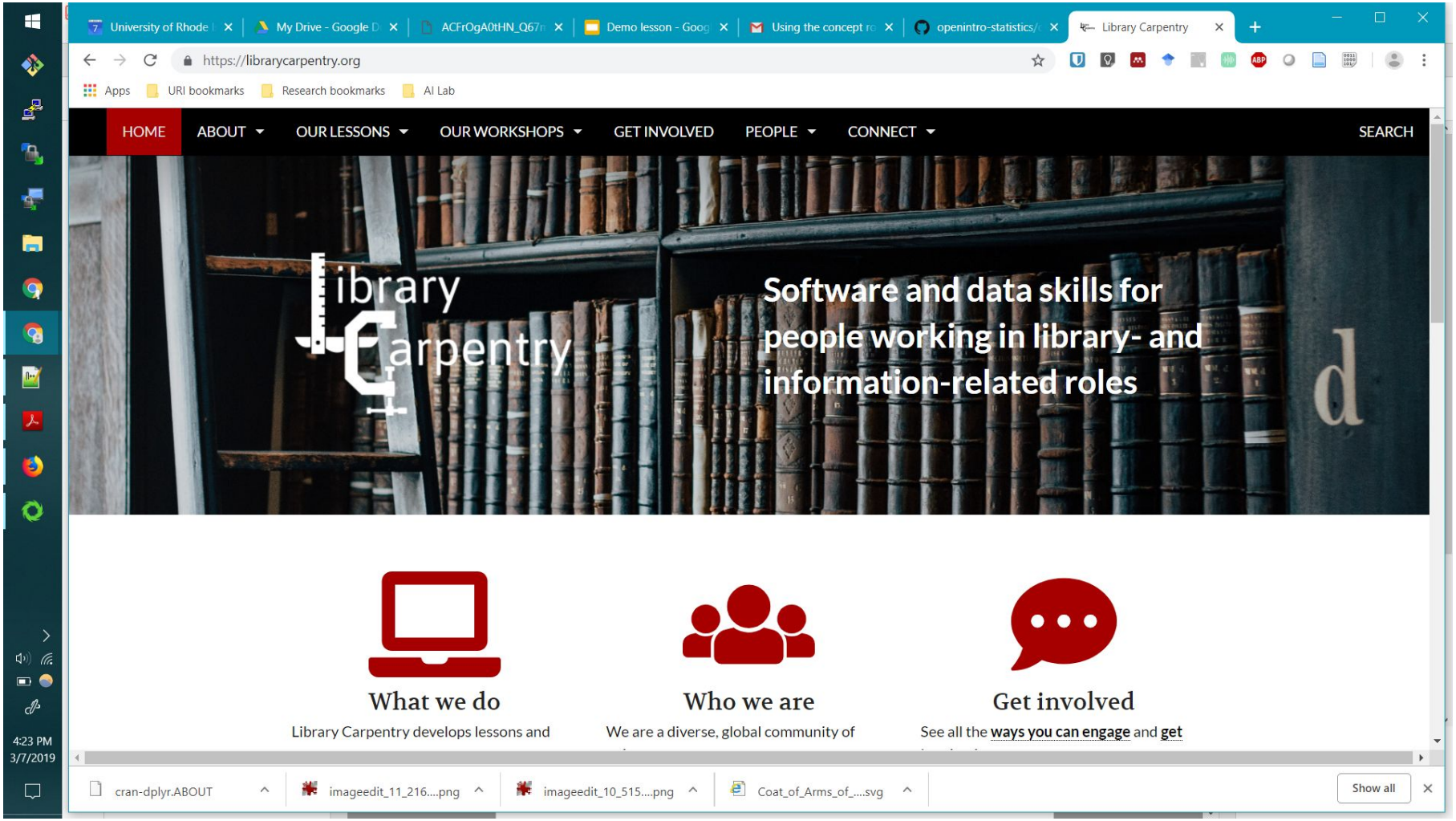
Data Literacy for Librarians (demo lesson)

Harrison Dekker


University of Rhode
My Drive - Google D
ACFrOgA0tHN_Q67r
Demo lesson - Goog
Using the concept ro
openintro-statistics/c
Library Carpentry
https://librarycarpentry.org
Apps
URI bookmarks
Research bookmarks
AI Lab
HOME
ABOUT
OUR LESSONS
OUR WORKSHOPS
GET INVOLVED
PEOPLE
CONNECT
SEARCH
Library Carpentry
Software and data skills for people working in library- and information-related roles
What we do
Library Carpentry develops lessons and
Who we are
We are a diverse, global community of
Get involved
See all the ways you can engage and get
cran-dplyr.ABOUT
imageedit_11_216....png
imageedit_10_515....png
Coat_of_Arms_of_....svg
Show all
4:23 PM
3/7/2019

Attribution

This lesson is derived from OpenIntro Statistics, Third Edition, an open source textbook, and accompanying video tutorials that can be found online at https://www.youtube.com/watch?v=nEHFF1ADpWE&list=PLkIselvEzpM6pZ76FD3NoCvvgkj_p-dE8 and at https://www.youtube.com/watch?v=Mjif8PTgzUs&index=2&list=PLkIselvEzpM6pZ76FD3NoCvvgkj_p-dE8

The role of data in scientific research

- Scientific research is all about seeking **answers to questions** by making **careful observations** and applying **rigorous methods**.
- **Field notes**, **surveys**, and **experiments** are some of the ways observations (i.e. data) are collected.
- **Statistics** is the study of how best to collect, analyze, and draw conclusions from data.

Case Study

Stents and risk of stroke

In this section we will consider an experiment that studies the effectiveness of stents in treating patients at risk of stroke.

- Stents are devices put inside blood vessels that assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death.
- Does the use of stents reduce the risk of stroke?

Treatment and control groups

The study involved 451 at risk patients. Each volunteer patient was randomly assigned to one of two groups.

- Patients in the **treatment group** received a stent and medical management, which included medications, management of risk factors, and help in lifestyle modification.
- Patients in the **control group** received the same medical management as the treatment group but they did not receive stents.

Data table

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment.

Patient	Group	0-30 days	0-365 days
1	treatment	No event	No event
. . .	. . .	. . .	. . .
451	control	No event	Stroke

Results from two patients from the stents study

Data summary

A statistical data analysis allows us to consider all of the data at once.\

	0 - 30 days		**0 - 365 days**	
	stroke	**no event**	**stroke**	**no event**
treatment	33	191	45	179
control	13	214	28	199
Total	**46**	**405**	**73**	**378**

Descriptive statistics for the stent study.

Practice

	0 - 30 days		0 - 365 days	
	stroke	**no event**	**stroke**	**no event**
treatment	33	191	45	179
control	13	214	28	199
Total	**46**	**405**	**73**	**378**

Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year.

Summary statistics

A summary statistic is a single number based on the sample that summarizes a large amount of data.

- Proportion who had a stroke in the treatment group:

$$\frac{\text{45 patients recieved a stent and had a stroke}}{\text{224 patients recieved a stent}} = 0.20 = 20\%$$

- Proportion who had a stroke in the control group:

$$\frac{\text{28 patients recieved a stent and had a stroke}}{\text{227 patients recieved a stent}} = 0.12 = 12\%$$

Random fluctuation

Do the data show a real difference between the groups?

- Fluctuation is part of almost any type of data generating process.
- It is possible that the 8% difference in the stent study is due to this natural variation.

Summary

Case study - using stents to prevent strokes

- Stents and risk of stroke
- Treatment and control groups
- Data table
- Summary statistics
- Random fluctuation

Data Basics

Outline

Data Basics

- Observations, variables, and data matrices
- Types of variables

The email data set

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
. . .	. . .	. . .	. . .	. . .	. . .
50	no	15,829	242	html	small

Variables and descriptions

variable	description
spam	Specifies whether the message was spam
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
format	Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format
number	Indicates whether the email contained no number, a small number (under 1 million), or a large number.