

David J. Stanley

PSYC6060 Course Notes

To my students,
from whom I've learn so much about teaching.

Contents

List of Tables	vii
List of Figures	ix
Preface	xi
About the Author	xv
1 Introduction	1
1.1 A focus on workflow	1
1.2 R works with plug-ins	1
1.3 Create an account at R Studio Cloud	2
1.4 Join the class workspace	2
1.5 Exploring the R Studio Interface	4
1.5.1 Console panel	4
1.5.2 Script Panel	5
1.6 Writing your first script	6
1.6.1 Create the script file	6
1.6.2 Add a comment to your script	6
1.6.3 Background about the tidyverse	7
1.6.4 Add library(tidyverse) to your script	8
1.6.5 Activate tidyverse auto-complete for your script	8
1.7 Loading your data	9
1.7.1 Use read_csv (not read.csv) to open files.	9
1.8 Checking out your data	10
1.8.1 view(): See a spreadsheet view of your data	10
1.8.2 print(): See you data in the Console	10
1.8.3 head(): Check out the first few rows of data	11
1.8.4 tail(): Check out the last few rows of data	12
1.8.5 summary(): Quick summaries	12
1.9 Run <i>vs.</i> Source with Echo <i>vs.</i> Source	13
1.9.1 Run select text	14
1.9.2 Source (without Echo)	14
1.9.3 Source with Echo	14
1.10 Trying Source with Echo	15
1.11 A few key points about	17
1.11.1 Lists	18

1.12 That's it!	19
2 Handling Data with the Tidyverse	21
2.1 Required	21
2.2 Objective	21
2.3 Using the Console	22
2.4 Basic tidyverse commands	22
2.4.1 select()	23
2.4.2 summarise()	25
2.4.3 filter()	26
2.4.4 group_by()	28
2.4.5 mutate()	29
2.5 Advanced tidyverse commands	29
2.5.1 select()	31
2.5.2 summarise()	34
2.5.3 mutate()	36
2.6 Using help	40
2.7 Base R vs tidyverse	41
2.7.1 Tibbles vs. data frames	41
2.7.2 read.csv() and data frames	42
2.7.3 read_csv() and tibbles	43
3 Making your data ready for analysis	45
3.1 Required Packages	45
3.2 Objective	46
3.3 Context	46
3.4 Begin with the end in mind	48
3.4.1 Structuring data: Obtaining tidy data	49
3.5 Data collection considerations	52
3.5.1 Naming conventions	53
3.5.2 Likert-type items	53
3.5.3 ANOVA between	55
3.5.4 ANOVA within	56
3.6 Following the examples	56
3.7 Entering data into spreadsheets	58
3.8 Experiment: Between	59
3.8.1 Creating factors	60
3.8.2 Factor screening	62
3.8.3 Numeric screening	63
3.9 Experiment: Within one-way	64
3.9.1 Creating factors	65
3.9.2 Factor screening	67
3.9.3 Numeric screening	68
3.9.4 Pivot to tidy data	69
3.10 Experiment: Within N-way	71

3.10.1	Creating factors	72
3.10.2	Factor screening	74
3.10.3	Numeric screening	75
3.10.4	Pivot to tidy data	76
3.11	Surveys: Single Occassion	78
3.11.1	Creating factors	80
3.11.2	Factor screening	82
3.11.3	Numeric screening	83
3.11.4	Scale scores	85
3.12	Surveys: Multiple Occasions	89
3.12.1	Creating factors	91
3.12.2	Factor screening	94
3.12.3	Numeric screening	95
3.12.4	Scale scores	97
3.12.5	Pivot to tidy data	100
3.13	Basic descriptive statistics	102
3.13.1	skim()	102
3.13.2	apa.cor.table()	103
3.13.3	tidyverse	103
3.13.4	Cronbach's alpha	105
4	Populations	107
4.1	Notation	107
4.2	Population vs samples	109
4.3	A small population	110
4.3.1	Mean (μ)	110
4.3.2	Variance (σ^2)	112
4.3.3	Standard Deviation (σ)	114
4.4	Visualizing populations	115
4.5	Population comparisons	115
4.5.1	Standardized units	116
4.5.2	Cohen's d units	117
4.5.3	Cohen's d advantages	119
4.5.4	Cohen's d caveats	119
4.5.5	Benchmarks	119
4.6	Key points	122
5	Samples	125
5.1	Notation	125
5.2	Estimating the mean	126
5.3	Estimating variance	130
5.4	Estimating standard deviation	133
5.5	Estimating SMD	133
5.5.1	Population SMD	133
5.5.2	Sample SMD	134

5.6 Overview	136
Appendix	137
A More to Say	137
Bibliography	139
Index	141

List of Tables

1.1	R packages are similar to smart phone apps (Kim, 2018) . . .	2
3.3	Between participant data entered one row per participant . .	50
3.4	Within participant data entered one row per participant . . .	50
3.5	A tidy data version of the within participant data	51



List of Figures

1.1	Screen shot of workspace join message.	3
1.2	Screen shot of welcome message	3
1.3	Screen shot of starting first assignment	3
1.4	R Studio interface	4
1.5	Source button options	14
3.1	Open science in an article header	47
3.2	Data science pipeline by Roger Peng.	49
3.3	Spreadsheet entry of running data	58
3.4	Text view of CSV data	59
3.5	Word document created by apa.cor.table	103
4.1	Data for understanding summation notation	107
4.2	Variance as a fit index for the mean	111
4.3	The difference between two population means can be expressed in the original units as indicated by ΔM . Alternatively, the difference can be expressed using a Standardized Mean Difference (SMD). The SMD index is also known as the population-level d -value and is represented by the symbol δ . The SMD is a way of expressing the difference between population means without using the original units.	118
4.4	An advantage of using the Standardized Mean Difference (SMD) to index the difference between two population means (i.e., δ) is that it takes the population standard deviation into account. In these three examples, the difference between the populations means is the same using the original/raw units of centimeters. However, the standard deviation of the populations varies across scenarios A, B, and C. The SMD illustrates that these three scenarios are different. If you only examined the difference in the original units (i.e., ΔM) you would conclude the effect is the same across the three scenarios. However, by using SMD, indexed by δ - the population d -value, you see that the effect is progressively stronger from scenario A, to B, to C. This is illustrated by the fact that there is progressively less overlap between the distributions as you move from scenario A to C.	120

4.5	The three scenarios in this figure illustrate that a Standardized Mean Difference (i.e., population d -value or δ) can remain constant across scenarios when there is a change in the raw difference (i.e., ΔM) between the population means. This SMD is consistent across the three scenario despite a change in the mean difference using original units; this occurs because the standard deviations also changes across the three scenarios. .	121
4.6	Cohen's (1988) effect size benchmarks.	122
5.1	Illustration of the standardized mean difference of 1.50 for height between males and females. The solid black vertical lines indicates the means (μ) for the two populations.	134
5.2	The male sample variance ($n-1$) is an estimate of male population variance. Likewise, the female sample variance ($n-1$) is an estimate of the female population variance.	135
5.3	We assume the population variances are the same. Therefore, the male and female sample variances are both estimates of the same population variance.	135
5.4	We create a single estimate of the population variance called pooled variance (s_{pooled}^2). When sample sizes are equal, the pooled variance is just the average of the two sample variances (both using $n-1$). When the sample sizes are unequal, we need to use a more sophisticated formula to obtain the pooled variance. .	136

Preface

R's emerging role in psychology will be described here..

Structure of the book

I'll put more about the structure of the book here in the future.

Software information and conventions

I used the **knitr** package (Xie, 2015) and the **bookdown** package (Xie, 2020) to compile my book. My R session information is shown below:

```
xfun::session_info()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.6, RStudio 1.3.1056
##
## Locale: en_CA.UTF-8 / en_CA.UTF-8 / en_CA.UTF-8 / C / en_CA.UTF-8 / en_CA.UTF-8
##
## Package version:
##   abind_1.4.5             acepack_1.4.1
##   arm_1.11.1             askpass_1.1
##   assertthat_0.2.1       backports_1.1.8
##   base64enc_0.1-3        BH_1.72.0.3
##   bookdown_0.19.1        boot_1.3.25
##   broom_0.7.0.9000       callr_3.4.3
##   cellranger_1.1.0       checkmate_2.0.0
##   cli_2.0.2              clipr_0.7.0
```

##	cluster_2.1.0	coda_0.19.3
##	colorspace_1.4-1	compiler_4.0.2
##	crayon_1.3.4	curl_4.3
##	data.table_1.12.8	DBI_1.1.0
##	dbplyr_1.4.3	desc_1.2.0
##	digest_0.6.25	dplyr_1.0.0
##	ellipsis_0.3.1	evaluate_0.14
##	fansi_0.4.1	farver_2.0.3
##	forcats_0.5.0	foreign_0.8-80
##	Formula_1.2-3	fs_1.4.1
##	generics_0.0.2	ggplot2_3.3.2
##	glue_1.4.1	graphics_4.0.2
##	grDevices_4.0.2	grid_4.0.2
##	gridExtra_2.3	gsl_2.1.6
##	gtable_0.3.0	haven_2.3.1
##	highr_0.8	Hmisc_4.4-0
##	hms_0.5.3	htmlTable_2.0.0
##	htmltools_0.4.0	htmlwidgets_1.5.1
##	httr_1.4.1	isoband_0.2.2
##	janitor_2.0.1	jpeg_0.1-8.1
##	jsonlite_1.6.1	knitr_1.28
##	labeling_0.3	lattice_0.20-41
##	latticeExtra_0.6-29	lavaan_0.6.6
##	learnSampling_0.2	lifecycle_0.2.0
##	lme4_1.1.23	lubridate_1.7.8
##	magrittr_1.5	markdown_1.1
##	MASS_7.3.51.6	Matrix_1.2-18
##	matrixcalc_1.0.3	MBESS_4.7.0
##	methods_4.0.2	mgcv_1.8.31
##	mi_1.0	mime_0.9
##	minqa_1.2.4	mnormt_2.0.0
##	modelr_0.1.8	munsell_0.5.0
##	mvtnorm_1.1.1	nlme_3.1-148
##	nloptr_1.2.2.1	nnet_7.3-14
##	numDeriv_2016.8.1.1	OpenMx_2.17.4
##	openssl_1.4.1	packrat_0.5.0
##	parallel_4.0.2	pbapply_1.4.2
##	pbivnorm_0.6.0	pillar_1.4.4
##	pkgbuild_1.0.8	pkgconfig_2.0.3
##	pkgload_1.1.0	png_0.1-7
##	praise_1.0.0	predictionInterval_1.0.0
##	prettyunits_1.1.1	processx_3.4.2
##	progress_1.2.2	ps_1.3.3
##	psych_1.9.12.31	purrr_0.3.4
##	R6_2.4.1	RColorBrewer_1.1-2

## Rcpp_1.0.4.6	RcppEigen_0.3.3.7.0
## RcppParallel_5.0.2	readr_1.3.1
## readxl_1.3.1	rematch_1.0.1
## repr_1.1.0	reprex_0.3.0
## rlang_0.4.6	rmarkdown_2.2
## rpart_4.1-15	rpf_1.0.4
## rprojroot_1.3.2	rsconnect_0.8.16
## rstudioapi_0.11	rvest_0.3.5
## scales_1.1.1	selectr_0.4.2
## sem_3.1.11	semTools_0.5.3
## skimr_2.1.1	snakecase_0.11.0
## splines_4.0.2	StanHeaders_2.21.0.5
## statmod_1.4.34	stats_4.0.2
## stats4_4.0.2	stringi_1.4.6
## stringr_1.4.0	survival_3.1-12
## sys_3.3	testthat_2.3.2
## tibble_3.0.1	tidyr_1.1.0
## tidyselect_1.1.0	tidyverse_1.3.0
## tinytex_0.23	tmvnsim_1.0-2
## tools_4.0.2	utf8_1.1.4
## utils_4.0.2	vctrs_0.3.1
## viridis_0.5.1	viridisLite_0.3.0
## whisker_0.4	withr_2.2.0
## xfun_0.14	xml2_1.3.2
## yaml_2.2.1	

Package names are in bold text (e.g., **rmarkdown**), and inline code and filenames are formatted in a typewriter font (e.g., `knitr::knit('foo.Rmd')`). Function names are followed by parentheses (e.g., `bookdown::render_book()`).



About the Author

David J. Stanley is an Associate Professor of Industrial and Organizational Psychology at the University of Guelph in Canada. He obtained his PhD from Western University in London, Ontario. David has published articles in *Advances in Methods and Practices in Psychological Science*, *Organizational Research Methods*, *Journal of Applied Psychology*, *Perspectives in Psychological Science*, *Journal of Business and Psychology*, *Journal of Vocational Behaviour*, *Journal of Personality and Social Psychology*, *Behavior Research Methods*, *Industrial and Organizational Psychology*, and *Emotion* among other journals. David also created the `apaTables` R package.



1

Introduction

Welcome! In this guide, we will teach you about statistics using the statistical software R with the interface provided by R Studio. The purpose of this chapter is to provide you with a set of activities that get you up-and-running in R quickly so get a sense of how it works. In later chapters we will revisit these same topics in more detail.

1.1 A focus on workflow

An important part of this guide is training you in a workflow that will avoid many problems than can occur when using R.

1.2 R works with plug-ins

R is a statistical language with many plug-ins called **packages** that you will use for analyses. You can think of R as being like your smartphone. To do things with your phone you need **an App** (R equivalent: a *package*) from the App Store (R equivalent: *CRAN*). Apps need to be **downloaded** (R equivalent: *install.packages*) before you can use them. To use the app you need **Open** it (R equivalent: *library command*). These similarities are illustrated in Table 1.1 below.

TABLE 1.1: R packages are similar to smart phone apps (Kim, 2018)

Smart Phone Terminology	R Terminology
App	package
App Store	CRAN
Download App from App Store	<code>install.packages("apaTables", dependencies = TRUE)</code>
Open App	<code>library("apaTables")</code>

1.3 Create an account at R Studio Cloud

R Studio Cloud¹ accounts are free and required for this guide. Please go to the website and set up a new account.

1.4 Join the class workspace

To do the assignment required for this class you need to join the class workspace on R Studio Cloud. To do so:

1. Log into R Studio Cloud (if you haven't already done so).
2. Go to your university email account and find the message with the subject "R Studio Workspace Invitation". In this message there is a link to the class R Studio Cloud workspace.
3. Click on the workspace link in the email or paste it into your web browser. You should see a screen like the one below in Figure 1.1. Click on the Join button.
4. Then you should see the welcome message illustrated in Figure 1.2. Above this message is the Projects menu option. Click on the word Project.
5. You should now see the First Project displayed as in 1.3. Click the Start button. You will then move to a view of R Studio.

¹<http://www.rstudio.cloud>

1.4 Join the class workspace

3



FIGURE 1.1: Screen shot of workspace join message.



FIGURE 1.2: Screen shot of welcome messag



FIGURE 1.3: Screen shot of starting first assignment

5. In R Studio it is essential you use projects to keep your files organized and in the same spot. For this course, when you start an assignment on R Studio Cloud and the project will already have been made for you. Later you will learn to make your own R Studio Projects.

1.5 Exploring the R Studio Interface

Once you have opened (or created) a Project folder, you are presented with the R Studio interface. There are a few key elements to the user interface that are illustrated in Figure 1.4 In the lower right of the screen you can see the a panel with several tabs (i.e., Files, Plots, Packages, etc) that I will refer to as the Files pane. You look in this pane to see all the files associated with your project. On the left side of the screen is the Console which is an interactive pane where you type and obtain results in real time. I've placed two large grey blocks on the screen with text to more clearly identify the Console and Files panes. Not shown in this figure is the Script panel where we can store our commands for later reuse.



FIGURE 1.4: R Studio interface

1.5.1 Console panel

When you first start R, the Console panel is on the left side of the screen. Sometimes there are two panels on the left side (one above the other); if so, the Console panel is the lower one (and labeled accordingly). We can use R

a bit like a calculator. Try typing the following into the Console window: $8 + 6 + 7 + 5$. You can see that R immediately produced the result on a line preceded by two hashtags (`##`).

```
8 + 6 + 7 + 5
```

```
## [1] 26
```

We can also put the result into a variable to store it. Later we can use the `print` command to see that result. In the example below we add the numbers 3, 0, and 9 and store the result in the variable `my_sum`. The text “`<-`” indicate you are putting what is on the right side of the arrow into the variable on the left side of the arrow. You can think of a variable as cup into which you can put different things. In this case, imagine a real-world cup with `my_sum` written on the outside and inside the cup we have stored the sum of 3, 0, and 9 (i.e., 12).

```
my_sum <- 3 + 0 + 9
```

We can inspect the contents of the `my_sum` variable (i.e., `my_sum` cup) with the `print` command:

```
print(my_sum)
```

```
## [1] 12
```

Variables are very useful in R. We will use them to store a single number, an entire data set, the results of an analysis, or anything else.

1.5.2 Script Panel

Although you can use R with just with the Console panel, it's a better idea to use scripts via the Script panel - not visible yet. Scripts are just text files with the commands you use stored in them. You can run a script (as you will see below) using the Run or Source buttons located in the top right of the Script panel.

Scripts are valuable because if you need to run an analysis a second time you don't have to type the command in a second time. You can run the script again and again without retyping your commands. More importantly though, the script provides a record of your analyses.

A common problem in science is that after an article is published, the authors can't reproduce the numbers in the paper. You can read more about the

important problem in a surprising article in the journal *Molecular Brain*². In this article an editor reports how a request for the data underlying articles resulted in the wrong data for 40 out of 41 papers. Long story short – keep track of the data and scripts you use for your paper. In a later chapter, it’s generally poor practice to manipulate or modify or analyze your data using any menu driven software because this approach does not provide a record of what you have done.

1.6 Writing your first script

1.6.1 Create the script file

Create a script in your R Studio project by using the menu File > New File > R Script.

Save the file with an appropriate name using the File menu. The file will be saved in your Project folder. A common, and good, convention for naming is to start all script names with the word “script” and separate words with an underscore. You might save this first script file with the name “script_my_first_one.R”. The advantage of beginning all script files with the word script is that when you look at your list of files alphabetically, all the script files will cluster together. Likewise, it’s a good idea to save all data files such that they begin with “data_”. This way all the data files will cluster together in your directory view as well. You can see there is already a data file with this convention called “data_okcupid.csv”.

You can see as discussed previously, we are trying to instill an effective workflow as you learn R. Using a good naming convention (that is consistent with what others use) is part of the workflow. When you write your scripts it’s a good idea to follow the tidyverse style guide³ for script names, variable name, file names, and more.

1.6.2 Add a comment to your script

In the previous section you created your first script. We begin by adding a comment to the script. A comment is something that will be read by humans rather than the computer/R. You make comments for other people that will read your code and need to understand what you have done. However, realize

²<https://molecularbrain.biomedcentral.com/articles/10.1186/s13041-020-0552-2>

³<https://style.tidyverse.org>

that you are also making comments for your future self as illustrated in an XKCD cartoon⁴.

A good way to start every script is with a comment that includes the date of your script (or even better when you installed your packages, more on this later). Like smartphone apps, packages are updated regularly. Sometimes after a package is updated it will no longer work with an older script. Fortunately, the checkpoint package⁵ lets users role back the clock and use older versions of packages. Adding a comment with the date of your script will help future users (including you) to use your script with the same version of the package used when you wrote the script. Dating your script is an important part of an effective and reproducible workflow.

```
# Code written on: YYYY/MM/DD
# By: John Smith
```

Note that in the above comment I used the internationally accepted date format order Year/Month/Day. Some people use the mnemonic *Your My Dream* to remember this order. Wikipedia provides more information about the Internationally Date Format ISO 8601⁶.

Moving forward, I suggest you use comments to make your own personal notes in your own code as your write it.

1.6.3 Background about the tidyverse

There are generally two broad ways of using R, the older way and the newer way. Using R the older way is referred to as using base R. A more modern approach to using R is the tidyverse. The tidyverse represents a collection of packages the work together to give R a modern workflow. These packages do many things to help the data analyst (loading data, rearranging data, graphing, etc.). We will use the tidyverse approach to R in this guide.

As noted the tidyverse is a collection of packages. Each package adds new commands to R. The number of packages and correspondingly the number of new commands added to R by the tidyverse is large. Below is a list of the tidyverse packages:

```
## [1] "broom"      "cli"        "crayon"     "dbplyr"
## [5] "dplyr"      "forcats"    "ggplot2"    "haven"
## [9] "hms"        "httr"       "jsonlite"   "lubridate"
## [13] "magrittr"   "modelr"     "pillar"     "purrr"
```

⁴<https://xkcd.com/1421/>

⁵<https://cran.r-project.org/web/packages/checkpoint/index.html>

⁶https://en.wikipedia.org/wiki/ISO_8601

```
## [17] "readr"      "readxl"      "reprex"      "rlang"  
## [21] "rstudioapi" "rvest"       "stringr"     "tibble"  
## [25] "tidyr"      "xml2"        "tidyverse"
```

Before you can use a package it needs to be installed – this is the same as downloading an app from the App Store. Normally, you can install a **single** packages with the `install.packages` command. Previously, you needed run an `install.package` command for every package in the tidyverse as illustrated below (though we no longer use this approach).

```
# The old way of installing the tidyverse packages  
# Like downloading apps from the app store  
  
install.packages("broom", dep = TRUE)  
install.packages("cli", dep = TRUE)  
install.packages("ggplot", dep = TRUE)  
# etc
```

Fortunately, the tidyverse packages can now be installed with a single `install.packages` command. Specifically, the `install.packages` command below will install all of the packages listed above.

Class note: For the “First Lab”, I’ve done the `install.packages` for you. So there is no need to use the `install.packages` command below in this first lab.

```
install.packages("tidyverse", dep = TRUE)
```

1.6.4 Add `library(tidyverse)` to your script

The tidyverse is now installed, so we need to activate it. We do that with the `library` command. Put the library line below at the top of your script file (below your comment):

```
# Code written on: YYYY/MM/DD  
# By: John Smith  
library(tidyverse)
```

1.6.5 Activate tidyverse auto-complete for your script

Select the `library(tidyverse)` text with your mouse/track-pad so that it is highlighted. Then click the Run button in the upper right of the Script panel.

Doing this “runs” the selected text. After you click the Run button you should see text like the following the Console panel:

When you use `library(tidyverse)` to activate the tidyverse you activate the most commonly used subset of the tidyverse packages. In the output you see checkmarks beside names of the tidyverse packages you have activated.

By activating these packages you have added new commands to R that you will use. Sometimes these packages replace older versions of commands in R. The “Conflicts” section in the output shows you where the packages you activated replaced older R commands with newer R commands. You can activate the other tidyverse package by running a library command for each package – if needed. No need to do so now.

Most importantly, running the `library(tidyverse)` prior to entering the rest of your script allows R Studio to present auto-complete options when typing your text. Remember to start each script with the `library(tidyverse)` command and then Run it so you get the autocomplete options for the rest of the commands your enter.

1.7 Loading your data

1.7.1 Use `read_csv` (not `read.csv`) to open files.

If you inspect the Files pane on the right of the screen you see the **data_okcupid.csv** data file in our project directory. We will load this data with the commands below. If you followed the steps above, you should have auto-complete for the tidyverse commands you type for now in – in the current R session. Enter the command below into your script. As your start to type `read_csv` you will likely be presented with an auto-complete option. You can use the arrow keys to move up and down the list of options to select the one you want - then press tab to select it.

Once your command looks like the one below select the text and click on the “Run” button.

Note: If you are not in the class, the data file is available from: <https://github.com/dstanley4/psyc6060bookdown>

```
okcupid_profiles <- read_csv(file = "data_okcupid.csv")
```

```
## Parsed with column specification:  
## cols(
```

```
## age = col_double(),  
## diet = col_character(),  
## height = col_double(),  
## pets = col_character(),  
## sex = col_character(),  
## status = col_character()  
## )
```

The output indicates that you have loaded a data file and the type of data in each column. The sex column is of type `col_character` which indicates it contains text/letters. Most of the columns are of the type character. The age and height columns contain numbers are correspondingly indicated to be the type `col_double`. The label `col_double` indicates that a column of numbers represented in R with high precision⁷. There are other ways of representing numbers in R but this is the type we will see/use most often.

1.8 Checking out your data

There many ways of viewing the actual data you loaded. A few of these are illustrated now.

1.8.1 `view()`: See a spreadsheet view of your data

You can inspect your data in a spreadsheet view by using the `view` command. Do NOT add this command to your script file – EVER. Adding it to the script can cause substantial problems. Type this command in the Console.

```
view(okcupid_profiles)
```

1.8.2 `print()`: See you data in the Console

You can inspect the first few rows of your data with the `print()` command. It is OK to add a print command to your script. Try the `print()` command below in the Console:

⁷https://en.wikipedia.org/wiki/Double-precision_floating-point_format

```
print(okcupid_profiles)
```

```
## # A tibble: 59,946 x 6
##   age diet      height pets      sex status
##   <dbl> <chr>    <dbl> <chr>    <chr> <chr>
## 1  22 strictly an~  75 likes dogs and l~ m   single
## 2  35 mostly other  70 likes dogs and l~ m   single
## 3  38 anything      68 has cats      m   availa~
## 4  23 vegetarian    71 likes cats      m   single
## 5  29 <NA>          66 likes dogs and l~ m   single
## 6  29 mostly anyt~  67 likes cats      m   single
## 7  32 strictly an~  65 likes dogs and l~ f   single
## 8  31 mostly anyt~  65 likes dogs and l~ f   single
## 9  24 strictly an~  67 likes dogs and l~ f   single
## 10 37 mostly anyt~  65 likes dogs and l~ m   single
## # ... with 59,936 more rows
```

1.8.3 head(): Check out the first few rows of data

You can inspect the first few rows of your data with the `head()` command. Try the command below in the Console:

```
head(okcupid_profiles)
```

```
## # A tibble: 6 x 6
##   age diet      height pets      sex status
##   <dbl> <chr>    <dbl> <chr>    <chr> <chr>
## 1  22 strictly any~  75 likes dogs and l~ m   single
## 2  35 mostly other  70 likes dogs and l~ m   single
## 3  38 anything      68 has cats      m   availa~
## 4  23 vegetarian    71 likes cats      m   single
## 5  29 <NA>          66 likes dogs and l~ m   single
## 6  29 mostly anyt~  67 likes cats      m   single
```

You can be even more specific and indicate you only want the first three row of your data with the `head()` command. Try the command below in the Console:

```
head(okcupid_profiles, 3)
```

```
## # A tibble: 3 x 6
##   age diet      height pets      sex status
##   <dbl> <chr>    <dbl> <chr>    <chr> <chr>
```

```
## 1    22 strictly any~    75 likes dogs and l~ m    single
## 2    35 mostly other    70 likes dogs and l~ m    single
## 3    38 anything        68 has cats            m    availa~
```

1.8.4 tail(): Check out the last few rows of data

You can inspect the last few rows of your data with the `tail()` command. Try the command below in the Console:

```
tail(okcupid_profiles)
```

```
## # A tibble: 6 x 6
##   age diet      height pets      sex  status
##   <dbl> <chr>    <dbl> <chr>    <chr> <chr>
## 1   31 <NA>      62 likes dogs      f    single
## 2   59 <NA>      62 has dogs      f    single
## 3   24 mostly anyt~  72 likes dogs and lik~ m    single
## 4   42 mostly anyt~  71 <NA>          m    single
## 5   27 mostly anyt~  73 likes dogs and lik~ m    single
## 6   39 <NA>      68 likes dogs and lik~ m    single
```

You can be even more specific and indicate you only want the last three row of your data with the `tail()` command. Try the command below in the Console:

```
tail(okcupid_profiles, 3)
```

```
## # A tibble: 3 x 6
##   age diet      height pets      sex  status
##   <dbl> <chr>    <dbl> <chr>    <chr> <chr>
## 1   42 mostly anyt~  71 <NA>          m    single
## 2   27 mostly anyt~  73 likes dogs and lik~ m    single
## 3   39 <NA>      68 likes dogs and lik~ m    single
```

1.8.5 summary(): Quick summaries

You can a short summary of your data with the `summary()` command. Note that we will use the `summary()` command in many places in the guide. The output of the `summary()` command changes depending on what you give it - that is put inside the brackets. You can give the `summary()` command many things such as data, the results of a regression analysis, etc.

Try the command below in the Console. You will see that `summary()` give the mean and median for each of the numeric variables (age and height).

```
summary(okcupid_profiles)
```

```
##      age      diet      height
## Min.   : 18.0   Length:59946   Min.    : 1.0
## 1st Qu.: 26.0   Class :character   1st Qu.:66.0
## Median : 30.0   Mode  :character   Median :68.0
## Mean   : 32.3                      Mean   :68.3
## 3rd Qu.: 37.0                      3rd Qu.:71.0
## Max.   :110.0                      Max.    :95.0
##                                     NA's    :3
##      pets      sex      status
## Length:59946   Length:59946   Length:59946
## Class :character Class :character Class :character
## Mode  :character Mode  :character Mode  :character
##
##
##
```

1.9 Run vs. Source with Echo vs. Source

There are different ways of running commands in R. So far you have used two of these. You can enter them into the Console as we have done already. Or you can put them in your script select the text and click the Run button. There are four ways of running commands in your script.

You can:

1. Console: Enter commands directly
2. Script: Select the command(s) and press the Run button.
3. Script: Source (Without Echo)
4. Script: Source With Echo

Two of these approaches involve using the Source button, see Figure 1.5. You bring up the options for the Source button, illustrated in this figure, by clicking on the small arrow to the right of the word Source.



FIGURE 1.5: Source button options

1.9.1 Run select text

The Run button will run the text you highlight and present the relevant output. You have used this command a fair amount already.

I strongly suggest you **ONLY** use the Run button when testing a command to make sure it works or to debug a script. Or to run `library(tidyverse)` as you start working on your script so that you get the autocomplete options.

In general, you should always try to execute your R Scripts using the Source with Echo command (preceded by a Restart, see below). This ensures your script will work beginning to end for you in the future and for others that attempt to use it. Using the Run button in an ad lib basis can create output that is not reproducible.

1.9.2 Source (without Echo)

Source (without Echo) is not designed for the typical analysis workflow. It is mostly helpful when you run simulations. When you run Source (without Echo) much of the output you would wish to read is suppressed. In general, avoid this option. If you use it, you often won't see what you want to see in the output.

1.9.3 Source with Echo

The Source with Echo command runs all of the contents of a script and presents the output in the R console. This is the approach you should use to running your scripts in most cases.

Prior to running Source with Echo (or just Source), it's always a good idea to restart R. This makes sure you clear the computer memory of any errors from any previous runs.

So you should do the following **EVERY** time you run your script.

1. Use the menu item: **Session > Restart R**

2. Click the down arrow beside the Source button, and click on Source With Echo

This will clear potentially problematic previous stats, run the script commands, and display the output in the Console. Moving forward we will use this approach for running scripts. Once you have used Source with Echo once, you can just click the Source button and it will use Source with Echo automatically (without the need to use the pull down option for selecting Source with Echo).



- Using Restart R before you run a script, or R code in general, is a critical workflow tip.

1.10 Trying Source with Echo

Put the `head()`, `tail()`, and `summary()` command we used previously into your script. Then save your script using the File > Save menu. Your script should appear as below.

```
# Code written on: YYYY/MM/DD
# By: John Smith
library(tidyverse)

okcupid_profiles <- read_csv(file = "data_okcupid.csv")

head(okcupid_profiles)

tail(okcupid_profiles)

summary(okcupid_profiles)
```

Now do the following:

1. Use the menu item: **Session > Restart R**
2. Click the down arrow beside the Source button, and click on Source With Echo

You should see the output below:

```
# Code written on: YYYY/MM/DD
# By: John Smith
library(tidyverse)

okcupid_profiles <- read_csv(file = "data_okcupid.csv")
```

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   diet = col_character(),
##   height = col_double(),
##   pets = col_character(),
##   sex = col_character(),
##   status = col_character()
## )
```

```
head(okcupid_profiles)
```

```
## # A tibble: 6 x 6
##   age diet          height pets          sex status
##   <dbl> <chr>        <dbl> <chr>        <chr> <chr>
## 1    22 strictly any~    75 likes dogs and l~ m   single
## 2    35 mostly other    70 likes dogs and l~ m   single
## 3    38 anything        68 has cats          m   availa~
## 4    23 vegetarian      71 likes cats        m   single
## 5    29 <NA>            66 likes dogs and l~ m   single
## 6    29 mostly anyth~    67 likes cats        m   single
```

```
tail(okcupid_profiles)
```

```
## # A tibble: 6 x 6
##   age diet          height pets          sex status
##   <dbl> <chr>        <dbl> <chr>        <chr> <chr>
## 1    31 <NA>            62 likes dogs        f   single
## 2    59 <NA>            62 has dogs          f   single
## 3    24 mostly anyt~    72 likes dogs and lik~ m   single
## 4    42 mostly anyt~    71 <NA>              m   single
## 5    27 mostly anyt~    73 likes dogs and lik~ m   single
## 6    39 <NA>            68 likes dogs and lik~ m   single
```

```
summary(okcupid_profiles)
```



```
##      age      diet      height
##  Min.   : 18.0   Length:59946   Min.    : 1.0
## 1st Qu.: 26.0   Class :character   1st Qu.:66.0
## Median : 30.0   Mode  :character   Median :68.0
## Mean   : 32.3                      Mean   :68.3
## 3rd Qu.: 37.0                      3rd Qu.:71.0
## Max.   :110.0                      Max.   :95.0
##                                     NA's   :3
##      pets      sex      status
## Length:59946   Length:59946   Length:59946
## Class :character Class :character Class :character
## Mode  :character Mode  :character Mode  :character
##
##
##
##
```

Congratulations you just ran your first script!

1.11 A few key points about

Sometimes you will need to send a command additional information. Moreover, that information often needs to be grouped together into a vector or a list before you can send it to the command. We'll learn more about doing so in the future but here is a quick over view of vectors and lists to provide a foundation for future chapters.

1.11.0.1 Vector of numbers

We can create a vector of only numbers using the “c” function - which you can think of as being short for “combine” (or concatenate). In the commands below we create a vector of a few even numbers called “even_numbers”.

```
even_numbers <- c(2, 4, 6, 8, 10)
```

```
print(even_numbers)
```

```
## [1]  2  4  6  8 10
```

We can obtain the second number in the vector using the following notation:

```
print(even_numbers[2])
```

```
## [1] 4
```

1.11.0.2 Vector of characters

We can also create vectors using only characters. Note that I use **SHIFT RETURN** after each comma to move to the next line.

```
favourite_things <- c("copper kettles",  
                      "woolen mittens",  
                      "brown paper packages")
```

```
print(favourite_things)
```

```
## [1] "copper kettles"      "woolen mittens"  
## [3] "brown paper packages"
```

As before, can obtain the second item in the vector using the following notation:

```
print(favourite_things[2])
```

```
## [1] "woolen mittens"
```

1.11.1 Lists

Lists are similar to vectors in that you can create them and access items by their numeric position. Vectors must be all characters or all numbers. Lists can be a mix of characters or numbers. Most importantly items in lists can be accessed by their label. Note that I use **SHIFT RETURN** after each comma to move to the next line in the code below.

```
my_list <- list(last_name = "Smith",  
                first_name = "John",  
                office_number = 1913)
```

```
print(my_list)
```

```
## $last_name
```

```
## [1] "Smith"
##
## $first_name
## [1] "John"
##
## $office_number
## [1] 1913
```

You can access an item in a list using double brackets:

```
print(my_list[2])
```

```
## $first_name
## [1] "John"
```

You can access an item in a list by its label/name using the dollar sign:

```
print(my_list$last_name)
```

```
## [1] "Smith"
```

```
print(my_list$office_number)
```

```
## [1] 1913
```

1.12 That's it!

Congratulations! You've reached the end of the introduction to R. Take a break, have a cookie, and read some more about R tomorrow!



2

Handling Data with the Tidyverse

2.1 Required

The following packages must be installed before starting this chapter.

Required Packages
tidyverse

The following data files are used in this chapter:

Required Data
data_okcupid.csv
data_experiment.csv

The files are available at: <https://github.com/dstanley4/psyc6060bookdown>

2.2 Objective

The objective of this chapter is to familiarize you with some key commands in the tidyverse. These commands are used in isolation of each other for the most part. In the next chapter we will use these commands in a more coordinated way as we load a data set and move it from raw data to data that is ready for analysis (i.e., analytic data). You can start this project by Starting the class assignment on R Studio Cloud that corresponds to the chapter name.

2.3 Using the Console

All of the commands in this chapter should be typed into the Console within R. If you see a command split over multiple lines, use SHIFT-RETURN (macOS) or SHIFT-ENTER (Windows) to move the next line that is part of the same command.

2.4 Basic tidyverse commands

If you inspect the Files tab on the lower-right panel in R Studio you will see the file `data_okcupid.csv`. The code below loads that file.

```
library(tidyverse)
okcupid_profiles <- read_csv("data_okcupid.csv")
```

You can see the first few rows of the data using the `print()` command. Each row presents a person whereas each column represents a variable. If you have a large number of columns you will only see the first several columns with this approach to viewing your data.

```
print(okcupid_profiles)
```

```
## # A tibble: 59,946 x 6
##   age diet      height pets      sex status
##   <dbl> <chr>    <dbl> <chr>    <chr> <chr>
## 1  22 strictly an~  75 likes dogs and l~ m single
## 2  35 mostly other  70 likes dogs and l~ m single
## 3  38 anything      68 has cats      m availa~
## 4  23 vegetarian    71 likes cats      m single
## 5  29 <NA>          66 likes dogs and l~ m single
## 6  29 mostly anyt~  67 likes cats      m single
## 7  32 strictly an~  65 likes dogs and l~ f single
## 8  31 mostly anyt~  65 likes dogs and l~ f single
## 9  24 strictly an~  67 likes dogs and l~ f single
## 10 37 mostly anyt~  65 likes dogs and l~ m single
## # ... with 59,936 more rows
```

But it's also helpful just to see a list of the columns in the data with the `glimpse()` command:

```
glimpse(okcupid_profiles)
```

```
## Rows: 59,946
## Columns: 6
## $ age      <dbl> 22, 35, 38, 23, 29, 29, 32, 31, 24, 37, ...
## $ diet     <chr> "strictly anything", "mostly other", "an...
## $ height   <dbl> 75, 70, 68, 71, 66, 67, 65, 65, 67, 65, ...
## $ pets     <chr> "likes dogs and likes cats", "likes dogs...
## $ sex      <chr> "m", "m", "m", "m", "m", "m", "f", "f", ...
## $ status   <chr> "single", "single", "available", "single..."
```

The `glimpse()` command is useful because it quickly allows you to see all of the columns. Moreover, it allows you to see the type for each column. Types were briefly discussed in the last chapter. Notice in the output beside each column name that some columns are labeled “dbl” which is short for double – a type of numeric column. Other columns are labeled “chr” which is short for character - meaning the columns contain characters. These designations will become important in the next chapter as we prepare data for analysis.

2.4.1 `select()`

The `select()` command allows you to obtain a subset of the columns in your data. The commands below can be used to obtain the age and height columns. You can read the command as: take the `okcupid_profiles` data and then select the age and height columns. The “%>%” symbol can be read as “and then”. You can see that this code prints out the data with just the age and height columns. Remember, use SHIFT-ENTER or SHIFT-RETURN to move to the next line in the block of code.

```
okcupid_profiles %>%
  select(age, height)
```

```
## # A tibble: 59,946 x 2
##   age height
##   <dbl> <dbl>
## 1    22     75
## 2    35     70
## 3    38     68
## 4    23     71
## 5    29     66
```

```
## 6      29      67
## 7      32      65
## 8      31      65
## 9      24      67
## 10     37      65
## # ... with 59,936 more rows
```

Of course, it's usually of little help to just print the subset of the data. It's better to store it in a new data. In the command below we store the resulting data in a new data set called `new_data`.

```
new_data <- okcupid_profiles %>%
  select(age, height)
```

The `glimpse()` command shows us that only the age and height columns are in `new_data`.

```
glimpse(new_data)
```

```
## Rows: 59,946
## Columns: 2
## $ age      <dbl> 22, 35, 38, 23, 29, 29, 32, 31, 24, 37, ...
## $ height   <dbl> 75, 70, 68, 71, 66, 67, 65, 65, 67, 65, ...
```

In the above example we indicated the columns we wanted to retain from the `okcupid_profiles` data using the `select()` command. However, we can also indicate the columns we want to drop from `okcupid_profiles` using a minus sign (-) in front of the columns we specify in the `select()` command.

```
new_data <- okcupid_profiles %>% select(-age, -height)
```

The `glimpse()` command shows us that we kept all the columns except the age and height columns when we created `new_data`.

```
glimpse(new_data)
```

```
## Rows: 59,946
## Columns: 4
## $ diet      <chr> "strictly anything", "mostly other", "an...
## $ pets      <chr> "likes dogs and likes cats", "likes dogs...
## $ sex        <chr> "m", "m", "m", "m", "m", "m", "f", "f", ...
## $ status     <chr> "single", "single", "available", "single..."
```


2.4.2 summarise()

The `summarise()` command can be used to generate descriptive statistics for a specified column. You can easily calculate column descriptive statistics using the corresponding commands for `mean()`, `sd()`, `min()`, `max()`, among others. In the example below we calculate the mean for the age column.

In the code below, `mean(age, na.rm = TRUE)`, indicates to R that it should calculate the mean of the age column. The `na.rm` indicates how missing values should be handled. The `na` stands for not available; in R missing values are classified as Not Available or NA. The `rm` stands for remove. Consequently, `na.rm` is asking: “Should we remove missing values when calculating the mean?” The `TRUE` indicates that yes, missing values should be removed when calculating the mean. The result of this calculation is placed into a variable labelled `age_mean`, though we could have used any label we wanted instead of `age_mean`. We see that the mean of the age column is, with rounding, 32.3.

```
okcupid_profiles %>%
  summarise(age_mean = mean(age, na.rm = TRUE))

## # A tibble: 1 x 1
##   age_mean
##   <dbl>
## 1     32.3
```

More than one calculation can occur in the same `summarise()` command. You can easily add the calculation for the standard deviation with the `sd()` command.

```
okcupid_profiles %>%
  summarise(age_mean = mean(age, na.rm = TRUE),
            age_sd = sd(age, na.rm = TRUE))

## # A tibble: 1 x 2
##   age_mean age_sd
##   <dbl>   <dbl>
## 1     32.3    9.45
```

Often this process does too much rounding. We can get more exact results by adding an `as.data.frame()` to the end of the commands.

```
okcupid_profiles %>%
  summarise(age_mean = mean(age, na.rm = TRUE),
            age_sd = sd(age, na.rm = TRUE)) %>%
  as.data.frame()
```

```
##   age_mean age_sd
## 1    32.34  9.453
```

2.4.3 filter()

The `filter()` command allows you to obtain a subset of the rows in your data. In the example below we create a new data set with just the males from the original data.

Notice the structure of the original data below in the `glimpse()` output. There is a column called `sex` that uses `m` and `f` to indicate male and female, respectively. Also notice that there are 59946 rows in the `okcupid_profiles` data.

```
glimpse(okcupid_profiles)
```

```
## Rows: 59,946
## Columns: 6
## $ age      <dbl> 22, 35, 38, 23, 29, 29, 32, 31, 24, 37, ...
## $ diet     <chr> "strictly anything", "mostly other", "an...
## $ height   <dbl> 75, 70, 68, 71, 66, 67, 65, 65, 67, 65, ...
## $ pets     <chr> "likes dogs and likes cats", "likes dogs...
## $ sex      <chr> "m", "m", "m", "m", "m", "m", "f", "f", ...
## $ status   <chr> "single", "single", "available", "single..."
```

We use the `filter` command to select a subset of the rows based on the content of any column. In this case the `sex` column is used to obtain a subset of the rows; the rows with the value “m” are obtained. Notice the double equals sign is used to indicate “equal to”. The reason a double equals sign is used here (instead of a single equals sign) is to distinguish it from the use of the single equals sign in the `summarise` command above. In the `summarise` command above, the single equal sign was used to indicate “assign to”. That is, assign to `age_mean` the mean of the column `age` after it is calculated. A single equals sign indicates “assign to” whereas a double equals sign indicates “is equal to”.

```
okcupid_males <- okcupid_profiles %>%
  filter(sex == "m")
```

We use `glimpse()` to inspect these all male data. Notice that only the letter `m` is in the `sex` column - indicating only males are in the data set. Also notice that there are 35829 rows in the `okcupid_males` data - fewer people because males are a subset of the total number of rows.

```
glimpse(okcupid_males)
```

```
## Rows: 35,829
## Columns: 6
## $ age      <dbl> 22, 35, 38, 23, 29, 29, 37, 35, 28, 24, ...
## $ diet     <chr> "strictly anything", "mostly other", "an...
## $ height   <dbl> 75, 70, 68, 71, 66, 67, 65, 70, 72, 72, ...
## $ pets     <chr> "likes dogs and likes cats", "likes dogs...
## $ sex      <chr> "m", "m", "m", "m", "m", "m", "m", "m", ...
## $ status   <chr> "single", "single", "available", "single..."
```

The filter command can be combined with the summarise command to get the descriptive statistics for males without the hassle of creating new data. This is again done using the %>% “and then” operator.

```
okcupid_profiles %>%
  filter(sex == "m") %>%
  summarise(age_mean = mean(age, na.rm = TRUE),
            age_sd = sd(age, na.rm = TRUE))
```

```
## # A tibble: 1 x 2
##   age_mean age_sd
##   <dbl>   <dbl>
## 1     32.0    9.03
```

We see that for the 35829 females the mean age is 32.0 and the standard deviation is 9.0.

Likewise, we can obtain the descriptive statistics for females with only a slight modification, changing m to f in the filter command:

```
okcupid_profiles %>%
  filter(sex == "f") %>%
  summarise(age_mean = mean(age, na.rm = TRUE),
            age_sd = sd(age, na.rm = TRUE))
```

```
## # A tibble: 1 x 2
##   age_mean age_sd
##   <dbl>   <dbl>
## 1     32.8   10.0
```

We see that for the 24117 females the mean age is 32.8 and the standard deviation is 10.0.

2.4.4 group_by()

The process we used with the filter command would quickly become onerous if we had many subgroups for a column. Consequently, it's often better to use the group() command to calculate descriptive statistics for the levels (e.g., male/female) of a variable. By telling the computer to group_by() sex the summarise command is run separately for every level of sex (i.e., m and f).

```
okcupid_profiles %>%
  group_by(sex) %>%
  summarise(age_mean = mean(age, na.rm = TRUE),
            age_sd = sd(age, na.rm = TRUE))

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 3
##   sex   age_mean age_sd
##   <chr>   <dbl> <dbl>
## 1 f       32.8  10.0
## 2 m       32.0   9.03
```

Fortunately, it's possible to use more than one grouping variable with the group_by() command. In the code below we group by sex and status (i.e., dating status).

```
okcupid_profiles %>%
  group_by(sex, status) %>%
  summarise(age_mean = mean(age, na.rm = TRUE),
            age_sd = sd(age, na.rm = TRUE))

## `summarise()` regrouping output by 'sex' (override with `.groups` argument)

## # A tibble: 10 x 4
## # Groups:   sex [2]
##   sex   status   age_mean age_sd
##   <chr> <chr>     <dbl> <dbl>
## 1 f     available  32.2   8.54
## 2 f     married    33.7   8.13
## 3 f     seeing someone 28.1   6.44
## 4 f     single     33.0  10.2
## 5 f     unknown    27.8   5.91
## 6 m     available  34.8   9.40
## 7 m     married    38.7  10.1
## 8 m     seeing someone 30.8   7.06
## 9 m     single     31.9   9.04
## 10 m    unknown    40.7   8.87
```

The resulting output provide for age the mean and standard deviation for every combination of sex and dating status. The first five rows provide output for females at every level of dating status whereas the subsequent five rows provide output for males at every level of dating status.

2.4.5 mutate()

The `mutate()` command can be used to calculate a new column in a data. In the example below we calculate a new column called `age_centered` which is the new version of the `age` column where the mean of the column has been removed from every value. This is merely an example of the many different types of calculations we can perform to create a new column using `mutate()`.

```
okcupid_profiles <- okcupid_profiles %>%
  mutate(age_centered = age - mean(age, na.rm = TRUE))
```

Notice that the `glimpse()` command reveals that after we use the `mutate()` command there is a new column called `age_centered`.

```
glimpse(okcupid_profiles)
```

```
## Rows: 59,946
## Columns: 7
## $ age      <dbl> 22, 35, 38, 23, 29, 29, 32, 31, 24...
## $ diet     <chr> "strictly anything", "mostly other...
## $ height   <dbl> 75, 70, 68, 71, 66, 67, 65, 65, 67...
## $ pets     <chr> "likes dogs and likes cats", "like...
## $ sex      <chr> "m", "m", "m", "m", "m", "m", "f",...
## $ status   <chr> "single", "single", "available", "...
## $ age_centered <dbl> -10.3403, 2.6597, 5.6597, -9.3403,...
```

2.5 Advanced tidyverse commands

In this advanced selection we revisit the commands from the basic tidyverse section but use more complicated code to either select or apply an action to more than one column at a time. We will indicate the columns that we want to select or apply an action to using: `starts_with()`, `ends_with()`, `contains()`, `matches()`, or `where()`. The first four of these are used to indicate columns based on column names. In contrast, the last command, `where()`, is used to

indicate the columns based on the column type (numeric, character, factor, etc.).

We will review all five commands for indicating the columns we want in the `select()` selection below. Following that we will, for brevity, typically use only one of the five commands when illustrating how they work with `summarise()` and `mutate()`.

We begin by loading a new data.

```
library(tidyverse)
data_exp <- read_csv("data_experiment.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   sex = col_character(),
##   t1_vomit = col_double(),
##   t1_aggression = col_double(),
##   t2_vomit = col_double(),
##   t2_aggression = col_double()
## )
```

The `glimpse()` command reveals that this is a small data set where every row represents one rat. The sex of the rat is recorded as well as, for each of two time points, a rating of vomiting and aggression.

```
glimpse(data_exp)

## Rows: 6
## Columns: 6
## $ id          <dbl> 1, 2, 3, 4, 5, 6
## $ sex         <chr> "male", "female", "male", "female..."
## $ t1_vomit     <dbl> 3, 2, 0, 3, 2, 1
## $ t1_aggression <dbl> 5, 6, 4, 7, 3, 8
## $ t2_vomit     <dbl> 2, 1, 1, 2, 1, 2
## $ t2_aggression <dbl> 6, 7, 6, 7, 5, 8
```

2.5.1 select()

2.5.1.1 select() using column name

2.5.1.1.1 starts_with()

`starts_with()` allows us to select columns based on how the column name begins. Here we put the columns that begin with “t1” into a new data called `data_time1`.

```
data_time1 <- data_exp %>%  
  select(starts_with("t1"))
```

The `glimpse` command shows us the new data only contains the columns that begin with “t1”

```
glimpse(data_time1)
```

```
## Rows: 6  
## Columns: 2  
## $ t1_vomit      <dbl> 3, 2, 0, 3, 2, 1  
## $ t1_aggression <dbl> 5, 6, 4, 7, 3, 8
```

2.5.1.1.2 ends_with()

`ends_with()` allows us to select columns based on how the column name ends. Here we put the columns that end with “aggression” into a new data set called `data_aggression`.

```
data_aggression <- data_exp %>%  
  select(ends_with("aggression"))
```

```
glimpse(data_aggression)
```

```
## Rows: 6  
## Columns: 2  
## $ t1_aggression <dbl> 5, 6, 4, 7, 3, 8  
## $ t2_aggression <dbl> 6, 7, 6, 7, 5, 8
```

2.5.1.1.3 *contains()*

`contains()` allows us to select columns based on the contents of the column name. Here we put the columns that have “_” in the name into a new data set called `new_data`.

```
new_data <- data_exp %>%  
  select(contains("_"))
```

```
glimpse(new_data)
```

```
## Rows: 6  
## Columns: 4  
## $ t1_vomit      <dbl> 3, 2, 0, 3, 2, 1  
## $ t1_aggression <dbl> 5, 6, 4, 7, 3, 8  
## $ t2_vomit      <dbl> 2, 1, 1, 2, 1, 2  
## $ t2_aggression <dbl> 6, 7, 6, 7, 5, 8
```

2.5.1.1.4 *matches()*

It's also possible to use *regex* (regular expression) to select columns. Regex is a powerful way to specify search/matching requirements for text - in this case the text of column names. An explanation of regex is beyond the scope of this chapter. Nonetheless the example below uses regex to select any column with an underscore in the column name followed by any character. The result is the same as the above for the `contains()` command. However, the `matches()` command is more flexible than the `contains()` command and can take into account substantially more complicated situations.

```
data_matched<- data_exp %>%  
  select(matches("_."))
```

You can see the columns selected using regex:

```
glimpse(data_matched)
```

```
## Rows: 6  
## Columns: 4  
## $ t1_vomit      <dbl> 3, 2, 0, 3, 2, 1  
## $ t1_aggression <dbl> 5, 6, 4, 7, 3, 8  
## $ t2_vomit      <dbl> 2, 1, 1, 2, 1, 2  
## $ t2_aggression <dbl> 6, 7, 6, 7, 5, 8
```


You can learn about regex at [RegexOne](https://regexone.com)¹ and test your regex specification at [Regex101](https://regex101.com)². Ideally though, as we discuss in the next chapter, you can use naming conventions that are sufficiently thoughtful that you don't need regex, or only rarely. The reason for this is that regex can be challenging to use. As Twitter user @ThatJenPerson noted “Regex is like tequila: use it to try to solve a problem and now you have two problems.” Nonetheless, at one or two points in the future we will use regex to solve a problem (but not tequila).

2.5.1.2 select() using column type

If many cases we will want to select or perform an action on a column based on whether the column is a numeric, character, or factor column (indicated in glimpse output as dbl, chr, and fct, respectively). We will learn more about factors later in this chapter. Each of these column types can be selected by using `is.numeric`, `is.character`, or `is.factor`, respectively, in combination with the `where()` command.

We can select numeric columns using `where()` and `is.numeric`:

```
data_numeric_columns <- data_exp %>%  
  select(where(is.numeric))
```

You can see the new data contains only the numeric columns:

```
glimpse(data_numeric_columns)  
  
## Rows: 6  
## Columns: 5  
## $ id          <dbl> 1, 2, 3, 4, 5, 6  
## $ t1_vomit     <dbl> 3, 2, 0, 3, 2, 1  
## $ t1_aggression <dbl> 5, 6, 4, 7, 3, 8  
## $ t2_vomit     <dbl> 2, 1, 1, 2, 1, 2  
## $ t2_aggression <dbl> 6, 7, 6, 7, 5, 8
```

We can select numeric columns using `where()` and `is.character`:

```
data_character_columns <- data_exp %>%  
  select(where(is.character))
```

You can see the new data contains only the character columns:

¹<https://regexone.com>

²<https://regex101.com>

```
glimpse(data_character_columns)
```

```
## Rows: 6
## Columns: 1
## $ sex <chr> "male", "female", "male", "female", "male", ...
```

If a future chapter you will see how we can select factors using `where(is.factor)`.

2.5.2 summarise()

The `summarise()` command can summarise multiple columns when combined with `starts_with()`, `ends_with()`, `contains()`, `matches()`, and `where()`. However, to use these powerful tools for indicating columns with the `summarise` command we need the help of the `across()` command (i.e., across multiple columns).

If we want to obtain the mean of all the columns that start with “t1” we use the commands below. The `across` command requires that we indicate the columns we want via the `.cols` argument and the command/function we want to run on those columns via the `.fns` argument. In the example below, we also add `na.rm = TRUE` at the end; this is something we send to the `mean` command to let it know how we want to handle missing data. We add `as.data.frame()` to get a larger number of decimals.

```
data_exp %>%
  summarise(across(.cols = starts_with("t1"),
                    .fns = mean,
                    na.rm = TRUE)) %>%
  as.data.frame()
```

```
##   t1_vomit t1_aggression
## 1      1.833           5.5
```

If you want to get more sophisticated, you can also add this `.names` argument below which tells R to call label each output mean by the column name followed by “_mean”.

```
data_exp %>%
  summarise(across(.cols = starts_with("t1"),
                    .fns = mean,
                    na.rm = TRUE,
                    .names = "{col}_mean")) %>%
  as.data.frame()
```

```
##   t1_vomit_mean t1_aggression_mean
## 1           1.833                5.5
```

Often you want to calculate more than one statistic for each column. For example, you might want the mean, standard deviation, min, and max. These statistics can be calculated via the `mean`, `sd`, `min`, and `max` commands, respectively. However, you need to create a list with the statistics you desire.

Below we create a list of the descriptive statistics we desire called `desired_statistics`, but you can use any name you want. This list only needs to be specified once, but we will repeat it in the examples below for clarity.

```
desired_descriptives <- list(
  mean = ~mean(.x, na.rm = TRUE),
  sd = ~sd(.x, na.rm = TRUE)
)
```

Once you have created the list of descriptive statistics you want you can run the command below to obtain those statistics. However, as you will see the output is too wide to be helpful.

```
data_exp %>%
  summarise(across(.cols = starts_with("t1"),
                  .fns = desired_descriptives)) %>%
  as.data.frame()
```

```
##   t1_vomit_mean t1_vomit_sd t1_aggression_mean
## 1           1.833        1.169                5.5
##   t1_aggression_sd
## 1           1.871
```

Consequently, we add the `t()` command (i.e., transpose command) to the end of the summarise request to get a more readable list of statistics:

```
desired_descriptives <- list(
  mean = ~mean(.x, na.rm = TRUE),
  sd = ~sd(.x, na.rm = TRUE)
)

data_exp %>%
  summarise(across(.cols = starts_with("t1"),
                  .fns = desired_descriptives)) %>%
  as.data.frame() %>%
  t()
```

```
##           [,1]
## t1_vomit_mean 1.833
## t1_vomit_sd   1.169
## t1_aggression_mean 5.500
## t1_aggression_sd 1.871
```

Note that in the `across` command above we could also have used: `ends_with()`, `contains()`, `matches()`, or `where()`.

2.5.3 `mutate()`

The `mutate()` command can also be applied to multiple columns using the `across()` command. However, sometimes we need to embed our calculation in a custom function. Below is a custom function called `make_centered`. This custom function takes the values in a column and subtracts the column mean from each value in the column. This is the same task we did previous using the `mutate()` command in the basic tidyverse section.

```
make_centered <- function(values) {
  values_out <- values - mean(values, na.rm = TRUE)
  return(values_out)
}
```

The `glimpse()` command shows us all the column names. Also notice the values in the aggression columns are integers.

```
glimpse(data_exp)
```

```
## Rows: 6
## Columns: 6
## $ id      <dbl> 1, 2, 3, 4, 5, 6
## $ sex     <chr> "male", "female", "male", "female..."
## $ t1_vomit <dbl> 3, 2, 0, 3, 2, 1
## $ t1_aggression <dbl> 5, 6, 4, 7, 3, 8
## $ t2_vomit <dbl> 2, 1, 1, 2, 1, 2
## $ t2_aggression <dbl> 6, 7, 6, 7, 5, 8
```

We combine the `mutate()` command, with the `across()` command, and our custom `make_centered()` command below. The command “centers” or subtracts the mean from any column that ends with “aggression”.

```
data_exp <- data_exp %>%
```


You can see the new column we created with the `glimpse()` command:

```
glimpse(data_exp)

## Rows: 6
## Columns: 7
## $ id          <dbl> 1, 2, 3, 4, 5, 6
## $ sex         <chr> "male", "female", "male", "female..."
## $ t1_vomit     <dbl> 3, 2, 0, 3, 2, 1
## $ t1_aggression <dbl> -0.5, 0.5, -1.5, 1.5, -2.5, 2.5
## $ t2_vomit     <dbl> 2, 1, 1, 2, 1, 2
## $ t2_aggression <dbl> -0.5, 0.5, -0.5, 0.5, -1.5, 1.5
## $ vomit_avg    <dbl> 2.5, 1.5, 0.5, 2.5, 1.5, 1.5
```

The `print()` command could make it easier to see that the new column is the average of the other two, but if we use the `print()` command below it wouldn't work. Why? There are too many columns in the data set, so only the first few columns are shown.

```
print(data_exp)
```

To see that the new column, `vomit_avg`, is the average of the other vomit columns we use the `select` command before `print()`. This prints only the relevant columns. When this is done, it's easy to see how the values in the `vomit_avg` column are the mean of the other two columns.

```
data_exp %>%
  select(contains("vomit")) %>%
  print()

## # A tibble: 6 x 3
##   t1_vomit t2_vomit vomit_avg
##   <dbl>    <dbl>    <dbl>
## 1       3       2       2.5
## 2       2       1       1.5
## 3       0       1       0.5
## 4       3       2       2.5
## 5       2       1       1.5
## 6       1       2       1.5
```

2.5.3.2 mutate() for factors

It is critical that you indicate to R that categorical variables are in fact categorical variables. In R, categorical variables are referred to as factors. For humans, a factor like sex has three possible levels: female, male, intersex.

An inspection of the glimpse() command output reveals that the sex column has the type character - as indicated by “chr”. Also notice, as you inspect this output, that we use words (e.g., female) to indicate the sex in the column rather than a number to represent a female participant (e.g., 2). This is the preferred, but less common, approach to entering data.

```
glimpse(data_exp)

## Rows: 6
## Columns: 7
## $ id          <dbl> 1, 2, 3, 4, 5, 6
## $ sex         <chr> "male", "female", "male", "female..."
## $ t1_vomit     <dbl> 3, 2, 0, 3, 2, 1
## $ t1_aggression <dbl> -0.5, 0.5, -1.5, 1.5, -2.5, 2.5
## $ t2_vomit     <dbl> 2, 1, 1, 2, 1, 2
## $ t2_aggression <dbl> -0.5, 0.5, -0.5, 0.5, -1.5, 1.5
## $ vomit_avg    <dbl> 2.5, 1.5, 0.5, 2.5, 1.5, 1.5
```

We need to convert the sex column to a factor in order for R to handle it appropriately in analyses. Failure to indicate the column is a factor could result in R conducting all the analyses and presenting incorrect results. Consequently, it is critical that we convert the column to a factor. Fortunately, that is easily done using the as_factor() command (there is also an as.factor command if as_factor won't work for some reason).

We convert the sex column to a factor with this code:

```
data_exp <- data_exp %>%
  mutate(sex = as_factor(sex))
```

You can confirm this worked with the glimpse() command:

```
glimpse(data_exp)

## Rows: 6
## Columns: 7
## $ id          <dbl> 1, 2, 3, 4, 5, 6
## $ sex         <fct> male, female, male, female, male, ...
## $ t1_vomit     <dbl> 3, 2, 0, 3, 2, 1
```

```
## $ t1_aggression <dbl> -0.5, 0.5, -1.5, 1.5, -2.5, 2.5
## $ t2_vomit      <dbl> 2, 1, 1, 2, 1, 2
## $ t2_aggression <dbl> -0.5, 0.5, -0.5, 0.5, -1.5, 1.5
## $ vomit_avg     <dbl> 2.5, 1.5, 0.5, 2.5, 1.5, 1.5
```

If you entered your data using words for each level of sex (e.g., male, female) you're done at this point. However, if you used numbers to represent each level of sex in your data, there is one more step. Imagine your data was entered in a poorly advised manner, such that 1 was used to indicate male, 2 was used to indicate female, and 3 was used to indicate intersex. If this was the case, you need to indicate to R what each of those values represent. We do that with the code below.

```
data_exp <- data_exp %>%
  mutate(sex = fct_recode(sex,
                           male = "1",
                           female = "2",
                           intersex = "3"))
```

2.6 Using help

In order to become efficient at analyzing data using R, you will need to become adapt at reading and understanding the help files associated with each command. After you have activated a package using the library command (e.g., library(tidyverse)) you can access the help page for every command in that package. To access the help page simply type a question mark followed by the command you want to know how to use (no space between them). The code below brings up the help page for the select() command. Notice that we put the library() command first - just a reminder that this needs to be done prior to using help for that package. Try the commands below in the Console:

```
library(tidyverse)
?select
```

Examine the page that appears on the Help tab in the panel in the lower right of your screen. Read through the help file comparing what you read there to what we have learned about the select command. Notice how the help file tells you about the argument that you sent into the select() command, and also what the select() command returns when it receives those commands. Pay particular attention to the examples near the bottom of the help page. At

the very bottom of the help page you will see [Package dplyr version 1.0.0 Index]. This tells you the `select()` command is from the dplyr package (part of the tidyverse). Notice that the word “Index” is underlined. Click on the word Index. You will be presented with list of other commands in the dplyr package.

As you become more experienced with R help pages, this is how you will learn to use new commands. Examine the help pages for the commands below by typing a question mark into the Console followed by the command name. Note that for `filter` and `starts_with` you will be presented with a menu instead of help page. This typically occurs because the command is in more than one package. If this does occur, read through the options you are presented with to try and figure out which one you wanted. Typically, you want the first option. If you’re not sure, try one. IF it’s not what you want, use the back arrow in the Help panel to go back and pick another one.

- `mutate`
- `filter`
- `starts_with`

2.7 Base R vs tidyverse

All of the commands used to this point in the chapter have been the tidyverse approach to using R. That is the approach we will normally use. However, it’s important to note that there is another way of using R, called base R.

Sometimes students have problems with their code when they mix and match these approaches using a bit of both. We will be using the tidyverse approach to using R but on the internet you will often see sample code that uses the older base R approach. A bit of background knowledge is helpful for understanding why we do things one way (e.g., `read_csv` with the tidyverse) instead of another (e.g., `read.csv` with base R).

2.7.1 Tibbles vs. data frames

When you load data into R, it is typically represented in one of two formats inside the computer - depending on the command you used. The original format for representing a data set in R is the data frame. You will see this term used frequently when you read about R. When you load data using `read.csv()`, your data is loaded into a data frame in the computer. That is, your data is represented in the memory of the computer in a particular format

and structure called a data frame. This contrasts with the newer tidyverse approach to representing data in the computer called a tibble - which is just a newer more advanced version of the data frame.

2.7.2 read.csv() and data frames

When you read data into R using the command `read.csv()` (with a period) you load the data into a data frame (base R).

```
my_dataframe <- read.csv(file = "data_okcupid.csv")
```

Notice that when you print a data frame it does not show you the number of rows or columns above the data like our example did with the `okcupid_profiles` data. Likewise, it does not show you the type of data in each column (e.g., `dbl`, `fct`, `chr`). It also presents all of your data rather than just the first few rows (as the tibble does). As a result, in the output below, we show only the first 10 rows of the output - because all the rows are printed in your Console with a data frame (too much to show here).

```
print(my_dataframe)
```

```
##      age      diet height      pets
## 1    22 strictly anything    75 likes dogs and likes cats
## 2    35      mostly other    70 likes dogs and likes cats
## 3    38      anything      68      has cats
## 4    23      vegetarian    71      likes cats
## 5    29      <NA>         66 likes dogs and likes cats
## 6    29      mostly anything    67      likes cats
## 7    32 strictly anything    65 likes dogs and likes cats
## 8    31      mostly anything    65 likes dogs and likes cats
## 9    24 strictly anything    67 likes dogs and likes cats
## 10   37      mostly anything    65 likes dogs and likes cats
##      sex      status
## 1     m      single
## 2     m      single
## 3     m available
## 4     m      single
## 5     m      single
## 6     m      single
## 7     f      single
## 8     f      single
## 9     f      single
## 10    m      single
```

2.7.3 read_csv() and tibbles

When you read data into R using the command `read_csv()` (with an underscore) you load the data into a tibble (tidyverse).

```
my_tibble <- read_csv(file = "data_okcupid.csv")
```

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   diet = col_character(),
##   height = col_double(),
##   pets = col_character(),
##   sex = col_character(),
##   status = col_character()
## )
```

The tibble is modern version of the data frame. Notice that when you print a tibble it DOES show you the number of rows and columns. As well, it shows you the type of data in each column. Importantly, the tibble only provides the first few rows of output so it doesn't fill your screen.

```
print(my_tibble)
```

```
## # A tibble: 59,946 x 6
##   age diet      height pets      sex status
##   <dbl> <chr>    <dbl> <chr>    <chr> <chr>
## 1  22 strictly an~  75 likes dogs and l~ m single
## 2  35 mostly other  70 likes dogs and l~ m single
## 3  38 anything      68 has cats      m availa~
## 4  23 vegetarian    71 likes cats      m single
## 5  29 <NA>          66 likes dogs and l~ m single
## 6  29 mostly anyt~  67 likes cats      m single
## 7  32 strictly an~  65 likes dogs and l~ f single
## 8  31 mostly anyt~  65 likes dogs and l~ f single
## 9  24 strictly an~  67 likes dogs and l~ f single
## 10 37 mostly anyt~  65 likes dogs and l~ m single
## # ... with 59,936 more rows
```

In short, you should always use tibbles (i.e., use `read_csv()` not `read.csv()`). The differences between data frames and tibbles run deeper than the superficial output provided here. On some rare occasions an old package or command may not work with a tibble so you need to make it a data frame. You can do so with the commands below. We will flag these rare occurrences to you when they occur.

```
# Create a data frame from a tibble  
new_dataframe <- as.data.frame(my_tibble)
```

3

Making your data ready for analysis

3.1 Required Packages

The following packages must be installed before starting this chapter.

Required Packages

apaTables
HMisc
janitor
psych
skimr
tidyverse

Important Note: You should NOT use `library(psych)` at any point! There are major conflicts between the `psych` package and the `tidyverse`. We will access the `psych` package commands by preceding each command with `psych::` instead of using `library(psych)`.

The following data files are used in this chapter:

Required Data

data_ex_between.csv
data_ex_within.csv
data_food.csv
data_item_scoring.csv
data_item_time.csv

The files are available at: <https://github.com/dstanley4/psyc6060bookdown>

3.2 Objective

In this chapter we strongly advocate that you use a naming convention for file, variable, and column names. This convention will save you hours of hassles and permit easy application of certain tidyverse commands. However, we must stress that although the naming convention we advocate is based on the tidyverse style guide, it is not “right” or “correct” - there are other naming conventions you can use. Any naming convention is better than no naming convention. The naming convention we advocate here will solve many problems. We encourage to use this system for for weeks or months over many projects - until you see the benefits of this system, and correspondingly its shortcomings. After you are well versed in the strengths/weaknesses of the naming conventions used here you may choose to create your own naming convention system.

3.3 Context

Due to a number of high profile failures to replicate study results (Nosek, 2015) it's become increasingly clear that there is a general crisis of confidence in many areas of science (Baker, 2016). Statistical (and other) explanations have been offered (Simmons et al., 2011) for why it's hard to replicate results across different sets of data. However, scientists are also finding it challenging to recreate the numbers in their own papers using their own data. Indeed, the editor of Molecular Brain asked authors to submit the data used to create the numbers in published papers and found that the wrong data was submitted for 40 out of 41 papers (Miyakawa, 2020).

Consequently, some researchers have suggested that it is critical to distinguish between replication and reproducibility (Patil P., 2019). Replication refers to trying to obtain the same result from a different data sets. Reproducibility refers to trying to obtain the same results from the same data set. Unfortunately, some authors use these two terms interchangeably and fail to make any distinction between them. I encourage you to make the distinction and the use the terms consist with use suggested by (Patil P., 2019).

It may seem that reproducibility should be a given - but it's not. Correspondingly, there is a trend for journals and authors to adopt Transparency and Openness Promotion (TOP) guidelines¹. These guidelines involve such things


¹<https://www.cos.io/our-services/top-guidelines>

as making your materials, data, code, and analysis scripts available on public repositories so anyone can check your data. A new open science journal rating system has even emerged called the TOP Factor².

The idea is not that open science articles are more trustworthy than other types of articles – the idea is that trust doesn’t play a role. Anyone can inspect the data using the scripts and data provided by authors. It’s really just the same as making your science available for auditing the way financial records can be audited. But just like in the world of business, some people don’t like the idea of making it possible for others to audit their work. The problems reported in *Molecular Brain* (doubtless common to many journals) are likely avoided with open science – because the data and scripts needed to reproduce the statistics in the articles are uploaded prior to publication.

The TOP open science guidelines have made an impact and some newer journals, such as *Meta Psychology*, have fully embraced open science. Figure 3.1 shows the header from an article³ in *Meta Psychology* that clearly delineates the open science attributes of the article that used computer simulations (instead of participant data). Take note that the header even indicates who verified that the analyses in the article were reproducible.

Meta-Psychology, 2020, vol 4, MP.2019.1630
<https://doi.org/10.15626/MP.2019.1630>
 Article type: Original Article
 Published under the CC-BY4.0 license



Open data: N/A
 Open materials: Yes
 Open and reproducible analysis: Yes
 Open reviews and editorial process: Yes
 Preregistration: N/A

Edited by: Rickard Carlsson
 Reviewed by: Thom Baguley, Julia Haaf,
 Paul-Christian Bürkner
 Analysis reproduced by: Erin Buchanan
 All supplementary files can be accessed at OSF:
<https://doi.org/10.17605/OSF.IO/3UZAM>

FIGURE 3.1: Open science in an article header

In Canada, the majority of university research is funded by the Federal Government’s Tri-Agency (i.e., NSERC, SSHRC, CIHR). The agency has a new draft Data Management Policy⁴ in which they state that “*The agencies believe that research data collected with the use of public funds belong, to the fullest extent possible, in the public domain and available for reuse by others.*” The perspective of the funding agency on data ownership differs substantially from that of some researchers who incorrectly believe “they own their data”. In Canada at least, the government makes it clear that when tax payers fund research (through the Tri-Agency) the research data is public property. Additionally the Tri-Agency Data Management policy clearly indicates the responsibilities of funded researchers:

”Responsibilities of researchers include:

- incorporating data management best practices into their research;

²<https://topfactor.org>

³<https://open.lnu.se/index.php/metapsychology/article/view/1630/2266>

⁴https://www.ic.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html

- developing data management plans to guide the responsible collection, formatting, preservation and sharing of their data throughout the entire life cycle of a research project and beyond;
- following the requirements of applicable institutional and/or funding agency policies and professional or disciplinary standards;
- acknowledging and citing data sets that contribute to their research; and
- staying abreast of standards and expectations of their disciplinary community.”

As a result of this perspective on data, it’s important that you think about structuring your data for reuse by yourself and others before you collect it. Toward this end, properly documenting your data file and analysis scripts is critical.

3.4 Begin with the end in mind

In this chapter we will walk you through the steps from data collection, data entry, loading raw data, and the creation of data you will analyze (analytic data) via pre-processing scripts. These steps are outlined in Figure 3.2. This figure makes a clear distinction between raw data and analytic data. Raw data refers to the data as you entered it into a spreadsheet or received it from survey software. Analytic data is the data that has been structured and processed so that it is ready for analysis. This pre-processing could include such things as identifying categorical variables to the computer, averaging multiple items into a scale score, and other tasks.

It’s critical that you don’t think of the analysis of your data as being completely removed from the data collection and data entry choices you make. Poor choices at the data collection and data entry stage can make your life substantially more complicated when it comes time to write the pre-processing script that will convert your raw data to analytic data. The mantra of this chapter is *begin with the end in mind*.

It’s difficult to begin with the end in mind when you haven’t read later chapters. So, here we will provide you with some general thoughts around different approaches to structuring data files and the naming conventions you can use when creating those data files.



FIGURE 3.2: Data science pipeline by Roger Peng.

3.4.1 Structuring data: Obtaining tidy data

When conducting analyses in R it is typically necessary to have data in a format called tidy data (Wickham, 2014). Tidy data⁵, as defined by Hadley, involves (among other requirements) that:

1. Each variable forms a column.
2. Each observation forms a row.

The tidy data format can be initially challenging for some researchers to understand because it is based on thinking about, and structuring data, in terms of observations/measurements instead of participants. In this section we will describe common approaches to entering animal and human participant data and how they can be done keeping the tidy data requirement in mind. It's not essential that data be entered in a tidy data format but it is essential that you enter data in a manner that makes it easy to later convert data to a tidy data format. When dealing with animal or human participant data it's common to enter data into a spreadsheet. Each row of the spreadsheet is typically used to represent a single participant and each column of the spreadsheet is used to represent a variable.

Between participant data. Consider Table 3.3 which illustrates between participant data for six human participants running 5 kilometers. The first

⁵<https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html>

TABLE 3.3: Between participant data entered one row per participant

id	sex	elapsed_time
1	male	40
2	female	35
3	male	38
4	female	33
5	male	42
6	female	36

TABLE 3.4: Within participant data entered one row per participant

id	sex	march	may	july
1	male	40	37	35
2	female	35	32	30
3	male	38	35	33
4	female	33	30	28
5	male	42	39	37
6	female	36	33	31

column is `id`, which indicates there are six unique participants and provides and identification number for each of them. The second column is `sex`, which is a variable, and there is one observation per for row, so `sex` also conforms to the tidy data specification. Finally, there is a last column `five_km_time` which is a variable with one observation per row – also conforming to tidy data specification. Thus, single occasion between subject data like this conforms to the tidy data specification. There is usually nothing you need to do to convert between participant data (or cross-sectional data) to be in a tidy data format.

Within participant data. Consider Table 3.4 which illustrates within participant data for six human participants running 5 kilometers - but on three different occasions. The first column is `id`, which indicates there are six unique participants and provides and identification number for each of them. The second column is `sex`, which is a variable, and there is one observation per for row, so `sex` also conforms to the tidy data specification. Next, there are three different columns (`march`, `may`, `july`) each of which contains elapsed time (in minutes) for the runner in a different month. Elapsed run times are spread out over three columns so `elapsed_time` is not in a tidy data format. Moreover, it's not clear from the data file that `march`, `may`, and `july` are levels of a variable called `occasion`. Nor is it clear that `elapsed_times` are recorded in each of those columns (i.e., the dependent variable is unknown/not labeled). Although this format is fine as a data entry format it clearly has problems associated with it when it comes time to analyze your data.

TABLE 3.5: A tidy data version of the within participant data

id	sex	occasion	elapsed_time
1	male	march	40
1	male	may	37
1	male	july	35
2	female	march	35
2	female	may	32
2	female	july	30
3	male	march	38
3	male	may	35
3	male	july	33
4	female	march	33
4	female	may	30
4	female	july	28
5	male	march	42
5	male	may	39
5	male	july	37
6	female	march	36
6	female	may	33
6	female	july	31

Thus, a major problem with entering repeated measures data in the one row per person format is that there are hidden variables in the data and you need insider knowledge to know what the columns represent. That said, this is not necessarily a terrible way to enter your data as long as you have all of this missing information documented in a data code book.

Disadvantages one row per participant	Advantages one row per participant
1) Predictor variable (<i>occasion</i>) is hidden and spread over multiple columns	1) Easy to enter this way
2) Unclear that each month is a level of the predictor variable <i>occasion</i>	
3) Dependent variable (<i>elapsed_time</i>) is not indicated	
4) Unclear that <i>elapsed_time</i> is the measurement in each month column	

Fortunately, the problems with Table 3.4 can be largely resolved by converting the data to the a tidy data format. This can be done with the `pivot_long()` command that we will learn about later in this chapter. Thus, we can enter the data in the format of Table 3.4 and later convert it to a tidy data format. After this conversion the data will be appear as in Table 3.5. For `elapsed_time` variable this data is now in the tidy data format. Each row corresponds to a single `elapsed_time` observed. Each column corresponds to a single variable. Somewhat problematically, however, `sex` is repeated three times for each person (i.e., over the three rows) - and this can be confusing. However, if the focus in on analyzing elapsed time this tidy data format makes sense. Importantly, there is an `id` column for each participant so R knows that this information is repeated for each participant and is not confused by repeating the sex designation over three rows. Indirectly, this illustrates the importance of having an `id` column to indicate each unique participant.

Why did we walk you through this technical treatment of structuring data at this point in time - so that you pay attention to the advice the follows. You can see at this point that you may well need to restructure your data for certain analyses. The ability to do so quickly and easily depends upon following the advice in this chapter around naming conventions for variables and other aspects of your analyses. You can imagine the challenges for converting the data in Figure 3.4 to the data in Figure 3.5 by hand. You want to be able to automate that process and others - which is made substantially easier if you follow the forthcoming advice about naming conventions in the tidyverse.

3.5 Data collection considerations

Data can be collected in a wide variety of ways. Regardless of the method of data collection researchers typically come to data in one of two ways: 1) a research assistant enters the data into a spreadsheet type interface, or 2) the data is obtained as the output from computer software (e.g., Qualtrics, SurveyMonkey, Noldus, etc.).

Regardless of the approach, it is critical to name your variables appropriately. For those using software, such as Qualtrics, this means setting up the software to use appropriate variable names PRIOR to data collection - so the exported file has desirable column names. For spreadsheet users, this means setting up the spreadsheet in which the data will be recorded with column names that are amenable to the future analyses you want to conduct.

Although failure to take this thoughtful approach at the data collection stage can be overcome - it is only overcome with substantial manual effort. There-

fore, as noted previously, we strongly encourage you to follow the naming conventions we espouse here when you set up your data recording regime. Additionally, we encourage you to give careful thought in advance to the codes you will use to record missing data.

3.5.1 Naming conventions

To make your life easier down the road, it is critical you set up your spreadsheet or online survey such that it uses a naming convention prior to data collection. The naming conventions suggested here are adapted from the tidyverse style guide⁶.

- Lowercase letters only
- If two word column names are necessary, only use the underscore (“_”) character to separate words in the name.
- Avoid short decontextualized variable names like q1, q2, q3, etc.
- Do use moderate length column names. Aim to achieve a unique prefix for related columns so that those columns can be selected using the `starts_with()` command discussed in the previous chapter. Be sure to avoid short two or three letter prefixes for item names. Instead, use unique moderate length item prefixes so that it will easy to select those columns using `start_with()` such that you don’t accidentally get additionally columns you don’t want - that have a similar prefix. See the Likert-type item section below for additional details.
- If you have a column name that represents the levels of two repeated measures variables only use the underscore character to separate the levels of the different variables. See within-participant ANOVA section below for details.

3.5.2 Likert-type items

A Likert-type item is typically composed of a statement with which participants are asked to agree or disagree. For example, participants could be asked to indicate the extent to which they agree with a number of statements such as “I like my job”. Then they would be presented with response scale such as: 1 - Strongly Disagree, 2 - Moderately Disagree, 3 - Neutral, 4, Moderately Agree, 5 - Strongly Agree. A common question is, how should I enter the data?

- **Enter numeric responses not labels.** You should enter the numeric value for each item response (e.g., 5) into your data - not the label (e.g., Strongly

⁶<https://style.tidyverse.org>

Agree). The labels associated with each value can be applied later in a script, if needed.

- **High numbers should be associated with more of the construct being measured.** When designing your survey or data collection tools, it is important that you set the response options appropriately. If your scale measures job satisfaction, it is important that you collect data in a manner that ensures high numbers on the job satisfaction scale indicate high levels of job satisfaction. Therefore, assigning numbers makes sense using the 5-point scale: 1 - Strongly Disagree, 2 - Moderately Disagree, 3 - Neutral, 4, Moderately Agree, 5 - Strongly Agree. With this approach high response numbers indicate more job satisfaction. However, using the opposite scale would not make sense: 1 - Strongly Agree, 2 - Moderately Agree, 3 - Neutral, 4, Moderately Disagree, 5 - Strongly Disagree. With this opposite scale high numbers on a job satisfaction scale would indicate lower levels of job satisfaction - a very confusing situation. Avoid this situation, assign numbers so that higher numbers are associated with more of the construct being measured.
- **Use appropriate item names.** As described in the naming convention section, use moderate length names with different labels for each subscale.
- **Use moderate length column names unique to each subscale.** Imagine you have a survey with an 18-item commitment scale (Meyer et al., 1993) composed of three 6-item subscales: affective, normative, and continuance commitment. It would be a poor choice to prefix the labels of all 18 columns in your data with “commit” such that the names would be commit1, commit2, commit3, etc. The problem with this approach is that it fails to distinguish among the three subscales in naming convention; making it impossible to select the items for a single subscale using `starts_with()`. A better, but still poor choice for a naming convention would be use use a two letter prefix for the three scale such ac, nc, and cc. This would result in names for the columns like ac1, ac2, ac3, etc. This is an improvement because you could apparently (but likely not) select the columns using `starts_with(“ac”)`. The problem with these short names is that there could be many columns in data set that start with “ac” beside the affective commitment items. You might want to select the affective commitment items using `starts_with(“ac”)`; but you would get all the affective commitment item columns; but also all the columns measuring other variables that also start with “ac”. Therefore, it’s a good idea to use a moderate length unique prefix for column names. For example, you might use prefixes like affectcom, normcom, and contincom for the three subscales. This would create column names like affectcom1, affectcom2, affectcom3, etc. These column prefixes are unlikely to be duplicated in other places in your column name conventions making it easy to select those columns using a command like `starts_with(“affectcom”)`.

Indicate in the item name if the item is reversed keyed. Sometimes

with Likert-type items, an item is reverse keyed. For example, on a job satisfaction scale, participants will typically respond to items that reflect job satisfaction using the scale: 1 - Strongly Disagree, 2 - Moderately Disagree, 3 - Neutral, 4, Moderately Agree, 5 - Strongly Agree. Higher numbers indicate more job satisfaction. Sometimes, however, some items will use the same 1 to 5 response scale but be worded in the opposite manner such as “I hate my job”. Responding with a 5 to this item would indicate high job *dissatisfaction*. But the columns for job satisfaction items should have high values indicate job satisfaction not job dissatisfaction. Consequently, we flag the names of columns with reversed responses (i.e., reverse-key items) so that we know to treat those column differently later. Columns with reverse-keyed items need to be processed by a script so that the values are flipped and scored in the right direction. The procedure for doing so is outlined in the next point.

Indicate in the item name the range for reverse-key items. If an item is reverse keyed, the process for the flipping the scores depends upon the range of a scale. Although 5-point scales are common, any number of points are possible. The process for correcting a reverse-key item depends upon: 1) the number of points on the scale, and 2) the range of the points on the scale. The reverse-key item correction process is different for an item that uses a 5-point scale ranging from 1 to 5 versus from 0 to 4. Both are 5-point scales but your correction process will be different. Therefore, for reverse-key items add a suffix at the end of each item name that indicates an item is reverse keyed and the range of the item. For example, if the third job satisfaction item was reversed keyed on scale using a 1 to 5 response format you might name the item: `jobsat3_rev15`. The suffix “_rev15” indicates the item is reverse keyed and the range of responses used on the item is 1 to 5. Be sure to set up your survey with this naming convention when you collect your data.

- If you collect items over multiple time points use a prefix with a short code to indicate the time followed by an underscore. For example, if you had a multi-item self-esteem scale you might call the column for the first time “t1_esteem1_rev15”. This indicate that you have for time 1 (t1), the first self-esteem item (esteem1) and that item is reverse keyed on a 1 to 5 scale.

3.5.3 ANOVA between

Avoid numerical representation of categorical variables. Don’t use 1 or 2 to represent a variable like sex. Use male and female in your spreadsheet - likewise in your survey program. Similarly, for between participant variables like `drug_condition` don’t use 1 or 2 use “drug” and “placebo” but the actually drug name would be even better than the word “drug.”

3.5.4 ANOVA within

If you have a study that involves within-participant predictors naming conventions can become examples. When you have a single repeated measures predictor like occasion in the previous running example, it is often necessary to spread the level of that variable over multiple columns (e.g., march, may, july).

When you have multiple repeated-measures predictors the situation is even more complicated. In this case, each column name needs to represent the levels of multiple repeated measures predictors at the time of data entry. For example, imagine you are a food researcher interested in taste ratings (the dependent variable) for various foods and contexts. You have a predictor, food type (i.e., `food_type`), with three levels (pizza, steak, burger). You have a second predictor, temperature, with two levels (hot, cold). All participants taste all foods at all temperatures. Thus, six columns are required to record taste rating for each participant: `pizza_hot`, `pizza_cold`, `steak_hot`, `steak_cold`, `burger_hot`, and `burger_cold`. Notice how each name contains one level of each predictor variable. The levels by the two predictor variables are separated by a single underscore. This should be the only underscore in the variable name because that underscore will be used by the computer when changing the data to the tidy format. If you had two underscores in a name like “italian_pizza_hot” you would confuse the `pivot_longer()` command when it attempts to create a tidy version of the data. The computer would think there were three repeated-measures predictors instead of two. Thus, when dealing with repeated measures predictors, only use underscores to separate levels of predictor variables in column names.

3.6 Following the examples

Below we present example scripts transforming raw data to analytic data for various study designs (experimental and survey). These examples illustrate the value of using the naming conventions outlined previously. Don’t just read the example - follow along with the projects by creating a separate script for each example. Resist the urge to cut and paste from this document - type the script yourself.

When first learning iPhone/Mac software development, I did so by taking a course at Big Nerd Ranch⁷ - yes, that’s a real place. They advised in their material (and now book) the following: “We have learned that “going through

⁷<https://www.bignerdranch.com>

the motions” is much more important than it sounds. Many times we will ask you to start typing in code before you understand it. We realize that you may feel like a trained monkey typing in a bunch of code that you do not fully grasp. But the best way to learn coding is to find and fix your typos. Far from being a drag, this basic debugging is where you really learn the ins and outs of the code. That is why we encourage you to type in the code yourself. You could just download it, but copying and pasting is not programming. We want better for you and your skills.”, p. xiv, (Keur and Hillegass, 2020). This is excellent advice for a beginning statistician or data scientist as well. And as an aside: if you want to learn iPhone programming you can’t go wrong with the Big Nerd Ranch guide!

As you work through this chapter, create your own new script for each example. In light of the above advice, avoid copying and pasting code - type it out; you will be the better for it.

Getting started:

The Class: R Studio in the Cloud Assignment

1. The data should be in the assignment project automatically. Just start the assignment.

For everyone in the class, that’s it.

For those of you not in the class, and reading this work, see the two options below:

R Studio Cloud, custom project

1. Create a new Project using the web interface
2. Upload all the example data files into the project. The data files needed are listed at the beginning of this chapter. The upload button can be found on the Files tab.

R Studio Computer, custom project

1. Create a folder on your computer for the example
2. Place all the example data files in that folder. The data files needed are listed at the beginning of this chapter.
3. Use the menu item File > New Project... to start the project
4. On the window that appears select “Existing Directory”
5. On the next screen, press the “Browse” button and find/select the folder with your data

6. Press the Create Project Button

Regardless of whether you are working from the cloud, or locally, you should now have an R Studio project with your data files in it.

We anticipate that many people will doubtless want to refer back to an encapsulated set of instructions for each design. Therefore the example for each design is written in a way that it stands alone. A consequence of this approach is that there is some redundancy in the code across examples. We see this a strength - because readers will see the commonalities across different types of designs.

As you make a script for each example:

- Recall the instruction from Chapter 1 about putting the date and your name in the script via comments.
- Recall the instruction from Chapter 1 about running `library(tidyverse)` before you type the rest of each script - this provides you with tidyverse auto-complete for the script.
- After you type each new block of code in an example, save your script.
- After you type each new block of code in an example, do two additional things: 1) Session Restart R, 2) Run your script using Source with Echo.

3.7 Entering data into spreadsheets

The first example uses a data file `data_ex_between.csv` that corresponds to a fictitious example where we recorded the run times for a number of male and female participants. How did we create this data file? We used a spreadsheet to enter the data, as illustrated in Figure 3.3. Programs like Microsoft Excel and Google Sheets are good options for entering data.

id	sex	elapsed_time
1	male	40
2	female	35
3	male	38
4	female	33
5	male	42
6	female	36

FIGURE 3.3: Spreadsheet entry of running data

The key to using these types of programs is to save the data as a `.csv` file when

you are done. CSV is short for Comma Separated Values. After entering the data in Figure 3.3 we saved it as `data_ex_between.csv`. There is no need to do so, but if you were to open this file in a text editor (such as TextEdit on a Mac or Notepad on Windows) you would see the information displayed in Figure 3.4. You can see there is one row per person and the columns are created by separating each values by a comma; hence, comma separated values.

```
id,sex,elapsed_time
1,male,40
2,female,35
3,male,38
4,female,33
5,male,42
6,female,36
```

FIGURE 3.4: Text view of CSV data

There are many ways to save data, but the CSV data is one of the better ones because it is a non-proprietary format. Some software, such as SPSS, uses a proprietary format (e.g., `.sav` for SPSS) this makes it challenging to access that data if you don't have that (often expensive) software. One of our goals as scientists is to make it easy for others to audit our work - that allows science to be self-correcting. Therefore, choose an open format for your data like `.csv`.

3.8 Experiment: Between

This section outlines a workflow appropriate for when you plan to conduct independent-groups t-test or a between-participants ANOVA.

To Begin:

- Use the Files tab to confirm you have the data: `data_ex_between.csv`
- Start a new script for this example. Don't forget to start the script name with "script_".

As noted previously, these data correspond to a design where the researcher is interested in comparing run times (`elapsed_time`) based on sex (male/female).

```
# Date: YYYY-MM-DD
# Name: your name here
# Example: Between-participant experiment

# Load data
```

```
library(tidyverse)

my_missing_value_codes <- c("-999", "", "NA")

raw_data_beween <- read_csv(file = "data_ex_between.csv",
                             na = my_missing_value_codes)
```

We load the initial data into a `raw_data_beween` data set but immediately make a copy that we will work with called `analytic_data_between`. It's good to keep a copy of the raw data for reference in the event that you encounter problems.

```
analytic_data_between <- raw_data_beween
```

After loading the data we do initial cleaning to remove empty row/columns and ensure proper naming for columns:

```
library(janitor)

# Initial cleaning
analytic_data_between <- analytic_data_between %>%
  remove_empty("rows") %>%
  remove_empty("cols") %>%
  clean_names()
```

You can confirm the column names follow our naming convention with the `glimpse()` command.

```
glimpse(analytic_data_between)

## Rows: 6
## Columns: 3
## $ id      <dbl> 1, 2, 3, 4, 5, 6
## $ sex     <chr> "male", "female", "male", "female"...
## $ elapsed_time <dbl> 40, 35, 38, 33, 42, 36
```

3.8.1 Creating factors

Following initial cleaning, we identify categorical variables as factors. If you plan to conduct an ANOVA - it's critical that all predictor variables are converted to factors. Inspect the `glimpse()` output - if you followed our data entry

naming conventions, categorical variables should be of the type character. We have one variable, sex, that is a categorical variable of type character (i.e., chr). The participant id column is categorical as well, but of type double (i.e., dbl) which is a numeric column.

```
glimpse(analytic_data_between)

## Rows: 6
## Columns: 3
## $ id      <dbl> 1, 2, 3, 4, 5, 6
## $ sex     <chr> "male", "female", "male", "female"...
## $ elapsed_time <dbl> 40, 35, 38, 33, 42, 36
```

You can quickly convert all character columns to factors using the code below. In this case, the code just converts the sex column to a factor. Because there is only one column (sex) being converted to a factor, we could have treated it the same way as the id column below. However, we use this code because of its broad applicability to many scripts.

```
analytic_data_between <- analytic_data_between %>%
  mutate(across(.cols = where(is.character),
    .fns = as_factor))
```

The participant identification number in the id column is a numeric column, so it was not converted by the above code. The id column is converted to a factor with the code below.

```
analytic_data_between <- analytic_data_between %>%
  mutate(id = as_factor(id))
```

You can ensure both the sex and id columns are now factors using the glimpse() command.

```
glimpse(analytic_data_between)

## Rows: 6
## Columns: 3
## $ id      <fct> 1, 2, 3, 4, 5, 6
## $ sex     <fct> male, female, male, female, male, ...
## $ elapsed_time <dbl> 40, 35, 38, 33, 42, 36
```

This example is so small it's clear you didn't miss converting any columns to factors. In general, however, at this point you should inspect the output

of the `glimpse()` command and make sure you have converted all categorical variables to factors - especially those you will use as predictors.

3.8.2 Factor screening

Inspect the levels of each factor carefully. Make sure the factor levels of each variable are correct. Examine spelling and look for additional unwanted levels. For example, you wouldn't want to have the following levels for `sex`: `male`, `mmale`, `female`. Obviously, `mmale` is an incorrectly typed version of `male`. Scan all the factors in your data for erroneous factor levels. The code below displays the factor levels:

```
analytic_data_between %>%
  select(where(is.factor)) %>%
  summary()
```

```
## id      sex
## 1:1  male  :3
## 2:1  female:3
## 3:1
## 4:1
## 5:1
## 6:1
```

The order of the levels influences how graphs are generated. In these data, the `sex` column has two levels: `male` and `female` in that order. The code below adjusts the order of the `sex` variable because we want the x-axis of a future graph to display columns in the left to right order: `female`, `male`.

```
analytic_data_between <- analytic_data_between %>%
  mutate(sex = fct_relevel(sex,
                           "female",
                           "male"))
```

You can see the new order of the factor levels with `summary()`:

```
analytic_data_between %>%
  select(where(is.factor)) %>%
  summary()
```

```
## id      sex
## 1:1  female:3
## 2:1  male  :3
```

```
## 3:1
## 4:1
## 5:1
## 6:1
```

3.8.3 Numeric screening

For numeric variables, you should search for impossible values. For example, in the context of this example you want to ensure that none of the `elapsed_time`s are impossible, or so large they appear to be data entry errors. One option for doing so is the `summary` command again. This time, however, we use “`is.numeric`” in the `where()` command.

```
analytic_data_between %>%
  select(where(is.numeric)) %>%
  summary()
```

```
## elapsed_time
## Min.      :33.0
## 1st Qu.:35.2
## Median :37.0
## Mean    :37.3
## 3rd Qu.:39.5
## Max.    :42.0
```

Scan the min and max values to ensure there are not any impossible values. If necessary, go back to the original data source and fix these impossible values. Alternatively, you might need to change them to missing values (i.e., NA values).

In this example all the values are reasonable values. However, if we discovered an out of range value (or values) for elapsed time, we could convert those values to missing values with the code below. This code changes (i.e., mutates) a value in the `elapsed_time` column to become NA (not available or missing) if that value is less than zero. If the value is greater than, or equal to zero, it stays the same. Note that when using this command we have to be very specific in terms of specifying our missing value. It usually needs to be one of `NA_real_` or `NA_character_`. For numeric columns use `NA_real_` and for character columns use `NA_character_`.

```
analytic_data_between <- analytic_data_between %>%
  mutate(elapsed_time = case_when(
    elapsed_time < 0 ~ NA_real_,
    elapsed_time >= 0 ~ elapsed_time))
```

Once you are done numeric screening, the data is ready for analysis.

3.9 Experiment: Within one-way

This section outlines a workflow appropriate for when you have a repeated measures design with a single repeated measures predictor. The data corresponds to a design where the researcher is interested in comparing run times (elapsed_time) across three different occasions (march/may/july).

To Begin:

- Use the Files tab to confirm you have the data: data_ex_within.csv
- Start a new script for this example. Don't forget to start the script name with "script_".

```
# Date: YYYY-MM-DD
# Name: your name here
# Example: Within-participant experiment

# Load data
library(tidyverse)

my_missing_value_codes <- c("-999", "", "NA")

raw_data_within <- read_csv(file = "data_ex_within.csv",
                             na = my_missing_value_codes)

## Parsed with column specification:
## cols(
##   id = col_double(),
##   sex = col_character(),
##   march = col_double(),
##   may = col_double(),
##   july = col_double()
## )
```

We load the initial data into raw_data_within but immediately make a copy that we will work with called analytic_data_within. It's good to keep a copy of the raw data for reference if you encounter problems.


```
analytic_data_within <- raw_data_within
```

After loading the data we do initial cleaning to remove empty row/columns and ensure proper naming for columns:

```
library(janitor)

# Initial cleaning
analytic_data_within <- analytic_data_within %>%
  remove_empty("rows") %>%
  remove_empty("cols") %>%
  clean_names()
```

You can confirm the column names following our naming convention with the `glimpse` command - and see the data type for each column.

```
glimpse(analytic_data_within)

## Rows: 6
## Columns: 5
## $ id    <dbl> 1, 2, 3, 4, 5, 6
## $ sex   <chr> "male", "female", "male", "female", "male..."
## $ march <dbl> 40, 35, 38, 33, 42, 36
## $ may   <dbl> 37, 32, 35, 30, 39, 33
## $ july  <dbl> 35, 30, 33, 28, 37, 31
```

3.9.1 Creating factors

Following initial cleaning, we identify categorical variables as factors. If you plan to conduct an ANOVA - it's critical that all predictor variables are converted to factors. Inspect the `glimpse()` output - if you followed our data entry naming conventions, categorical variables should be of the type character.

```
glimpse(analytic_data_within)

## Rows: 6
## Columns: 5
## $ id    <dbl> 1, 2, 3, 4, 5, 6
## $ sex   <chr> "male", "female", "male", "female", "male..."
## $ march <dbl> 40, 35, 38, 33, 42, 36
## $ may   <dbl> 37, 32, 35, 30, 39, 33
```

```
## $ july <dbl> 35, 30, 33, 28, 37, 31
```

We have one variable, sex, that is a categorical variable of type character (i.e., chr). The participant id column is categorical as well, but of type double (i.e., dbl) which is a numeric column.

You can quickly convert all character columns to factors using the code below. In this case, this just converts the sex column to a factor. Because there is only one column (sex) being converted to a factor, we could have treated it the same as the id column below. However, we use this code because of its broad applicability to many scripts.

```
analytic_data_within <- analytic_data_within %>%  
  mutate(across(.cols = where(is.character),  
                .fns = as_factor))
```

The participant identification number in the id column is a numeric column, so it was not converted by the above code. The id column is converted to a factor with the code below.

```
analytic_data_within <- analytic_data_within %>%  
  mutate(id = as_factor(id))
```

You can ensure both the sex and id columns are now factors using the glimpse() command.

```
glimpse(analytic_data_within)
```

```
## Rows: 6  
## Columns: 5  
## $ id    <fct> 1, 2, 3, 4, 5, 6  
## $ sex   <fct> male, female, male, female, male, female  
## $ march <dbl> 40, 35, 38, 33, 42, 36  
## $ may   <dbl> 37, 32, 35, 30, 39, 33  
## $ july  <dbl> 35, 30, 33, 28, 37, 31
```

This example is so small it's clear you didn't miss converting any columns to factors. In general, however, at this point you should inspect the output of the glimpse() command and make sure you have converted all categorical variables to factors - especially those you will use as predictors.

3.9.2 Factor screening

Inspect the levels of each factor carefully. Make sure the factor levels of each variable are correct. Examine spelling and look for additional unwanted levels. For example, you wouldn't want to have the following levels for sex: male, mmale, female. Obviously, mmale is an incorrectly typed version of male. Scan all the factors in your data for erroneous factor levels. The code below displays the factor levels:

```
analytic_data_within %>%  
  select(where(is.factor)) %>%  
  summary()
```

```
## id      sex  
## 1:1  male  :3  
## 2:1 female:3  
## 3:1  
## 4:1  
## 5:1  
## 6:1
```

The order of the levels influences how graphs are generated. In these data, the sex column has two levels: male and female in that order. The code below adjusts the order of the sex variable because we want the x-axis of a future graph to display columns in the left to right order: female, male.

```
analytic_data_within <- analytic_data_within %>%  
  mutate(sex = fct_relevel(sex,  
                           "female",  
                           "male"))
```

You can see the new order of the factor levels with `summary()`:

```
analytic_data_within %>%  
  select(where(is.factor)) %>%  
  summary()
```

```
## id      sex  
## 1:1 female:3  
## 2:1  male  :3  
## 3:1  
## 4:1  
## 5:1  
## 6:1
```

3.9.3 Numeric screening

For numeric variables, it is important to find and remove impossible values. For example, in the context of this example you want to ensure none of the `elapsed_times` are impossible (i.e., negative) or clearly data entry errors.

Because we have several numeric columns that we are screening, we use the `skim()` command from the `skimr` package. The `skim()` command quickly provides basic descriptive statistics. In the output for this command there are also several columns that begin with `p`: `p0`, `p25`, `p50`, `p75`, and `p100` (`p25` and `p75` omitted in output due to space). These columns correspond to the 0th, 25th, 50th, 75th, and 100th percentiles, respectively. The minimum and maximum values for the data column are indicated under the `p0` and `p100` labels. The median is the 50th percentile (`p50`). The interquartile range is the range between `p25` and `p75`.

```
library(skimr)

analytic_data_within %>%
  select(where(is.numeric)) %>%
  skim()

##   skim_variable n_missing  mean   sd p0 p50 p100
## 1      march          0 37.33 3.33 33  37   42
## 2       may          0 34.33 3.33 30  34   39
## 3      july          0 32.33 3.33 28  32   37
```

Scan the minimum and maximum values (`p0` and `p100`) to ensure there are not any impossible values. If necessary, go back to the original data source and fix these impossible values. Alternatively, you might need to change them to missing values (i.e., `NA` values).

In this example all the values are reasonable values. However, if we discovered an out of range value (or values) for elapsed time we could convert those values to missing values with the code below. This code changes (i.e., mutates) a value in the `march` column to become `NA` (not available or missing) if that value is less than zero. If the value is greater than or equal to zero, it stays the same. Note that when using this command we have to be very specific in terms of specifying our missing value. It usually needs to be one of `NA_real_` or `NA_character_`. For numeric columns use `NA_real_` and for character columns use `NA_character_`.

```
analytic_data_within <- analytic_data_within %>%
  mutate(march = case_when(
```

```
march < 0 ~ NA_real_,
march >= 0 ~ march))
```

3.9.4 Pivot to tidy data

The analytic data in its current form does not conform to the tidy data specification. Inspect the data with the `print()` command. Notice that there is not a column for occasion (with levels march/may/july). Instead, there are three columns each of which represents a level of occasion. The levels of occasion are spread across three columns called march, may, and july. Each of these columns contains elapsed time for participants in that month.

```
print(analytic_data_within)
```

```
## # A tibble: 6 x 5
##   id    sex    march    may    july
##   <fct> <fct>   <dbl> <dbl> <dbl>
## 1 1    male     40     37     35
## 2 2    female   35     32     30
## 3 3    male     38     35     33
## 4 4    female   33     30     28
## 5 5    male     42     39     37
## 6 6    female   36     33     31
```

The `pivot_longer()` command below converts our data to the tidy data format. In this command we specify the columns march, may, and july are all levels of a single variable called occasion. We specify the columns involved with the `cols` argument. The code `march:july` after the `cols` argument selects the march column, the july column, and all the columns in-between. Each column contains elapsed times at level of the variable occasion. The `names_to` argument is used to indicate that a new column called occasion should be created to hold the different months. The `value_to` argument is used to indicate that a new column called elapsed_time should be created to hold all the values from the march, may, and july columns.

```
analytic_data_within_tidy <- analytic_data_within %>%
  pivot_longer(cols = march:july,
               names_to = "occasion",
               values_to = "elapsed_time"
  )
```

You can see the data in the new format below.

```
print(analytic_data_within_tidy)
```

```
## # A tibble: 18 x 4
##   id    sex occasion elapsed_time
##   <fct> <fct> <chr>          <dbl>
## 1 1    male  march           40
## 2 1    male  may             37
## 3 1    male  july            35
## 4 2    female march           35
## 5 2    female may             32
## 6 2    female july            30
## 7 3    male  march           38
## 8 3    male  may             35
## 9 3    male  july            33
## 10 4    female march           33
## 11 4    female may             30
## 12 4    female july            28
## 13 5    male  march           42
## 14 5    male  may             39
## 15 5    male  july            37
## 16 6    female march           36
## 17 6    female may             33
## 18 6    female july            31
```

Notice that the new column occasion is of the type character. We need it to be a factor. So use the code below to do so:

```
analytic_data_within_tidy <- analytic_data_within_tidy %>%
  mutate(occasion = as_factor(occasion))
```

You can confirm that occasion is now a factor with the `glimpse()` command. Once this is complete, you are done preparing your one-way within participant analytic data.

```
glimpse(analytic_data_within_tidy)
```

```
## Rows: 18
## Columns: 4
## $ id          <fct> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4...
## $ sex         <fct> male, male, male, female, female, ...
## $ occasion    <fct> march, may, july, march, may, july...
## $ elapsed_time <dbl> 40, 37, 35, 35, 32, 30, 38, 35, 33...
```

You now have two data sets `analytic_data_within` and `ana-`

lytic_data_within_tidy. You can calculate descriptive statistics, correlations and general cross-sectional analyses using the analytic_data_within data set. If you want to conduct a repeated measures ANOVA you use the analytic_data_within_tidy data set. Both data sets are now ready for analysis.

3.10 Experiment: Within N-way

This section outlines a workflow appropriate for when you have a repeated measures design with multiple repeated measures predictors. The data corresponds to a design where the researcher is interested in assessing the taste of food as a function of food type (pizza/steak/burger) and temperature (hot/cold).

To Begin:

- Use the Files tab to confirm you have the data: data_food.csv
- Start a new script for this example. Don't forget to start the script name with "script_".

```
# Date: YYYY-MM-DD
# Name: your name here
# Example: 2-way within-participant experiment

# Load data
library(tidyverse)

my_missing_value_codes <- c("-999", "", "NA")

raw_data_within_nway <- read_csv(file = "data_food.csv",
                                na = my_missing_value_codes)
```

We load the initial data into raw_data_within_nway but immediately make a copy that we will work with called analytic_data_within_nway. It's good to keep a copy of the raw data for reference if you encounter problems.

```
analytic_data_within_nway <- raw_data_within_nway
```

After loading the data we do initial cleaning to remove empty row/columns and ensure proper naming for columns:

```
library(janitor)

# Initial cleaning
analytic_data_within_nway <- analytic_data_within_nway %>%
  remove_empty("rows") %>%
  remove_empty("cols") %>%
  clean_names()
```

You can confirm the column names following our naming convention with the `glimpse` command - and see the data type for each column.

```
glimpse(analytic_data_within_nway)

## Rows: 6
## Columns: 8
## $ id          <dbl> 1, 2, 3, 4, 5, 6
## $ sex         <dbl> 1, 2, 1, 2, 1, 2
## $ pizza_hot   <dbl> 7, 8, 7, 8, 7, 9
## $ pizza_cold  <dbl> 6, 7, 5, 7, 6, 7
## $ steak_hot   <dbl> 6, 6, 7, 7, 8, 7
## $ steak_cold  <dbl> 3, 3, 4, 5, 7, 8
## $ burger_hot  <dbl> 7, 8, 7, 8, 7, 8
## $ burger_cold <dbl> 4, 3, 3, 3, 2, 5
```

3.10.1 Creating factors

Following initial cleaning, we identify categorical variables as factors. If you plan to conduct an ANOVA - it's critical that all predictor variables are converted to factors. In this example, there are two categorical variables `id` and `sex`, but both are represented numerically. As revealed by the `glimpse()` output.

```
glimpse(analytic_data_within_nway)

## Rows: 6
## Columns: 8
## $ id          <dbl> 1, 2, 3, 4, 5, 6
## $ sex         <dbl> 1, 2, 1, 2, 1, 2
## $ pizza_hot   <dbl> 7, 8, 7, 8, 7, 9
## $ pizza_cold  <dbl> 6, 7, 5, 7, 6, 7
## $ steak_hot   <dbl> 6, 6, 7, 7, 8, 7
## $ steak_cold  <dbl> 3, 3, 4, 5, 7, 8
```



```
## $ burger_hot <dbl> 7, 8, 7, 8, 7, 8
## $ burger_cold <dbl> 4, 3, 3, 3, 2, 5
```

We convert both the sex and id columns to factors with the `mutate()` command below:

```
analytic_data_within_nway <- analytic_data_within_nway %>%
  mutate(id = as_factor(id),
         sex = as_factor(sex))
```

The sex column is a factor but we have to tell the computer that 1 indicates male and 2 indicates female.

```
analytic_data_within_nway <- analytic_data_within_nway %>%
  mutate(sex = fct_recode(sex,
                          male = "1",
                          female = "2"))
```

You can ensure all of these columns are now factors using the `glimpse()` command.

```
glimpse(analytic_data_within_nway)
```

```
## Rows: 6
## Columns: 8
## $ id      <fct> 1, 2, 3, 4, 5, 6
## $ sex     <fct> male, female, male, female, male, f...
## $ pizza_hot <dbl> 7, 8, 7, 8, 7, 9
## $ pizza_cold <dbl> 6, 7, 5, 7, 6, 7
## $ steak_hot <dbl> 6, 6, 7, 7, 8, 7
## $ steak_cold <dbl> 3, 3, 4, 5, 7, 8
## $ burger_hot <dbl> 7, 8, 7, 8, 7, 8
## $ burger_cold <dbl> 4, 3, 3, 3, 2, 5
```

Inspect the output of the `glimpse()` command and make sure you have converted all categorical variables to factors - especially those you will use as predictors. As noted in the previous examples, its common to have additional columns that are categorical predictors but appear in the `glimpse()` output as being of the type character. That is not the case in these data, but if it were the command below would turn them into factors:

```
analytic_data_within_nway <- analytic_data_within_nway %>%
  mutate(across(.cols = where(is.character),
                 .fns = as_factor))
```

3.10.2 Factor screening

Inspect the levels of each factor carefully. Make sure the factor levels of each variable are correct. Examine spelling and look for additional unwanted levels. For example, you wouldn't want to have the following levels for sex: male, mmale, female. Obviously, mmale is an incorrectly typed version of male. Scan all the factors in your data for erroneous factor levels. The code below displays the factor levels:

```
analytic_data_within_nway %>%  
  select(where(is.factor)) %>%  
  summary()
```

```
## id      sex  
## 1:1  male  :3  
## 2:1 female:3  
## 3:1  
## 4:1  
## 5:1  
## 6:1
```

The order of the levels influences how graphs are generated. In these data, the sex column has two levels: male and female in that order. The code below adjusts the order of the sex variable because we want the x-axis of a future graph to display columns in the left to right order: female, male.

```
analytic_data_within_nway <- analytic_data_within_nway %>%  
  mutate(sex = fct_relevel(sex,  
                           "female",  
                           "male"))
```

You can see the new order of the factor levels with `summary()`:

```
analytic_data_within_nway %>%  
  select(where(is.factor)) %>%  
  summary()
```

```
## id      sex  
## 1:1 female:3  
## 2:1  male  :3  
## 3:1  
## 4:1  
## 5:1  
## 6:1
```

3.10.3 Numeric screening

For numeric variables, it's important find and remove impossible values. For example, in the context of this example you want to ensure none of the taste ratings the six columns (pizza_hot, pizza_cold, steak_hot, steak_cold, burger_hot, and burger_cold) are outside the range of the 1 to 10 rating scale.

Because we have several numeric columns that we are screening, we use the `skim()` command from the `skimr` package. The `skim()` command quickly provides basic descriptive statistics. In the output for this command there are also several columns that begin with `p`: `p0`, `p25`, `p50`, `p75`, and `p100` (`p25` and `p75` omitted in output due to space). These columns correspond to the 0th, 25th, 50th, 75th, and 100th percentiles, respectively. The minimum and maximum values for the data column are indicated under the `p0` and `p100` labels. The median is the 50th percentile (`p50`). The interquartile range is the range between `p25` and `p75`.

```
library(skimr)

analytic_data_within_nway %>%
  select(where(is.numeric)) %>%
  skim()

##   skim_variable n_missing mean    sd  p0  p50  p100
## 1   pizza_hot      0 7.67 0.82   7  7.5    9
## 2   pizza_cold      0 6.33 0.82   5  6.5    7
## 3   steak_hot       0 6.83 0.75   6  7.0    8
## 4   steak_cold      0 5.00 2.10   3  4.5    8
## 5   burger_hot      0 7.50 0.55   7  7.5    8
## 6   burger_cold     0 3.33 1.03   2  3.0    5
```

Scan the minimum and maximum values (`p0` and `p100`) to ensure there are not any impossible values. That is, ensure all values are inside the 1 to 10 range for the dependent variable. If necessary, get back to the original data source and fix these impossible values. Alternatively, you might need to change them to missing values (i.e., NA values).

In this example all the values are reasonable values. However, if we discovered an out-of-range value (or values) for elapsed time we could convert those values to missing values with the code below. This code changes (i.e., mutates) a value in the `pizza_hot` column to become NA (not available or missing) if that value is outside the 1 to 10 range of the rating scale. Note that when using this command we have to be very specific in terms of specifying our missing value. It usually needs to be one of `NA_real_` or `NA_character_`. For numeric columns use `NA_real_` and for character columns use `NA_character_`.

```
# Values lower than 1 are converted to missing values
analytic_data_within_nway <- analytic_data_within_nway %>%
  mutate(pizza_hot = case_when(
    pizza_hot < 1 ~ NA_real_,
    pizza_hot >= 1 ~ pizza_hot))

# Values greater than 10 are converted to missing values
analytic_data_within_nway <- analytic_data_within_nway %>%
  mutate(pizza_hot = case_when(
    pizza_hot > 10 ~ NA_real_,
    pizza_hot <= 10 ~ pizza_hot))
```

3.10.4 Pivot to tidy data

The analytic data in its current form does not conform to the tidy data specification. Inspect the data with the `print()` command. Notice that columns do not exist for temperature (with levels hot/cold) or food type (with levels pizza/steak/burger). Instead, there are six columns that are combinations of the levels of these variables (i.e., `pizza_hot`, `pizza_cold`, `steak_hot`, `steak_cold`, `burger_hot`, and `burger_cold`). Each of these columns contains taste ratings on a 1 to 10 point scale.

```
print(analytic_data_within)
```

```
## # A tibble: 6 x 5
##   id    sex  march    may    july
##   <fct> <fct>  <dbl> <dbl> <dbl>
## 1 1    male    40    37    35
## 2 2    female  35    32    30
## 3 3    male    38    35    33
## 4 4    female  33    30    28
## 5 5    male    42    39    37
## 6 6    female  36    33    31
```

We need to restructure the data into the tidy data format so that we have a `food_type` column and a temperature column to properly represent these predictors. As well, we need a column that contains all of the taste ratings that is clearly labeled taste. Doing all of these things will ensure we have one variable per column and one observation per row - consistent with the requirements of tidy data. The `pivot_longer()` command below converts our data to the tidy data format.

In this command we specify the columns `march`, `may`, and `july` are all levels

of a single variable called occasion. We specify the columns involved with the `cols` argument. The code `pizza_hot:burger_cold` after the `cols` argument selects the `pizza_hot` column, the `burger_cold` column, and all the columns in between. Each column contains taste ratings at a combination of the levels for the variables `food_type` and `temperature`. The `names_to` argument is used to indicate that two new columns should be created to represent food and temperature. Notice the order food then temperature. This is consistent with our naming convention; in the column name `pizza_hot` the `food_type` is specified before temperature. The `value_to` argument is used to indicate that a new column called `taste` should be created to hold all the values from the `pizza_hot`, `pizza_cold`, `steak_hot`, `steak_cold`, `burger_hot`, and `burger_cold` columns.

```
analytic_data_nway_tidy <- analytic_data_within_nway %>%
  pivot_longer(cols = pizza_hot:burger_cold,
               names_to = c("food_type", "temperature"),
               names_sep = "_",
               values_to = "taste"
  )
```

You can see the data in the new format below.

```
print(analytic_data_nway_tidy)

## # A tibble: 36 x 5
##   id    sex  food_type temperature taste
##   <fct> <fct> <chr>      <chr>      <dbl>
## 1 1    male  pizza      hot          7
## 2 1    male  pizza      cold          6
## 3 1    male  steak      hot          6
## 4 1    male  steak      cold          3
## 5 1    male  burger     hot          7
## 6 1    male  burger     cold          4
## 7 2    female pizza      hot          8
## 8 2    female pizza      cold          7
## 9 2    female steak      hot          6
## 10 2   female steak      cold          3
## # ... with 26 more rows
```

Notice that the new column `occasion` is of the type character. We need it to be a factor. Use the code below to do so:

```
analytic_data_nway_tidy <- analytic_data_nway_tidy %>%
  mutate(food = as_factor(food_type),
         temperature = as_factor(temperature))
```

You can confirm that occasion is now a factor with the `glimpse()` command. Once this is complete, you are done preparing your one-way within participant analytic data.

```
glimpse(analytic_data_nway_tidy)
```

```
## Rows: 36
## Columns: 6
## $ id      <fct> 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2,...
## $ sex     <fct> male, male, male, male, male, male,...
## $ food_type <chr> "pizza", "pizza", "steak", "steak",...
## $ temperature <fct> hot, cold, hot, cold, hot, cold, ho...
## $ taste    <dbl> 7, 6, 6, 3, 7, 4, 8, 7, 6, 3, 8, 3,...
## $ food     <fct> pizza, pizza, steak, steak, burger,...
```

You now have two data sets `analytic_data_within_nway` and `analytic_data_nway_tidy`. You can calculate descriptive statistics, correlations and general cross-sectional analyses using the `analytic_data_within_nway` data set. If you want to conduct a repeated measures ANOVA you use the `analytic_data_nway_tidy` data set. Both data sets are now ready for analysis.

3.11 Surveys: Single Occasion

This section outlines a workflow appropriate for when you have cross-sectional single occasion survey data. The data corresponds to a design where the researcher has measured, age, sex, eye color, self-esteem, and job satisfaction. Two of these, self-esteem and job satisfaction, were measured with multi-item scales with reverse-keyed items.

To Begin:

- Use the Files tab to confirm you have the data: `data_item_scoring.csv`
- Start a new script for this example. Don't forget to start the script name with "script_".

```
# Date: YYYY-MM-DD
# Name: your name here
# Example: Single occasion survey

# Load data
library(tidyverse)
```

```
my_missing_value_codes <- c("-999", "", "NA")

raw_data_survey <- read_csv(file = "data_item_scoring.csv",
                             na = my_missing_value_codes)

## Parsed with column specification:
## cols(
##   id = col_double(),
##   age = col_double(),
##   sex = col_character(),
##   eye_color = col_character(),
##   esteem1 = col_double(),
##   esteem2 = col_double(),
##   esteem3 = col_double(),
##   esteem4 = col_double(),
##   esteem5_rev15 = col_double(),
##   jobsat1 = col_double(),
##   jobsat2_rev15 = col_double(),
##   jobsat3 = col_double(),
##   jobsat4 = col_double(),
##   jobsat5 = col_double()
## )
```

We load the initial data into a `raw_data_survey` but immediately make a copy we will work with called `analytic_data_survey`. It's good to keep a copy of the raw data for reference if you encounter problems.

```
analytic_data_survey <- raw_data_survey
```

Remove empty row and columns from your data using the `remove_empty_cols()` and `remove_empty_rows()`, respectively. As well, clean the names of your columns to ensure they conform to tidyverse naming conventions.

```
library(janitor)

# Initial cleaning
analytic_data_survey <- analytic_data_survey %>%
  remove_empty("rows") %>%
  remove_empty("cols") %>%
  clean_names()
```

You can confirm the column names following our naming convention with the `glimpse` command - and see the data type for each column.

```
glimpse(analytic_data_survey)
```

```
## Rows: 300
## Columns: 14
## $ id          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11...
## $ age         <dbl> 23, 22, 18, 23, 22, 17, 23, 22, 1...
## $ sex         <chr> "male", "female", "male", "female...
## $ eye_color   <chr> "blue", "brown", "hazel", "blue",...
## $ esteem1     <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, 3, ...
## $ esteem2     <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 2, ...
## $ esteem3     <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3, 4,...
## $ esteem4     <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, NA,...
## $ esteem5_rev15 <dbl> 2, 2, 2, 2, 2, NA, NA, 2, 2, 2, 3...
## $ jobsat1     <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, 3, ...
## $ jobsat2_rev15 <dbl> 1, 1, 1, NA, 1, 1, 2, 1, 2, 2, 3,...
## $ jobsat3     <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ jobsat4     <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4, N...
## $ jobsat5     <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5, 4,...
```

3.11.1 Creating factors

Following initial cleaning, we identify categorical variables as factors. If you plan to conduct an ANOVA - it's critical that all predictor variables are converted to factors. Inspect the `glimpse()` output - if you followed our data entry naming conventions, categorical variables should be of the type character.

```
glimpse(analytic_data_survey)
```

```
## Rows: 300
## Columns: 14
## $ id          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11...
## $ age         <dbl> 23, 22, 18, 23, 22, 17, 23, 22, 1...
## $ sex         <chr> "male", "female", "male", "female...
## $ eye_color   <chr> "blue", "brown", "hazel", "blue",...
## $ esteem1     <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, 3, ...
## $ esteem2     <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 2, ...
## $ esteem3     <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3, 4,...
## $ esteem4     <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, NA,...
## $ esteem5_rev15 <dbl> 2, 2, 2, 2, 2, NA, NA, 2, 2, 2, 3...
## $ jobsat1     <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, 3, ...
```



```
## $ jobsat2_rev15 <dbl> 1, 1, 1, NA, 1, 1, 2, 1, 2, 2, 3,...
## $ jobsat3      <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ jobsat4      <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4, N...
## $ jobsat5      <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5, 4,...
```

We have two variables, `sex` and `eye_color`, that are categorical variable of type character (i.e., `chr`). The participant id column is categorical as well, but of type double (i.e., `dbl`) which is a numeric column. You can quickly convert all character columns to factors using the code below:

```
analytic_data_survey <- analytic_data_survey %>%
  mutate(across(.cols = where(is.character),
    .fns = as_factor))
```

The participant identification number in the `id` column is a numeric column, so we have to handle that column on its own.

```
analytic_data_survey <- analytic_data_survey %>%
  mutate(id = as_factor(id))
```

You can ensure all of these columns are now factors using the `glimpse()` command.

```
glimpse(analytic_data_survey)
```

```
## Rows: 300
## Columns: 14
## $ id      <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11...
## $ age     <dbl> 23, 22, 18, 23, 22, 17, 23, 22, 1...
## $ sex     <fct> male, female, male, female, male,...
## $ eye_color <fct> blue, brown, hazel, blue, NA, haz...
## $ esteem1 <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, 3, ...
## $ esteem2 <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 2, ...
## $ esteem3 <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3, 4,...
## $ esteem4 <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, NA,...
## $ esteem5_rev15 <dbl> 2, 2, 2, 2, 2, NA, NA, 2, 2, 2, 3...
## $ jobsat1  <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, 3, ...
## $ jobsat2_rev15 <dbl> 1, 1, 1, NA, 1, 1, 2, 1, 2, 2, 3,...
## $ jobsat3  <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ jobsat4  <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4, N...
## $ jobsat5  <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5, 4,...
```

Inspect the output of the `glimpse()` command and make sure you have converted all categorical variables to factors - especially those you will use as predictors.

Note: If you have factors like sex that have numeric data in the column (e.g., 1 and 2) instead of male/female you need to handle the situation differently. The preceding section, Experiment: Within N-way, illustrates how to handle this scenario.

3.11.2 Factor screening

Inspect the levels of each factor carefully. Make sure the factor levels of each variable are correct. Examine spelling and look for additional unwanted levels. For example, you wouldn't want to have the following levels for sex: male, mmale, female. Obviously, mmale is an incorrectly typed version of male. Scan all the factors in your data for erroneous factor levels. The code below displays the factor levels:

```
analytic_data_survey %>%
  select(where(is.factor)) %>%
  summary()
```

```
##           id           sex           eye_color
##  1      :  1    male      :147    blue : 99
##  2      :  1   female   :149   brown: 98
##  3      :  1  intersex:  2    hazel:100
##  4      :  1    NA's     :  2    NA's :  3
##  5      :  1
##  6      :  1
## (Other):294
```

Also inspect the output of the above `summary()` command paying attention to the order of the levels in the factors. The order influences how text output and graphs are generated. In these data, the sex column has two levels: male and female in that order. Below we adjust the order of the sex variable because we want the x-axis of a future graph to display columns in the left to right order: female, male.

```
analytic_data_survey <- analytic_data_survey %>%
  mutate(sex = fct_relevel(sex,
                           "intersex",
                           "female",
                           "male"))
```

For eye color, we want a future graph to have the most common eye colors on the left so we reorder the factor levels:

```
analytic_data_survey <- analytic_data_survey %>%
  mutate(eye_color = fct_infreq(eye_color))
```

You can see the new order of the factor levels with `summary()`:

```
analytic_data_survey %>%
  select(where(is.factor)) %>%
  summary()
```

```
##          id          sex      eye_color
##  1      :  1  intersex:  2    hazel:100
##  2      :  1   female :149    blue : 99
##  3      :  1    male   :147   brown: 98
##  4      :  1   NA's    :  2   NA's :  3
##  5      :  1
##  6      :  1
## (Other):294
```

3.11.3 Numeric screening

For numeric variables, it's important to find and remove impossible values. For example, in the context of this example you want to ensure none of the Likert responses are impossible (e.g., outside the 1- to 5-point rating scale) or clearly data entry errors.

Because we have several numeric columns that we are screening, we use the `skim()` command from the `skimr` package. The `skim()` command quickly provides basic descriptive statistics. In the output for this command there are also several columns that begin with `p`: `p0`, `p25`, `p50`, `p75`, and `p100` (`p25` and `p75` omitted in output due to space). These columns correspond to the 0th, 25th, 50th, 75th, and 100th percentiles, respectively. The minimum and maximum values for the data column are indicated under the `p0` and `p100` labels. The median is the 50th percentile (`p50`). The interquartile range is the range between `p25` and `p75`.

Start by examining the range of non-scale items. In this case it's only age. Examine the output to see if any of the age values are unreasonable. As noted, in the output `p0` and `p100` indicate the 0th percentile and the 100th percentile; that is the minimum and maximum values for the variable. Check to make sure none of the age values are unreasonably low or high. If they are, you may need to check the original data source or replace them with missing values.

```
library(skimr)
analytic_data_survey %>%
  select(age) %>%
  skim()
```

```
##   skim_variable n_missing mean   sd p0 p50 p100
## 1             age          3 20.52 2.05 17  20   24
```

With respect to the multi-item scales, it makes sense to look at sets of items rather than all of the items at once. This is because sometimes items from different scales use different response ranges. For example, one measure might use a response scale with a range from 1 to 5; whereas another measure might use a response scale with a range from 1 to 7. This is undesirable from a psychometric point of view, as discussed previously, but if it happens in your data - look at the scale items separately to make it easy to see out of range values.

We begin by looking at the items in the first scale, self-esteem. Possible items responses for this scale range from 1 to 5; make sure all responses are in this range. If any values fall outside this range, you may need to check the original data source or replace them with missing values - as described previously.

```
analytic_data_survey %>%
  select(starts_with("esteem")) %>%
  skim()
```

```
##   skim_variable n_missing mean   sd p0 p50 p100
## 1      esteem1         24 3.39 0.54  3  3    5
## 2      esteem2         28 2.35 0.48  2  2    3
## 3      esteem3         31 3.96 0.37  3  4    5
## 4      esteem4         15 3.54 0.50  3  4    4
## 5 esteem5_rev15         35 2.22 0.47  1  2    3
```

Follow the same process for the job satisfaction items. Write that code on your own now.

Possible item responses for the job satisfaction scale range from 1 to 5, make sure all responses are in this range. If any values fall outside this range, you may need to check the original data source or replace them with missing values - as described previously.

```
analytic_data_survey %>%
  select(starts_with("jobsat")) %>%
  skim()
```

```
##   skim_variable n_missing mean   sd p0 p50 p100
## 1      jobsat1         25 3.34 0.51  3  3    5
## 2 jobsat2_rev15         27 1.51 0.61  1  1    3
## 3      jobsat3         28 2.84 0.37  2  3    3
## 4      jobsat4         35 4.29 0.70  3  4    5
## 5      jobsat5         24 4.57 0.61  3  5    5
```

3.11.4 Scale scores

For each person, scale scores involve averaging scores from several items to create an overall score. The first step in the creation of scales is correcting the values of any reverse-keyed items.

3.11.4.1 Reverse-key items

The way you deal with reverse-keyed items depends on how you scored them. Imagine you had a 5-point scale. You could have scored the scale with the values 1, 2, 3, 4, and 5. Alternatively, you could have scored the scale with the values 0, 1, 2, 3, and 4. The mathematical approach you use to correcting reverse-keyed items depends upon whether the scale starts with 1 or 0.

In this example, we scored the data using the value 1 to 5; so that is the approach illustrated here. See the extra information box for details on how to fixed reverse-keyed items when the scale begins with zero.

In this data file all the reverse-keyed items were identified with the suffix “_rev15” in the column names. This suffix indicates the item was reverse keyed and that the original scale used the response points 1 to 5. We can see those items with the `glimpse()` command below. Notice that there are two reverse-keyed items - each on difference scales.

```
analytic_data_survey %>%
  select(ends_with("_rev15")) %>%
  glimpse()
```

```
## Rows: 300
## Columns: 2
## $ esteem5_rev15 <dbl> 2, 2, 2, 2, 2, NA, NA, 2, 2, 2, 3...
## $ jobsat2_rev15 <dbl> 1, 1, 1, NA, 1, 1, 2, 1, 2, 2, 3...
```

To correct a reverse-keyed item where the lowest possible rating is 1 (i.e, 1 on a 1 to 5 scale), we simply subtract all the scores from a value one more than the highest point possible on the scale (i.e., one more than 5). For example, if

a 1 to 5 response scale was used we subtract each response from 6 to obtain the recoded value.

Original value	Math	Recoded value
1	6 - 1	5
2	6 - 2	4
3	6 - 3	3
4	6 - 4	2
5	6 - 5	1

The code below:

- selects columns that end with "_rev15" (i.e., both esteem and jobsat scales)
- subtracts the values in those columns from 6
- renames the columns by removing "_rev15" from the name because the reverse coding is complete

```
analytic_data_survey <- analytic_data_survey %>%
  mutate(6 - across(.cols = ends_with("_rev15"))) %>%
  rename_with(.fn = str_remove,
    .cols = ends_with("_rev15"),
    pattern = "_rev15")
```

You can use the `glimpse()` command to see the result of your work. If you compare these new values to those obtained from the previous `glimpse()` command you can see they have changed. Also notice the column names no longer indicate the items are reverse keyed.

```
glimpse(analytic_data_survey)
```

```
## Rows: 300
## Columns: 14
## $ id      <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
## $ age     <dbl> 23, 22, 18, 23, 22, 17, 23, 22, 17, N...
## $ sex     <fct> male, female, male, female, male, fem...
## $ eye_color <fct> blue, brown, hazel, blue, NA, hazel, ...
## $ esteem1 <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, 3, 4, N...
## $ esteem2 <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 2, 2, N...
## $ esteem3 <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3, 4, 4, ...
## $ esteem4 <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, NA, 4, ...
## $ esteem5 <dbl> 4, 4, 4, 4, 4, NA, NA, 4, 4, 4, 3, 4, ...
## $ jobsat1 <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, 3, 4, 4...
## $ jobsat2 <dbl> 5, 5, 5, NA, 5, 5, 4, 5, 4, 4, 3, 5, ...
```

```
## $ jobsat3 <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ jobsat4 <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4, NA, 5...
## $ jobsat5 <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5, 4, NA,...
```



- If your scale had used response options numbered 0 to 4 the math is different. For each item you would use subtract values from the highest possible point (i.e, 4) instead of one larger than the highest possible point.

Original value	Math	Recoded value
0	4 - 0	4
1	4 - 1	3
2	4 - 2	2
3	4 - 3	1
4	4 - 4	0

Thus, the mutate command would instead be:

```
mutate(4 - across(.cols = ends_with("__rev15")) )
```

3.11.4.2 Creating scores

The process we use for creating scale scores deletes item-level data from `analytic_data_survey`. This is a desirable aspect of the process because it removes information that we are no longer interested in from our analytic data. That said, before we create scale score, we create a backup on the item-level data called `analytic_data_survey_items`. We will need to use this backup later to compute the reliability of the scales we are creating.

```
analytic_data_survey_items <- analytic_data_survey
```

We want to make a `self_esteem` scale and plan to select items using `starts_with("esteem")`. But prior to doing this we make sure the `start_with()` command only gives us the items we want - and not additional unwanted items. The output below confirms there are not problems associated with using `starts_with("esteem")`.

```
analytic_data_survey %>%
```

```
select(starts_with("esteem")) %>%
  glimpse()
```

```
## Rows: 300
## Columns: 5
## $ esteem1 <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, 3, 4, NA,...
## $ esteem2 <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 2, 2, NA,...
## $ esteem3 <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3, 4, 4, 4,...
## $ esteem4 <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, NA, 4, 3,...
## $ esteem5 <dbl> 4, 4, 4, 4, 4, NA, NA, 4, 4, 4, 3, 4, 4...
```

Likewise, we want to make a `job_sat` scale and plan to select items using `starts_with("jobsat")`. The code and output below using `starts_with("jobsat")` only returns the items we are interested in.

```
analytic_data_survey %>%
  select(starts_with("jobsat")) %>%
  glimpse()
```

```
## Rows: 300
## Columns: 5
## $ jobsat1 <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, 3, 4, 4, ...
## $ jobsat2 <dbl> 5, 5, 5, NA, 5, 5, 4, 5, 4, 4, 3, 5, 3,...
## $ jobsat3 <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ jobsat4 <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4, NA, 5, ...
## $ jobsat5 <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5, 4, NA, 4...
```

We calculate the scale scores using the `rowwise()` command. The `mean()` command provides the mean of columns by default - not people. We use the `rowwise()` command in the code below to make the `mean()` command work across columns (within participants) rather than within columns. The `mutate` command calculates the scale score for each person. The `c_across()` command combined with the `starts_with()` command ensures the items we want averaged together are the items that are averaged together. Notice there is a separate `mutate` line for each scale. The `ungroup()` command turns off the `rowwise()` command. We end the code block by removing the item-level data from the data set.

Important: Take note of how we name the scale variables (e.g., `self_esteem`, `job_sat`). We use a slightly different convention than our items. That is, these scale labels were picked so that they would *not* be selected by a `starts_with("esteem")` or `starts_with("jobsat")`. Why - because we later use those commands to remove the item-level data. We would want the command designed to remove the item-level data to also remove the scale we just cal-

culated! This example illustrates how carefully you need to think about your naming conventions.

```
analytic_data_survey <- analytic_data_survey %>%
  rowwise() %>%
  mutate(self_esteem = mean(c_across(starts_with("esteem")),
                             na.rm = TRUE)) %>%
  mutate(job_sat = mean(c_across(starts_with("jobsat")),
                         na.rm = TRUE)) %>%
  ungroup() %>%
  select(-starts_with("esteem")) %>%
  select(-starts_with("jobsat"))
```

We can see our data now has the `self_esteem` column, a `job_sat` column, and that all of the item-level data has been removed.

```
glimpse(analytic_data_survey)
```

```
## Rows: 300
## Columns: 6
## $ id      <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, ...
## $ age     <dbl> 23, 22, 18, 23, 22, 17, 23, 22, 17, ...
## $ sex     <fct> male, female, male, female, male, f...
## $ eye_color <fct> blue, brown, hazel, blue, NA, hazel...
## $ self_esteem <dbl> 3.200, 3.800, 3.800, 3.000, 3.400, ...
## $ job_sat  <dbl> 4.00, 5.00, 4.40, 3.50, 4.00, 3.80, ...
```

You now have two data sets `analytic_data_survey` and `analytic_data_survey_items`. You can calculate descriptive statistics, correlations and most analyses using the `analytic_data_survey`. To obtain the reliability of the scales you just created though you will need to use the `analytic_data_survey_items`. Both sets of data are ready for analysis.

3.12 Surveys: Multiple Occasions

This section outlines a workflow appropriate for when you have multiple occasion survey data. The data corresponds to a design where the researcher has measured, age, sex, eye color, self-esteem, and job satisfaction at each of two times points. Self-esteem and job satisfaction were measured with multi-item scales with reverse-keyed items.

To Begin:

- Use the Files tab to confirm you have the data: data_item_time.csv
- Start a new script for this example. Don't forget to start the script name with "script_".

```
# Date: YYYY-MM-DD
# Name: your name here
# Example: Multiple occasion survey

# Load data
library(tidyverse)

my_missing_value_codes <- c("-999", "", "NA")

raw_data_occasions <- read_csv(file = "data_item_time.csv",
                               na = my_missing_value_codes)

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   sex = col_character(),
##   eye_color = col_character()
## )

## See spec(...) for full column specifications.
```

We load the initial data into a raw_data_occasions but immediately make a copy we will work with called analytic_data_occasions. It's good to keep a copy of the raw data for reference if you encounter problems.

```
analytic_data_occasions <- raw_data_occasions
```

Remove empty rows and columns from your data using the remove_empty("rows") and remove_empty("cols"), respectively. As well, clean the names of your columns to ensure they conform to tidyverse naming conventions.

```
library(janitor)

# Initial cleaning
analytic_data_occasions <- analytic_data_occasions %>%
  remove_empty("rows") %>%
```

```
remove_empty("cols") %>%
clean_names()
```

You can confirm the column names following our naming convention with the `glimpse` command - and see the data type for each column.

```
glimpse(analytic_data_occasions)

## Rows: 300
## Columns: 24
## $ id          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,...
## $ age         <dbl> 23, 22, 18, 23, 22, 17, 23, 22...
## $ sex         <chr> "male", "female", "male", "fem...
## $ eye_color   <chr> "blue", "brown", "hazel", "blu...
## $ t1_esteem1  <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, ...
## $ t1_esteem2  <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, ...
## $ t1_esteem3  <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3,...
## $ t1_esteem4  <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, ...
## $ t1_esteem5_rev15 <dbl> 2, 2, 2, 2, 2, NA, NA, 2, 2, 2...
## $ t1_jobsat1  <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, ...
## $ t1_jobsat2_rev15 <dbl> 1, 1, 1, NA, 1, 1, 2, 1, 2, 2,...
## $ t1_jobsat3  <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ t1_jobsat4  <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4...
## $ t1_jobsat5  <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5,...
## $ t2_esteem1  <dbl> 4, 5, 5, 4, NA, 4, 4, 5, 5, 5,...
## $ t2_esteem2  <dbl> 3, 4, 4, 3, 3, 4, 3, 4, 4, 4, ...
## $ t2_esteem3  <dbl> 5, 5, 5, 4, 5, 5, 3, 5, 5, 4, ...
## $ t2_esteem4  <dbl> 4, 5, 5, 4, 5, 5, 5, 5, 4, 5, ...
## $ t2_esteem5_rev15 <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ t2_jobsat1  <dbl> 4, 6, 5, 4, 4, 4, 4, 6, 4, NA,...
## $ t2_jobsat2_rev15 <dbl> 2, 2, 2, 3, 2, 2, 3, 2, 3, 3, ...
## $ t2_jobsat3  <dbl> 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, ...
## $ t2_jobsat4  <dbl> 3, 6, 6, 5, 5, 5, 5, 6, 3, 5, ...
## $ t2_jobsat5  <dbl> 6, 3, 6, 5, NA, 5, 5, 6, 6, 6,...
```

3.12.1 Creating factors

Following initial cleaning, we identify categorical variables as factors. If you plan to conduct an ANOVA - it's critical that all predictor variables are converted to factors. Inspect the `glimpse()` output - if you followed our data entry naming conventions, categorical variables should be of the type character.

```
glimpse(analytic_data_occasions)
```

```
## Rows: 300
## Columns: 24
## $ id          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,...
## $ age         <dbl> 23, 22, 18, 23, 22, 17, 23, 22...
## $ sex         <chr> "male", "female", "male", "fem...
## $ eye_color   <chr> "blue", "brown", "hazel", "blu...
## $ t1_esteem1  <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, ...
## $ t1_esteem2  <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, ...
## $ t1_esteem3  <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3,...
## $ t1_esteem4  <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, ...
## $ t1_esteem5_rev15 <dbl> 2, 2, 2, 2, 2, NA, NA, 2, 2, 2...
## $ t1_jobsat1  <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, ...
## $ t1_jobsat2_rev15 <dbl> 1, 1, 1, NA, 1, 1, 2, 1, 2, 2,...
## $ t1_jobsat3  <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ t1_jobsat4  <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4...
## $ t1_jobsat5  <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5,...
## $ t2_esteem1  <dbl> 4, 5, 5, 4, NA, 4, 4, 5, 5, 5,...
## $ t2_esteem2  <dbl> 3, 4, 4, 3, 3, 4, 3, 4, 4, 4, ...
## $ t2_esteem3  <dbl> 5, 5, 5, 4, 5, 5, 3, 5, 5, 4, ...
## $ t2_esteem4  <dbl> 4, 5, 5, 4, 5, 5, 5, 5, 4, 5, ...
## $ t2_esteem5_rev15 <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ t2_jobsat1  <dbl> 4, 6, 5, 4, 4, 4, 4, 6, 4, NA,...
## $ t2_jobsat2_rev15 <dbl> 2, 2, 2, 3, 2, 2, 3, 2, 3, 3, ...
## $ t2_jobsat3  <dbl> 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, ...
## $ t2_jobsat4  <dbl> 3, 6, 6, 5, 5, 5, 5, 6, 3, 5, ...
## $ t2_jobsat5  <dbl> 6, 3, 6, 5, NA, 5, 5, 6, 6, 6,...
```

We have two variables, `sex` and `eye_color`, that are categorical variable of type character (i.e., `chr`). The participant id column is categorical as well, but of type double (i.e., `dbl`) which is a numeric column. You can quickly convert all character columns to factors using the code below:

```
analytic_data_occasions <- analytic_data_occasions %>%
  mutate(across(.cols = where(is.character),
    .fns = as_factor))
```

The participant identification number in the `id` column is numeric, so we have to handle that column on its own.

```
analytic_data_occasions <- analytic_data_occasions %>%
  mutate(id = as_factor(id))
```

You can ensure all of these columns are now factors using the `glimpse()` command.

```
glimpse(analytic_data_occasions)
```

```
## Rows: 300
## Columns: 24
## $ id          <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,...
## $ age         <dbl> 23, 22, 18, 23, 22, 17, 23, 22...
## $ sex         <fct> male, female, male, female, ma...
## $ eye_color   <fct> blue, brown, hazel, blue, NA, ...
## $ t1_esteem1  <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, ...
## $ t1_esteem2  <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, ...
## $ t1_esteem3  <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3,...
## $ t1_esteem4  <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, ...
## $ t1_esteem5_rev15 <dbl> 2, 2, 2, 2, 2, NA, NA, 2, 2, 2...
## $ t1_jobsat1  <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, ...
## $ t1_jobsat2_rev15 <dbl> 1, 1, 1, NA, 1, 1, 2, 1, 2, 2,...
## $ t1_jobsat3  <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ t1_jobsat4  <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4...
## $ t1_jobsat5  <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5,...
## $ t2_esteem1  <dbl> 4, 5, 5, 4, NA, 4, 4, 5, 5, 5,...
## $ t2_esteem2  <dbl> 3, 4, 4, 3, 3, 4, 3, 4, 4, 4, ...
## $ t2_esteem3  <dbl> 5, 5, 5, 4, 5, 5, 3, 5, 5, 4, ...
## $ t2_esteem4  <dbl> 4, 5, 5, 4, 5, 5, 5, 5, 4, 5, ...
## $ t2_esteem5_rev15 <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ t2_jobsat1  <dbl> 4, 6, 5, 4, 4, 4, 4, 6, 4, NA,...
## $ t2_jobsat2_rev15 <dbl> 2, 2, 2, 3, 2, 2, 3, 2, 3, 3, ...
## $ t2_jobsat3  <dbl> 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, ...
## $ t2_jobsat4  <dbl> 3, 6, 6, 5, 5, 5, 5, 6, 3, 5, ...
## $ t2_jobsat5  <dbl> 6, 3, 6, 5, NA, 5, 5, 6, 6, 6,...
```

Inspect the output of the `glimpse()` command and make sure you have converted all categorical variables to factors - especially those you will use as predictors.

Note: If you have factors like `sex` that have numeric data in the column (e.g, 1 and 2) instead of male/female you need to handle the situation differently. The preceding section, Experiment: Within N-way, illustrates how to handle this scenario.

3.12.2 Factor screening

Inspect the levels of each factor carefully. Make sure the factor levels of each variable are correct. Examine spelling and look for additional unwanted levels. For example, you wouldn't want to have the following levels for sex: male, mmale, female. Obviously, mmale is an incorrectly typed version of male. Scan all the factors in your data for erroneous factor levels. The code below displays the factor levels:

```
analytic_data_occasions %>%
  select(where(is.factor)) %>%
  summary()
```

```
##           id           sex      eye_color
##  1      :  1    male      :147    blue : 99
##  2      :  1   female  :149    brown: 98
##  3      :  1   intersex:  2    hazel:100
##  4      :  1    NA's    :  2    NA's :  3
##  5      :  1
##  6      :  1
## (Other):294
```

Also inspect the output of the above `summary()` command paying attention to the order of the levels in the factors. The order influences how text output and graphs are generated. In these data, the sex column has two levels: male and female in that order. Below we adjust the order of the sex variable because we want the x-axis of a future graph to display columns in the left to right order: female, male.

```
analytic_data_occasions <- analytic_data_occasions %>%
  mutate(sex = fct_relevel(sex,
                           "intersex",
                           "female",
                           "male"))
```

For eye color, we want a future graph to have the most common eye colors on the left so we reorder the factor levels:

```
analytic_data_occasions <- analytic_data_occasions %>%
  mutate(eye_color = fct_infreq(eye_color))
```

You can see the new order of the factor levels with `summary()`:

```
analytic_data_occasions %>%
  select(where(is.factor)) %>%
  summary()
```

```
##           id           sex      eye_color
##  1      : 1   intersex: 2   hazel:100
##  2      : 1   female  :149   blue : 99
##  3      : 1   male   :147   brown: 98
##  4      : 1   NA's    : 2   NA's : 3
##  5      : 1
##  6      : 1
## (Other):294
```

3.12.3 Numeric screening

For numeric variables, it's important to find and remove impossible values. For example, in the context of this example you want to ensure none of the Likert responses are impossible (e.g., outside the 1- to 5-point rating scale) or clearly data entry errors.

Because we have several numeric columns that we are screening, we use the `skim()` command from the `skimr` package. The `skim()` command quickly provides basic descriptive statistics. In the output for this command there are also several columns that begin with p: p0, p25, p50, p75, and p100 (p25 and p75 are omitted in output due to space). These columns correspond to the 0th, 25th, 50th, 75th, and 100th percentiles, respectively. The minimum and maximum values for the data column are indicated under the p0 and p100 labels. The median is the 50th percentile (p50). The interquartile range is the range between p25 and p75.

Start by examining the range of non-scale items. In this case it's only age. Examine the output to see if any of the age values are unreasonable. As noted, p0 and p100 in the output indicate the 0th percentile and the 100th percentile; that is the minimum and maximum values for the variable. Check to make sure none of the age values are unreasonably low or high. If they are, you may need to check the original data source or replace them with missing values.

```
library(skimr)
analytic_data_occasions %>%
  select(age) %>%
  skim()
```

```
##   skim_variable n_missing mean   sd p0 p50 p100
```

```
## 1          age          3 20.52 2.05 17  20  24
```

With respect to the multi-item scales, it makes sense to look at sets of items rather than all of the items at once. This is because sometimes items from different scales use different response ranges. For example, one measure might use a response scale with a range from 1 to 5; whereas another measure might use a response scale with a range from 1 to 7. This is undesirable from a psychometric point of view, as discussed previously, but if it happens in your data - look at the scale items separately to make it easy to see out of range values.

We begin by looking at the items in the first scale, self-esteem. Possible item responses for this scale range from 1 to 5. Make sure all responses are in this range. If any values fall outside this range, you may need to check the original data source or replace them with missing values - as described previously.

Because we want to select the self-esteem items from both time 1 and time 2 we cannot use the `starts_with()` command. Instead we use the `contains()` command in the code below.

```
analytic_data_occasions %>%
  select(contains("esteem")) %>%
  skim()
```

```
##      skim_variable n_missing mean    sd p0 p50 p100
## 1      t1_esteem1      24 3.39 0.54  3  3    5
## 2      t1_esteem2      28 2.35 0.48  2  2    3
## 3      t1_esteem3      31 3.96 0.37  3  4    5
## 4      t1_esteem4      15 3.54 0.50  3  4    4
## 5 t1_esteem5_rev15      35 2.22 0.47  1  2    3
## 6      t2_esteem1       5 4.27 0.64  3  4    6
## 7      t2_esteem2       5 3.33 0.47  3  3    4
## 8      t2_esteem3       6 4.77 0.69  3  5    6
## 9      t2_esteem4       3 4.46 0.59  3  5    5
## 10 t2_esteem5_rev15     4 3.19 0.45  2  3    4
```

Follow the same process for the job satisfaction items. Possible item responses for the job satisfaction scale range from 1 to 5, make sure all responses are in this range. If any values fall outside this range, you may need to check the original data source or replace them with missing values - as described previously.

```
analytic_data_occasions %>%
  select(contains("jobsat")) %>%
  skim()
```


##	skim_variable	n_missing	mean	sd	p0	p50	p100
## 1	t1_jobsat1	25	3.34	0.51	3	3	5
## 2	t1_jobsat2_rev15	27	1.51	0.61	1	1	3
## 3	t1_jobsat3	28	2.84	0.37	2	3	3
## 4	t1_jobsat4	35	4.29	0.70	3	4	5
## 5	t1_jobsat5	24	4.57	0.61	3	5	5
## 6	t2_jobsat1	2	4.23	0.62	3	4	6
## 7	t2_jobsat2_rev15	3	2.54	0.59	2	2	4
## 8	t2_jobsat3	5	3.76	0.43	3	4	4
## 9	t2_jobsat4	3	5.03	0.99	3	5	6
## 10	t2_jobsat5	3	5.36	0.92	3	6	6

3.12.4 Scale scores

For each person, scale scores involve averaging scores from several items to create an overall score. The first step in the creation of scales is correcting the values of any reverse-keyed items.

3.12.4.1 Reverse-key items

The way you deal with reverse-keyed items depends on how you scored them. Imagine you had a 5-point scale. You could have scored the scale with the values 1, 2, 3, 4, and 5. Alternatively, you could have scored the scale with the values 0, 1, 2, 3, and 4. The mathematical approach you use to correcting reverse-keyed items depends upon whether the scale starts with 1 or 0.

In this example, we scored the Likert items using the values 1 to 5. Therefore, we use the reverse keying approach for scales that begin with 1. The preceding section, “Surveys: Single occasion”, describes how the math differs when the response scale starts with 0. We encourage you to read that section before going further if you have not done so already.

In this data file all the reverse-keyed items were identified with the suffix “_rev15” in the column names. This suffix indicates the item was reverse keyed and that the original scale used the response points 1 to 5. We can see those items with the `glimpse()` command below. Notice that there are two reverse-keyed items - each on difference scales.

```
analytic_data_survey %>%
  select(ends_with("_rev15")) %>%
  glimpse()
```

```
## Rows: 300
## Columns: 0
```

To correct reverse-keyed items where the lowest possible rating is 1 (i.e., 1 on a 1 to 5 scale), we simply subtract all the scores from a value one more than the highest possible rating (i.e., 6).

The code below:

- selects columns that end with "_rev15" (i.e., both esteem and jobsat scales)
- subtracts the values in those columns from 6
- renames the columns by removing "_rev15" from the name because the reverse coding is complete

```
analytic_data_occasions <- analytic_data_occasions %>%
  mutate(6 - across(.cols = ends_with("_rev15")) ) %>%
  rename_with(.fn = str_remove,
             .cols = ends_with("_rev15"),
             pattern = "_rev15")
```

You can use the `glimpse()` command to see the result of your work. If you compare these new values to those obtained from the previous `glimpse()` command you can see they have changed. Also notice the column names no longer indicate the items are reverse keyed.

3.12.4.2 Creating scores

The process we use for creating scale scores deletes item-level data from `analytic_data_survey`. This is a desirable aspect of the process because it removes information we no longer need. That said, before we create scale scores, we create a backup on the item-level data in `analytic_data_survey` called `analytic_data_survey_items`. We will need to use this backup later to compute the reliability of the scales we are creating.

```
analytic_data_occasions_items <- analytic_data_occasions
```

We want to make a self-esteem scale and plan to select items using `starts_with("t1_esteem")`. Prior to doing this we make sure the `start_with()` command only gives us the items we want - and not additional unwanted items. The output below confirms there are not problems associated with using `starts_with("t1_esteem")`.

```
analytic_data_occasions %>%
  select(starts_with("t1_esteem")) %>%
  glimpse()
```

```
## Rows: 300
```

```
## Columns: 5
## $ t1_esteem1 <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, 3, 4, ...
## $ t1_esteem2 <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 2, 2, ...
## $ t1_esteem3 <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3, 4, 4, ...
## $ t1_esteem4 <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, NA, 4, ...
## $ t1_esteem5 <dbl> 4, 4, 4, 4, 4, NA, NA, 4, 4, 4, 3, 4, ...
```

Repeat this set of commands using `t2_esteem`, `t1_jobsat`, and `t2_jobsat`. Make sure that those `start_with()` terms select only the relevant items and not others.

We calculate the scale scores using the `rowwise()` command. The `mean()` command provides the mean of columns by default - not people. We use the `rowwise()` command in the code below to make the `mean()` command work across columns (within participants) rather than within columns. The `mutate` command calculates the scale score for each person. The `c_across()` command combined with the `starts_with()` command ensures the items we want averaged together are the items that are averaged together. Notice that there is a separate `mutate` line for each scale. The `ungroup()` command turns off the `rowwise()` command. We end the code block by removing the item-level data from the data set.

Important: Take note of how the names of the scale variables (e.g., `esteem_t1`, `jobsat_t1`) use a slightly different convention than our items. That is, these scale labels were picked so that they would *not* be selected by a `starts_with("t1_esteem")` or `starts_with("t1_jobsat")`. Why - because we later use those commands to remove the item-level data. We would want the command designed to remove the item-level data to also remove the scale we just calculated! This example illustrates how carefully you need to think about your naming conventions.

```
analytic_data_occasions <- analytic_data_occasions %>%
  rowwise() %>%
  mutate(esteem_t1 = mean(c_across(starts_with("t1_esteem")),
                           na.rm = TRUE)) %>%
  mutate(esteem_t2 = mean(c_across(starts_with("t2_esteem")),
                           na.rm = TRUE)) %>%
  mutate(jobsat_t1 = mean(c_across(starts_with("t1_jobsat")),
                           na.rm = TRUE)) %>%
  mutate(jobsat_t2 = mean(c_across(starts_with("t2_jobsat")),
                           na.rm = TRUE)) %>%
  ungroup() %>%
  select(-starts_with("t1_esteem")) %>%
  select(-starts_with("t2_esteem")) %>%
  select(-starts_with("t1_jobsat")) %>%
  select(-starts_with("t2_jobsat"))
```

We can see our data now has the columns `t1_esteem`, `t2_esteem`, `t1_jobsat`, and `t2_jobsat`. As well, we can see that all of the item-level data has been removed from the data set.

```
glimpse(analytic_data_occasions)
```

```
## Rows: 300
## Columns: 8
## $ id      <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
## $ age     <dbl> 23, 22, 18, 23, 22, 17, 23, 22, 17, N...
## $ sex     <fct> male, female, male, female, male, fem...
## $ eye_color <fct> blue, brown, hazel, blue, NA, hazel, ...
## $ esteem_t1 <dbl> 3.200, 3.800, 3.800, 3.000, 3.400, 3...
## $ esteem_t2 <dbl> 3.8, 4.4, 4.4, 3.6, 4.0, 4.2, 3.6, 4...
## $ jobsat_t1 <dbl> 4.00, 5.00, 4.40, 3.50, 4.00, 3.80, 3...
## $ jobsat_t2 <dbl> 4.20, 4.40, 5.00, 4.20, 4.25, 4.40, 4...
```

3.12.5 Pivot to tidy data

The analytic data in its current form does not conform to the tidy data specification. Inspect the data with the `print()` command. Notice that a single column is not used to represent esteem, jobsat, or time. Rather, there are four columns that are a mix of these variables. The consequence of this is that there are two esteem ratings/observations on each row and two jobsat ratings/observations on each row. Tidy data is structured so that each variable is represented in a single column and each observation has its own row.

```
print(analytic_data_occasions)
```

```
## # A tibble: 300 x 8
##   id      age sex  eye_color esteem_t1 esteem_t2 jobsat_t1
##   <fct> <dbl> <fct> <fct>         <dbl>     <dbl>     <dbl>
## 1 1      23 male  blue           3.2        3.8        4
## 2 2      22 fema~ brown          3.8        4.4        5
## 3 3      18 male  hazel          3.8        4.4       4.4
## 4 4      23 fema~ blue           3         3.6       3.5
## 5 5      22 male  <NA>          3.4        4         4
## 6 6      17 fema~ hazel          3.5        4.2       3.8
## 7 7      23 male  blue           3         3.6       3.6
## 8 8      22 fema~ brown          3.8        4.4       4.6
## 9 9      17 male  hazel          3.6        4.2       3.75
```

```
## 10 10      NA fema~ blue      3.6      4.2      3.8
## # ... with 290 more rows, and 1 more variable:
## #   jobsat_t2 <dbl>
```

The `pivot_longer()` command below converts our data to the tidy data format. In this command we specify the columns with data by using `esteem_t1:jobsat_t2`; this selects these two columns and all of the columns between them. Each of these columns represents a dependent variable at a particular time in the format “variable_time” (e.g., `esteem_t1`). The code `names_to = c(“value”, “time”)` explains this format to R. It indicates that the first part of the column name (e.g., `esteem`) contains the name of the variable (expressed in the code as “value”). It also indicates that the second part of the column name represents time. The line `names_sep = “_”` tells the R that the underscore character is used to separate the first part of the name from the second part of the name. When this code is executed it creates a tidy version of data set stored in `analytic_survey_tidy`.

```
analytic_occasion_tidy <- analytic_data_occasions %>%
  pivot_longer(esteem_t1:jobsat_t2,
    names_to = c(“value”, “time”),
    names_sep = “_”)
```

You can see the new data with the `print()` command. Notice that each participant has multiple rows associated with them.

```
print(analytic_occasion_tidy)
```

```
## # A tibble: 600 x 7
##   id    age sex   eye_color time  esteem jobsat
##   <fct> <dbl> <fct>   <fct>    <chr>   <dbl> <dbl>
## 1 1      23 male   blue     t1       3.2    4
## 2 1      23 male   blue     t2       3.8    4.2
## 3 2      22 female brown    t1       3.8    5
## 4 2      22 female brown    t2       4.4    4.4
## 5 3      18 male   hazel    t1       3.8    4.4
## 6 3      18 male   hazel    t2       4.4    5
## 7 4      23 female blue     t1       3      3.5
## 8 4      23 female blue     t2       3.6    4.2
## 9 5      22 male   <NA>     t1       3.4    4
## 10 5     22 male   <NA>     t2       4     4.25
## # ... with 590 more rows
```

You now have three data sets. The data `analytic_occasion_tidy` is appropriate for conducting a repeated measures ANOVA or more complicated analyses. The data `analytic_data_occasions` is appropriate for calculating descriptive

statistics and correlations. The data `analytic_occasions_items` is appropriate for calculating the reliability of the scales you constructed. These data are ready for analysis.

3.13 Basic descriptive statistics

Regardless of the design of the study, most researchers want to see descriptive statistics for the variables in their study. We offer three approaches for obtaining descriptive statistics below. For convenience we use the recent data set `analytic_data_occasions`. But recognize the commands below can be used with all the analytic data sets we created for the various designs.

3.13.1 `skim()`

One approach is the `skim()` command from the `skimr` package. The `skim()` command quickly provides the basic descriptive statistics. In the output for this command there are also several columns that begin with `p`: `p0`, `p25`, `p50`, `p75`, and `p100` (`p25` and `p75` are omitted in output due to space). These columns correspond to the 0th, 25th, 50th, 75th, and 100th percentiles, respectively. The minimum and maximum values for the data column are indicated under the `p0` and `p100` labels. The median is the 50th percentile (`p50`). The interquartile range is the range between `p25` and `p75`. Notice that we run this command on the “wide” version of the data (`analytic_data_occasions`) rather than tidy version of the data (`analytic_occasion_tidy`).

```
library(skimr)
skim(analytic_data_occasions)
```

```
##   skim_variable n_missing mean   sd   p0  p50  p100
## 1          age         3 20.52 2.05 17.0 20.0 24.00
## 2    esteem_t1         0  3.40 0.32  2.5  3.4  4.25
## 3    esteem_t2         0  3.93 0.34  3.2  4.0  4.80
## 4   jobsat_t1         0  3.91 0.43  2.0  4.0  5.00
## 5   jobsat_t2         0  4.37 0.42  3.0  4.4  5.25
```

3.13.2 apa.cor.table()

Another approach is the `apa.cor.table()` command from the `apaTables` package. This quickly provides the basic descriptive statistics as well as correlations among variable. As well, it will even create a Word document with this information, see Figure 3.5. Notice that we run this command on the “wide” version of the data (`analytic_data_occasions`) rather than tidy version of the data (`analytic_occasion_tidy`).

```
library(apaTables)
analytic_data_survey %>%
  select(where(is.numeric)) %>%
  apa.cor.table(filename = "apa_descriptives.doc")
```

Means, standard deviations, and correlations with confidence intervals

Variable	<i>M</i>	<i>SD</i>	1	2	3	4
1. age	20.52	2.05				
2. esteem_t1	3.40	0.32	-.04 [-.15, .08]			
3. esteem_t2	3.93	0.34	.01 [-.10, .13]	.84** [.80, .87]		
4. jobsat_t1	3.91	0.43	-.00 [-.12, .11]	.64** [.56, .70]	.56** [.48, .63]	
5. jobsat_t2	4.37	0.42	-.02 [-.13, .10]	.58** [.50, .65]	.52** [.43, .60]	.82** [.77, .85]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

FIGURE 3.5: Word document created by `apa.cor.table`

3.13.3 tidyverse

A final approach uses tidyverse commands. This approach is oddly long - and we won’t describe how it works in detail. But, based on the information in the previous chapter you should be able to work out how this code works. Even though this code is long - it provide the ultimate in flexibility. If a new statistic is developed that you want to use, you can simply include the command for it in the `desired_descriptives` list and it will be included in your table. Notice that we run this command on the “wide” version of the data (`analytic_data_occasions`) rather than tidy version of the data (`analytic_occasion_tidy`).

```

library(tidyverse)
# HMisc package must be installed.
# Library command not needed for HMisc package.

desired_descriptives <- list(
  mean = ~mean(.x, na.rm = TRUE),
  CI95_LL = ~Hmisc::smean.cl.normal(.x)[2],
  CI95_UL = ~Hmisc::smean.cl.normal(.x)[3],
  sd = ~sd(.x, na.rm = TRUE),
  min = ~min(.x, na.rm = TRUE),
  max = ~max(.x, na.rm = TRUE),
  n = ~sum(!is.na(.x))
)

row_sum <- analytic_data_occasions %>%
  summarise(across(.cols = where(is.numeric),
    .fns = desired_descriptives,
    .names = "{col}___{fn}"))

long_summary <- row_sum %>%
  pivot_longer(cols = everything(),
    names_to = c("var", "stat"),
    names_sep = c("___"),
    values_to = "value")

summary_table <- long_summary %>%
  pivot_wider(names_from = stat,
    values_from = value)

# round to 3 decimals
summary_table_rounded <- summary_table %>%
  mutate(across(.cols = where(is.numeric),
    .fns= round,
    digits = 3)) %>%
  as.data.frame()

print(summary_table_rounded)

```

	var	mean	CI95_LL	CI95_UL	sd	min	max	n
## 1	age	20.522	20.288	20.756	2.048	17.0	24.00	297
## 2	esteem_t1	3.403	3.366	3.440	0.324	2.5	4.25	300
## 3	esteem_t2	3.927	3.889	3.966	0.337	3.2	4.80	300
## 4	jobsat_t1	3.905	3.856	3.955	0.435	2.0	5.00	300
## 5	jobsat_t2	4.368	4.320	4.416	0.425	3.0	5.25	300

3.13.4 Cronbach's alpha

If you want Cronbach's alpha to estimate the reliability of the scale, you can use the `alpha` command from the `psych` package with the code below. Note we have to use the item-level data we previously created a copy of called `analytic_data_survey_items`. The `glimpse()` command illustrates this data set has all the original items (after reverse-key coding has been fixed).

```
analytic_data_survey_items %>%
  glimpse()

## Rows: 300
## Columns: 14
## $ id      <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
## $ age     <dbl> 23, 22, 18, 23, 22, 17, 23, 22, 17, N...
## $ sex     <fct> male, female, male, female, male, fem...
## $ eye_color <fct> blue, brown, hazel, blue, NA, hazel, ...
## $ esteem1 <dbl> 3, 4, 4, 3, 3, 3, 3, 4, 4, 4, 3, 4, N...
## $ esteem2 <dbl> 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 2, 2, N...
## $ esteem3 <dbl> 4, 4, 4, 3, 4, 4, NA, 4, 4, 3, 4, 4, ...
## $ esteem4 <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 3, 4, NA, 4, ...
## $ esteem5 <dbl> 4, 4, 4, 4, 4, NA, NA, 4, 4, 4, 3, 4, ...
## $ jobsat1 <dbl> 3, 5, 4, 3, 3, 3, 3, 5, 3, 3, 3, 4, 4...
## $ jobsat2 <dbl> 5, 5, 5, NA, 5, 5, 4, 5, 4, 4, 3, 5, ...
## $ jobsat3 <dbl> 3, NA, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ jobsat4 <dbl> NA, 5, 5, 4, 4, 4, 4, 5, NA, 4, NA, 5...
## $ jobsat5 <dbl> 5, NA, 5, 4, 5, 4, 4, 5, 5, 5, 4, NA, ...
```

We calculated reliability using `psych::alpha()` command. Cronbach's alpha is labeled "raw alpha" in the output. Cronbach's alpha is an estimate of the proportion of variability in observed scores that is due to actual differences among participants (rather than measurement error). Remember, never use `library(psych)`, it will break the tidyverse packages. Instead, precede all `psych` package commands with `psych::` as we do below with `psych::alpha()`.

```
rxx_alpha <- analytic_data_occasions_items %>%
  select(starts_with("t1_esteem")) %>%
  psych::alpha()

print(rxx_alpha$total)

## raw_alpha std.alpha G6(smc) average_r S/N ase mean
## 0.6622 0.6634 0.6173 0.2827 1.97 0.03035 3.403
## sd median_r
## 0.3239 0.2927
```



4

Populations

4.1 Notation

In this chapter we will use summation notation. If you are not familiar with summation notation, we present a brief overview here.

Consider a scenario where we have the IQ data for three participants. We use the N symbol to represent the number of participants. Because we have three participants $N = 3$. The data for these participants is illustrated in Figure 4.1.

Notice how each person in the data set can be represented by the variable X : the first person by X_1 , the second by X_2 , and the third by X_3 . Often we refer to individuals in a data set by using the variable X accompanied by a subscript (e.g., 1, 2, 3, etc.).

Data Looks Like This		But Think of It Like This	
id	iq	X_1	110
1	110	X_2	120
2	120	X_3	100
3	100		

Each row represents a person

FIGURE 4.1: Data for understanding summation notation

Referring to participants using the variable X and subscript is valuable because it can be used in conjunction with the sigma (i.e., Σ) symbol for summation. Consider the example below in which we use the summation notation to indicate that we want to add all the X values (representing IQ) for the participants. We use a lower case i to represent all possible subscript values. The notation, $i = 1$, below the Σ symbol indicates that we should start with participant 1. The notation, N , above the Σ symbol indicates that we should

iterate i up to the value indicated by N ; in this case 3, because there are three participants.

$$\begin{aligned}\sum_{i=1}^N X_i &= X_1 + X_2 + X_3 \\ &= 110 + 120 + 100 \\ &= 330\end{aligned}$$

Sometimes, to simplify the notation, the numbers above and below the Σ symbol are omitted. Likewise, the i subscript is omitted. There is a general understanding that when these components of the notation are omitted the version of the notation above is implied.

$$\begin{aligned}\sum X &= X_1 + X_2 + X_3 \\ &= 110 + 120 + 100 \\ &= 330\end{aligned}$$

Calculating a mean. The full version of the notation can be used to indicate how an average/mean is calculated.

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^N X_i}{N} \\ &= \frac{X_1 + X_2 + X_3}{3} \\ &= \frac{110 + 120 + 100}{3} \\ &= \frac{330}{3} \\ &= 110\end{aligned}$$

Likewise, the concise version of the notation can be used to indicate how an average/mean is calculated.

$$\begin{aligned}\bar{X} &= \frac{\sum X}{N} \\ &= \frac{X_1 + X_2 + X_3}{3} \\ &= \frac{110 + 120 + 100}{3} \\ &= \frac{330}{3} \\ &= 110\end{aligned}$$

Calculating squared differences. A common task in statistics is to calculate

1) the squared difference between each person and the mean, and 2) add up those squared differences. This calculation is easily expressed with the full version of the notation.

$$\begin{aligned}
 \sum_{i=1}^N (X_i - \bar{X})^2 &= (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2 \\
 &= (110 - 110)^2 + (120 - 110)^2 + (100 - 110)^2 \\
 &= (0)^2 + (10)^2 + (-10)^2 \\
 &= 0 + 100 + 100 \\
 &= 200
 \end{aligned}$$

Likewise, the sum of the squared differences from the mean can be expressed using the concise version of the notation.

$$\begin{aligned}
 \sum (X - \bar{X})^2 &= (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2 \\
 &= (110 - 110)^2 + (120 - 110)^2 + (100 - 110)^2 \\
 &= (0)^2 + (10)^2 + (-10)^2 \\
 &= 0 + 100 + 100 \\
 &= 200
 \end{aligned}$$

4.2 Population vs samples

As we move closer to conducting our own research it is critical to make a distinction between populations and samples. A population is the complete set of people/animals about which we want to make conclusions. A sample is a randomly selected subset of the population. In most scenarios it is impractical to work with an entire population and, for practical reasons, we study a subset of the population called a sample.

Researchers, and consumers of research, typically have little interest in making conclusions at the sample level. In general, we care about conclusions that generalize to the population but not conclusions that only apply to specific individuals in the sample. Consider the case of COVID-19. Imagine a research team creates a vaccine that they hope generates immunity to COVID-19. We care very little if the immunity only works for the specific individuals in the study. However, we care a great deal if the immunity works, or is likely to work, for all Canadians or all humans. Thus, we study samples but typically wish to make conclusions that apply to the population.

Statistic tests are a means of helping researchers use sample data to make

conclusions at the population level. When we calculate a number that summarises an attribute of all of the people/animals in the population we refer to it as a **parameter**.

4.3 A small population

In this section we review how to calculate three commonly used population parameters (mean, variance, and standard deviation). Populations are typically quite large but for simplicity we focus on a population composed of the weights of just three chocolate chip cookies. We refer to the three cookies as X_1 , X_2 , and X_3 . The cookies have the weights of 8, 10, and 12 grams, respectively.

4.3.1 Mean (μ)

It can be helpful to create a model that describes our data. Of course, the model won't describe every participant perfectly and each participant will differ to some extent from the model.

Model: To create a model we first need data, which in this example will be the weight of three different chocolate chip cookies. As mentioned previously, the weights of the three cookies are designated by X_1 , X_2 , and X_3 . A simple model for our cookie weight data is the mean. At the population level the mean is represented by the symbol μ ; at the sample level a different notation is used.

$$\begin{aligned}\mu &= \frac{\sum X}{N} \\ &= \frac{X_1 + X_2 + X_3}{3} \\ &= \frac{8 + 10 + 12}{3} \\ &= \frac{30}{3} \\ &= 10\end{aligned}$$

We can think of the “mean cookie” as our model for our cookie weight data, see Figure 4.2. The “mean cookie” is represented by μ in equations.

Error: As mentioned previously, each participant (i.e., cookie) differs to some extent from our model (“mean cookie”). In general this can be conceptualized as:

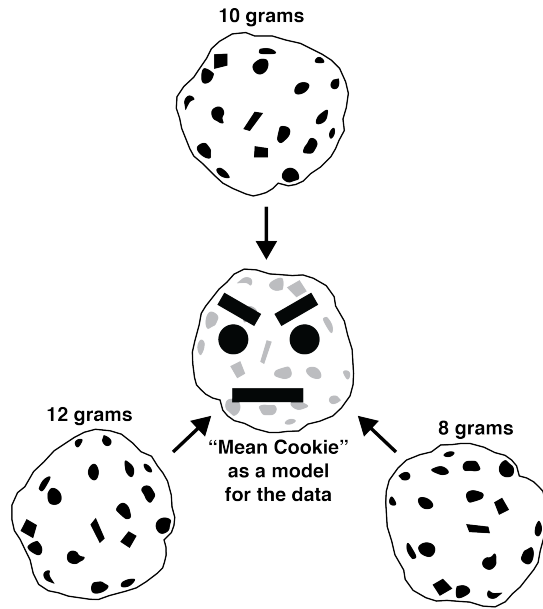


FIGURE 4.2: Variance as a fit index for the mean

$$X_i = model + error_i$$

More specifically, the difference between the weight of any individual cookie (X_i) and the model (μ) is indicated by $error_i$ as shown below.

$$X_i = \mu + error_i$$

The model, above is just a concise way of describing the following:

$$X_1 = \mu + error_1$$

$$X_2 = \mu + error_2$$

$$X_3 = \mu + error_3$$

That is the weights of the three cookies ($X_1 = 8$, $X_2 = 10$, and $X_3 = 12$) can be conceptualized as:

$$X_1 = 10 + (-2)$$

$$X_2 = 10 + 0$$

$$X_3 = 10 + 2$$

The mean/average of the population, $\mu = 10$, is a parameter that serves as a

model for the cookie weight data. However, it's helpful to have an index, known as variance, that indicates the extent to which the data do not correspond to the model from the model.

4.3.2 Variance (σ^2)

Variance is a simple way of calculating a single number to represent how data differ from a model. It is represented, at the population level, by the symbol σ^2 ; a different notation is used at the sample level.

Previously, how we expressed the difference/deviation of cookie weights (data) from the model (i.e., mean) with an error term in the equation $X_i = \mu + \text{error}_i$, see Figure 4.2. The model for all the cookies is $\mu = 10$. If we consider a single cookie weight of 8 grams (a data point represented by X_1), the difference between the cookie from the model is -2 (i.e., error):

$$X_1 = 10 + (-2)$$

We want a number that indicates the quality of the cookie model. Specifically, we want a single number that indexes overall how the data (i.e., cookie weights) differ from the model (i.e., the mean cookie). We refer to that index as variance (σ^2).

Calculating Squared Errors. To calculate variance (σ^2), we use the errors for the cookies – how the cookies differ from the mean/model. The first step is to square the errors/differences. Those squared numbers are referred to as the “squared differences” or “squared errors”. The calculation of the squared error for each cookie weight is shown below. The squared errors (or squared differences) are 4, 0, and 4.

Cookie Weight	Model	Squared Difference
$X_1 = 8$	$\mu = 10$	$(X_1 - \mu)^2 = (8 - 10)^2 = 4$
$X_2 = 10$	$\mu = 10$	$(X_2 - \mu)^2 = (10 - 10)^2 = 0$
$X_3 = 12$	$\mu = 10$	$(X_3 - \mu)^2 = (12 - 10)^2 = 4$

Averaging Squared Errors. To obtain variance we calculate the average of the squared errors. This is shown below:

$$\begin{aligned}
\sigma^2 &= \frac{\sum (X - \mu)^2}{N} \\
&= \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + (X_3 - \mu)^2}{N} \\
&= \frac{(8 - 10)^2 + (10 - 10)^2 + (12 - 10)^2}{3} \\
&= \frac{(-2)^2 + (0)^2 + (2)^2}{3} \\
&= \frac{4 + 0 + 4}{3} \\
&= \frac{8}{3} \\
&= 2.67 \text{ grams}^2
\end{aligned}$$

The resulting variance is 2.67 grams². The cookie weights were measured in grams. The unit for variance, however, is grams² because we squared the errors as part of the calculation. Recall the formula for calculating an average (shown below) and compare it to the variance calculation (above). Notice that variance is just an average – an average of squared errors. Correspondingly, in some areas of statistics they don't use the term variance, they use a synonym - **mean squared error**.

$$\bar{X} = \frac{\sum X}{N}$$

It probably strikes you as an odd choice to square the difference between each data point and the model. Why not just use the difference (e.g., $(8 - 10) = -2$) when calculating variance? Why not use the absolute difference (e.g., $|8 - 10| = 2$) when calculating variance? The answer is somewhat complex, but it relates to the more general situation in statistics of trying to find models that best fit the data (which occurs by minimizing errors). When we use squared errors it is easier to apply calculus, via derivatives, to calculate a model that minimizes the errors (i.e., obtains the best fit). Long story short, for complex mathematical reasons, we use squared errors when calculating the fit (or lack of fit) of a model.

Interpretation. A variance of zero indicates that the model fits the data perfectly. In the cookie case, if the variance was zero, that would indicate that all the cookies had the same weight as the model, exactly 10 grams. To the extent that the variance is larger than zero it implies the data points (i.e., cookie weights) differ from the model (i.e., the mean cookie). By implication, a larger variance indicates larger differences among the observations (e.g., cookie weights). That is, when the variance is small, cookie weights tend to be similar to the model – and each other. In contrast, when the variance is large, cookie weights tend to be different from the model – and each other.

4.3.3 Standard Deviation (σ)

An alternative index for how data differ from the mean/model is the standard deviation. To understand standard deviation you have to understand variance. Variance is a single number that indexes how data differ from a model. The interpretation of variance is straight forward. It is the average of the squared differences between the data and the model.

Standard deviation is represented by the symbol σ and can be calculated as the square root of variance:

$$\begin{aligned}\sigma &= \sqrt{\sigma^2} \\ &= \sqrt{2.67} \\ &= 1.63\text{grams}\end{aligned}$$

One reason that people like standard deviation is presents the difference between the data and the model in the original units (e.g., grams). This is in contrast to variance which presents the difference between the data and the model in squared units (e.g., 2.67 grams²).

Interpretation. Unfortunately, although variance has a straight forward interpretation, standard deviation does not. Sometimes standard deviation is, incorrectly, described as how much data points differ on average from the mean. A quick calculation of the average difference reveals a number (1.33) that does not correspond to the standard deviation (1.63):

$$\begin{aligned}\overline{diff} &= \frac{\sum |X - \mu|}{N} \\ &= \frac{|8 - 10| + |10 - 10| + |12 - 10|}{3} \\ &= \frac{2 + 0 + 2}{3} \\ &= \frac{4}{3} \\ &= 1.33\end{aligned}$$

As illustrated above, standard deviation is not equal to the average of the deviations from the mean. Because standard deviation is not an average, it's much harder to describe how to interpret it. In our view, the best way to think of standard deviation is simply as the square root of variance; because variance has a straight forward interpretation.

Therefore, we encourage you to think primarily in terms of variance rather than standard deviation due to the fact the interpretation of variance is more straightforward. Additionally, variance is foundational in the language used to describe regression and analysis of variance. That said, standard deviation

is used in the calculation of some standardized effect sizes - so it is important to know and understand both indices.

Overall, the rules for interpreting standard deviation are similar to those for variance; but the standard deviation values are smaller than variance values. In the cookie case, if the standard deviation was zero, that would indicate that all the cookies had the same weight as the model, exactly 10 grams. To the extent that the standard deviation is larger than zero it implies the data points (i.e., cookie weights) differ from the model (i.e., the mean cookie). By implication, a larger standard deviation indicates larger differences among the observations (e.g., cookie weights). That is, when the standard deviation is small, cookie weights tend to be similar to the model – and each other. In contrast, when the standard deviation is large, cookie weights tend to be different from the model – and each other.

4.4 Visualizing populations

Three different ways

- histotext
- histogram
- density curve

4.5 Population comparisons

To facilitate comparing two populations we use the heights of males and females, measured in centimeters, as an example. There is variability in the heights of both males and females. However, on average, males are taller than females. That is, the mean of the male population is higher than the mean of the female population.

In this section we present a large number of hypothetical populations representing the heights of males and females. Each of the following figures contains three scenarios (labeled A, B, C) in which we manipulate the mean height and variability of the populations. In each scenario the standard deviation of heights is the same for the male and female populations.

4.5.1 Standardized units

4.5.1.1 Individual scores

Often when we compare two means we use the original metric. In the case of the male and female heights that metric is centimeters. The original units are a useful way to convey information about the difference between two populations.

In addition to the original unit it is also possible, and sometimes desirable, to use the standardized mean difference. The word *standardized* is used to indicate that the comparison is relative to the standard deviation.

Imagine a population of male heights ($\mu = 170$, $\sigma = 10$) from which we have obtained a single individual, Ian, whose height is 185 cm.

$$X_{Ian} = 185cm$$

The original units are useful for describing Ian's height but it doesn't tell us about his height relative to the other people in the population. We have to know the mean and standard deviation of the population to know if Ian is shorter or taller than the average height - and by how much. We use a *z*-score calculation for this purpose:

$$\begin{aligned} z_{Ian} &= \frac{X_{Ian} - \mu_{males}}{\sigma_{males}} \\ &= \frac{185 - 170}{10} \\ &= \frac{15}{10} \\ &= 1.50 \end{aligned}$$

The above calculation is a ratio. Ratios are used to compare two numbers. The numerator (number on the top) is compared to the denominator (number on bottom) through division. The resulting number tells you how much larger the numerator is than the denominator.

In this case, the numerator is the extent to which Ian is taller than the mean height for males ($X_{Ian} - \mu_{males}$). This numerator is compared to the denominator - which is the standard deviation for males (σ_{males}). The resulting number is 1.50 which indicates the numerator is 1.50 times larger than the denominator. In other words, Ian is 1.50 standard deviations taller than the average male. This is a standardized score for Ian's height - it expresses the difference between his height and the mean height in standard deviation units.

4.5.1.2 Population means

The same approach to generating standardized scores can be applied to population means. Consider a situation where we have population of male heights ($\mu = 170$, $\sigma = 10$) and a population of female heights ($\mu = 165$, $\sigma = 10$). Notice that both populations have the same standard deviation.

We can calculate a standardized value to compare these heights. This standardized value is called the standardized mean difference (SMD). Alternatively, it is also known as Cohen's d which is represented at the population level with the symbol δ . Calculation of the standardized mean difference is based on the premise that both populations have the same standard deviation.

In this calculation the numerator represents the difference between the two population means. The denominator represents the population standard deviation - which is the same for both populations. The resulting division of this ratio reveals that the numerator is 0.50 times as large (i.e., half as large) as the denominator. That is, the difference between the populations is half as large as the standard deviation. Therefore, the population mean for males is 0.50 standard deviations larger than the population mean for female; $\delta = 0.50$:

$$\begin{aligned}\delta &= \frac{\mu_{males} - \mu_{females}}{\sigma} \\ &= \frac{170 - 165}{10} \\ &= \frac{5}{10} \\ &= 0.50\end{aligned}$$

4.5.2 Cohen's d units

Because the unit for the standardized mean difference is the standard deviation it can be easy to interpret if you are unfamiliar with the original units. It is usually a good idea to report the difference between populations in both the original units (cm) and standardized units (δ). Figure 4.3 illustrates three different population difference scenarios (A through C). The population standard deviation is held constant across the three scenarios. You can see that as the difference between the population means increases in raw units - it does the same in δ (i.e., Cohen's d) units. In raw units (i.e., cm), the difference between the population means for scenarios A through C are 5 cm, 10 cm, and 20 cm, respectively. In standardized units (i.e., standard deviations), the difference between the population means for scenarios A through C are 0.50 standard deviations, 1.0 standard deviations, and 2.0 standard deviations, respectively. In other words, the population-level Cohen's d -values are 0.50, 1.0, and 2.0 for scenarios A through C.

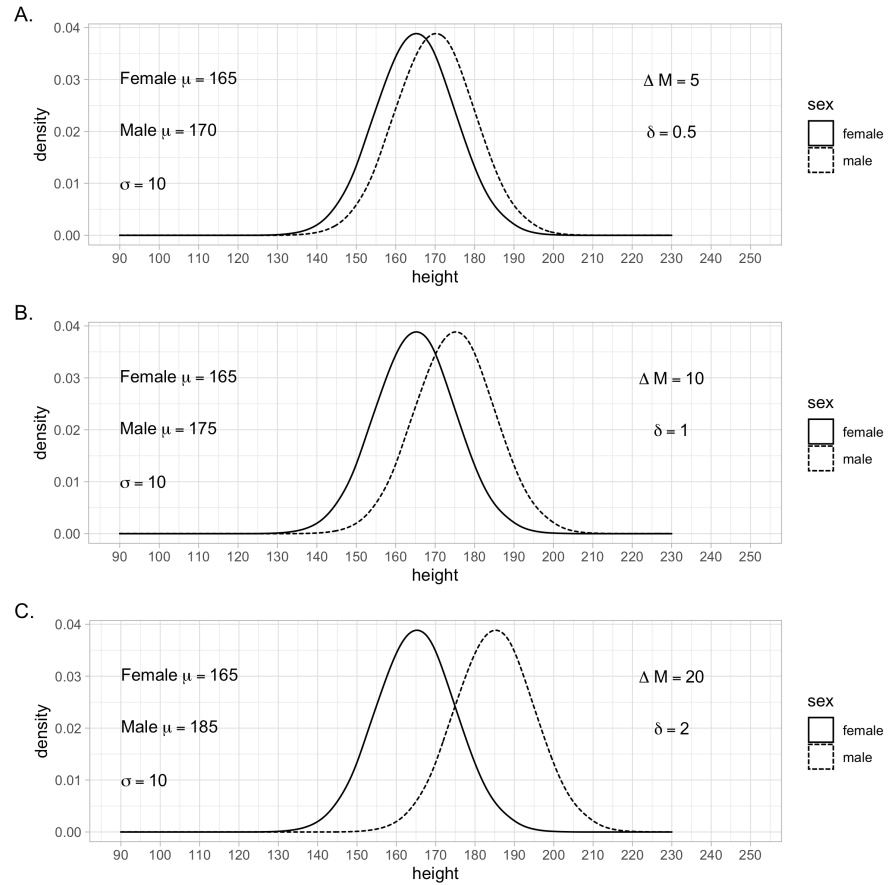


FIGURE 4.3: The difference between two population means can be expressed in the original units as indicated by ΔM . Alternatively, the difference can be expressed using a Standardized Mean Difference (SMD). The SMD index is also known as the population-level d -value and is represented by the symbol δ . The SMD is a way of expressing the difference between population means without using the original units.

4.5.3 Cohen's d advantages

The standardized mean difference takes into account the variability of heights around each population mean. This means that the same difference between two population means can produce different standardized mean difference values if the population standard deviation varies. In the scenarios depicted in Figure 4.4 the population standard deviation becomes increasingly small - resulting in larger standardized mean difference values (i.e., δ). This larger δ value corresponds to progressively less overlap between the two populations. Thus, taking into account the standard deviation of the populations can be viewed as a strength of using the standardized mean difference.

4.5.4 Cohen's d caveats

It is important to also look at the original units when interpreting results - not just the standardized mean difference. Examine the scenarios in Figure 4.5. Notice how the δ value stays constant across scenarios - as does the overlap of the two distributions. However, inspect the shape of the curves and the original units to see how the scenarios vary. Both the original units and the standardized mean difference (i.e., Cohen's d) provide important interpretational information - don't rely on just one of them.

4.5.5 Benchmarks

Cohen suggested that the standardized mean difference values of 0.20, 0.50, and 0.80 correspond to the effect size labels of small, medium, and large, respectively (Cohen, 1988). These effect sizes are illustrated in Figure 4.6. As described in an interesting blog post¹, these benchmark values came from reviewing a single issue of the *Journal of Abnormal and Social Psychology*. Basing benchmarks on such a small number of studies is potentially problematic - as is the fact that at that time all the studies were prone to publication bias. A recent investigation (see Schäfer and Schwarz, 2019) of effect sizes in pre-registered studies, with no publication bias, suggests substantially lower benchmark values.

You can visualize any δ value (i.e., population d -value) using the rpsychologist website². This website also provides a number of interesting statistics such as the percentage of overlapping values in two populations for a given δ/d value.

¹<https://replicationindex.com/2015/09/22/the-statistical-power-of-abnormal-social-psychological-research-a-review-by-jacob-cohen/>

²<https://rpsychologist.com/d3/cohend/>

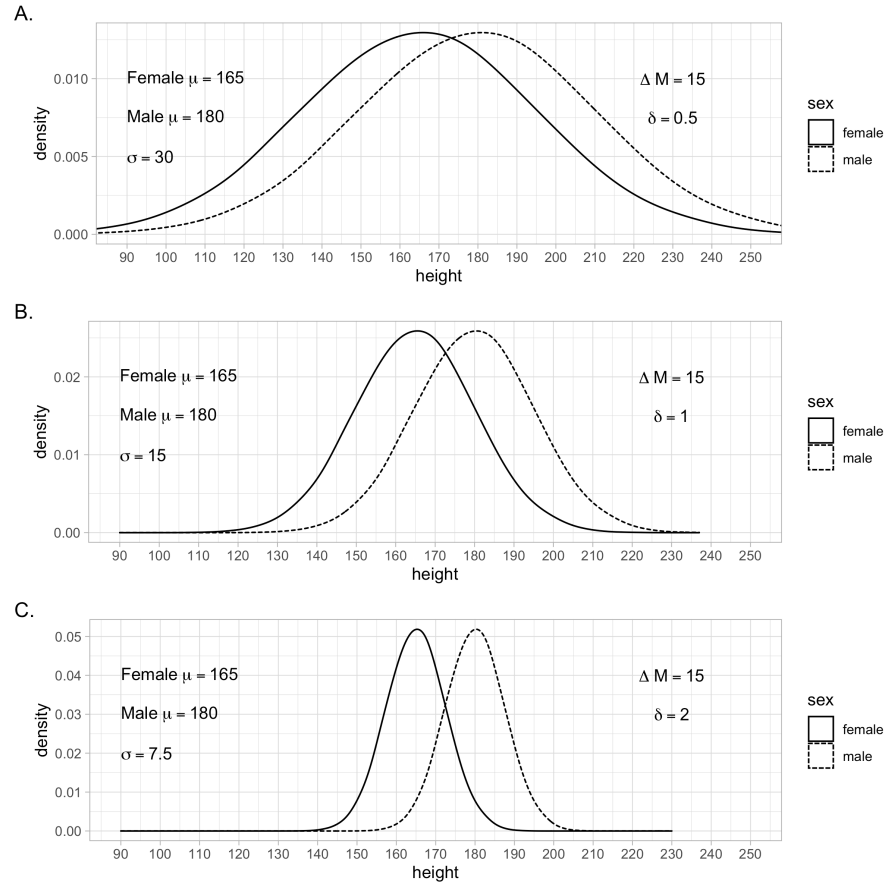


FIGURE 4.4: An advantage of using the Standardized Mean Difference (SMD) to index the difference between two population means (i.e., δ) is that it takes the population standard deviation into account. In these three examples, the difference between the populations means is the same using the original/raw units of centimeters. However, the standard deviation of the populations varies across scenarios A, B, and C. The SMD illustrates that these three scenarios are different. If you only examined the difference in the original units (i.e., ΔM) you would conclude the effect is the same across the three scenarios. However, by using SMD, indexed by δ - the population d -value, you see that the effect is progressively stronger from scenario A, to B, to C. This is illustrated by the fact that there is progressively less overlap between the distributions as you move from scenario A to C.

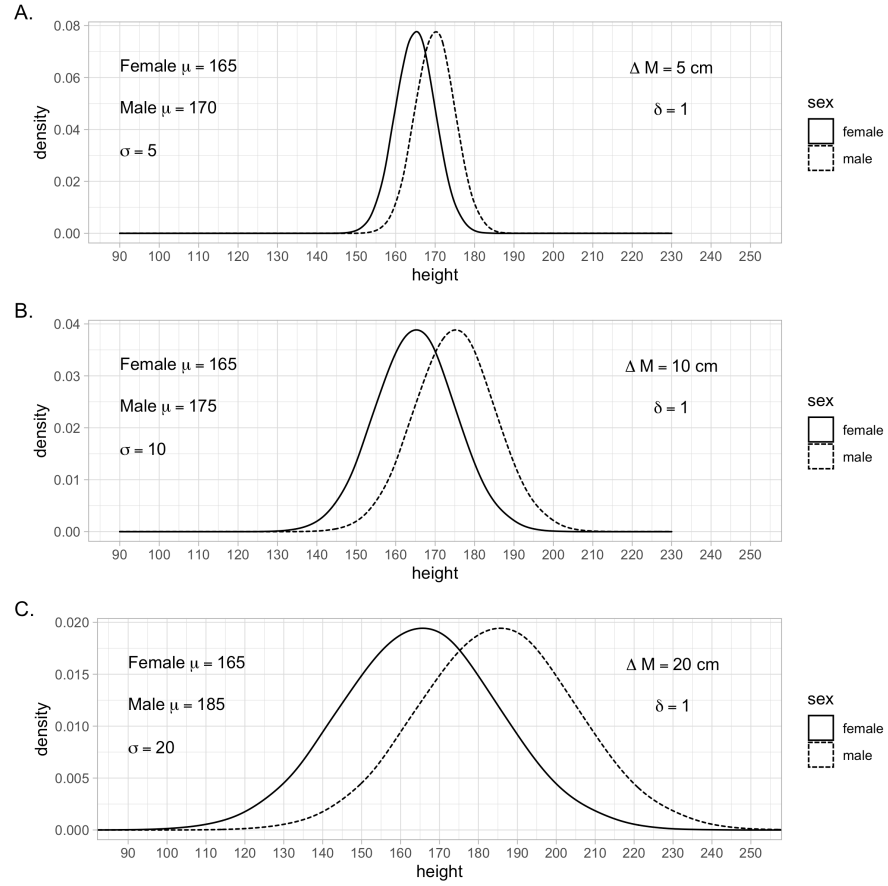


FIGURE 4.5: The three scenarios in this figure illustrate that a Standardized Mean Difference (i.e., population d -value or δ) can remain constant across scenarios when there is a change in the raw difference (i.e., ΔM) between the population means. This SMD is consistent across the three scenario despite a change in the mean difference using original units; this occurs because the standard deviations also changes across the three scenarios.

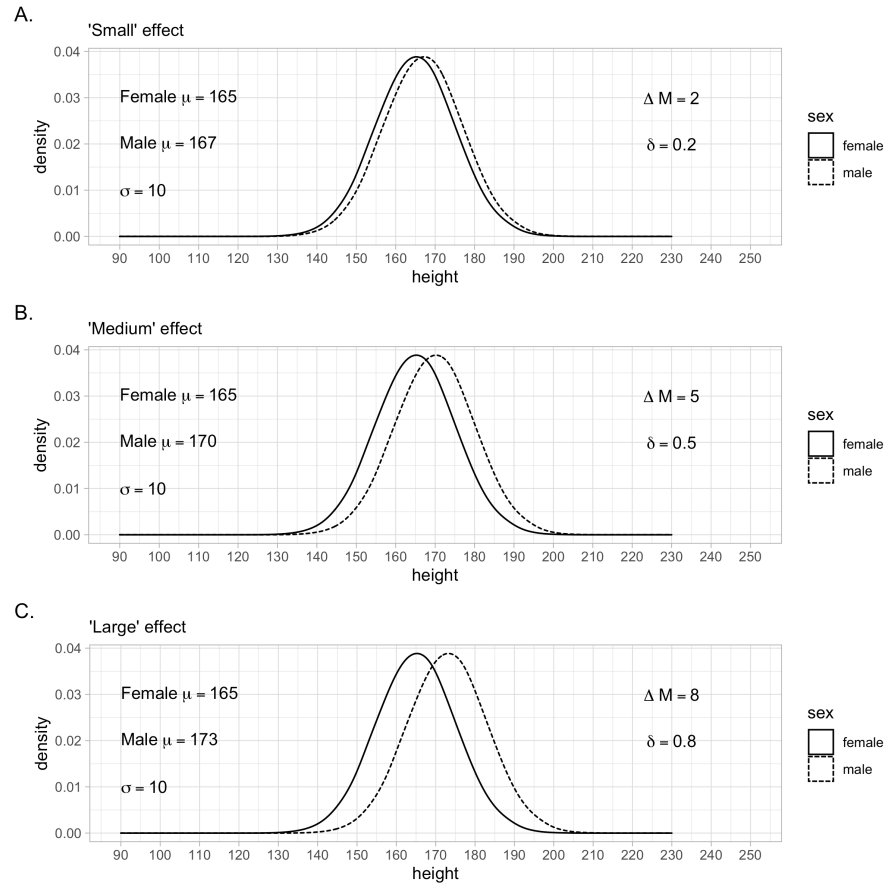


FIGURE 4.6: Cohen's (1988) effect size benchmarks.

4.6 Key points

1. Populations are described using numbers called parameters.
2. Population-level parameters are often represented using Greek letter.
3. Commonly used parameters include mean (μ or \bar{X}), variance (σ^2), or standard deviation (σ).
4. Individual scores can be expressed in standardized units.

5. Population differences can be described in original raw units or standardized units called the standardized mean difference (SMD).
6. SMD is based on the premise that the two population being compared have the same standard deviation.
7. SMD is a ratio that compares two numbers (the numerator and the denominator).
8. Make sure you understand what is represented by both the numerator and denominator in the SMD ratio.
9. SMD (i.e., Cohen's d) represents the number of standard deviations between two population means. Recall both populations have the same standard deviation.
10. SMD is indicated at the population level using the Greek letter delta (δ). At the sample level we tend to use the term " d -value" or "Cohen's d ".



5

Samples

Researchers are usually interested in describing the attributes of a population; numbers that describe the population are called parameters. Two parameters that are frequently of interest are the mean and variance of the population. Unfortunately, it's rarely possible to obtain information from every member of a population to calculate a parameter. Consequently, researchers use subsets of the population called samples to estimate parameters. Numbers calculated from sample data are called statistics. Typically, sample statistics are used to estimate population parameters.

Sample statistics, however, often differ from population parameters. The difference between a sample statistic and the population parameter occurs because the sample data is random subset of the population data — with correspondingly fewer observations. Sometimes the sample statistic will be higher than the population parameter; other times the sample statistic will be lower than the population parameter. Because random sampling is used to select the sample data the direction and magnitude of the difference between the sample statistic the population parameter will vary randomly.

Further complicating matters is the fact that the formula used for a sample statistic may or may not be the same as the formula used for the corresponding population parameter. This occurs because the purpose of the sample statistic is typically not to describe the sample. Rather the purpose of the sample statistic is to estimate the population parameter. Depending on the parameter, you may or may not be able to use the same formula with sample data as you would with population data.

5.1 Notation

In the formulas below, when we refer the population, we use uppercase letters when indicating members (X) or the size (N). In contrast, when we refer to the sample, we use lowercase letters when indicating members (x) or the size (n). Make sure you notice the similarities between population and sample formulas even when the notation differs.

5.2 Estimating the mean

We are interested in sample means (\bar{x}) to the extent that they provide an estimate of the population mean (μ).

The population mean, μ , is a parameter calculated using Formula (5.1)

$$\mu = \frac{\sum X}{N} \quad (5.1)$$

The sample mean, \bar{x} , is a statistic calculated using the using Formula (5.2). The bar of above the x , indicates that it is a mean. Notice that the two formulas are the same - even though they use different notation.

$$\bar{x} = \frac{\sum x}{n} \quad (5.2)$$

A sample mean is calculated using sample data which is a random subset of the population. Consequently, the sample mean (a statistic) is likely to differ from the population mean (a parameter). If you, inaccurately, believe that you can learn something meaningful from a single study, the fact that sample means differ from the population mean could potentially be concerning. Statisticians know, however, that you can rarely learn anything from a single study or even a small set of studies. Consequently, they are more interested in what is true, on average, over a large number of studies.

One way of assessing the quality of the sample mean as an estimate of the population mean is by examining bias. If the average of many sample means is equal to the population mean, then we consider the sample mean to be an unbiased estimator of the population mean.

Fortunately, the average of many sample means does equal the population mean - so the sample mean is an unbiased estimator of the population mean. This fact can be illustrated with a computer simulation.

We begin by activating the required packages:

```
library(tidyverse)
library(learnSampling)
```

Next, we create a large population with 100,000 people using the `get_height_population()` command:

```
pop_data <- get_height_population()
```

The `glimpse()` command can be used to confirm that the population contains 100,000 people.

```
glimpse(pop_data)
```

```
## Rows: 100,000
## Columns: 3
## $ id      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1...
## $ sex     <chr> "male", "female", "female", "male", "mal...
## $ height  <dbl> 177, 150, 171, 157, 169, 187, 163, 173, ...
```

We can use the `head()` command to illustrate that each row in `pop_data` represents a single person.

```
head(pop_data, 10)
```

```
## # A tibble: 10 x 3
##       id sex    height
##   <int> <chr>   <dbl>
## 1     1 male     177
## 2     2 female   150
## 3     3 female   171
## 4     4 male     157
## 5     5 male     169
## 6     6 male     187
## 7     7 female   163
## 8     8 male     173
## 9     9 female   172
## 10    10 male     193
```

We can use tidyverse commands to determine the mean of the population (μ):

```
pop_data %>%
  summarise(pop_mean = mean(height)) %>%
  as.data.frame()
```

```
##   pop_mean
## 1    172.5
```

We see that the mean of the population is $\mu = 172.48$. We can illustrate how random sampling works, with respect to estimating this population mean, by taking a large number of samples from the population.

In the code below, we use the `get_M_samples()` command to obtain 5000 samples, each composed of 10 people. For each of these 5000 samples we calculate a variety of statistics and place them in the `many_samples` data frame. Each row of `many_samples` represents a sample of 10 people. Each column of `many_samples` indicates a sample statistic or the population parameter the statistic is estimating.

```
# set.seed ensures you get the same random samples
set.seed(1)

many_samples <- get_M_samples(pop.data = pop_data,
                              data.column.name = height,
                              n = 10,
                              number.of.samples = 5000)
```

We use the `head()` command to see the first 10 rows (i.e., 10 samples of 5000 samples):

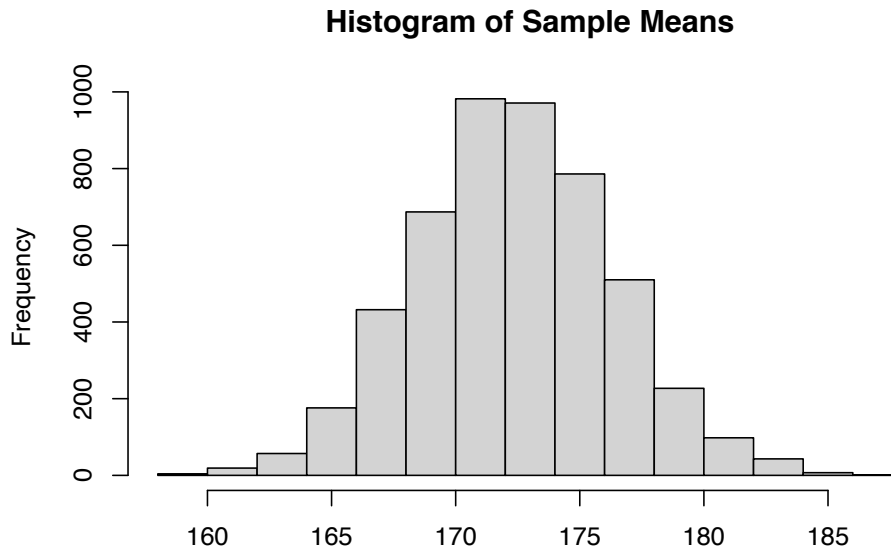
```
head(many_samples, 10)
```

##	n	pop.M	sample.M	pop.var	sample.var.n	sample.var.n_1
## 1	10	172.5	174.3	157.5	27.81	30.90
## 2	10	172.5	161.8	157.5	83.36	92.62
## 3	10	172.5	173.2	157.5	139.16	154.62
## 4	10	172.5	180.0	157.5	321.80	357.56
## 5	10	172.5	172.8	157.5	160.96	178.84
## 6	10	172.5	172.0	157.5	159.60	177.33
## 7	10	172.5	166.5	157.5	86.45	96.06
## 8	10	172.5	173.5	157.5	139.45	154.94
## 9	10	172.5	177.7	157.5	167.01	185.57
## 10	10	172.5	168.5	157.5	262.65	291.83

You can see that for each sample we indicate `n` (the sample size), `pop.M` (the mean of the population), `sample.M` (the mean of the sample), and a few other statistics. Because all the samples came from the same population you can see that the population mean (`pop.M`) is constant across rows/samples. In contrast, the sample mean (`sample.M`) varies across samples due to random sampling error.

You can see the full extent to which the sample means vary by creating a graph with the code below. In this code, we use the `pull()` command to extract the value from the `sample.M` column and then we send them to the base R histogram command, `hist()`.


```
many_samples %>%  
  pull(sample.M) %>%  
  hist(main = "Histogram of Sample Means")
```



You can see the 5000 sample means plotted in this graph. Recall the population mean for heights is $\mu = 172.48$ cm. Notice that most of the sample means cluster around this value. Also notice there is considerable variability about this value. Any given sample mean (\bar{x}) may differ substantially from the population mean ($\mu = 172.48$). This variability illustrates the challenges with learning something from a single study - particularly a study with a small sample size.

Statisticians, recognizing the limitations of a single study, are not particularly concerned if a single sample deviates from the population mean. That said, statisticians are very concerned whether the results of a large number of studies are correct – on average. That is, does the average of many sample means correspond to the population mean. If, on average, the sample mean is accurate we refer to it as an unbiased estimator.

In the code below we calculate the average of the 5000 sample means to determine if the sample mean is an unbiased estimator.

```
many_samples %>%  
  summarise(mean_of_sample.M = mean(sample.M))
```

```
## mean_of_sample.M
## 1 172.4
```

We find that the average of the 5000 sample means is 172.47 which is very close to the population mean of 172.48. Note that when we did this, we used the same formula to calculate the sample mean as we did the population mean. The average of the sample means was not identical to the population mean but it was very close - it would have been exactly the same with many more samples. Consequently, the sample mean provides an unbiased estimate of the population mean. In other words, it makes sense to use the sample mean as an estimate of the population mean. If we try to estimate the population mean with a sample mean we will, on average, be correct; although any given sample/study mean might be “wrong”.

5.3 Estimating variance

We are interested in sample variances (s^2) to the extent that they provide an estimate of the population mean (σ^2). Recall, the population variance, σ^2 , is a parameter calculated using Formula (5.3).

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \quad (5.3)$$

The sample variance, s^2 , is a statistic calculated using Formula (5.4). The formula is the same as at the population level but we use a different notation because it is a sample.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n} \quad (5.4)$$

Examine the output for 5000 samples we created using the `head()` command. Notice in this output we see the population variance is 157.46 (see the `pop.var` column). The values in this column were calculated, based on everyone in the population, using Formula (5.3).

```
head(many_samples, 10)
```

```
##      n pop.M sample.M pop.var sample.var.n sample.var.n_1
## 1  10 172.5   174.3   157.5     27.81      30.90
## 2  10 172.5   161.8   157.5     83.36      92.62
## 3  10 172.5   173.2   157.5    139.16     154.62
## 4  10 172.5   180.0   157.5    321.80     357.56
```

```
## 5  10 172.5    172.8    157.5        160.96        178.84
## 6  10 172.5    172.0    157.5        159.60        177.33
## 7  10 172.5    166.5    157.5         86.45         96.06
## 8  10 172.5    173.5    157.5        139.45        154.94
## 9  10 172.5    177.7    157.5        167.01        185.57
## 10 10 172.5    168.5    157.5        262.65        291.83
```

Now inspect the `sample.var.n` column the values in this column were calculated based on sample data using Formula (5.4). In this column we present the sample variance calculated, using the above formula, with n in the denominator. That is, in this column we have calculated sample variance using the same formula (but with different notation) as we used at the population level. Notice the values in the `sample.var.n` column vary due to random sampling.

As we did with the mean, we can take the average of the 5000 sample variances (each using n in the denominator) to determine if the sample variance provides a biased or unbiased estimate of the population variance.

```
many_samples %>%
  summarise(mean_of_var.n = mean(sample.var.n))
```

```
##   mean_of_var.n
## 1          141.1
```

We find that the average of the 5000 sample variances is 141.86 which is not close to the population variance of 157.46. Note that when we did this, we used the same formula to calculate the sample variance as we did the population variance.

The average of the sample variances (141.86) was different from the population variance (157.46). The average of the sample variance was much smaller than the population variance. Consequently, the sample variance provides a biased estimate of the population variance. If we try to estimate the population variance with a sample variance we will, on average, be wrong.

Fortunately, there is a way around the fact that the sample variance provides a biased estimate of the population variance. We can derive a formula at the sample level that estimate the population variance that is not biased (see Hayes). An unbiased estimate of the population variance can be obtained if we calculate the sample variance but divide by $n-1$ instead of n . The unbiased estimate is calculated using Formula (5.5).

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \quad (5.5)$$

The values in the `sample.var.n_1` column were calculated, based on sample data, using Formula (5.5). Check out this column using the `head()` command:

```
head(many_samples, 10)
```

```
##      n pop.M sample.M pop.var sample.var.n sample.var.n_1
## 1  10 172.5   174.3   157.5      27.81      30.90
## 2  10 172.5   161.8   157.5      83.36      92.62
## 3  10 172.5   173.2   157.5     139.16     154.62
## 4  10 172.5   180.0   157.5     321.80     357.56
## 5  10 172.5   172.8   157.5     160.96     178.84
## 6  10 172.5   172.0   157.5     159.60     177.33
## 7  10 172.5   166.5   157.5      86.45      96.06
## 8  10 172.5   173.5   157.5     139.45     154.94
## 9  10 172.5   177.7   157.5     167.01     185.57
## 10 10 172.5   168.5   157.5     262.65     291.83
```

We can evaluate the quality of Formula (5.5), using $n - 1$, by averaging over values in the `sample.var.n_1` column.

```
many_samples %>%
  summarise(mean_of_var.n_1 = mean(sample.var.n_1))
```

```
##   mean_of_var.n_1
## 1             156.7
```

We see that the average of the 5000 values using $n - 1$ in the denominator is 157.62 which is very close to the population variance of 157.46. Consequently, when we use $n - 1$ in the denominator we have an unbiased estimate of the population variance. If we try to estimate the population variance with a sample variance, using $n - 1$ in the denominator, we will, on average, be right.

You may wonder at this point, when we use $n - 1$ in the denominator of the sample variance, can we still think of it as the average of the squared deviations from the mean. The short answer is yes. When you use $n - 1$ in the denominator of the sample variance you are not calculating the variance for the group of people in the sample. Rather, you are estimating the variance for the much larger group of people in the population. Consequently, it makes sense to think of sample variance, using $n - 1$, as an estimate of the average of the squared errors in the population. That is, it makes sense to think of sample variance, using $n - 1$, as an estimate of the average of the squared differences between each person in the population and the population mean.

5.4 Estimating standard deviation

Due to the above findings for variance, we tend to estimate the population standard deviation using Formula (5.6).

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad (5.6)$$

5.5 Estimating SMD

```
female_pop_data <- pop_data %>% filter(sex == "female")
male_pop_data <- pop_data %>% filter(sex == "male")
```

For females:

$$\begin{aligned}\mu_{female} &= 165 \\ \sigma_{female} &= 10\end{aligned}$$

For males:

$$\begin{aligned}\mu_{male} &= 180 \\ \sigma_{male} &= 10\end{aligned}$$

5.5.1 Population SMD

In original units:

$$\begin{aligned}\Delta M &= (\mu_{male} - \mu_{female}) \\ &= (180 - 165) \\ &= 15\end{aligned}$$

Formula (5.7) below illustrates....

$$\delta = \frac{(\mu_{male} - \mu_{female})}{\sigma} \quad (5.7)$$

Calculation:

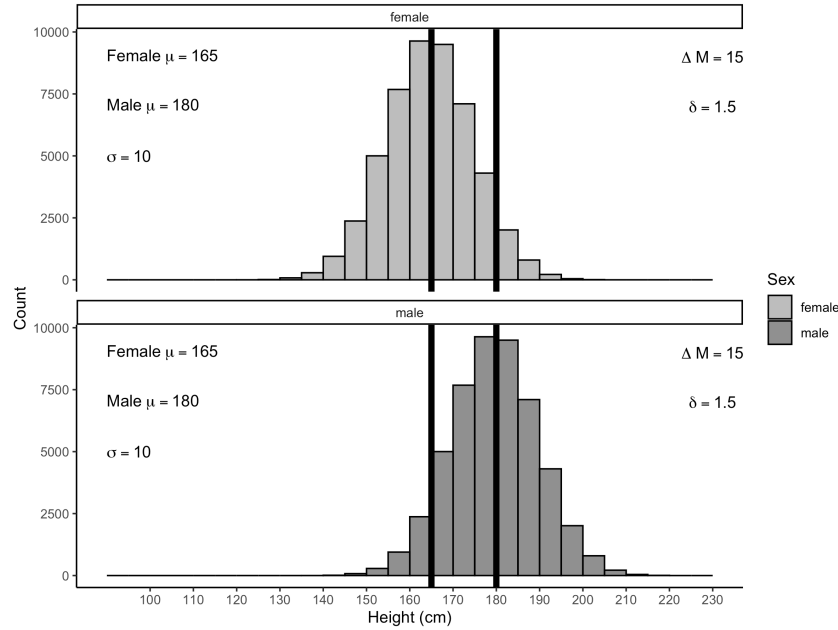


FIGURE 5.1: Illustration of the standardized mean difference of 1.50 for height between males and females. The solid black vertical lines indicates the means (μ) for the two populations.

$$\begin{aligned}
 \delta &= \frac{(\mu_{male} - \mu_{female})}{\sigma} \\
 &= \frac{(180 - 165)}{10} \\
 &= \frac{15}{10} \\
 &= 1.50
 \end{aligned}$$

5.5.2 Sample SMD

Formula (5.8) below illustrates...

$$d = \frac{(\bar{x}_{male} - \bar{x}_{female})}{s_{pooled}} \quad (5.8)$$

5.5.2.1 Estimating means

5.5.2.2 Estimating variance

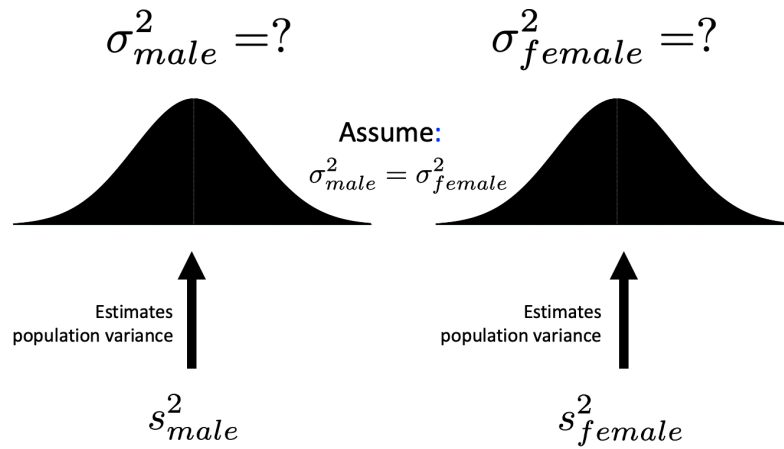


FIGURE 5.2: The male sample variance (n-1) is an estimate of male population variance. Likewise, the female sample variance (n-1) is an estimate of the female population variance.

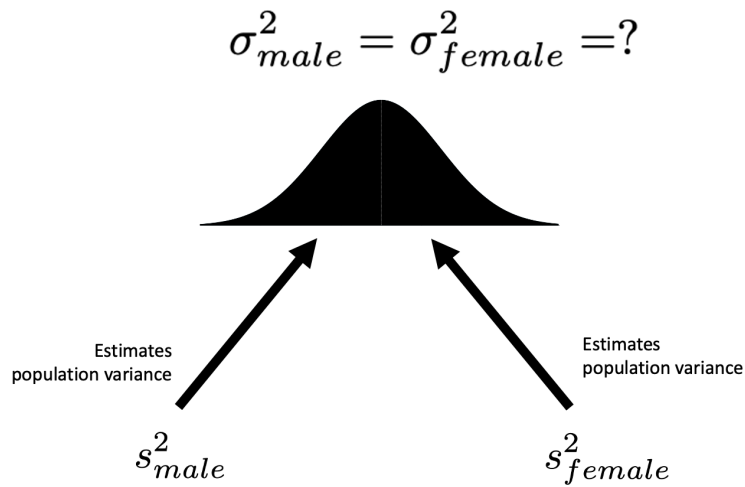


FIGURE 5.3: We assume the population variances are the same. Therefore, the male and female sample variances are both estimates of the same population variance.

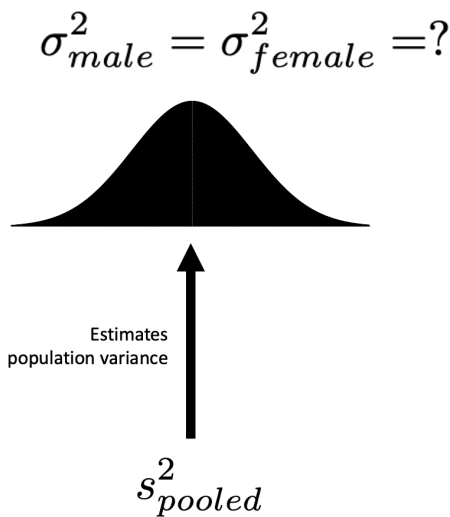


FIGURE 5.4: We create a single estimate of the population variance called pooled variance (s^2_{pooled}). When sample sizes are equal, the pooled variance is just the average of the two sample variances (both using $n-1$). When the sample sizes are unequal, we need to use a more sophisticated formula to obtain the pooled variance.

5.5.2.3 Combining everything

$$d = \frac{(\bar{x}_{male} - \bar{x}_{female})}{s_{pooled}}$$

5.6 Overview

Parameter	Estimated by this statistic
$\mu = \frac{\sum X}{N}$	$\bar{x} = \frac{\sum x}{n}$
$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$	$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$
$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$
$\delta = \frac{\mu_1 - \mu_2}{\sigma}$	$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$

A

More to Say

Yeah! I have finished my book, but I have more to say about some topics. Let me explain them in this appendix.

To know more about **bookdown**, see <https://bookdown.org>.



Bibliography

- Baker, M. (2016). 1500 scientists lift the lid on reproducibility. *Nature*, 533.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic press.
- Keur, C. and Hillegass, A. (2020). *iOS programming: The Big Nerd Ranch guide*. Pearson Technology Group.
- Meyer, J. P., Allen, N. J., and Smith, C. A. (1993). Commitment to organizations and occupations: Extension and test of a three-component conceptualization. *Journal of applied psychology*, 78(4):538.
- Miyakawa, T. (2020). No raw data, no science: another possible source of the reproducibility crisis. *Mol Brain*, 13(24).
- Nosek (2015). Estimating the reproducibility of psychological science. *Science*, 349.
- Patil P., Peng R.D., . L. J. (2019). A visual tool for defining reproducibility and replicability. *Nat Hum Behav*, 3:650–652.
- Schäfer, T. and Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10:813.
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.
- Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59.
- Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.
- Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.19.1.



Index

bookdown, [ix](#)

date format, [7](#)

knitr, [ix](#)

R Studio Cloud, [2](#), [7](#)

tidyverse, [7](#)