STAT 113
Introduction to Statistics

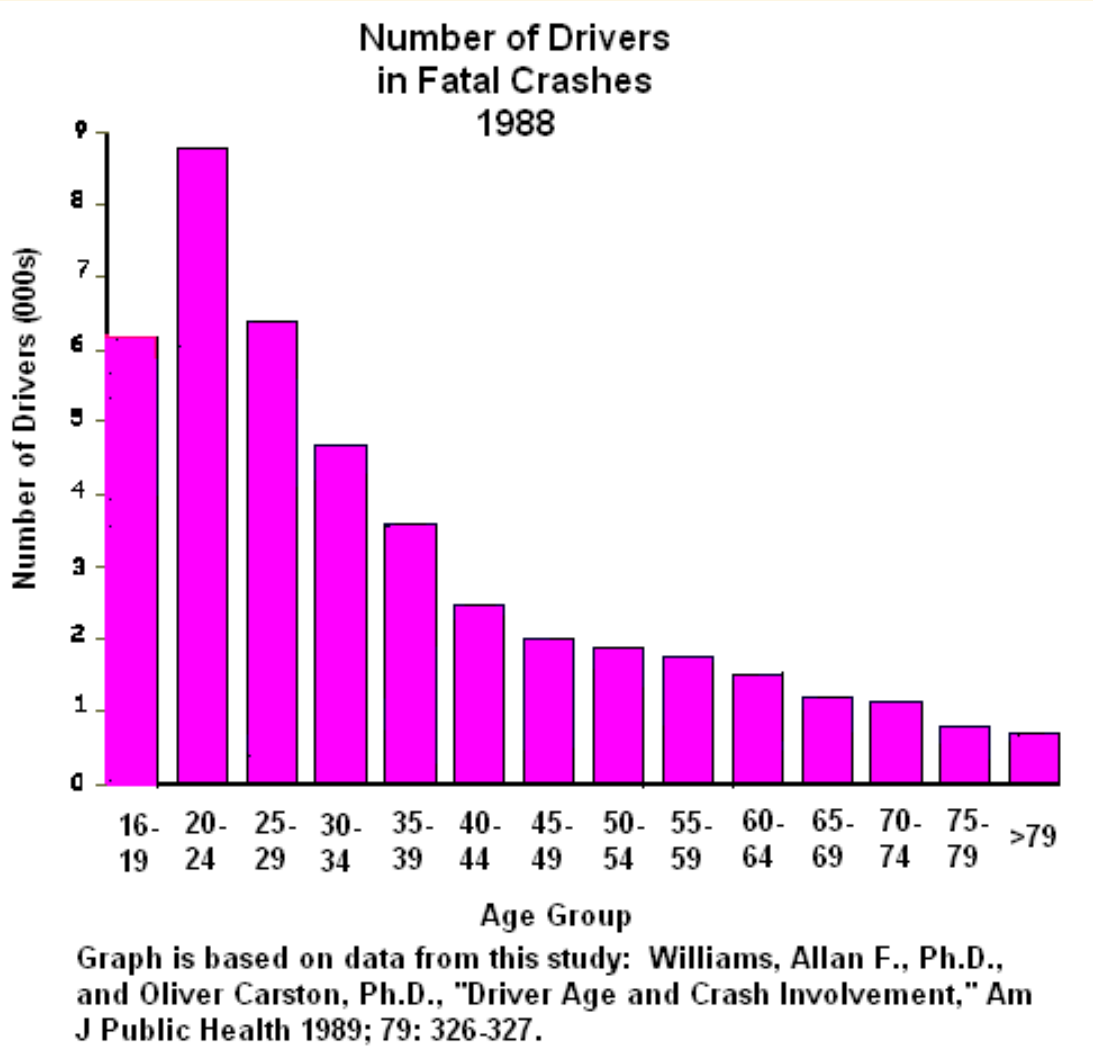
Colin Reimer Dawson

Oberlin College

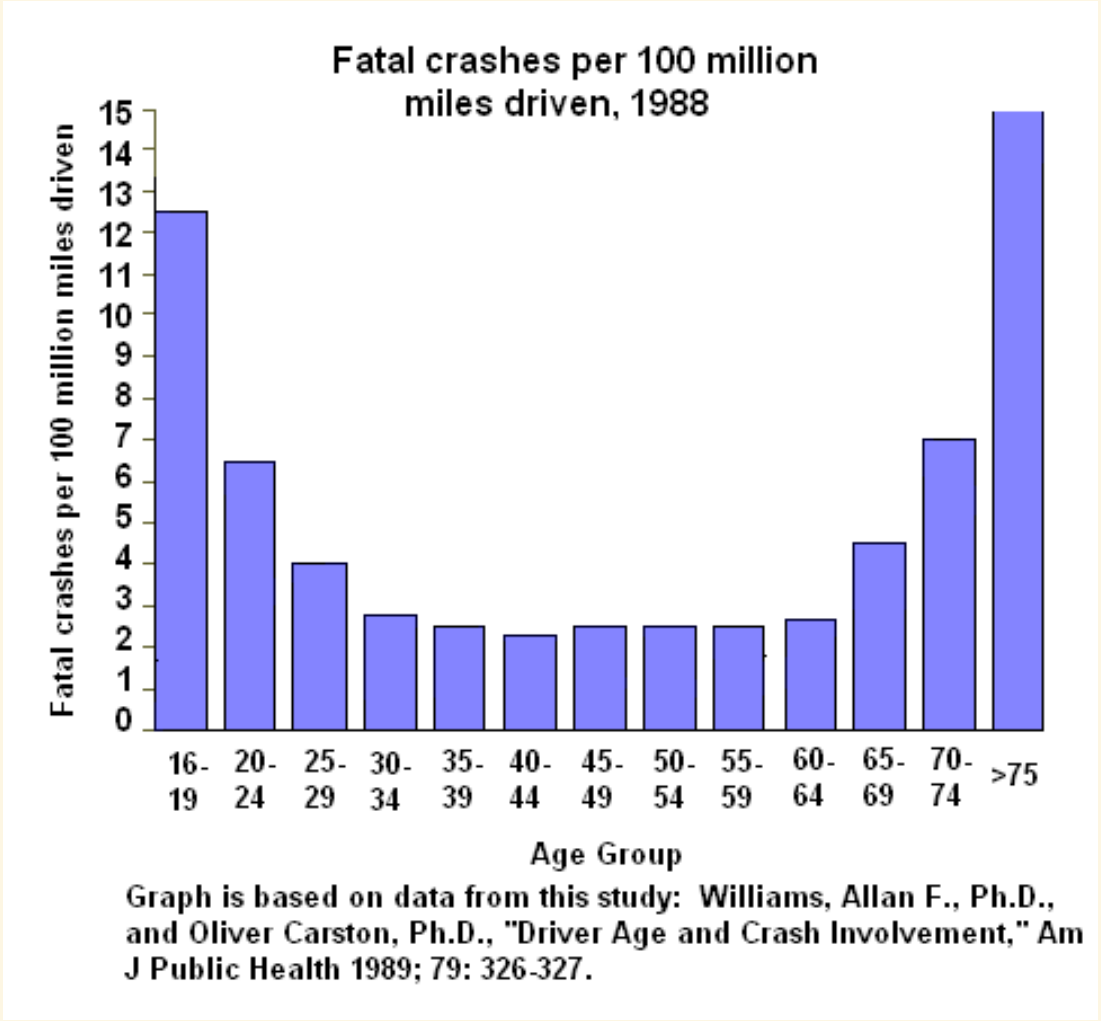
August 29, 2017

Outline

Bar Chart: Fatal Crashes



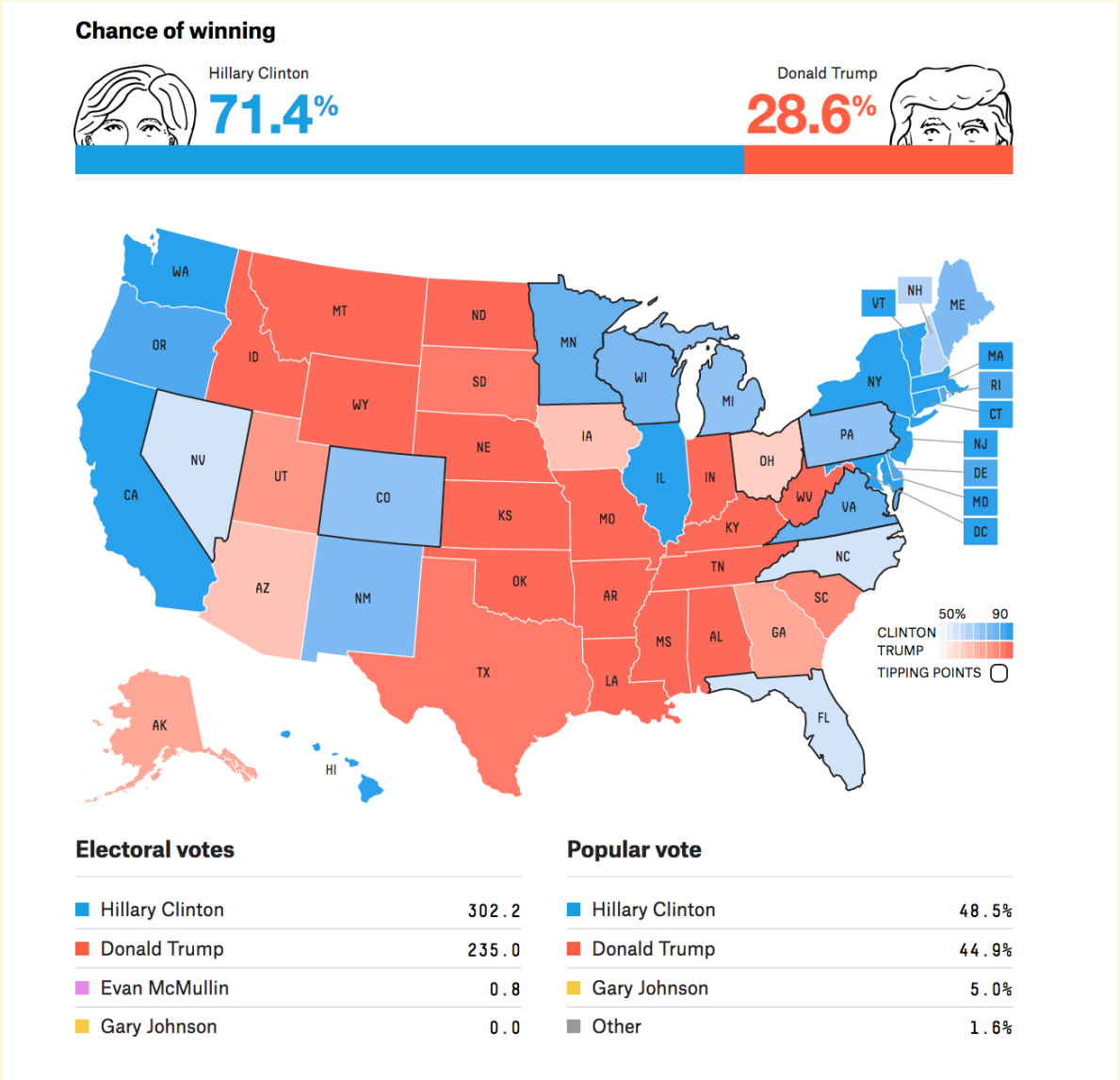
Bar Chart, same data



The Importance of an Accurate Baseline

- A theme of this course: "If there were nothing interesting going on, what data would we expect to see?"
- Interpret actual results relative to this "baseline"

Election Prediction: A Failure of Statistics?



Sampling Error

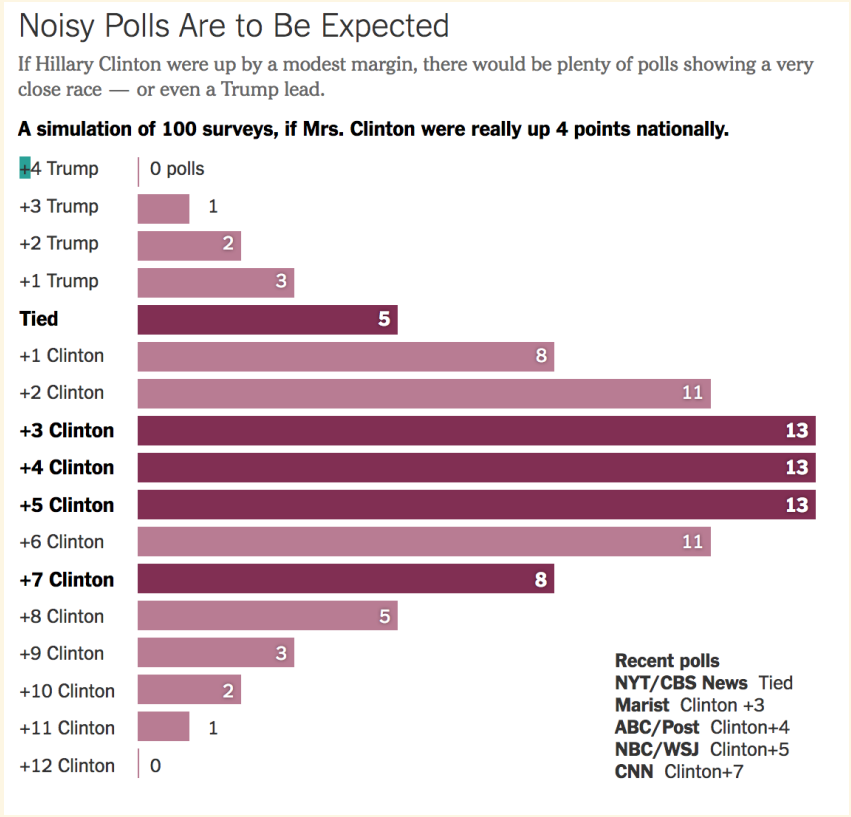


Figure: Polls come from samples, not the population, and are subject to sampling error. Source: The New York Times "The Upshot" 7/18/16.

Sampling and Non-response Bias?

STATE	WHITE NON-COLLEGE SHARE	TRUMP MARGIN OF VICTORY: ADJ. POLLING AVERAGE	TRUMP MARGIN OF VICTORY: ACTUAL	TRUMP MARGIN OF VICTORY: ACTUAL VS. POLLS
West Virginia	65.7%	+27.9	+42.2	+14.3
Kentucky	62.2	+18.4	+29.8	+11.4
Iowa	62.0	+3.4	+9.4	+6.0
Maine	61.6	-6.9	-2.7	+4.2
Idaho	60.4	+19.7	+31.8	+12.1
North Dakota	59.6	+24.7	+35.7	+11.0
Wisconsin	57.2	-5.4	+0.7	+6.1
Montana	56.7	+16.4	+20.5	+4.1
New Hampshire	56.5	-3.5	-0.4	+3.1
Wyoming	55.0	+36.3	+46.3	+10.0
Virginia	36.7	-5.4	-5.3	+0.1
Georgia	34.2	+4.0	+5.2	+1.2
New Jersey	32.9	-11.2	-14.1	-2.9
Texas	31.4	+8.5	+9.1	+0.6
New York	29.8	-18.7	-21.2	-2.5
Maryland	29.2	-26.3	-26.6	-0.3
New Mexico	27.5	-5.3	-8.2	-2.9
California	26.4	-23.0	-30.0	-7.0
Hawaii	15.2	-20.8	-32.2	-11.4
D.C.	2.2	-69.3	-86.8	-17.5

Adjusted polling average is from the FiveThirtyEight polls-only forecast as of Nov. 8, 2016. Actual results are accurate as of Nov. 30.

Figure: Is the sample *representative* of the population? Polling errors strongly correlated with state share of whites w/o college degrees. Source: fivethirtyeight.com

Moving Targets

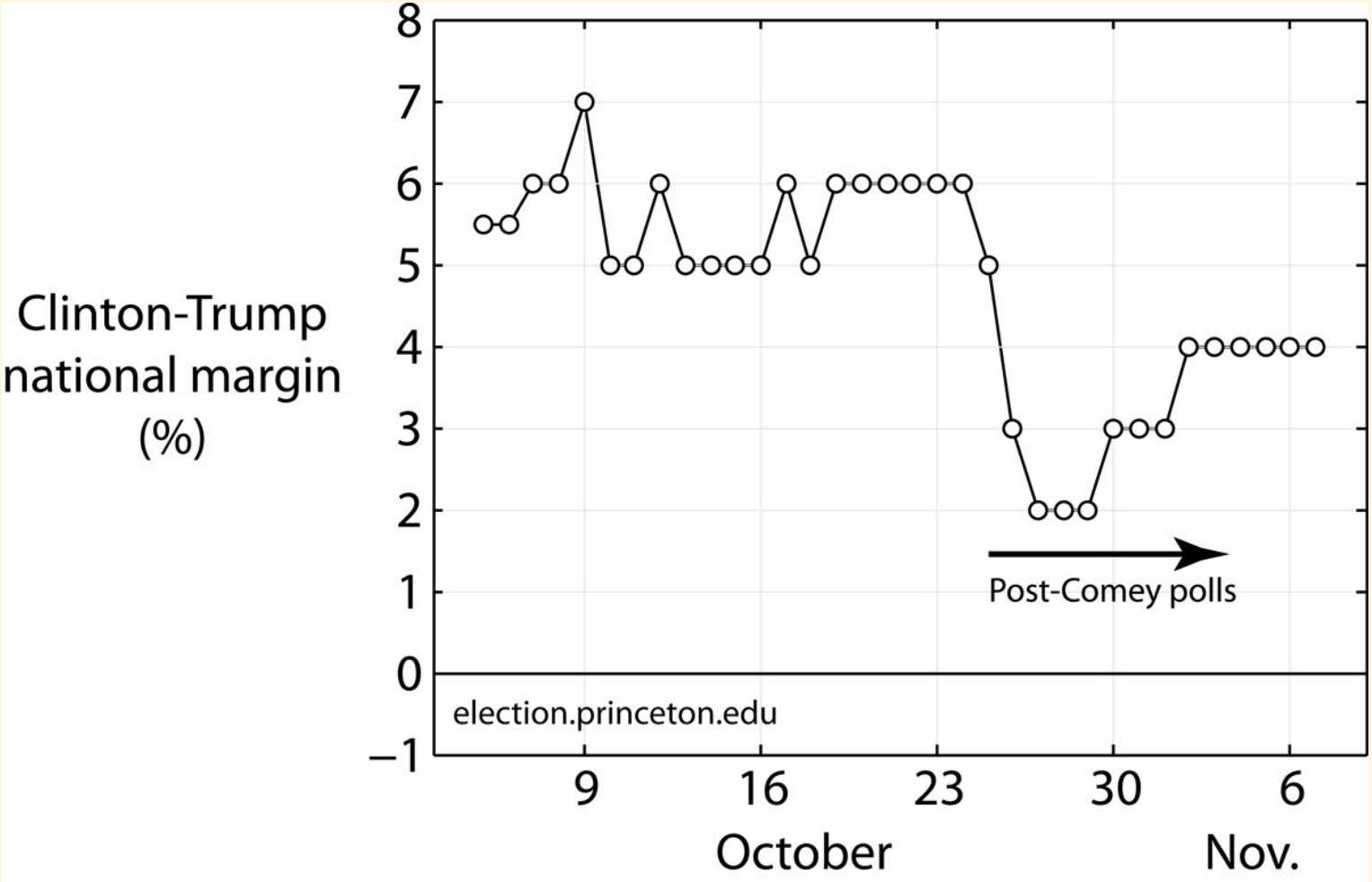


Figure: Is our data about what we want it to be about? Source: Princeton Election Consortium `http://election.princeton.edu`

Race and the Death Penalty

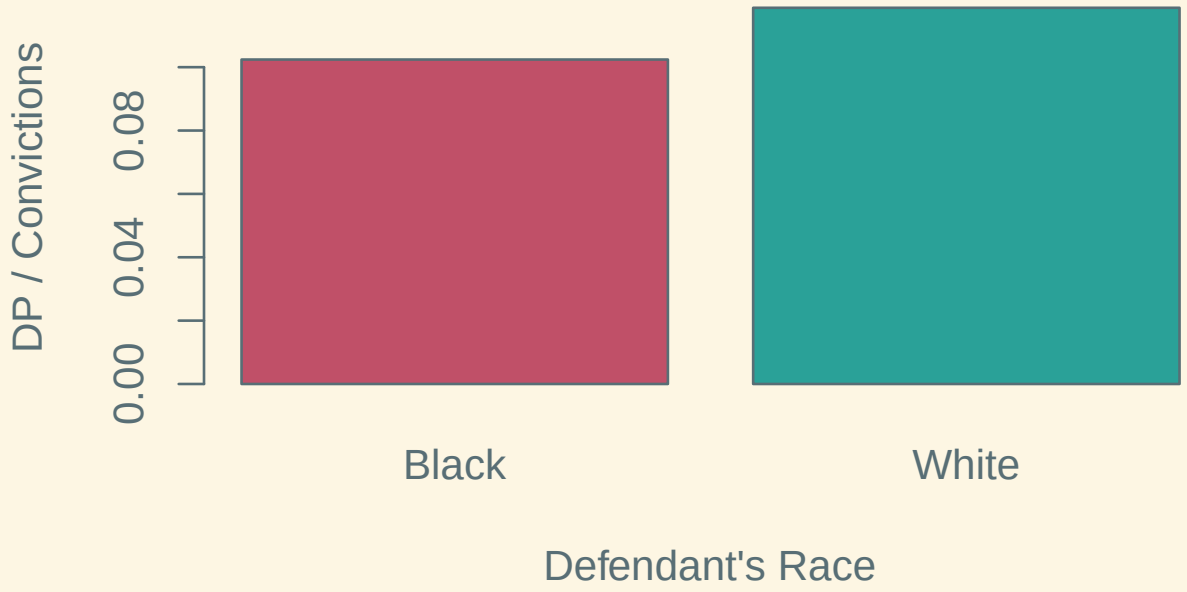


Figure: Proportions of Death Sentences out of Total Convictions (Florida, 1981), Grouped Race of Defendant. Source: Agresti (2002)

Race and the Death Penalty

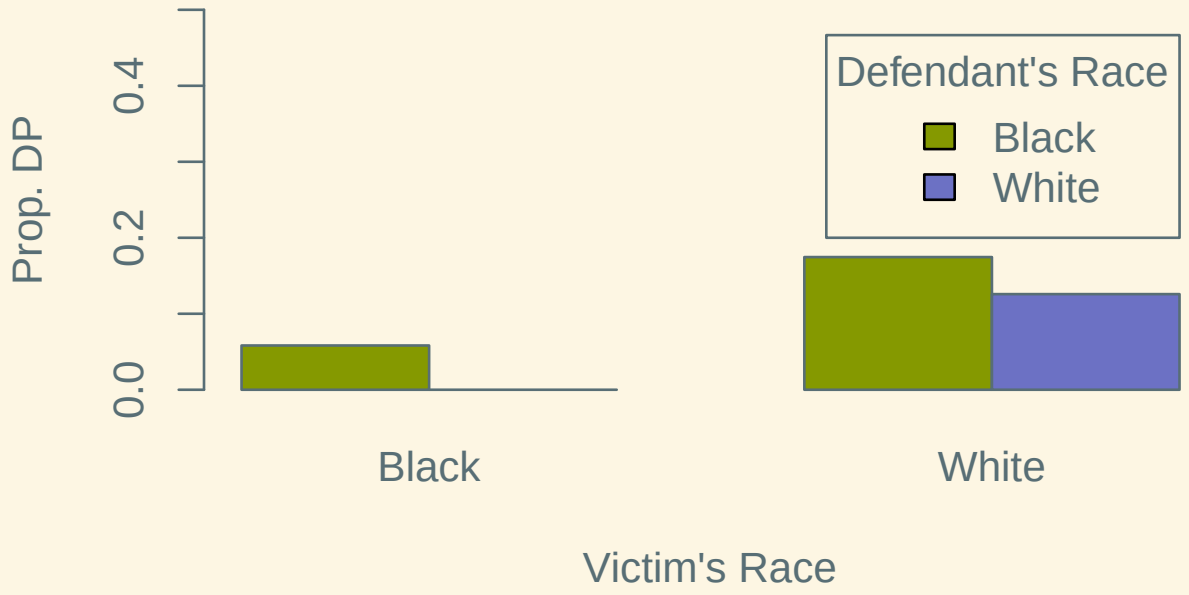


Figure: Proportions of Death Sentences out of Total Convictions (Florida, 1981), Grouped by both Victim's and Defendant's Race. Source: Agresti (2002)

Statistics: An Alternative to "Alternative Facts"

"You're saying it's a falsehood. ... Sean Spicer, our press secretary, gave alternative facts to that." – Kellyanne Conway, White House counselor, on *Meet the Press*, 1/22/2017

"You know, the very powerful and the very stupid have one thing in common," the Doctor said. "They don't alter their views to fit the facts. They alter the facts to fit their views." – Doctor Who, *The Face of Evil, Part 4, aired 1977*

"In God we trust, all others must bring data." – attributed to statistician W. Edwards Deming, 1900-1993.

Statistics: A Highly Employable Skill

"[A] new analysis of help-wanted postings for entry-level jobs suggests that [liberal arts] graduates can improve their job prospects markedly by acquiring a small level of proficiency in one of eight specific skill sets, such as social media or data analysis. For example, the analysis found an additional 137,000 entry-level jobs for liberal-arts graduates who had data-analysis or management skills. It also found that such data-analysis jobs paid an average of $12,700 above the average salary for jobs traditionally open to liberal-arts graduates without such skills." – Chronicle of Higher Ed., 6/9/16

Defensive Statistical Literacy

"There are three kinds of lies: lies, damned lies, and statistics." – Unknown (questionably attributed by Mark Twain to Benjamin D'Israeli)

- An overarching course goal: become a literate *consumer* of statistics

On the Web

- Course Website: `http://colindawson.net/stat113`
- Syllabus, slides, handouts, homework, labs, demos available there
- Exception: HW/Lab Solutions (on Blackboard)
- Also on Blackboard: electronic submission of assignments

Course Outline

- Data Collection, Structure of Data ($\sim$1.5 weeks)
- Data Description and Visualization ($\sim$2.5 weeks)
- Statistical Inference: Conceptual Foundations ($\sim$3.5 weeks)
- Specific Methods for Data Analysis ($\sim$5.5 weeks)
- Computational Skills for Data Analysis (throughout)

Graded Components

- Weekly(ish) Quizzes
- Homework problems
- Lab assignments
- 3 In-Class Exams
- Group Final Project

"Specifications"-Based Grading System

- Each HW/Quiz/Exam question is associated with one or more "Specific Learning Objective" (SLO).
- Each item/SLO graded "M" for "Mastery" "M-" for near-Mastery, "P" for progress (or N for "not assessable")
- The grade for each SLO is the average of the top few item-grades for that SLO (see handouts for details)
- Reassessments for "concepts" SLOs are available by taking a "reassessment quiz" in office hours (max. 2 SLOs per week), but must have made a good faith attempt on related homework first.
- Project grade comes from proposal, "pilot analysis", presentation, and final paper

A Note on Software

- We will use R (the "engine") via RStudio (the "control panel")
- Two options: Access via a log-in on your browser (`rstudio.oberlin.edu`), or install on your own computer (see below)
- Browser version a bit less smooth at times, getting data and work in and out is a bit clunkier at times, but less you need to manage
- If you want to use your own computer in lab, please install both R/RStudio on your computer before you get there.

R: http://www.r-project.org
RStudio: http://www.rstudio.com

HW 0(a) Everyone enrolled should come to my office hours sometime in the first two weeks, just for an intro. Book a 10-minute slot via my Google calendar (link on the course website)

HW 0(b) Please fill out the background survey here:
`https://goo.gl/forms/56Im3gGuPdYxnzLv1`
before next class

First graded HW due electronically

- Due Tues. 9/5: Lab 1 (we will start this in lab on Friday)
- Due Tues. 9/12: Textbook Problems (see website)