Changes over Time

.....

Overflow from lab 5 – use for lab 6?

Note: It's very interesting that in 2010, Americans 65 and over are 13% of the population, but are a much larger proportion of the 2010 data in brfss. A good opportunity to discuss sampling methods, over/under sampling, etc.?

Height

14. Draw a basic histogram of the height variable using the brfss data. Is this a good view of the data? Explain.

15. In order to focus in on the typical values for the height variable, we can pipe the data through a filter to remove the tails. Let's keep the middle 99% of the values. First find the quantiles for 0.005 (the lowest 1/2-percent) and 0.995 (the highest 1/2-percent) and then them to form a subset of the heights:

```
lowCutoff <- quantile(~height, data=brfss, na.rm=TRUE, probs=.005)
highCutoff <- quantile(~height, data=brfss, na.rm=TRUE, probs=.995)
heightSubset <- brfss %>%
filter(height >= lowCutoff & height <= highCutoff)
```

    Now draw the filtered histogram:

```
histogram(~height, data=heightSubset)
```

    Describe what you see. Can you explain why it looks like that?

16. There is not *really* a gap in the middle of the data. The appearance of a break is coming from integer data interacting badly with the breakpoints for creating the histogram bins. One way to fix this is to put breaks specifically at all the ".5" marks on the horizonatal axis:

```
myBreaks <- seq(from=lowCutoff - 0.5, to=highCutoff + 0.5, by=1)
myBreaks
histogram(~height, data=heightSubset, breaks=myBreaks)
```

    Describe the height distribution. Be sure to discuss, what are the typical heights for these respondents.

Height and Sex

17. Because of the difference in average heights for males as compared to females, we might have expected the histogram to be clearly bimodal. Indeed, with a boxplot we can see this difference (include a bwplot of sex~height in your response to this question). Can you explain why the histogram does not show a clear bimodal pattern?

    As a companion to the bwplot, let's also make a histogram which is paneled by sex. Notice the use of the formula ~height|sex for height versus sex, and the layout=c(1,2) option for forcing the panels to line up vertically (1 column, 2 rows):

```
histogram(~height|sex, data=heightSubset, breaks=myBreaks, layout=c(1,2))
```

Height and Weight

In this section we investigate the relationship between height and weight for our respondents. Because the data file is huge (over 1.3 million rows), we can save computing time by working with a random sample of the rows.

18. Use the following commands to make a scatterplot of weight versus height for a sample of 30,000 rows from the brfss dataset (notice the use of the slice command to use a subset of the rows):

```
mySample <- sample(1:nrow(brfss), 30000)
summary(mySample)
brfssSample <- brfss %>% slice(mySample)
xyplot(weight~height, data=brfssSample)
```

Describe the overall pattern in the data. Is there a relationship between height and weight for these subjects?

19. One way to summarize the data in a scatterplot is to add a smooth fit line (notice that the "line" in this context might be curved). Add the fit line with the panel.loess command (lwd= is an option for setting line width):

```
ladd(panel.loess(x, y, col="darkgreen", lwd=2))
```

The smooth fit line is almost straight. Does this mean there is a "strong" linear relationship between the two variables? Explain.

TODO: more options: pch, cex

.....

The data in brfss span 15 years; we have data for 2000, 2005, 2010, and 2015.

14. Make a stacked bar graph showing age7 (age in 10-year increments) vs. year (survey year). What do you learn about the U.S. population over the last 15 years?

15. Use tally and barchart to make a graph of the general health level against the survey year, and write a conclusion. What do you learn about general health levels in the U.S. over the time span 2000 to 2015? Provide a plausible explanation for any pattern you see.

Whenever we notice an association between two variables, we should ask if there are "lurking" variables that might help to explain the association. Age and sex are possible lurking variables in this situation.

16. Use the code provided below to make a graph of general health vs. survey year, broken down (paneled) by age group.

```
myColors = brewer.pal(5, "PuBuGn")
myKey = list(text=list(levels(brfss$genhealth)), columns=4,
    rectangles=list(col=myColors))
healthVsYearAndAge <- tally(~genhealth|year+age7, data=brfss,
    format="percent", useNA="no")
healthVsYearAndAge %>% aperm(c(2,3,1)) %>% barchart(col=myColors, key=myKey)
```

Describe what you see in this final graph. How does it compare with your observations in the previous question? How does this make sense?

=======

Need to introduce mosaicplot at some point.

12. As an alternative to the stacked bar graph, we can draw a mosaicplot. A basic 2-variable example shows the relationship between income and health for the brfss participants:

```
healthVsIncome <- tally(~income+genhealth, data=brfss,useNA="no")
healthVsIncome %>% mosaicplot(color=brewer.pal(5,"RdPu"))
```

What do we see in this plot? What is the meaning of the varying bar widths? Etc.