

# Introduction and R

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 506: Introduction to Experimental Design

- Book is *A First Course in Design and Analysis of Experiments* by Gary Oehlert. Free PDF file on webpage.
- Grading based solely on homeworks; mostly data analyses but some thought problems.
- Lecture notes will be notes taken from text melded w/ Dr. Oehlert's notes and my own experience; posted the night before.
- We will follow the text pretty much in order, covering about 15 pages per lecture. Keep up!
- Computing will be done in R; I will show you how to install R and some necessary packages today.
- Let's first go through the syllabus.

# Prerequisites

The official requirements are one of:

- MATH 122 - Calculus for Business Administration & Social Sciences
- MATH 142 - Calculus II
- MATH 201 - Elementary Statistics

For this course, calculus is not needed at all. However, I am hoping that you have had some statistics and probability, *somewhere*.

Otherwise there is too much to introduce up front. In particular you should have seen, e.g., the normal distribution, simple hypothesis testing, and p-values.

If you have not and still want to take this course, I suggest taking a morning or afternoon and reading through some of my STAT 205 notes; see the syllabus.

The homework will mostly be data analyses and projects using the computing, so having a great math/stat background is not necessary to get something out of the course and do well.

## 1.1 Why experiment?

- Researchers use experiments to answer questions.
- Example: what type of fertilizer grows the biggest radishes in a month? Three *treatments*: Miracle Grow, organic compost juice, or nothing. Why include nothing?
- An experiment is comprised of *treatments*, *experimental units*, and *responses*.
- Experimental units here are radish plants.
- The response could be many measurements; for this example it was the weight of each plant after four weeks.

# Third place out of 80 students!



Compost juice, Miracle Grow, water only.

# Advantages of experiments

- 1 Experiments allow us directly compare the affects of the treatment(s) of interest.
- 2 Experiments can be designed to minimize bias in the comparisons.
- 3 Experiments can be designed so that error in making comparisons among treatments is small, e.g. high precision.
- 4 Experiments allow us to (carefully) infer causation.

## Experiment vs. observational study

An *observational study* also has treatments, units, and responses. However, the treatments are not controlled by us, the experimenter. Example 1.1 (p. 2) Does spanking hurt?

Observational study useful in that data are cheap and easy to get; often starts as anecdotal evidence. Observational studies often lead to prospective studies and/or experiments.

Example: kava and lung cancer.

<http://www.healthtalk.umn.edu/2014/01/07/u-m-research-finds-kava-plant-may-prevent-cigarette-smoke-induced-lung-cancer/> Why mice?

Highlights importance of ethics (p. 4). Cannot ethically induce lung cancer in humans.

# Establishing causal relationship

Mosteller and Tukey (1977) suggest that causation can be ascribed when 1 and 2 below are satisfied:

- ① Consistency: relationship is roughly same strength and direction across populations.
- ② Responsiveness: changing causal variable changes response as expected.

Experiment allows changing causal variable.

May want to hypothesize a **mechanism** behind the causal relationship.



## 1.2 Components of an experiment

An *experimental design* has treatments, experimental units, and a method to assign treatments to units.

Should keep a statistical model in mind when designing an experiment. NIH/NSF grant proposals expect this.

Good design avoids systematic error, is precise, allows estimation of pure error, and has broad validity.

## 1.2 Components of an experiment

- Systematic error is reduced through randomization of treatments, measuring devices, doctors, etc.
- Precision is increased by increasing the sample size.
- Estimation of pure error is possible by considering an appropriate statistical model for analyzing the experiential data with appropriate sample size.
- Broad validity is accomplished by sampling as large and diverse population as possible, e.g. both men and women, different ages, etc. Kava example has population of mice.

## 1.3 Terms and concepts

**Treatments** are what we manipulate, or control in the experiment: types and/or amounts of fertilizer; amount of kava extract, temperature of an industrial process, etc.

The treatments are applied to **experimental units**: radish plants, mice, humans, lightbulbs, etc. When humans are the experimental units they are called **subjects**.

**Responses** are what is measured on the experimental unit and recorded as data: weight of radish plant, how long cancer is in remission, how long light bulb works, etc. The responses (data) will let us explain differences among treatments and perform statistical inference.

## 1.3 Terms and concepts

**Randomization** is the use of a given probabilistic mechanism for the assignment of treatments, measuring devices, laboratories, etc. to experimental units.

**Experimental error** is random variation present in all responses beyond the treatment. Different experimental units will give different responses to the same treatment; even repeatedly measuring the same unit will often result in different responses, e.g. DPOAE testing.

**Blinding** occurs when the evaluators of a response do not know which treatment was given to which unit; double blinding occurs when both the evaluators and the subjects do not know the assignment of treatments. Reduces bias.

## 1.3 Terms and concepts

The **Control treatment** is “standard” treatment used as baseline or basis of comparison for the other treatments. Might be treatment in common use, or might be no treatment.

**Placebo** is a control treatment used when the act of simply applying a treatment affects the response. Examples: reduction in headache pain when given a sugar pill, sham surgery, tractor compressing soil.

A treatment level may be comprised of several **factors**, e.g. time and temperature baking bread. Various fixed settings for each factor are called levels.

## 1.3 Terms and concepts

**Confounding** occurs when the effect of one factor or treatment cannot be distinguished from that of another factor or treatment. The two factors or treatments are said to be confounded.

Consider planting red burgundy okra in fields in Richland County and silver queen okra in Charleston County. Location effects cannot be distinguished from variety effects: the variety factor and the location factor are confounded.

## 1.4 Outline of course

- Chapter 2: randomization.
- Chapters 3–7: completely randomized designs.
- Chapters 8–10: factorial treatment structure (more than one treatment).
- Chapters 11–12: random effects designs.
- Chapter 13: complete block designs.
- Chapter 14: incomplete block designs.
- Chapters 15–16: incomplete block designs with multiple treatments, split plots.
- Chapter 17: adding covariates (ANCOVA).
- Chapter 18: fractional factorials .
- Chapter 19: response surfaces (we may not get to).

## 1.5 More on experimental units

An experimental unit is often made up of **measurement units**. An experimental unit can receive any of the treatments under consideration.

Radish experiment: each row of radishes (about 5 to 10 radishes in each row) was given one of the three fertilizer treatments from a watering can. A row of radishes (often called a “plot” in agricultural experiments) is an experimental unit; each radish is a measurement unit.

Mice: mice are caged together, with different cages receiving different nutritional supplements. The cage is the experimental unit, and the mice are the measurement units.

Sometimes all measurement units from an experimental unit are combined into an overall summary, e.g. mean, median, max, etc.



## 1.5 More on experimental units

Prof. Oehlert discusses a few more issues concerning units:

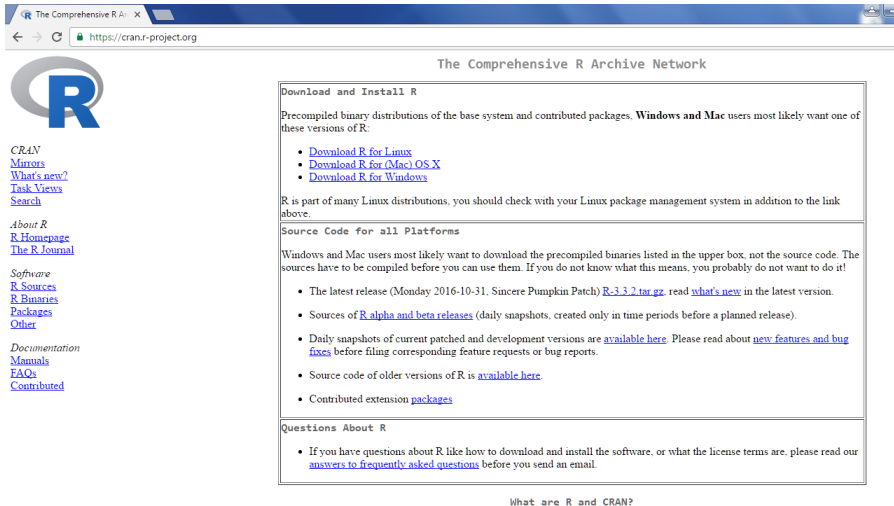
- Figuring out a size or shape of the experimental unit.
- Not all units are equivalent: edge effects, grabbing mouse closest to cage door, timing of measurement, etc.
- How to split up experimental unit into measurement units?
- Independence? Measurement units need to be well-separated in time or space so that treatments don't overlap.
- Sample size?

## 1.6 More on responses

- May be several **primary responses** of interest, each answering a different question.
- Often primary response not measurable, e.g. lifetime. **Surrogate response** is proxy for primary response, e.g. proportion of those who lived past 5 years. Another term: **biomarkers**.
- Another example: CD4, viral load counts for HIV patients; article *Comparison of CD4 Cell Count, Viral Load, and Other Markers for the Prediction of Mortality among HIV-1 Infected Kenyan Pregnant Women* at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2758232/>
- **Predictive responses** are related to the primary response (or surrogate), called “covariates.”
- **Audit responses**: were treatments applied correctly?

- R is a powerful, free statistical computing and graphics package.
- Popular with many researchers due to contributed packages: R functions to do specialized, advanced, & often complex statistical analyses.
- R can also do many important, routine calculations, analyses, and provide common graphical displays used in this course.
- Installed in several of the computing labs across campus, e.g. Sloan 108 & 109, Gambrell 003.
- You can download it and install it from CRAN:  
<http://cran.r-project.org/>

# The Comprehensive R Archive Network



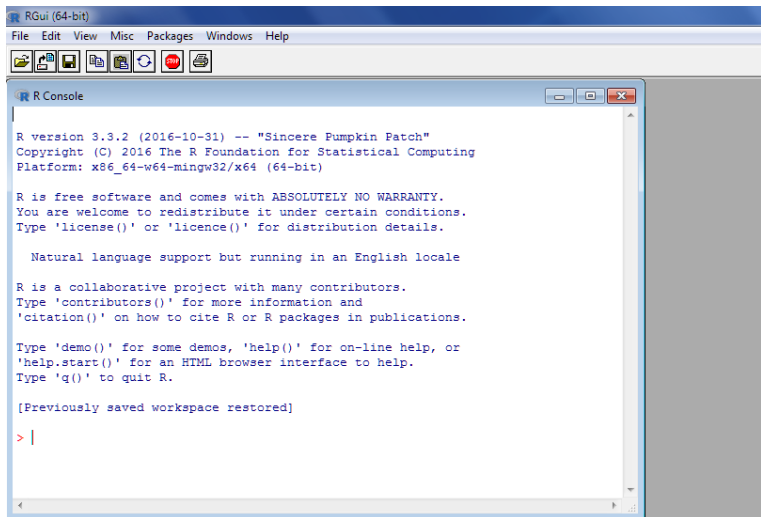
The screenshot shows a web browser window with the address bar displaying <https://cran.r-project.org>. The page title is "The Comprehensive R Archive Network". On the left side, there is a navigation menu with links for CRAN, Mirrors, What's new?, Task Views, Search, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Other, Documentation, Manuals, FAQs, and Contributed. The main content area is titled "Download and Install R" and contains the following text: "Precompiled binary distributions of the base system and contributed packages. **Windows and Mac** users most likely want one of these versions of R." followed by a bulleted list of links: "Download R for Linux", "Download R for (Mac) OS X", and "Download R for Windows". Below this, it states "R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above." The next section is "Source Code for all Platforms" with the text: "Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!" followed by a bulleted list: "The latest release (Monday 2016-10-31, Sincere Pumpkin Patch) [R-3.3.2.tar.gz](#), read [what's new](#) in the latest version.", "Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).", "Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.", "Source code of older versions of R is [available here](#).", and "Contributed extension [packages](#)". The final section is "Questions About R" with a bulleted list: "If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email." At the bottom of the page, there is a link "What are R and CRAN?".

Here is where you download R.

# Installing R

- From <http://cran.r-project.org/>, under Download and Install R click on your platform (Windows or MacOS X).
- click on base and on the next page click on Download R 3.3.2 for Windows (this is the latest release as of January 2016).
- Click  and when it's done downloading run the executable by clicking on it – alternatively you can choose to  directly after downloading from the web.
- The installation program will ask you a series of questions; choose the defaults. (e.g. English language, the suggested installation folder, the checked selected components to install, not to customize startup options, shortcut in the Start Menu, and additional tasks).
- When it's done, click on the new R desktop icon. Click on the console. This is where you will type commands to R.

# The R interface



The screenshot shows the RGui (64-bit) application window. The title bar reads "RGui (64-bit)". The menu bar includes "File", "Edit", "View", "Misc", "Packages", "Windows", and "Help". Below the menu bar is a toolbar with icons for file operations (New, Open, Save, Print, Refresh, Stop, Run). The main window contains an "R Console" pane with the following text:

```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"  
Copyright (C) 2016 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
[Previously saved workspace restored]  
  
> |
```

# Some code to try

Note that the # sign is a “comment” – R ignores anything after #.

```
# generate some random normal data
data=rnorm(100)
# look at a histogram and a boxplot
hist(data)
boxplot(data)
# compute the sample mean, median, variance, standard deviation
mean(data)
median(data)
var(data)
sd(data)
# if you have a question about a command, preface it with ?
?hist
```

Note: when it comes to R, Google is your friend.

# Installing packages needed for STAT 506

```
install.packages("lme4")
install.packages("pbkrtest")
install.packages("car")
install.packages("perm")
install.packages("effects")
install.packages("tseries")
install.packages("combinat")
install.packages("FrF2")
install.packages("RLRsim")
install.packages("rsm")
install.packages("mvtnorm")

install.packages("http://people.stat.sc.edu/hansont/stat506/oehlert_1.02.tar.gz",
  repos=NULL,type="source")

install.packages("http://people.stat.sc.edu/hansont/stat506/cfcdae_0.8-4.zip",
  repos=NULL,type="source") # Windows
install.packages("http://people.stat.sc.edu/hansont/stat506/cfcdae_0.8-4.gz",
  repos=NULL,type="source") # Mac
```



- R will allow you to do all analyses covered in this course, and beyond.
- There are some tutorials, both installed in R and on the web. Under Help choose Manuals (in PDF) and choose An introduction to R. This can get you started.
- For homework, I'll give you a skeleton set of commands to get the basic job done with no frills.
- R's error messages can be cryptic and therefore R is not as "user friendly" as some other packages such as Minitab.
- However it is free; now being used by millions of people.