Intro to Social Science Data Analysis

Lecture 4: Replication!

Christopher Gandrud

September 24, 2012

Last class we discussed:

- populations & samples,
- random samples & convenience samples,
- response, explanatory, and control variables,
- importing data into R,
- merging data sets.

Last class we discussed:

- populations & samples,
- random samples & convenience samples,
- response, explanatory, and control variables,
- importing data into R,
- merging data sets.

Last class we discussed:

- populations & samples,
- random samples & convenience samples,
- response, explanatory, and control variables,
- importing data into R,
- merging data sets.

Last class we discussed:

- populations & samples,
- random samples & convenience samples,
- response, explanatory, and control variables,
- importing data into R,
- merging data sets.

Last class we discussed:

- populations & samples,
- random samples & convenience samples,
- response, explanatory, and control variables,
- importing data into R,
- merging data sets.

Review Quiz (1)

Imagine you have a data set in a `.csv` file on your desktop.

Describe the steps you would take to import the data set into R.

Review Quiz (2)

1. What is the difference between a population and a sample?
2. What do you need to be honest about when using convenience samples?

Review Quiz (2)

1. What is the difference between a population and a sample?
2. What do you need to be honest about when using convenience samples?

Review Quiz (3)

Comment the code:

```
#
library(reshape)

#
Data <- rename(Data, c(country_name = "Country"))

#
Data$Country[Data$Country
             == "Dem. Rep. Congo"] <- "DRC"
```

Now we know how to get data into R in a format we can use for statistical analysis.

Before we start analysing the data, it is important to learn an important computational research skill . . .

Reproducible Research

Replicable Research

Research is replicable if *there is sufficient information for independent researchers to make the same findings using the same procedures* (King 1995, 444).

For example,

A team of scientists clone a sheep.

The team documents the prodedures they use to clone the sheep and make these procedures available on their website.

Another team of scientists is able to use the information on the website to clone another sheep.

Reproducible Research

Reproducible reasearch is when *the data and code used to make a finding are available and they are sufficient for an independent researcher to make the same findings (see Peng 2011).*

Why reproducible, not replicable?

Sometimes with observational data it may require too many resources to gather a new set of data or, especially in the case of cross-country data it may not be possible to gather other data.

So, reproducibility can be a **second-best** if replicability is difficult.

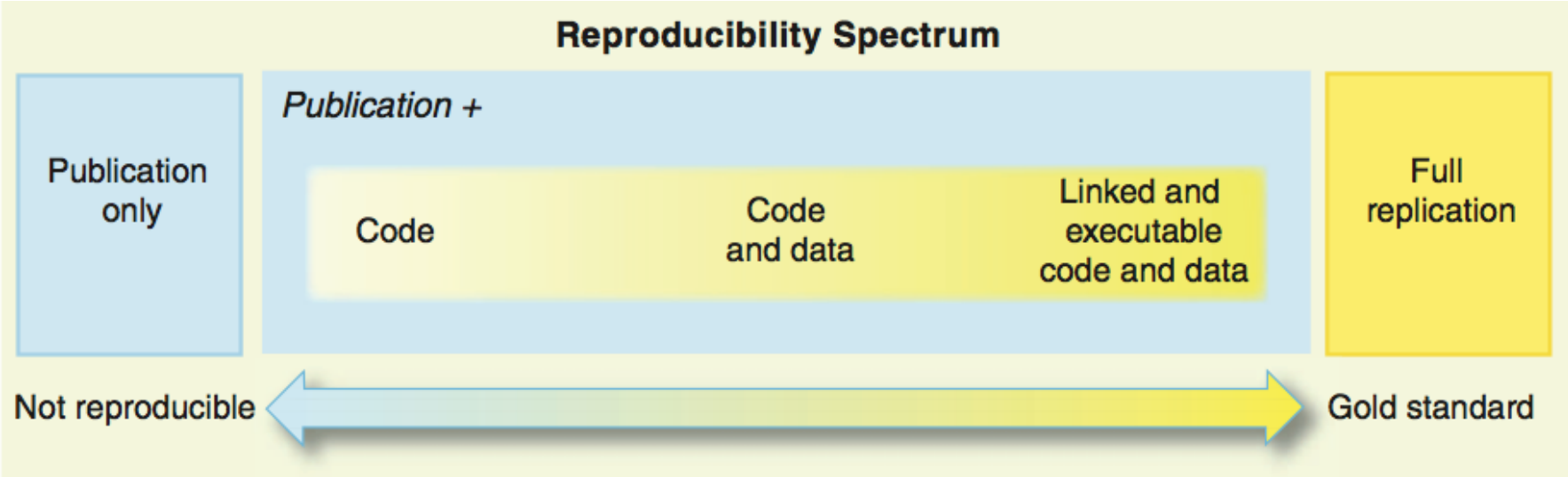
What is research?

A book, an article, or a research paper is **not research**.

It is an **advertisment** for the research.

The research is "the **full software environment, code, and data** that produced the results" (Buckheit and Donoho, 1995; Donoho, 2010, 385)

Doing reproducible computational research.



Peng (2011, 1226)

Why do reproducible research?

In this class I require you to.

Reproducible research is a key part of science.

Reproducible research & science

Braude (1979, 2) has even called replication the "demarcation between science and non-science."

Science

The "systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into **testable** laws and theories"

(see APA `http://www.aps.org/policy/statements/99_6.cfm`)

Replication & Science

Scientists need to "**expose** their ideas and results to independent testing and replication by others.

This requires the **open exchange of data, procedures and materials**."

(Emphasis added, see APA
`http://www.aps.org/policy/statements/99_6.cfm`)

Avoiding effort duplication

Making your data and procedures available also helps avoid **effort dublication**.

- Others don't have to gather the same data or figure out the same analysis that you have already done.

This helps build **cumulative scientific knowledge**.

Reproducible research is good for you.

As you will see, making your research reproducible requires an extra **upfront** investment.

Beyond the benefits for science, why should you make your research more reproducible?

Why reproducible research benefits you.

Reproducible research benefits for you:

- Better work habits.
- Better teamwork.
- Making changes is easier.
- Higher research impact.

Reproducible research benefits for you:

- Better work habits.
- Better teamwork.
- Making changes is easier.
- Higher research impact.

Why reproducible research benefits you.

Reproducible research benefits for you:

- Better work habits.
- Better teamwork.
- Making changes is easier.
- Higher research impact.

Reproducible research benefits for you:

- Better work habits.
- Better teamwork.
- Making changes is easier.
- Higher research impact.

We have *already started* doing reproducible research.

- You have already been making your code human readable with comments after the #.
- You have been compiling notebooks of your code and output.

We have *already started* doing reproducible research.

- You have already been making your code human readable with comments after the #.
- You have been compiling notebooks of your code and output.

So far we have mostly focused on the source code.

Today we will learn how to **weave** the source code and presentation documents together.

First, we will learn the basics of the **Markdown** markup language.

Second, then we will learn how to use the *knitr* package to weave our *source code* and *Markdown presentation documents*.

Doing reproducible computational research.

To be "really reproducible" we need a way to **link the data, code, and presentation** documents.

The *knitr* package for R is this link.

$$\text{R Source Code} \Longleftrightarrow \text{knitr package} \Longleftrightarrow \text{Markdown}$$

The Markdown Markup Language.

Markup Language

Instructions for how to format a presentation document.

You are probably familiar with the HTML markup language used to create websites.

HTML is a Pain!

```
<style>
    li, p { font-size: 11pt; line-height: 125%; margin: 20px; }
</style>

<h2>Posts</h2>
<table class="table table-striped">
  <tbody>
        {% for post in site.posts %}
    <tr>
      <td>
                  <h3><a href="{{ post.url }}">{{ post.title }}</a></h3>
          <i>{{ post.summary }}</i>
                  <p><small>{{ post.date | date: "%B %e, %Y" }} . {{ post.category }} . {% fo
          </td>
    </tr>
        {% endfor %}
  </tbody>
</table>
```

Markdown

The Markdown markup language *simplifies* the process of creating HTML markup files.

Markdown files have the file extension: `.md` or `.markdown`.

Markdown → HTML

```
# INT3042: Introduction to Social Science Data Analysis
## Yonsei University, Wonju
### Fall 2012
### Version: 13 September 2012

---

**Instructor:** Dr. Christopher Gandrud

- Email: <gandrud@yonsei.ac.kr>
- Website: <http://christophergandrud.blogspot.com/>

**Office Hours:** 15:00-17:00 Wednesday (정208)

- You can also send me an email or come to my office whenever you need to.

**Time:** 11:00-12:15 Tuesday & Thursday (정215)

**Objectives:** This course's main objective is to: *learn how to take raw social science data, explore it, and present the results in a useful way*. In this course you will learn all of the basic skills needed to do each of these steps with the statistical language **R**. *Part I* of the course introduces you to both basic data structures and **RStudio** (a program that makes using **R** easier). *Part II* introduces basic data analysis and visualizations techniques. *Part III* covers slightly more advanced statistical tools, primarily linear regression. Finally, in *Part IV* we will apply all of these skills in a pair research project.

As part of achieving this straightforward objective, the course is intended to also do the following:

The course is intended to be *useful*. I hope that the course will be one of the more useful courses you take in university. It is intended to be useful for students who want to go on to do graduate-level *academic research*. It is also intended to be useful for students who want to go directly into the *non-
```

```
<h1 id="int3042:introductiontosocialsciencedataanalysis">INT3042: In
Social Science Data Analysis</h1>

<h2 id="yonseiuniversitywonju">Yonsei University, Wonju</h2>

<h3 id="fall2012">Fall 2012</h3>

<h3 id="version:13september2012">Version: 13 September 2012</h3>

<hr />

<p><strong>Instructor:</strong> Dr. Christopher Gandrud </p>

<ul>
<li>Email: <a
href="&#109;&#97;&#105;&#108;&#116;&#111;&#58;&#x67;&#97;&#110;&#x64
x64;&#x40;&#121;&#111;&#x6e;&#115;&#x65;&#x69;&#x2e;&#x61;&#99;&#x2e
&#x67;&#x61;&#110;&#x64;&#x72;&#x75;&#x64;&#x40;&#121;&#x6f;&#110;&#
;&#46;&#97;&#99;&#46;&#x6b;&#114;</a></li>
<li>Website: <a href="http://christophergandrud.blogspot.com/">http:
christophergandrud.blogspot.com/</a></li>
</ul>

<p><strong>Office Hours:</strong> 15:00&#8211;17:00 Wednesday (정208)

<ul>
<li>You can also send me an email or come to my office whenever you
</ul>

<p><strong>Time:</strong> 11:00&#8211;12:15 Tuesday &amp;amp; Thursday (

<p><strong>Objectives:</strong> This course&#8217;s main objective i
how to take raw social science data, explore it, and present the res
useful way</em>. In this course you will learn all of the basic skil
each of these steps with the statistical language <strong>R</strong>
em> of the course introduces you to both basic data structures and
<strong>RStudio</strong> (a program that makes using <strong>R</stro
<em>Part II</em> introduces basic data analysis and visualizations t
<em>Part III</em> covers slightly more advanced statistical tools, p
regression. Finally, in <em>Part IV</em> we will apply all of these
pair research project. </p>
```

Final

```
# INT3042: Introduction to Social
Science Data Analysis
## Yonsei University, Wonju
### Fall 2012
### Version: 13 September 2012

---

**Instructor:** Dr. Christopher Gandrud

- Email: <gandrud@yonsei.ac.kr>
- Website: <http://christophergandrud.blogspot.com/>

**Office Hours:** 15:00-17:00 Wednesday (정208)

- You can also send me an email or come to my office whenever you need to.

**Time:** 11:00-12:15 Tuesday & Thursday (정215)

**Objectives:** This course's main objective is to: *learn how to take raw
social science data, explore it, and present the results in a useful way*. In
this course you will learn all of the basic skills needed to do each of these
steps with the statistical language **R**. *Part I* of the course introduces
you to both basic data structures and **RStudio** (a program that makes using
**R** easier). *Part II* introduces basic data analysis and visualizations
techniques. *Part III* covers slightly more advanced statistical tools,
primarily linear regression. Finally, in *Part IV* we will apply all of these
skills in a pair research project.

As part of achieving this straightforward objective, the course is intended to
also do the following:

The course is intended to be *useful*. I hope that the course will be one of
the more useful courses you take in university. It is intended to be useful for
students who want to go on to do graduate-level *academic research*. It is also
```

INT3042: Introduction to Social Science Data Analysis

Yonsei University, Wonju

Fall 2012

Version: 13 September 2012

Instructor: Dr. Christopher Gandrud

- Email: gandrud@yonsei.ac.kr
- Website: http://christophergandrud.blogspot.com/

Office Hours: 15:00-17:00 Wednesday (정208)

- You can also send me an email or come to my office whenever you ne

Time: 11:00-12:15 Tuesday & Thursday (정215)

Objectives: This course's main objective is to: *learn how to take raw socia science data, explore it, and present the results in a useful way*. In this cou will learn all of the basic skills needed to do each of these steps with the st language **R**. *Part I* of the course introduces you to both basic data structur

Basic Markdown Syntax

You can find a guide to basic Markdown Syntax at:

- `http://daringfireball.net/projects/markdown/basics`

Also, **anything** you can do in HTML you can do in Markdown.

Free programs for editing markdown documents

- **Mac:** Mou (`http://mouapp.com/`).
- **Windows:** MarkdownPad (`http://markdownpad.com/`).

These programs allow you to create webpages & PDFs.

You can host HTML pages from your Dropbox Public Folder

Free programs for editing markdown documents

- **Mac:** Mou (`http://mouapp.com/`).
- **Windows:** MarkdownPad (`http://markdownpad.com/`).

These programs allow you to create webpages & PDFs.

You can host HTML pages from your Dropbox Public Folder

R Markdown

If we want to combine R code and Markdown we can use RStudio.

To create a new R Markdown document:

`File` → `New` → `R Markdown`

R Markdown documents have the file extension `.Rmd`.

The RStudio Markdown Top Bar, Left Side

Open a new R Markdown document & play with these buttons in RStudio.

What do they do?

Now that we understand the basics of the Markdown markup language, lets start "**knitting**" R code into our presentation documents.

The “Knit HTML” Button

CLicking on the “Knit HTML” button ():

- Runs the R **code chunks**,
- Puts the output into a new plain Markdown file (`.md`),
- Converts the Markdown file to HTML.

The "Knit HTML" Button

CLicking on the "Knit HTML" button ():

- Runs the R **code chunks**,
- Puts the output into a new plain Markdown file (`.md`),
- Converts the Markdown file to HTML.

The "Knit HTML" Button

CLicking on the "Knit HTML" button ():

- Runs the R **code chunks**,
- Puts the output into a new plain Markdown file (`.md`),
- Converts the Markdown file to HTML.

Code Chunks

We place R code inside of code chunks, this seperates them from the markup and text.

Knitr looks for code chunks and runs the code in them.

Knitr/Markdown Code Chunk Syntax

````
# A Knitr/Markdown code chunk starts like this
```{r}
````

````
# A Knitr/Markdown code chunk ends like this
```
````

Chunk Options

You can add options to your chunks to change how they behave.

Chunk Options

````
```{r}
summary(cars)
```

You can also embed plots, for example:

```{r fig.width=7, fig.height=6}
plot(cars)
```
````

Chunk Options

The options:

```
```{r fig.width=7, fig.height=6}
```

Set how wide and tall the output figure is.
```

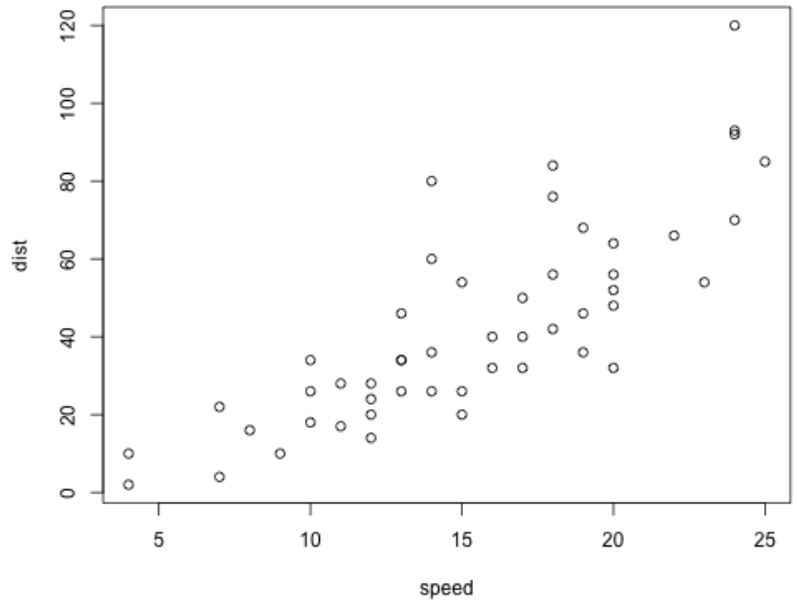
R Markdown Final

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2
##  1st Qu.:12.0   1st Qu.: 26
##  Median :15.0   Median : 36
##  Mean   :15.4   Mean   : 43
##  3rd Qu.:19.0   3rd Qu.: 56
##  Max.   :25.0   Max.   :120
```

You can also embed plots, for example:

```
plot(cars)
```



Useful Chunk Options

Chunk Option	Description
`eval=FALSE`	Does not run the code
`echo=FALSE`	Does not include the code
`error=FALSE`	Does not include error messages
`warning=FALSE`	Does not include warning messages
`message=FALSE`	Does not include message messages
`fig.align='center'`	Centers a figure

More Chunk Options

All chunk options can be found at:

`http://yihui.name/knitr/options`.

If you want well formatted PDF files and slide shows (like this one), especially if you plan to go to graduate school.

You can learn how to use knitr and LaTeX.

See me for more details.

Note: LaTeXmarkup syntax is more complicated than Markdown. You also do not need to know LaTeXfor this course.

References I

Braude, S.E. ESP and Psychokinesis. A philosophical examination. Temple University Press, Philadelphia, PA, 1979.

Buckheit, J. B. and Donoho, D. L. (1995). Wavelab and Reproducible Research, pages 5581. Springer, New York.

Donoho, D. L. (2010). An Invitation to Reproducible Computational Research. Biostatistics, 11(3):385388.

King, Gary. 1995. Replication, Replication. PS: Political Science and Politics 28(3): 444452.

Peng, Roger D. 2011. Reproducible Research in Computational Science. Science 334:1226-1227.