Complementary and Alternative Medicine: Testing Weighted Data

Part A. An Introduction to Weighted Data

Weighting is commonly used in survey analysis to take into account the sampling design used in the collection of data, along with patterns of response or non-response among the individuals surveyed. With weighted data, each subject in the survey is assigned a value (called a **weight**) according to how representative they are of the total population. When subjects are multiplied by their weights, the adjusted sample is more reflective of the actual population.

When dealing with weighted data, these weights must be taken into account in order to perform an appropriate analysis.

Testing Weighted Data

There is no universally accepted norm for analyzing weighted data; statisticians are still working to determine the best way to account for weights in data analysis. However, some methods are more appropriate than others. Today's lab will focus on examining three different methods for analyzing weighted data, in the hopes of determining which is most appropriate to use, given the information available.

The three methods you will examine today are:

- The **Simple Random Sample (SRS) Method** assumes that the sample is an unweighted sample that is representative of the population, and does not include adjustments based on the weights that are assigned to each entry in the data set.
- The **Raw Weight (RW) Method** multiplies each entry by their respective weight and runs the analysis on this adjusted weighted sample.
- The **Rao-Scott Method** takes into account both sampling variability and varibility among the assigned weights to adjust the chi-square from the RW method.

Introduction to the CAM Data Set

One example of a data set which incorporates a weight variable is the **Complementary and Alternative Medicine (CAM) Survey**, which was conducted by the National Center for Health Statistics (NCHS) in 2012.

For the CAM survey, NCHS researchers gathered information on race, sex, employment, marital status, whether each individual surveyed chose to use CAM and if so why they chose to use it, as well as whether the individual had undergone physical therapy or surgery, had seen a chiropractor or used one of many other CAM methods. Weights were assigned based on race, sex, and age.

Part B. Importing the Data

Require the necessary packages using the code below:

```r
if(!require(ggplot2)) install.packages("ggplot2"); require(ggplot2)
if(!require(weights)) install.packages("weights"); require(weights)
if(!require(survey)) install.packages("survey"); require(survey)
if(!require(scales)) install.packages("scales"); require(scales)
```

First, 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":

```r
CAM <- read.csv("filepath")
```

The code below will allow you to see the first 5 rows of the data set to get a better idea of the information it contains. You can change the number of rows displayed by changing the number in the command.

```r
head(CAM, 5)
```

1. What is the sex of the 4th subject surveyed?
2. Is the 2nd subject a CAM user?

Part C.Two-Way Tables Manually

To make visualizations, you first need to create two-way tables of the data you are analyzing.

Below is an example comparing the proportion of subjects by region that have had physical therapy among the *Black/African American Wellness users* in the sample.

Unweighted Observed:

Region	Yes	No	Row Total
South	16	113	129
Midwest	3	54	57
Northeast	5	42	47
West	5	31	36
Column Total	29	240	269

Unweighted Expected:

To calculate the expected cell values, use the equation $\frac{\text{row total}\cdot\text{column total}}{\text{total}}$. For example, to calculate the expected count for the "south" row, "yes" column, you would use $\frac{129\cdot 29}{269} = 13.91$.

3. Using this formula, fill in the missing cells below.

Region	Yes	No
South	13.91	
Midwest		
Northeast		41.93
West	3.881	32.12

Weighted Observed:

Remember, the weighted tables are created by multiplying each subject by their assigned weight and then summing the total of the weights for each cell in the table.

Notice that the observed values added together always equal the row total or column total. For example, the "South Yes" and the "South No" cells added together equal the "South Total" cell.

4. Using this information, finish filling out the table below.

Region	Yes	No	Row Total
South	68,312	658,296	726,608
Midwest		297,578	313,571
Northeast	20,496	276,689	297,185
West		110,336	143,523
Column Total	137,988		1,480,887

Weighted Expected:

The weighted expected frequencies are calculated using the same formula as earlier. Below is the completed table.

Region	Yes	No
South	67,705	658,903
Midwest	29,218	284,353
Northeast	27,691	269,494
West	13,373	130,150

Part D. Using R Code to Create the Two-Way Tables

The code that follows will allow you to create the two tables above, one of the unweighted counts and one of the weighted frequencies. Once you have created these tables, you can use them to make visualizations.

Unweighted

To create the unweighted table in R (rather than by hand), we need to subset the data so that we can create totals based on only our subset of the sample.

Below is an example of how to subset the data to look at specifically Wellness users. Note that you need to include both the specific column you are interested in, as well as the subgroup within the column that you wish to subset by. For instance, *Wellness* is a factor in the *use_type* column so you would need to include both that specific column and *Wellness* in the code below.

```
CAMsubset <- CAM[CAM$use_type == "Wellness", ]
```

Each time after you have subsetted your data, open it from your global environment to confirm that the subset command was effective. Make sure to use caution in regards to capitalization and spelling.

If you want to look at an even smaller portion of your data, you can subset it further. Below is an example of how to subset your data to look at a specific race.

```
CAMsubset <- CAMsubset[CAMsubset$race == "Black/African Am. Only", ]
```

You can also subset based on gender, location and other factors, but be mindful of the size of your subsetted data set; if you do not include enough individual entries, this will affect the validity of your chi-square test.

Once you have subsetted the data, you can create your table using the code below. Note that for the ftable() command you will need to specify the two columns you'd like to compare.

```
#Creates two way table of counts by region and physical therapy
unweightedtable <- ftable(CAMsubset$region, CAMsubset$physical_therapist)
```

Weighted

To create a table of the weighted frequencies, you will use slightly different code. First, you will need to specify a survey design (which will be used in the Rao-Scott correction). This survey design will indicate which column of your data contains the assigned weights. For the CAM data set, this will always be the *weight* column of the data frame. Next, the svytable() function creates a two-way table, taking the weights into account and subsetting by the variables that you specify.

```
#Specify survey design
camdesign <- svydesign(~0, data=CAM, weights=CAM$weight)

#Create your two-way table
weightedtable <- svytable(~region + physical_therapist, subset(camdesign, use_type
=="Wellness" & race=="Black/African Am. Only"))
```

The unweighted table represents the information used in a SRS test, while the weighted frequencies table corresponds to the RW and the Rao-Scott tests.

Part E. Visualization Code

Unweighted Counts

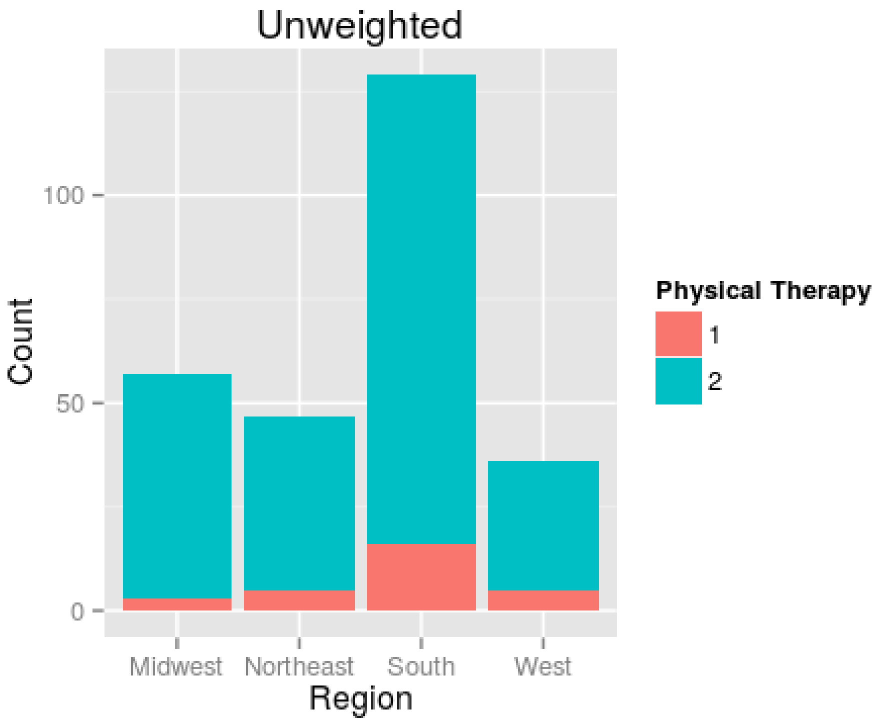
First you need to convert the unweighted table to a data frame that is compatable with the ggplot() function in the {ggplot2} package, which visualizes our data.

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

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

To create a graph of the unweighted data, use the code below. If you wish to analyze variables other than *region* and *physical therapist*, be sure to change the graph labels.

```
ggplot(unweighteddata, aes(x = Region, y = Count, fill = Physical_Therapy)) +
       geom_bar(
         #position = "fill",
         stat = "identity") +
       #scale_y_continuous(labels = percent_format()) +
       labs(x = "Region", y = "Count", fill = "Physical Therapy") +
       ggtitle("Unweighted")
```



Note: This graph corresponds to how you would treat the data in an SRS test.

Weighted Counts

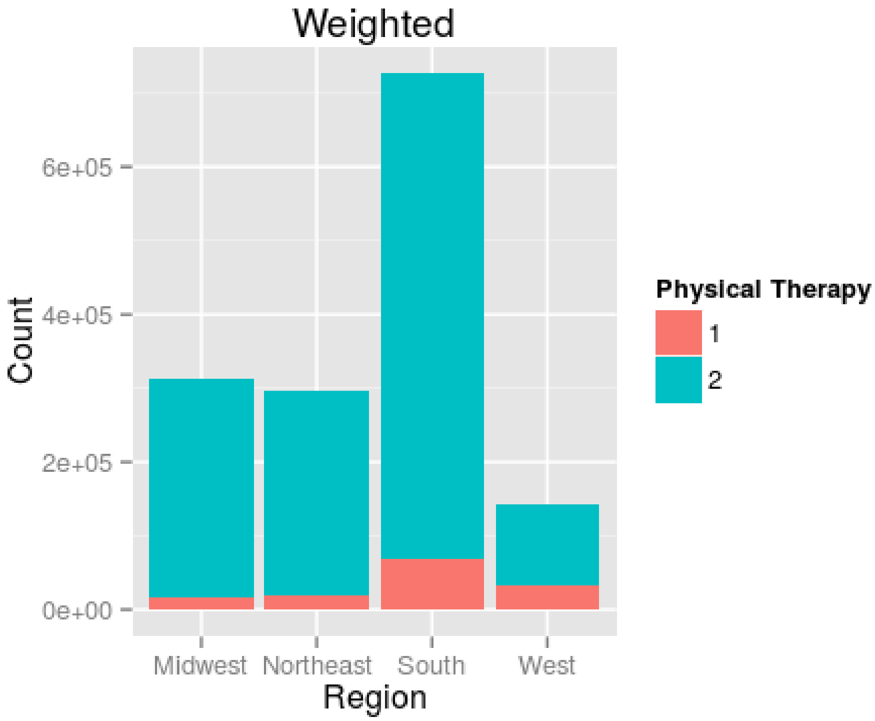
Now, you need to convert the weighted table to a data frame that is compatible with the ggplot() function.

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

#Assigns the column names Region, Physical_Therapy, and Couns to the data frame
colnames(weighteddata) <- c("Region", "Physical_Therapy", "Count")
```

To create a graph of the weighted data, use the code below. If you wish to analyze variables other than *region* and *physical therapist*, be sure to change the graph labels.

```
ggplot(weighteddata, aes(x = Region, y = Count, fill = Physical_Therapy)) +
       geom_bar(
         #position = "fill",
         stat = "identity") +
       #scale_y_continuous(labels = percent_format()) +
       labs(x = "Region", y = "Count", fill = "Physical Therapy") +
       ggtitle("Weighted")
```



Note: This graph corresponds to how you would treat the data in a multiplicative test or before a Rao-Scott correction.

The graph codes above contain two commmented out lines of code, the *position* argument and the *scale_y()* argument. Uncomment these lines in both the weighted and unweighted graphs, change the *y* argument in the labs() function to "Percentage", and run the code again to get graphs displaying percentage comparisons of the data, rather than comparison by counts.

5. What does the unweighted graph describe? How about the weighted graph?

6. When might the percentage graph be more appropriate to examine?

7. When might the counts graph be more appropriate to examine?

Part F. Testing the Data

Chi-Square Testing

Continuing with the example from above (comparing the proportion of Black Wellness users by region that have had physical therapy), here is the code to perform the three types of Chi-Square analysis from the introduction.

For both the SRS and RW chi-square tests in R, you will use the wtd.chi.sq() function from the {weights} package. For the Rao-Scott correction, you will need to take into account the design correction mentioned earlier in the lab.

SRS Chi-Square

To calculate the SRS chi-square test by hand, use the equation below. The observed and expected counts come from the unweighted two-way summary tables.

$$[1]\chi^2_{\text{unweighted}} = \sum \frac{(\text{observed counts} - \text{expected counts})^2}{\text{expected counts}}$$

.

As stated before, you will use the wtd.chi.sq() function to calculate the chi-square test in R. In this function, you have the option to specify a weight column. For the SRS test you do not need to specify a weight column and the wtd.chi.sq() function performs a simple two-way table chi-square analysis on the two variables listed in the function.

```
wtd.chi.sq(CAMsubset$physical_therapist, CAMsubset$region)
```

8. Try switching the order of the variables for the wtd.chi.sq() function and running the code in your console. Does it impact the test output? If so, how?

Raw Weight Chi-Square

To calculate the RW chi-square test, use the same formula for the SRS test, shown again below. However, for the RW test use the weighted observed and expected values from the weighted data summary tables above.

$$[2]\chi^2_{\text{weighted}} = \sum \frac{(\text{weighted observed frequencies} - \text{weighted expected frequencies})^2}{\text{weighted expected frequencies}}$$

.

To calculate the weighted chi-square using R, you can again use the wtd.chi.sq() function from the {weights} package, this time assigning the weight column to be *CAMsubset$Weight*. Use the code below to run a RW chi-square test:

```
wtd.chi.sq(CAMsubset$physical_therapist, CAMsubset$region, weight=CAMsubset$weight
)
```

Rao-Scott Chi-Square

The first two tests used the general equation

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

.

For the Rao-Scott correction, we use the weighted tables from the RW section and run a chi-square test using equation [2]. However, we also take into accound weight variability by doing a few more calculations. We first multiply the chi-square test statistic from [2] by $\frac{\text{n}}{\text{N}}$, where n is the sample size (for example, 34,525 in this data set) and N is the sum of the weights (for example, 234,920,670 in this data set). We then divide by a design correction, D, found by taking into account the weights as well as the clustering and stratification involved in the collection of the sample. This gives us a final equation of:

$$\chi^2_{\text{Rao-Scott}} = \frac{n}{\hat{N}D} \cdot \sum \frac{(\text{weighted observed frequencies} - \text{weighted expected frequencies})^2}{\text{weighted expected frequences}}$$

.

To perform the Rao-Scott method in R, simply run summary() on the weighted table you made earlier for the weighted graph.

```
summary(weightedtable, statistic = "Chisq")
```

Conclusion

Now that you have performed each of the three types of tests used in weighted data analysis, let's briefly revisit the assumptions for each test. The SRS method assumes that you are analyzing a simple random sample, and that this simple random sample is representative of the population. However, because you know that the sample is neither a SRS nor is it representative of the population, this method is inappropriate. The RW method multiplies each entry by their weight giving a slightly more representative sample, however this method still assumes you are testing a SRS. In this example you do not have a SRS, as is the case with most surveys. Thus, both the SRS and RW methods are inaccurate methods for testing this data set. The Rao-Scott method involves adjustments for non-SRS sample designs as well as accounting for the weights, resulting in a better representation of the population.

Part F. Try it On Your Own

Go to the CAM Data shiny app.

First select *region* as the **X Axis Variable** and *physical_therapist* as the **Color By** variable. Click the **Test Outputs** Tab and make sure that your test results above match the output displayed.

9. Next, select the **X Axis Variable** to be *Sex* and the **Color By** variable to be *Surgery*. Examine the chi-square values from each of the three types of tests. Which test gives the most extreme p-value? The least extreme?

10. Select the **Graphs** tab and click the **Percentage** button. Do these graphs appear to agree with any of your chi-square values from the previous question?

11. The SRS method only takes sampling variablity into account. Why might the output from an SRS chi-square test be misleading or inappropriate?

12. The RW method uses an over-inflated sample size. What are the consequences of an extremely large sample size?

References:

Winship, Christopher and Larry Radbill (1994). "Sampling Weights and Regression Analysis." *Sociological Methods and Research.* 23(2), 230-257. Web. 9 June 2015.

Additional information about the CAM data set can be found at: http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm