# Statistics 1A Practicum Series

Academic Year 2018-2019

dr. M.E. Timmerman
dr. E.M.L.A. van Krimpen-Stoop
prof. dr. H.A.L. Kiers

Adapted from Dutch
and additional material added by
dr. D. van Ravenzwaaij
dr. R. D. Morey and
dr. I.J.L. Egberink

Updated September 4, 2018

university of
groningen

# Contents

# Chapter 0

# Preliminary Material

## 0.1 The importance of statistics

It is no secret that many students of psychology approach statistics nervously. They may not be comfortable with math, and may see statistics as a hurdle to be overcome in their psychological training. This viewpoint is unfortunate, and as you will see in this course, incorrect. As you begin the sequence of statistics courses in this Psychology program, you should understand clearly the importance of statistics in modern psychology.

Psychology is an empirical science. If a psychological researcher wants to know about something, they have to *observe* it, either directly or indirectly. In astronomy, if you want to know about the movement of planets, you can observe the planets. In chemistry, if you want to know what happens when two chemicals react, you can combine the two chemicals. In psychology, we are often interested in specific populations of people. However, it is nearly impossible to directly observe entire populations.

For instance, if I am interested in the effect of a drug on performance on a test, I know that test scores are likely to vary for all kinds of reasons. People vary in ability, tests vary in difficulty, and so on. Let's say I give two people the test, and one person is given the drug and the other a placebo. The person given the drug scores 75%, and the person given the placebo scores 70%. I might then argue that the drug works to increase test scores. But we aren't interested in only the observed test takers. We are interested in what would happen, on average, for everyone.

In my example, it is not in dispute that the person given the drug scored higher than the person given the placebo. The argument that the drug must work, however, should strike you as weak. You might ask, "what would happen if we chose two *different* people?" or "what would happen if we repeated the test?" These are good questions, and they are based on the idea that test scores are samples from populations. Many times they vary for reasons that are not important. If I am interested in the effect of a drug, I do not care that individual people have different abilities, or that the person who received the placebo was having a bad day.

This leads us to an interesting question: if we cannot see the whole population, how can we ever know anything? This *fundamental* philosophical question is important in every science that deals with measurement error; however, in psychological science, the question is even more critical, because people are very different from one another. Statistics provides a principled way to answer the question of how we may "observe" a population, making statistics a central, and indispensable, part of modern psychology. Statistics forms the epistemological foundation for the science of psychology; without statistics, psychology would lose its status as a science.

This is why it is critical that psychological researchers understand statistics. If a researcher makes a scientific claim without understanding the *basis* of that claim, then the claim has no foundation; it may as well have been made through recourse to magic, astrology, or religion. It is therefore critical for you to understand statistics so that you can both perform and evaluate scientific research.

Finally, you should realize that statistics is not a static body of knowledge. Statistical methods are a primary tool by which psychological science is carried out. New areas in psychology, needing new tools, continue to open every year. Clever ways to analyze psychological data are invented all the time, making the research area of quantitative psychology a fertile source of new tools for psychology. If you are interested in quantitative psychological research, please let me know.

## 0.2 Organization of the course

### 0.2.1 Literature

- Agresti: Statistical Methods for the Social Sciences (5th Edition)
- Navarro: Learning Statistics with R (First Edition)
- Reader, electronic document

### 0.2.2 Lectures

Lectures will be held twice a week. In the lectures, I will discuss the topics necessary for you to complete the assignments in this book. Although attending the lectures is not mandatory, there are several reasons why attending is important: first, the lectures contain content that will be on the exam. Second, there are some topics that are not covered in the text, but will be covered in the lectures. Finally, there is no reason not to attend the lectures. They will improve your understanding of the material.

Prepare for each lecture by reading the materials that are suggested in the reading box at the beginning of each chapter in this reader. If you read before the lectures, you will understand the material from the lectures better.

### 0.2.3 Practicals

Aside from the two lectures per week, you are required to attend the practicals every week (seven in total). In the practicals, you will work on applying what you learned in the text and lectures by working on the exercises in this electronic reader.

There are several important components to each week's practical.

- *Reading*
  At the beginning of each chapter of this document is a box containing the reading assignment for that week. This reading is required to understand the topic(s) of the practical. All of the readings come from the Agresti book and onlinestatbook.com, the required texts for the course.

- *Practice problems*
  At the back of each chapter are exercises. The publisher of the book is presently working to put worked out answers to the exercises online, but they are not available yet at the time of this writing. Another resource for practice problems are those at onlinestatbook.com. Practice often, it is the best way to prepare for the exam!

- *Homework*
  The first section in each chapter of this document is called "Homework Exercises". These exercises are required, and you should turn them in on Student Portal as a word or pdf file. The file should be named "[s-number] - week [weeknumber].pdf". The deadline is every week on Wednesday, 3PM **Note that this includes the week of the first lecture!** Homework will be graded pass/fail, depending on whether you attempted/completed an ample amount of the assigned material, showed your work, and explained your answers. This means that empty boxes are not allowed and the homework will then be graded 'fail'. The answers should also be legibly written and clearly formulated. Otherwise, it might be graded 'fail'. The correct answers to these homework exercises will be discussed during the second lecture in that week. Since the homework needs to be turned in, it is useful to make a copy of the homework and to take it with you to the lectures.

- *In-class exercises*
  Besides the homework exercises, each chapter also contains “in-class exercises”. You will complete these during the practical. Because these often involve using SPSS (or, in the last two weeks, R), computers will be provided for each class. It is required that you are present at all practicals and that you are actively working on the in-class exercises. If you are not working actively during the practical, your practical instructor might decide to grade your in-class exercises as “fail”.

All seven practicals are mandatory and you cannot miss one practical. When you are not able to attend a practical or when you have failed one, you must make up for that practical. There is a special make-up session for all students, which will be on Monday from 3PM to 5PM (check rooster.rug.nl for the room). You must enroll in the first make-up group following your absence. Enrollment will be via Student Portal. If you have not enrolled, you are not on the list and the assistant will send you away and your attendance will not be registered.

You may enroll for the make-up session once. If you are ill for a longer period or for a second time, you must contact the practical coordinator Karin Siebenga at k.siebenga@rug.nl.

### 0.2.4 Exams

There will be one exam for the course. It will be two hours long and will consist of multiple-choice questions. The same counts for the resit exam. Please check the example exam on Student Portal to see what the exams are like.

#### Allowed Calculator

Scientific calculators (and also simple calculators) are allowed during the exam, but *no graphical/graphing calculators*. Any calculator or anything else you can store text on, such as a cell phone, is not allowed.

#### Formula sheet and tables

During the exam, you need to know the formulas by heart, there will not be a formula sheet. $Z$-score and binomial Tables will be provided.

#### Test/exam review

The exam and resit exam of Statistics 1a will be discussed during test review meetings. During such a meeting the exam questions will be shown, together with the correct answers and sometimes a short explanation. Check Student Portal for the exact date, time and room of these test reviews.

You need to register for the test reviews via Student Portal. Recall from the regulations with regard to examinations (which you can find in the course catalogue) that the test review “is limited to only those students who actually took part in the exam”. Therefore, registration via Student Portal is mandatory.

### 0.2.5 Your final grade

In order to pass the course you must pass all components of the practicals. This includes:

- homework exercises graded “pass”, you are allowed to have one “fail” (you do not have to make up for it)
- attending all practicals (i.e., it is not allowed to miss one practical)
- actively working on all in-class exercises (i.e., it is not allowed to have a “fail” for the in-class exercises)

If you have passed all the practical components, then your grade for Statistics 1a is based solely on your performance on the exam. Passing all the practical components is only valid this academic year and the following academic year.

### 0.2.6 Getting the most out of the practicals

The most important part of learning about statistics is building intuition. This may seem counterintuitive; after all, statistics is a branch of mathematics, and mathematics seems to be about formalisms, the opposite of intuition. But intuition plays a large part in learning statistics; so large, in fact, that if you do not foster intuition you will find this class very difficult.

How can you foster statistical intuition? First, make sure you do the readings and exercises assigned. The readings and exercises are assigned to help you learn and build intuition. Second, take the practice exams on `www.socrative.com` to test your knowledge and statistical skills mid-way or at the end of the lecture series. Third, you can foster intuition by communicating with your peers. Often when people learn something, they will use a metaphor. Different people will use different metaphors, and you may find that a peer has a metaphor that works for you. For instance, I may say that sampling a person's opinion (yes/no) about an issue is like flipping a coin. Someone else may think about it as drawing colored marbles out of a bag. Talk with your peers, and ask how they think about a statistical concept. One way to do this is through the Discussion Forum on Student Portal. I would like to encourage you to use this forum when you have questions regarding the content of the book and to encourage you to help each other by answering each other's questions. I will monitor the questions and provided answers, but will not answer the questions myself (only when certain answers are incorrect). In this way, you will learn from each other.

If you still have trouble understanding a concept, ask your practical instructor, ask me before/during/after class or send me an email (see Contact information).

### 0.2.7 Contact information

- If you have questions about the content of the practical, ask your practical instructor in class.

- If you have a question and/or remark regarding the organization of the course, you can send me, dr. Don van Ravenzwaaij, an email at d.van.ravenzwaaij@rug.nl.

- If you have a question regarding the content of the course, ask me before/during/after class. If you prefer, you can send me an email with your question as well. However, in that case I ask that you provide me with a specific question together with your own attempt to handle the problem. This will help you to get insight into your difficulties and I can see what part of the problem creates the difficulties so that I can best help you towards the solution.

- If you have questions about the organization of the practicals, contact the practical coordinator Karin Siebenga through email (k.siebenga@rug.nl); use Statistics 1a as subject; working days: Mon-Thu).

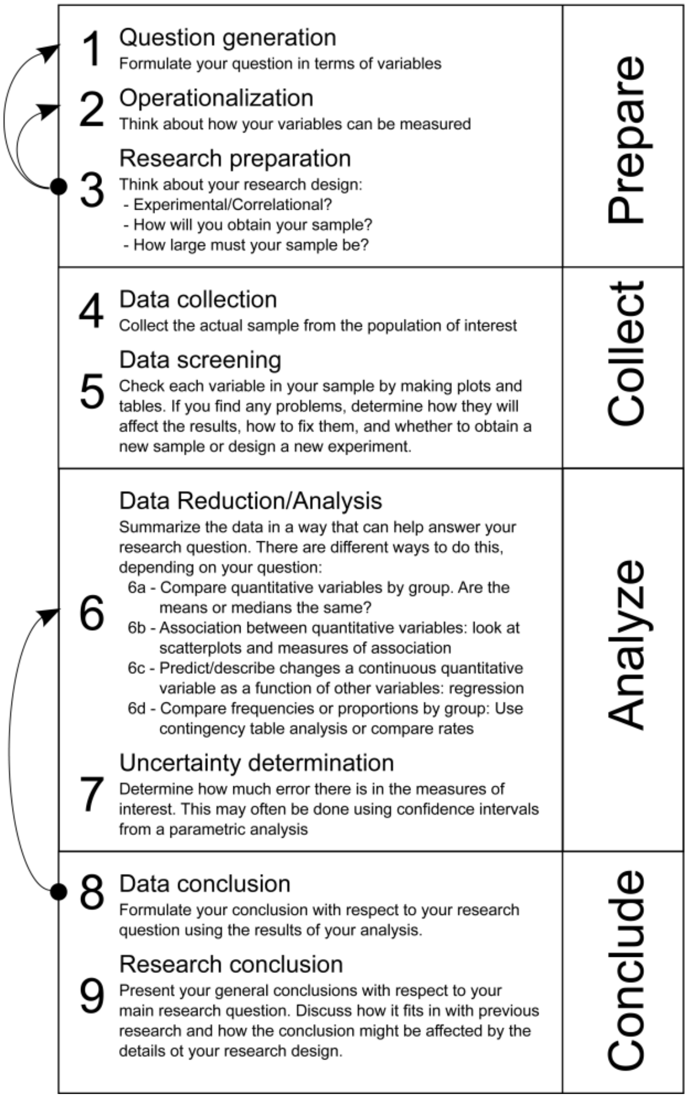
## 0.3 The research cycle

One of the things that separates the sciences, including psychology, from other methods of inquiry is that the sciences are systematic. Being systematic about research is one way scientists increase their objectivity: if researchers approach all research topics systematically, it is more difficult to inject subjectivity into research.

The research process in psychology has a general form which we will call the research cycle, shown in Figure 1. Of course, this exact approach is not set in stone. Instead, it is a general guideline that will help you understand how to approach research in a systematic manner.

Every chapter in this reader starts with the research cycle. Because every chapter in this reader covers different topics, the research cycle will have certain topics highlighted to let you know how the topics of the week are related to the research path.

*Don van Ravenzwaaij*



1 Question generation
Formulate your question in terms of variables

2 Operationalization
Think about how your variables can be measured

3 Research preparation
Think about your research design:
- Experimental/Correlational?
- How will you obtain your sample?
- How large must your sample be?

Prepare

4 Data collection
Collect the actual sample from the population of interest

5 Data screening
Check each variable in your sample by making plots and tables. If you find any problems, determine how they will affect the results, how to fix them, and whether to obtain a new sample or design a new experiment.

Collect

6 Data Reduction/Analysis
Summarize the data in a way that can help answer your research question. There are different ways to do this, depending on your question:
6a - Compare quantitative variables by group. Are the means or medians the same?
6b - Association between quantitative variables: look at scatterplots and measures of association
6c - Predict/describe changes a continuous quantitative variable as a function of other variables: regression
6d - Compare frequencies or proportions by group: Use contingency table analysis or compare rates

7 Uncertainty determination
Determine how much error there is in the measures of interest. This may often be done using confidence intervals from a parametric analysis

Analyze

8 Data conclusion
Formulate your conclusion with respect to your research question using the results of your analysis.

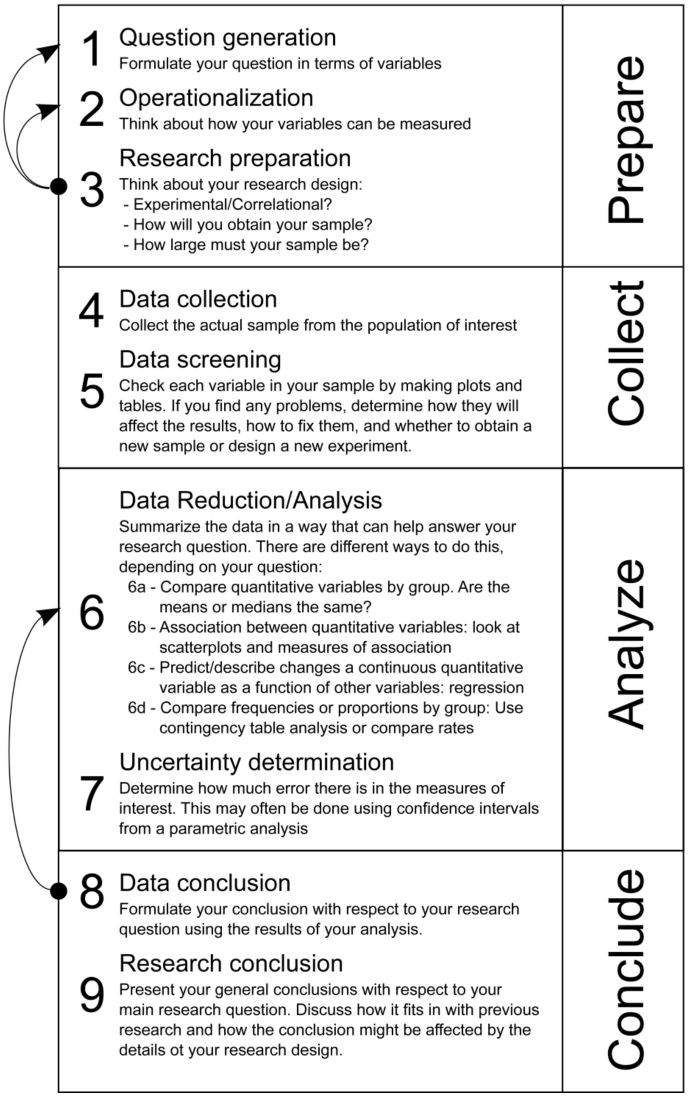
9 Research conclusion
Present your general conclusions with respect to your main research question. Discuss how it fits in with previous research and how the conclusion might be affected by the details ot your research design.

Conclude

Figure 1: The research path.

# Chapter 1

# Week 1: Descriptive Statistics



Prepare

1 Question generation
Formulate your question in terms of variables

2 Operationalization
Think about how your variables can be measured

3 Research preparation
Think about your research design:
- Experimental/Correlational?
- How will you obtain your sample?
- How large must your sample be?

Collect

4 Data collection
Collect the actual sample from the population of interest

5 Data screening
Check each variable in your sample by making plots and tables. If you find any problems, determine how they will affect the results, how to fix them, and whether to obtain a new sample or design a new experiment.

Analyze

6 Data Reduction/Analysis
Summarize the data in a way that can help answer your research question. There are different ways to do this, depending on your question:
6a - Compare quantitative variables by group. Are the means or medians the same?
6b - Association between quantitative variables: look at scatterplots and measures of association
6c - Predict/describe changes a continuous quantitative variable as a function of other variables: regression
6d - Compare frequencies or proportions by group: Use contingency table analysis or compare rates

7 Uncertainty determination
Determine how much error there is in the measures of interest. This may often be done using confidence intervals from a parametric analysis

Conclude

8 Data conclusion
Formulate your conclusion with respect to your research question using the results of your analysis.

9 Research conclusion
Present your general conclusions with respect to your main research question. Discuss how it fits in with previous research and how the conclusion might be affected by the details ot your research design.

Figure 1.1: Research topics for week 1

| Agresti | onlinestatbook.com |
|---|---|
| Agr: 1.1-1.4, 2.1, 3.1-3.4 | OSB: 1-3 |
| Practice exercises are in the back of each chapter in the book. | |

## 1.1 Homework exercises

*Give a full answer to each question, justifying your answer and showing work, if necessary.*

1. The number of deaths due to cancer has grown over the years in the Netherlands. In 1970, it was reported that 25 217 people died of cancer. In 2002, the number of deaths reported was 37 975. A politician uses these figures to claim that no progress has been made in treating cancer; in fact, the politician claims that treatment has gotten worse. Explain how it is possible that the number of deaths increases even though treatment improves. Describe a variable which is a better summary measure than the number of deaths for examining the effectiveness of treatment of cancer in the Netherlands.

2. There are indications that an increase in the quantity of calcium in one's diet can help reduce blood pressure. In an experiment, additional calcium was added to the diet of an experimental group, while a placebo was added to a control group's diet. Each participant's systolic blood pressure while resting was measured before the experiment, and again 12 weeks into the experiment[1]. Thus, each group's blood pressure after the experiment can be compared with the group's blood pressure before the experiment 1 . The measurements for each participant at the beginning of the experiment are given below:

| Calcium group | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 107 | 110 | 123 | 129 | 112 | 111 | 107 | 112 | 136 | 102 | |
| Placebo group | | | | | | | | | | |
| 123 | 109 | 112 | 102 | 98 | 114 | 119 | 112 | 110 | 117 | 130 |

   (a) Why is it important that scores in the experimental (calcium) and control (placebo) groups be similar?

[1] This is called a within-subjects design, because each participant's score after is compared with their score before.

(b) Make a back-to-back stemplot plot of the data.

(c) Does this plot show any important differences between the two groups before the experiment began?

(d) Are the centers of the two groups near one another?

3. A student wants to know if the students in his study year are about the same age. To answer this question, he asks six different students their age. The data is shown in the following table:

| Student # | Age |
|---|---|
| 1 | 18 |
| 2 | 21 |
| 3 | 19 |
| 4 | 45 |
| 5 | 24 |
| 6 | 20 |

(a) Compute the mean, the median, and the standard deviation of the age of the six students.

(b) There is an outlier in the data. Compute the mean, median, and standard deviation without the outlier.

(c) What can you conclude about the difference between the means, medians, and standard deviations in problems 3a and 3b?

(d) The student wondered if "the students in his study year are about the same age." What summary statistic is appropriate for answering this question?

4. The Questionnaire of Study Habits and Attitudes (QSHA) is a psychological test designed to measure student motivation, study habits, and attitudes. A University gives the QSHA to a sample of 18 female first-year students. Their scores are given below:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 154 | 109 | 137 | 115 | 152 | 140 | 154 | 178 | 101 |
| 103 | 126 | 126 | 137 | 165 | 165 | 129 | 200 | 148 |

(a) Make a histogram of these data

(b) Can you identify any outliers?

(c) Compute the mean and median for these data.

(d) Discuss the relationship between the mean and the median with respect to the general distribution of the test scores shown in problem 4a.

The University now samples 20 male first-year students. Their scores are below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 108 | 140 | 114 | 91 | 180 | 115 | 126 | 92 | 169 | 146 |
| 104 | 109 | 132 | 75 | 88 | 113 | 151 | 70 | 115 | 187 |

(e) Compute the five-number summary for both the males' data and the females' data.

(f) Following the 1.5 × $IQR$ rule, are there outliers?

(g) Make boxplots of the two score distributions. Use modified boxplots whereby the outliers are shown and marked clearly.

## In-Class Exercises (week1)

## 1.2 Using SPSS

### Goals for this section

- Learning the basic operation of SPSS
- Learning how to make a few clear, simple plots using SPSS
- Drawing the right conclusions from plots
- Choosing the correct summary statistics to characterize the central tendency and spread of data
- Drawing the right conclusions from summary statistics

### 1.2.1 A general introduction to using SPSS

SPSS (Statistical Package for the Social Sciences) is a fully-featured statistical software package. SPSS is designed to make statistical calculations easy. In this section, you will learn to do a number of basic things in SPSS, such as opening a data file. Mastering these techniques, and being comfortable with SPSS, will make it easier for you to learn more complex SPSS skills later on.

We will first address the differences between Microsoft Excel, with which you may be somewhat familiar, and SPSS. Excel was designed explicitly to manipulate data. In Excel, it is very easy to apply a simple function to data, such as the mean or standard deviation. To do this in SPSS is a bit more involved. However, SPSS has the distinct advantage over Excel that a large number of advanced statistical techniques are available. In Excel, many of these advanced techniques would be difficult, if not impossible, to use. It is possible to use Excel and SPSS side-by-side, using Excel for data quality control and then SPSS for statistical analyses. As you learn SPSS, you will decide what approach works best for you.

Although the usage of SPSS is not explicitly tested in this class, the interpretation of the output from an SPSS analysis is tested. As such, try to get insight into the structure and function of SPSS. This will make the following practica much easier, doing exercises will take less time, and your SPSS skills will be invaluable in your future studies.

If there is something about the use of SPSS you do not understand, first try looking it up in the SPSS reference book or in the SPSS help (Help → Topics). Or try asking another student, before asking your practical instructor.

### 1.2.2 The two views in SPSS

Start SPSS now. There are multiple ways to do this:

- Double-click on the SPSS icon on your desktop.
- Click on the Windows Start menu, then under programs find SPSS. Click on “SPSS for Windows”.
- If neither of first two options are available, click on the NAL icon on your desktop and then select “SPSS for Windows” under “Mathematics & Statistics”. Then double-click on SPSS.

.

**Data View**

After a brief information window to let you know what version of SPSS you are using, SPSS will open. You may get a window asking “What would you like to do?” Click “Cancel” to close this window. Your SPSS window should now look like Figure 1.2. This window is called the SPSS data editor.

From the data editor, it is possible to view and edit data. It looks like a spreadsheet. Like a spreadsheet, each column represents a variable, and each row represents an observation.

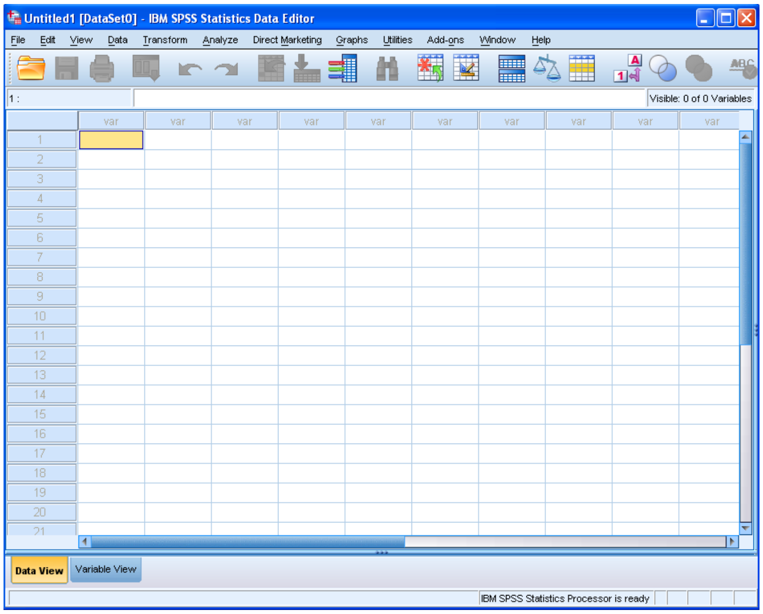


Figure 1.2: SPSS data editor with no data loaded.

Note several things about the data editor window:

- The title bar of the window says "Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor". "SPSS Data Editor" indicates that the window is the data editor, and "Untitled1 [DataSet0]" indicates that we have not loaded any data yet, so SPSS chose a default name for our data.
- The menu bar has some typical options (File, Edit, View,. . . ,Help) and some options unique to SPSS (Data, Transform, Analyze,. . . ).
- The tool bar under the menu bar contains useful buttons to make some tasks faster. Move your mouse over the buttons for a description of each one.
- The status bar at the bottom of the window tells you what SPSS is doing. Currently, it should say "SPSS Processor is ready."

At the bottom of the window, on the left side, notice the two tabs labeled "Data View" and "Variable View". In data view you can examine your data, like in a spreadsheet. In Variable view, you can define variables, rename variables, and change their properties. We will discuss the Variable view in more detail later. In order to use SPSS to analyze your data, you must first load the data into SPSS. Typically your data will be stored in one of three ways:

- in an SPSS file.
- in another type of file, such as an Excel file or a text file
- in a hard copy, on paper

We will deal with the first situation in this course. In Statistics 2, you will learn how to deal with the later two situations.

**How to open an SPSS file**

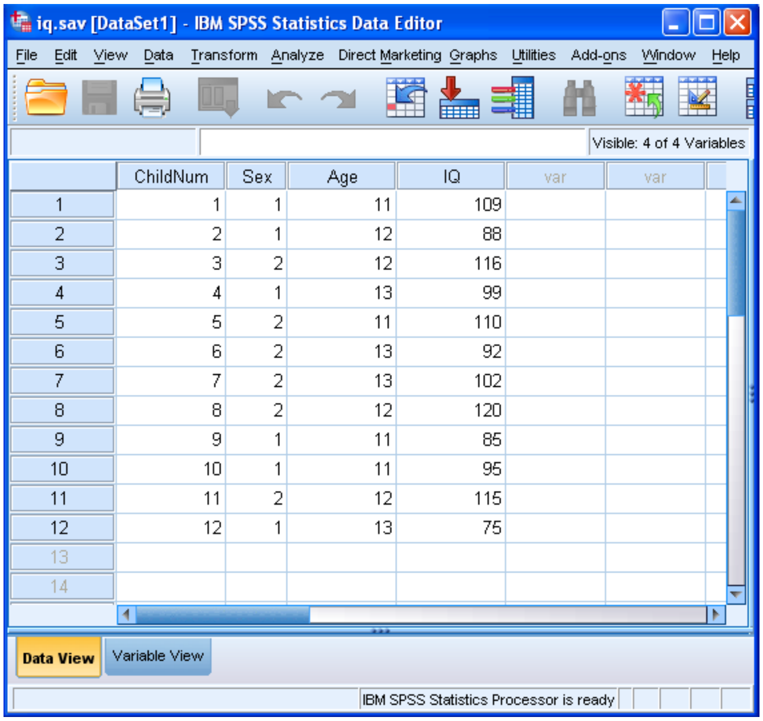
SPSS files always end in the extension .sav. In the dataset "1 - IQ.sav", you will find (hypothetical) IQ scores from 12 children, along with their sex and age. There are multiple ways to open a datafile:

- Double-click on the file.

- Drag the file onto your open SPSS data editor window.

- Use the File → Open → Data. . . menu and select the file you want to open.

Open the "1 - IQ.sav" dataset now. When you have opened the file properly, your data editor should look like Figure 1.3. The rows in SPSS always represent an individual case, or observation. The columns each represent a variable pertaining to the case.


iq.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Visible: 4 of 4 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

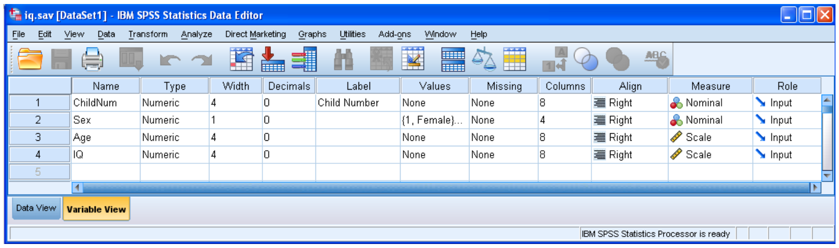
| | ChildNum | Sex | Age | IQ | var | var |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 11 | 109 | | |
| 2 | 2 | 1 | 12 | 88 | | |
| 3 | 3 | 2 | 12 | 116 | | |
| 4 | 4 | 1 | 13 | 99 | | |
| 5 | 5 | 2 | 11 | 110 | | |
| 6 | 6 | 2 | 13 | 92 | | |
| 7 | 7 | 2 | 13 | 102 | | |
| 8 | 8 | 2 | 12 | 120 | | |
| 9 | 9 | 1 | 11 | 85 | | |
| 10 | 10 | 1 | 11 | 95 | | |
| 11 | 11 | 2 | 12 | 115 | | |
| 12 | 12 | 1 | 13 | 75 | | |
| 13 | | | | | | |
| 14 | | | | | | |

Figure 1.3: SPSS with the "1 - IQ.sav" dataset loaded, in *Data view*.

1. Look at the data. How many children have an age of 12 years? What is the lowest IQ score measured for these 12 children?

**Variable View**

In the column for the variable sex, there are two scores: 1 and 2. This is a coding of the sex for each case. The code 1 stands for female, and the code 2 stands for male. It is often convenient to use codings like this in data. In order to examine the coding in use in a data set, we need to use the SPSS's "Variable view". Click on "Variable view" on the bottom of the screen. Your SPSS window should look like Figure 1.4.


iq.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

| | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ChildNum | Numeric | 4 | 0 | Child Number | None | None | 8 | Right | Nominal | Input |
| 2 | Sex | Numeric | 1 | 0 | | {1, Female}... | None | 4 | Right | Nominal | Input |
| 3 | Age | Numeric | 4 | 0 | | None | None | 8 | Right | Scale | Input |
| 4 | IQ | Numeric | 4 | 0 | | None | None | 8 | Right | Scale | Input |
| 5 | | | | | | | | | | | |

Figure 1.4: SPSS with the "1 - IQ.sav" dataset loaded, in *Variable view*.

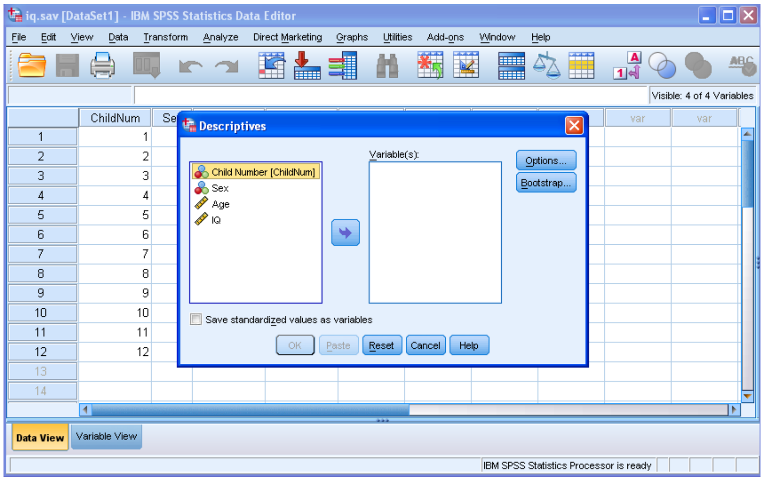
In variable view, you can look at the properties of each variable. Each row corresponds to a variable, and each column represents a property. For example, look at the row corresponding to the variable "Gender", and then look under the "Values" column. If you click on this cell, then click on the three dots in the right side of the cell, you will be able to examine the coding used for this variable. You can also add new codings for data sets whose codings are not explicit.

Also, notice the "Label" column. Here, you can add descriptions to your variables. For this simple data set it is unnecessary, but in more complicated data sets this feature is very useful.

On the far right side of the variable view, there is a column named "Measure". This column determines what type of variable the row represents. For example, the measure type of Gender is "Nominal", indicating that the actual numbers are just names, or codes, and should not be interpreted as numbers per se. It is meaningless to say, for instance, that 1 (Female) is less than 2 (Male). However, Age is a "scale" variable, indicating that the numbers have meaning. It is perfectly reasonable to say that 12 (years) is greater than 11 (years) for the Age variable. Check the other variables to see if their measure setting makes sense.

### 1.2.3 A first calculation in SPSS

Previously, we have only looked at data. Now we will perform a simple data analysis: computing summary statistics of the IQ scores. Examples of summary scores include the mean, median, minimum and maximum. Click back to the data view. From the menu, click Analyze → Descriptive Statistics → Descriptives. Your SPSS window should look like Figure 1.5. From this screen, you can tell SPSS which variables you would like descriptive statistics for. We will compute descriptive statistics for the IQ variable, so select IQ in the left window, then click the symbol between the two windows to add IQ to the list of variables we want.


iq.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 4 of 4 Variables
Se
var
var
Descriptives
Variable(s):
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

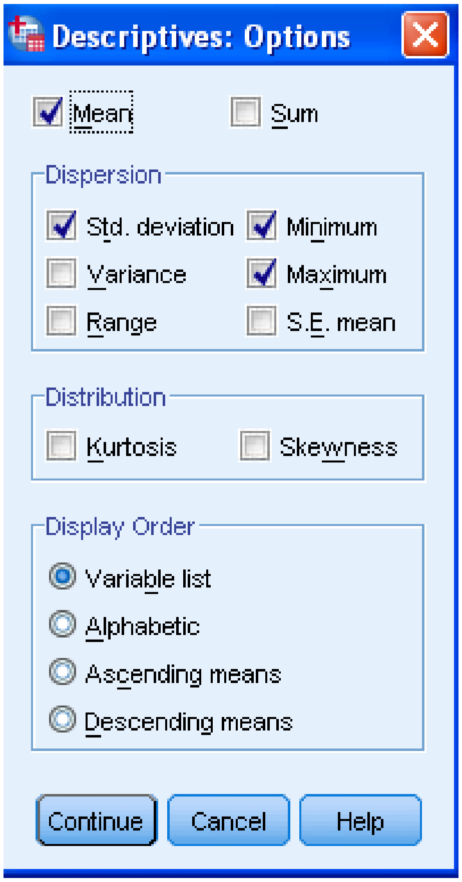
| | ChildNum |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 6 |
| 7 | 7 |
| 8 | 8 |
| 9 | 9 |
| 10 | 10 |
| 11 | 11 |
| 12 | 12 |
| 13 | |
| 14 | |

Child Number [ChildNum]
Sex
Age
IQ

□ Save standardized values as variables

Figure 1.5: SPSS with the “1 - IQ.sav” dataset loaded, after clicking Analyze → Descriptive Statistics → Descriptives.

Before we tell SPSS to compute our descriptive statistics, we may want to modify what SPSS computes for us. Click on the “Options…” button in the lower right of the Descriptives window. Your window should look like Figure 1.6. From this window, we can control what descriptive statistics SPSS will output. This menu is not very useful for us right now, but nearly every analysis has a similar Options menu; the options will become very useful later. Click “Continue” to return to the Descriptives window, and then click “OK” to compute the descriptive statistics.



☑ Mean ☐ Sum

Dispersion

☑ Std. deviation ☑ Minimum
☐ Variance ☑ Maximum
☐ Range ☐ S.E. mean

Distribution

☐ Kurtosis ☐ Skewness

Display Order

● Variable list
○ Alphabetic
○ Ascending means
○ Descending means

Figure 1.6: The SPSS Descriptive Statistics options menu.

SPSS will now compute your descriptive statistics. The results will be opened in a new window, which SPSS calls the “output viewer”. The output viewer will contain the output of all your analyses in an easy-to-read format. The output from SPSS may be saved or printed for later use. Use the table output by SPSS to answer the following questions.

2. What are the minimum, maximum, and the mean IQ scores?

3. In addition to summary statistics, it is also useful to know the frequencies of certain variables in our data. For instance, we may wish to know how many males and how many females there are in our sample. Click on your data editor window on the Windows task bar, then click Analyze → Descriptive Statistics → Frequencies from the SPSS menu. Add the variable Gender to the Variable window, and click "OK".

   Notice that SPSS adds the output to your current Output window.

   How many females are there in the sample? How many males? What percentages of the sample do each represent?

You have now performed a number of simple tasks in SPSS. You should feel comfortable with the interface while you are learning SPSS; feel free to open menus you have never opened and to click option boxes you have not used before. If you do not understand the output, do not worry - we will probably learn about it later. The most important thing to realize is that you cannot hurt anything by experimenting. Experimenting with the options in SPSS is the best way to learn.

### 1.2.4 Saving output and data in SPSS

It is often useful to save what you have done in SPSS. The first thing you might want to save is the output of an analysis. To to this, click back to the Output window. There are two ways you might want to save data:

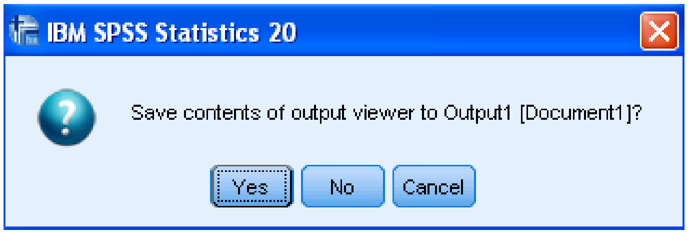
- In a format that can only be read by SPSS. If you are saving the output for later use by you or colleagues who all use SPSS, it might be useful to exchange only SPSS files.

- In a generic format like PDF or HTML. If you want to exchange results with other researchers who do not use SPSS, it is polite to send them a more generic file type. Also, if you are writing about the output in a report, exporting to a Word or text document might be useful.

- By printing the output. This is useful if you want to make notes on the output by hand.

Printing the output is the same as in every other Windows program, so we will not cover it here. To save the output in SPSS format, select File → Save from the Output window. Select a file name and location, and save the file. SPSS output files are always saved with the .spo extension. In order to save the output in a non-SPSS format, click File → Export. In the menu that opens, choose a File Name (where you want to save the file, and what you want to call it). Then choose an Export Format and click "OK". Your exported file should be saved in the location you selected. Now, close the Output window. If you haven't saved your output, you might get a warning like in Figure 1.7. If you have already saved the output, click "No". If you wish to save the output, click "Yes".

If you have made changes to the data itself, such as renaming variables, adding new variables, or editing the data itself, you might want to save the data. To do this, click back to the data editor and click File → Save or File → Save As... and choose a location and name. Be careful when saving data, however; always make sure that you have a backup of the original data somewhere, in case you make a mistake. As a researcher, your data is the most important resource you have. Treat it with care, and do not overwrite old data without being absolutely sure. Of course, in this class, we keep copies of all the data sets, so if you accidentally lose some data you can just download it again.


IBM SPSS Statistics 20

Save contents of output viewer to Output1 [Document1]?

Figure 1.7: When you close the output screen, SPSS will ask whether you want to save the output.

## 1.3 What is a normal body temperature?

For the past century, a body temperature of 37° has been considered "normal". The data file "1 - Temperature.sav" (based on [Mackowiak et al., 1992]) contains measurements of the body temperatures of many adult men and women.

| *Variable* | *Description* |
|---|---|
| Gender | 1=man, 2=woman |
| Heart | Heart rate in beats per minute |
| Temp | Body temperature in Celsius degrees |

1. Using SPSS, make a histogram for the variable body temperature. Ignore the variable Gender for now, and only make one histogram for body temperature. Describe the histogram.

2. What is the "normal" temperature for these data? What do you mean by "normal"? (There are several acceptable answers to this question.)

3. How much do people vary in their temperatures? How did you measure how much temperatures vary? (Again, there are several acceptable answers to this question.)

4. What can you conclude about the "normal" temperature of 37 ° Celsius from your histogram?

5. Using SPSS, make a histogram and a boxplot showing the body temperature separately for men and women. (Hint: Create a histogram like you did before, click on "Groups/Point ID", then "Rows panel variable", then drag Gender to the "Panel" box. For the boxplot, use Graphs→Legacy Dialogs→Boxplot. Choose "Summaries for groups of cases" and "Simple". Click "Define", and choose Temp as your variable and Gender as your "Category Axis".) Describe the histograms and the boxplots.

6. Do the "normal" temperatures appear to be the same for men and women? Why or why not? (Warning: look carefully at the axes.)

## 1.4 Life expectancy in the world

These data come from two sources: UNESCO's 1990 Demographic Year Book, and Day's Annual Register 1992. The file "1 - Poverty.sav" contains data from 97 countries [Rouncefield, 1995]. Different characteristics of the country are described which may be used as indices of poverty.

| **Variable** | **Description** |
|---|---|
| Births | Number of births per 1000 people |
| Deaths | Deaths per 1000 people |
| Infantd | Infant mortality per 1000 people |
| Malelife | Life expectancy at birth for men |
| Femlife | Life expectancy at birth for women |
| GNP | Gross National Product |
| Group | 1 Eastern Europe |
| | 2 South America and Mexico |
| | 3 Western Europe, North America, Japan, Australia, New Zealand |
| | 4 Middle East |
| | 5 Asia |
| | 6 Africa |
| Country | Country name |

Choose either life expectancy of men or life expectancy of women. Using SPSS, make a plot showing the different life expectancy of the different regions (Group). It is up to you to pick the type of plot.

1. In which region is life expectancy the highest? Where is it the lowest?

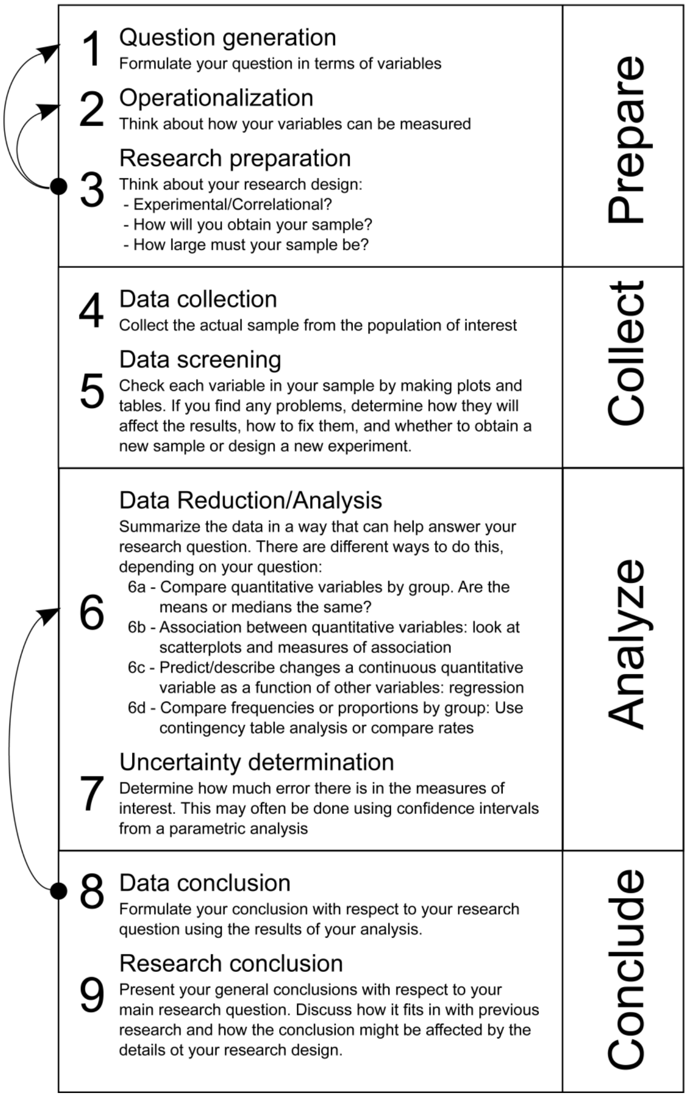
2. In which region is variance the greatest? Can you come up with an explanation for this?

### Important points for this week

- You learned how to open and save SPSS data and output
- You learned how to make a histogram, boxplot, and frequency table in SPSS
- You learned how to compute summary statistics (min, max, mean) in SPSS
- Try to remember how to do each of these things in SPSS. It will make later classes much easier

# Chapter 2

# Week 2: Sampling error



1 Question generation
Formulate your question in terms of variables

2 Operationalization
Think about how your variables can be measured

3 Research preparation
Think about your research design:
- Experimental/Correlational?
- How will you obtain your sample?
- How large must your sample be?

Prepare

4 Data collection
Collect the actual sample from the population of interest

5 Data screening
Check each variable in your sample by making plots and tables. If you find any problems, determine how they will affect the results, how to fix them, and whether to obtain a new sample or design a new experiment.

Collect

6 Data Reduction/Analysis
Summarize the data in a way that can help answer your research question. There are different ways to do this, depending on your question:
6a - Compare quantitative variables by group. Are the means or medians the same?
6b - Association between quantitative variables: look at scatterplots and measures of association
6c - Predict/describe changes a continuous quantitative variable as a function of other variables: regression
6d - Compare frequencies or proportions by group: Use contingency table analysis or compare rates

7 Uncertainty determination
Determine how much error there is in the measures of interest. This may often be done using confidence intervals from a parametric analysis

Analyze

8 Data conclusion
Formulate your conclusion with respect to your research question using the results of your analysis.

9 Research conclusion
Present your general conclusions with respect to your main research question. Discuss how it fits in with previous research and how the conclusion might be affected by the details ot your research design.

Conclude

Figure 2.1: Research topics for week 2

| Agresti | onlinestatbook.com |
|---|---|
| Agr: 3.5-3.7, 4.3 | OSB: 4, 7 |
| Practice exercises are in the back of each chapter in the book. | |

## 2.1 Homework exercises

*Give a full answer to the question, justifying your answer and showing work, if necessary.*

1. A group of 201 Statistics 1A students have taken their first statistics test. The five-number summary for their test scores is the following:

   2.3 5.1 6.2 7.8 9.2

   (a) The student who received the 9.2 found an error in her grade. Her corrected grade is 9.5. Assuming there are no other errors, what is the new five-number summary?

   (b) Create the box plot using this new five-number summary.

2. The distribution of durations of pregnancies, from conception to birth, has (roughly) a normal distribution with a mean of 266 days and a standard deviation of 16 days. Use the 68-95-99.7 rule to answer the following questions.

   (a) How long are the middle 95% of pregnancy durations? Give two numbers, a lower bound and an upper bound.

(b) How short are the shortest 2.5% of pregnancies? How long are the longest 2.5% of pregnancies?

3. The Questionnaire of Study Habits and Attitudes (QSHA) is a psychological test designed to measure student motivation, study habits, and attitudes. Scores on the QSHA for psychology students are normally distributed with a mean of 130 and a standard deviation of 5. In contrast, scores of anthropology students are normally distributed with a mean of 120 and a standard deviation of 10

   (a) Marissa studies psychology and has a score of 140 on the QSHA. Her roommate Anton studies anthropology and has a score of 130 on the QHSA. Compute the z-scores for each student, with respect to their own area of study.

   (b) Find the percentage of psychology students with a score higher than 140.

   (c) Find the percentage of anthropology students with a score higher than 130.

   (d) Which student, Marissa or Anton, has a higher score on the QSHA compared with classmates in their own area of study?

4. In order to make the interpretation and comparison of scores easier, scores are often standardized. A good example is IQ scores. The intent is that the IQ-scores for the entire population, if they were to take an IQ test, would be normal with a mean of 100 and a standard deviation of 15. Of course, it is impossible to give an IQ test to everyone, so samples must suffice. A researcher has devised a new type of IQ test, consisting of 80 questions. Each question is worth 1 point. The researcher gives the test to a representative sample of 423 adults. In this sample, the observed mean score is 50 points and the standard deviation is 10 points. In order to compare scores on the new test with standard IQ scores, the researcher would like to apply a linear transformation so that the new mean is 100 and the new standard deviation is 15.

   (a) Give the general formula for a linear transformation of some variable x.

   (b) How can the researcher apply the general formula to his problem? In particular, what must he multiply and what must he add to the scores of his sample to yield a mean of 100 and a standard deviation of 15?

## In-Class Exercises (week 2)

### Goals for this section

- Gain insight into the amount of uncertainty in the sampling of summary statistics
- Become familiar with density curves
- Gain insight into the relationship between the normal distribution and the standard normal distribution
- Learn about samples from normal distributions
- Application of normal distributions to psychological variables
- ...Practice finding proportions from normal distributions
- ...Practice finding intervals which correspond to proportions from normal scores

### 2.1.1 Graphically representing uncertainty

The following table gives eighteen scores from a hypothetical sample of children on an IQ test. It is known that IQ scores in the population are distributed normally with a mean of 100 and a standard deviation of 15.

| | | | | | |
|---|---|---|---|---|---|
| 93.0 | 85.1 | 115.5 | 124.0 | 105.0 | 100.3 |
| 101.0 | 107.0 | 113.0 | 129.0 | 114.0 | 97.8 |
| 101.8 | 112.0 | 91.0 | 91.0 | 90.0 | 113.0 |

1. What is the minimum score in the sample?

2. What proportion of the scores in the population are less than the minimum score in the sample? Hint: find the $z$-transform of the minimum score

3. Is the minimum value observed in the sample an unlikely value? Why or why not?.

4. Between what values will the middle 50% of IQ scores in the population fall? (Hint: determine the appropriate tail-probabilities, and determine the $z$-scores corresponding to those probabilities. Then, transform them to IQ scores using a linear transformation).

5. What proportion of your sample falls outside of the area where the middle 50% of scores fall in the population?

6. Do you expect the sample properties to exactly match the population properties? Why or why not?

7. Suppose that a therapist claims that Cognitive Behavioral Therapy (CBT) is not appropriate for children with very low IQ. The therapist will not consider CBT for any child in the lowest 3% of population IQs. What IQ scores ensure that a child will not receive CBT?

## 2.2 Tufte's principles

Edward Tufte [Tufte, 2001] is a researcher famous for describing principles to follow when creating graphical displays of data. He emphasizes how important it is for the graphic to clearly communicate to the reader the intended information. For this exercise, please look at the three plots presented below and do the following:

- Describe the conclusion you draw from the plot.

- Describe any difficulties you had in interpreting the plot.

- Describe how you would make the plot better. Would you use the same plot type? Would you use the same information?

- Sketch a plot that you think would be better than the plot shown. What conclusions can you draw from your plot?

The last 15 minutes of this class, your tutor will discuss this exercise with the entire group.

1. Figure 1 shows life expectancy in the United States as a function of time, income, and sex.

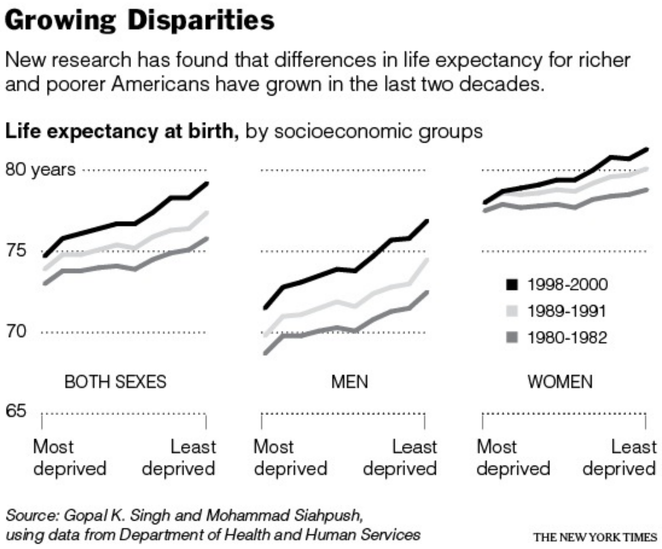


Figure 2.2: Source:`http://www.nytimes.com/2008/03/23/us/23health.html`

The plot is from the New York Times.

2. Figure 2.3 shows the official results of the 2007 election alongside exit poll results.

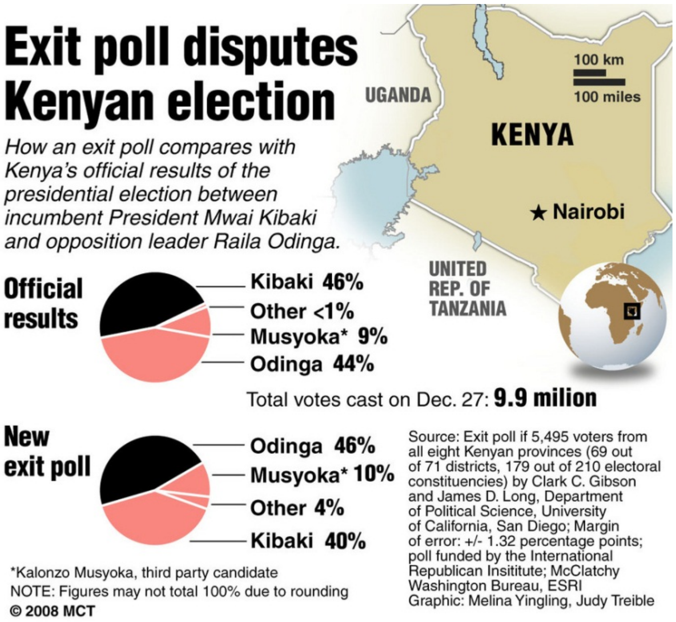


Figure 2.3: Source: `http://www.mcclatchydc.com/homepage/story/43442.html`

The plot is from McClatchy's Washington Bureau, a well-known publication in the United States.

3. Figure 2.4 shows a (not serious) plot of the difficulty of eating different fruit against the tastiness of each fruit.

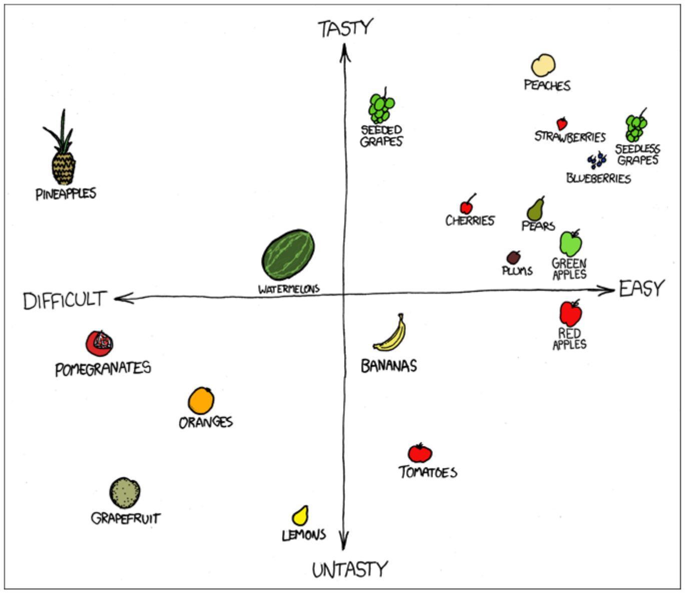


Figure 2.4: Source: `http://xkcd.com/388/`. Use permitted under the Creative Commons license.

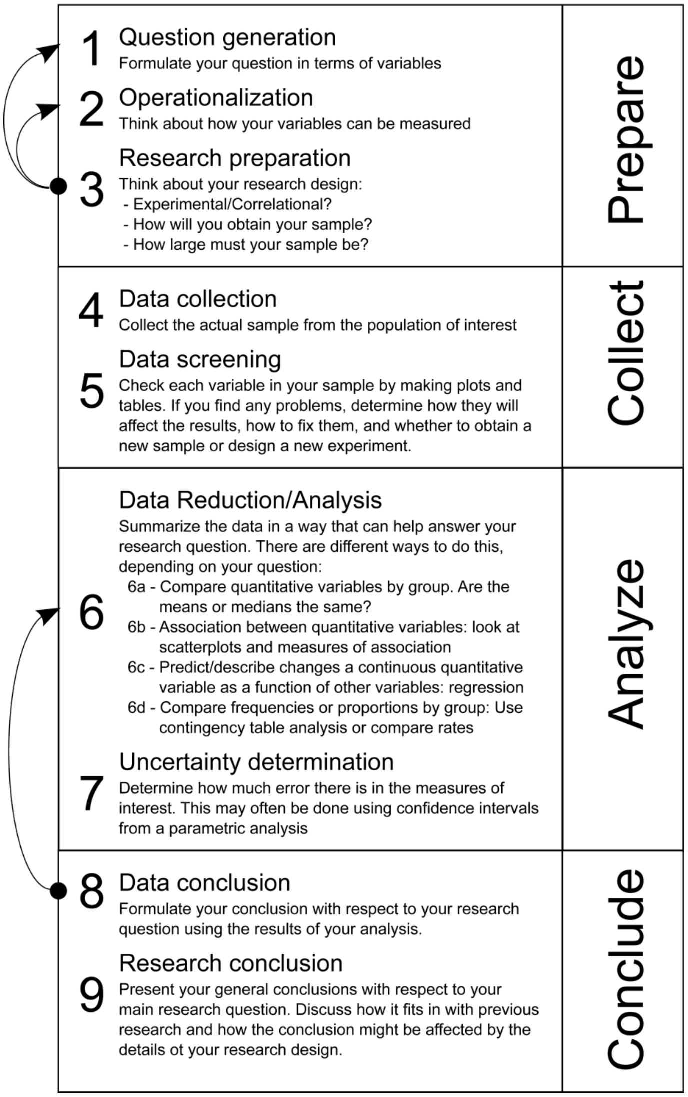
The plot is from the web comic xkcd

### Insights for this week

- The normal distribution is a good model of the population of many psychological variables of interest.
- It is possible that a sample of scores taken from a normal distribution does not *look* normal. However …
- …the larger the population, the more like the population the sample will be.
- With the help of a model of the population (i.e. the normal distribution), one can determine the proportion of scores in a range of values.

# Chapter 3

# Week 3: Correlation and Regression



Prepare

1 Question generation
Formulate your question in terms of variables

2 Operationalization
Think about how your variables can be measured

3 Research preparation
Think about your research design:
- Experimental/Correlational?
- How will you obtain your sample?
- How large must your sample be?

Collect

4 Data collection
Collect the actual sample from the population of interest

5 Data screening
Check each variable in your sample by making plots and tables. If you find any problems, determine how they will affect the results, how to fix them, and whether to obtain a new sample or design a new experiment.

Analyze

6 Data Reduction/Analysis
Summarize the data in a way that can help answer your research question. There are different ways to do this, depending on your question:
- 6a - Compare quantitative variables by group. Are the means or medians the same?
- 6b - Association between quantitative variables: look at scatterplots and measures of association
- 6c - Predict/describe changes a continuous quantitative variable as a function of other variables: regression
- 6d - Compare frequencies or proportions by group: Use contingency table analysis or compare rates

7 Uncertainty determination
Determine how much error there is in the measures of interest. This may often be done using confidence intervals from a parametric analysis

Conclude

8 Data conclusion
Formulate your conclusion with respect to your research question using the results of your analysis.

9 Research conclusion
Present your general conclusions with respect to your main research question. Discuss how it fits in with previous research and how the conclusion might be affected by the details ot your research design.

Figure 3.1: Research topics for week 3

| Agresti | onlinestatbook.com |
|---|---|
| Agr: 9.1, 9.4 | OSB: 6.6, 14.1-14.3, 14.6 |
| Practice exercises are in the back of each chapter in the book. | |

## 3.1 Homework exercises

*Give a full answer to each question, justifying your answer and showing work, if necessary.*

1. A student wants to know if there is a linear relationship between heights of husbands (H) and heights of their wives (W). The student collects the following data, with heights of husbands and wives in centimeters:

| | $W(x)$ | $H(y)$ | $(x-\bar{x})$ | $(x-\bar{x})^2$ | $(y-\bar{y})$ | $(y-\bar{y})^2$ | $zx$ | $zy$ | $zxzy$ |
|---|---|---|---|---|---|---|---|---|---|
| | 166 | 182 | | | | | | | |
| | 164 | 178 | | | | | | | |
| | 166 | 180 | | | | | | | |
| | 165 | 178 | | | | | | | |
| | 170 | 181 | | | | | | | |
| | 165 | 181 | | | | | | | |
| $\Sigma$ | | | | | | | | | |

(a) Use SPSS to make a scatterplot of these data. To do this, you will need to type in the data in SPSS (one column for each variable). Include the plot with your homework.

(b) Based on the scatterplot, do you expect a positive or negative correlation coefficient? And how large or small do you expect the relationship to be?

(c) Compute the correlation coefficient between the lengths of husbands and wives. Fill in the above table to help you.

(d) How would the correlation coefficient change if all the husbands in the table were 10cm taller?

(e) How would the value change if all the men were 5% taller?

2. A psychologist is interested in the relationship between watching violent television programs and aggressive behavior in young children. During a period of several weeks 10 children were observed. For each of these children, the number of hours spent watching violent television programs and the number of aggressive episodes towards other children are recorded. The following table shows the results:

| Child # | Hours violent TV | # aggressive episodes |
|---|---|---|
| 1 | 14 | 9 |
| 2 | 8 | 6 |
| 3 | 6 | 1 |
| 4 | 12 | 8 |
| 5 | 10 | 4 |
| 6 | 9 | 5 |
| 7 | 9 | 6 |
| 8 | 13 | 1 |
| 9 | 4 | 2 |
| 10 | 5 | 1 |
| mean | 9 | 4.3 |
| SD | 3.37 | 2.98 |

(a) Use SPSS to make a scatterplot of these data. To do this, you will need to type in the data in SPSS (one column for each variable). Include the plot with your homework.

(b) Does there appear to be a relationship between the two variables? If so, is the relationship positive or negative?

(c) There is a linear relationship in the data, but there is one point that is clearly outside this relationship. Circle this point in your scatterplot (you can use a pencil to draw a circle on the print). What is the correlation coefficient without this point? (hint: use SPSS)

The correlation between the hours of violent television watched per week and number of aggressive episodes is 0.60 (you can use SPSS to confirm).

(d) Use these data to produce a regression equation to predict the number of aggressive episodes for a child given the number of hours of violent television the child watches.

(e) Using the equation you generated above, how many aggressive episodes would you predict for a child who watches 5 hours per week of violent television?

3. A group of educators wants to know if learning foreign languages causes students to have a better grasp of their native language. Using data from high school students' files, they collect data, including scores on a (native) language test that all students are required to take, and how many foreign language classes the students have taken. The educators find that students who have taken more foreign language classes have much higher scores on the test of native language skill. They would like to conclude that learning foreign languages improves native language skills. What third-variable explanation should prevent them from concluding this?

## In-Class Exercises

**Goals for this section**

- Practice interpreting scatterplots and corresponding correlations
- Learn to make scatterplots and calculate correlations using SPSS
- Learn how to implement a least-square regression in SPSS
- Learn the interpretation of the results of regression analyses

## 3.2 Correlation and Regression

### 3.2.1 Do correlations say it all?

The statistician Frank Anscombe [Anscombe, 1973] made a dataset to illustrate different types of associations between two variables. Use the data set "3 - Anscombe.sav" for this set of exercises.

1. Use SPSS to make scatterplots of each of the following associations. Use the boxes in Figure 3.2 to recreate the scatterplots, and estimate the correlation coefficient using the scatterplots.

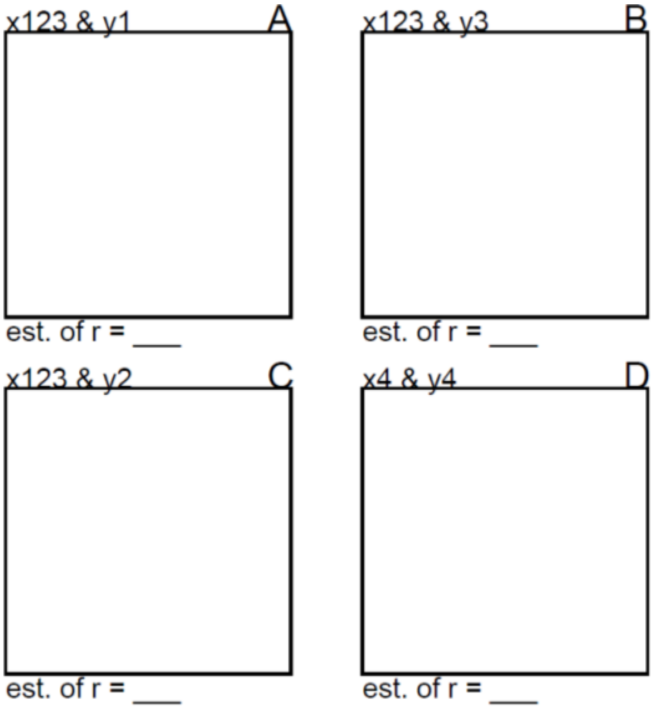


Figure 3.2: Space for drawing scatterplots in problem 1.

2. Using SPSS, compute the correlation coefficients between each of the pairs of variables above.

| Pair | r |
|---|---|
| A | |
| B | |
| C | |
| D | |

3. Did you estimate the correlation coefficients correctly? If not, what properties of the scatterplots fooled you?

4. Is the correlation coefficient always a good indicator of the strength of association between two variables? Why or why not?

5. Which of the associations in problem 1 is well-described by the correlation coefficient?

**The experts say...**

"Always plot your data before calculating common statistical measures such as correlation."

- [Moore et al., 2017], p. 104

6. Do you agree with [Moore et al., 2017]? What is the important lesson from Anscombe's data?

### 3.2.2 Measuring body fat and other characteristics

Psychologists sometimes use body fat as a measure of health. However, precisely measuring body fat is not easy. In order to measure body fat, a person is placed underwater and body density is measured. For a large study, it would be helpful to find a way to predict body fat from other easily-measured characteristics. Use the data file "3 - Fat.sav". This dataset gives two precise estimates of body fat ("fat1" and "fat2"), and a number of easier-to-measure characteristics from a sample of 252 men. It is not known which of the two precise estimates of body fat is best. Ensure that your results in this exercise are accurate, because you will need them in another exercise.

| | |
|---|---|
| Case | Participant # |
| fat1 | Precise estimate of body fat proportion (Brozek) |
| fat2 | Precise estimate of body fat proportion (Siri) |
| Age | Age |
| weight | Weight (kg) |
| height | Height (m) |
| adiposit | Body Mass Index (weight/height$^2$ ) |
| neck | Neck circumference (cm) |
| chest | Chest circumference (cm) |
| abdomen | Abdomen circumference (cm) |
| hip | Hip circumference (cm) |
| thigh | Thigh circumference (cm) |
| knee | Knee circumference (cm) |
| ankle | Ankle circumference (cm) |
| biceps | Biceps circumference (cm) |
| forearm | Forearm circumference (cm) |
| wrist | Wrist circumference (cm) |

1. You want to predict body fat proportion from other easy-to-measure variables.

   (a) Which variable(s) is (are) the dependent variables?

   (b) Which variable(s) is (are) the independent variables?

2. Use one of the body fat measures as the dependent variables. Choose an independent variable that you think will best predict body fat (do not use Weight, however). Using SPSS, make a plot of the association between the two variables you chose.

   (a) Sketch the plot using the following space and describe the relationship in words.

(b) Add your estimate of the best fitting linear regression line to the plot.

(c) Is a straight line appropriate to describe the relationship between the two variables? If not, find another variable that is linearly related. What variable did you find? Repeat the previous questions using this variable, and use the variable for the remaining problems.

3. For the following items, we will more precisely describe the relationship between the two variables you chose. Double-click on the scatterplot, click on “Elements”, then click on “Fit Line at Total” and show the regression line.

   (a) Discuss what the meaning of “least-square linear regression line” is.

   (b) Compare the regression line you drew in Problem 2b to the line that SPSS draws. Was your estimate a good estimate of the least-square regression line?

4. Using SPSS, find the linear regression parameters that correspond to the least-square regression line.

   (a) What are the slope, intercept, and corresponding regression equation? (Hint: look for the coefficients table, under “Unstandardized coefficients.”)

   (b) Does the intercept of this particular regression equation have any meaning? If so, what is the meaning? If not, why not? (Hint: think about the meaning of the intercept in general, and then consider it in light of the variables you are using.)

(c) What is the interpretation of the slope parameter of the regression equation?

(d) In the SPSS output, find the correlation coefficient and the $R^2$ . Calculate $R^2$ from the correlation coefficient to make sure the value is correct.

(e) How good is the independent variable you have chosen in predicting body fat proportion? Do you think your chosen independent variable would make a good alternative to the difficult-to-measure, precise body fat measures?

### 3.2.3 Measuring body fat using other measures (II)

Suppose that a number of students gave a short presentation regarding the results of their exploration into the relationship between body fat and other measures, including the results of the regression analysis. Read the report below, and think about these questions:

1. What information do you find useful?
2. Can you make comparisons between the different independent variables with respect to their usefulness in predicting body fat? Why or why not?
3. What further information would you find useful that is *not* presented?
4. How can the presentation of data be improved?

**Presentation Example**

Bodyweight (in kilograms) was used to predict body fat proportion (using Brozek's measure). A scatterplot, with the corresponding regression line, is shown below in Figure 3.3. It appears to be an approximately linear relationship. The analysis was based on a sample of 252 men with a range of weights is between 53.8 and 164.7 kg. Thus, the regression equation cannot be used to predict body fat proportion for children or women. The regression equation was:

$$F\% = 0.36 \times W - 10$$

here $F\%$ is body fat percentage, $W$ is weight in kilograms. For example, a man weighs 100kg. The regression prediction is:

$$F\% = 0.36 \times 100 - 10 = 26\%.$$

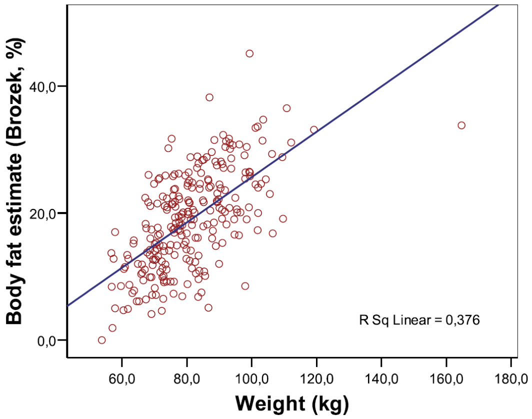
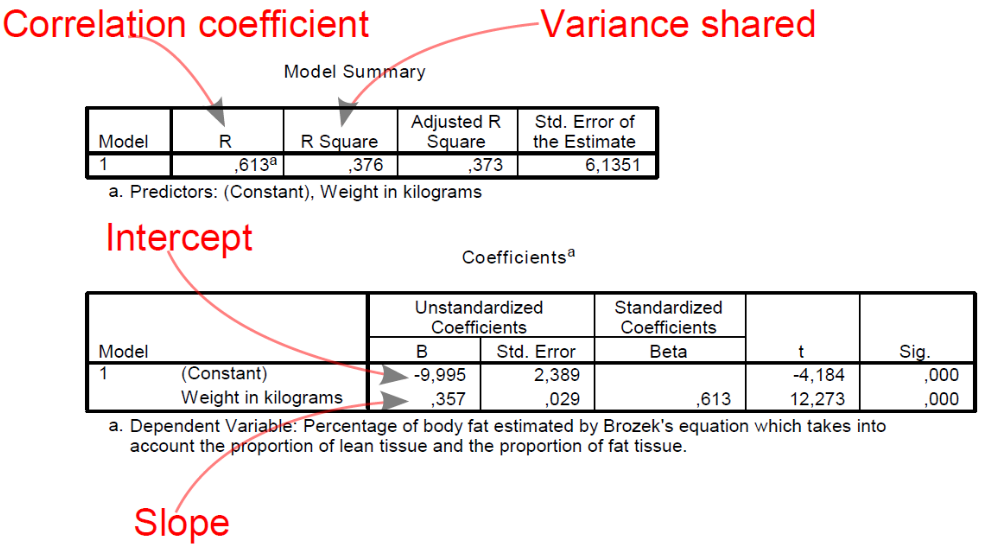


Figure 3.3: Scatterplot with regression line for the regression of *fat1* (a measure of body fat proportion) onto *weight* (body weight).

The correlation between the percentage of body fat and weight was 0.61, with $R^2 = 0.38$. This means that variance in body weight and variance in body fat share 38% common variance. A great deal of variance (more than 60%) is not shared. Figure 3.3 shows excerpts from the output from the regression analysis in SPSS.



Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,613[a] | ,376 | ,373 | 6,1351 |

a. Predictors: (Constant), Weight in kilograms

Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | -9,995 | 2,389 | | -4,184 | ,000 |
| | Weight in kilograms | ,357 | ,029 | ,613 | 12,273 | ,000 |

a. Dependent Variable: Percentage of body fat estimated by Brozek's equation which takes into account the proportion of lean tissue and the proportion of fat tissue.

Figure 3.4: SPSS output for the regression of *fat1* (a measure of body fat proportion) onto *weight* (body weight in kilograms).

**Questions about the presentation**

1. Which other information could be added to the report to make it more informative? What information could be added about the individual variables? How could the fit of the model be improved?

2. The presentation listed some limits of the usefulness of the regression equation. What were they, and what other limits would you add?

### Important points about correlation

- Association between two continuous variables can be explored graphically using a scatterplot.
- Association between variables can be described in terms of form, strength, and direction.
- The association between two variables can have different forms, one of which is the linear relationship.
- The correlation coefficient $(r)$ is a measure of the strength and direction of a linear relationship.
- The identification of outlying points in the relationship can be informative.
- A strong association between two tests does not necessarily mean they measure the same thing.
- If there is a causal link between two variables, there will be a relationship between them. However, this relationship does not have to be linear!

### Important points about regression

- A regression model is useful to predict one variable from other variables.
- If an independent variable is a useful predictor of a dependent variable, that does not mean that the dependent variable causes the independent variable.
- Linear regression is useful if a straight line can be used to describe the data. A least-square regression line minimizes the squared prediction error from every point.
- The intercept of the regression equation gives a prediction of the value of the dependent variable when the independent variable is exactly 0.
- Sometimes the intercept of the regression line is not interpretable, for various reasons (for example, the independent variable cannot meaningfully be 0)
- The slope of the regression line gives a measure of how the dependent variable changes as the independent variable increases or decreases in value.
- It is often incorrect to interpret the regression line outside of the range of your independent variable.
- The squared correlation coefficient, $R^2$ , gives a measure of how well the regression model fits.
- Scatterplots can be useful for looking at the form of the relationship between two variables and to check that the size of the residuals does not depend on the size of the independent variable (often a sign of nonlinearity).
- The quality of the prediction of the regression equation can (often) be increased by adding more independent variables to the equation.
- A regression with more than one independent variable is called a *multiple regression*.

# Chapter 4

# Week 4: More association

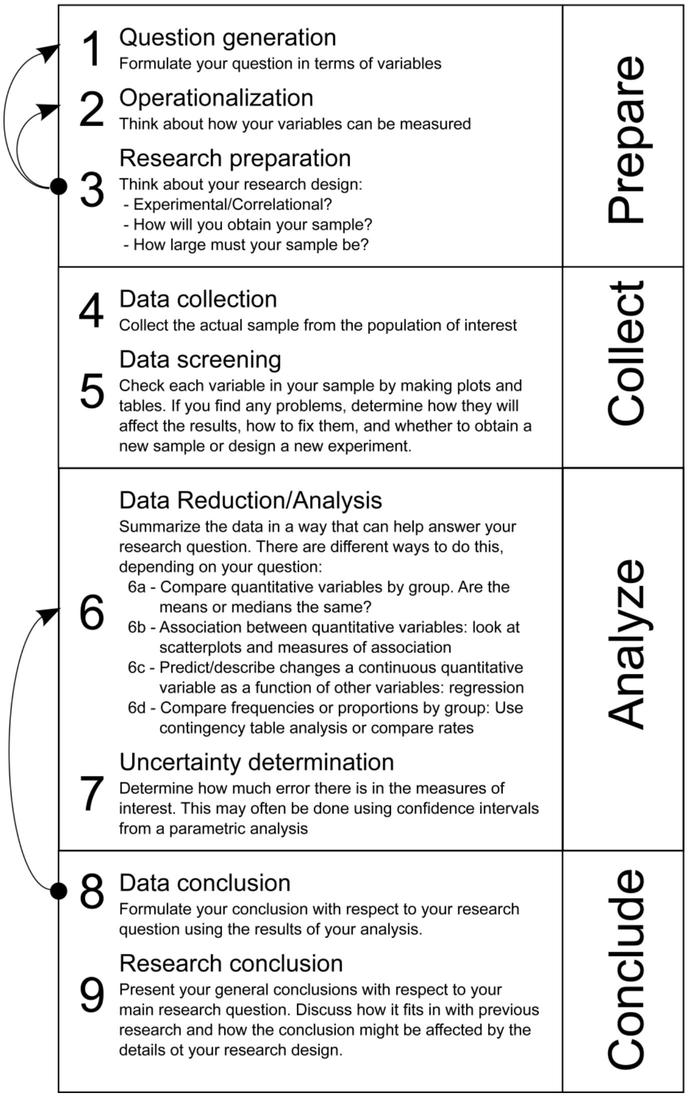

1 Question generation
Formulate your question in terms of variables
2 Operationalization
Think about how your variables can be measured
3 Research preparation
Think about your research design:
- Experimental/Correlational?
- How will you obtain your sample?
- How large must your sample be?
Prepare
4 Data collection
Collect the actual sample from the population of interest
5 Data screening
Check each variable in your sample by making plots and tables. If you find any problems, determine how they will affect the results, how to fix them, and whether to obtain a new sample or design a new experiment.
Collect
6 Data Reduction/Analysis
Summarize the data in a way that can help answer your research question. There are different ways to do this, depending on your question:
6a - Compare quantitative variables by group. Are the means or medians the same?
6b - Association between quantitative variables: look at scatterplots and measures of association
6c - Predict/describe changes a continuous quantitative variable as a function of other variables: regression
6d - Compare frequencies or proportions by group: Use contingency table analysis or compare rates
7 Uncertainty determination
Determine how much error there is in the measures of interest. This may often be done using confidence intervals from a parametric analysis
Analyze
8 Data conclusion
Formulate your conclusion with respect to your research question using the results of your analysis.
9 Research conclusion
Present your general conclusions with respect to your main research question. Discuss how it fits in with previous research and how the conclusion might be affected by the details ot your research design.
Conclude

Figure 4.1: Research topics for week 4

| **Agresti** | **onlinestatbook.com** |
|---|---|
| Agr: 2.2-2.5 | OSB: 6.1-6.4, 18.8 |
| Practice exercises are in the back of each chapter in the book. | |

## 4.1 Homework exercises

*Give a full answer to each question, justifying your answer and showing work, if necessary.*

1. During the lectures, we discussed different statistics which measure the strength of association between two variables. Which is the right statistic to use depends on what question you are asking about the data. There is sometimes little discussion about what the right statistic to use is, or which statistics best represent the strength of an association.

   The following data contain the results of 26 children with neurological impairments on two different tests for spatial reasoning [Efron and Tibshirani, 1993]. Use the data file "4 - Spatial Tests.sav" for this exercise. There are three variables of interest in the data file. "child" is each child's unique number, "testa" is the child's score on Test A, and "testb" is the child's score on Test B. The range of the two tests is similar: both have a possible range from 0 to 50. Psychologists like to use two versions of tests, hoping that the two tests measure the same characteristic. If the tests measure the same characteristic, researchers can test a person multiple times without giving them the same test. If researchers gave participants the same test twice, performance might vary for trivial reasons, such as memory for the first test they took. The question is to what extent the scores on the two tests are associated.

   (a) Use the following box to draw a scatterplot that represents perfect agreement between the two test scores. Make sure to label the x- and y-axis.

   (b) Describe the relationship you have drawn in problem 1a.

(c) Use SPSS to create a scatterplot showing the relationship between the scores in Test A and Test B.

(d) Examine the resulting scatterplot. How would you describe the relationship between scores on Test A and Test B?

(e) If the tests measure the same property, what would you expect the correlation to be?

(f) Compute and report the correlation coefficient using SPSS.

(g) What other summary statistics would you expect to be the same if the two tests measure the same characteristic?

(h) Use SPSS to compute the summary statistics of Test A and Test B and report them. What do you conclude about the similarity of scores between Test A and Test B?

2. Some researchers want to know what the opinions of the population of the Netherlands are with respect to euthanasia. After a television program addressing the deaths of terminal cancer patients, a phone survey of 1750 people who watched the program is taken. Of this sample, 70% say that they believe euthanasia should be allowed. The same question is asked of a random sample of 250 Dutch citizens. Of this sample, 35% believe that euthanasia should be allowed. Explain how the smaller sample of 250 Dutch citizens is nonetheless a better measure of the opinions of the Dutch people than the larger sample.

3. There is a researcher who is interested in whether anxiety during a pregnancy causes more anxiety for the child later in life. A total of 71 pregnant women were recruited to take part in the survey. In week 12 of pregnancy, each woman is given an anxiety questionnaire. When the children are 8 years of age, the mothers filled out an anxiety questionnaire for the children, and the children also filled out a separate anxiety questionnaire. The two scores taken at age 8 were combined to make one number. Of interest is the association between the score taken at age 8 and the score during pregnancy.

   (a) Is this research an experiment? Explain your answer.

   (b) What are the dependent and independent variables?

   (c) How might a third variable explain any association between the two scores?

4. Two friends of yours, Betty and Tim, have decided to calculate Kendall's $\tau$ for their preferences of four snacks: chips, chocolate, licorice, and icecream. Betty rates them in the following order of preference: icecream, chocolate, licorice, chips. Tim rates them in the following order of preference: chips, icecream, chocolate, licorice. Can you help them out and provide Kendall's $\tau$?

# In-Class Exercises

**Goals for this section**

- Gain insight into the selection and interpretation of different measures of association.

## 4.2 Choice of association statistic

In this section a number of example data sets are used so that you can practice choosing a good measure of association.

### 4.2.1 Physicians and life expectancy

These data come from the World Almanac and the 1993 Book of Facts [Rossman, 1994]. The data are from different countries with population greater than 20 million in 1990. The data file “4 - Televisions.sav” contains data on life expectancy and a number of other characteristics.

| Variable | Description |
|---|---|
| country | Name of the country |
| life_exp | Life expectancy |
| per_tv | Number of people per television |
| per_phys | Number of people per physician |
| femlife | Female life expectancy |
| malelife | Male life expectancy |

1. Using SPSS, make a figure showing the relationship between life expectancy and the number of people per physician.

   (a) Describe the association in the plot. What is the form? Is it positive or negative? How strong does the association look?

   (b) Is the sample correlation coefficient a good measure of the association between life expectancy and number of people per physician? Why or why not?

(c) Someone claims that more doctors in a country causes life expectancy to rise. Does your plot support that view?

2. Use SPSS to make a figure of the relationship between life expectancy and the number of people per television.

   (a) Describe the association in your plot.

   (b) Someone claims that more televisions in a country causes life expectancy to rise. Does your plot support that view? What do you think of their view?

   (c) What third variable might explain the relationship of both number of people per television and number of people per physician with life expectancy?

### 4.2.2 Fertility rate and life expectancy

In this section, we will use the “1 - Poverty.sav” dataset, which we have already worked with.

| Variable | Description |
|---|---|
| births | Number of births per 1000 people |
| deaths | Deaths per 1000 people |
| infantd | Infant mortality per 1000 people |
| malelife | Life expectancy at birth for men |
| femlife | Life expectancy at birth for women |
| gnp | Gross National Product |
| group | 1 Eastern Europe |
| | 2 South America and Mexico |
| | 3 Western Europe, North America, Japan, Australia, New Zealand |
| | 4 Middle East |
| | 5 Asia |
| | 6 Africa |
| country | Country name |

We will now explore a different aspect of these data: the relationship between fertility rates and life expectancy.

1. Using SPSS, create a figure that shows the relationship between life expectancy of women and number of births per 1000 people (fertility rate).

   (a) Describe the form, direction, and strength of the association in the plot.

   (b) Is it appropriate to use the correlation coefficient to describe the association? Estimate the correlation coefficient, then use SPSS to compute it.

   (c) What can explain the relationship between life expectancy of women and the fertility rate? Does having more children cause women to live shorter lives?

(d) Make a new plot, this time showing the association between the life expectancy of men and the fertility rate. In light of this new plot, would you change your answer to the previous problem?

(e) Make a new scatterplot, but now group the countries by regions. You can do this by creating a simple scatterplot, as usual, but this time add "Country grouping" to "Set Markers By:" in the Simple Scatterplot setup. How does the country grouping help explain the relationship between life expectancy of women and the fertility rate?

### 4.2.3 Subliminal priming

Cognitive psychologists often talk about the separation of unconscious processes and conscious processes. Unconscious processes are all those that happen outside your awareness. The amount of influence that unconscious processes have on conscious ones is a matter of debate in cognitive psychology.

Some psychologists have argued that stimuli that are flashed so fast that participants cannot see them can affect subsequent decisions nonetheless. In a typical task, a "prime" digit is displayed for a very short duration, and then another "target" digit is presented directly after. The so–called prime is presented so quickly that they are often not consciously perceived. After the target digit is displayed, participants are asked to decide whether the target digit is less than 5 or greater than 5. The interesting finding is that when the prime digit is on the same side of 5 as the target digit, participants classify the target faster, despite the fact that the prime was barely visible. The phenomenon of identifying a target more quickly due to the effect of some prime is called *priming*. The question is whether primes that are presented so fast that participants cannot see them still affect the decision to the target.

[Morey et al., 2008] were interested in the question of how to decide whether or not primes are possible to consciously see. They presented digits (2, 3, 4, 6, 7, or 8) to 22 participants at different speeds, from 17ms to 100ms, and asked participants to classify them as less-than- or greater-than-5. The resulting data is in the file "4 - Subliminal Priming.sav", open that file.

| **Variable** | **Description** |
|---|---|
| partnum | Participant number |
| ncor17ms | Number correct, out of 90, to 17ms stimuli |
| ncor25ms | Number correct, out of 90, to 25ms stimuli |
| ncor42ms | Number correct, out of 90, to 42ms stimuli |
| ncor58ms | Number correct, out of 90, to 58ms stimuli |
| ncor75ms | Number correct, out of 90, to 75ms stimuli |
| ncor100ms | Number correct, out of 90, to 100ms stimuli |
| pcor17ms | Proportion correct to 17ms stimuli |
| pcor25ms | Proportion correct to 25ms stimuli |
| pcor42ms | Proportion correct to 42ms stimuli |
| pcor58ms | Proportion correct to 58ms stimuli |
| pcor75ms | Proportion correct to 75ms stimuli |
| pcor100ms | Proportion correct to 100ms stimuli |

1. If a participant were completely unable to see the digits at a particular speed, what proportion correct would you expect? Remember, there are two possible responses, less-than-5 or greater-than-5. Explain your answer.

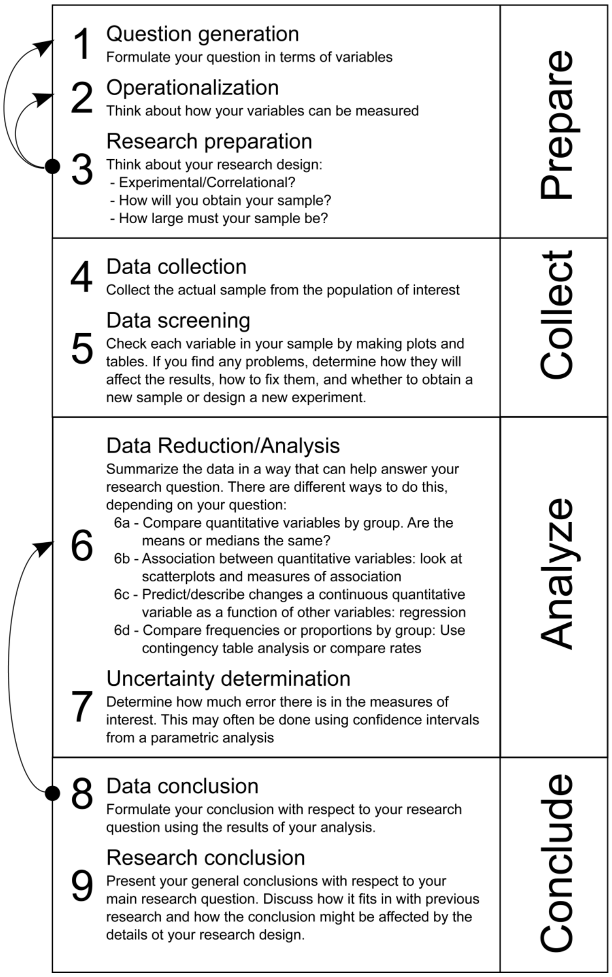
2. Using SPSS, find the mean proportion correct for each digit duration. Which duration is the best? Which duration is the worst? How does the proportion correct change as a function of duration?

3. Using SPSS, create a scatterplot of participants' proportion correct in the easiest condition (pcor100ms) against their proportion correct in the next–easiest condition (pcor75ms). Describe the association with respect to form, direction, and strength. What is the $r^2$?

4. Now, create a scatterplot of participants' proportions correct in the easiest condition (pcor100ms) against their proportion correct in the hardest condition (pcor17ms). Describe the association with respect to form, direction, and strength and comment on the difference with the result from the previous question.

# Chapter 5

# Week 5: Sampling design, chance, and modeling chance



1 Question generation
Formulate your question in terms of variables

2 Operationalization
Think about how your variables can be measured

3 Research preparation
Think about your research design:
- Experimental/Correlational?
- How will you obtain your sample?
- How large must your sample be?

4 Data collection
Collect the actual sample from the population of interest

5 Data screening
Check each variable in your sample by making plots and tables. If you find any problems, determine how they will affect the results, how to fix them, and whether to obtain a new sample or design a new experiment.

6 Data Reduction/Analysis
Summarize the data in a way that can help answer your research question. There are different ways to do this, depending on your question:
6a - Compare quantitative variables by group. Are the means or medians the same?
6b - Association between quantitative variables: look at scatterplots and measures of association
6c - Predict/describe changes a continuous quantitative variable as a function of other variables: regression
6d - Compare frequencies or proportions by group: Use contingency table analysis or compare rates

7 Uncertainty determination
Determine how much error there is in the measures of interest. This may often be done using confidence intervals from a parametric analysis

8 Data conclusion
Formulate your conclusion with respect to your research question using the results of your analysis.

9 Research conclusion
Present your general conclusions with respect to your main research question. Discuss how it fits in with previous research and how the conclusion might be affected by the details ot your research design.

Figure 5.1: Research topics for week 5

| **Agresti** | **onlinestatbook.com** |
|---|---|
| Agr: 4.1, 4.2, 4.4-4.7 | OSB: 5.1-5.3 |
| Practice exercises are in the back of each chapter in the book. | |

## 5.1 Homework exercises

*Give a full answer to each question, justifying your answer and showing work, if necessary.*

1. A psychologist is interested in whether meditation reduces anxiety. The level of anxiety in each person was determined by an interview with the psychologist. After the interview, the participants were randomly assigned to one of two groups. In the first group, the participants were trained in techniques of meditation. They meditated every day for one hour during the one month experimental period. For participants in the second group, it was *suggested* that the participants take some quiet time every day. At the end of the month the psychologist again interviewed the participants and the psychologist rated their level of anxiety. It appeared that the anxiety levels of the experimental (meditation) group were lower on average than the control (suggested quiet time) group.

   (a) What are the dependent and independent variables in this study?

   (b) The outcome in this experiment may be due to the fact that the interviews were not blind. What is meant by blind?

   (c) How might this research be affected by the lack of blindness?

   (d) What variables might explain the results in this study?

2. Read the following paragraph. Indicate whether each underlined number is a parameter or a statistic.

A telemarketing company in Groningen uses a computer system to choose phone numbers at random in the Netherlands. Of the first 100 numbers chosen, 26 are not listed in the phone directory. This is not surprising, since 19% of the phone numbers in the Netherlands are unlisted.

3. The type of medical care that a patient receives can vary with the age of the patient. In a large study of women who had tumors in their breasts, women were asked whether they had previously had a mammogram. The results are expressed as proportions of the sample in the table below. For example, 0.321 is the proportion of the women in the sample group who are younger than 65 years old and who had had a mammogram. All the proportions sum to 1 because the table lists every possible combination of age and mammogram status.

| | **Mammogram?** | | |
|---|---|---|---|
| **Age** | Yes | No | |
| $< 65$ | 0.321 | 0.124 | __ |
| $\geq 65$ | 0.365 | 0.190 | __ |
| | — | — | — |

(a) What is the probability that a randomly selected participant in this study is younger than 65 years old? What is the probability that a randomly selected participant is 65 years old or older?

(b) What is the probability that a randomly selected participant had a mammogram? What is the probability that they did not?

(c) For a randomly selected participant in this sample, are the events A = {the patient is 65 or older} and B = {A mammogram has been done} independent? Discuss why or why not.

4. A survey of 400 Dutch adults presented the following question: “What, in your opinion, is the most serious problem for education in our schools?” Suppose that in the population, the true proportion of those who would answer “violence” is 0.25. The proportion of those answering “violence,” however, will vary from sample to sample. Suppose that the sample proportion of those answering “violence” is a normal with a mean of 0.25 and a standard deviation of 0.022. What is the probability that . . .

   (a) . . . at least half the sample answers “violence?”

   (b) . . . less than 25% of the sample answers ”violence?”

   (c) . . . the sample proportion is between 0.25 and 0.30?

   (d) . . . the sample proportion is less than 0.05 or greater than 0.30?

## In-Class Exercises

**Goals for this section**

- To learn about different sampling designs.
- Get an overview of the advantages and disadvantages of the different methods of collecting data.

## 5.2 Sampling designs and sampling distributions

### 5.2.1 The Groningen 4 mile

Suppose that you would like to do research on the participants of the four-mile race in Groningen (the Groningen 4-mile). The organization would like to know how the participants found out about the race and how satisfied they were with the organization of the race. The participants can be divided into different categories. Men and women are in different categories, as well as professional runners, teams, individuals, and so on. Read the following scenarios for the selection of the sample of participants for the research.

A. Samples are drawn from every category in which racers compete. The participants get a number in every category when they register for the race. Random numbers from every category are then drawn from each category. A questionnaire is sent to the selected participants' home.

B. A large sign at the finish line of the race asks everyone who would like to fill out a questionnaire to go to a nearby tent. Those who volunteer fill out the questionnaire after the race.

C. Every tenth person that finishes the race is approached by an interviewer, who has a questionnaire. A random number is used to determine who among the first ten to finish is given the questionnaire; after that, every tenth person is asked.

D. Participants are able to register from two months before the race up until the day of the race. Questionnaires are sent to racers who register on certain days. All registrants' names are placed in a hat, and eight names are drawn. Questionnaires are sent to every fifth person who registered on the same day as one of the drawn names. Random numbers are used to decide who of the first five registrations gets the questionnaire.

E. Every racer gets a unique number when they register. A computer is used to select 250 registration numbers at random. Questionnaires are sent to these 250 racers.

Answer the following questions about the sampling schemes.

1. What is the population in which the organizers of the race are interested?

2. Describe the advantages of each of the sampling schemes above.

3. Which scheme would you use for your research? Why do you think that is the best scheme?

4. Are any of these schemes likely to have an under-representation problem? If so, which schemes?

5. Are any of these schemes likely to have a non-response problem? If so, which schemes? Are some schemes more susceptible to non-response bias than others?

6. Are some schemes more susceptible to response bias than others? Which schemes, and why?

### 5.2.2 Assessing a research design

Read the abstract of the article *Cognitive-Behavioral Therapy for Children With Anxiety Disorders in a Clinical Setting: No Additional Effect of a Cognitive Parent Training* [Nauta et al., 2003].
"Objective: To evaluate a 12-week cognitive-behavioral treatment program for children with anxiety disorders and the additional value of a seven-session cognitive parent training program. Method: Seventy-nine children with an anxiety disorder (aged 7-18 years) were randomly assigned to a cognitive behavioral treatment condition or a wait-list control condition. Families in the active treatment condition were randomly assigned to an additional seven-session cognitive parent training program. Semistructured diagnostic interviews were conducted with parents and children separately, before and after treatment and at 3 months follow-up. Questionnaires included child self-reports on anxiety and depression and parent reports on child's anxiety and behavioral problems. Results: Children with anxiety disorders showed more treatment gains from cognitive-behavioral therapy than from a wait-list control condition. These results were substantial and significant in parent measures and with regard to diagnostic status, but not in child self-reports. In the active treatment condition, children improved on self-reported anxiety and depression, as well as on parent reports on their child's anxiety problems. These results were equal for clinically referred and recruited children. Child self-reports decreased to the normal mean, whereas parents reported scores that were lower than before treatment but were still elevated from the normal means. No significant outcome differences were found between families with or without additional parent training. Conclusions: Children with anxiety disorders profited from cognitive-behavioral therapy. Children improved equally whether or not additional parent training was offered."

1. Using a drawing, describe the research design in the abstract. What type of design is used? What are the independent variables and what are the levels of each independent variable? Also, how is the population of interest defined, and how is it sampled?

2. What are the dependent variables in this research?

3. Is it possible to set up this research in a double blind manner? If so, how? If not, why not?

4. What conclusion can you draw about the effectiveness of individual CBT? What conclusion can you draw about the added effect of parent training on top of individual CBT?

5. What shortcomings are there in this research design?

### 5.2.3 Sampling distributions

A researcher wants to determine the cognitive effect of general anesthesia. The researcher has a sample of 80 participants who have undergone hip surgery under general anesthesia. Among other measures, the IQ of the 80 participants is measured. In this section, you will examine the sampling distribution of the median with sample size 5 and 15. In order to do this, *we will consider the 80 participants as the population of interest*. You will draw samples of 5 and 15 from the population of 80, and then compute the median of the sample. By repeating this a number of times, you can draw a histogram of the distribution of the sample medians. Open the datafile "5 - Narc.sav". We are only interested in the "iq" variable.

1. Using SPSS, compute and report the median IQ for the 80 participants [1]

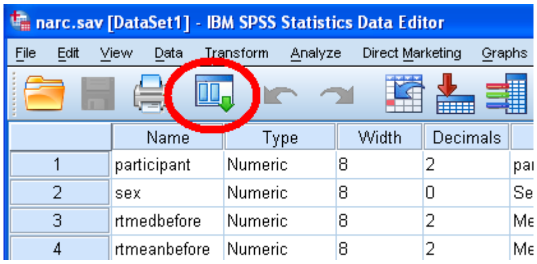
2. Using SPSS, select 5 random observations from the 80.[2] For this sample of 5, compute the median and add it to the appropriate cell (sample 1, size 5) of the table below.

[1] To do this, use Analyze→Reports→Case Summaries. Deselect "Display cases" and put "iq" under variables. Click on "Statistics…" and choose to compute only the median. Click "Continue", then "OK".

[2] To do this, select Data→Select Cases... from the menu. Select "Random sample of cases", then under "Output" ensure that "Filter out unselected cases" is selected. Then click "Sample…". In the box that pops up, choose "Exactly 5 cases from the first 80 cases" and click "Continue," then "OK". This will filter out all except 5 of the observations.

| Sample number | Sample size 5 | Sample size 15 |
|---|---|---|
| 1 | __ | __ |
| 2 | __ | __ |
| 3 | __ | __ |
| 4 | __ | __ |
| 5 | __ | __ |
| 6 | __ | __ |
| 7 | __ | __ |
| 8 | __ | __ |
| 9 | __ | __ |
| 10 | __ | __ |
| 11 | __ | __ |
| 12 | __ | __ |
| 13 | __ | __ |
| 14 | __ | __ |
| 15 | __ | __ |

3. Repeat the last step until you have completed the "Sample size 5" column in the table, then use a sample size of 15 out of 80 to complete the "Sample size 15" column. *Important tip*: use the SPSS feature "dialog recall", shown in Figure 5.2 to save time. Pull down the dialog recall, select the appropriate step, and click "OK" (the options you previously used will still be selected).



| | Name | Type | Width | Decimals | |
|---|---|---|---|---|---|
| 1 | participant | Numeric | 8 | 2 | par |
| 2 | sex | Numeric | 8 | 0 | Se |
| 3 | rtmedbefore | Numeric | 8 | 2 | Me |
| 4 | rtmeanbefore | Numeric | 8 | 2 | Me |

Figure 5.2: The dialog recall button in SPSS.

4. Are the median values in the table statistics or parameters? Explain your answer.

5. Create a new SPSS dataset using File→New→Data. Enter the data in your table into SPSS. You will have two scale variables: "Medians sample size 5" and "Medians sample size 15."

6. Use SPSS to make histograms of the two variables, and sketch them in Figure 5.3. Make sure that the x-axes on your two histograms have the same range, so that you can compare them easily.

7. Describe the two sampling distributions of the median (one for sample size 5, and one for sample size 15) you sketched.

8. Use SPSS to compute the minimum, maximum, median, mean, and standard deviation of the sampled medians for the two sample sizes. Fill them in the table below.

| Statistic | Sample size 5 | Sample size 15 |
|---|---|---|
| Minimum | __ | __ |
| Median | __ | __ |
| Mean | __ | __ |
| Maximum | __ | __ |
| Standard deviation | __ | __ |

9. Compare the two distributions of sample medians using both the summary statistics and your histograms. What do you find?

10. Is the sample median unbiased? Explain your answer.

11. We have learned that variability in a statistic can be reduced by increasing the sample size. Is the data you collected in this exercise consistent with that fact?

# Chapter 6

# Week 6: Probability

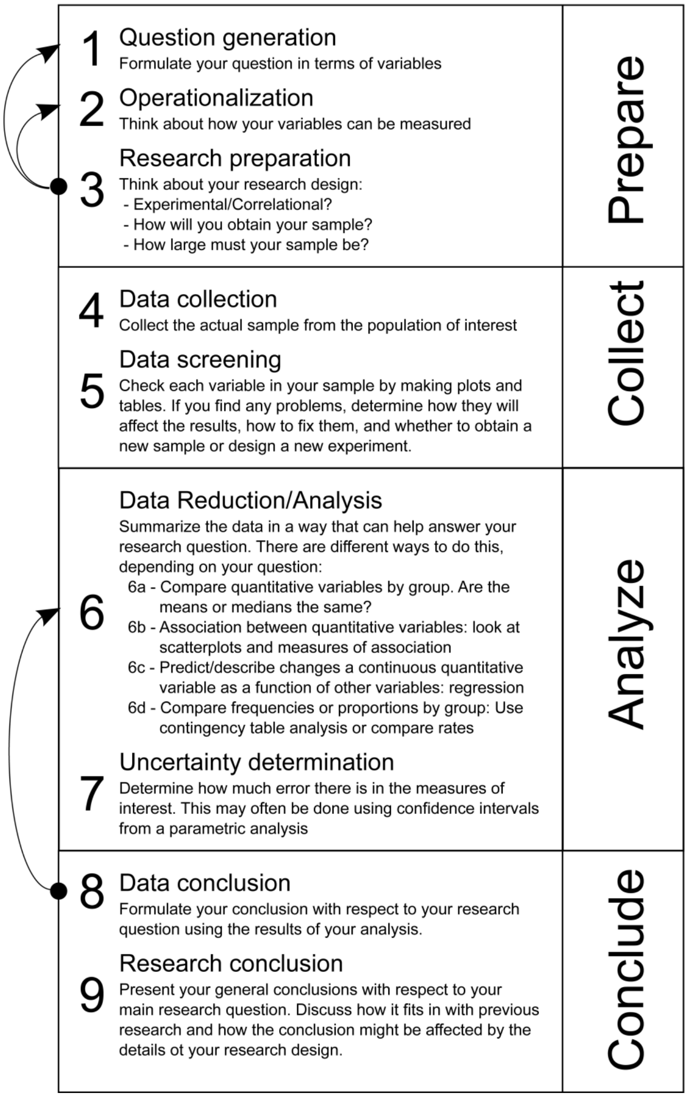

1 Question generation
Formulate your question in terms of variables
2 Operationalization
Think about how your variables can be measured
3 Research preparation
Think about your research design:
- Experimental/Correlational?
- How will you obtain your sample?
- How large must your sample be?
Prepare
4 Data collection
Collect the actual sample from the population of interest
5 Data screening
Check each variable in your sample by making plots and tables. If you find any problems, determine how they will affect the results, how to fix them, and whether to obtain a new sample or design a new experiment.
Collect
6 Data Reduction/Analysis
Summarize the data in a way that can help answer your research question. There are different ways to do this, depending on your question:
6a - Compare quantitative variables by group. Are the means or medians the same?
6b - Association between quantitative variables: look at scatterplots and measures of association
6c - Predict/describe changes a continuous quantitative variable as a function of other variables: regression
6d - Compare frequencies or proportions by group: Use contingency table analysis or compare rates
7 Uncertainty determination
Determine how much error there is in the measures of interest. This may often be done using confidence intervals from a parametric analysis
Analyze
8 Data conclusion
Formulate your conclusion with respect to your research question using the results of your analysis.
9 Research conclusion
Present your general conclusions with respect to your main research question. Discuss how it fits in with previous research and how the conclusion might be affected by the details ot your research design.
Conclude

Figure 6.1: Research topics for week 6

| **Agresti** | **onlinestatbook.com** |
|---|---|
| Agr: - | OSB: 5.4, 9.1-9.3 |
| Practice exercises are in the back of each chapter in the book. | |

## 6.1 Homework exercises

*Give a full answer to each question, justifying your answer and showing work, if necessary.*

1. A researcher is interested in studying the play behavior of children. He is specifically interested in how many different toys children will play with if given the option. In the research, children are placed in a room with six toys. For each child, the number $(X)$ of different toys they play with is recorded. Suppose that over the course of many similar studies, the distribution of the number of toys children play with is shown to be the following:

| Number of toys played with $(X)$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Probability | 0.05 | 0.16 | 0.30 | 0.23 | 0.17 | 0.07 | 0.02 |

(a) Find the mean $\mu_X$ and standard deviation $\sigma_X$ of the random variable $X$.

(b) What is the mean$\mu_Y$ and standard deviation $\sigma_Y$ of the variable $Y = X - 2.6$?

(c) What is the meaning of the variable $Y = X - 2.6$?

2. It has been shown that student motivation (as measured by the Questionnaire of Study Habits and Attitudes) differs between men and women. Specifically, female students score higher on average than male students. The distribution of motivation scores for female students $(F)$ is normal with a mean of 122 and a standard deviation of 28; the distribution of motivation scores for male students $(M)$ is normal with a mean of 106 and a standard deviation of 36. In notation,

F $\sim$ Normal(122, 28)
M $\sim$ Normal(106, 36)

A female and a male student are chosen at random and these two students take the QSHA.

(a) Argue why these two scores should be independent of one another.

(b) Compute the mean and standard deviation of the distribution of the difference of scores $(F - M)$.

(c) Can you use this information to determine the probability that the random female scores higher than the random male? If yes, tell how and compute the probability. If not, explain why not and tell what extra information is needed.

3. A researcher is interested in the relationship between the effectiveness of psychotherapy and the educational level attained by the patient. The researchers obtain a sample of 2500 patients of psychotherapy, and classify them according to the level of education they have: education up to age 17 but no university, some university but no degree, and bachelor's degree. The results are in the following table.

| | | Education | | |
|---|---|---|---|---|
| Psychotherapy | No University | Some University | Bachelor's degree | Total |
| Effective | 300 | 250 | 500 | 1050 |
| Ineffective | 700 | 500 | 250 | 1450 |
| Total | 1000 | 750 | 750 | 2500 |

(a) What is the probability that a randomly selected participant from this study had found psychotherapy to be effective?

(b) Given that a patient had no university education, what is the probability that psychotherapy was effective for them?

(c) Given that a patient had only some university education, what is the probability that psychotherapy was effective for them?

(d) Given that a patient had a bachelor's degree, what is the probability that psychotherapy was effective for them?

(e) Given that psychotherapy was effective for the patient, what is the probability that they had a bachelor's degree?

4. People who have a particular gene (call it M) have a high chance of clinical depression as an adult. It has been shown that 80% of depressed people have gene M, and 20% do not. Of healthy people, it is also known that 10% have the gene M and 90% do not. Additionally, suppose that 30% of the population will experience depression as an adult.

(a) Make a tree diagram to clearly show the relationship between depression and the gene M.

(b) Compute the probability that a person from the population who we know has gene M experiences depression.

(c) Compute the probability that a person from the population who we know does not have gene M experiences depression.

## In-Class Exercises

## 6.2 Distribution of the population, samples, and the sample mean

**Goals for this section** In this section, we will explore . . .

- the relationship between the population distribution and sample distribution.
- the relationship between the population distribution and the sampling distribution of the mean.
- the effect of the sample size on 1 and 2.

### 6.2.1 Distribution of a variable: binomial or normal?

In the lectures, you have learned about normal random variables. Random variables are normally distributed if and only if the density follows a particular curve.

**Definition: Normal density curve** The normal density curve for a Normal $(\mu, \sigma)$ distribution is the function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(x-\mu)^2}{2\sigma^2}$$

where exp is the exponential function (i.e., $e^x$ )

A plot of this curve is shown in Figure 6.2. Only continuous distributions whose density function match the normal density function are truly normal.

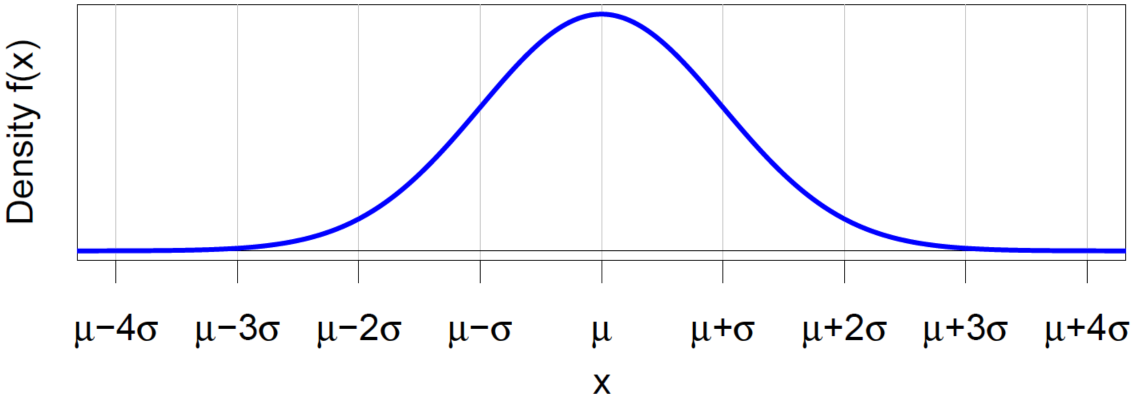


Figure 6.2: The normal density curve

A binomially distributed random variable, on the other hand, is the discrete distribution representing the number of successes out of $n$ attempts, when the probability of success is $p$. When $n$ is large and $p$ is not extreme, the binomial distribution can be approximated by the normal distribution. However, it must be emphasized that this is only an approximation. A binomial random variable can never actually be normal. These are the properties of a binomial random variable:

- Fixed number of observations $n$
- All $n$ observations are independent of one another
- Every observation is in one of two categories (e.g. male/female, or yes/no)
- The probability $p$ of being in a category is always the same

1. Describe why a *normal* random variable must be a *continuous* random variable.

2. In the problems below regarding the difference between the normal and binomial distributions, a number of situations are described. For each situation, describe the variables and, for each, explain whether the binomial distribution or the normal distribution is a more appropriate model. Wherever possible, give the values of the parameters of the binomial or normal distribution, in the form Binomial$(n = \cdots, p = \cdots)$ or Normal$(\mu = \cdots, \sigma = \cdots)$.

   (a) Based on an extensive survey of every employee, a large corporation knows that 60% of employees are happy with their job. They have decided to ask additional questions to a random sample of 15 employees. What distribution is most appropriate for modeling “the number of employees who are happy with their job in the sample of 15?”

   (b) In order to be safe when driving long distances, it is recommended that a driver takes a break after two hours of driving. However, drivers do not take a break after exactly two hours of driving. It is thought that 95% of drivers take a break after 100 to 140 minutes of driving. What distribution is most appropriate for modeling ”the number of minutes of driving before drivers stop for a break?”

   (c) A researcher is interested in how toys capture the attention of babies. For each baby, a toy is placed in front of the baby, and it is noted whether the baby looks at it. This is repeated 10 times. What is the appropriate distribution form modeling “the number of times that a baby looks at the toy?”

(d) The researcher mentioned in the previous problem is also interested in the length of time (in seconds) that a baby looks at the toy, when they do look at the toy. What distribution is most appropriate for modeling "the number of seconds a baby looks at the toy?"

(e) A researcher wants to study how a particular genetic mutation is related to behavior. Out of every 1000 people in the population, it is known that about 50 have the mutation. Because it is not obvious who has the mutation and who does not, the researcher takes a small random sample of 5 people to check their DNA for the mutation. What is the appropriate distribution for modeling "the number of people in the researcher's sample who have the mutation?"

It is important that you can understand the difference between the following terms:

- *Population distribution:* the distribution of a population of scores. This is the same from sample to sample. It cannot be observed directly, except in unusual cases when you know what it is.

- *Sample distribution:* the distribution of a sample drawn from the population. When a sample is drawn, the sample distribution can be observed, and is often depicted using a histogram. When the sample size is very large, the sample distribution will begin to look more and more like the population distribution.

- *Sampling distribution of the mean:* the distribution of the mean ($\bar{x}$) for a given sample size drawn from the population.[1] This depends on the population distribution and the sample size. As the sample size gets large, this becomes more and more independent of the population distribution, and more normal.

  The rest of the exercises in this chapter will help to clarify these differences. As you finish the rest of the chapter, whenever a distribution is mentioned, think about which of the three distributions is meant and why.

### 6.2.2 Sample distributions and the sampling distribution of the mean

In this section, we will use the statistical software program package R. Double–click on the blue circle with a white R symbol on your desktop or open the "RStudio" program from the start menu to launch R. Once in the program, click on "File → Open File . . . " and open the script "6 - Sampling Distributions.r".

[1] There can be sampling distributions of other statistics, too. The most important sampling distribution for now is the mean, but there are also sampling distributions of standard deviation, maximum, minimum, and others. In general, a sampling distribution is the distribution of some function of your sample.

You will notice two boxes with text on your screen. One is the "R Console", the other one is the "6 - Sampling Distributions.r" file you just opened, this is your "script". In your script, text that is preceded by the # symbol are "comments". They are meant to help you understand what the lines of code are for. The lines of code that are not preceded by the # symbol are the meat of the code.
At first, code may look intimidating, but do not worry: the exercises below will take you through it step by step.

1. Select the first four lines of code, the first line being the one that starts with "Size = 5". Once you have selected this code, you need to "execute" the code. To do this, it needs to be put in the console. You can either copy the code you selected, put the cursor in the console (the bottom line that starts with >) and paste it. Alternatively, you can press CTRL + R. This action is called "sourcing code". You have now created three variables: "Data5", "Data100", and "Data10000".

   (a) In the console, type "Data5" (without parentheses) and press "enter". If you have done this correctly, five numbers should appear. What do you think these numbers mean? Hint: the line above the one that starts with "Size = 5" describes what you are doing.

   (b) Now, type "Data100" (without parentheses) and press "enter". If you have done this correctly, one hundred numbers should appear. How do these numbers relate to the previous ones?

   (c) Change the variable "Size" from 5 to 2. To do this, simply change the number after the "=" sign to 2. Source the four lines again and type in Data5 in the console once more. There should still be five numbers, so what have you changed? Hint: there is a difference between the number of observations in a sample and the number of samples.

   (d) Change the variable "Size" back from 2 to 5. Now source all the code that belongs to Exercise 1, a total of 9 lines excluding comments (15 including comments). After a short delay, a third box pops up with four different plots. The top-left plot displays the population distribution, the other three display sampling distributions. Please sketch the distributions below and comment on the differences between the distributions.

2. In this exercise, we will examine what happens if we sample from a different population distribution. The uniform distribution has a minimum and a maximum value. Within this “range”, each value has an equal probability. Therefore, the density function looks like a straight line. This is different from the Normal distribution, where values further away from the mean have a lower probability of occurring. You can examine what this distribution looks like by sourcing the line that begins with “curve (dunif . . . ”.

   (a) Why is this distribution called “the uniform distribution”? Is this distribution skewed?

   (b) In the previous exercise, we have examined what the sampling distribution of the mean looked like when we drew 5, 100, or 10,000 samples from a normally distributed population. What do you think these sampling distributions will look like for the uniform distribution? Sketch your prediction here.

   (c) Check your answer using R. Select all the lines of code that belong to exercise 2 and source it. Then examine the four plots. Was your prediction correct? How does the sampling distribution of the mean of a uniform population relate to that of a normal population?

3. Last R exercise, we will now examine what happens if we sample from a skewed population distribution: the exponential distribution. You can examine what this distribution looks like by sourcing the line that begins with “curve (dexp . . . ”.

   (a) Is this distribution left-skewed or right-skewed?

(b) In the previous exercise, we have examined what the sampling distribution of the mean looked like when we drew 5, 100, or 10,000 samples from a normally distributed population. What do you think these sampling distributions will look like for the exponential distribution? Sketch your prediction here.

(c) Check your answer using R. Select all the lines of code that belong to exercise 3 and source it. Then examine the four plots. Was your prediction correct? How does the sampling distribution of the mean of an exponentially distributed population relate to that of a normal population?

(d) Had you noticed that in this exercise, we set the sample size to 25? Set the sample size to 2 and source the code again. What changed?

(e) Set the sample size to 1, source the code again, and compare the top-left and the bottom-right panels. What do you notice?

## Important points about sampling distributions

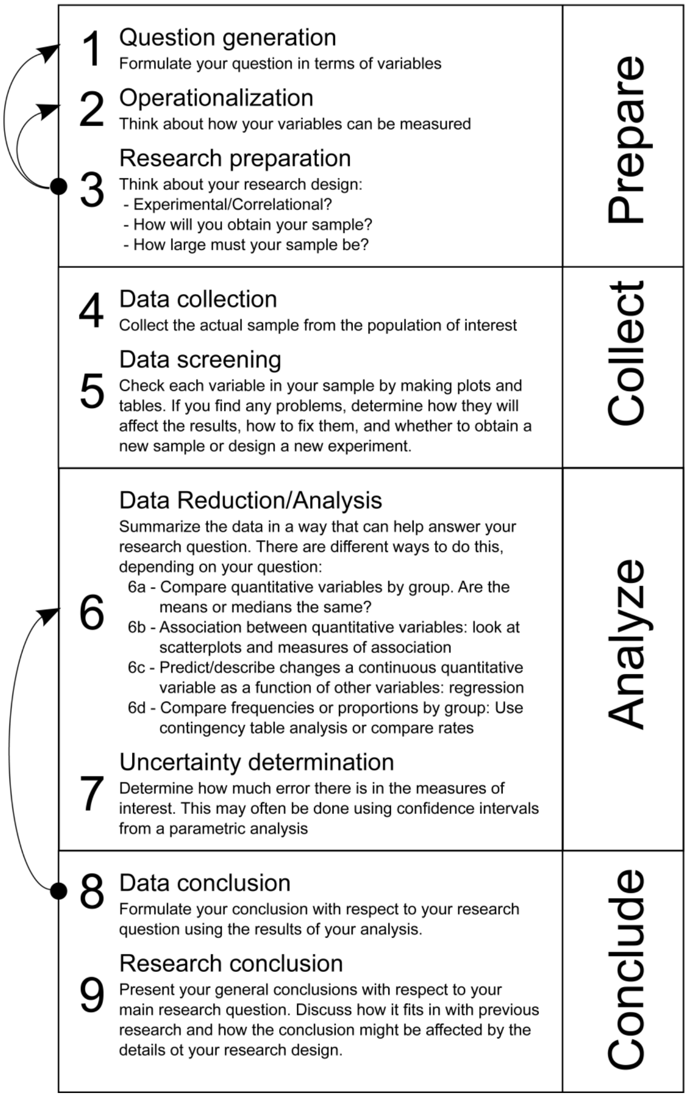
- If the mean of a random variable $X$ is $\mu_X$, the mean of the sampling distribution of $\bar{x}$ will be $\mu_X$, the same as the population mean.
- If the standard deviation of a random variable $X$ is $\sigma_X$, the standard deviation of the sampling distribution of $\bar{x}$ will be $\frac{\sigma_X}{\sqrt{n}}$ where n is the sample size.
- These relationships are the same *regardless of the distribution of the random variable*!

## New R functions

- *?x* Brings up help-function for a formula (try "?mean").
- *mean(x)* Calculates the mean of vector $x$.
- *rnorm(x,y,z)* Generates $x$ random values, normally distributed with mean $y$ and sd $z$.
- *dnorm(x,y,z)* Gives density for value $x$ under a normal distribution with mean $y$ and sd $z$.
- *curve(x,...)* Plots a density curve of density function $x$.
- *hist(x,...)* Plots a histogram of vector $x$.
- *runif(x,y,z)* Generates $x$ random values, uniformly distributed with minimum $y$ and maximum $z$.
- *dunif(x,y,z)* Gives density for value $x$ under a uniform distribution with minimum $y$ and maximum $z$.
- *rexp(x,y)* Generates $x$ random values, exponentially distributed with rate parameter $y$.
- *dexp(x,y)* Gives density for value $x$ under an exponential distribution with rate parameter $y$.

## Chapter 7

# Week 7: Sampling distributions



1 Question generation
Formulate your question in terms of variables

2 Operationalization
Think about how your variables can be measured

3 Research preparation
Think about your research design:
- Experimental/Correlational?
- How will you obtain your sample?
- How large must your sample be?

Prepare

4 Data collection
Collect the actual sample from the population of interest

5 Data screening
Check each variable in your sample by making plots and tables. If you find any problems, determine how they will affect the results, how to fix them, and whether to obtain a new sample or design a new experiment.

Collect

6 Data Reduction/Analysis
Summarize the data in a way that can help answer your research question. There are different ways to do this, depending on your question:
6a - Compare quantitative variables by group. Are the means or medians the same?
6b - Association between quantitative variables: look at scatterplots and measures of association
6c - Predict/describe changes a continuous quantitative variable as a function of other variables: regression
6d - Compare frequencies or proportions by group: Use contingency table analysis or compare rates

7 Uncertainty determination
Determine how much error there is in the measures of interest. This may often be done using confidence intervals from a parametric analysis

Analyze

8 Data conclusion
Formulate your conclusion with respect to your research question using the results of your analysis.

9 Research conclusion
Present your general conclusions with respect to your main research question. Discuss how it fits in with previous research and how the conclusion might be affected by the details ot your research design.

Conclude

Figure 7.1: Research topics for week 7

| Agresti | onlinestatbook.com |
|---|---|
| - | - |
| Practice exercises are in the back of each chapter in the book. | |

## 7.1 Homework exercises

*Give a full answer to each question, justifying your answer and showing work, if necessary.*

1. A university in the United States more famous for its basketball team than its academics claims that 80% of its basketball players graduate. In order to validate these claims, a study is conducted. A poll is taken of 20 players who played on the basketball team during the past years. Of these 20, 11 graduated and 9 did not. If the university is correct, then these 20 players should be a sample from a binomial distribution with parameters $N = 20$ and $p = 0.8$.

   (a) Let Y be the random variable representing the number of graduates found to have graduated in a sample of basketball players. Write the university's claim in formal terms. What is the probability of what we found in our sample ($Y = 11$, exactly) if the University is correct?

   (b) What is the probability that we would find 11 or fewer graduates in our sample (i.e. $Y \leq 11$)?

   (c) Based on your answer to problem 1b, what do you think about the university's claim?

2. According to demographic data, 12% of rural Dutch children under the age of 6 live in families with an income under the official poverty line. In order to study rural childhood learning, a random sample of 100 children under the age of 6 is taken.

   (a) What is the expected number of children who live in poverty in our sample of 100?

(b) What is the standard deviation of the number of children who live in poverty in our sample of 100?

(c) Use the normal approximation to the binomial to compute the probability that at least 18 children in our sample live in poverty. Write this probability in formal terms, using the correct notation.

3. A college is considering implementing a requirement that all students take several years of a foreign language in order to graduate. The student newspaper decides to poll professors about the new plan. Suppose for the following questions that in reality, 60% of professors support the new requirement. Use correct notation in your answers.

(a) If the student newspaper polls 5 professors, what is the probability that a majority (3 or more) will support the requirement, assuming that they are randomly sampled?

(b) If the student newspaper polls 99 professors, what is the probability that a majority (50 or more) will support the requirement, assuming that they are randomly sampled?

4. The Wechsler Intelligence Scale for Children (WISC) is an intelligence test for children. Scores (represented by random variable X) on this test are normally distributed with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$. In the following questions, make sure you use proper formal notation.

(a) What is the probability that a randomly selected child obtains a score of greater than 121?

(b) Suppose that a random sample of 50 children is drawn, and they all take the WISC. What is the mean and standard deviation of the distribution of the sample mean $\bar{x}$ ?

(c) What is the probability that $\bar{x} > 121$ given the random sample of 50 children?

5. Suppose that scores on some test in a population are clearly not normal, but come from an unfamiliar, very skewed distribution with a mean of 80 and a standard deviation of 10. Answer the following questions, and explain your answers

   (a) Suppose that you draw a very large number of samples (say, $N$) with size $n$ from this distribution, and you compute the mean $\bar{x}$. What would you expect the distribution of these $\bar{x}$ samples to look like?

   (b) Will the mean of the $\bar{x}$ distribution be 80, $\frac{80}{N}$, or $\frac{80}{n}$?

   (c) Will the standard deviation of the $\bar{x}$ distribution be 10, $\frac{10}{N}$, $\frac{10}{\sqrt{N}}$, $\frac{10}{n}$, or $\frac{10}{\sqrt{n}}$?

**In-Class Exercises**

## 7.2 Probability and working with the binomial distribution

### 7.2.1 Probability and computing probability

1. Explain the similarity between flipping a coin and drawing a sample from a population.

2. Suppose you flip a coin twice, and note for each flip which side is up, heads (H) or tails (T).

   (a) What is the sample space S?

   (b) What is the probability of every element in S? In working out your answer, did you use the rule for independent events or the rule for disjoint events? Explain your answer.

   (c) Let event $A$ be defined as the event that the two flips are the same. Rewrite event $A$ in formal notation

   (d) What is the probability of event A? In working out your answer, did you use the rule for independent events or the rule for disjoint events? Explain your answer.

### 7.2.2 Random variables

For exercises 1 through 2, choose/mark the best answer.

1. A density curve can be used to determine the probability . . .
   a. . . . of getting a "head" when flipping a coin.
   b. . . . of an outcome from a discrete random variable.
   c. . . . of an outcome from a continuous random variable.
   d. . . . all of the above.

2. The law of large numbers is only appropriate for random variables that . . .
   a. . . . are distributed discretely.
   b. . . . are distributed continuously.
   c. . . . are distributed normally.
   d. . . . have a (finite) mean.

3. Describe in simple language what the law of large numbers implies.

## 7.3 More probability

1. It is known that 10% of the people in a certain population are HIV-positive. It is also known for a specific HIV test that if a person is HIV-positive, the test will give a positive result 90% of the time. When the person is HIV-negative, the test will give a positive result 20% of the time. What is the probability that a random person selected who has a positive test result actually has HIV? Show your work.

2. In the preceding problem, three numbers were given. Change one of these numbers until the chance that a randomly selected person with a positive test result actually has HIV is 9/10. Check that your answer is correct.

## 7.4 Working with the binomial distribution in R

If the number of successes in an experiment is binomially distributed, then you can use R to determine the probability of y successes out of N trials. Also, you can easily find the chance of getting y or fewer successes (the *cumulative probability*).

For example, suppose a basketball player is going to shoot 8 free-throws, and we know that this player scores a point on 50% of all his throws. What is the probability that the player will get exactly 3 points? Let us call the random variable that represents the number of successful shots Y. Then Y is binomially distributed with a probability of .5. In formal notation, $Y \sim B(N = 8, p = .5)$. The question then becomes: what is the probability that $Y = 3$ when $Y \sim B(N = 8, p = .5)$, or $P(Y = 3|Y \sim B(N = 8, p = .5))$.

We will now use R to compute the probability. In R, you can use *functions* to access the probabilities that so far, we have retrieved from tables. To get the probability of exactly X successes out of Y attempts, we will use the function "dbinom (# successes, # attempts, success rate)". For instance, if we type "dbinom (4, 10, .5)" in the console (without quotation marks), we ask R for the probability of exactly 4 successes out of 10 attempts when the success rate is .5.

To get the probability of X or fewer successes out of Y attempts, we will use the function "pbinom (# successes, # attempts, success rate)". For instance, if we type "pbinom (5, 8, .5)" in the console (without quotation marks), we ask R for the probability of 5 or fewer successes out of 8 attempts when the success rate is .5.

1. What is the probability that the basketball player in our example will get exactly 3 baskets out of 8 attempts?

2. What is the probability that the player will get 3 or fewer baskets in his 8 attempts? First, write the question in formal terms. Then, use R to compute it.

## 7.5 Approximating the binomial distribution using a normal distribution

In section 7.4, you learned how to compute binomial probabilities using R. This method gives exact probabilities. In some cases, however, it is helpful to *approximate* the binomial distribution with the normal distribution. Most of the methods that we will use in the future are designed for use with normally distributed data, so understanding how to approximate the binomial with the normal is very useful. Note that this approximation only works when the binomial number of trials, $N$, is more than about 20. The greater the number of trials, the better the approximation. The purpose of this section is to compare the probabilities we obtain using the binomial table, using the binomial distribution in R, and using the normal approximation.

So far, we have looked up probabilities for the normal approximation in Table A. For instance, in the previous section with the example of the basketball player, the normal approximation for getting 3 or fewer baskets out of 8 attempts when the success rate is .5 starts by working out the mean ($np = 4$) and the standard deviation ($\sqrt{(np(1-p))} \approx 1.414$, then the $Z$-score is $\frac{3-4}{1.414} \approx -0.7071$. Looking up $p(Z \leq -.7071)$ in Table A yields $p = .2389$. We can get that probability in R by using the function "pnorm". If we type "pnorm (-.7071)" in the console (without quotation marks), we get $p = .2397501$, a more precise answer than the table provides. For the next exercise, you can execute the normal approximation with the help of Table A or with the help of R, your choice.

1. Fill in the following table:

| Probability | Using the binomial table | Using "pbinom" | Using normal approx. |
|---|---|---|---|
| $P(Y \leq 4 \mid Y \sim Binomial(N = 10, p = 0.1)$ | ____ | ____ | ____ |
| $P(Y \leq 4 \mid Y \sim Binomial(N = 10, p = 0.5)$ | ____ | ____ | ____ |
| $P(Y \leq 4 \mid Y \sim Binomial(N = 20, p = 0.1)$ | ____ | ____ | ____ |
| $P(Y \leq 4 \mid Y \sim Binomial(N = 20, p = 0.5)$ | ____ | ____ | ____ |

2. Compare the probabilities you calculated using the binomial table with the probabilities you computed using "pbinom" in R. Are they different? If so, why?

3. Compare the probabilities you calculated using "pbinom" with the probabilities you calculated using the normal approximation.

   (a) Why are these values different?

(b) The differences are relatively small when the probabilities are extreme, but we see that for the second problem, the normal approximation is off by quite a bit. Why do you think that is?

(c) Which column of values is the most accurate? Why?

**New R functions**

- *dbinom(x,y,z)* Gives density for value $x$ under a binomial distribution with size $y$ and success rate $z$.
- *pbinom(x,y,z)* Gives cumulative distribution for value $x$ under a binomial distribution with size $y$ and success rate $z$.
- *pnorm(x,y,z)* Gives cumulative distribution for value $x$ under a normal distribution with mean $y$ and sd $z$.

# List of Figures

# List of Tables

# Bibliography

[Anscombe, 1973] Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27:17–21. 3.2.1

[Efron and Tibshirani, 1993] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London. 1

[Mackowiak et al., 1992] Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992). A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association*, 268:1578–1580. 1.3

[Moore et al., 2017] Moore, D. S., McCabe, G. P., and Craig, B. A. (2017). *Introduction to the practice of statistics, ninth edition*. W. H. Freeman and Company, New York. 5, 6

[Morey et al., 2008] Morey, R. D., Rouder, J. N., and Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, 52:21–36. 4.2.3

[Nauta et al., 2003] Nauta, M., Scholing, A., Emmelkamp, P., and Minderaa, R. (2003). Cognitive-behavioral therapy for children with anxiety disorders in a clinical setting: No additional effect of a cognitive parent training. *Journal of the American Academy of Child and Adolescent Psychiatry*. 5.2.2

[Rossman, 1994] Rossman, A. (1994). Televisions, physicians, and life expectancy. *Journal of Statistics Education*, 2. 4.2.1

[Rouncefield, 1995] Rouncefield, M. (1995). The statistics of poverty and inequality. *Journal of Statistics Education*. 1.4

[Tufte, 2001] Tufte, E. R. (2001). *The Visual Display of Quantitative Information, 2nd edition*. Graphics Press. 2.2