NHANES

December 2017

Introduction

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, Kaplan, and Horton 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.[1] The figure and analysis below are done with a proxy simple random sample.

Data information & loading data

We download data on demographic information and body image. The data are in SAS format, but R has no trouble scraping the data from the NHANES website and uploading it into R.

[1] 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).

```
demourl <- "http://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/DEMO_G.XPT"
bodyurl <- "http://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/BMX_G.XPT"
NHANES.demo <- read_xpt(demourl)
NHANES.body <- read_xpt(bodyurl)
NHANES.demo$gender <- ifelse(NHANES.demo$RIAGENDR==1, "male", "female")

comb <- dplyr::inner_join(NHANES.body, NHANES.demo, by = "SEQN")
```

Additionally, the NHANES data were collected using a cluster sampling scheme, so it is important to use the variables which describe the weights on the sampling to create a sample which is reflective of the population. See the following for more information: http://web.grinnell.edu/individuals/kuipers/stat2labs/weights.html, ?NHANES (within R, using the NHANES packages), http://www.cdc.gov/nchs/data/series/sr_02/sr02_162.pdf.

```
set.seed(4747)
numobs = 2000
SRSsample <- sample(1:nrow(comb), numobs, replace=FALSE, prob=comb$WTMEC2YR/sum(comb$WTMEC2YR))
comb <- comb[SRSsample,]
adults = comb %>%
  dplyr::filter(RIDAGEYR >=18, BMXBMI>1) %>%
  dplyr::filter(DMDMARTL>0 & DMDMARTL < 10) %>%
  dplyr::mutate(rel=ifelse(DMDMARTL==6|DMDMARTL==1, "committed", "not")) %>%
  dplyr::mutate(bmi=BMXBMI)
```

Using dynamic data within a typical classroom

One research question of interest is whether people in a committed relationship have a higher BMI than those who are not. Note that a causal connection cannot be made here, but we are justified in thinking about the data as a good random sample from the US population. We filter only the adults out of the sample and also created a relationship variable as to whether or not the individual is in a committed relationship.

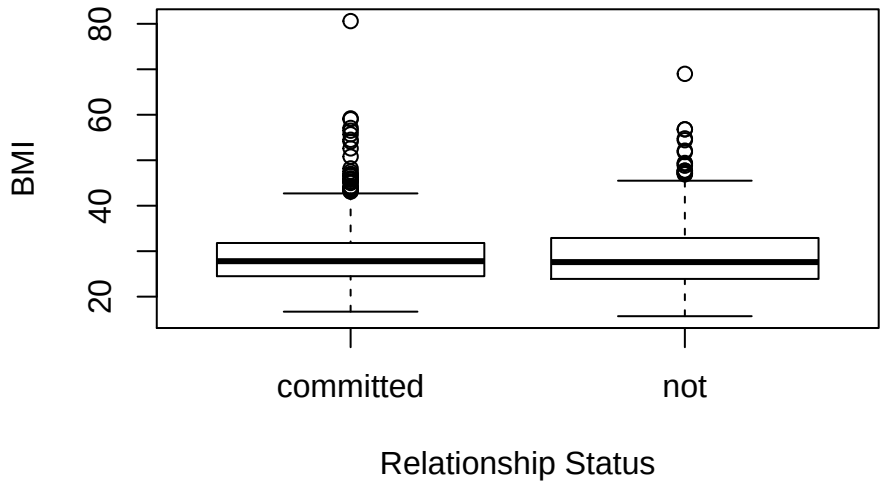
The boxplots both demonstrate that there is not a substantial difference between the BMI for those in committed relationships versus those who are not. The tests of significance validate the ideas from the descriptive statistics.

It is worth noting here that the sample size is quite large. If students repeat this analysis with different variables, it should be noted that very small effect sizes can be seen with large datasets. A small p-value might indicate that there are significant effects, but an extra interpretation as to whether the effect is a practical difference warrants consideration. It does not seem that the small *average* effect on BMI of being in a relationship is of any particular note when considering the large standard deviation across individual BMIs of both groups.

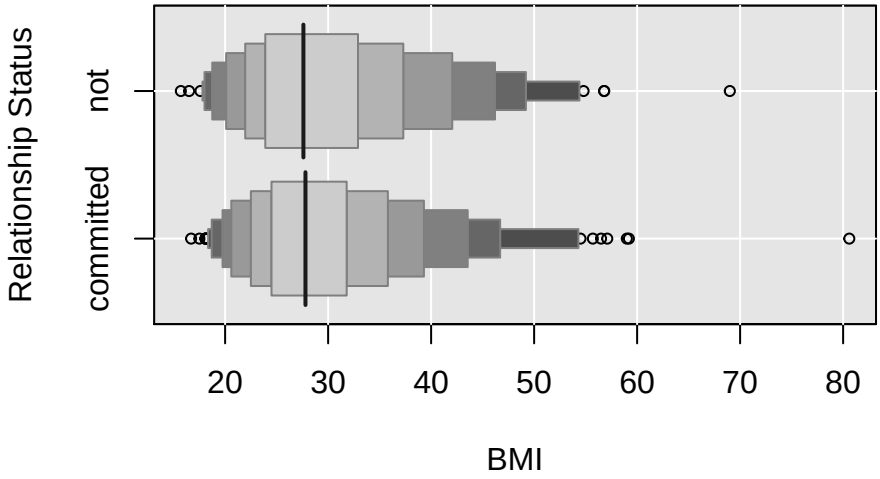
Additionally, again we point out that although these data are likely a good representation of the population, they cannot be used to find causal relationships. Indeed, even if BMI had been different on average across the two groups, we do not know if lower BMI causes one to be more likely in a committed relationship or whether a committed relationship leads to a lower BMI. Asking your students how one could gather such information would be a productive class discussion (e.g., paired observations, measurements over time, etc.).

```
adults = comb %>%
  filter(RIDAGEYR >=18, BMXBMI>1) %>%
  filter(DMDMARTL>0 & DMDMARTL < 10) %>%
  mutate(rel=ifelse(DMDMARTL==6|DMDMARTL==1, "committed", "not")) %>%
  mutate(bmi=BMXBMI)
```

```
boxplot(bmi ~ rel, data=adults, xlab="Relationship Status", ylab="BMI")
```




```
with(adults, LVboxplot(bmi ~ rel, xlab="BMI", ylab="Relationship Status"))
```




```
t.test(bmi ~ rel, data=adults)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by rel
## t = -0.59232, df = 1055.2, p-value = 0.5538
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9899362  0.5308616
## sample estimates:
## mean in group committed       mean in group not
##                28.84089                29.07043
```

```
dim(adults)
```

```
## [1] 1397   76
```

Thinking outside the box

The data we have downloaded has many variables, some of which have meanings that are not immediately obvious. The variable names are listed at the NHANES website, for example, the demographic data is at http://wwwn.cdc.gov/nchs/nhanes/search/variablelist.aspx?Component=Demographics&CycleBeginYear=2011.

```
names(adults)
```

```
##  [1] "SEQN"     "BMDSTATS" "BMXWT"    "BMIWT"    "BMXRECUM" "BMIRECUM"
##  [7] "BMXHEAD"  "BMIHEAD"  "BMXHT"    "BMIHT"    "BMXBMI"   "BMDBMIC"
## [13] "BMXLEG"   "BMILEG"   "BMXARML"  "BMIARML"  "BMXARMC"  "BMIARMC"
## [19] "BMXWAIST" "BMIWAIST" "BMXSAD1"  "BMXSAD2"  "BMXSAD3"  "BMXSAD4"
## [25] "BMDAVSAD" "BMDSADCM" "SDDSRVYR" "RIDSTATR" "RIAGENDR" "RIDAGEYR"
## [31] "RIDAGEMN" "RIDRETH1" "RIDRETH3" "RIDEXMON" "RIDEXAGY" "RIDEXAGM"
## [37] "DMQMILIZ" "DMQADFC"  "DMDBORN4" "DMDCITZN" "DMDYRSUS" "DMDEDUC3"
## [43] "DMDEDUC2" "DMDMARTL" "RIDEXPRG" "SIALANG"  "SIAPROXY" "SIAINTRP"
## [49] "FIALANG"  "FIAPROXY" "FIAINTRP" "MIALANG"  "MIAPROXY" "MIAINTRP"
## [55] "AIALANGA" "WTINT2YR" "WTMEC2YR" "SDMVPSU"  "SDMVSTRA" "INDHHIN2"
## [61] "INDFMIN2" "INDFMPIR" "DMDHHSIZ" "DMDFMSIZ" "DMDHHSZA" "DMDHHSZB"
## [67] "DMDHHSZE" "DMDHRGND" "DMDHRAGE" "DMDHRBR4" "DMDHREDU" "DMDHRMAR"
## [73] "DMDHSEDU" "gender"   "rel"      "bmi"
```

For this analysis, however, we use height and weight. We start with a simple scatterplot of height and weight with expected results (there is a correlation between height and weight, and men tend to be taller on average than women). When adding a smoothed curve to the data, however, we are able to discuss how smooth curves are created, how to find the SE of the smooth curve, why there is extra variability due to extremes and also due to fewer data points on the ends, extrapolation (note that the two curves have different ranges), and the outcome that slopes of the two curves are not substantially different (no interaction) though might warrant further study.

```
ggplot(adults, aes(x=BMXHT, y=BMXWT, group=gender, color=gender)) +
  xlab("Height") + ylab("Weight") + geom_point(alpha=.5) +
  ggtitle("Height vs Weight by Gender")
```

```
ggplot(adults, aes(x=BMXHT, y=BMXWT, group=gender, color=gender)) +
  xlab("Height") + ylab("Weight") + geom_point(alpha=.5) +
  stat_smooth(alpha=1) +
  ggtitle("Height vs Weight by Gender with Smooth Regression Fit")
```

Additional ideas for analysis:

With many continuous and categorical variables, the data can be used for both standard statistical regression (e.g., linear, logistic, etc.) or machine learning predictive modeling (e.g., LASSO, support vector machines, regression trees).

References

Pruim, R, D Kaplan, and NJ Horton. 2014. *Mosaic: Project MOSAIC (Mosaic-Web.org) Statistics and Mathematics Teaching Utilities*. http://CRAN.R-project.org/package=mosaic.

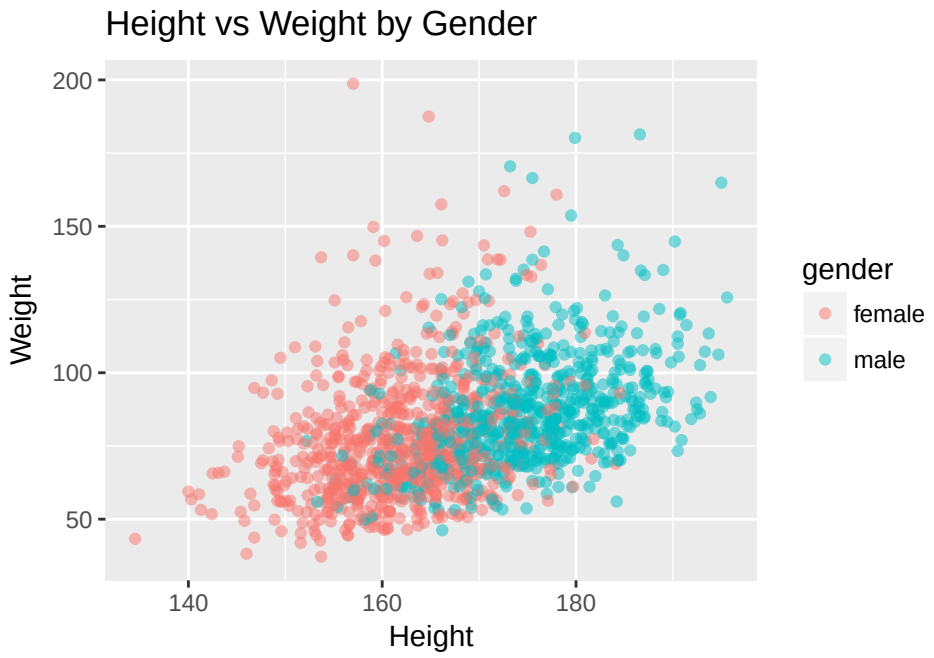


Figure 1: A scatterplot of weight and height broken down by gender. A smooth regression is fit to the points, and the confidence interval for the smooth curve as well as the increase in variability for small and large values of height are part of an important discussion in a second level regression course.

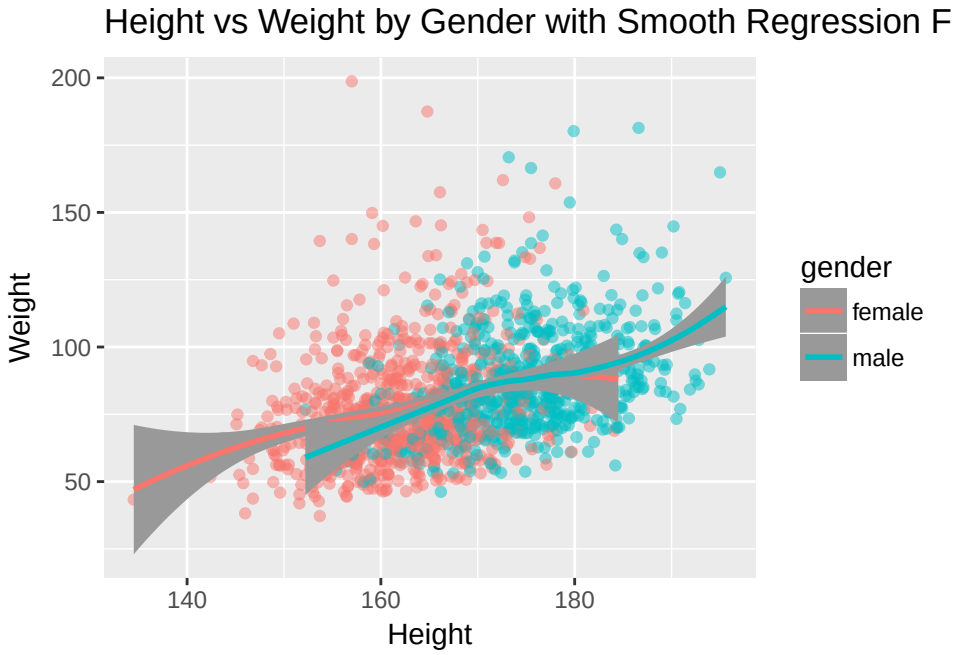


Figure 2: A scatterplot of weight and height broken down by gender. A smooth regression is fit to the points, and the confidence interval for the smooth curve as well as the increase in variability for small and large values of height are part of an important discussion in a second level regression course.