# Reproducible Research

## knitr, LaTeX and R

Statistical Consulting Group

San Diego State University

# What is R?

- Statistical Programming Environment
- Interpreted Language
- (Comparatively) easy to work/prototype in
- (can be) slower than other programming languages
- Amazing Graphic Capabilities
- Incredible power through packages

# What is LaTeX?

- TEXis a typesetting environment
  - Developed by Donald Knuth to typeset books he was writing
- LATEXis an extension to that environment, to
  - Automate tedious tasks of writing a paper
  - Generate Professional looking output
- Produces clean text with proper kerning
- Consistent and (comparatively) easy writing of mathematical formulae

# What is knitr?

- Replacement for Sweave
- Runs *R* Code, encapsulates results and prepares for typesetting in LaTeX
- Automatically formats results to look good

# Reproducible Research

- Reproducibility is fundamental to good science. The Scientific method hinges on the reproducibility of results.
- Use of knitr to automate final document preparation down to a single button (that will produce the same output regardless of who presses it) contributes to this idea of reproducible research
- Advantages to you: You don't have to copy and paste results from R into Word, or Powerpoint. If you find some bug or decide to do something different, a small change in the code and recompiling the document (1 button) averts the crisis.

# Downloads

| | |
|---|---|
| R | `http://www.r-project.org/` |
| Rstudio | `http://www.rstudio.com/` |
| MikTeX | `http://miktex.org/download` |

# How to Use

- Install all three downloads
- Install knitr: install.packages('knitr')
- Set Options > Sweave > Weave 'Rnw' Files using knitr
- Rstudio allows the use of Projects
  - Keep contained everything relating to a specific analysis
  - One Project per Assignment
- New → R Sweave Document

# A Basic Document

- Preamble
  - Document Class *(article, beamer)*
  - Package Inclusions *(extend functionality)*
  - Define New Commands
  - Author/Title Declaration
- Document
  - Title?
  - Abstract?
  - Table of Contents?
  - Sections
    - Subsections... and deeper
  - References

```
\documentclass[10pt,letterpaper]{article}
\usepackage[margin=1in]{geometry}
\usepackage{amsmath}
\usepackage{amssymb}
\title{The New Document}
\date{}
\author{Peter Trubey}
\newcommand{\iid}{\stackrel{iid}{\sim}}
\newcommand{\eqd}{\stackrel{d}{=}}
\begin{document}
\maketitle
\abstract{An abstract goes here}
\section{In the Beginning}
Some Stuff here
\subsection{things happened}
More stuff here
\subsubsection{And more things happened}
3 levels of sections?  The horrors.
\end{document}
```

# Environments

- An environment allows you to treat something differently than free text
- There are various types of environments
  - *Math* - For displaying single and multi-line mathematical formulae
    - *equation* - Single line mathematical formulae
    - *align* - Multi-line formulae - Use & to align
  - Floats
    - *figure* - For displaying Images
    - *table* - For displaying Tables
    - *verbatim* - For displaying really ugly code
- With *float* package, you can also make custon environments to suit your needs.
  - E.g., psuedocode to describe an algorithm.

# Math Environments

- Equation

$$z_i = \frac{\pi\phi_{\theta_1}(x_i)}{\pi\phi_{\theta_1}(x_i) + (1-\pi)\phi_{\theta_2}(x_i)}$$

- Align

$$\begin{aligned}
\mu_1 &= \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i} \\
\mu_2 &= \frac{\sum_{i=1}^{n}(1-z_i)x_i}{\sum_{i=1}^{n}(1-z_i)} \\
\sigma^2 &= \frac{\sum_{i=1}^{n} z_i(x_i-\mu_1)^2 + (1-z_i)(x_i-\mu_2)^2}{n} \\
\pi &= \frac{\sum_{i=1}^{n} z_i}{n}
\end{aligned}$$

# Math Environments : equation

```
\begin{equation*}
z_i = \frac{\pi\phi_{\theta_1}(x_i)}{\pi\phi_{\theta_1}(x_i)
  + (1-\pi)\phi_{\theta_2}(x_i)}
\end{equation*}
```

$$z_i = \frac{\pi\phi_{\theta_1}(x_i)}{\pi\phi_{\theta_1}(x_i) + (1-\pi)\phi_{\theta_2}(x_i)}$$

# Math Environments : align

```
\begin{align*}
\mu_1 &= \frac{\sum_{i=1}^nz_ix_i}{\sum_{i=1}^nz_i}\\
\mu_2 &= \frac{\sum_{i=1}^n(1-z_i)x_i}{\sum_{i=1}^n(1-z_i)}\\
\sigma^2 &= \frac{\sum_{i=1}^nz_i(x_i-\mu_1)^2
  + (1-z_i)(x_i-\mu_2)^2}{n}\\
\pi &= \frac{\sum_{i=1}^nz_i}{n}
\end{align*}
```

$$
\begin{aligned}
\mu_1 &= \frac{\sum_{i=1}^nz_ix_i}{\sum_{i=1}^nz_i}\\
\mu_2 &= \frac{\sum_{i=1}^n(1-z_i)x_i}{\sum_{i=1}^n(1-z_i)}\\
\sigma^2 &= \frac{\sum_{i=1}^nz_i(x_i-\mu_1)^2 + (1-z_i)(x_i-\mu_2)^2}{n}\\
\pi &= \frac{\sum_{i=1}^nz_i}{n}
\end{aligned}
$$

# A Float Environment

A figure environment, specifically

```
\begin{figure}[h]
\centering
\caption{In this figure environment, there is a fig!}
\includegraphics[width=.8\textwidth]{fig.png}
\label{fig:fig}
\end{figure}
```

- Figures automatically take number labels
- Use positioning flag to control where the figure appears
- declare reference label after caption!
- R code chunks can be declared inside figures

# Table Environments

- table
  - A wrapper for your table
  - Allows you to caption, label, and float the table
- tabular
  - The actual table
  - various types of tabular for different things
    - tabular
    - tabularx - evenly spaced columns
    - tabulary - even whitespace around column contents

# A LaTeX Table

```
\begin{table}
    \begin{tabular}{lllll}
    X  & Y  & Z & W & T \\ \hline
    1  & 3  & 1.386294361 & 1.921812056 & 1.243283885 \\
    2  & 4  & 1.791759469 & 3.210401996 & 1.475207592 \\
    3  & 5  & 2.079441542 & 4.324077125 & 1.629162771 \\
    4  & 6  & 2.302585093 & 5.30189811  & 1.743721514 \\
    \end{tabular}
\end{table}
```

| X | Y | Z | W | T |
|---|---|---|---|---|
| 1 | 3 | 1.386294361 | 1.921812056 | 1.243283885 |
| 2 | 4 | 1.791759469 | 3.210401996 | 1.475207592 |
| 3 | 5 | 2.079441542 | 4.324077125 | 1.629162771 |
| 4 | 6 | 2.302585093 | 5.30189811 | 1.743721514 |

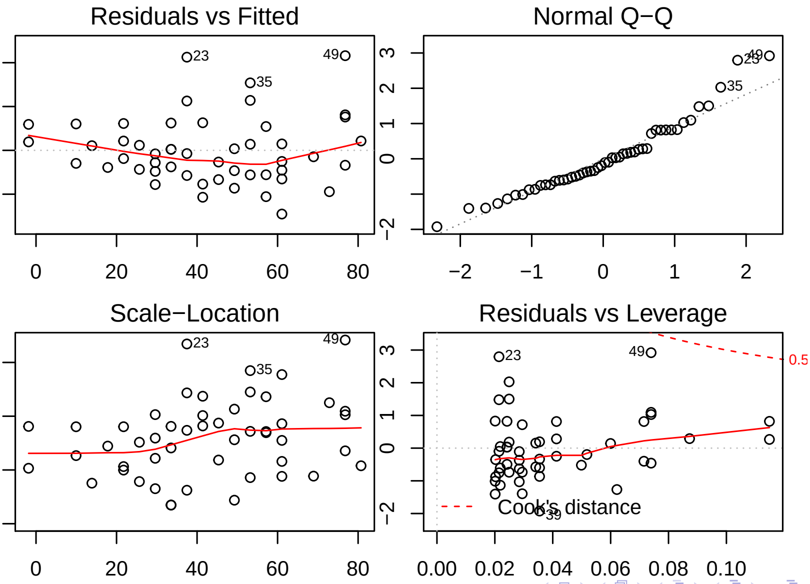
# R in LaTeX

We can dynamically run R code in LaTeX using Sweave or knitr.
We package the R code in chunks. For instance, we can run analysis in LaTeX

```
fit1 = lm(dist ~ speed, data = cars)
fit2 = lm(dist ~ speed + I(speed^2), data = cars)
anova(fit1, fit2)

## Analysis of Variance Table
##
## Model 1: dist ~ speed
## Model 2: dist ~ speed + I(speed^2)
##   Res.Df   RSS Df Sum of Sq   F Pr(>F)
## 1     48 11354
## 2     47 10825  1       529 2.3   0.14
```

We can plot these results as well. If contained within a chunk is the code to generate a plot, then knitr will generate the plot.


Residuals vs Fitted
Normal Q−Q
Scale−Location
Residuals vs Leverage
23
49
35
39
0.5
Cook's distance
0
20
40
60
80
−2
−1
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10

# Tables

Certain packages make our lives easier. For instances, xtable allows high quality automatic table generation.

```
library(xtable)
print(xtable(anova(fit1, fit2)))
```

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 48 | 11353.52 | | | | |
| 2 | 47 | 10824.72 | 1 | 528.81 | 2.30 | 0.1364 |

# Bibliography

At the end of the document, we can place the bibliography. If you set it to, LaTeX will automatically generate a bibiliography from sources you cite.

```
\bibliographystyle{plain}
\bibliography{refs}
```

- bibliographystile determines what fields from the reference file are displayed
- bibliography names the reference file (.bib file extension expected)

# A Reference File

The Reference file allows you to define referenced works. It is filled with entries such as:

```
@article{blackholes,
  author="Rabbert Klein",
  title="Black Holes and Their Relation to Hiding Eggs",
  journal="Theoretical Easter Physics",
  publisher="Eggs Ltd.",
  year="2010",
  note="(to appear)"
}
```

- The document format decides the standard fields. You won't always be able to fill every relevant field.
- There are many tools online that help to manage the bibilography.
- Of all entries in the .bib file, BibTeX will only put in the bibliography those entries who are cited in the paper.

# Some Friendly Advice...

- Make sure all R code runs properly first, before putting it in document.
- Isolate all running code into an R script file (.R), and source it in first chunk.
  - Then cache the chunk. You won't have to run the code again, making document compiling faster.
- Make all subsequent chunks output only.
  - store all outputs in variables, so you need only print the variable to get the output.
- Don't be afraid to define new commands. That's what they're there for.
  - Math equations are tedious to type. Less so with custom commands.
- Asterisks change the standard behavior of many environments/items
  - Math Environments don't get line numbers
  - sections don't show up in TOC, don't get numbers
- Easy formatting of LaTeX Tables can be accomplished with online websites: `http://truben.no/latex/table/`