10)
```

```
## Power of Test = 93.4%
```

```
t.ps(n = 100, diff = 53.75-50, sigma = 10)
```

```
## Power of Test = 96%
```

```
t.ps(n = 100, diff = 53.65-50, sigma = 10)
```

```
## Power of Test = 95.1%
```

good enough!

Example: for testing $H_0 : \mu = 50.0$ vs. $H_a : \mu \neq 50.0$ with true $\mu = 52$ and sigma = 10.0, how large would the sample size need to be to have $\alpha = \beta = 0.05$?

we could do this again with trial and error, but actually the routine t.ps will calculate whatever argument is missing, so

```
 t.ps(diff = 52-50, sigma = 10, power = 95)
```

```
## Sample size required is  328
```

Here are some interesting cases:

22.5.1 Effect of the true mean

n	True Mean	μ_0	Difference	Power
100	50.0	50	0.0	5.0
100	50.5	50	0.5	7.8
100	51.0	50	1.0	16.5
100	51.5	50	1.5	31.5
100	52.0	50	2.0	50.6
100	52.5	50	2.5	69.6
100	53.0	50	3.0	84.4

so the further the true mean is from the one specified in H_0, the less likely we are to commit the type II error, or

The more wrong the null hypothesis is, the more likely we are to make the right decision

Analogy: Criminal Trial

If there is a very clear evidence that the accused is guilty it should be easy to find him guilty. For example, if we have a video of the person committing the crime it is much easier to find them guilty than if we just have some circumstancial evidence.

22.5.2 Effect of the standard deviation

σ	Power
4	99.8
5	97.7
6	91.0
7	80.8
8	69.6
9	59.4
10	50.6

so the smaller the standard deviation, the less likely we are to commit the type II error, or

the closer together the data is, the easier it is to find a small difference between the true and the hypothesized mean

22.5.3 Effect of α

α	Power
0.001	8.4
0.010	26.6
0.050	50.6
0.100	63.3

so the smaller the α, the smaller the power, or

the harder we make it for the test to reject the null hypothesis, the lower the power.

Analogy: Criminal Trial:

If we change the rules of a trial to make it harder to find an innocent person guilty, we make it easier that a guilty person goes free. For example, if it were decided that the prosecutor is no longer allowed to use fingerprint evidence, some people who were wrongly accused because of their fingerprints would no longer face jail (α goes down) but some criminals will now go free (β goes up, the power of finding a guilty person to be guilty goes down)

22.5.4 Effect of Sample Size n

n	Power
50	27.8
100	50.6
150	68.2
200	80.4
250	88.3
300	93.2
350	96.2

so the large the sample size, the higher the power, or

the more information (data) we have the better a job we can do

Analogy: Criminal Trial:

The more evidence for guilt is presented, the more likely we are to get a guilty verdict (= reject the null hypothesis of innocence)

22.6 Power Curve

Example Let's return to the textbook example. There we we have the test

1. Parameter: mean μ
2. Method: 1-sample t test
3. Assumptions: data comes from normal distribution, or n large. Checked boxplot
4. $\alpha = 0.05$
5. $H_0 : \mu = 73$ (mean score is still 73)
6. $H_a : \mu > 73$ (mean score is higher than 73)

What can be said about the power of the test? Let's assume we have not done the experiment yet, we are going to do it next semester. We already know the class will have 27 students (they just registered) and we also know the standard deviation will be 7.1 (maybe because

this is what it was in the past and we don't expect this to change, at least not much). Now, if we knew that the true population score with the new textbook is going to be 75.5, we could calculate the power:

```
t.ps(n = 27, diff = 75.5-73, sigma = 7.1,
     alternative = "greater")
```

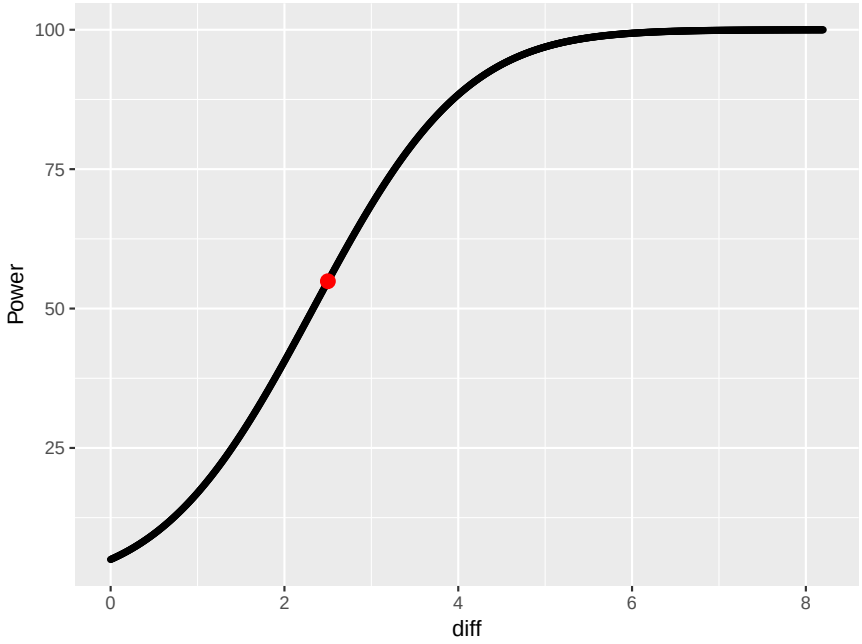
```
## Power of Test = 54.9%
```

So in this case we have a 54.9% chance of correctly rejecting the null hypothesis.

But why 75.5? If we knew that this is the mean score with the new textbook, we would be done, $75.5 > 73$ and H_0 is false!

So instead of calculating the power of a test for just one, or even a few values of the true mean, what we can do is calculate it for all of them, and display the result as a curve:

```
t.ps(n = 27, diff = 75.5-73, sigma = 7.1,
     alternative = "greater")
```




```
## Power of Test = 54.9%
```

With this we can consider different scenarios: if the true mean is 77.0, the difference is 77.0-73=4.0, and the power is about 90%

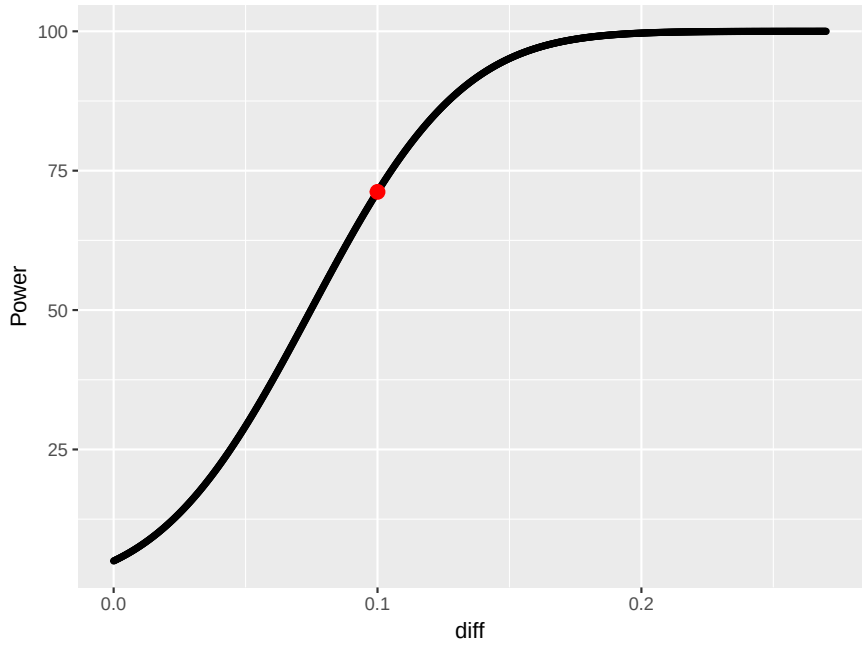
22.6.1 Practice Example

Let's say we are planning a survey of the students at the Colegio. We will interview 100 randomly selected students and ask them what their GPA is. Then we will do the test at the 5% level of

$H_0 : \mu = 2.7$ vs $H_a : \mu > 2.7$

If the true mean GPA of all the students at the Colegio 2.8, what is the power of this test? What is the meaning of this power? Use a standard deviation of 0.45.

```
t.ps(n = 100, diff = 2.8-2.7, sigma = 0.45,
     alternative = "greater")
```




```
## Power of Test = 71.2%
```

with a sample of 100 students we will correctly reject the null hypothesis with a probability of 71.2%

22.7 Importance of Sample Size

Example Using the best currently available treatment the mean survival time of patients with a certain type of terminal cancer is 122 days. A pharmaceutical company has just developed a new drug for these patients which they believe will lead to longer survival times.

To test this they randomly select 13 patients and give them this treatment. The mean survival time of these patients turns out to be 127 days with a standard deviation of 45 days. So they carry out the following hypothsis test.

1. Parameter: mean μ
2. Method: 1-sample t test
3. Assumptions: data come from a normal distribution (Checked boxplot)
4. $\alpha = 0.05$

5. $H_0 : \mu = 122$ (same survival times as with old treatment, new treatment is **not** better)

6. $H_a : \mu > 122$ (longer survival times than with old treatment, new treatment is better)

7. $p = 0.3479$

```
one.sample.t(y = 127, shat = 45, n = 13,
     mu.null = 122, alternative = "greater")
```

```
## p value of test H0: mu=122 vs. Ha: mu > 122:  0.3479
```

8. $p = 0.3479 > \alpha$, so we fail to reject the null hypothesis

9. There is not enough evidence to conclude that this new treatment is better than the old one.

So far, so good. Now let's say that instead of 13 patients the company did the study with 1300 patients. They find:

1. Parameter: mean μ

2. Method: 1-sample t test

3. Assumptions: data come from a normal distribution (Checked boxplot) 4. $\alpha = 0.05$
4. $H_0 : \mu = 122$ (same survival times as with old treatment, new treatment is **not** better)

5. $H_a : \mu > 122$ (longer survival times than with old treatment, new treatment is better)

6. $p = 0.000$

```
one.sample.t(y = 127, shat = 45, n = 1300,
       mu.null = 122, alternative = "greater")
```

```
## p value of test H0: mu=122 vs. Ha: mu > 122:  0.000
```

8. $p = 0.000 < \alpha$, so we reject the null hypothesis

9. The new treatment is statistically significantly better than the old one.

As you see, whether a difference of 5 days is statistically significant depends on the sample size of the study! This is true no matter what the difference is. Let's do this example again, but now say that the mean survival time in the study was 122.12 days, just 2 hours more. Even this difference is statistically significant, although we would need a sample size of about 4million!

Example

Recall the coin tossing example from before. There we considered the p-values if we tossed a coin 100 time, and got 60% Heads. We saw:

Tosses	Heads	Percentage	p value	Reject H0?
10	6	60%	0.344	No
20	12	60%	0.263	No
30	18	60%	0.200	No
40	24	60%	0.154	No
50	30	60%	0.119	No
60	36	60%	0.092	No
70	42	60%	0.072	No
80	48	60%	0.057	No
90	54	60%	0.045	Yes
100	60	60%	0.035	Yes

So whether or not we reject the null hypothesis of a fair coin depends not only on whether the coin is really fair or not, it also depends on the sample size! With less than 90 flips we can not reject the null.

So, after we carried out a hypothesis test, what can we conclude? There are always the following possibilities:

- If we rejected the null hypothesis:

  we reject H_0 because H_0 is false

  we committed the type I error (but we know the probability of doing so - α)

- If we failed to reject the null hypothesis:

  we failed to reject H_0 because H_0 is true

  we committed the type II error

  we failed to reject H_0 because our sample size was to small!

In real live we never know what the correct reason is!

Example So in the case of the company in real live they would not (yet) give up on the new drug, but understanding that n=13 is very small they would repeat the clinical trial (if possible) with a larger sample size.

22.8 "Accept H_0" vs "Fail to reject H_0"

When we do a hypothesis test and find $p > \alpha$, we say we **failed to reject** the null hypothesis. Why is it wrong to say we **accept** the null hypothesis?

Example

let's say we have the following theory: in Puerto Rico nobody is over six feet tall. So we carry out an experiment. We randomly select 10 people and measure their height. None of the 10 is over six feet. Now we carry out a hypothesis test with

H_0: Everybody is six feet tall or less vs H_a: Some people are over six feet tall

given that we have not found anyone over six feet we clearly won't reject the null hypothesis. But should we actually accept it? Of course not.

Even if we had measured 10000 people we still could not be certain that none of the almost 4 million people in PR is over six feet tall.

Actually, even if we had measured every person in Puerto Rico **except one,** we still could not be completely certain that none of the almost 4 million people in PR is over six feet tall.

The way hypothesis testing works we can prove that a null hypothesis is wrong (by rejecting it) but we can never prove that a null hypothesis is right.

Analogy: Criminal Trial

We do actually say: the jury found the accused "not guilty". A jury aquits a person if there is not enough evidence to find them guilty. That is not to say they are innocent, maybe they are - maybe they are not. We just don't have sufficient proof of guilt.

22.9 Statistical vs. Practical Significance

Often you read something like: the new drug was shown to be statistically significantly better than previous drugs. What does that mean? First of all it (usually) means that somebody carried out a hypothesis test and rejected the null hypothesis of no difference between the drugs. But should you care?

Example Say you have to go to a hospital for some checkups. Nothing complicated or dangerous, but you will need to be in the hospital for a few days. You have a choice of hospital A here in Mayaguez, or hospital B in San Juan (assume for a moment you are from Mayaguez). You recently read in the newspaper about a survey done in both hospitals were patients were asked to rate the hospital on things such as: Where the doctors nice? Was the food ok? Did they let you watch TV? In this survey hospital A got a score of 57% and hospital B got 61%. This difference turned out to be statistically significant.

Where will you go?

Example Say you have to go to a hospital for some dangerous surgery. You have a choice of hospital A here in Mayaguez, or hospital C in Miami (again assume you are from Mayaguez). You recently read in the newspaper about a study done in both hospitals on how patients who had that surgery did. In this study hospital A had a survival rate of 57% and hospital C had 61%, but this difference turned out **not** to be statistically significant (**?**).

Where will you go?

Just because something is statistically significant does not automatically mean it is important, and just because something is not statistically significant does not mean you should not care.

22.10 The Silly Hypothesis Test

Consider the following research question: Is the median income of men and women in Puerto Rico the same? Say to answer this question we do a survey of 1000 randomly selected men and women and find out their income. Then we do the hypothesis test with

H_0: Median Incomes are the same vs. H_0: Median Incomes are not the same

But why do this survey and test at all? After all, we already know the answer: there is absolutly no chance at all that the true population median incomes of men and women are **exactly** the same! In many fields generally it is known apriori that the null hypothesis is wrong, so why do a test?

There are two answers to this:

a. The real question should be: can we reject the null hypothesis **at this sample size**?

b. Maybe we really should not do a test, but instead find a confidence interval (here for the difference in median incomes) There are of course null hypotheses that could really be true:

H_0: nothing can move faster than the speed of light.

22.11 Warning

There is one common misuse of hypothesis testing you should be aware of. It concerns searching for something, anything significant:

Example: There is a famous (infamous?) case of three psychiatrists who studied a sample of schizophrenic persons and a sample of nonschizophrenic persons. They measured 77 variables for each subject - religion, family background, childhood experiences etc. Their goal was to discover what distinguishes persons who later become schizophrenic. Using their data they ran 77 hypothesis tests of the significance of the differences between the two groups of subjects, and found 2 significant at the 2% level.They immediately published their findings.

What's wrong here? Remember, if you run a hypothesis test at the 2% level you expect to reject the null hypothesis of no relationship 2% of the time, but 2% of 77 is about 1 or 2, so just by random fluctuations they could (should?) have rejected that many null hypotheses! This is not to say that the variables they found to be different between the two groups were not really different, only that their method did not proof that.

In its general form this is known as the problem of **simultaneous inference** and is one of the most difficult issues in Statistics today.

23 The Lady tasting tea

In 1935 Sir R.A. Fisher wrote a book with the title *The Design of Experiments*. This book and several others that he wrote were so important that today Fisher is often called the father of Statistics.

In the book he tells the following story: one day one of his collegues at the Rothemstead Experimental Station, Muriel Bristol (Ph.D), claimed she could tell whether in a cup of tea the tea had been poured into the cup before the milk, or vice versa.

Fisher devised an experiment to test that claim as follows: He filled eight identical cups with milk and tea, four with the milk first and four with the tea first. Then he randomly put them

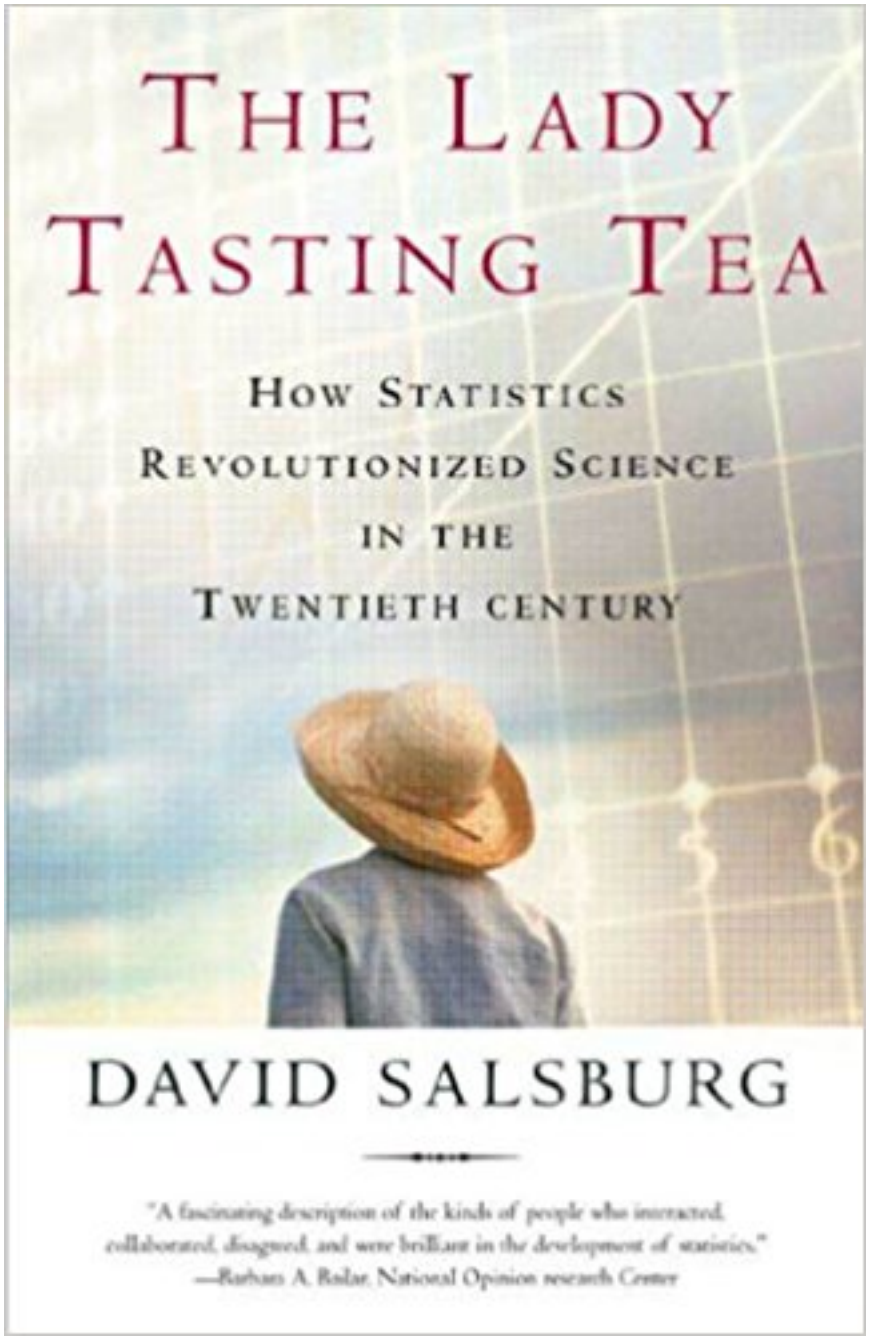

THE LADY TASTING TEA
HOW STATISTICS REVOLUTIONIZED SCIENCE IN THE TWENTIETH CENTURY
DAVID SALSBURG
"A fascinating description of the kinds of people who interacted, collaborated, disagreed, and were brilliant in the development of statistics."
—Barbara A. Bailar, National Opinion research Center

Figure 19:

on a table and asked Muriel to pick the four with the tea poured first. Muriel was told the experimental setup, so she knew there were four cups of each kind.

What can we say about this experiment? Let's write down one possible arrangement. Here T is a cup where the tea has been poured first (of course without Muriel knowing this!), whereas M is one with the milk first:

T M M T T T M M

Let's also say that the first four cups are those the Lady has identified as the one with the milk poured first. So in this case she got two correct and two wrong. Not very good!

Of course this is what we would expect to see if indeed the Lady knows what she is doing:

M M M M T T T T

Now Fisher decided to only accept Muriels claim if indeed she could identify all four cups with milk poured first correctly. If she was just guessing, how likely was it she would get that lucky? Well, how many possible arrangements of the cups she picks are there? Here they are:

M M M M	M M &
M M M T	M M &
M M M T	M M &
M M M T	M M &
M M M T	M M &
M M M T	M M &
M M M T	M M &
M M M T	M T &
M M M T	M T &
M M T T	M T &
M M T T	M T &
M M T T	M M &
M M T T	M M &
M M T T	M M &
M M T T	M M &
M M M T	M M &
M M M T	M M &
M M M T	M M &
M M M T	M M &
M M T T	M M &
M M T T	M M &
M M T T	M M &
M M T T	M M &
M M T T	M M &

Want to count them? Sometimes it helps to know some math: if there are $2n$ cups there are $\binom{2n}{n}$ such arrangements, or here:

$$\binom{8}{4} = \frac{8!}{4!4!} = \frac{40320}{24 \times 24} = 70$$

only the first one (M M M M) is correct, so her chances of getting it right if she were to just guess randomly were $\frac{1}{70} = 0.0143$.

Let's view this experiment in light of our previous discussion on hypothesis testing:

23.0.1 1) Parameter of Interest

One can view this as an experiment with two variables (actually tea first/actually milk first and identified as tea first/identified as milk first) and whether the two are independent or not. In this sense the parameter is a correlation, or as we say for categorical data, an association.

23.0.2 2) Method of Analysis

this idea of counting the total number of possible answers is now know as Fisher's Exact test.

23.0.3 3) Assumptions of the Method

the experiment has to be set up as described above. For example, it is very important that Muriel knew that exactly four cups had the milk poured first.

23.0.4 4) Type I error probability α

$\alpha = 0.05$

23.0.5 5) Null hypothesis H_0

H_0: Muriel is just guessing.

Notice that common feature of many hypothesis tests, namely to pick the "negative option" as the null.

23.0.6 6) Alternative hypothesis H_a

This is where it gets interesting, because there isn't one!

The idea of an alternative hypothesis was invented a bit later by Jerzy Neyman and Egon Pearson (son of Karl Pearson of correlation fame). Fisher never liked it. They had some very good fights over this!

One consequence of not having an alternative hypothesis is that one can not find the power of the test.

23.0.7 7) p-value

So, how did Muriel do? In fact she was perfect, she got all eight cups correct! Therefore we have $p = 1/70 = 0.0143$.

23.0.8 8) Decision of the test

$p = 0.0143 < 0.05 = \alpha$, and so we reject the null hypothesis.

23.0.9 9) Conclusion

Muriel certainly proved her claim.

23.0.10 Type I error:

Type I error =

reject the null hypothesis although it is true =

conclude that Muriel knows what she is doing although she was just guessing.

23.0.11 Type II error:

Type II error =

fail to reject the null hypothesis although it is false =

conclude that Muriel was lying although she actually knows her tea (but unfortunately made a mistake).

23.1 Historical Importance

Using this simple experiment, Fisher established most of the fundamental principles for hypothesis testing, which contributed to major advances across biological and physical sciences. A careful read of the original text shows a precise use of terms, in a concise and unambiguous presentation, in contrast with many textbooks written later that were more confusing than helpful.

24 Inference for the Mean

Assumptions
Confidence Interval
Hypothesis Test
Power

Sample Size

After all the theory, here are some examples. Actually, we have discussed almost everything here already.

24.1 Method

1-sample t

24.2 Assumptions

The methods discussed here work if:

- the data comes from a simple random sample
- the data comes from a normal distribution or the sample size is large enough

The last assumption is a bit vague, just how large is "large enough"? The basic principle here is that we need a balance:

- If the distribution of the data is almost normal, a sample size as small as 10 is ok.
- If the distribution of the data is very non-normal (large outliers etc..), a sample size as large as 100 might be needed.

24.3 R Routines

one.sample.t - test and confidence interval

t.ps - power and sample size

24.4 Confidence Interval

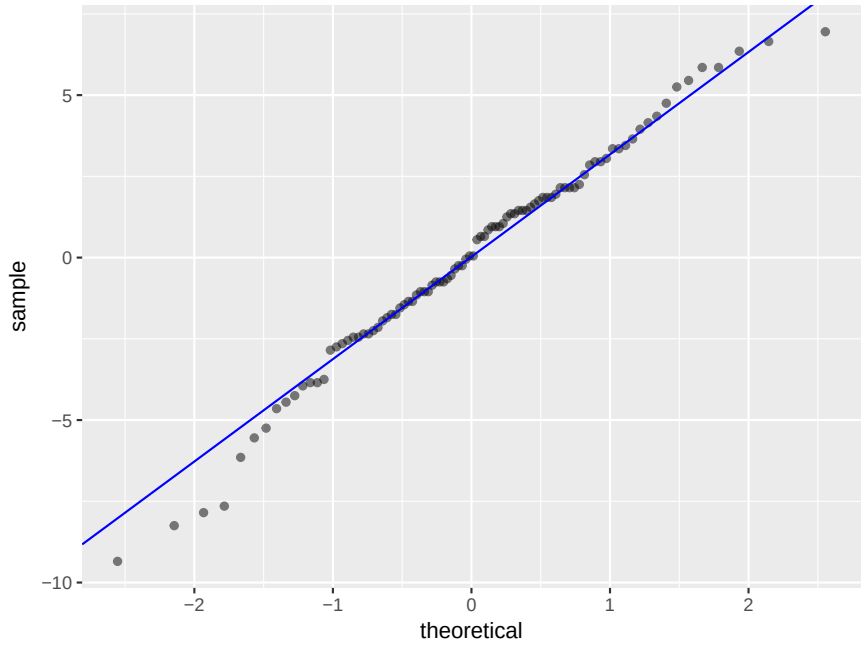
24.4.1 Case Study: Drug Use of Mothers and the Health of the Newborn

Consider again the data set for newborn babies and the drug status of their mothers. Find a 90% confidence interval for the length of babies:

```
attach(mothers)
sort(Length)
```

```
##  [1] 40.2 41.3 41.7 41.9 43.4 44.0 44.3 44.9 45.1 45.3 45.6 45.7 45.7 45.8
## [15] 46.7 46.8 46.9 47.0 47.1 47.1 47.2 47.2 47.3 47.4 47.6 47.7 47.8 47.8
## [29] 48.0 48.1 48.2 48.2 48.4 48.5 48.5 48.5 48.7 48.8 48.8 48.8 48.9 49.0
## [43] 49.2 49.3 49.3 49.5 49.6 49.6 50.1 50.2 50.2 50.4 50.5 50.5 50.5 50.6
## [57] 50.8 50.9 50.9 51.0 51.0 51.0 51.1 51.2 51.3 51.4 51.4 51.4 51.5 51.7
## [71] 51.7 51.7 51.7 51.8 52.1 52.4 52.5 52.5 52.6 52.9 52.9 53.0 53.2 53.5
## [85] 53.7 53.9 54.3 54.8 55.0 55.4 55.4 55.9 56.2 56.5
```

```
one.sample.t(Length, conf.level = 90)
```




```
## A 90% confidence interval for the population mean is (49, 50.1)
```

Assumptions: normal plot ok

Example In a survey 150 people leaving a mall were asked how much money they spent. The mean was \$45.60 with a standard deviation of \$12.70. Find a 95% confidence interval for the true mean.

```
one.sample.t(y = 45.60, shat = 12.70, n = 150)
```

```
## A 95% confidence interval for the population mean is (43.6, 47.6)
```

24.5 Hypothesis Test

The details of the hypothesis test for a population mean are as follows:

Null Hypothesis: $H_0 : \mu = \mu_0$

Note: μ_0 is not "μ_0" but a specific number which you need to get from the problem.

Alternative Hypothesis: Choose one of the following, depending on the problem:

a. $H_a : \mu < \mu_0$

b. $H_a : \mu > \mu_0$

c. $H_a : \mu \neq \mu_0$

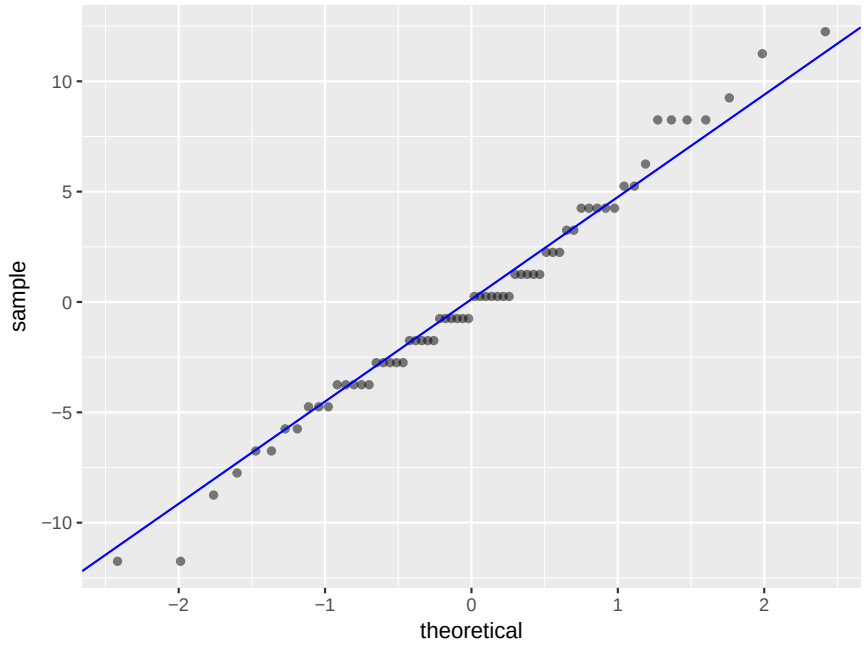
24.5.1 Case Study: Simon Newcomb's Measurements of the Speed of Light

We have previously seen that (after eliminating the outliers -44 and -2) the mean of Newcombs measurements of the speed of light is 27.75, whereas using modern instruments the equivalent measurement is 33.02. Does this say his measuring method was bad? The question is whether this sample mean is statistically significantly dfferent from the population mean.

Let's answer this question now:

1. Parameter: mean μ
2. Method: 1-sample t
3. Assumptions: normal data or large sample, normalplot is ok
4. $\alpha = 0.05$
5. $H_0 : \mu = 33.02$ (Newcomb's experiment measured correct value)
6. $H_a : \mu \neq 33.02$ (Newcomb's experiment did not measure correct value)
7. p = 0.000

```
attach(newcomb)
one.sample.t(Deviation[Deviation > 0], mu.null = 33.02)
```




```
## p value of test H0: mu=33.02 vs. Ha: mu <> 33.02:  0.000
```

8. $p < \alpha$, so we reject the null hypothesis
9. Newcomb's experiment did not measure correct value
   Assumptions: normal plot ok

Figure 20:

24.5.2 Case Study: Resting Period of Monarch Butterflies

Some Monarch butterflies fly early in the day, others somewhat later. After the flight they have to rest for a short period. It has been theorized that the resting period (RIP) of butterflies flying early in the morning is shorter because this is a thermoregulatory mechanism, and it is cooler in the mornings. The mean RIP of all Monarch butterflies is 133 sec. Test the theory at the 10% level.

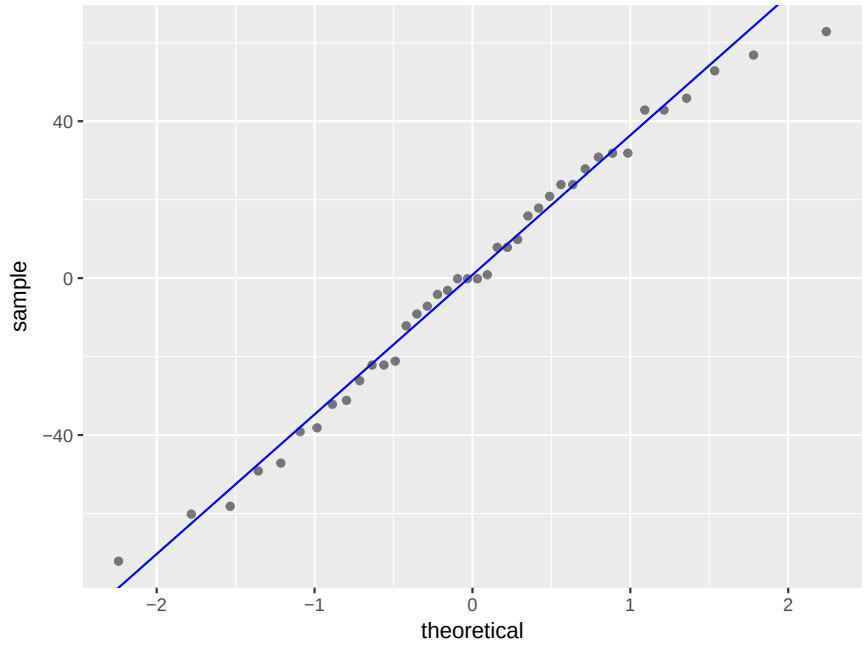
Research by Anson Lui, Resting period of early and late flying Monarch butterflies Danaeus plexippus, 1997

1. Parameter: mean μ
2. Method: 1-sample t
3. Assumptions: normal data or large sample
4. $\alpha = 0.1$
5. $H_0 : \mu = 133$ (RIP is the same for early morning flying butterflies as all others)
6. $H_0 : \mu < 133$ (RIP is the shorter for early morning flying butterflies)
7. $p = 0.056$

```
attach(butterflies)
sort(RIP.sec.)
```

```
##  [1]  52  64  66  75  77  85  86  92  93  98 102 102 103 112 115 117 120
## [18] 121 124 124 124 125 132 132 134 140 142 145 148 148 152 155 156 156
## [35] 167 167 170 177 181 187
```

```
one.sample.t(RIP.sec., mu.null=133, alternative = "less")
```




```
## p value of test H0: mu=133 vs. Ha: mu < 133:  0.0558
```

8. $p = 0.0558 < \alpha = 0.1$, so we reject the null hypothesis

9. It appears the resting time is somewhat shorter, but the conclusion is not a strong one.

Assumtions: normal plot ok

Example In the past the average purchase of a customer in a certain store was \$55 . The store just ran an ad in the newspaper and wants to know whether it increased sales. In week following the ad 43 customers spent an average of \$63 with a standard devation of \$18. Test at the 10% level whether the promotion was a success.

1. Parameter: mean μ

2. Method: 1-sample t

3. Assumptions: assumed to be ok

4. $\alpha = 0.1$

5. $H_0 : \mu = 55$ (same mean sales as before, ad did not work)

6. $H_a : \mu > 55$ (higher mean sales than before, ad did work)

7. $p = 0.0028$

```
one.sample.t(y = 63, shat = 18, n = 43,
             mu.null = 55, alternative = "greater")
```

```
## p value of test H0: mu=55 vs. Ha: mu > 55:  0.0028
```

8. $p < \alpha$, so we reject the null hypothesis
9. higher mean sales than before, ad did work.

24.6 Power

Recall that the power of a test is the probability to reject the null hypothesis when the null hypothesis is indeed wrong.

Calculating the power of a test usually means making a guess what the true value of the parameter might be.

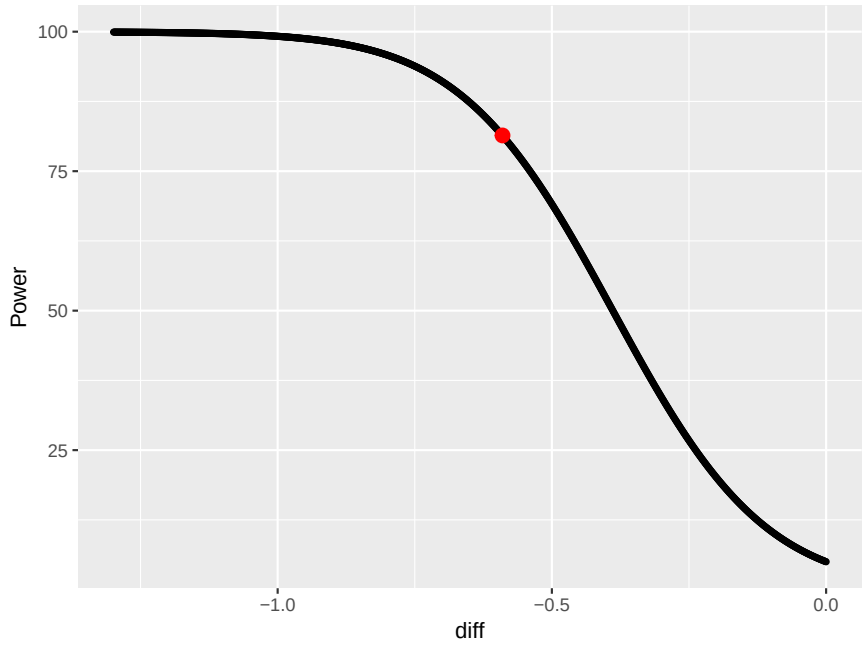
Example Over many years the mean number of accidents per month on a street was 2.15 with a standard deviation of 0.75. The city council is considering to install traffic lights at a number of intersections. After that they will monitor the number of accidents for one year. If it turned out that the lights lower the number of accidents to 1.56 per month, what is the probability that they would detect this drop? Use $\alpha = 0.05$.

The test they will eventually do will have the following:

4. $\alpha = 0.05$
5. $H_0 : \mu = 2.15$ (Same number of accidents with the traffic lights)
6. $H_a : \mu < 2.15$ (Lower number of accidents with the traffic lights)

Now to calculate the power:

```
t.ps(n=12, diff = 1.56-2.15, sigma = 0.75,
     alternative="less")
```



```
## Power of Test = 81.4%
```

so there is an 81.4% chance to correctly conclude that the traffic lights lowered the number of accidents.

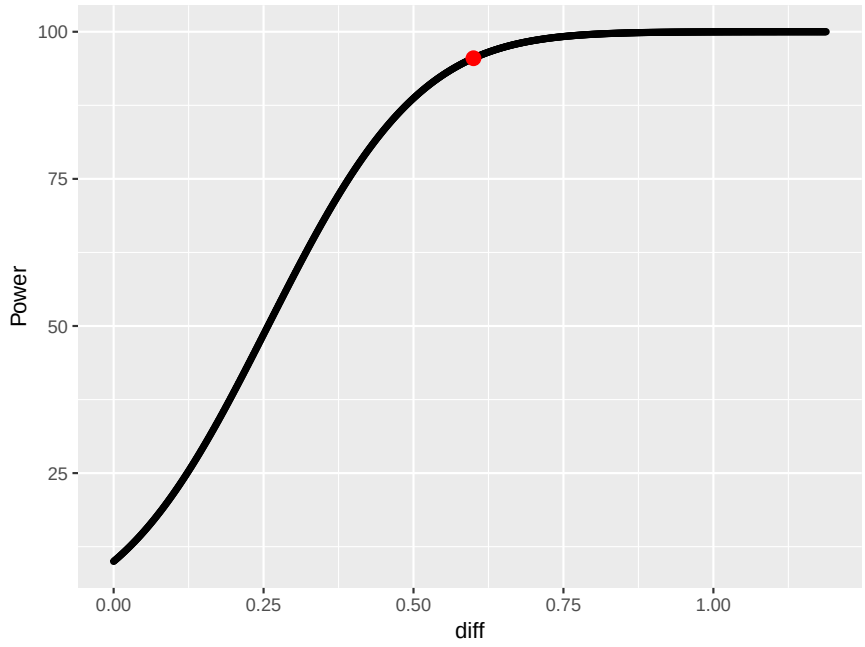
But why 1.56? After all, we have not even installed the traffic lights, so we can't know what will happen once we do. So we really should look at the whole Power Curve.

Example We are planning a survey of the employees of a large company. In the survey we will ask them how happy they are to work there, on a scale of 1 to 10. Eventually we will test at the 10% level whether

$H_0 : \mu = 5.0$ vs $H_0 : \mu > 5.0$

If we randomly select 250 employees and if the true mean happiness is 5.6, what is the power of this test? Assume $\sigma = 1.4$.

```
t.ps(n=50, diff=5.6-5.0, sigma=1.4,
     alpha=0.1, alternative = "greater")
```




```
## Power of Test = 95.5%
```

24.7 Sample Size Calculations

One of the most important questions facing a researcher is how large a sample he needs to be able to draw valid conclusions. If the goal is to do a hypothesis test the *t.ps* command is again the way to go:

Example A company has been making "widgets" which have a mean life time of 127 days with a standard deviation of 45.5 days. They have recently redesigned the production process, and believe that now the lifetime is 145 days. They want to test that hypothesis. How many

widgets do they need to test to have a 95% chance of detecting this difference? They will carry out the test at the 10% level.

```
t.ps(diff = 145-127, sigma = 45.5, power = 95,
     alpha = 0.1, alternative = "greater")
```

```
## Sample size required is  57
```

Let's say that instead of a hypothesis test we want to find a confidence interval. We have seen that one effect of the sample size is to make the confidence interval shorter:

```
one.sample.t(10, shat=1, n=20, ndigit=3)
```

```
## A 95% confidence interval for the population mean is (9.532, 10.468)
```

and so the length of the interval is

$$10.486 - 9.535 = 0.951$$

but if the sample size is 40 we have

```
one.sample.t(10, shat=1, n=40, ndigit=3)
```

```
## A 95% confidence interval for the population mean is (9.68, 10.32)
```

$$10.32 - 9.68 = 0.64$$

and so this interval is shorter.

A sample size calculation starts with a decision on how large an interval we are willing to accept. Let's call this length L. Usually one specifies the error E, which is

$$E = L/2$$

The error E is equivalent to what power we want in the hypothesis testing case above.

Notice that the caclulation of the interval also involves *shat*. This of course is the sample standard deviation, an estimate of the population standard deviation σ. Here are several possible ideas:

- Is there already an estimate of σ we can use, maybe from a previous or from a similar study?
- If not maybe we can do a pilot study (something that is very often a good idea anyway)

Example: We found that a 90% confidence interval for the mean length of babies to be $(49.0, 50.1)$, or 49.55 ± 0.55, so the error on this estimate is 0.55. What sample size would be needed to find a 90% confidence interval with an error of 0.25?

We can use the sample standard deviation as a guess for the population standard deviation.

```
t.ps(sigma= sd(Length), E = 0.25, conf.level = 90)
```

```
## [1] "Sample size required is  497"
```

Example We want to do a survey of the students of the Colegio. One question will be their GPA, and we want to find a 99% confidence interval with a length of 0.25. A pilot study of 25 students had a sample standard deviation of 0.45. How many students will we need in our survey?

length of interval = L = 0.25, so E = L/2 = 0.25/2 = 0.125

```
t.ps(sigma= 0.45, E = 0.125, conf.level = 99)
```

```
## [1] "Sample size required is  86"
```

But what if we did not do a pilot study and therefore do not know the standard deviation? Sometimes we can make an educated guess.

Remember our old rule of thumb:

$$\text{Range}/4 = s$$

For GPA a likely range is 2-4, so

Range/4 = (4-2)/4 = 0.5 = s = σ, so

```
t.ps(sigma= 0.5, E = 0.125, conf.level = 99)
```

```
## [1] "Sample size required is  107"
```

Example We want to do a study of the age at which students graduate from the Colegio. We will find a 90% confidence interval with an error of 1 month. A pilot study showed that the standard deviation of the ages is 0.8 years. What sample size is needed?

```
t.ps(sigma= 0.8, E = 1/12, conf.level = 90)
```

```
## [1] "Sample size required is  250"
```

25 Inference for a Proportion (Percentage) π

Assumptions
Confidence Interval
Hypothesis Test
Power

Sample Size

In this section we will discuss inference for proportions (or percentages) such as the percentage of people who prefer Coke over Pepsi, who will vote PNP in the next election, who earn more than $50,000 per year etc.

Say we do a survey of n people and ask them "Do you prefer Coke over Pepsi?" Then if we only allow "Yes"and "No" answers we have the Bernoulli trial with success probability π. The object of interest here is π, the proportion of people in the whole population who prefer Coke over Pepsi. Obviously the proportion in the sample who prefer Coke over Pepsi will be our point estimate of π.

Notation: often in this context we use $\widehat{p}$

Example Say in a survey of 500 people 312 say they prefer Coke over Pepsi. Then a point estimate for the proportion of people who prefer Coke over Pepsi is

$$\widehat{p} = \frac{312}{500} = 0.624$$

Note Most often problems are stated in terms of **percentages** instead of **proportions** but all the methods use proportions. Simply multiply by 100% at the end.

Example A point estimate for the percentage of people who prefer Coke over Pepsi is 62.4%

Note Sometimes problems are for **probabilites**, that is the same as proportion in this context.

Example The probability of a six on a fair die is 16%

25.1 Method

Exact Binomial

R commands:

- *one.sample.prop* confidence intervals and hypothesis testing
- *prop.ps* power and sample size

25.2 Assumptions

None!

25.3 Confidence Interval

A $100(1-\alpha)\%$ confidence interval for the population proportion π is found with the **one.sample.prop** command.

25.3.1 Case Study: Binge Drinking in College

Alcohol on college campuses is a very serious problem. But how common is it? A survey of 17,096 students in US four-year colleges collected information on drinking behavior and alcohol-related problems. (Henry Wechsler et al., "Health and Behavioral Consequences of Binge

Drinking in College", *Journal of the American Medical Association*, 272 (1994).

The researchers defined "frequent binge drinking" as having five or more drinks in a row three or more times in the past two weeks. According to this definition 3,314 students were classified as frequent binge drinkers.

Problem: Find a point estimate for the percentage of frequent binge drinkers.

Solution: A point estimate for the proportion of frequent binge drinkers is

$$\widehat{p} = \frac{3314}{17096} = 0.194$$

therefore a point estimate for the percentage is 19.4%

Problem: Find a 99% confidence interval for the percentage of frequent binge drinkers.

```
one.sample.prop(x = 3314, n = 17096, conf.level = 99)
```

```
## A 99% confidence interval for the population proportion is (0.186, 0.202)
```

Note unlike the *one.sample.t* command the *one.sample.prop* command has no argument *shat*.

25.3.2 Case Study: Vacations of Puerto Ricans

The website of the Puerto Rico Tourism Company had the results of a survey of Puerto Ricans and their vacation travel. The study measures short trip leisure travel habits of average Puerto Rican families and allows for the monitoring of consumer preferences on a continuous basis. According to the report for July - September 2009 10% of the respondents had made a trip to Cabo Rojo, the highest number of any place in PR.

Find a 95% CI for the true percentage of PR travelers who visit Cabo Rojo. The survey was based on 400 interviews.

```
one.sample.prop(x = 40, n = 400, conf.level = 95)
```

```
## A 95% confidence interval for the population proportion is (0.073, 0.135)
```

Example In a sample of 200 people entering a store, 61 actually bought something. Find a 90% confidence interval for the percentage of "buyers".

```
one.sample.prop(x = 61, n = 200, conf.level = 90)
```

```
## A 90% confidence interval for the population proportion is (0.252, 0.363)
```

25.4 Hypothesis Test

Null Hypothesis: $H_0 : \pi = \pi_0$

Alternative Hypothesis: Choose one of the following:

a. $H_a : \pi < \pi_0$

b. $H_a : \pi > \pi_0$

c. $H_a : \pi \neq \pi_0$

Again we can use the one.sample.prop command. To get the p value of a test we need to use the argument pi.null

25.4.1 Case Study: Jon Kerrichs Coin

Test at the 5% level of significance whether 5067 heads in 10000 flips are compatible with a fair coin.

1. Parameter: proportion π
2. Method: exact binomial
3. Assumptions: None
4. $\alpha = 0.05$
5. $H_0 : \pi = 0.5$ (50% of flips result in "Heads", coin is fair)
6. $H_a : \pi \neq 0.5$ (coin is not fair)
7. $p = 0.1835$

```
one.sample.prop(x = 5067, n = 10000, pi.null = 0.5)
```

```
## p value of test H0: pi=0.5 vs. Ha: pi <> 0.5:  0.1835
```

8. $p = 0.1835 > \alpha = 0.05$, so we fail to reject the null hypothesis.
9. it appears Jon Kerrich's coin was indead fair.

Example Let's assume for a moment that Jon Kerrichs coin was actually **not** a fair coin but one with $\pi = 0.505$. How often would he have had to flip his coin to reject the null hypothesis?

Of course now we don't have any data, so we have to guess what $\hat{\pi}$ might have been. For example if he had flipped this coin 10000 times we would expect him to get about $10000 \times 0.505 = 5050$ heads. Running the test with these numbers we find:

```
n <- 10000
one.sample.prop(x = 0.505*n, n = n, pi.null = 0.5)
```

```
## p value of test H0: pi=0.5 vs. Ha: pi <> 0.5:  0.3222
```

```
n <- 20000
one.sample.prop(x = 0.505*n, n = n, pi.null = 0.5)
```

```
## p value of test H0: pi=0.5 vs. Ha: pi <> 0.5:  0.1594
```

```
n <- 30000
one.sample.prop(x = 0.505*n, n = n, pi.null = 0.5)
```

```
## p value of test H0: pi=0.5 vs. Ha: pi <> 0.5:  0.0843
```

```
n <- 40000
one.sample.prop(x = 0.505*n, n = n, pi.null = 0.5)
```

```
## p value of test H0: pi=0.5 vs. Ha: pi <> 0.5:  0.046
```

so if he had flipped his coin about 40000 times he would have rejected the null hypothesis of a fair coin at the 5% level.

Remember: even small differences (0.5 vs 0.505) will be rejected if the sample size is large enough

25.4.2 Practice Example

Say we roll a die 500 times and got 100 "sixes". Is this compatible with a fair die? Test at the 5% level.

1. Parameter: proportion π
2. Method: exact binomial
3. Assumptions: None
4. $\alpha = 0.05$
5. $H_0 : \pi = 1/6$ (die is fair)
6. $H_a : \pi \neq 1/6$ (die is not fair)
7. $p = 0.0524$

```
1/6
```

```
## [1] 0.1666667
```

```
one.sample.prop(x = 100, n = 500, pi.null = 0.16667)
```

```
## p value of test H0: pi=0.16667 vs. Ha: pi <> 0.16667:  0.0524
```

8. $p = 0.0524 > \alpha = 0.05$, so we fail to reject the null hypothesis.
9. it appears this is a fair die

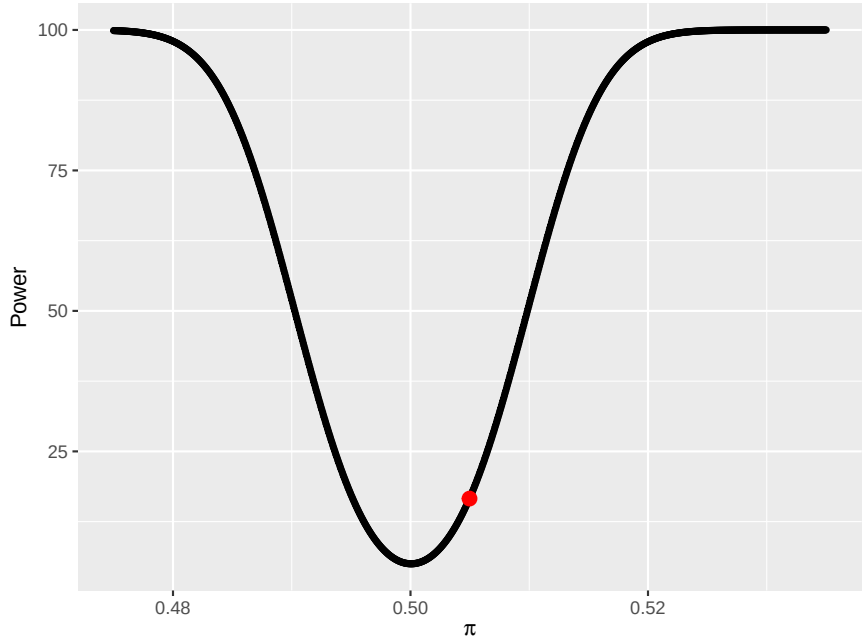
25.5 Power of the Test

Again we need to worry about the power of our test.

25.5.1 Case Study: Jon Kerrichs Coin

let's assume his coin had a probability of 0.505 to come up heads. What was the power of the test we did above? To find out we can use the *prop.ps* command:

```
prop.ps(n = 10000, phat = 0.505, pi.null = 0.5)
```




```
## [1] "Power of Test = 16.6%"
```

so with "just" 10000 flips there was only a very small chance of detecting that his coin was a little unfair.

Again one would probably do this for many values of π and draw a graph.

Note unlike in the *t.ps* command in the *prop.ps* command we need need both phat and pi.null, not just the difference:

```
prop.ps(n = 10000, phat = 0.5, pi.null = 0.505)
```

```
## [1] "Power of Test = 17.4%"
```

25.6 Sample Size Calculation

As with the mean the sample size calculation is different depending on whether we want to do a hypothesis test or find a confidence interval

25.6.1 Case Study: Jon Kerrichs Coin

If indeed his coin had a probability of Heads of 0.505, how often would he have to flip his coin to have a power of 90%?

```
prop.ps(power = 90, phat = 0.505, pi.null = 0.5)
```

```
## [1] "Sample size required is  105281"
```

How about if we want to find a confidence interval? As with the mean we have to decide on the error E (twice the length of the interval) we want. Then we can use the prop.ps command.

We have the same problem as with the mean here, the sample size depends on the true π, but we are trying to estimate π!

The same ideas as with the mean such as doing a pilot study work here as well. In addition we have something else we can do here. It turns out that using phat $= 1/2$ will lead to a sample size that is always sufficient. prop.ps does this unless another phat is given.

Example You want to do a survey of likely voters for the next election. You want to find a 95% confidence interval for the percentage of voters for the PNP, with an arror of E = 0.03. What sample size is required?

```
prop.ps(E = 0.03)
```

```
## [1] "Sample size required is  1068"
```

Example same as above, but for the PIP. Here we already know that π is around 5%, so

```
prop.ps(phat = 0.05, E = 0.03)
```

```
## [1] "Sample size required is  203"
```

Look again at the prop.ps command for the sample size. There is something truely amazing about what is **not** part of command!

Example

We want to do a study on the percentage of students in some class that are female. We want to find a 95% confidence interval with an error of 2%. What sample size will we need?

```
prop.ps(E = 0.02)
```

```
## [1] "Sample size required is  2401"
```

Example A company regularly receives a shipment of electronic parts. Their contract with the supplier says that the shipment can contain up to 5% faulty parts. They suspect that the current shipment has 10% faulty parts. If they plan on randomly selecting parts, testing them and then do a hypothesis test at the 10% level, how many parts do they need to select so that the hypothesis test has a power of 90%?

```
prop.ps(power = 90, phat = 0.1, pi.null = 0.05,
    alpha = 0.1, alternative = "greater")
```

```
## [1] "Sample size required is  187"
```

26 Bayesian Statistics

Say you pick a coin from your pocket. It's just any coin, nothing special. You flip it 10 times and get 3 heads . What can we conclude about this coin?

Now each flip is a Bernoulli trial with success parameter π. We have previously seen that the standard estimator for π is the ratio of successes to trials, so we find $\widehat{\pi} = x/n = 3/10 = 0.3$.

But wait just a minute! This is a regular coin, we all know that coins are (almost) fair, so we know that really $\pi = 0.5$! 3 head in 10 flips of a fair coin is a perfectly fine outcome, in fact the probability of 3 or less heads in 10 flips of a fair coin is 0.172, so this will happen easily.

What's going on? The problem is that the formula $\widehat{\pi} = x/n$ is completely general, it is the same whether we flip a coin (head vs tails), survey people (male vs female), check students in a class (pass vs fail) or do anything else that is a Bernoulli trial. It does not take into account that we know a lot about this experiment "flip a coin" **a priori**, that is before we ever do it, namely that (almost always) $\pi = 0.5$.

Of course there is also the issue that 10 flips is very few, just 300 heads in 1000 flips would be a very different thing. But situations with little data are quite common, and it would still be nice to have a more sensible answer than 0.3.

In fact, it is possible to include such a priori information in a statistical analysis, applying what is called **Bayesian Statistics**. The principle idea is this:

- "encode" your knowledge of the experiment before it is done in what is called a **prior distribution**.
- do the experiment and collect the data
- combine the data and the prior to get to the **posterior distribution**, which now encodes our updated knowledge of what we know after having done the experiment.

Note: the prior and the posterior are regular probability distributions like those we discussed before.

The science of Statistics comes in two flavours: Bayesian and Frequentist. There are a number of fundamental differences between them. One we have already seen: a Bayesian analysis not only can but has to begin by specifying a prior distribution . This can be a strength (as in the coin flip example above) or a weakness, mostly in situations were we really don't have much prior knowledge. A Frequentist analysis on the other hand doesn't need a prior, but also can't use one if there is one!

There are deeper differences as well, for example the very definition of what a probability is. Those issues are quite fundamental to doing Statistics but unfortunately much beyond what we can discuss in an introductory class!

So, how do we do this "combine data and prior" step? It uses something called **Bayes formula** (which is where the name comes from) and a lot of heavy math, calculus and more. This is one reason why Bayesian statistics is not yet as widely used as most Statisticians think it should be! But more and more computer programs can take care of the calculations for us.

I have written an "Interactive Bayesian Calculator for Percentages", which we can use for our problem. Run it with

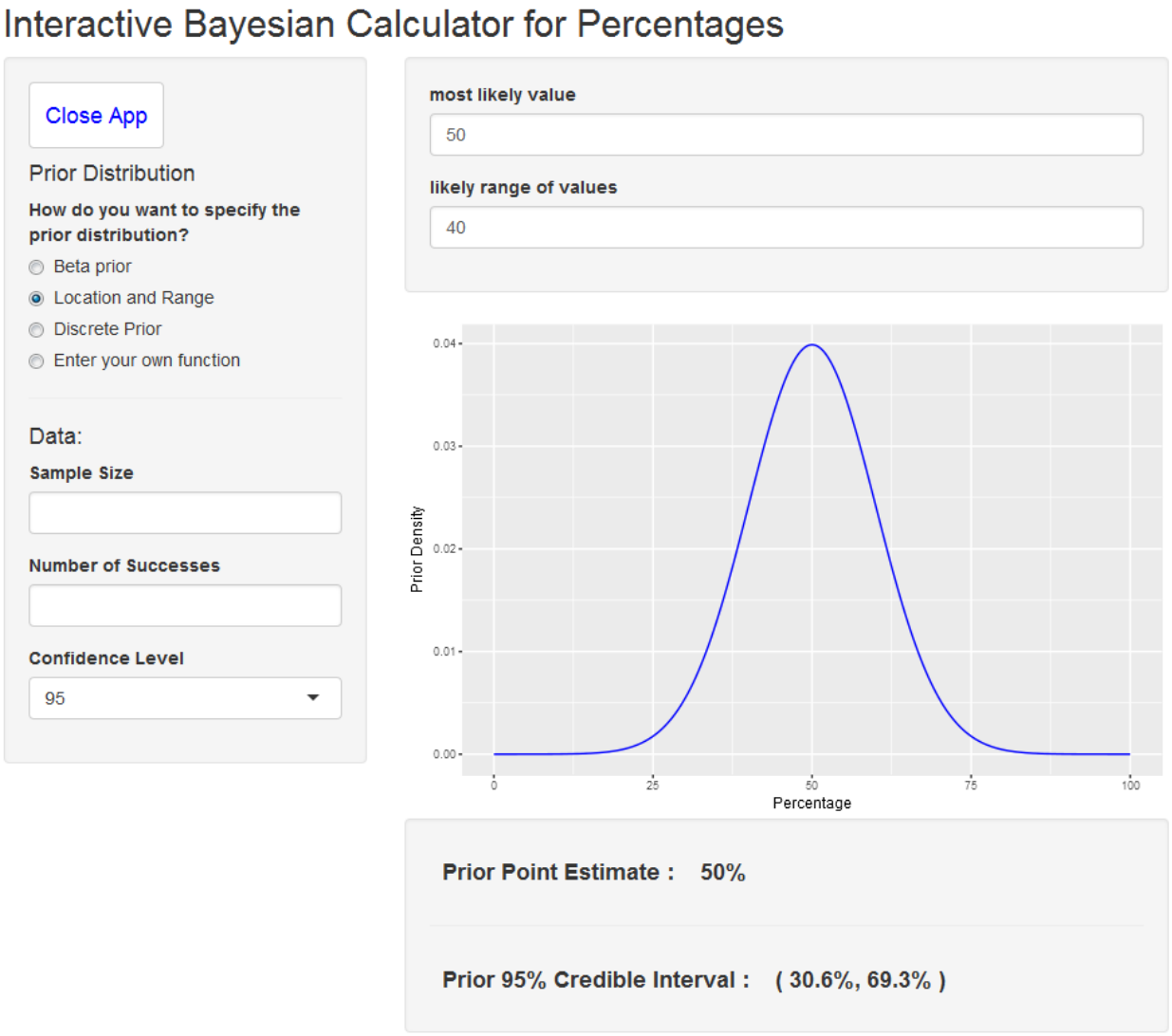


Figure 21:

```
ibayesprop()
```

when it opens it looks like this:

The first thing we need to do is specify the prior distribution.

There are several ways to do this, listed on the left side. The default option is to specify what we think the most likely value is and what the range might be. We do think this is a fair coin, so 50% is ok. The graph shows that any value between about 25% and 75% is ok. You can use the box above the graph to change that if you want.

Below the graph we see the interval (30.6%, 69.3%). If our prior distribution is reasonable for our problem than the true percentage should be inside.

Now let's enter our data, Sample Size = 10 and Number of Successes = 3:

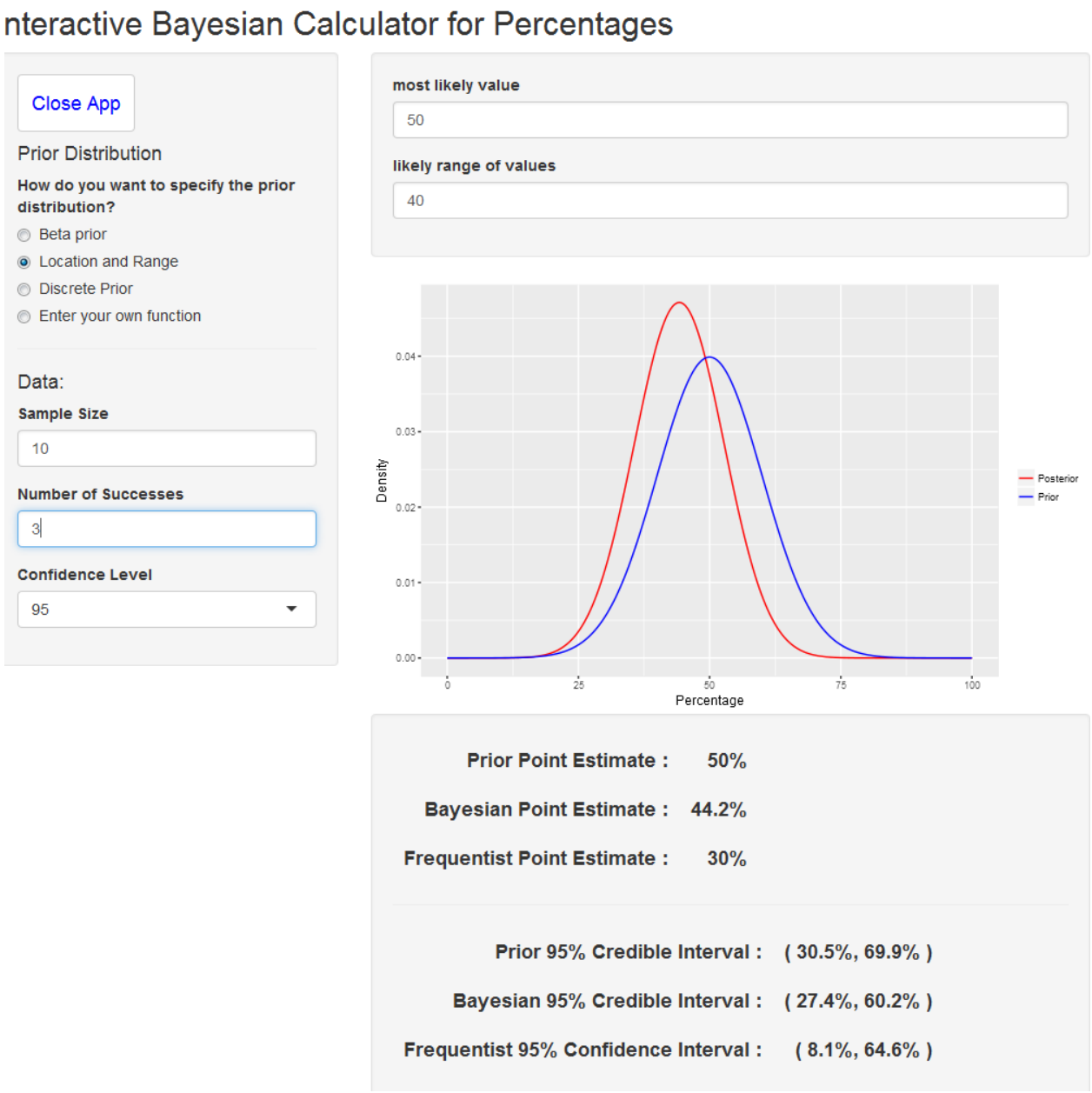

nteractive Bayesian Calculator for Percentages
Close App
Prior Distribution
How do you want to specify the prior distribution?
Beta prior
Location and Range
Discrete Prior
Enter your own function
Data:
Sample Size
10
Number of Successes
3
Confidence Level
95
most likely value
50
likely range of values
40
Density
Percentage
Posterior
Prior
Prior Point Estimate : 50%
Bayesian Point Estimate : 44.2%
Frequentist Point Estimate : 30%
Prior 95% Credible Interval : (30.5%, 69.9%)
Bayesian 95% Credible Interval : (27.4%, 60.2%)
Frequentist 95% Confidence Interval : (8.1%, 64.6%)

Figure 22:

The blue curve is the same prior distribution as before, the red is the posterior distribution, that is our best guess after having seen the result of the experiment. Notice that because there were fewer heads than we expected from a fair coin it has shifted a bit to the left.

On the bottom we see the 95% Bayesian credible interval (27.4%, 60.2%), which is our best guess after having done the experiment.

For comparison we also have the 95% confidence interval (8.1%, 64.6%).

Let's see what would happen if we actually had 300 heads in 1000 flips:

Again the red posterior curve has shifted to the left, by now it is far away from the blue prior one. The Bayesian interval is (27.2%, 33.2%).

Notice that it is quite similar to the Frequentist confidence interval (27.2%, 33.0%). This is something we see a lot: in cases were there is a lot of data (100 flips) the answers from a Bayesian and a Frequentist analysis tend to be very similar. This of course makes good sense: whatever our expectation was before the experiment (as encoded in the prior distribution), we will certainly change that expectation in the face of a lot of evidence (aka data).

26.0.0.1 Specifying a Prior Distribution

There is a vast literature on how to go about encoding our prior knowledge. In the app I have included four ways to do so:

- Location and Range: just as it says, decide what the most likely value is and in what range the answer should be.

Here are four examples:

1) we think the true percentage is around 50%, but it could be as low as 25% and as high as 75%
2) we think the true percentage is around 50%, but it could be as low as 0% and as high as 100%
3) we think the true percentage is small, maybe even 0, and no larger than 20%
4) we are quite certain that the true percentage is around 20%.

- Beta prior: this is a class of distributions which has a number of advantages. It has two parameters α and β, and you can use sliders to get a shape that works for your experiment.

1) we really have no idea where π might be.

This one is the default for the Beta. It looks a little funny but has some good theoretical features (for the specialists: it is the Jeffrey's prior for the binomial)

2) we really have no idea where π might be.

Another favorite, what is called a flat prior.

3) we think the true percentage is around 50% but we are not to sure of that.

Note that here we have $\alpha = \beta$, which will always put the peak of the curve at 50%

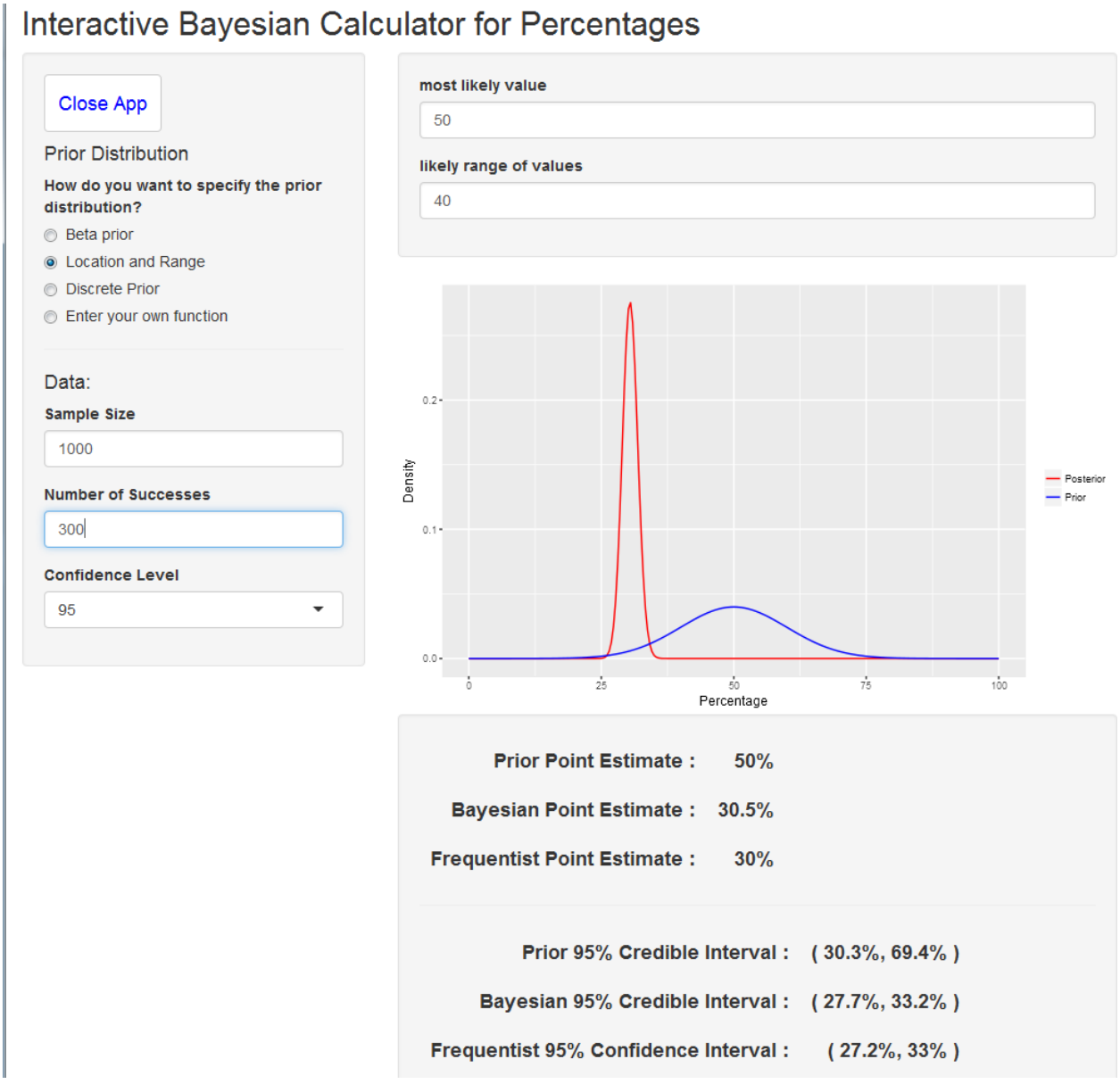

Interactive Bayesian Calculator for Percentages
Close App
Prior Distribution
How do you want to specify the prior distribution?
Beta prior
Location and Range
Discrete Prior
Enter your own function
Data:
Sample Size
1000
Number of Successes
300
Confidence Level
95
most likely value
50
likely range of values
40
Density
0.2
0.1
0.0
0
25
50
75
100
Percentage
Posterior
Prior
Prior Point Estimate : 50%
Bayesian Point Estimate : 30.5%
Frequentist Point Estimate : 30%
Prior 95% Credible Interval : (30.3%, 69.4%)
Bayesian 95% Credible Interval : (27.7%, 33.2%)
Frequentist 95% Confidence Interval : (27.2%, 33%)

Figure 23:

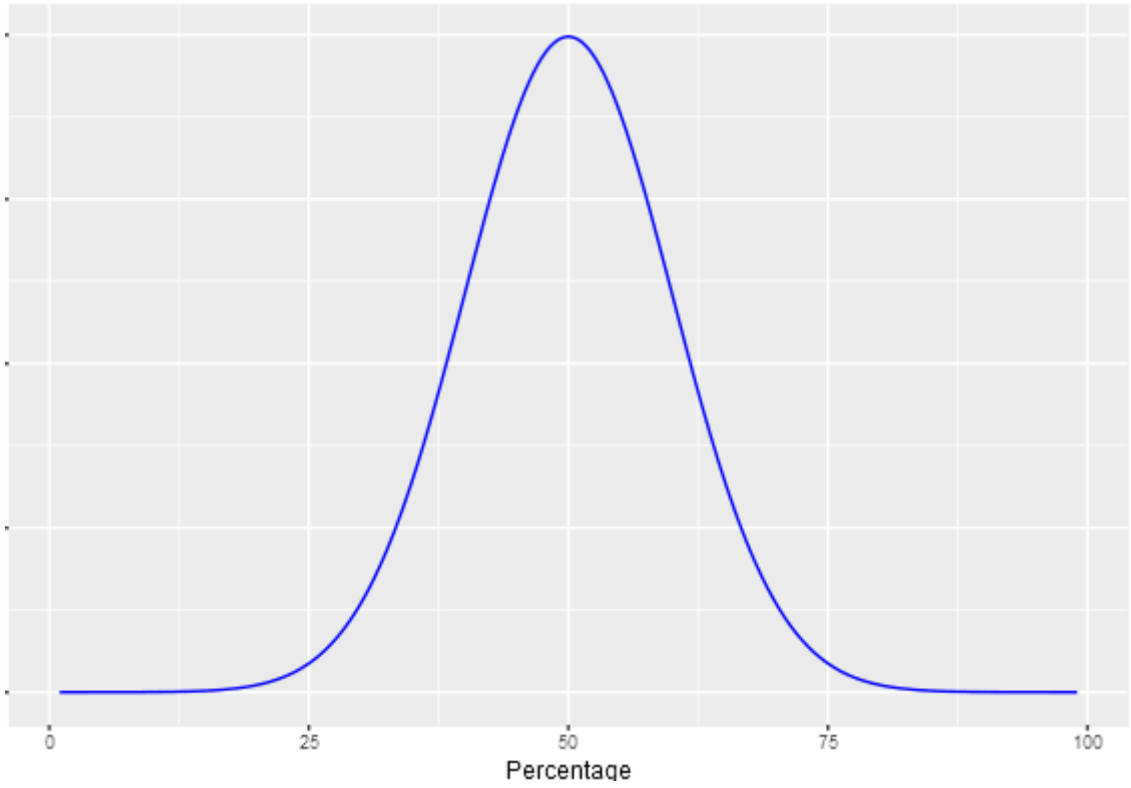


Figure 24:

4) we are quite certain that the true percentage is greater than 50%.

- Discrete prior: here we can specify the (relative) probabilities for 10 points in some interval.

1) we really have no idea where π might be.

2) we really have no idea where π might be, but is not likely that it is either very close to 0 or very close to 100%

3) we think the true percentage is around 40%. Moreover, we are vey sure it is not less than 30% and not higher than 50%.

- Enter your own function: here you can enter any R expression for any function you like! (and that could e a prior, of course). we think the true percentage is is either around 25%

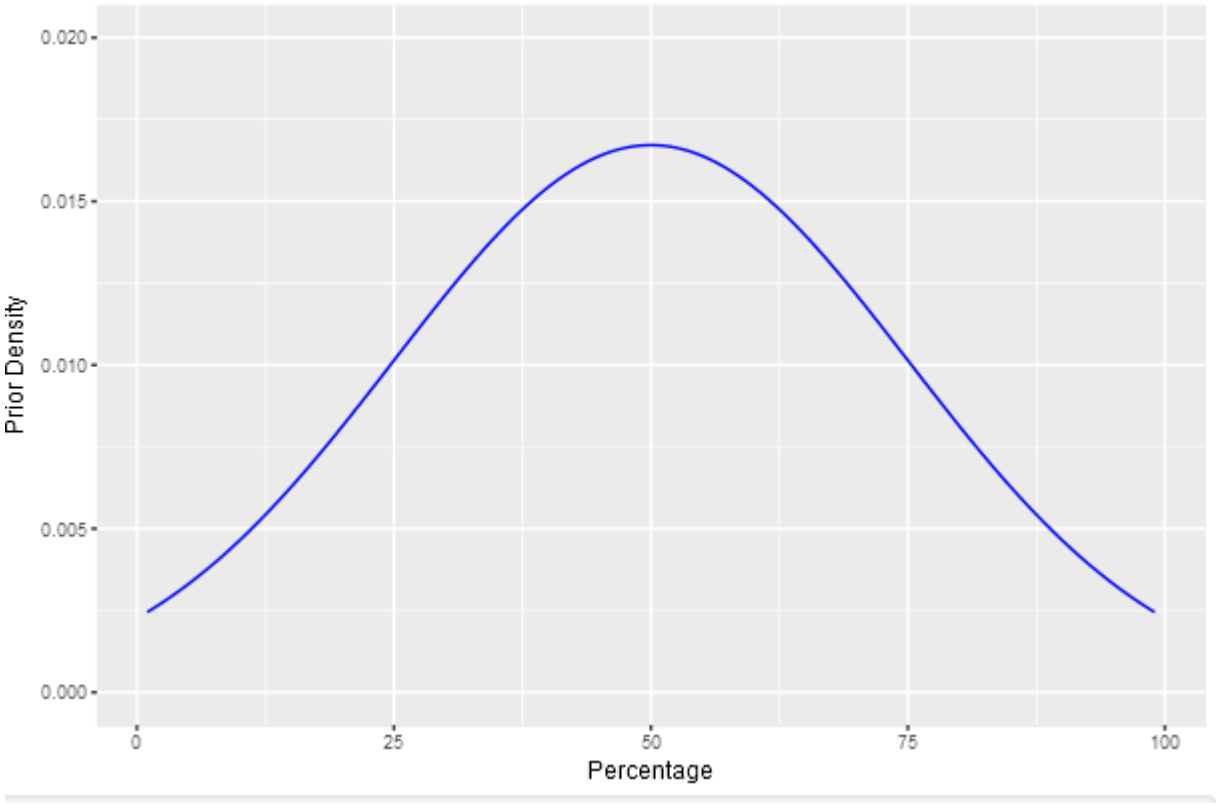

0.020
0.015
0.010
0.005
0.000
Prior Density
0
25
50
75
100
Percentage

Figure 25:

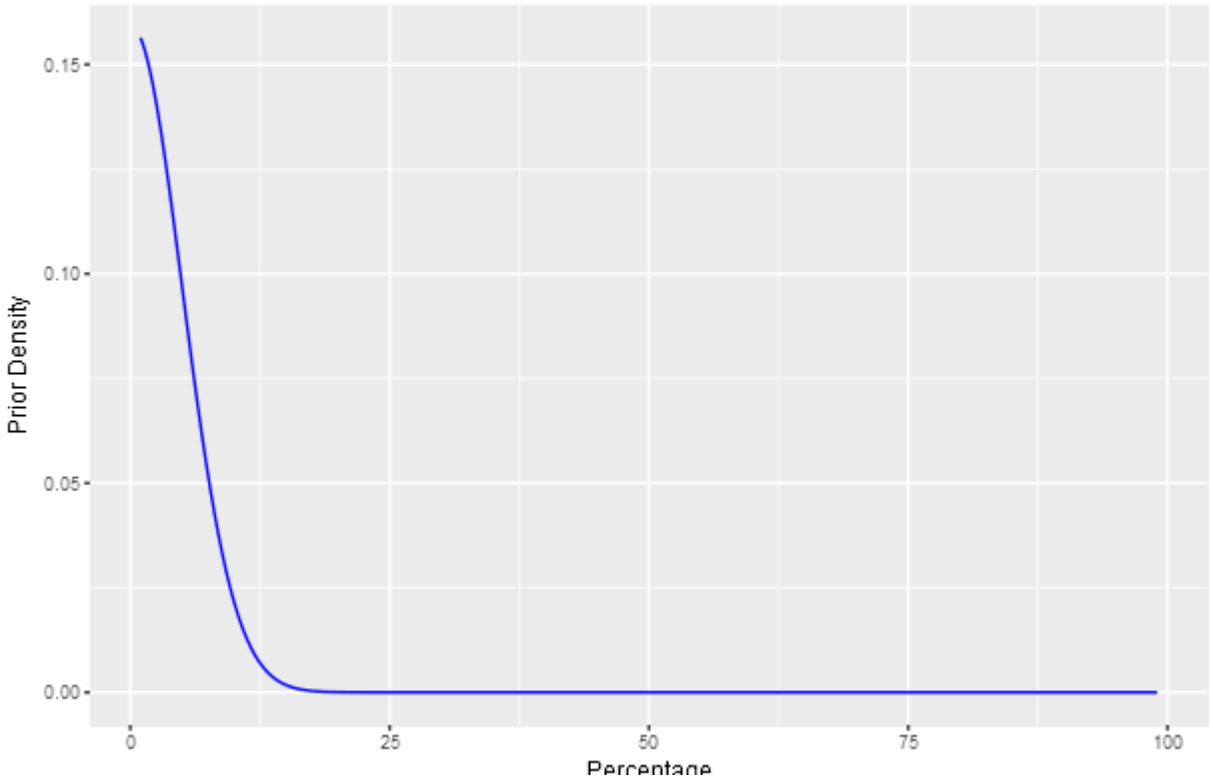
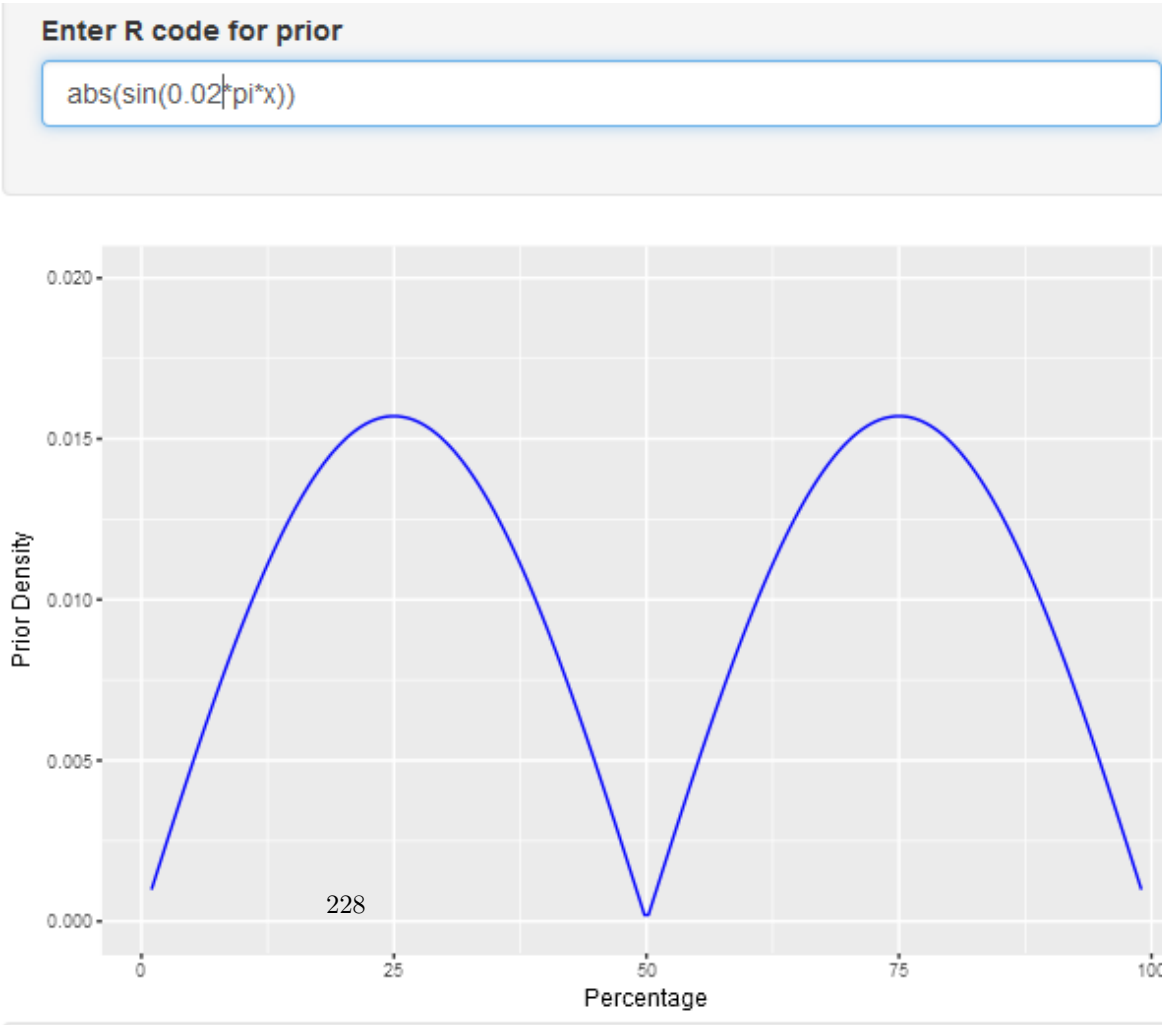


Figure 26:



or around 75%

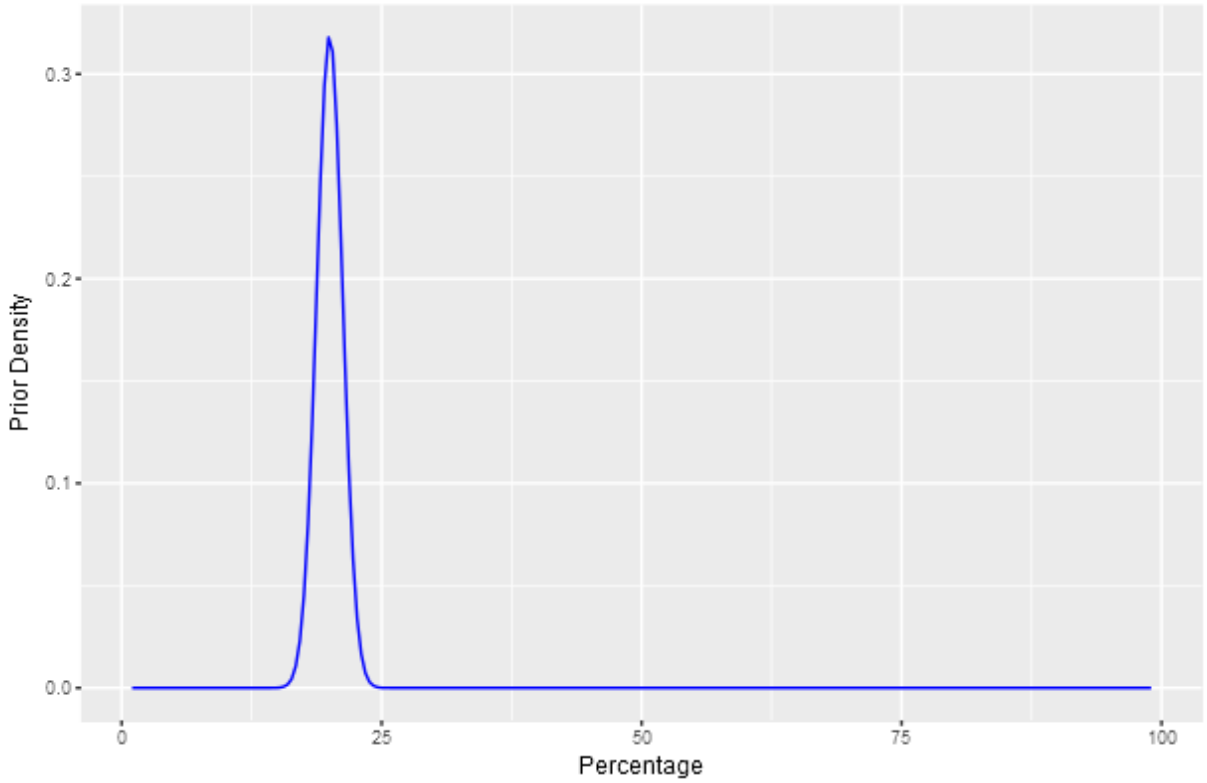


Figure 27:

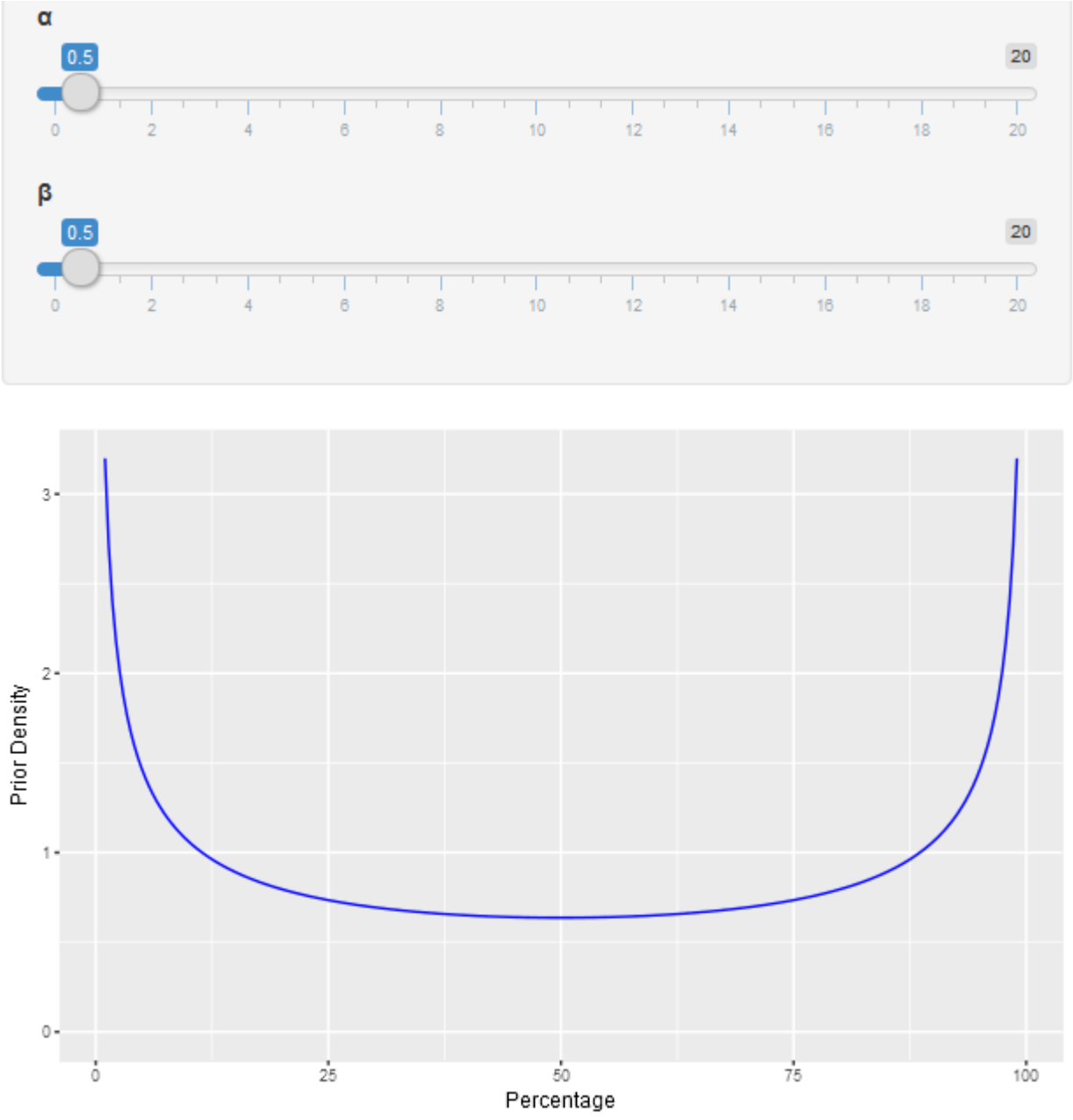

α
0.5
20
0
2
4
6
8
10
12
14
16
18
20
β
0.5
20
0
2
4
6
8
10
12
14
16
18
20
Prior Density
3
2
1
0
0
25
50
75
100
Percentage

Figure 28:

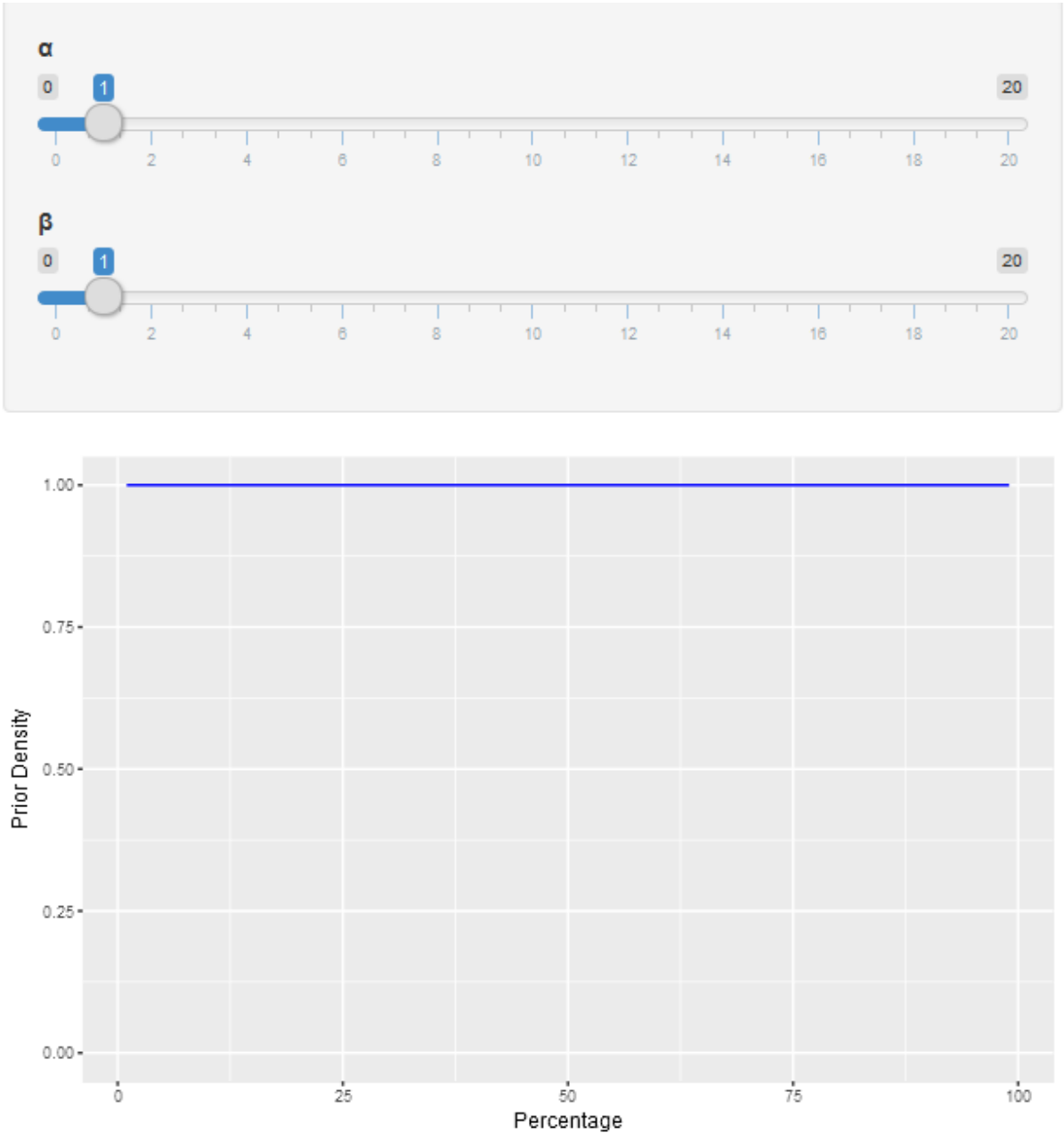

α
0
1
20
0
2
4
6
8
10
12
14
16
18
20
β
0
1
20
0
2
4
6
8
10
12
14
16
18
20
1.00
0.75
0.50
0.25
0.00
Prior Density
0
25
50
75
100
Percentage

Figure 29:

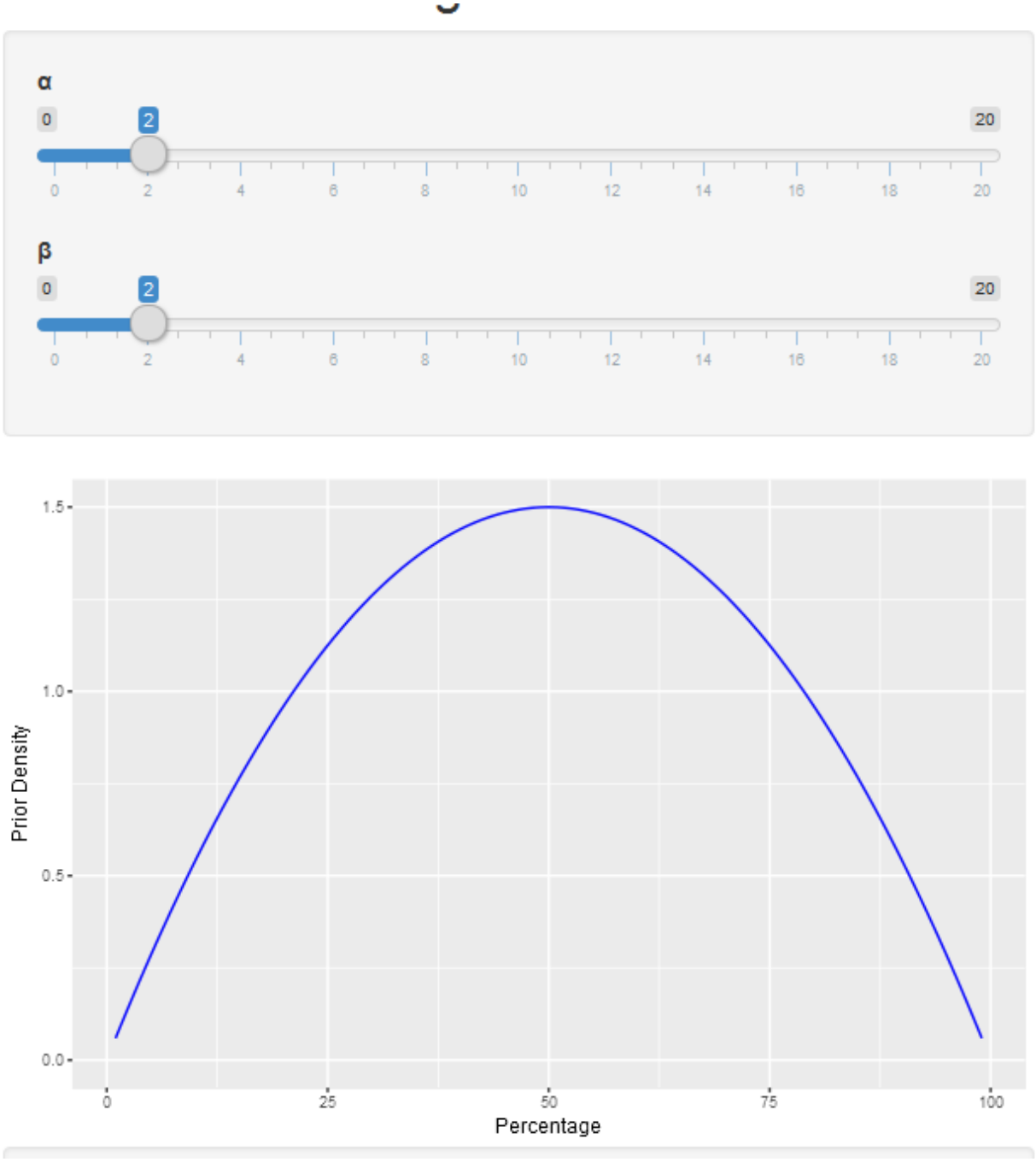

α
0
2
20
0 2 4 6 8 10 12 14 16 18 20
β
0
2
20
0 2 4 6 8 10 12 14 16 18 20
1.5
1.0
0.5
0.0
Prior Density
0
25
50
75
100
Percentage

Figure 30:

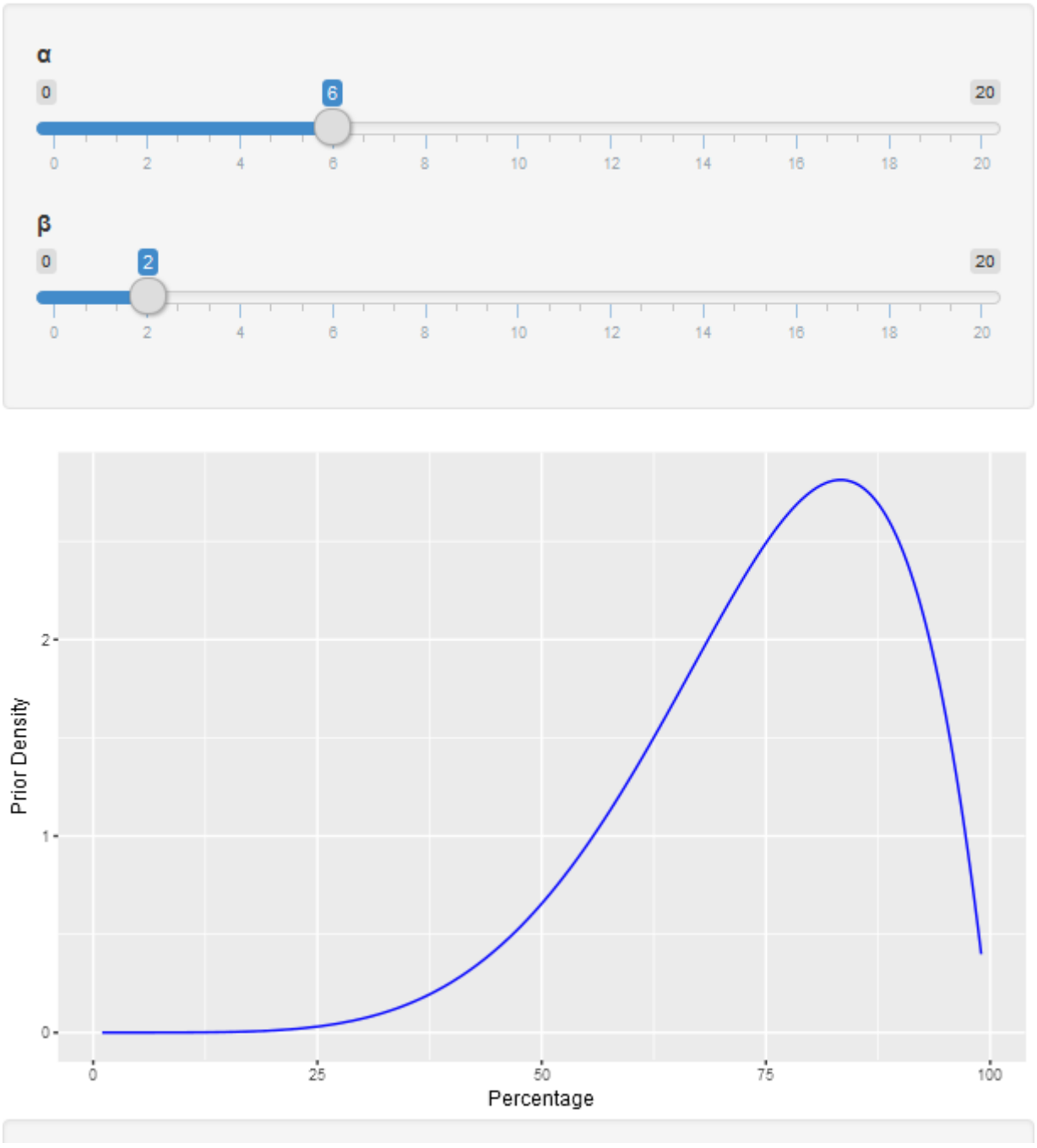

α
0
6
20
0 2 4 6 8 10 12 14 16 18 20
β
0
2
20
0 2 4 6 8 10 12 14 16 18 20
Prior Density
2
1
0
0
25
50
75
100
Percentage

Figure 31:

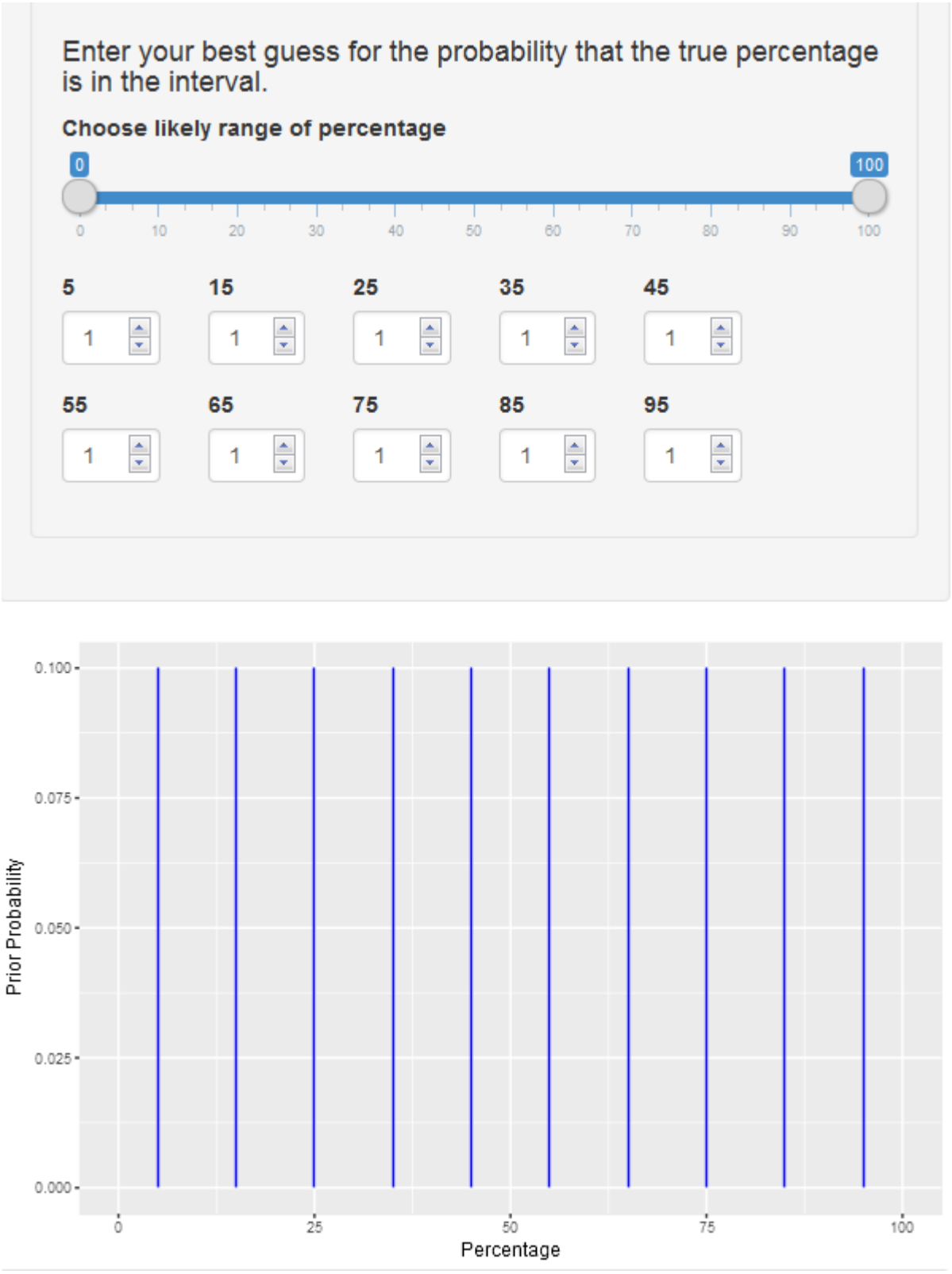
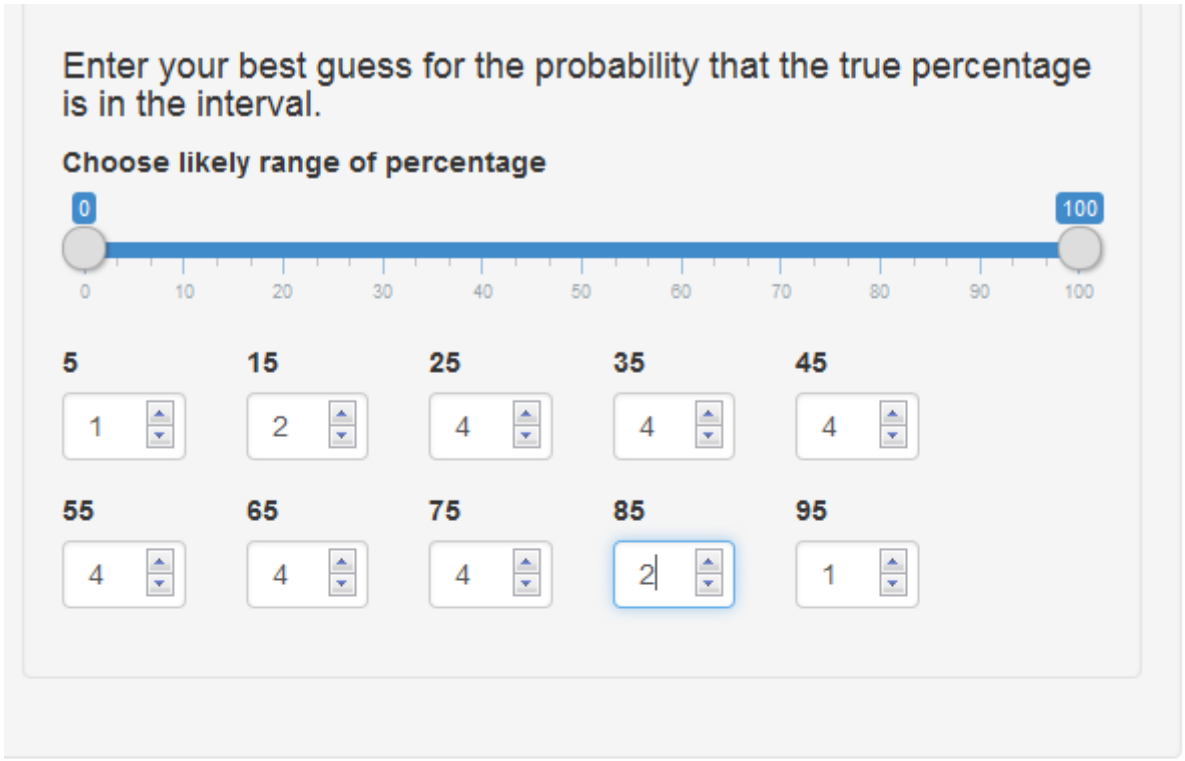

Enter your best guess for the probability that the true percentage is in the interval.
Choose likely range of percentage
0
100
0
10
20
30
40
50
60
70
80
90
100
5
15
25
35
45
1
1
1
1
1
55
65
75
85
95
1
1
1
1
1
0.100
0.075
0.050
0.025
0.000
Prior Probability
0
25
50
75
100
Percentage

Figure 32:


Enter your best guess for the probability that the true percentage is in the interval.
Choose likely range of percentage
0
100
0
10
20
30
40
50
60
70
80
90
100
5
15
25
35
45
1
2
4
4
4
55
65
75
85
95
4
4
4
2
1

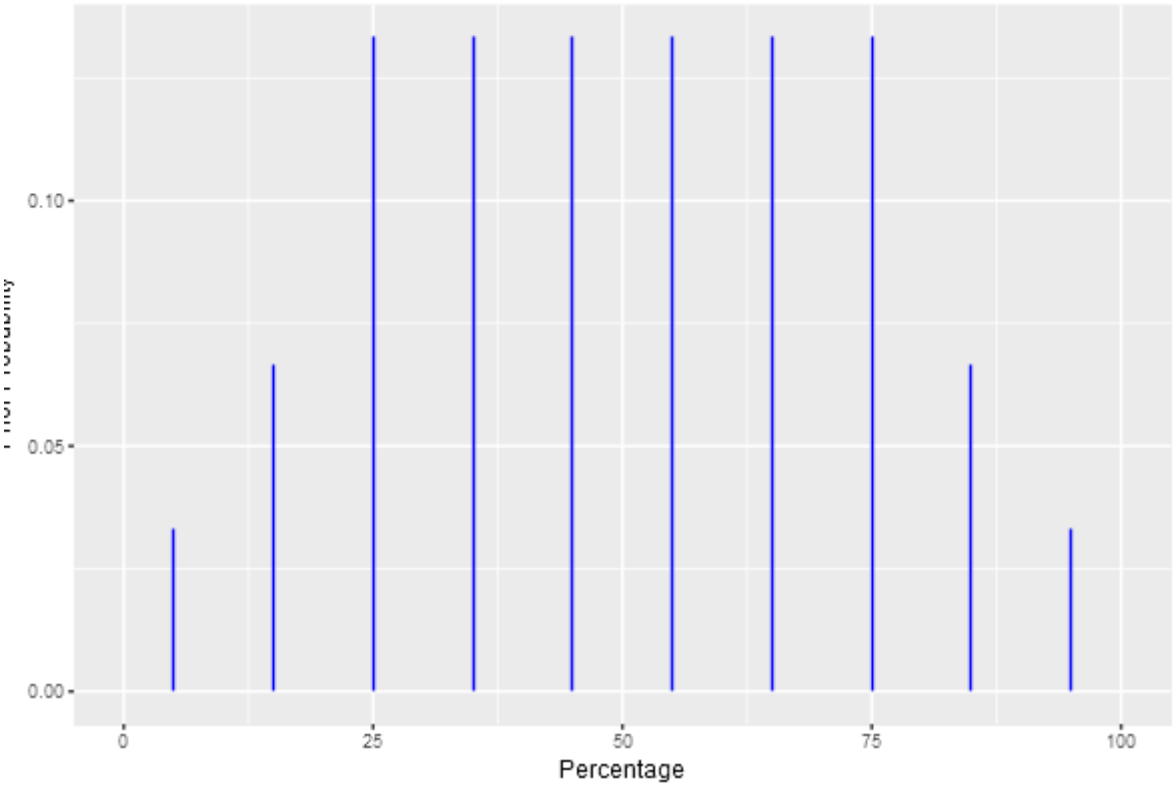

0.10
0.05
0.00
0
25
50
75
100
Percentage

Figure 33:

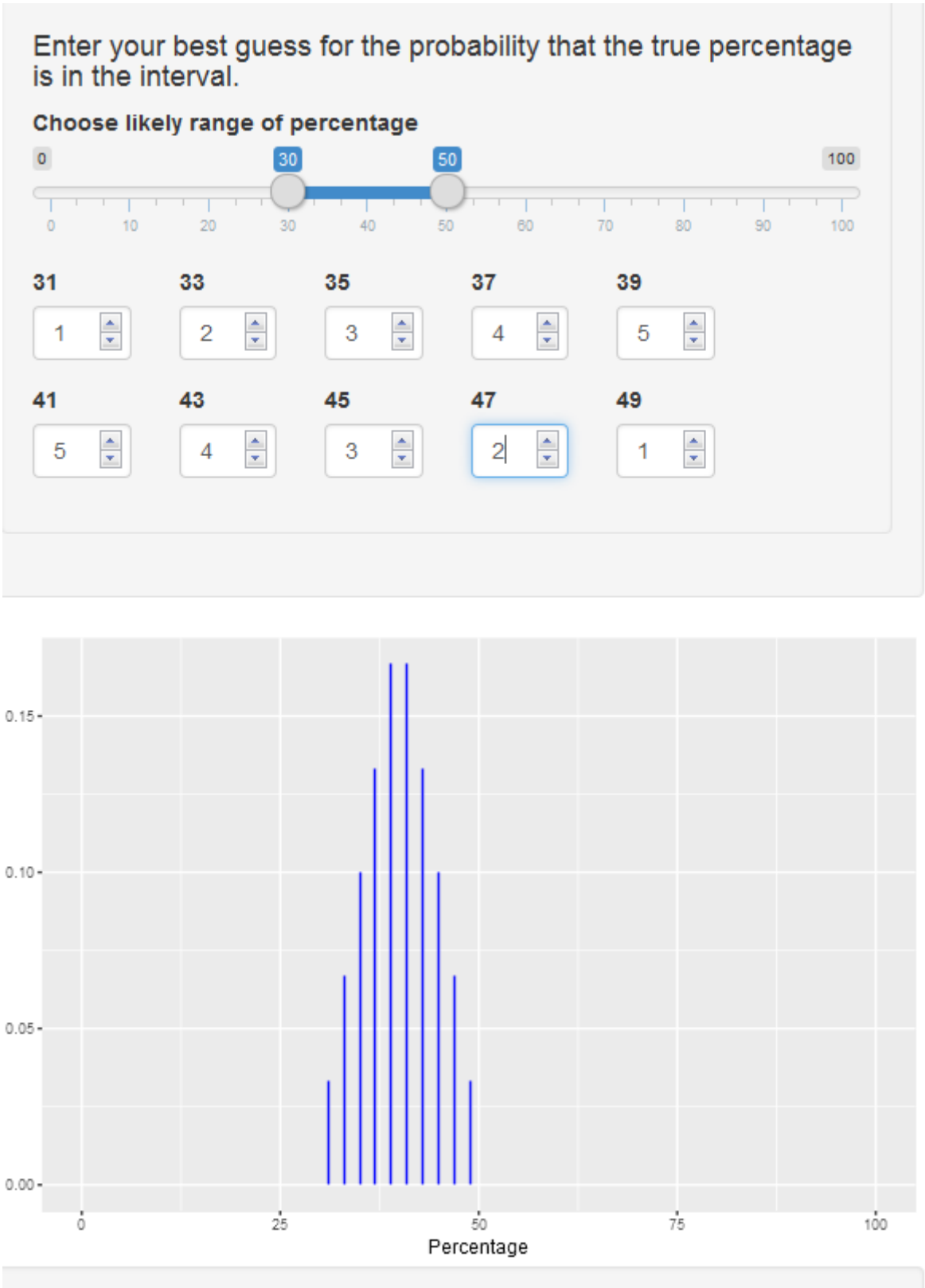

Enter your best guess for the probability that the true percentage is in the interval.
Choose likely range of percentage
0
30
50
100
0 10 20 30 40 50 60 70 80 90 100
31
1
33
2
35
3
37
4
39
5
41
5
43
4
45
3
47
2
49
1
0.15
0.10
0.05
0.00
0
25
50
75
100
Percentage

Figure 34:

Can you think of any situation where this might actually be an apprioriate prior?

26.0.0.2 Example

We have collected data from some recent classes. For each student in each class we we found their gender. What would be an appropriate prior to use here?

Actually it will depend on the class. For example, if this is a class in engineering, the percentage of females is likely smaller than 50% but if it is a course in nursing it likely larger than 50%. If we don't know what class it is we should use a prior which allows for some range. So maybe Location and Range with most likely value = 50 and likely range of values = 60

26.0.0.3 Example

We have collected data from some recent introductory statistics classes. For each student in each class we we know whether they got an A or not. What would be an appropriate prior to use here?

Here a prior with a peak at 10% seems appropriate. Moreover, any number above 20% is highly unlikely.

Try Beta prior with $\alpha = 2$ and $\beta = 20$.

26.0.0.4 Example

We have collected data from some experiment. We know the following:

- the percentage is definitely between 70% and 90%
- the percentage is most likely between 78% and 82%
- the percentage is twice as likely to be less than 78% than it is to be over 82%.

Here is one way to encode this prior knowledge with the Discrete prior option:

26.0.0.5 Example

So, how about our coin? What should we do here?

There are really two possibilities: either the coin is fair, so π is just about 0.5, and that is most likely the case. Or it is not fair, and then π could really be anything at all.

Here is one way to encode this:

The prior (blue) curve is flat from 0 to 100 but moves up sharply between 48 and 52. (This is often called Lincoln's hat function!) Under the posterior (red) curve this is still most likely a fair coin, but there is little higher chance that it has a bias towards tails.

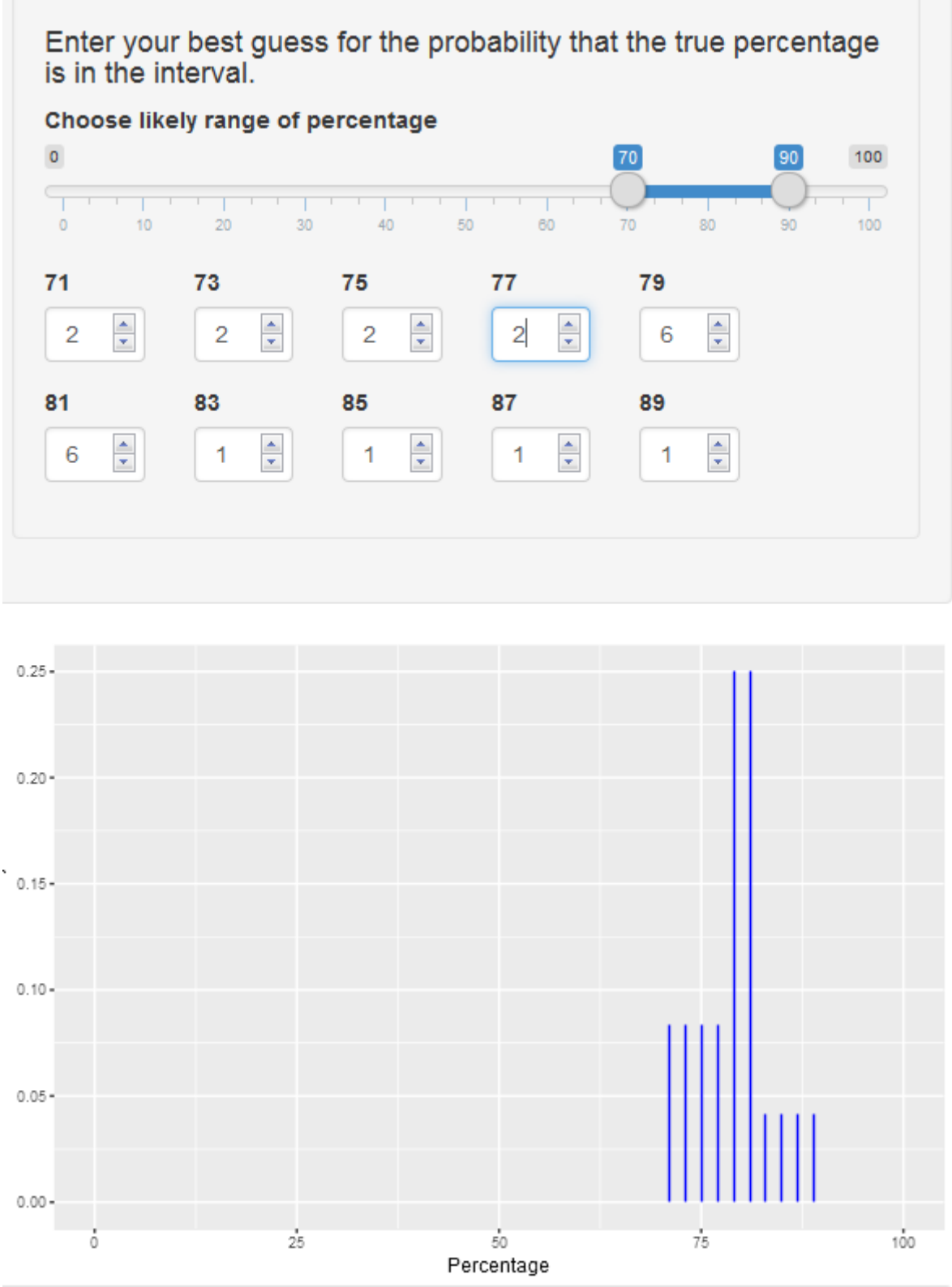

Enter your best guess for the probability that the true percentage is in the interval.
Choose likely range of percentage
0
70
90
100
0 10 20 30 40 50 60 70 80 90 100
71
73
75
77
79
2
2
2
2
6
81
83
85
87
89
6
1
1
1
1
0.25
0.20
0.15
0.10
0.05
0.00
0
25
50
75
100
Percentage

Figure 35:

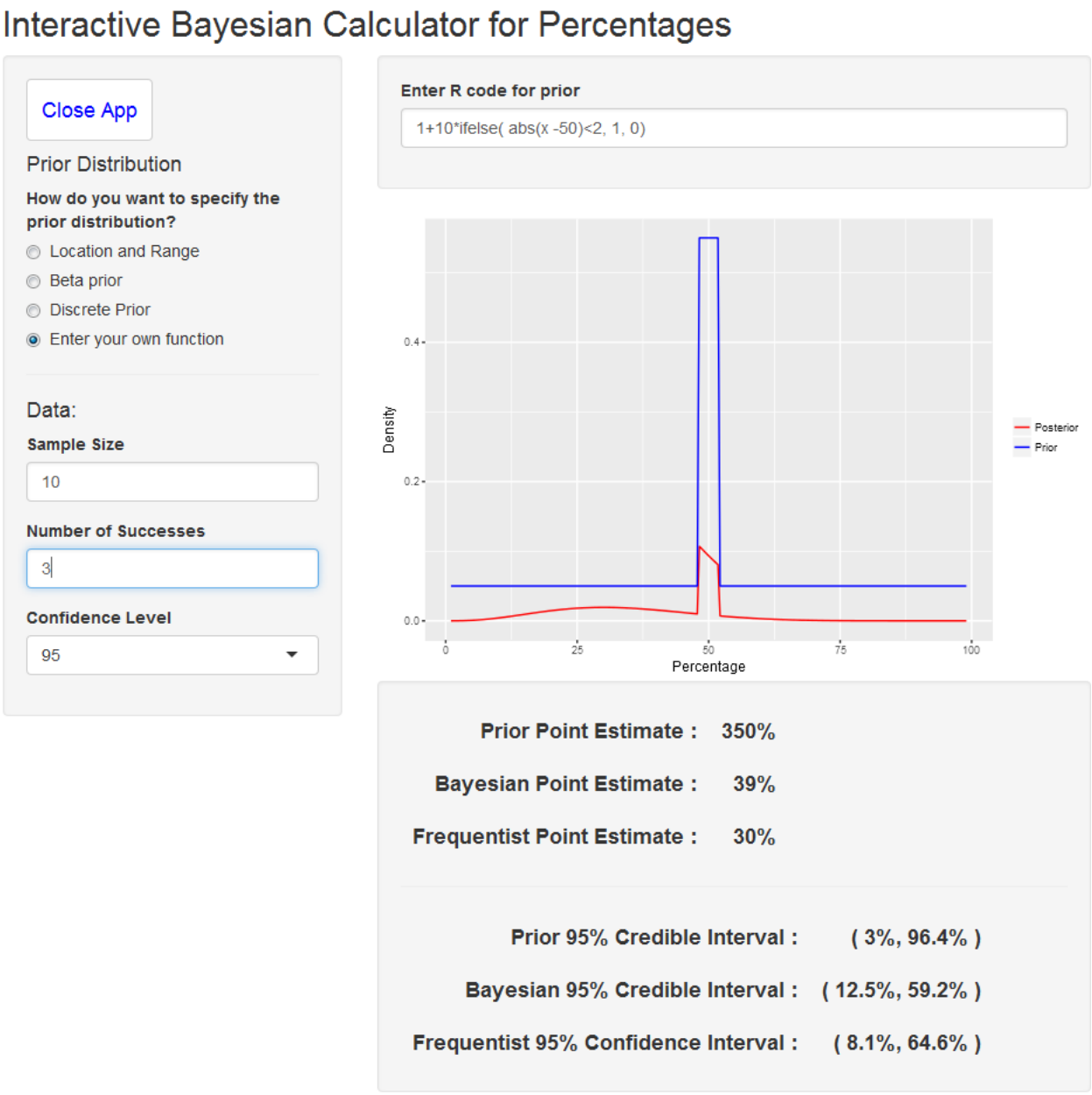

Interactive Bayesian Calculator for Percentages
Close App
Prior Distribution
How do you want to specify the prior distribution?
Location and Range
Beta prior
Discrete Prior
Enter your own function
Data:
Sample Size
10
Number of Successes
3
Confidence Level
95
Enter R code for prior
1+10*ifelse(abs(x -50)<2, 1, 0)
Density
0.4
0.2
0.0
0
25
50
75
100
Percentage
Posterior
Prior
Prior Point Estimate : 350%
Bayesian Point Estimate : 39%
Frequentist Point Estimate : 30%
Prior 95% Credible Interval : (3%, 96.4%)
Bayesian 95% Credible Interval : (12.5%, 59.2%)
Frequentist 95% Confidence Interval : (8.1%, 64.6%)

Figure 36:

27 Exercise 3: Inference

27.0.1 Problem 1:

In a sample of 130 birds of a certain species 24 were found to have specific genetic variation. Find a 90% confidence interval for the population percentage of birds with this variation. If the true percentage is 19.5%, how large a sample would be needed to get a 90% confidence interval with a width of 6%?

27.0.2 Problem 2

A researcher has a theory that predicts the mean number of sunspots per hour to be 4.5. Over a 40 hour time period he observes the following number of sunspots each hour:

5 7 5 11 12 8 5 4 0 3 4 7 2 3 2 8 6 8 3 2 2 6 5 3 6 5 6 10 7 4 3 4 6 5 8 4 3 4 5 9

Test at the 5% level whether his theory is correct.

27.0.3 Problem 3

Over some years the mean number of sick days per week taken by the employees of a company was 0.28. The company has recently paid for an education campaign regarding things like healthy eating, regular exercise etc. They now want to see whether this has improve their employees health. How many weeks do they need to wait if they want the do a hypothesis test at the 10% level and if they hope the mean number of sick days has gone down to 0.2 and they want the power of the test to be 90%. (assume the standard deviation to be 0.15)

27.0.4 Problem 4

Say we have a coin we suspect is not a fair coin. We have time to flip the coin 500 times but we only want to do this if the power of the test for a fair coin is 80%. How "unfair" does the coin have so we can do the experiment ?

27.0.5 Problem 5

In a study on the number of germs present on a square inch of floor in an average house researchers found the mean number to be 2560 with a standard deviation of 530. The study was done in 35 houses. Find a 90% confidence interval for the mean number of germs per square inch.

27.0.6 Problem 6

In a survey 37 of 110 respondents said they would be interested in buying a certain new type of kitchen appliance. Test at the 5% level whether more than one quarter of the population might be interested in buying this appliance.

27.1 Solutions

27.1.1 Problem 1:

In a sample of 130 birds of a certain species 24 were found to have specific genetic variation. Find a 90% confidence interval for the population percentage of birds with this variation.

Parameter: percentage Problem: confidence interval

Method: one.sample.prop

```
one.sample.prop(x = 24, n = 130, conf.level = 90)
```

```
## A 90% confidence interval for the population proportion is (0.132, 0.251)
```

so a 90% confidence interval for the population percentage of birds with this variation is (13.2%, 24.1%)

If the true percentage is 19.5%, how large a sample would be needed to get a 90% confidence interval with a width of 6%?

Problem: sample size
Method: prop.ps

width 6% = width 0.06 for proportions, so E = 0.03

```
prop.ps(pi.null = 0.195, E = 0.03, conf.level = 90)
```

```
## [1] "Sample size required is  752"
```

27.1.2 Problem 2

A researcher has a theory that predicts the mean number of sunspots per hour to be 4.5. Over a 40 hour time period he observes the following number of sunspots each hour:

5 7 5 11 12 8 5 4 0 3 4 7 2 3 2 8 6 8 3 2 2 6 5 3 6 5 6 10 7 4 3 4 6 5 8 4 3 4 5 9

Test at the 5% level whether his theory is correct.

Parameter: mean

Problem: hypothesis test

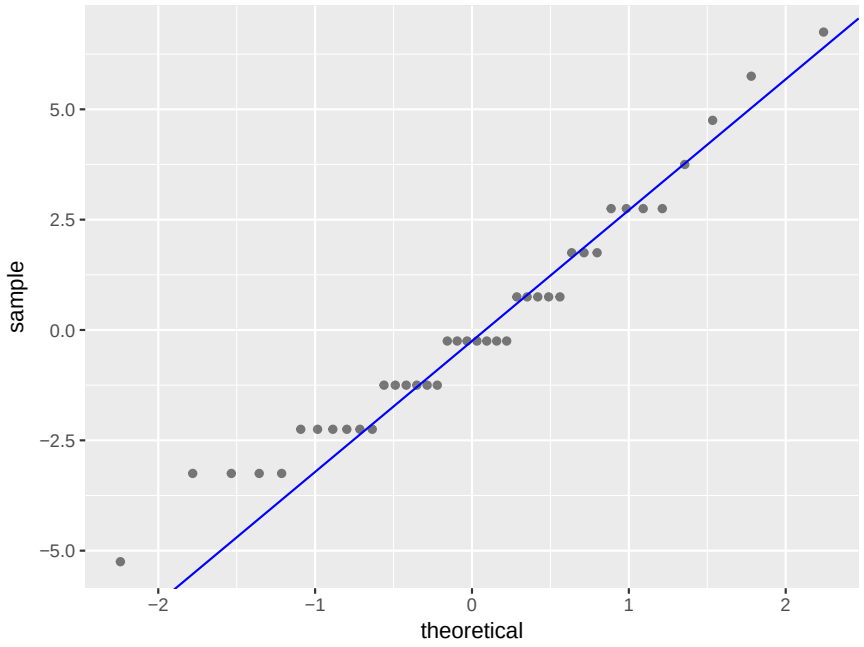
Method: one.sample.t

1. Parameter: mean μ

2. Method: 1-sample t

3. Assumptions: normalplot is ok
4. $\alpha = 0.05$
5. $H_0 : \mu = 4.5$ (Theory correct)
6. $H_a : \mu] \neq 4.5$ (Theory false)
7. $p = 0.0778$

```
x <-
c(5, 7, 5, 11, 12, 8, 5, 4, 0, 3, 4, 7, 2, 3, 2, 8, 6, 8, 3,
2, 2, 6, 5, 3, 6, 5, 6, 10, 7, 4, 3, 4, 6, 5, 8, 4, 3, 4, 5,
9)
one.sample.t(x, mu.null = 4.5)
```




```
## p value of test H0: mu=4.5 vs. Ha: mu <> 4.5:  0.0778
```

8. $p > \alpha$, so we fail to reject the null hypothesis
9. the theory appears to be correct

Assumptions: normal plot ok

27.1.3 Problem 3

Over some years the mean number of sick days per week taken by the employees of a company was 0.28. The company has recently paid for an education campaign regarding things like healthy eating, regular exercise etc. They now want to see whether this has improve their employees health. How many weeks do they need to wait if they want the do a hypothesis

test at the 10% level and if they hope the mean number of sick days has gone down to 0.2 and they want the power of the test to be 90%. (assume the standard deviation to be 0.15)

Parameter: mean

Problem: sample size

Method: t.ps

```
t.ps(diff = 0.2-0.28, sigma = 0.15, power = 90,
     alpha = 0.1, alternative = "less")
```

```
## Sample size required is  25
```

27.1.4 Problem 4

Say we have a coin we suspect is not a fair coin. We have time to flip the coin 500 times but we only want to do this if the power of the test for a fair coin is 80%. How "unfair" does the coin have so we can do the experiment ?

Parameter: proportion

Problem: power

Method: prop.ps

```
prop.ps(n = 500, pi.null = 0.5, power = 80)
```

```
## [1] "Number of Successes required for power: < 219  or > 281"
```

```
c(219, 281)/500
```

```
## [1] 0.438 0.562
```

so if the true probability of heads is either less than 0.438 or higher than 0.562 the test will have a power of 80% or higher.

27.1.5 Problem 5

In a study on the number of germs present on a square inch of floor in an average house researchers found the mean number to be 2560 with a standard deviation of 530. The study was done in 35 houses. Find a 90% confidence interval for the mean number of germs per square inch.

Parameter: mean

Problem: confidence interval

Method: one.sample.t

```
one.sample.t(y = 2560, shat = 530,
             n = 35, conf.level = 90)
```

```
## A 90% confidence interval for the population mean is (2408.5, 2711.5)
```

27.1.6 Problem 6

In a survey 37 of 110 respondents said they would be interested in buying a certain new type of kitchen appliance. Test at the 5% level whether more than one quarter of the population might be interested in buying this appliance.

Parameter: proportion

Problem: hypothesis test

Method: one.sample.prop

1. Parameter: proportion π
2. Method: exact binomial
3. Assumptions: None
4. $\alpha = 0.05$
5. $H_0 : \pi = 0.25$
6. $H_a : \pi > 0.25$
7. $p = 0.0264$

```
one.sample.prop(x = 37, n = 110, pi.null = 0.25,
        alternative = "greater")
```

```
## p value of test H0: pi=0.25 vs. Ha: pi > 0.25:  0.0238
```

8. $p = 0.0264 < 0.05$, so we reject the null hypothesis.
9. it appears over one quarter of the population might be interested in buying this appliance.

28 Correlation Test

We can now return to the case of two quantitative variables and the question of whether or not they are related. Specifically, we have a test of

$H_0 : \rho = 0$ (no relationship) vs $H_a : \rho \neq 0$ (some relationship)

The assumptions of the test are that the relationship is linear and that there are no outliers. We can use the **mplot** command to check them. The command to find the p value is **pearson.cor.**

28.0.1 Case Study: The 1970's Military Draft

We have previously used simulation to see that a sample correlation r=-0.226 is very unusual (for n=366). Now we can do the formal test:

1) Parameter: Pearson's correlation coefficient ρ
2) Method: Test for Pearson's correlation coefficient ρ
3) Assumptions: relationship is linear and that there are no outliers.
4) $\alpha = 0.05$

5) $H_0 : \rho = 0$ (no relationship between Day of Year and Draft Number)
6) $H_a : \rho \neq 0$ (some relationship between Day of Year and Draft Number)
7) p = 0.0000

```
pearson.cor(Draft.Number, Day.of.Year, rho.null = 0)
```

```
## p value of test H0: rho=0 vs. Ha: rho <> 0:  0.000
```

8) $p = 0.0000 < \alpha = 0.05$, so we reject the null hypothesis,
9) There is a statistically significant relationship between Day of Year and Draft Number.

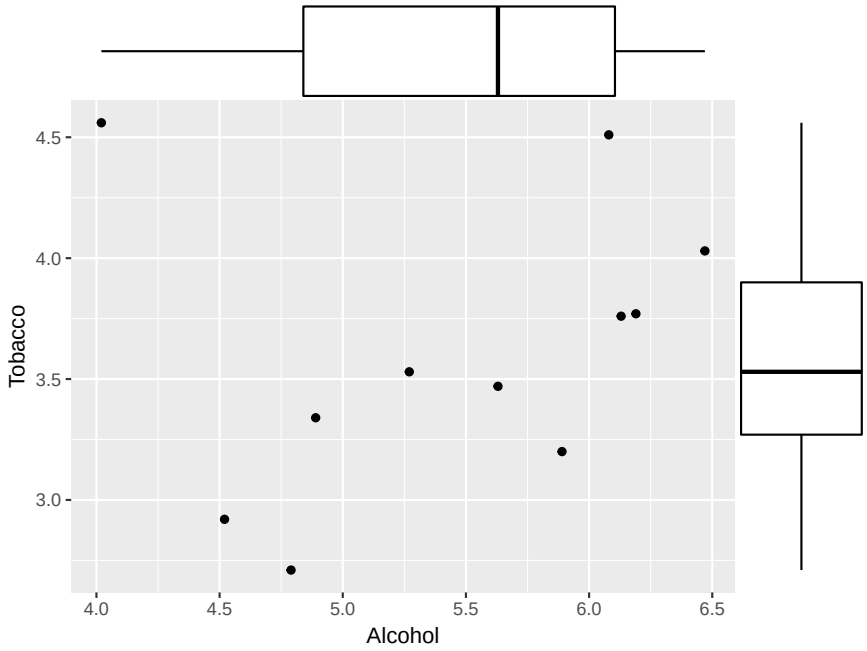
Assumptions: boxplots and scatterplot show no outliers. No non-linear relationship.

28.0.2 Case Study: Alcohol vs. Tobacco Expenditure

Data from a British government survey of household spending may be used to examine the relationship between household spending on tobacco products and alcoholic beverages. The numbers are the average expenditure for each of the 11 regions of England.

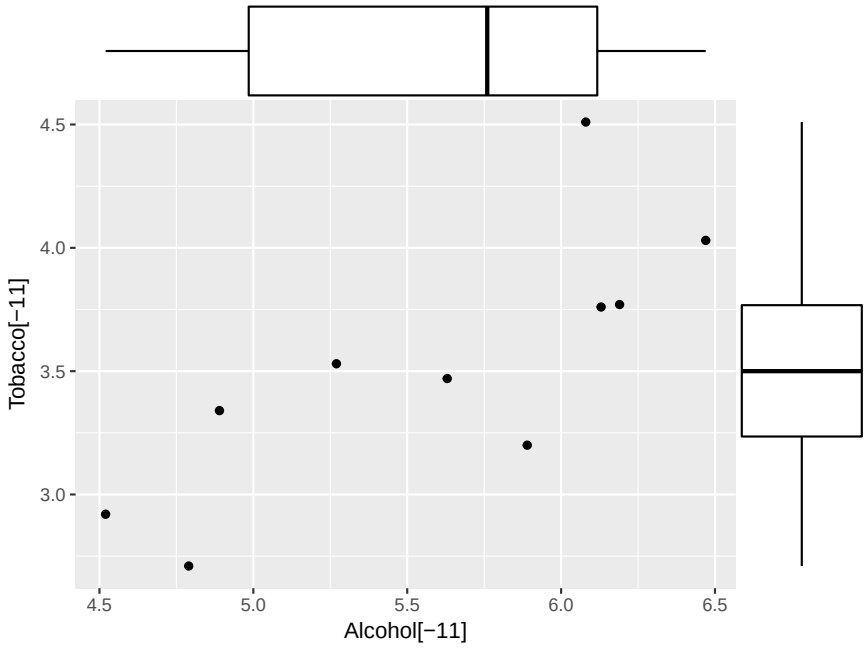
The marginal plot shows one outlier:

```
attach(alcohol)
mplot(Tobacco, Alcohol)
```



This is Northern Ireland, observation #11. Eliminating this observations show no more outleirs, and a linear relationship:

```
mplot(Tobacco[-11], Alcohol[-11])
```



So the test is:

1) Parameter: Pearson's correlation coefficient ρ
2) Method: Test for Pearson's correlation coefficient ρ
3) Assumptions: relationship is linear and that there are no outliers.
4) $\alpha = 0.05$

5) $H_0 : \rho = 0$ (no relationship between Alcohol and Tobacco)
6) $H_a : \rho \neq 0$ (some relationship between Alcohol and Tobacco)
7) $p = 0.0072$

```
pearson.cor(Tobacco[-11], Alcohol[-11], rho.null = 0)
```

```
## p value of test H0: rho=0 vs. Ha: rho <> 0:  0.0072
```

8) $p = 0.0072 < \alpha = 0.05$, so we reject the null hypothesis,
9) There is a statistically significant relationship between Alcohol and Tobacco

Note: Running the test with Northen Ireland would have given the wrong answer:

```
pearson.cor(Tobacco, Alcohol, rho.null = 0)
```

```
## p value of test H0: rho=0 vs. Ha: rho <> 0:  0.5087
```

p value of test for now is: $0.5087 > 0.05$

29 Regression

If there is a relationship between variables "x" and "y", can we describe it?

We do that by finding a **model**, that is an equation y=f(x) Here we keep it very simple and consider only **linear** relationships, that is equations of the form

$$y = mx + b$$

In Statistics though we use a slightly different notation:

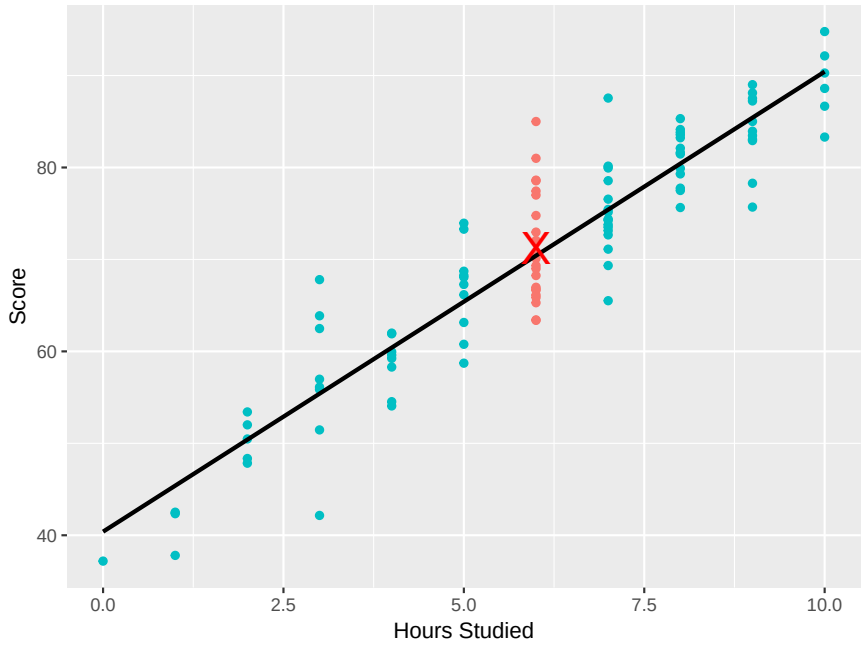
$$y = \beta_0 + \beta_1 x$$

The logic here is this: if we know x, we can compute y. Unfortunately there are always "errors" in this calculation, so the answer y varies even for the same x.

For example, let x be the number of hours a student studies for an exam, and y the score on the exam. Say we know from long experience that y=50+5x. So even if the student doesn't study at all (x=0) he/she would still get around 50 points, and for every hour studied the score goes up by about 5 points.

But of course there are many other factors influencing the grade such as general ability, previous experience, being healthy on the day of the exam, exam anxiety etc, so for any specific student the score will not be exactly what the equation predicts. So if three students all study 6 hours, the equation predicts a score of 50+5*6=80 for all of them but one might get a 69, the next a 78 and the third a 99. What the equation predicts is actually their mean score.

This is illustrated in the next graph:



where the scores of the people who studied 6 hours are in red, and their mean score is marked by an X.

29.1 App

```
run.app(lsr1)
```

this app illustrates the meaning of the line as the mean response.

β_0 and β_1 are numbers that depend on the population from which the data (X,Y) is drawn. Therefore they are **parameters** just like the mean or the median.

A standard problem is this: we have a dataset and we believe there is a linear relationship between x and y. We would like to know the equation

$$y = \beta_0 + \beta_1 x$$

that is we need to "guess" what β_0 and β_1 are. We will estimate them by a method called **least squares regression**. This is done by the R command **slr**:

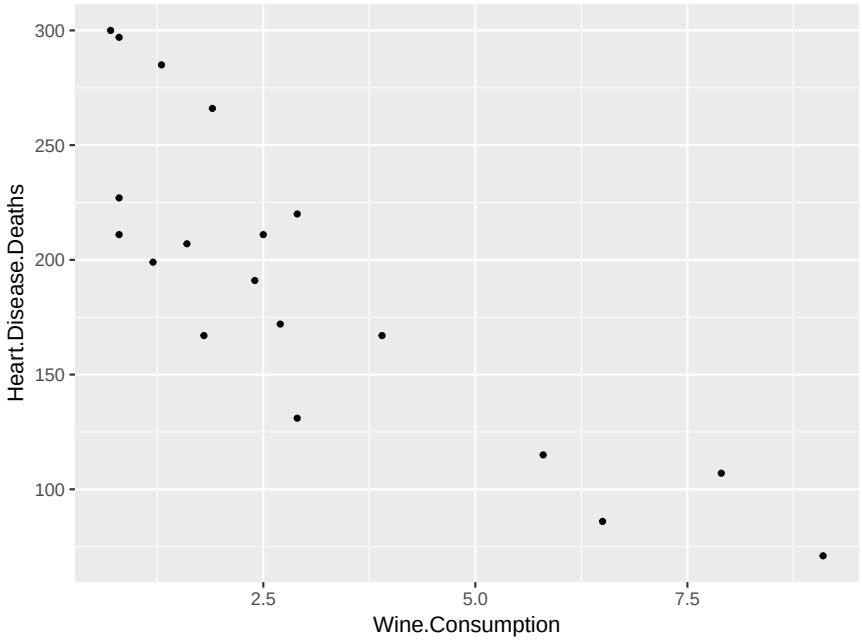
29.1.1 Wine Consumption and Heart Disease

Data from a study on consumption of wine (in liters per person) and heart disease rates (in per 100000) in 19 countries.

```
wine
```

```
##           Country Wine.Consumption Heart.Disease.Deaths
## 1       Australia              2.5                  211
## 2         Austria              3.9                  167
## 3         Belgium              2.9                  131
## 4          Canada              2.4                  191
## 5         Denmark              2.9                  220
## 6         Finland              0.8                  297
## 7          France              9.1                   71
## 8         Iceland              0.8                  211
## 9         Ireland              0.7                  300
## 10          Italy              7.9                  107
## 11    Netherlands              1.8                  167
## 12    New Zealand              1.9                  266
## 13         Norway              0.8                  227
## 14          Spain              6.5                   86
## 15         Sweden              1.6                  207
## 16    Switzerland              5.8                  115
## 17 United Kingdom              1.3                  285
## 18  United States              1.2                  199
## 19        Germany              2.7                  172
```

```
attach(wine)
splot(Heart.Disease.Deaths, Wine.Consumption)
```



so we see a clear negative correlation, the higher the wine consumption, the lower the heart disease rate.

Careful: this was an observational study, here correlation does NOT imply causation!

So, what can we say about the actual relationship?

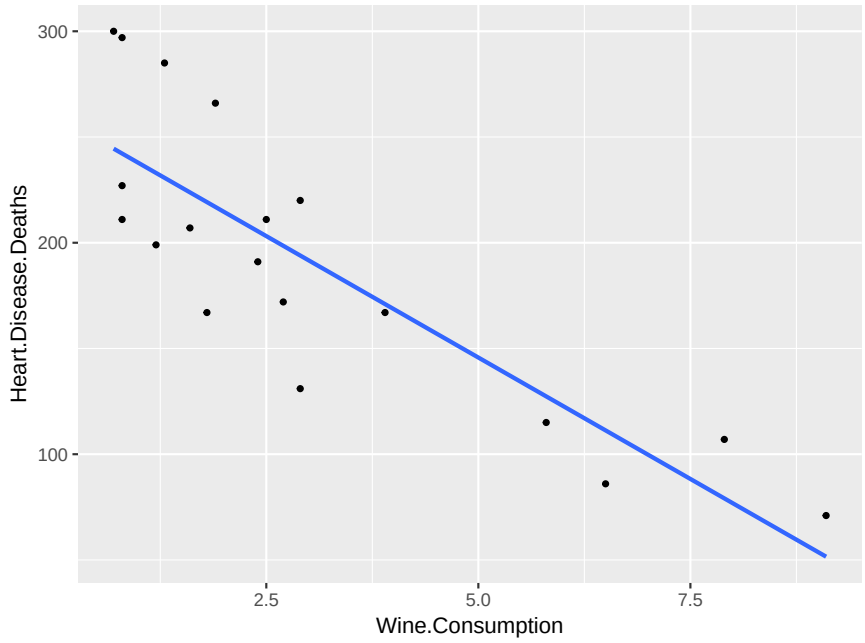
```
slr(Heart.Disease.Deaths, Wine.Consumption)
```

```
## The least squares regression equation is:
##  Heart.Disease.Deaths  = 260.563 - 22.969 Wine.Consumption
## R^2 = 71.03%
```

Note the slr command also draws a graph, which we can ignore.

There is a nice graph called the fitted line plot, which is the scatterplot with the least square regression line added to it:

```
splot(Heart.Disease.Deaths, Wine.Consumption, add.line=1)
```



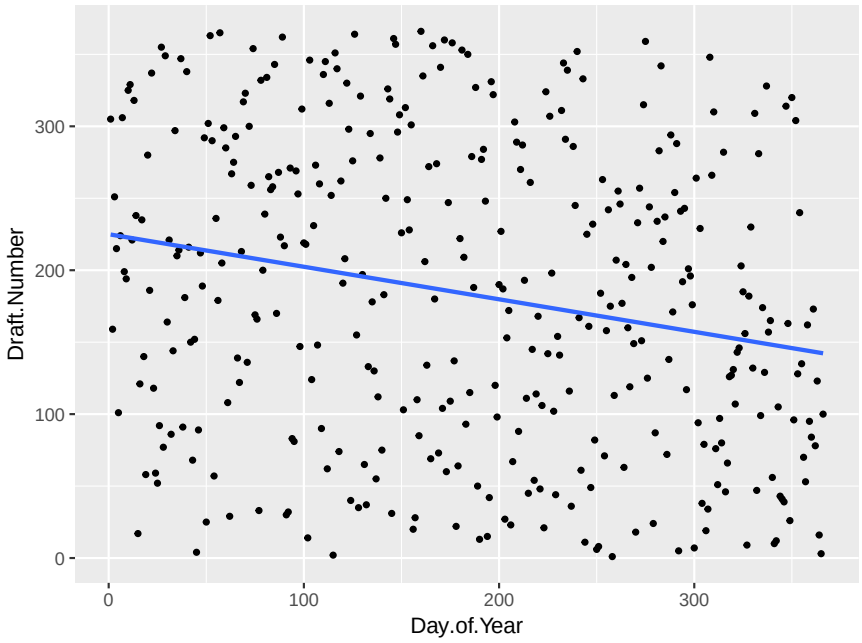
29.1.2 Case Study: 1970's Draft

```
attach(draft)
```

```
slr(Draft.Number, Day.of.Year)
```

```
## The least squares regression equation is:
##  Draft.Number  = 225.009 - 0.226 Day.of.Year
## R^2 = 5.11%
```

```
splot(Draft.Number, Day.of.Year, add.line = 1)
```



29.1.3 App

```
run.app(lsr)
```

this app illustrates the least squares regression line.

Just play around with different slopes and intercepts and see how the fitted line plot and the regression equation change

---

Here are two important facts about least squares regression:

1) say $\overline{X}$ is the mean of the x vector and $\overline{Y}$ is the mean of the y vector, then ($\overline{X}$, $\overline{Y}$) is **always** a point on the line.
2) We have seen previously that for the correlation coefficient it does not matter what variable we choose as X and which as Y, that is we have

cor(x,y) = cor(y,x)

Now let's see what happens in regression

```
slr(Heart.Disease.Deaths, Wine.Consumption)
```

```
## The least squares regression equation is:
##  Heart.Disease.Deaths  = 260.563 - 22.969 Wine.Consumption
## R^2 = 71.03%
```

The least squares regression equation is:

Heart.Disease.Deaths = 260.563 - 22.969 Wine.Consumption

so

22.969 Wine.Consumption = 260.563 - Heart.Disease.Deaths

and

Wine.Consumption = 260.563/22.969 - 1/22.969 Heart.Disease.Deaths

Wine.Consumption = 11.34 - 0.044 Heart.Disease.Deaths

BUT

```
slr(Wine.Consumption, Heart.Disease.Deaths)
```

```
## The least squares regression equation is:
##  Wine.Consumption  = 8.935 - 0.031 Heart.Disease.Deaths
## R^2 = 71.03%
```

and that is not the same equation!

So in regression it is important to distinguish between the **predictor or independent variable (x)**

and the

response or **dependent variable (y)**.

29.2 Regression towards the Mean

29.2.1 Case Study: The Sports Illustrated Jinx

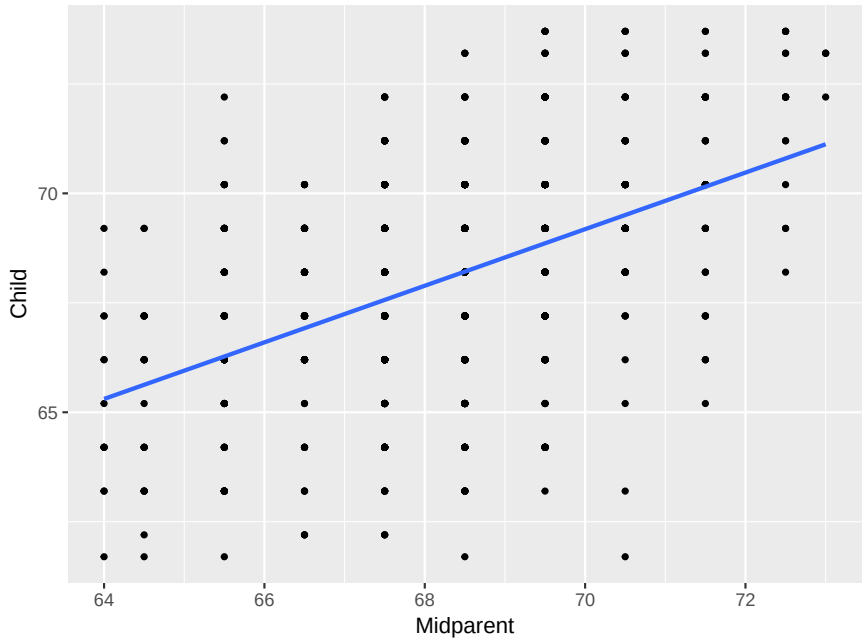
It has often been noted that anyone featured with his/her picture on the cover of Sports Illustrated is then jinxed, that is their performance goes down. Some people have tried to find an explanation for this, for example that an athlete gets lazy after being successful, or maybe that they have to many media days and can't practise enough. In reality this is just an example of **regression to the mean**.

29.2.2 Case Study: Galton's Data on the Heights of Parents and their Children

At the end of the 19th century Sir Francis Galton collected the height of the parent (actually fathers) and the height of the oldest child (son) of almost 1000 families.

The fitted line plot is

```
attach(galton)
splot(Child, Midparent, add.line= 1)
```



and the least squares regression equation is

```
slr(Child, Midparent)
```

```
## The least squares regression equation is:
##  Child  = 23.942 + 0.646 Midparent
## R^2 = 21.05%
```

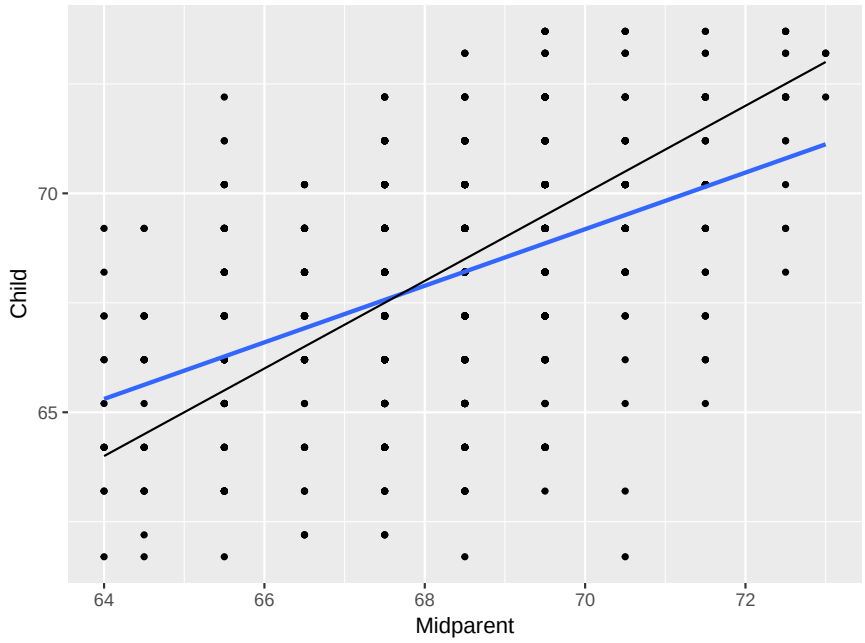
Notice: the slope of the line (0.646) is

a) $\beta_1 > 0$
b) $\beta_1 < 1$

so we see that

a) children of small (tall) parents tend to be small (tall) also
b) but not as small (tall) as their parents!

Let's add the line with slope 1 to the graph:

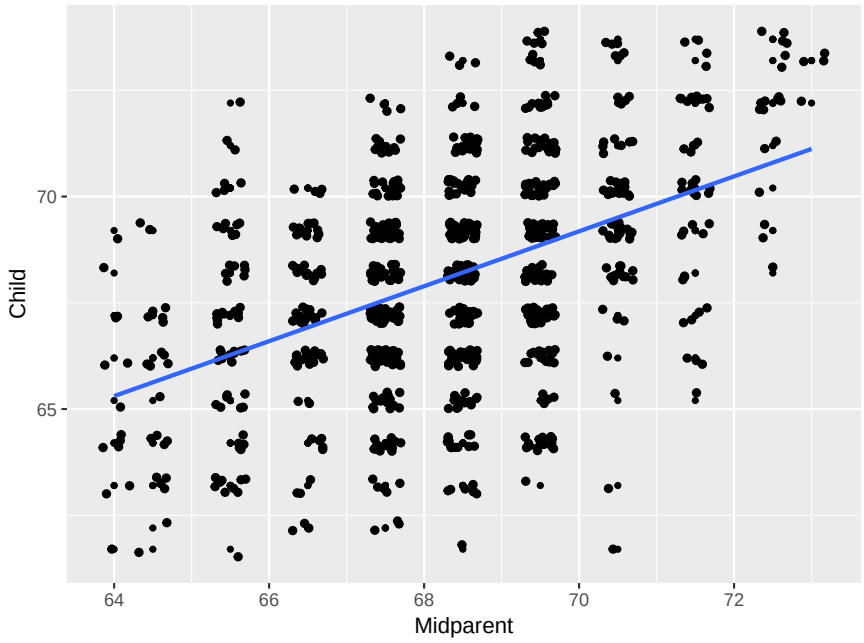


so we see that those observations on the extreme tend to "regress" (come back) to the "middle".

Of course this makes good sense: if a person is tall because of their genes (and the child has half of those) but also because of a lot of other factors, many of which the child does not have.

There is something about the graph that is not very nice. Because the heights were recorded only to within the nearest .2 inches lot's of data points repeat, but they appear only once in the graph. We can fix that by **jittering** the points, that is moving them randomly around just a little bit:

```
splot(Child,  Midparent, jitter=TRUE, add.line= 1)
```



Example Ever notice that often those students that do very well on the first exam do not quite so well on the second? Is it that they got lazy, thought the exams are easy, that they could do well without studying?

Maybe, but maybe it also just an example of regression to the mean!

Example say we have the following data: a group of subjects is participating in a weight loss program. They are weighted before and after the program. Now we pick out the (say) 10% people that were the heaviest at the beginning, and we notice that their average weight was 207 pounds then but is only 201 pounds at the end of the program. Can we conclude the program worked (at least for heavy people)?

Maybe, but not necesserily. Again the same outcome could be due to regression to the mean.

29.2.3 Case Study: The Sports Illustrated Jinx

This also explains the Jinx: an athlete gets on the cover after having done very well, likely a bit better than is normal even for them, after a while (Cover of no Cover) they will **regress to the mean**.

Regression towards the mean is one of those Statistical phenomena that is often misunderstood, with people looking for an explanation were this is none!

29.2.4 Praise vs Punishment

The psychologist Daniel Kahneman, winner of the 2002 Nobel Memorial Prize in Economic Sciences, pointed out that regression to the mean might explain why rebukes can seem to improve performance, while praise seems to backfire:

I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning. When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand and made his own short speech, which began by conceding that positive reinforcement might be good for the birds, but went on to deny that it was optimal for flight cadets. He said, "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not, because the opposite is the case." This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them. I immediately arranged a demonstration in which each participant tossed two coins at a target behind his back, without any feedback. We measured the distances from the target and could see that those who had done best the first time had mostly deteriorated on their second try, and vice versa. But I knew that this demonstration would not undo the effects of lifelong exposure to a perverse contingency.

For more on regression to the mean go here.

30 Exercise 4: Correlation and Regression

30.0.1 Problem 1:

Consider round 1 and and 2 of the Sony open golf tournament (data set **golfscores**). Is there a statistically significant relationship between the scores?

30.0.2 Problem 2:

Consider round 1 and and 2 of the Sony open golf tournament (data set **golfscores**). What is the least squares regression equation with Sony 1 as the predictor variable? Draw the fitted line plot. Is there an indication of "regression to the mean"? Why?

30.0.3 Problem 3:

Consider the men's long jump in the Olympics (**longjump**). How strong is the relationship between Year and LongJump?

30.0.4 Problem 4:

Consider the following data set:

x	y
10	58
11	54
12	51
13	52
14	62
15	57
16	63
17	64
18	69
19	71
20	70

Find the least squares regression equation and use it to predict the y value for an observation with x=15

30.1 Solutions

30.1.1 Problem 1:

Consider round 1 and and 2 of the Sony open golf tournament (data set **golfscores**). Is there a statistically significant relationship between the scores?

Parameter: correlation coeffcient

Problem Test for independence

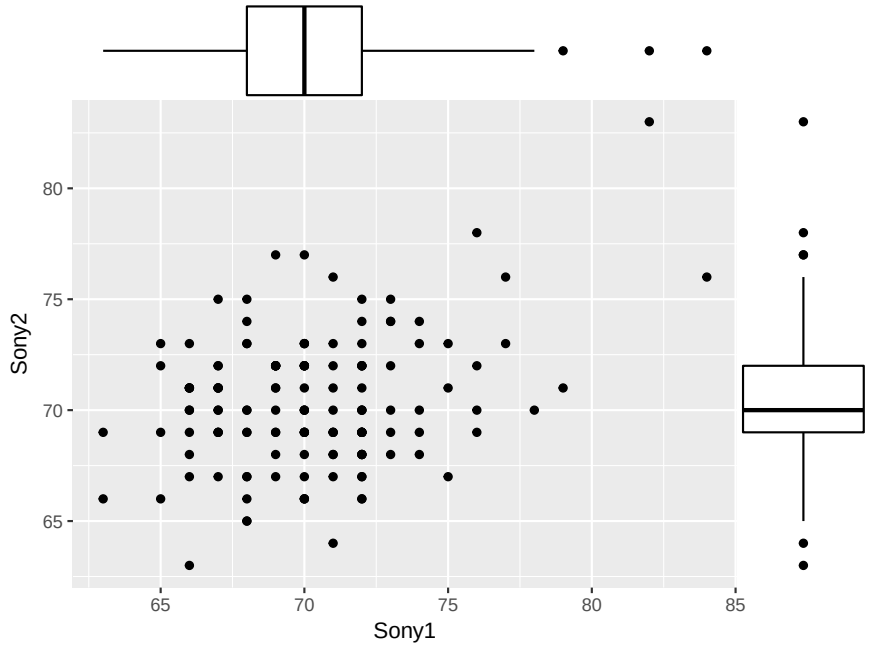
Method: pearson.test

```
attach(golfscores)
```

1) Parameter: Pearson's correlation coefficient ρ
2) Method: Test for Pearson's correlation coefficient ρ
3) Assumptions: relationship is linear and that there are no outliers.
4) $\alpha = 0.05$

5) H_0: $\rho = 0$ (no relationship between Day of Year and Draft Number)
6) H_a: $\rho \neq 0$ (some relationship between Day of Year and Draft Number)
7) p = 0.000

```
pearson.cor(Sony1, Sony2, rho.null=0)
```




```
## p value of test H0: rho=0 vs. Ha: rho <> 0:  0.000
```

8) $p < \alpha = 0.05$, so we reject the null hypothesis,
9) There is a statistically significant relationship between Day of Year and Draft Number.

Assumptions: boxplots and scatterplot show no outliers. No non-linear relationship.

30.1.2 Problem 2:

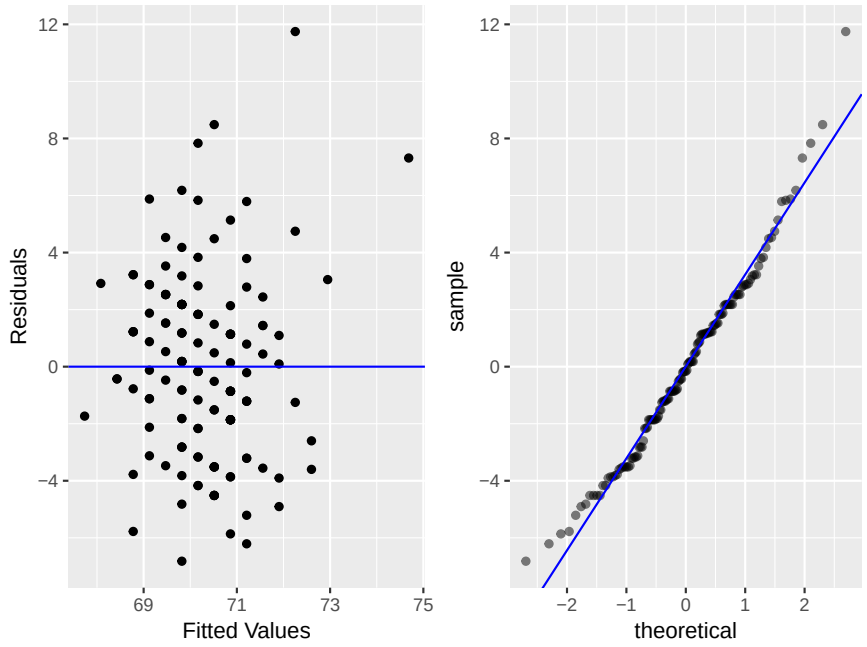
Consider round 1 and and 2 of the Sony open golf tournament (data set **golfscores**). What is the least squares regression equation with Sony 1 as the predictor variable? Draw the fitted line plot. Is there an indication of "regression to the mean"? Why?

Parameter: regression coefficients

Problem: find model

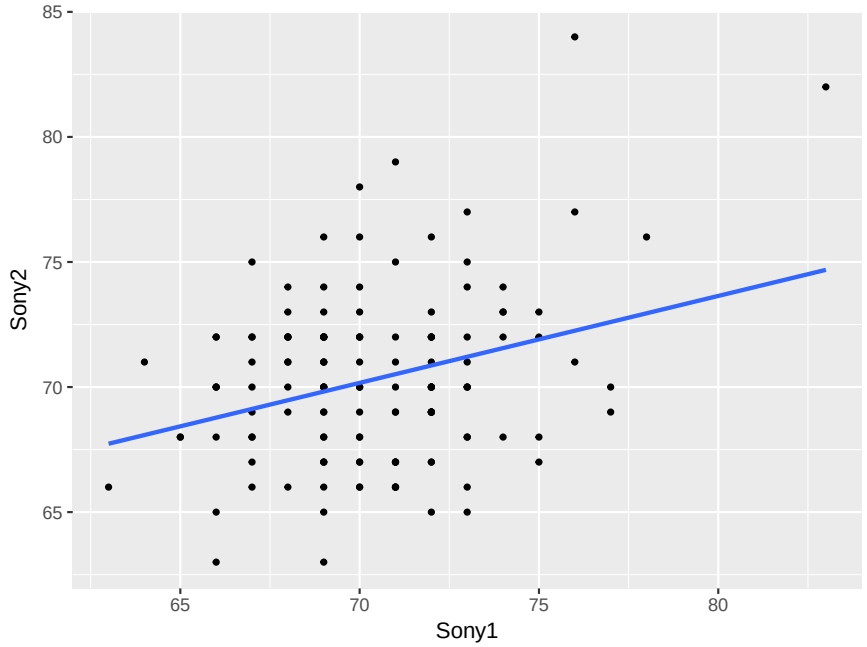
Method: slr

```
slr(Sony2, Sony1)
```




```
## The least squares regression equation is:
##  Sony2  = 45.836 + 0.348 Sony1
## R^2 = 9.27%
```

```
splot(y=Sony2, x=Sony1, add.line=1)
```



the slope of the line (0.348) is between 0 and 1, so yes, there is an indication of regression to the mean.

30.1.3 Problem 3:

Consider the men's long jump in the Olympics (**longjump**). How strong is the relationship between Year and LongJump?

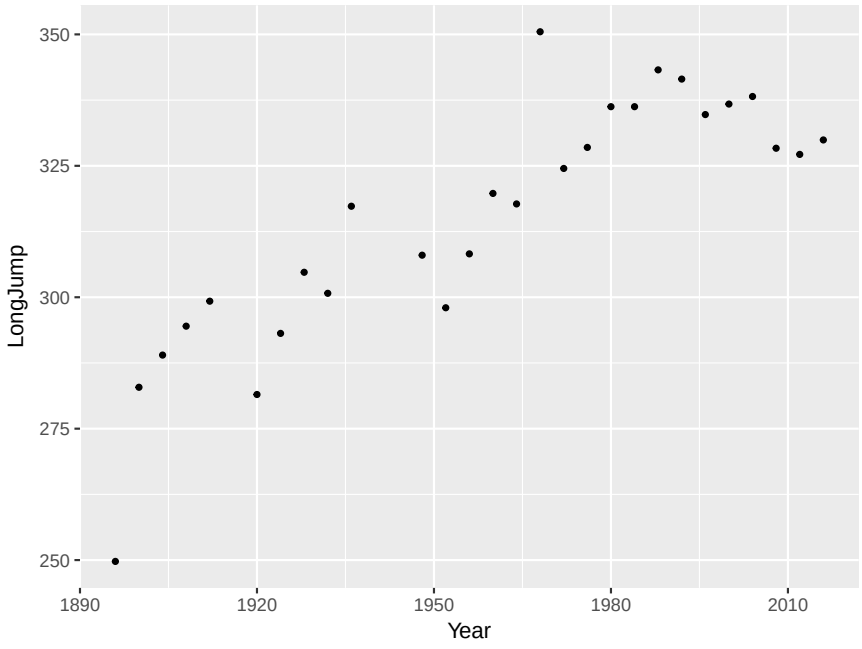
Parameter: correlation coeffcient

Problem: find correlation

Method: ????

the scatterplot of LongJump by Year shows a non-linear relationship, so we can't answer this question (want to know? come to ESMA3102!)

```
attach(longjump)
splot(LongJump, Year)
```



30.1.4 Problem 4:

Consider the following data set:

```
kable(p4data)
```

x	y
10	58
11	54
12	51
13	52
14	62
15	57
16	63
17	64
18	69
19	71
20	70

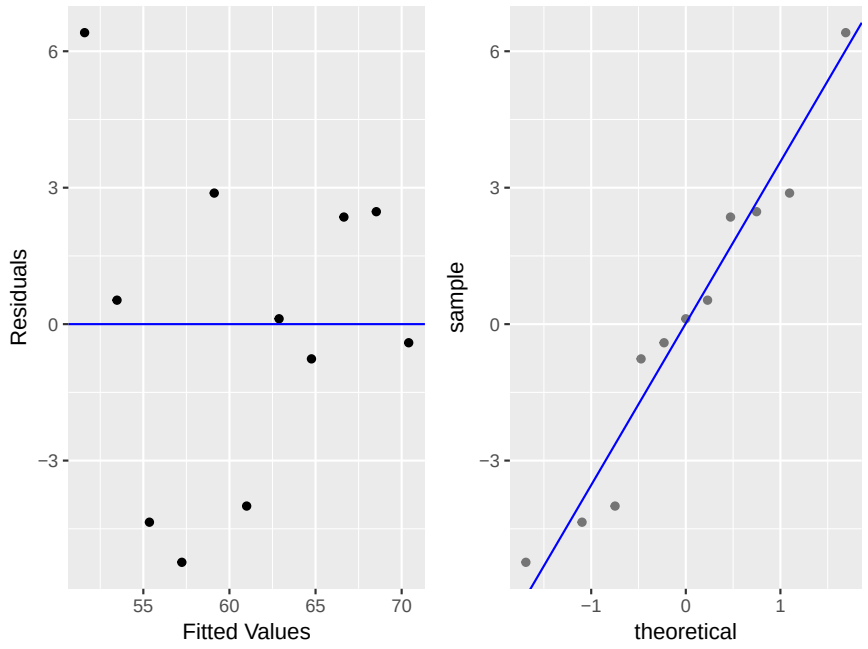
Find the least squares regression equation and use it to predict the y value for an observation with x=15

Parameter: regression coefficients

Problem: find model

Method: slr

```
slr(y=y, x=x)
```




```
## The least squares regression equation is:
##  y  = 32.773 + 1.882 x
## R^2 = 75.79%
```

```
32.773 + 1.882*15
```

```
## [1] 61.003
```

so y=61 is the prediction.