Lab 5: Reproducible Data Analysis using RMarkdown

1 Introduction

In this lab you will learn to write a document using R markdown, integrate live R code into a literate statistical program, compile R markdown documents using knitr and related tools, and organize a data analysis so that it is reproducible and accessible to others.

1.1 Why Reproducible Research?

From the Coursera Reproducible research course description: Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them. The need for reproducibility is increasing dramatically as data analyses become more complex, involving larger datasets and more sophisticated computations. Reproducibility allows for people to focus on the actual content of a data analysis, rather than on superficial details reported in a written summary. In addition, reproducibility makes an analysis more useful to others because the data and code that actually conducted the analysis are available. This course will focus on literate statistical analysis tools which allow one to publish data analyses in a single document that allows others to easily execute the same analysis to obtain the same results.

1.1.1 RR in current day science

- http://simplystatistics.org/2014/06/06/the-real-reason-reproducible-research-is-important/
- Special articles in nature http://www.nature.com/news/reproducibility-1.17552 that talk about the need to share code and reproducibility in the sciences.
- The spreadsheet error and austerity - as seen through The Colbert Report's eyes. Moral of this story: ask for the data and question results that look too good to be true.

1.1.2 The cancer research scandal at Duke.

Simple and not so simple erorrs, combined with some fradulent coverup puts cancer patient's lives at risk.

- 60 minutes story
- 60 min interview with Baggerly
- A biostatisticans telling of the tale Skip to 1:12 for the start.
- Kieth Baggerly talking about reproducible research and big data (such as genomic type data), and what can go wrong. This is the long version, but straight from the man himself.

2 Markdown

- Follow along with this tutorial by Roger Peng. Reproduce what he does in the video on your own machine.
- Review the basics of the markdown syntax in this video from The Data Scientist's Toolbox from the Coursera course.

2.1 Practice

1. Download the Lab5a.Rmd RMarkdown file and open the Lab5a Helper.html web page from the course website. The helper HTML will guide you in writing your R and Markdown code.
2. Follow the lab instructions.

2.2 Additional Resources

R Studio provides cheat sheets for commonly used commands.