Lab 2: Categorical Data and Multi-Variable Investigations

Overall Goals

In Lab 1[1] we were introduced to the basic operations of RStudio. In this lab we expand on the ways R helps us analyze datasets. We will learn how to:

- aggregate data based on a categorical variable
- work with individual variables (e.g. sorting, filtering)
- visualize categorical information with barcharts and pareto charts
- visualize "labeled" quantitative information via dotplots
- visualize categorical-quantitative relationships via boxplots

Specific R commands introduced: tally, sort, barchart, dotplot, tail, head, bwplot, median

Recall that before we start any work, we need to load our main functions package:

```
library(hanoverbase)
```

We will continue working with the counties dataset introduced in Lab 1. If the counties set is not visible in the Environment pane, you will need to "reload" it by running:

```
data(counties)
View(counties)
```

Working with Categorical Variables

Recall that categorical variables take values from a limited set of options, one value for each individual case. In our counties data set the "individual cases" are the U.S. counties.

1. What are the categorical variables in the counties dataset? You may need to bring up the help file for counties using the help(counties) command.

We start by asking for the number of counties in each state. This is what we would commonly call the *frequencies* of the "state" variable. The tally command is used to compute frequencies:

```
tally(~state, data=counties)
```

[1] Lab1Instructions.html

It would be nice to have these values sorted. Use the UP arrow to recall your last command, and add a "%>%" pipe step to sort the values, calling the sort command:

```
tally(~state, data=counties) %>% sort()
```

2. List the five states with the highest number of counties, including how many counties they each have.

We can get a visual representation of these counts using a barchart:

```
tally(~state, data=counties) %>% barchart()
```

Click the **Zoom** button to get a longer view of the barchart.

The default barchart makes no attempt to order the categories, which can make the barchart hard to read. We can combine tally, sort and barchart to achieve a barchart ordered by the frequencies. This is typically called a **Pareto chart**:

```
tally(~state, data=counties) %>% sort() %>% barchart()
```

3. Which states have fewer than 10 counties? Can you see this directly from the barchart?

On a lighter note, let us use the county name as a categorical variable, and determine the most popular county names are. We will combine tally and sort as above but using the name variable instead of the state variable.

```
tally(~name, data=counties) %>% sort()
```

The result of that command is pretty overwhelming. Since there are so many counties whose names only appear once, we do not get to see the end of the list. To do that, we can use the tail command, which shows a number of entries from the end:

```
tally(~name, data=counties) %>% sort() %>% tail(10)
```

There is similarly a head command that would show us a number of entries from the top. It is not useful in this particular instance.

4. What are the top 5 most frequent county names? Why does it make sense that these are popular county names?

5. We are curious to find out which states have a "Lincoln County". We will use the data table on the top left of the screen for this. If the county data is not visible there, double-click it in the Environment pane to bring up the data table. Use the filter option at the top right of the data table to only show the rows that contain the name "Lincoln County".

   Which of the original confederate states (South Carolina, Mississippi, Florida, Alabama, Georgia, Louisiana and Texas) have a "Lincoln" County? Which of these are actually named in honor of Abraham Lincoln? (Hint: include the word "namesake" in your internet search string.)

Working with Two Variables of a Dataset

In most situations in statistics we deal with more than one variable at a time, and more specifically we are interested in the relationships between variables. For example, we may look at the percent of female population for each county as a variable on its own (one variable analysis), or we may look at how these percents vary across states (two variable analysis).

6. Let us start by looking at the female variable. Look at the help file for the counties dataset. Exactly what is being measured by the female variable?

We can start our investigation by making a histogram and a five-number summary for the female variable. **Note**: The two lines are two different commands. Be sure to press enter on the first before typing the second.

```
histogram(~female, data=counties, breaks=40)
favstats(~female, data=counties)
```

7. Describe the distribution in plain English. Include a discussion of the range, shape (number of modes, skewness) and center.

8. Does the mode's value make sense? Why does it make sense that the distribution is unimodal?

9. Using the data table and sorting the female column, identify the three counties with the smallest percent of females. Possible reasons for a small female percentage could be the presence of a large male prison or military base, relative to the overall county population. Does this explanation seem to apply to the three counties (Hint: Wikipedia or Google search)?

10. Which county has the highest female percentage? Can you explain the probable reason for that?

We would like to compare these female percentages across states. Are there states whose counties tend to have unusual gender disparity? Or is the gender distribution quite uniform across states? A first step to answering this would be a boxplot of the female percentage against the states:

```
bwplot(state~female, data=counties)
```

11. Which state stands out in that most of its counties have a relatively low percentage of females?

It is difficult from this boxplot to identify anything more than the most extreme cases. We can provide "less" information per state to enable us to compare the states more easily. For instance we can ask what the median is, for each state, of the female percentage by county:

```
median(~female|state, data=counties) %>% sort()
```

For a visual depiction of this table, we can use a **dotplot**:

```
median(~female|state, data=counties) %>% sort() %>% dotplot()
```

We should point out that the values in this table are NOT the percent of females in the state. They are instead the medians of the percentages of the various counties in the state. We will see a bit later how to compute the state-wide percentages by *aggregating* the data.

Digging Deeper We might be interested in a single state and the distribution of female percentages in its counties.

We can get a quick listing of the counties for a single state by typing the state name in the "filter" box at the top right of the data table.

12. Use this method to find which county in Indiana has the highest female percentage. Is there a good explanation for why this county has such a high percentage?

Alternatively, we can make a visual display by using a filter and dotplot. We will try this for the counties in Alaska. The first line below stores the Alaska county data under the name alaskaCounties. **Note**: The two lines are two different commands. Be sure to press enter on the first before typing the second.

```
alaskaCounties <- counties %>% filter(state=="Alaska")
dotplot(name~female, data = alaskaCounties)
```

You should also use the following two commands for alternative views of the same distribution (don't forget that you can use the arrow buttons in the **plots** pane to scroll back and forth through the graphs you have made):

```
bwplot(~female, data=alaskaCounties)
histogram(~female, data=alaskaCounties, breaks=20, type="count")
```

13. Describe the pattern of the distribution of female percentages on the various counties in Alaska (typical values, outliers). Also provide the county name for each outlier and its percent female population (2010).

Please **click the red button (upper right)** to close your RStudio session when you are getting ready to leave the lab.