NHANES

June 2016

Introduction

The NHANES data come from the National Health and Nutrition Examination Survey, surveys given nationwide by the Center for Disease Controls (CDC). The data are collected to assess the health and well being of adults and children throughout the United States. The survey is one of the only to combine both survey questions and physical examinations.

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.

```
NHANES.demo <- sasxport.get("http://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/DEMO_G.XPT")
```

```
## Processing SAS dataset DEMO_G     ..
```

```
NHANES.body <- sasxport.get("http://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/BMX_G.XPT")
```

```
## Processing SAS dataset BMX_G      ..
```

```
NHANES.demo <-
  mutate(NHANES.demo, gender = ifelse(NHANES.demo$riagendr==1, "male", "female"))

comb <-
  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.

```
numobs = 2000
SRSsample <- sample(1:nrow(comb), numobs, replace=FALSE,
       prob=comb$wtmec2yr/sum(comb$wtmec2yr))
comb <- comb[SRSsample,]
```

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.

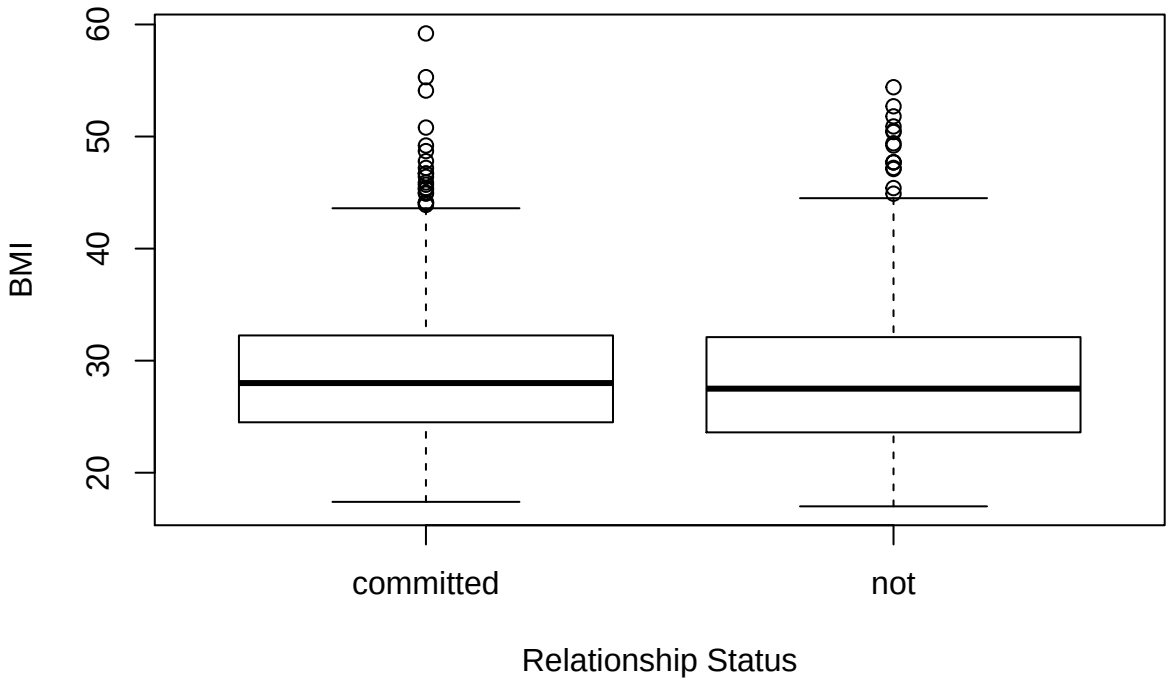
The boxplots and violin plots both demonstrate that there is not a substantial difference between the BMI for those in committed relationships versus those who are not. The tests of significance validate the ideas from the descriptive statistics.

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.

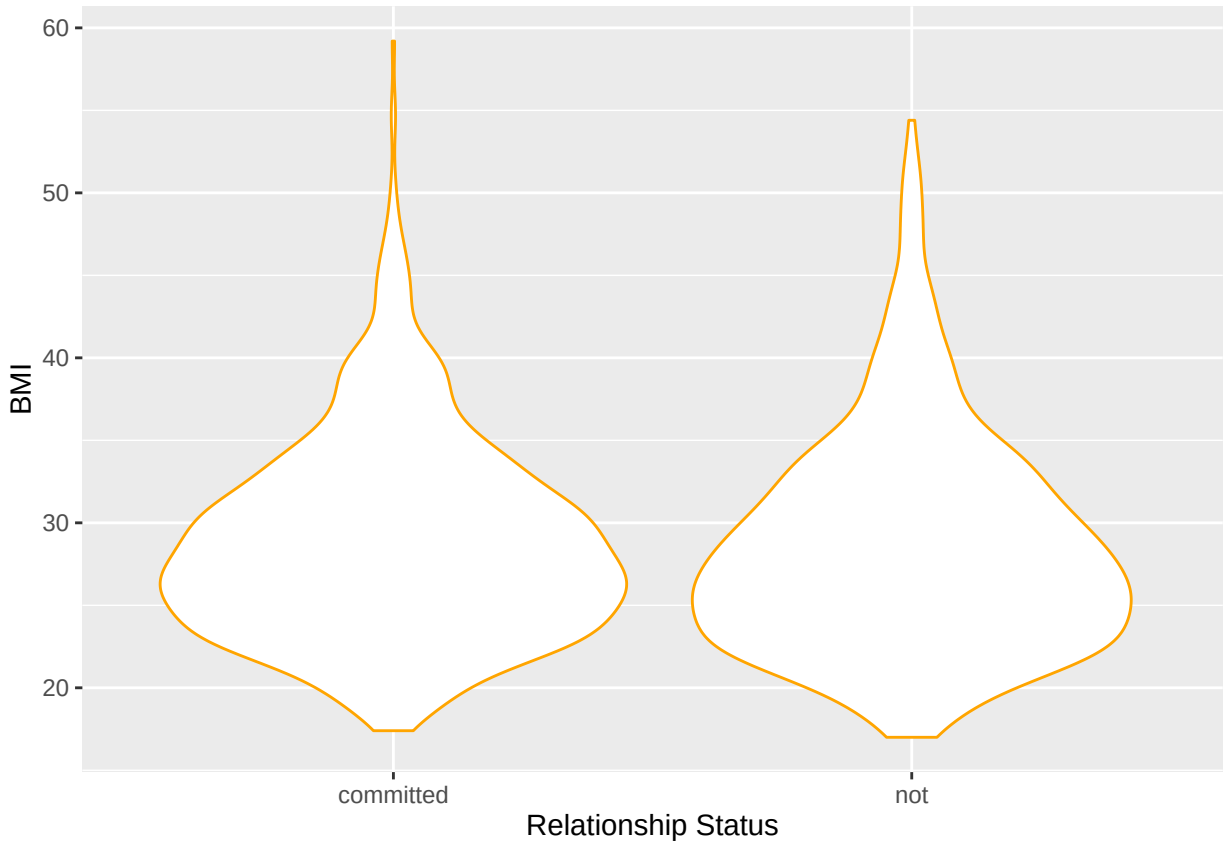
Additionally, again we point out that although these data are likely a good representation of the population, they cannot be used to find causal relationships. Indeed, even if BMI had been different on average across the two groups, we do not know if lower BMI causes one to be more likely in a committed 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")
```




```
ggplot(adults, aes(rel, bmi))+ geom_violin(color="orange")+
  xlab("Relationship Status") + ylab("BMI")
```




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

```
##
##  Welch Two Sample t-test
##
## data:  bmi by rel
## t = 0.82169, df = 1143.4, p-value = 0.4114
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.400355  0.977311
## sample estimates:
## mean in group committed       mean in group not
##                28.89678                28.60830
```

```
dim(adults)
```

```
## [1] 1405   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.

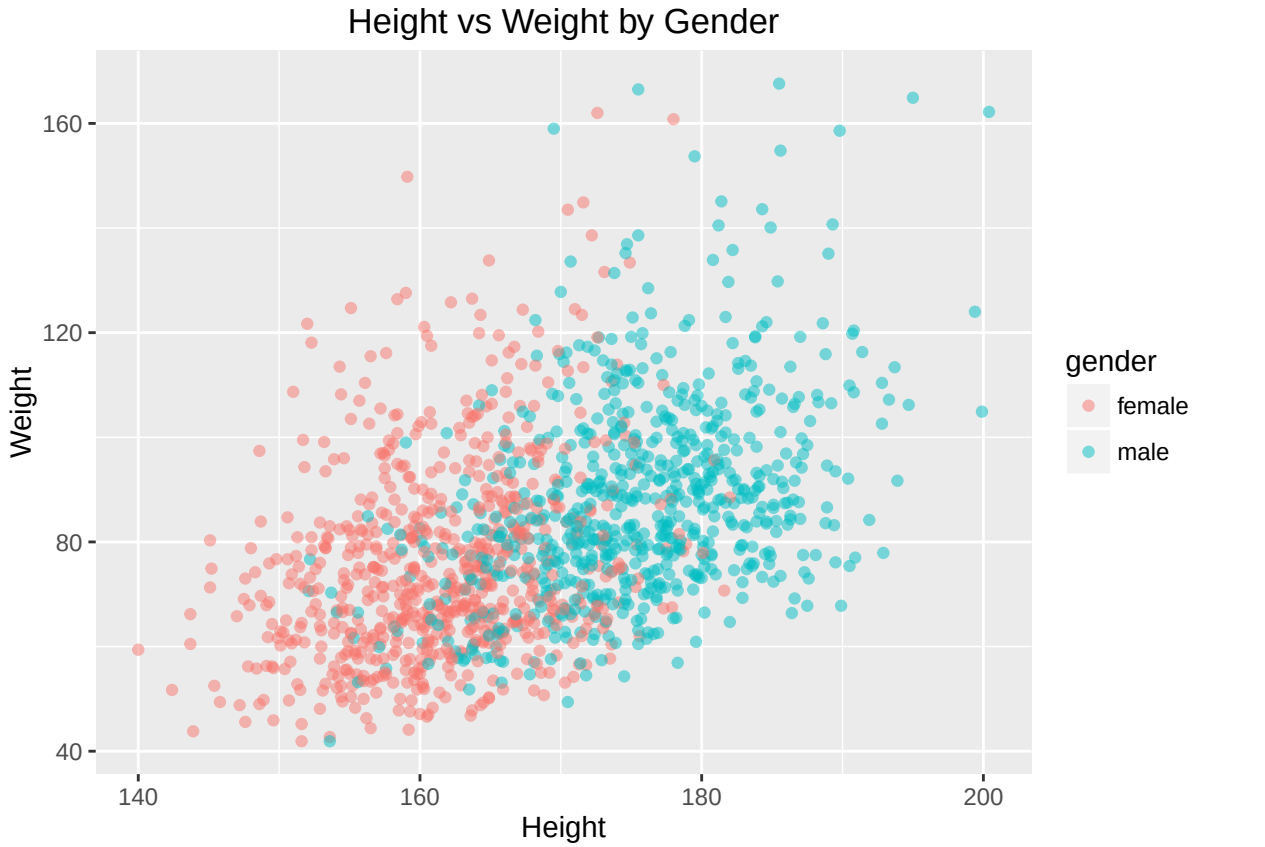
```
names(adults)
```

```
##  [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)) + geom_point(alpha=.5)+
  xlab("Height") + ylab("Weight") + 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).