MIE237 January 19-20 Labs

Neil Montgomery

January 17, 2016

What you will do in this lab

In this lab you will use some textbook datasets to make small "reports" in R markdown. Read the question to understand the background of the data, but otherwise don't bother with what the book asks. Instead do what I ask here.

1. 9.40 - produce a 95% confidence interval for the difference between the groups, making no equal variance assumption. *I did this one for you as an example.*
2. 9.46 - perform the hypothesis test with the null hypothesis that there is no difference between the groups, making no equal variance assumption.
3. 9.49 - produce a 95% confidence interval for the difference between the two groups, assuming equal variances.
4. 9.95 - perform the hypothesis test with the null hypothesis that there is no difference between the groups, assuming equal variances.

This document itself was written in R markdown and contains hints and examples of how to construct your own reports.

Preliminaries

Textbook Data

For the first and last time I'll tell you the textbook data will always be available at https://github.com/mie237labs/textbook_data

It's up to you how you manage the location of textbook datasets, but I would recommend for each lab you copy the required files to the fresh new directory and Rstudio project that you'll start for each lab.

Note that the textbook datasets are mostly very small and uninteresting, so I'll try to find some other datasets to work with as well from time to time.

Working with data in R

R is a full-fledged programming language so it's always possible to work with data at a very low level, but it's much easier to use packages that have been written to make working with data easier. These are the packages I use most often (especially the first three).

Package	Reason	More information
`dplyr`	filtering, grouping, summarizing	Introduction to dplyr
`ggplot2`	make nice plots	ggplot2
`rio`	data import and export	Import, Export, and Convert Data Files
`tidyr`	convert data to proper structure	Introducing tidyr
`lubridate`	working with dates	Do more with dates and times in R...

You'll need to install the packages (only once) before using them.

You'll start to pick up the usage by example in the lecture notes and in the source files of the labs.

R markdown

I suggest completing each lab in the form of a report written in R markdown, for practice since your assignments will have to be written that way. When you open a new R markdown document a skeleton example document comes up. You can also examine the R markdown documents I produce to make slides and lab instructions for more ideas.

Something to remember about R markdown. I will write the R code in the document and run the code, so that objects and things appear in the working environment for playing around with. But when it comes time to actually render the document, all the R code in the document is run in a new process independently of the working environment.

One additional package that I use in R markdown documents is the `knitr` package, mainly for the `kable` function that makes nice looking tables.

"Two-sample" t procedures

The analysis of two independent numerical samples is ultimately very straightforward. You need to:

1. Perform some exploratory data analysis (numerical and graphical summaries).
2. Do the calculations (such as using the `t.test` function in R).
3. Verify model assumptions.
4. Report on the results, which might be a confidence interval or a hypothesis test.

Fully worked example

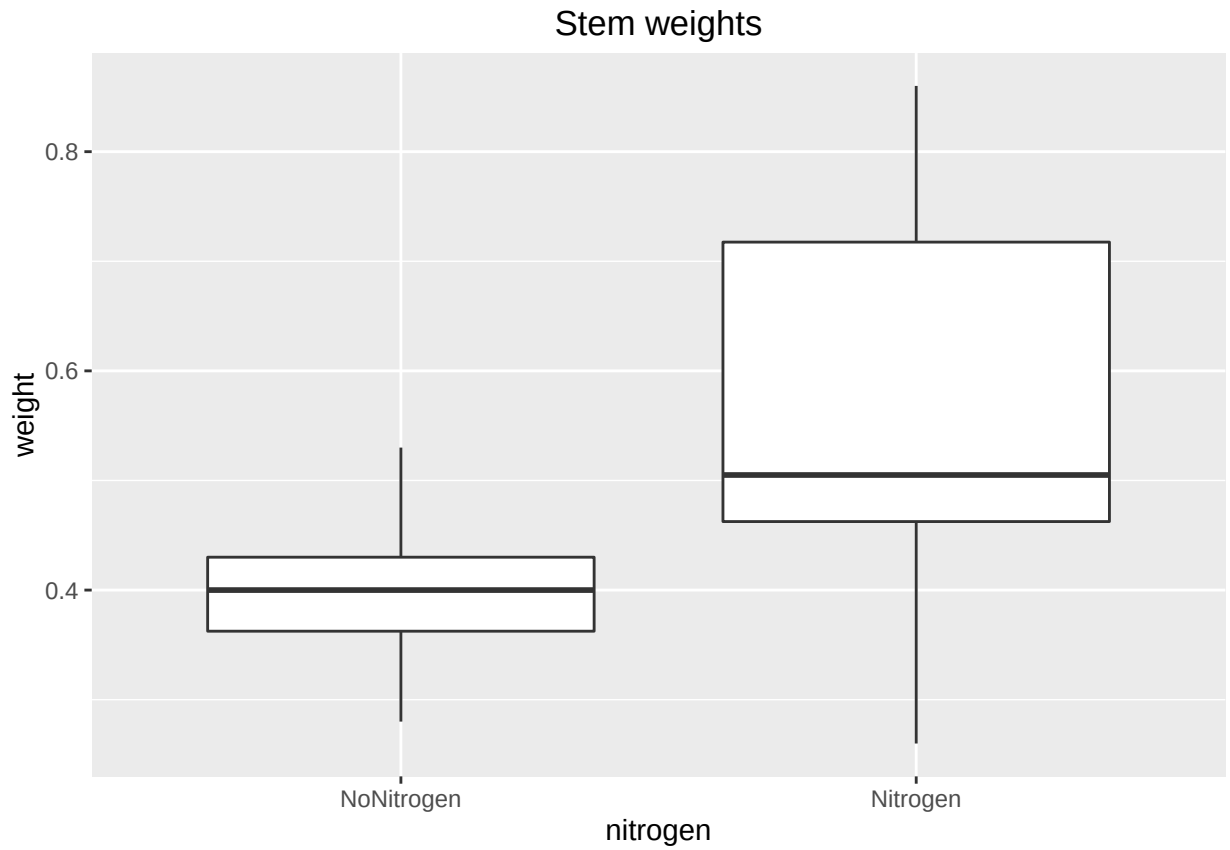
Let's look at the data for question 9.40 from the book. We did this one in class. I will write a little "report" as I might were I actually analysing this dataset for some reason. The pdf of the lab doesn't show all the underlying code. You'll need to look at the R markdown source for hints.

Example "report" using 9.40 data

This dataset concerns stem weights for trees that have been exposed, or not, to nitrogen.

Here is a numerical summary of the dataset and side-by-side boxplots of weights for the two groups of trees.

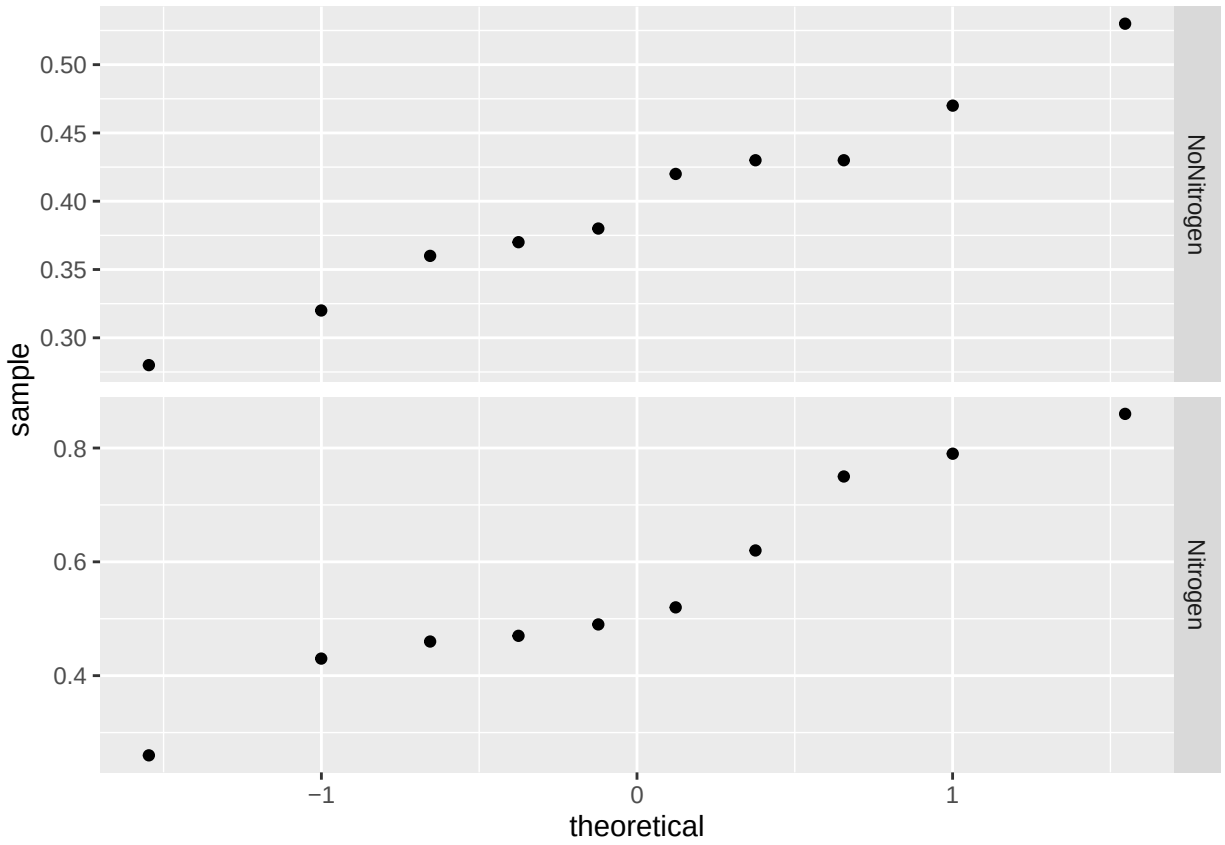
nitrogen	n	mean	sd
NoNitrogen	10	0.399	0.0727935
Nitrogen	10	0.565	0.1867411



Here is the analysis for the differene in mean weights between the groups. No equal variance assumption is being made.

```
##
##  Welch Two Sample t-test
##
## data:  weight by nitrogen
## t = -2.6191, df = 11.673, p-value = 0.02286
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.30452438 -0.02747562
## sample estimates:
## mean in group NoNitrogen   mean in group Nitrogen
##                    0.399                    0.565
```

The 95% confidence interval is [-0.305, -0.027]. The following normal quantile plots of the groups don't show evidence of a serious deviation from normality, although it is hard to tell with such small sample sizes.


0.50
0.45
0.40
0.35
0.30
NoNitrogen
0.8
0.6
0.4
Nitrogen
sample
−1
0
1
theoretical