**Getting Started With R When You Know Absolutely Nothing**

Version Date: September 1, 2020

Moin Syed, University of Minnesota

Associated Materials Available at: <https://osf.io/9gq4a/>

Thanks to Linh Nguyen for helpful comments and checking that the sample code worked. All questions and complaints should be directed to: moin@umn.edu

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

I mean it, this tutorial is really for if you don’t know anything about R, and even if you don’t have any experience coding/scripting at all. Of course, many of you have likely dabbled with learning R previously and have found many, many tutorials online. Given this, do we really need another one? Good news: this is not actually a new tutorial at all, but a *collection* of tutorials bundled together. Totally different.

R is a free, open source computing language that is used for statistical analysis, data manipulation, generating plots, and preparing reports, along with many other useful functions. Many students (and faculty…) in psychology have been exposed to SPSS for data analysis, where the user relies on a series of window-based options to “point-and-click” to select and run their analyses. However, in addition to the point-and-click method of using SPSS, users can alternatively rely on *scripting*, where they write out commands based on the rules of the program. When using R, scripting is your only option ([that is a lie](https://drive.google.com/file/d/1MaRADweSx67ecUiAQdXgNvp6bIkkZQ85/view), but for now just pretend it is true). This can be a little intimidating for students (and professors!) who do not have any experience with scripting. Nevertheless, we all have to confront intimidating activities at some point, and scripting/computing has increasingly become a standard component of psychological science. So how do you learn?

First off, it is important to be clear that I know very little about R. I do not even know enough to modestly claim “I’m no expert.” I really don’t know much of what I am doing. But I know a little, and I work to figure out what I do not know. I think my position is useful in this context because I have pretty good insight into how you might go about learning so that you can know just as little as I do. Hopefully, over time, you will know more.

From my perspective, you need to learn some basics about how the program works, some foundational rules, and what it is capable of. From there, you are best off working on actual data-based problems and figuring out what you need to as you go. Over time, some of those lessons will sink in, you will consolidate your knowledge, and then you will take it further. Some people call this learning! Indeed, you cannot just suddenly program in R. You have to learn how, and learning takes time. A lot of time. And Google.

That brings me to the first two laws of learning R:

*First Law of Learning R: Most people do not really know what they are doing*

*Second Law of Learning R: Using R involves spending more time on Google looking for solutions than actually coding.*

For all lessons you will use RStudio and for most of them you will use tidyverse. You have no idea what those mean, fine, but you soon will, and I wanted to mention it here for those who have heard of them. As mentioned R is freely available, so does not require any purchases on your part. The first lesson below will walk you through how to download and install R and RStudio.

You may be wondering how long all of this will take. That, like most things, depends. You could certainly work through all of this in about a week of concentrated effort, but most people would likely be most comfortable with spreading the work out over several weeks with regular effort or a full academic term for more sporadic effort. If you are some random person trying this out, then you do you. If you are a PI, group leader, or instructor, then do whatever fits with your training goals. This tutorial would certainly be more than sufficient to structure learning for an undergraduate research methods course (or similar) so that students could analyze their project data using R.

**Learning Phase 1: A Suuuper Gentle Introduction to R with the “R advent calendaR”**

Conservation Biologist [Kiirsti Owen](https://kiirstio.wixsite.com/kowen) created the “R advent calendaR” as an easy and fun way to learn the basics of R. Before I say more, I want to introduce the third law of learning R:

*Third Law of Learning R: when referring to anything related to R, and there is an “r” or r sound in a word, do something weird with it.* For example, Calendar is wrong, calendaR is correct. Adventure is wrong, adventr is correct.

The advent calendaR is divided into 25 lessons to complete in 25 days, but each lesson is quite short, especially those at the beginning, so you can and should do several lessons each day. You can find more information about and download the calendar here:

<https://kiirstio.wixsite.com/kowen/post/the-25-days-of-christmas-an-r-advent-calendar>

Be sure to scroll to the bottom of the page, where there are several important tips and notes about some errors in the lessons. I actually think these errors are quite useful for learning.

As you will see, this tutorial assumes you know *absolutely nothing* *at all* about R or coding. Like seriously nothing; you will see. You should be able to complete all of the lessons in a couple hours of work. Once you do, you will have learned the basic landscape of R/RStudio, how to read in data, manipulate data, create simple plots, and do basic analyses. You won’t learn much, but you will know more than you did before.

**Learning Phase 2: A Still Gentle but More In-depth Exploration of R**

The advent calendaR is great as an initial foray into R, but it is clearly very limited. Because we are *learning*, the next step is to use a training program that builds upon the previous while growing your skills. Professor [Andy Field](https://www.discoveringstatistics.com/about/), known for his many helpful statistics texts and resources, created “adventr: An Adventure in Statistics interactive tutorials,” and it perfectly fits the bill.

The adventr package is designed to accompany Field’s book, [An Adventure in Statistics](https://www.discoveringstatistics.com/books/an-adventure-in-statistics/), but the tutorial can be used without the book, assuming you have some background in statistics. Because it is designed with the book in mind, the content and progression looks similar to what you might find in an undergraduate statistics course: summarizing data, creating plots, comparing means, etc. There is some overlap with advent calendaR at the beginning, but repetition is a good thing during the learning process!

You can access the package here:

<http://milton-the-cat.rocks/home/adventr.html>

Note that you have to install the package in R and run it from there, but then all of the work is actually done in a web interface. You will see what I mean once you begin. Just be sure that you keep R open and that the package is loaded with library(adventr), otherwise the tutorial will not work properly.

There are several smallish tutorials within adventr. I recommend that you work your way through them all, doing 1-2 per day (probably about 30 minutes of work). If you do, you will have been exposed to a solid set of basic functions in R and will thus have an excellent springboard for further learning and development. After all, Google is truly effective only when you *actually know* what you are searching for. The adventr tutorials will get you there.

**Optional Detour: Learning Phase 2a – Keep Learning via Tutorials**

If after completing the adventr tutorials you want to continue learning in a tutorial environment, then I recommend working through the swirl lessons, which is an extensive set of lessons that move from beginning to advanced. Each lesson takes around 15 minutes and is completed entirely in R. It follows a similar rhythm and progression as the adventr tutorial, but runs through R rather than the web. Most of the early lessons will be repeats of adventr, so you can skip those if you really feel like you have the basics down.That said, the approach is slightly different and I always find that I pick up something new when going over the basics. Remember the first law, *most people don’t actually know what they are doing*, which means you will always learn something. You can access swirl here:

<https://swirlstats.com/students.html>

If you feel like you have learned the basics and want to start working with some real data that are similar to what you will actually work with in psychological research, then forget swirl and move on to Phase 3.

**Learning Phase 3 – Application with Real Data**

Ok, enough with the tutorials, it is *time for action*. One of the downsides of most R tutorials is that the data and examples do not closely resemble data you will typically work with when doing psychological research. Moreover, application to a new and realistic setting is a great way to push your learning further. In this phase you will continue to use RStudio and work with some real psychological data.

I have curated a simple data set, named EID\_Data\_MCAE2016.csv, and posted it on the companion OSF page (<https://osf.io/9gq4a/>). These are real data that come from a much larger data set (for details on procedure and so on, see [Fish et al., 2020](https://psyarxiv.com/k7na5/)). A total of 316 participants were recruited as part of a multicultural orientation event at a “large public university in the U.S. Midwest.” Thus, the sample consists of mostly 18-year-olds who identify as ethnic/racial minorities.

The dataset includes the following variables (20 in all): 1) a unique ID for each participant, 2) whether the participant was born in the U.S. (yes/no), 3) whether the participant was the first in their family to attend college (yes/no), 4) the 12-item Multigroup Ethnic Identity Measure (MEIM; Roberts et al., 1999), and 5) the 5-item Satisfaction with Life Scale (SWL; Diener et al., 1999).

For the MEIM and SWL only the raw items are included, not scale scores. Scoring instructions are included on the OSF page. There you will also find a codebook that describes all variables and possible values, as well as a R script file that includes a key for the exercises (but don’t look at that now!). So you now have a dataset and a description of that dataset. You are ready to go. Follow the instructions below to do some basic data wrangling, analyses, and plots.

1. Begin by installing (as needed) and loading the following packages: *summarytools*, *dplyr*, *psych*, *ggplot2*. You will also see references below to “base R.” This means that the function is built into R and does not require a specific package to run. That reminds me of another law:

*Fourth Law of Learning R: There is no “right” way to do anything in R. There are always at least two—and sometimes many more—different packages that can perform the same function.*

Indeed, you could complete this lesson using an entirely different set of packages. Good to know for now, and play around as you want, but if you are just getting started I suggest using the ones I listed above (besides, I only made a key for using those packages, so you can’t cheat if you use different ones).

1. Read in the raw .csv data file. Do some checks to make sure that everything looks as it should and that the data were read in properly.
2. Get to know the basics of your data by using *summarytools* to examine the frequencies and descriptive statistics for all items.
3. Use *dplyr* to calculate scale scores for the MEIM and SWL according to the scoring instructions and store them as new variables in a new dataset. Use *summarytools* to check the descriptive statistics of the new variables.
4. Use the *psych* package to create a correlation matrix consisting of the three scale scores. Then run a command to get only the correlation between MEIM exploration and SWL.
5. Run a multiple regression analysis in base R with the two MEIM subscales predicting SWL. Request confidence intervals for the estimates.
6. Create two scatterplots using *ggplot2*, one for MEIM exploration and SWL and one for MEIM commitment and SWL. And labels to the axes, change the scales, and add a fit line.
7. Run two sets of independent samples t-tests in base R on the three scales scores, one set for Born in the U.S and one set for First Generation College Student (so six t-tests total).
8. Create boxplots using *ggplot2* showing the three scale score distributions for Born in the U.S and First Generation College Student (so again, six plots).
9. Run a 2-way ANOVA in base R for Born in the U.S x First Generation College Student predicting the three scales scores (so three ANOVAs).
10. Create boxplots using *ggplot2* showing the three scale score distributions for Born in the U.S x First Generation College Student (so this time only three plots!).
11. Write a summary report describing everything you did in steps 1-11, embedding figures and using APA style to report the results of your tests. Feel free to make tables as well.

***Congratulations, now you know how to use R (sort of)***