Dynamic Data in the Statistics Classroom

Johanna Hardin

1. Introduction

There has been a recent push to change the way we - as statisticians - engage pedagogically with complex real-world data analysis problems. Two parallel forces have directed us toward embracing more complex-data real-world problems in the classroom.

The first driver is a call from educators to make statistics more relevant to students' experiences. In order to address the perspective of students who previously have been reported to "exhibit remarkably little curiosity about the material they are analyzing", Brown and Kass (2009) suggest the importance of "real-world problem solving" to get students engaged in their analyses. Gould (2010) wants introductory statistics students to leave the course with "a set of ... attitudes about data that are immediately applicable to their lives." Gould and Çetinkaya-Rundel (2013) suggest putting "data at the center of the curriculum." Horton, Baumer, and Wickham (2015) assert that "statistics students need to develop the capacity to make sense of the staggering amount of information collected in our increasingly data-centered world." And Zhu, Hernandez, Mueller, Dong, and Forman (2013) remind us that "data pre-processing bridges the gap from data acquisition to statistical analysis but has not been championed as a relevant component in statistics curricula."

The second driver arises from students and other stakeholders and is harder to document with references. Certainly my own experience (and that of my colleagues) is that students engage at the deepest level when they (a) care about the problem, and (b) understand the data collection, study design, or motivation of the problem. Kuiper and Sturdivant (2015) report, "Our student evaluations of these materials support Gould's (2010) comments suggesting that previously collected and cleaned data were considered abstract to the student ... We have found that unless students 1) have collected the data themselves or 2) clearly see where and how the data were collected, they often fail to appreciate it."

Indeed, even if the work is done for them, as Grimshaw (2015) argues, there is much benefit in the students seeing how the data were procured:

> Using the vocabulary of Wickham (2014), teachers hide the 'messy data' aspects and provide 'tidy data' - even when students possess the data skills required to work with the messy data. It is valuable for students to not only have many authentic data experiences but also to have the professor model the correct application of statistics by showing work with messy data in lectures...What is good for statistics majors can also be applied to introductory courses. There may be no data skills on the learning outcomes for these courses, but some examples and homework in an introductory course may be modified and/or updated to use the original source data instead of a curated dataset. The data skills required would certainly be modest and need to fit student backgrounds. The objective would be for students to see that data skills are

required in an analysis. Students may rely on code provided to them that results in their own copy of the dataset.

There have been repeated calls to use *real data* in the classroom (Cobb 1991, 1992, 2007, 2011; Workgroup on Undergraduate Statistics 2000; American Statistical Association Undergraduate Guidelines Workgroup 2014; GAISE College Group 2005; Carver, Everson, Gabrosek, Rowell, Horton, Lock, Mocko, Rossman, Velleman, Witmer, and Wood 2016). There are benefits to using real data sets (as opposed to made up numbers) in textbooks and in the classroom. Indeed, there are virtually no textbooks of any kind (AP Statistics, Introductory Statistics, second course in statistics) being written today without the vast majority of examples taken from actual studies or databases. Additionally, there has been a push to infuse R with datasets that are relevant, recent, and sophisticated (e.g., nycflights13 and other mosaic datasets, (Pruim, Kaplan, and Horton 2014)).

Unfortunately, by nature, however, all of the data given in a textbook or an R package is static. Currently, there do not exist mechanisms to continually update any dataset provided by the course materials. The most comprehensive baseball dataset compiled and provided to the students will be out of date (and less interesting) by next fall. The good news, however, is that, outside of the statistics classroom, data of all kinds are being updated in real time publicly and accessibly. And even better news is that R developers are continually improving interfaces to the vast amounts of public data. For example, the `tidyverse` package imports a handful of packages that make downloading data straightforward (e.g., see Wickham (2014) and more recently the packages: `rvest`, `readr`, `readxl`, `haven`, `httr`, and `xml2`; expect more to come).

Grimshaw (2015) delineates data along two axes: the first axis describes the source and format for the data; the second axis describes the amount of wrangling required on the data. Both axes are scored as good/better/best. Indeed, Grimshaw gives examples, sources, and R code for several examples. However, he included only one dataset from the "best/best" category that worked well in his classroom. Out of the three "best/best" datasets he used, two were not well received because of excessive time, computing expertise, or domain knowledge needed to wrangle the data into usable format and extract meaning from the dataset.

In this manuscript, we seek to address the difficulty of using dynamic data in the classroom by curating additional best/best datasets (including full R Markdown files to facilitate reproducible analysis). The available resources aim to open the world of dynamic data to those who have not previously worked directly with downloaded data. We hope our work can act as a starting point for those interested in building their data scraping skills. Additionally, the R Markdown files can be used as scaffolding for assignments designed to help students engage with online resources.

In the next section we work through an entire example including showing the steps and R code for downloading the data, providing example code for using it in a typical introductory statistics classroom, and suggesting ideas for expanded use beyond introductory statistics. Section 3 gives summaries of additional dynamic datasets that are fully curated as R Markdown files in the supplementary resources on Github (https://github.com/hardin47/DynamicData) as well as a list of other places to look for dynamic data, and we provide closing thoughts and ideas for future work in the Conclusion. The Appendix describes a few technical details that are helpful for making dynamic data accessible to students at all levels.

2. Complete Example: College Scorecard Data

Supplementary materials provide full R Markdown files for complete analysis on five different dynamic datasets. We start with a complete description of the College Scorecard materials as a way of illustrating the available resources.

Data on characteristics of US institutions of higher education were collected in an effort to make more transparent issues of cost, debt, completion rates, and post-graduation earning potential. An undertaking of the U.S. Department of Education, the College Scorecard data represent a compilation of institutional reporting, federal financial aid reports, and tax information. The process of gathering and compiling the data is well documented on the College Scorecard website https://collegescorecard.ed.gov/data/documentation/. One caveat is that some of the variables have only been collected on students receiving federal financial aid. Biases inherent to analyses done on data collected from a subgroup should be considered.

The College Scorecard dataset is incredibly rich. The individual institutions are broken down by total share of enrollment of various races, family income levels, first generation status, age of students, etc. Additionally, the data give matriculation information like SAT scores as well as graduation information like completion rate and income level. The dataset allows for a student to investigate political or personal hypotheses about college education and the costs and benefits within. The variables are described in a data dictionary given at https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary.xlsx.

2.1. Downloading the Data

For each of the five fully curated dynamic datasets, there is an R Markdown file (available at https://github.com/hardin47/DynamicData) which scrapes the data from an outside web source (presumably kept current and public by some other organization). The downloading step is shown here in the manuscript only for the College Scorecard data. Note, as discussed above, the data are downloaded directly from the website managed by the US Department of Education (and not stored locally as a csv file). It is worth pointing out that the code given here is more complicated than what is standard in an introductory statistics course (especially at the beginning of the semester!). However, both the GapMinder and Wikipedia examples can be adjusted in very simple ways, allowing each student to work with their own dataset. That is, you can scaffold an assignment by providing the majority of the downloading code and having the student fill in the URL (Wikipedia example) or variable names (GapMinder example). The R code for downloading the data for all of the examples are given in the supplementary R Markdown files, and each uses slightly different functions and syntax.

First, load the data into R.

```
college_url <-
  "https://s3.amazonaws.com/ed-college-choice-public/Most+Recent+Cohorts+(All+Data+Elements).csv"
```

```
college_data <- readr::read_csv(college_url)
dim(college_data)
```

```
## [1] 7804 1728
```

Next, use data wrangling methods to clean and organize the variables.

```
college_debt = college_data %>%
  dplyr::select(region, HBCU, DEBT_MDN, md_earn_wne_p10) %>%
  mutate(DEBT_MDN = readr::parse_number(DEBT_MDN),
         md_earn_wne_p10 = readr::parse_number(md_earn_wne_p10)) %>%
  mutate(HBCU = ifelse(HBCU=="NULL", NA, HBCU)) %>%
  mutate(region2 = ifelse(region=="0", "Military",
                   ifelse(region=="1", "New England",
                   ifelse(region=="2", "Mid East",
                   ifelse(region=="3", "Great Lakes",
                   ifelse(region=="4", "Plains",
                   ifelse(region=="5", "Southeast",
                   ifelse(region=="6", "Southwest",
                   ifelse(region=="7", "Rocky Mnts",
                   ifelse(region=="8", "Far West", "Outlying"))))))))))

summary(college_debt)


##      region              HBCU              DEBT_MDN        md_earn_wne_p10
##  Min.   :0.000   Length:7804        Min.   :   333   Min.   :  8400
##  1st Qu.:3.000   Class :character   1st Qu.:  7710   1st Qu.: 24200
##  Median :5.000   Mode  :character   Median :  9833   Median : 31200
##  Mean   :4.621                      Mean   : 11830   Mean   : 33233
##  3rd Qu.:6.000                      3rd Qu.: 15462   3rd Qu.: 39200
##  Max.   :9.000                      Max.   :131335   Max.   :250000
##                                     NA's   :1163     NA's   :2168
##    region2
##  Length:7804
##  Class :character
##  Mode  :character
##
##
##
##
```

2.2. Using dynamic data within a typical classroom

Using the downloaded data, we start by applying a technique from the introductory curriculum to a research question of interest based on the College Scorecard data. College debt is of particular interest to many college students, but debt can be mediated by post-graduation income. To fully investigate the relationship between the variables, we provide both confidence and prediction intervals for both variables.

After calculating a few individual intervals, we show all intervals represented graphically and broken down by geographic region. Note that the visual representations do not represent a simple summary plot of the data, and we leave it open to the instructor to have the students

engage more deeply with the many available variables.

Using the two variables measuring amount of debt of a typical (i.e., median) college graduate and median earning 10 years after matriculation, we create both confidence intervals and prediction intervals – keeping in mind that the observational unit is an academic institution. Note that the calculations below are for both confidence and prediction intervals. The confidence interval agglomerates institutions over the entire dataset; however, the prediction value is for a single *institution* (which is the observational unit). The analysis lends itself nicely to a conversation about confidence vs. prediction intervals as well as observational units as institution vs. as individual student. It is worth pointing out to the students that the prediction intervals likely hold more information related to their individual experiences than the confidence intervals. However, the unit of prediction is for an *institution*, and so the individual student debt and income is likely even more variable than shown here. Additionally, Figure 1 demonstrates the effect of samples size: consider the comparison of the Military intervals (one school) to the intervals for all of the US institutions (about 6000 schools). (Note: the intervals given in Figure 1 were created using an ANOVA model where the within variance is calculated across all regions, which is how the interval for military schools can be calculated. You may or may not want to bring that up with your students.)

The following R code uses the `mosaic` package to directly calculate both prediction and confidence intervals. Note the formula interface given by the tilde is described in detail here: http://rpruim.github.io/eCOTS2014/Workshop/Modeling.html.

```
debt_mod <- lm(DEBT_MDN~1, data = college_debt)
debt_fun <- mosaic::makeFun(debt_mod)
debt_fun()


##        1
## 11829.78


debt_fun(interval="confidence")


##       fit      lwr      upr
## 1 11829.78 11692.44 11967.13


debt_fun(interval="prediction")


##       fit      lwr      upr
## 1 11829.78 636.2383 23023.33


earn_mod <- lm(md_earn_wne_p10~1, data = college_debt)
earn_fun <- mosaic::makeFun(earn_mod)
earn_fun()


##        1
## 33232.59
```

```
earn_fun(interval="confidence")
```

```
##        fit      lwr      upr
## 1 33232.59 32864.78 33600.41
```

```
earn_fun(interval="prediction")
```

```
##        fit      lwr      upr
## 1 33232.59 5616.893 60848.29
```

The intervals are interesting, but they might be even more interesting if broken down by region and shown visually. Note how much smaller the confidence intervals are from the prediction intervals! The difference indicates lots of variability across institutions and large sample sizes.
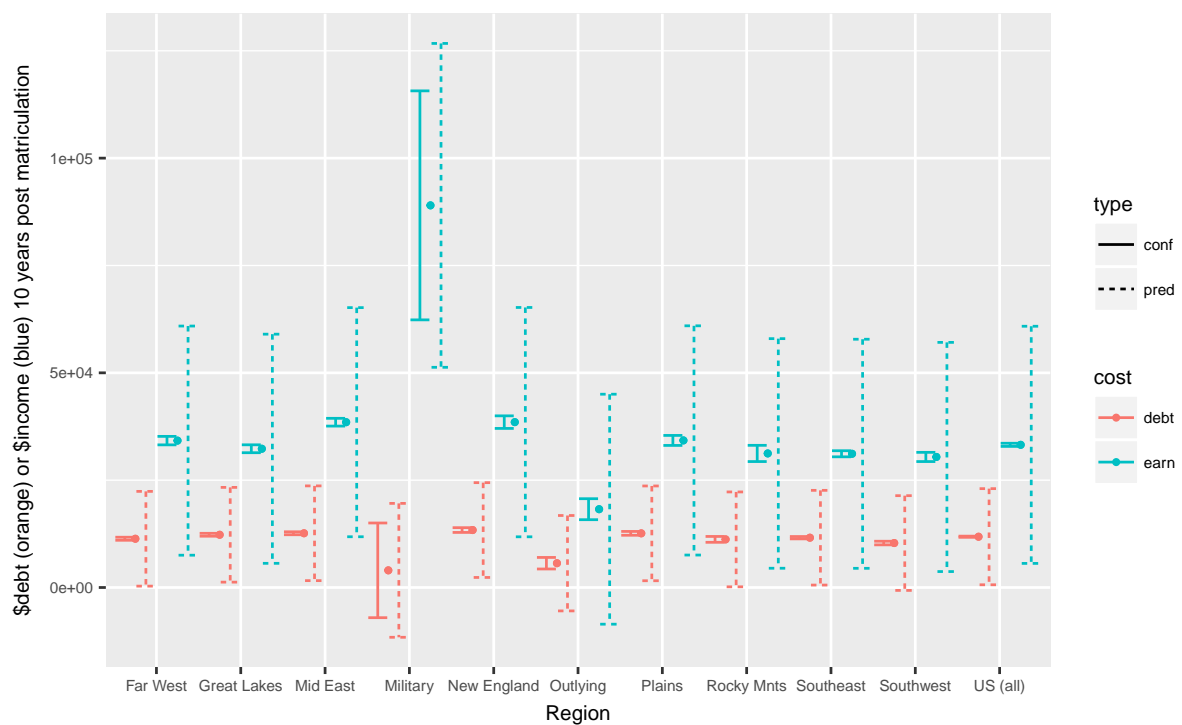

Figure 1. The x-axis represents the region of the institution. The y-axis represents either the amount of debt 10 years after matriculation (orange) or the amount of income 10 years after matriculation (blue). Confidence intervals for the average values (within region) are given by the solid lines. Prediction intervals for individual institutions are given by the dashed lines. The solid dot represents the center of both types of intervals (broken down by debt and income).

2.3. Thinking outside the box

Additionally, for each of the supplementary dynamic data R Markdown files, we add a section for

each analysis which is based on topics that are not traditionally taught in introductory statistics classes. The additional analysis is done not only to expand the tool box of the students but also to teach the students that they can often think about the problem in sophisticated ways even if all their tools come only from the introductory course.

The College Scorecard dataset is incredibly rich and can be used for many different types of model building: linear, logistic, machine learning. Indeed, thinking about interaction terms could be particularly insightful. Here, we give an example of regressing earnings on debt with the interaction term as whether or not the institution is one of the Historically Black Colleges and Universities (HBCU). Figure 2 displays the separate regression lines for the two distinct types of institutions.
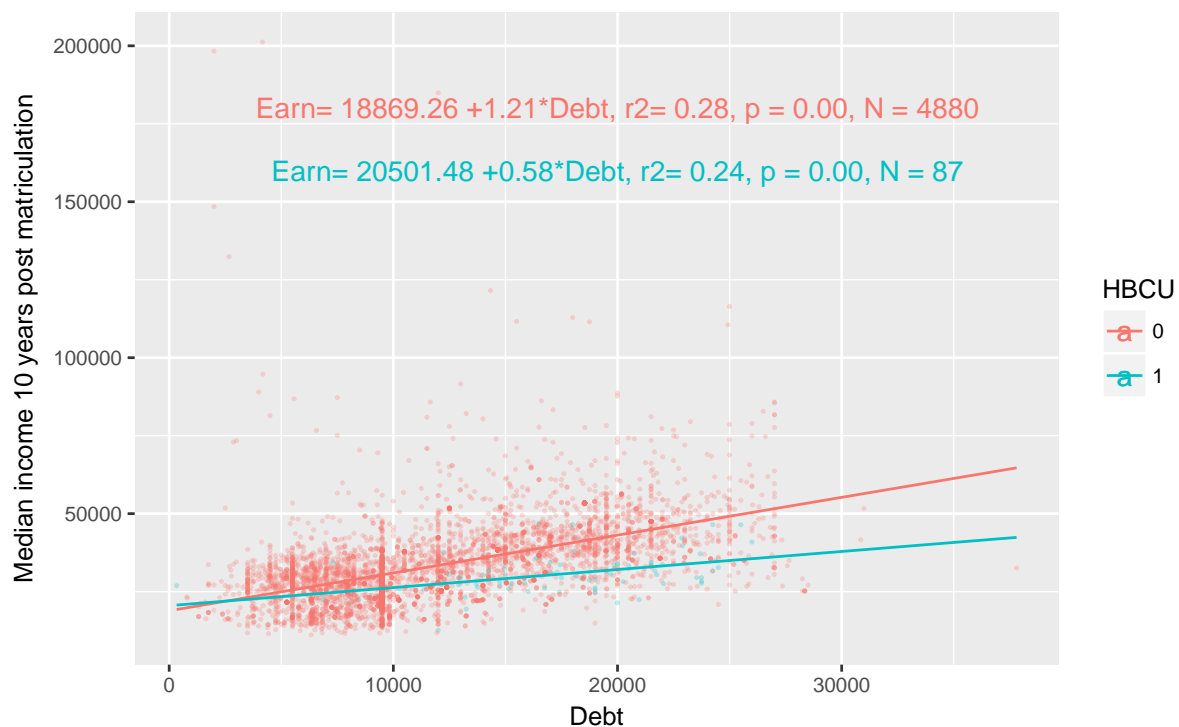

Figure 2. Median income regressed on debt. For the analysis, HBCU is interacted with debt to provide two distinct (and not parallel) regression lines. HBCU institutions are given in blue, and non-HBCU institutions are given in orange.

Many interesting conversations can ensue based on the regression of income on debt. Reminding the students that each observation is an institution is an important starting point. Additionally, students should be able to volunteer the dangers of using a model like this to suggest causality. Last, there might be room to discuss an inferential analysis of whether HBCUs are statistically different from non-HBCUs (noting the substantial differences in sample sizes).

It is not hard to come up with additional questions to investigate with the College Scorecard data. Indeed, because the data relate directly to college students, they should be able to find ways to engage with the data. We recommend continued conversations about how the data are valuable to the larger community, but that the information is not always complete (e.g., many variables are collected only on students who fill out financial aid forms) and not causative.

3. Dynamic Data Projects

For each of four additional dynamic datasets, we describe the source of the data and relevant variables & research questions, some standard and graphical techniques ("using dynamic data within a typical introductory statistics classroom"), and a statistical analysis appropriate for a course after introductory statistics ("thinking outside the box"). We also provide source information for an additional nine dynamic datasets.

3.1. Wikipedia Data

Wikipedia stores most of its tabular data in HTML tables. To scrape HTML tables from any website (or HTML file), use the R function *XML::readHTMLTable*. As an initial foray into downloading data directly from the internet into R, Wikipedia tables provide a nice introduction. In the supplementary R Markdown file associated with the Wikipedia data analysis, we also walk through some of the useful aspects of using `dplyr` to wrangle the data[1].
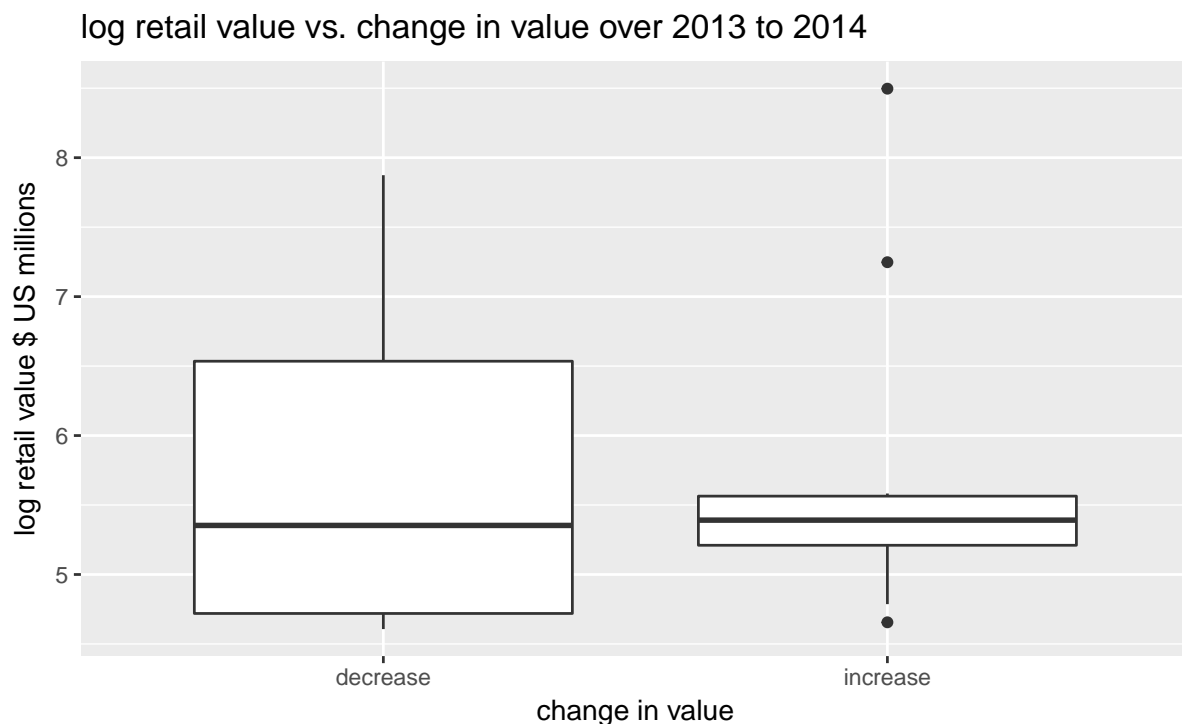

Figure 3. The log of the retail value of music sales broken down by whether or not the sales have increased or decreased (presumably over the previous year, although the Wikipedia documentation does not specify the time period over which the change is measured).

Using dynamic data within a typical classroom

The Wikipedia analysis given in the fully curated files explores an HTML table on sales of music (physical and digital) in 2014, `https://en.wikipedia.org/wiki/Music_industry`. One

---

[1]Original idea for this example provided by Nick Horton, Amherst College.

variable gives an indication of whether the retail value of the music sales has increased or decreased. Using the country-level music data, we perform a t-test, a Wilcoxon rank sum test, data transformations, and boxplots to investigate music retail sales (analysis given in supplementary materials, not shown here). By grouping the data into two categories we can investigate whether there is any statistical difference between the total average retail sales (in US\$) between those countries for whom retail sales increased versus those that decreased. The p-value for the initial t-test is reasonably large, but the boxplot shows that the difference in variability across the two groups is also large with a sample that either has large outliers or a long skewed right tail. Because the technical assumptions do not appear to be met, a log transformation of the data or a non-parametric test might be better assessments of the data (see Figure 3). The analysis leads to conversations about the source of the data and the reasons why p-values are non-significant. The example extends easily to each student choosing their own Wikipedia page and data table, graphical representations, and statistical analyses.

Thinking outside the box

Among the variables in the Wikipedia music dataset are the breakdown (percentages) of how the retail sales are distributed across physical, digital, performance rights, and synchronization. We might want to see whether there is a dependency of total retail sales on the breakdown of types of products. The problem is not well suited to introductory statistics as there is not an obvious statistic we can use within a sampling distribution (to create a p-value, etc.). Because there does not seem to be an obvious mechanism for evaluating the breakdown of products (and how "different" they are), we consider an ad-hoc measure and perform a permutation test to assess significance. The average breakdown of retail sales is given in the R Markdown file in the supplementary materials. One way to measure a discrepancy between the retail sales and the consistency of product breakdown is to correlate the retail sales with the sum of squared distances from the average breakdown of product types. We see that the metric we created to find a relationship between retail sales and breakdown of types of product does not show significance. A student project could be to think about different ways to measure how the breakdowns can be considered to be different.

3.2. NHANES Data

The NHANES data come from the National Health and Nutrition Examination Survey, surveys given nationwide by the Center for Disease Controls (CDC). The CDC adopted the following sampling procedure:

1. Selection of primary sampling units (PSUs), which are counties or small groups of contiguous counties.

2. Selection of segments within PSUs that constitute a block or group of blocks containing a cluster of households.

3. Selection of specific households within segments.

4. Selection of individuals within a household

About 12,000 persons per 2-year cycle were asked to participate in NHANES. Response rates

varied by year, but an average of 10,500 persons out of the initial 12,000 agreed to complete a household interview. Of these, about 10,000 then participated in data collection at the mobile exam center. The persons (observational units) are located in counties across the country. About 30 selected counties were visited during a 2-year survey cycle out of approximately 3,000 counties in the United States. Each of the four regions of the United States and metropolitan and non-metropolitan areas are represented each year. As such, the data collection is ongoing, and the data are updated on the NHANES website periodically, `http://www.cdc.gov/nchs/nhanes.htm`. This manuscript uses the 2011-2012 NHANES data, but we expect the data to be updated regularly, and the URL should simply change to 2013-2014 when it becomes available.

Note that the NHANES data are available in the `mosaic` package (Pruim *et al.* 2014), but the `mosaic` version of the NHANES data is static (from 2011-2012), and the data has been cleaned with pre-selected variables. Additionally, the variables can be downloaded directly using the `nhanesA` package in R. `https://cran.r-project.org/web/packages/nhanesA/vignettes/Introducing_nhanesA.html`. By accessing the data directly from the CDC's website, students become more involved in the data analysis process, understanding what they can and cannot get from the data.

The variables in the CDC's online NHANES dataset are virtually limitless. We use a few different datasets, merging them based on an individual identifier in the dataset. The variable information is all given online, but each at a different webpage. For example, the demographic data is at `http://wwwn.cdc.gov/nchs/nhanes/search/variablelist.aspx?Component=Demographics&CycleBeginYear=2011`.

A further important aspect to the example is that (as described above) the data do not constitute a simple random sample (it is a weighted sample) from a population. The sampling scheme can be part of the statistical inquiry into the data analysis, or it can be set by the instructor in the template used to download the data. The data which is directly downloaded from the CDC website includes variables on the weighting scheme. In the R Markdown file provided with this manuscript, we demonstrate how to create a dataset which *can* act as a simple random sample from the population.[2] The figure and analysis below are done with a proxy simple random sample.

Using dynamic data within a typical classroom

In the supplementary materials (not shown here), we start with a comparison of body mass index (BMI) for those in committed relationships and those not in committed relationships. The graphs of the two BMI distributions look quite similar, and the t-test shows a non-significant difference in means. The results prompt a discussion about averages versus individual results, causation, and sample size.

Thinking outside the box

Because of the large sample size and the ability to determine the functional form of a non-linear relationship, smoothing techniques can be used to model quantitative variables. Adding

---

[2]The weighting analysis was motivated by work done by Shonda Kuiper (`http://web.grinnell.edu/individuals/kuipers/stat2labs/weights.html`) as well as the Project Mosaic Team (`https://cran.r-project.org/web/packages/NHANES/NHANES.pdf`).

a smooth curve to a standard scatterplot leads (see Figure 4) to discussions about how smooth curves are estimated, SE of the smooth curve, extra variability and instability due to extremes and fewer data points on the ends. However, extrapolation (note that the two curves have different ranges?) and the slopes of the two curves not seeming to be different (no interaction?) might warrant further study.
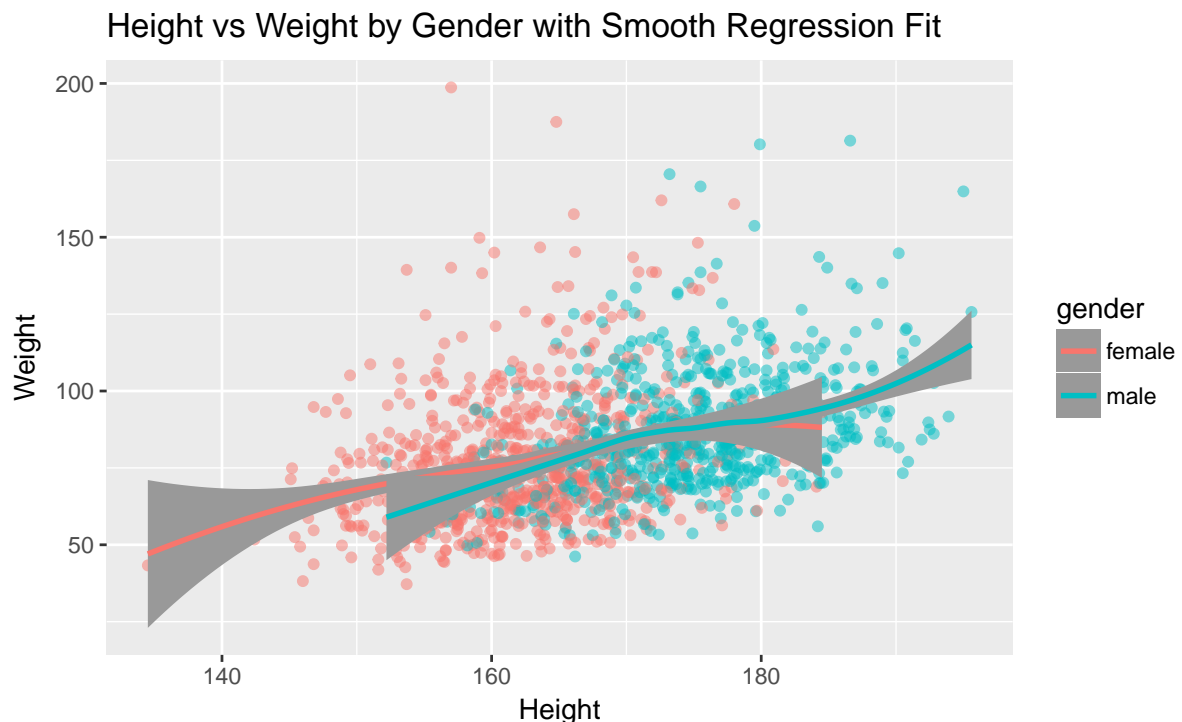

Figure 4. A scatterplot of weight and height broken down by gender. A smooth regression is fit to the points, and the confidence interval for the smooth curve as well as the increase in variability for small and large values of height are part of an important discussion in a second level regression course.

3.3. Global Literacy Rate Data

The next dynamic dataset comes from GapMinder http://www.gapminder.org/. GapMinder also has a plethora of variables from which students can choose (according to their own interests), but here we work with literacy rates measured at the country level. The analysis (and more importantly, the data scraping from GapMinder) could easily be extended for students interested in all sorts of political, social, environmental, or demographic data available. Understanding political and demographic trends across both time and location can provide very interesting insight into economic or political science questions. Alternatively the GapMinder data can be perused in a descriptive or graphical manner.

Using dynamic data within a typical classroom

The introductory analysis considers gender differences in literacy rates and uses a linear model on the difference between female and male literacy rates across time (the analysis is available in the

supplementary materials and not shown here). We show a graphical representation and discuss model assumptions including sampling and independence of residuals. The model indicates that the difference between male and female literacy rates is shrinking over time. However, we worry about the effects of other variables and encourage a more complete analysis. Indeed, there may be large biases in our model if important explanatory variables have been left out.

The data provided are ideal for a fantastic classroom conversation about causation, causal mechanisms, and confounding variables.

Thinking outside the box

In the second analysis, we work with the additional variable `continent`. The trends observed in the first analysis hold up in the second analysis (i.e., the difference declines over time in each of the continents). However, there are additional considerations to be made, for example, the differences between the slopes across the continents (the analysis is available in the supplementary materials and not shown here). We suggest additional explorations into the independence of the residuals and more advanced spatio-temporal patterns of literacy.

3.4. Weather Data from NOAA Buoys

The National Oceanic and Atmospheric Administration (NOAA) is the American federal agency in charge of collecting information and making decisions related to the oceans and the atmosphere. Throughout North America, they supply weather stations which are located both along the coast as well as in the middle of the ocean (on buoys). Among other variables, the weather stations collect information on wind, humidity, temperature, visibility, and atmospheric pressure. The data is all publicly available on NOAA's website, http://www.ndbc.noaa.gov/.

Using dynamic data within a typical classroom

Although the data do not constitute a random sample, they are very likely to be quite representative with respect to the difference in wind and air temperature at that location over the year. In the supplementary files (not shown here), we used a paired analysis (i.e., subtract the two variables and treat them as a single variable) to find a confidence interval for the true difference in temperature between wind and air. Also, we find a prediction interval for the difference in temperatures across individual measurements.

Thinking outside the box

Although a full analysis of the data would warrant multiple years of data (so as to understand yearly trends), we can estimate the spectral density of the time series using a smoothed periodogram (the data below represent measurements every hour for all of 2014). In the smoothed periodogram (see Figure 5) the x-axis is the frequency (one over the period) and y-axis represents the correlation (normalized) between the cosine wave at that frequency and the time series. We can see that wind speed has strong correlation at period 12 hours and period 24 hours. A more sophisticated analysis or longer project could include collecting data from multiple buoys, extended years, and/or additional information on storms https:

Wind Speed, Smoothed Periodogram
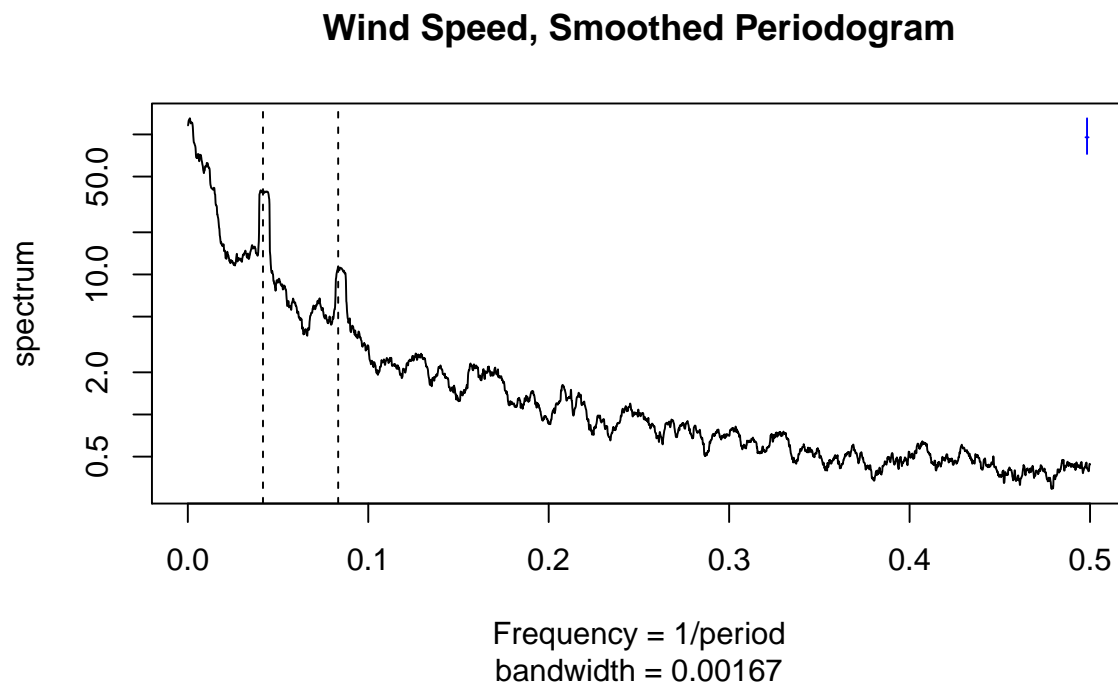

Frequency = 1/period
bandwidth = 0.00167

Figure 5. A smoothed periodogram of Wind Speed for buoy #46025 off the coast of Santa Monica. The dashed lines are at 1/12 and 1/24, representing 12-hour and 24-hour periodicity.

3.5. Other Dynamic Datasets

There are myriad sources of dynamic data which are public and accessible. We list a few additional resources here.

Baseball Data

Many students are interested in sports statistics. In particular, statisticians and sabermatricians have worked over the years to compile datasets and to think carefully about statistical methods applied to baseball data. Today's datasets can answer questions like: What is the probability of a particular event happening given certain real time situations?

Albert (2010) provides curated seasonal batting data from 1871 to 2009 at http://www.amstat.org/publications/jse/v18n3/mlb_batting.dat.txt. With the passage of time, the examples of Derek Jeter and Alex Rodriguez as current players and batting trends up to 2009 are less relevant to students interested in sports, and what was once a current research question has become stagnant and ossified. However, Albert (2010) provides sufficient detail about the collection of files on the Lahman Baseball Database (http://www.seanlahman.com/baseball-archive/statistics/ with data available as a .zip file which can be read directly into R http://seanlahman.com/files/database/lahman-csv_2015-01-24.zip) and

tasks required to filter and merge the master and batting files that the research questions and teaching notes can be updated to reflect current players and trends. Also see Jim Albert's blog *Exploring Baseball Data with R*, https://baseballwithr.wordpress.com/.

Cherry Blossom Ten Mile Run Data

Each year in April, the Cherry Blossom Ten Mile Run happens in Washington, DC (http://www.cherryblossom.org/theraces/tenmile.php). The race results (including finishing time, name, age, hometown, and pace) are available from 1999 to the present (although admittedly, the results after 2012 are more difficult to scrape off the web due to a change in the format of how the results are now posted). Kaplan and Nolan (2015) provide step-by-step instructions (http://rdatasciencecases.org/) for scraping the data and wrangling it into a format which can be used to investigate race results over time, by age, or by gender.

Fatal Accidents and US Census Data

The National Highway Traffic Safety Administration collects data on all fatalities suffered in motor vehicle traffic crashes. The data is posted yearly and publicly as part of the Fatal Accident Reporting System, http://www.nhtsa.gov/FARS. Each year is posted as a separate set of files, available in different formats. For example, download the SAS data (readable into R via the *read.sas7bdat* function in the sas7bdat package, the *sasxport.get* function in the Hmisc package, or *read.SAS* in the haven package) at ftp://ftp.nhtsa.dot.gov/fars/2014/SAS/.

Another dataset which is easily accessible is the census data (http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml). By looking at the *Decennial Census data* and the *Profile of General Population and Housing Characteristics*, students can study the ages and other demographic characteristics of the general population as compared to those individuals involved in fatal accidents[3].

Climate Data

Witt (2013) describes using data in class to reveal important insights on climate for statistics students. The data and analyses describe both decline of Arctic sea ice and global temperature increase. The (static) data used in the analyses is available through *The Journal of Statistics Education*, http://www.amstat.org/publications/jse/v21n1/witt.pdf, but the authors also provide links and information about the original data sources which are dynamic.

Notably, realclimate.org provides a catalogue of many different types of climate data and relevant source information (http://www.realclimate.org/index.php/data-sources/). Witt (2013) also reports:

> NASA's Global Change Master Directory is available at http://gcmd.gsfc.nasa.gov/. NOAA's National Climate Data Center maintains an extensive data directory available at http://www.ncdc.noaa.gov/oa/ncdc.html. Yet another good climate data source is the Data Guide maintained by the National Center for Atmospheric Research Climate at https://climatedataguide.ucar.edu/.

---

[3]Original idea for this example from Laura Kapitula, Grand Valley State University.

Iowa Liquor Sales Data

Iowa recently released an 800MB+ dataset containing all the weekly liquor sales from January 1, 2014 to the present (https://data.iowa.gov/Economy/Iowa-Liquor-Sales/m3tr-qhgy). The data seems to be updated monthly. Dan Nguyen has provided some scripting code and SQL analysis for downloading and wrangling the data (https://gist.github.com/dannguyen/18ed71d3451d147af414). Because the data are available in a csv format, they can be read into R directly using read.csv from the data.iowa.gov URL. The SQL code from the Nguyen's github site can be translated directly into R using the `dplyr` package. Note that the rectangular structure and comma delimited format of the data make it tempting to work with. The challenge for this particular dynamic dataset comes with its size, which can make it unwieldy on many computers.

Medicare Inpatient Charges Data

Medicare inpatient charge data is available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient2013.html which links directly to a zipped csv file. An analysis of the costs for Medicare Severity Diagnosis Related Group (MS-DRG) (i.e., the procedure) can be done over multiple different covariates.

World Bank Data

The R package `wbstats` connects directly to the World Bank API, providing a trove of global economic data. The package documentation and vignette provide an easy start for students to download basic data of their choosing, https://cran.r-project.org/web/packages/wbstats/vignettes/Using_the_wbstats_package.html.

Energy Consumption Data

The US Department of Energy collates energy usage at http://www.eia.gov/. For example, at http://www.eia.gov/coal/data.cfm they provide complete data on "coal production, consumption, exports, imports, stocks, mining, and prices."

Government Survey Data

Anthony Damico has compiled and documented a repository of data from many dozens of different government surveys http://www.asdfree.com/, including NHANES. Additionally, Damico espouses the merits of reproducible analysis done in R with GitHub. As one example, the General Social Survey (GSS) contains information on what Americans think of policies, issues, and priorities in the US (http://gss.norc.org/).

3.6. Using the Examples

We try to communicate to the students that there is information in most data sources. We want

to be wary and attentive to issues of experimental design and systematic biases. However, we do not want to leave our students feeling stuck every time they encounter a dataset which has not been gathered from a large randomized trial. Instead, we try to think of the pieces of the analysis that can elicit information which is interesting or possibly hypothesis generating.

Students at every level are ready to examine complicated relationships using tools which are accessible before taking multiple advanced statistics courses. "Unfortunately, many instructors teach the sections on data analysis as descriptive statistics, perhaps because this is what they experienced in their first course. They emphasize the process of calculating numerical summaries and making graphical displays, rather than using these as tools to explore what the data are saying." (Notz 2015)

To keep our students from getting stuck, "we must ask questions such as whether the data allow generalization to a larger population, whether their structure can be meaningfully described with the models we wish to fit, and whether important subgroups or individuals were excluded from the data. Exceptions, anomalies, outliers, and subgroups are best recognized and understood in the context of the question being addressed." (De Veaux and Velleman 2015)

4.   Conclusion

The few examples provided here give a glimpse into how to incorporate real and dynamic data into introductory statistics and and courses beyond. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report implores instructors to make one small change in our courses to keep up with changing technology and data (Carver *et al.* 2016). In this manuscript we have provided tools for helping educators seamlessly incorporate dynamic data into a standard and modern introductory (or beyond) statistics course. We also hope that statistics students will see the different ways of accessing data and feel empowered to gather more information on their own. Our courses must remain relevant to both the student experience and the research community or we run the risk of become irrelevant. Small steps will get us where we need to be.

5. Acknowledgements

6. Appendix

The Appendix contains technical details of interest to students who want to learn more about directly downloading data. R Studio allows for the teacher to provide template R Markdown files which download data directly from the Internet into the R Studio environment. Additionally, reproducible R Markdown files teach good science and analysis. Horton *et al.* (2015) discuss the merits of reproducible research as well as the ease of using R Markdown in the classroom.

As part of the supplementary materials (https://github.com/hardin47/DynamicData), we have provided R Markdown files giving both R code and related pedagogical commentary associated with different dynamic data examples for use in a statistics classroom. The data have all been collected directly from outside sources - websites that are updated periodically (and with information the students can access on their own).

6.1. What is an API?

An Application Programming Interface (API) is a set of programming instructions (written in code, giving the appropriate algorithm) for downloading data from a website that contains - typically - vast amounts of data. For example, Twitter has an API (https://dev.twitter.com/overview/api) which tells programmers how to access tweets (and related information) directly from the Twitter website. At any level, but particularly useful at an introductory level, it is recommended to use an R (or other software) package or function that allows R to speak to the API in order to download the data. For example, the R package `wbstats` connects directly from R to the World Bank API and `twitteR` provides an R interface to the Twitter API. Though the examples provided do *not* generally rely on APIs or a related R interface, it is good to be aware of APIs (and to tell your students!) in order to greatly increase the sources of data available to you and your students.

6.2. R Markdown

The R Markdown files provided as supplementary materials make it easy for instructors and students to perform reproducible analyses on data collection through analysis and synthesis of the results. The students get the important practice of tracking each and every step of their work. Reproducible analysis has recently gotten a lot of press (Johnson 2014; Stodden, Leisch, and Peng 2014), and R Studio (RStudio Team 2015) has produced a user friendly format for combining R code with html (or LATEX) word processing that has been used with introductory statistics (Baumer, Çetinkaya-Rundel, Bray, Loi, and Horton 2014).

R Markdown has a short learning curve and will be straightforward for your students to implement. One of the key considerations is that when running an R Markdown file, the file does not pay attention to anything running locally. R Markdown essentially restarts R, so the current state of your R session is totally irrelevant. Recall that the purpose of R Markdown is to create reproducible files, so the Markdown file should run on any computer anywhere (regardless of the local environment).

Important note: Any package used in R Markdown needs to be installed prior to compiling the .Rmd file. That is, for a line of code such as `library(dplyr)`, before running the Markdown file, `install.packages("dplyr")` must be run from *within the console* one time (ever).

6.3. Useful R functions

Some important R *functions* (in *italics*) and `packages` (in `typeface`) that will help you and your students navigate importing and using data in R include:

- `dplyr` package for data wrangling in general; cheat sheet at https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

- The *glimpse* function in `dplyr`

- `tidyr` for converting between wide and long formats and for the very useful *extract_numeric()* function (or *readr::parse_numeric()*)

- `ggplot2` for faceted graphing (also, `ggvis`) (Wickham and Sievert 2016); cheat sheet at https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

- `openintro` (or packages that come with the textbooks you use) which are great for pulling up any dataset from the text and building on it in class (Diez, Barr, and Çetinkaya-Rundel 2012)

- `mosaic` for consistent syntax and helpful functions used in introductory statistics (Pruim *et al.* 2014); cheat sheet at https://cran.r-project.org/web/packages/mosaic/vignettes/MinimalR.pdf

- `googlesheets` for loading data directly from Google spreadsheets

- `lubridate` if you ever need to work with any date fields

- `stringr` for text parsing and manipulation

- `rvest` for scraping data off the web; `readxl` for reading excel data

- `readr` (and `fread` with *data.table*) for loading large datasets with default *stringsAsFactors = FALSE*

- `tables` for nice looking summary tables.

References

Albert J (2010). "Baseball Data at Season, Play-by-Play, and Pitch-by-Pitch Levels." *Journal of Statistics Education*, **18**(3).

American Statistical Association Undergraduate Guidelines Workgroup (2014). "2014 curriculum guidelines for undergraduate programs in statistical science." *Technical report*, American Statistical Association, Alexandria, VA. URL http://www.amstat.org/education/curriculumguidelines.cfm.

Baumer B, Çetinkaya-Rundel M, Bray A, Loi L, Horton N (2014). "R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics." *Technology Innovations in Statistics Education*, **8**(1). URL http://escholarship.org/uc/item/90b2f5xh.

Brown EN, Kass RE (2009). "What is Statistics?" *The American Statistician*, **63**(2), 105–110.

Carver R, Everson M, Gabrosek J, Rowell GH, Horton N, Lock R, Mocko M, Rossman A, Velleman P, Witmer J, Wood B (2016). "Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report." *Technical report*, American Statistical Association. URL http://www.amstat.org/education/gaise/collegeupdate/GAISE2016_DRAFT.pdf.

Cobb GW (1991). "Teaching Statistics: More Data, Less Lecturing." *UME Trends*, pp. 3–7.

Cobb GW (1992). "Teaching Statistics." *In Lynn A. Steen (ed), Heeding the call for change: suggestions for curricular action (MAA Notes No. 22)*, pp. 3–43.

Cobb GW (2007). "The Introductory Statistics Course: A Ptolemaic Curriculum?" *Technology Innovations in Statistics Education*, **1**(1). URL https://escholarship.org/uc/item/6hb3k0nz.

Cobb GW (2011). "Teaching statistics: some important tensions." *Chilean Journal of Statistics*, **2**(1), 31–62.

De Veaux R, Velleman P (2015). "Teaching Statistics Algorithmically or Stochastically Misses the Point: Why not Teach Holistically? (Online discussion of "Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum From the Ground Up," by George W Cobb, *The American Statistician*)." *The American Statistician*, **69**(4).

Diez DM, Barr CD, Çetinkaya-Rundel M (2012). *openintro: OpenIntro data sets and supplemental functions*. R package version 1.4, URL http://CRAN.R-project.org/package=openintro.

GAISE College Group (2005). "Guidelines for Assessment and Instruction in Statistics Education." *Technical report*, American Statistical Association. URL http://www.amstat.org/education/gaise.

Gould R (2010). "Statistics and the Modern Student." *International Statistical Review*, **78**(2), 297–315.

Gould R, Çetinkaya-Rundel M (2013). "Teaching Statistical Thinking in the Data Deluge." In T Wassong, D Frischemeier, PR Fischer, R Hochmuth, P Bender (eds.), *Using Tools for Learning Statistics and Mathematics*, pp. 377–391. Springer.

Grimshaw S (2015). "A Framework for Infusing Authentic Data Experiences Within Statistics Courses." *The American Statistician*, **69**(4), 307–314. URL http://arxiv.org/abs/1507.08934.

Horton NJ, Baumer B, Wickham H (2015). "Setting the stage for data science: integration of data management skills in introductory and second courses in statistics." *CHANCE*, **28**(2), 40–50. URL http://arxiv.org/abs/1502.00318.

Johnson G (2014). "New Truths That Only One Can See." URL http://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html.

Kaplan D, Nolan D (2015). "Modeling Runners' Times in the Cherry Blossom Race." In *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*, pp. 45–104. CRC Press.

Kuiper S, Sturdivant RX (2015). "Using Online Game-Based Simulations to Strengthen Students' Understanding of Practical Statistical Issues in Real-World Data Analysis." *The American Statistician*, **69**(4), 354–361.

Notz W (2015). "Vision or Bad Dream? (Online discussion of "Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum From the Ground Up," by George W Cobb, *The American Statistician*)." *The American Statistician*, **69**(4).

Pruim R, Kaplan D, Horton N (2014). *mosaic: Project MOSAIC (mosaic-web.org) Statistics and Mathematics Teaching Utilities*. R package version 0.9-1-3, URL http://CRAN.R-project.org/package=mosaic.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. URL http://www.rstudio.com/.

Stodden V, Leisch F, Peng RD (eds.) (2014). *Implementing Reproducible Research*. Chapman and Hall / CRC Press.

Wickham H (2014). "Tidy Data." *Journal of Statistical Software*, **59**(10). URL http://www.jstatsoft.org/v59/i10/.

Wickham H, Sievert C (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer New York. URL http://had.co.nz/ggplot2/book.

Witt G (2013). "Using Data from Climate Science to Teach Introductory Statistics." *Journal of Statistics Education*, **21**(1).

Workgroup on Undergraduate Statistics (2000). "Guidelines for Undergraduate Statistics Programs, http://www.amstat.org/education/curriculumguidelines.cfm, accessed August 18, 2013." *Technical report*, American Statistical Association.

Zhu Y, Hernandez LM, Mueller P, Dong Y, Forman MR (2013). "Data Acquisition and Preprocessing in Studies on Humans: What is Not Taught in Statistics Classes?" *The American Statistician*, **67**, 235–241.