Political Preferences

Creating Population Estimates Using Weighted Data

Many statistics classes talk about data that comes from a **simple random sample** (SRS), used to accurately estimate the population. A SRS minimizes any potential bias by ensuring that each individual (and each combination of individuals) in a population has an equal chance of being selected.

When a researcher is interested in examining distinct subgroups within a population, it is often best to use a **stratified random sample** to better represent the entire population. A **stratified random sample** involves dividing the population of interest into several smaller groups, called "strata" and then taking a simple random sample from each of these smaller groups. This method is commonly used when we want to guarantee a large enough sample from each subgroup. When this type of sampling method is used, it is important to use **weights** to take the relative size of each subgroup into account.

Weights are values assigned to each individual within a sample. These assigned values take into account how representative the individual is of the total population.[1] When subjects are multiplied by their weights, the adjusted sample is more reflective of the actual population.

Political Preferences: Do voters Tend to Lean Towards the Republican or the Democrat Party?

In 2010, CBS and the New York Times conducted a national phone survey of 1189 subjects as part of "a continuing series of monthly surveys that solicit[ed] public opinion on a range of political and social issues" (ICPSR, 2012 March 15). In addition to *Political Preference*, they gathered information on *Race, Sex, Age,* and *Region of Residence*. Every individual in this data set was assigned a weight based on *Age, Sex, Race, Education,* and *Region*.

Part A. Looking at the Data Set

In this lab activity, you will need to use the following packages. The following code will install the packages if they are not already installed, and then use the require() function to queue them in your console for the lab.

```
#The {ggplot2} package creates graphs to visualize our data
if(!require(ggplot2)) install.packages("ggplot2"); require(ggplot2)

#The {survey} package factors the weights into our estimations and visualizations
if(!require(survey)) install.packages("survey"); require(survey)

#The {weights} package aids in creating our estimations and visualizations
if(!require(weights)) install.packages("weights"); require(weights)
```

After requiring these packages, read in the data set using the following command. You will need to specify the file path in the read.csv() function where it says "filepath". This code will assign the name "Political" to the data that we read in.

```
Political <- read.csv("filepath")
```

The head() function shows the variable names and the first 5 rows of the data set that you specify. To get an idea of the variables contained in the Political data set, use the code below to look at the first few rows of observations.

```
head(Political, 5)
```

Notice that the data set contains the following 8 variables:

- Weight - The assigned survey weight of each subject
- Race - Black, White, Asian, Other
- Sex - Female, Male
- Income (in thousands of dollars)
- Age group (18-25, 26-35, ..., 66-75, 76+)
- Region - Northeast, Northcentral, West, South
- Preference - Democrat, Republican, Alternative choice
- X- A column designating row number

Questions

1. What is the sex of the 3rd entry?
2. What is the political preference of the 2nd entry?
3. Due to nonresponse/missing values our dataset has fewer observations than the total number of individuals surveyed. Use the dim(Political) function to determine how many observations are in the data set.

Part B. Making Summary Tables

Suppose you want to see if there is a difference in political preference by race. This lab will investigate how failing to take the weights into account might affect this potential difference. The example below demonstrates one method of examining the data to answer this question by creating visualizations.

Unweighted Data Table

To make a two-way summary table you first need to calculate the counts for each subgroup of the sample. For example, suppose you first want to determine the number of Asian Republicans in the sample. The code below creates a new column in the data frame to count the number of Asian Republicans.

```
#Create a new column where a 1 indicates that Race = Asian and Preference = Republ
ican and a 0 if both conditions aren't satisfied
Political$ARIndicate <- ifelse(Political$Race == "Asian" &  Political$Preference =
= "Republican", 1, 0)
```

To calculate the total number of Asian Republicans in the sample, use the sum() function on the *ARIndicate* column, as demonstrated below.

```
sum(Political$ARIndicate)
```

Rather than continuing to count the table values this way, you can use the ftable() function in R. The code below creates a summary table (called *unweighted*) of political preference by race for

the observed sample. Note that *Political$Preference* and *Political$Race* refer to the *Preference* and *Race* columns in the *Political* data set.

```
#Create a summary table of the data
unweighted<-ftable(Political$Race, Political$Preference)
```

Print the unweighted table to see the counts by typing the name of the table into R, like the code below.

```
unweighted
```

Questions

3. Both the sum() command and the ftable() should give the same number of Asian Republicans in the sample. What value did you find?
4. How many of the observed individuals were both White and indicated that they preferred the Democratic Party?
5. What percentage of the sample are African Americans?

Weighted Data Table

It appears African Americans were oversampled at nearly twice the rate of the actual population proportion of Afircan Americans in the U.S.

Because CBS used weights to account for population estimates based on *Race, Sex* and *Gender*, it is likely that the weights assigned will address this crucial disparity in African Americans surveyed versus the true proportion of African Americans in the U.S.

In order to incorporate the weights into the analysis, you can multiply each entry by their assigned weights and then use these new values to calculate the totals. For example, to calculate the cell count for the Asian Republican cell in a weighted two-way summary table, multiply the *ARIndicate* column by the *Weight* column and then sum those values.

To calculate the values for the weighted two-way summary table, you will need to use the *ARIndicate* column created earlier. This time, you will need to multiply the column by the *Weight* column and then sum those values.

```
#Create a column called ARWeight of ARIndicate multiplied by the Weight column
Political$ARWeights <- Political$ARIndicate*Political$Weight
#Sum the new column
sum(Political$ARWeights)
```

You can also create this weighted table using the svydesign() and svytable() commands from the {survey} package. First, you must specify the survey design and designate the column in the data set that contains the weights. The "~0" in the code below indicates that there is no clustering in this data set. If a data set does use clustering or stratification, you should consult the {survey} package for more information about what is the most appropriate code to use.

```
#Modifies the data to be used in the {survey} package
polidesign <- svydesign(~0, data=Political, weights=Political$Weight)
```

```
#Creates a table of the estimated weighted frequencies of the data
weighted <- svytable(~Race + Preference, polidesign)
```

Print the table to see the weighted frequencies using the code below.

```
weighted
```

Questions

6. Both the sum(Political$ARWeights) and the weighted table should give the same estimated count of Asian Republicans. What value did you find?
7. How many of the weighted individuals were both White and indicated that they preferred the Democratic Party?
8. What percentage of the weighted data represent African Americans?

Part C. Visualizing the Political Data

Unweighted Bar Graph

In order to visualize the unweighted data, convert the unweighted table to a data frame using the code below:

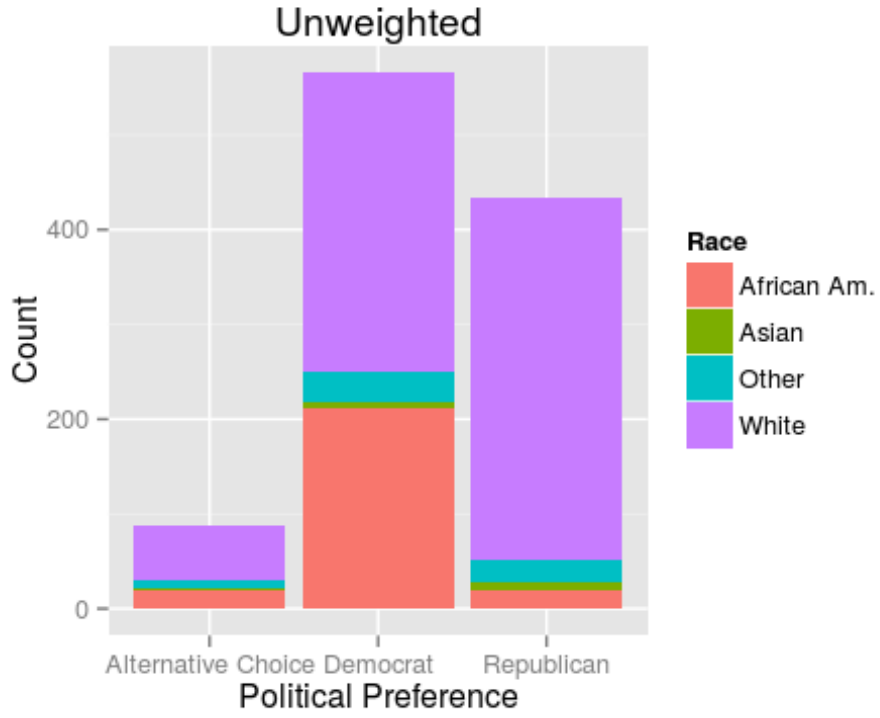
```
#Converts unweighted into data frame to be used with ggplot()
unweighteddata <- as.data.frame(unweighted)

#Assigns the column names Race, Preference, and Counts to the data frame
colnames(unweighteddata) <- c("Race", "Preference", "Count")
```

The code below specifies which column is on the x-axis, which column is on the y-axis and which column determines the fill of each bar. Since you want to look at the counts of each political party, set *Preference* as the x-axis variable and *Count* as the y-axis variable. Then, since you are intested in comparing the political parties by racial make-up, specify *Race* as the fill variable.

The ggplot() function creates the graph, and the geom_bar() function specifies how the bars are filled (in this case, "identity" means show the graph as a bar graph of counts). The labs() function lets us choose labels for the axes and legend, and ggtitle() allows you to title the graph. All of these functions are found in the {ggplot2} package.

```
#Generates a barplot of the unweighted table created above
 ggplot(unweighteddata, aes(x = Preference, y = Count, fill = Race)) +
     geom_bar(stat = "identity") +
  labs(x = "Political Preference", y = "Count", fill = "Race") +
     ggtitle("Unweighted")
```



Weighted Bar Graph

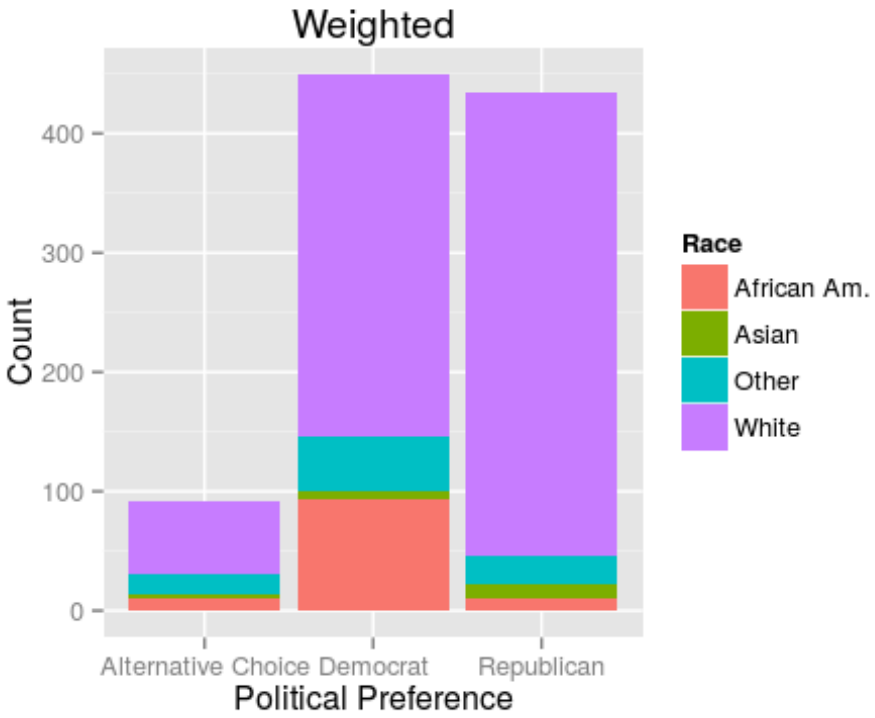
In order to visualize the weighted data, convert the weighted table to a data frame using the code below:

```
#Converts weighted into a data frame to be used with ggplot()
weighteddata <- as.data.frame(weighted)

#Assigns the column names Race, Preference, and Counts to the data frame
colnames(weighteddata) <- c("Race", "Preference", "Count")
```

In the code below, use the column names to specify which columns to appear on the x-axis, y-axis as well as the fill variable. As with the unweighted data graph, you can specify the labels and title using labs() and ggtitle() respectively, as well as indicate that you want the graph to be a bar graph of counts.

```
#This generates a barplot of the weighted table created above.
ggplot(weighteddata, aes(x = Preference, y = Count, fill = Race)) +
       geom_bar(stat = "identity") +
       labs(x = "Political Preference", y = "Count", fill = "Race") +
       ggtitle("Weighted")
```



When comparing the weighted and unweighted graphs, it is important to understand the context in which the data was collected. The sample proportion of African Americans was 23%, while the actual proportion of African Americans in the US is closer to 14%. The African American population of the US tends to vote largely Democrat. So, because they were over-represented in the sample data, it appears that voters greatly prefer Democrat policies over Republican policies. However, once population estimates are taken into account via weights, we see that the divide between Republican and Democrat is much smaller than it initially seemed.

Questions

9. Does the weighted data visualization have a different estimate of Democrats? How about Republicans?
10. Which graph most accurately reflects the population? Explain why you think this is.

Part D. Try it On Your Own

Go to the Political Data shiny app.

Select *Political Preference* as the **X-Axis Variable** and *Race* as the **Color By** Variable. Click the counts button above the graphs. Notice that these two graphs are the same ones you made in this lab. So far, this lab has focused on how weighting affects the distribution of political preference. Suppose you are now interested in looking at the racial breakdown within each political preference.

11. What can you conclude about the racial make up of those who prefer the Democrat party?

Now choose the **Percentage** button, instead of **By Counts**.

12. How do the weighted and unweighted graphs compare? What can you conclude about the racial make up of those who prefer the Democrat party from each graph?

Select *Sex* instead of *Race* as your **Color By** variable. This will give you the sex breakdown within each party.

13. How do the weighted and unweighted graphs compare?

14. Do you notice a more or less extreme difference between the two compared to the difference seen when looking at *Race*? Why do you think this is?

Conclusion

When studying a weighted survey it is important to note that weights affect not only visualizations of the data but also impact test statistics. Therefore it is essential to use caution when testing weighted data to be sure to account for the weight adjustments.

More information on this data set available at:

http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/11/studies/33183

Note: the weights for this data set were adjusted to account for non-response rates and the likelihood of being called (ie. households with more than one phone are more likely to be called than households with one phone).

References:

Winship, Christopher and Larry Radbill (1994). Sampling Weights and Regression Analysis. *Sociological Methods and Research.* 23(2), 230-257.

---

[1] For an introduction to weighted survey data, see the activity, "Should it Pass? An Introduction to Weighted Data" available at http://web.grinnell.edu/individuals/kuipers/stat2labs/weights.html