Political Preferences

Conducting Hypothesis Tests Using Weighted Data

Part A. An Introduction to Weighted Data

There is no universally accepted norm for analyzing weighted data; statisticians are still working to determine the best way to account for this issue. However, some methods are more appropriate than others.

In today's lab, you will examine three different methods for analyzing weighted data:

- 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.

Political Preferences: Republican or Democrat?

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." They gathered information on *Race, Sex, Age, Political Preference* and *Region of Residence*. For *political preference*, each individual was given the choice between *Republican, Democrat*, or several alternative options, grouped together in this lab for simplicity's sake as *Alternative Choice*. Every individual in this data set was assigned a weight based on *Age, Sex, Race, Education,* and *Region*. The weights were then adjusted to account for non-response and likelihood of being called (ie. households with more than one phone are more likely to be called than households with one phone).

Note: Due to nonresponse/missing values our dataset has only 1087 observations

Part B. Importing the Data Set

Require the necessary packages using the code below:

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

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

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":

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

To get an idea of the variables contained in this data set, use the head() function to show the first 5 rows of the data set you input into the function.

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

Part C. Testing the Data

Let's assume that you want to test whether there is a significant difference in the proportion of male versus female support for each of the political parties.

Simple Random Sample (SRS) Method

First, perform a test using the SRS method, assuming that the data was collected via a simple random sample. To do this, you first need to create a table of observed counts for the data. The table() command in R will create a two-way summary table of counts for any two columns you specify. Use the code below to create a table of *Political Preference* by *Sex*.

```
#Creates a two-way table of counts called unweighted of the Preference column by t
he Sex column in the Political data frame
unweighted <- table(Political$Preference, Political$Sex)
```

Print the table to see the observed counts:

```
unweighted
```

Next, you need to create a table of the expected counts.

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 *Republican* row, *Female* column, you would use $\frac{433 \cdot 603}{1087} = 240.2$.

1. Using this formula, fill in the missing cell below.

Political Preference	Male	Female
Alternative Choice		48.82
Republican	192.8	240.2
Democrat	252	314

Once you have calculated the observed and expected counts, you can continue with the chi-square test manually, using the formula below. When calculating the SRS test using this equation, the observed and expected counts come from the unweighted tables above.

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

.

You can also use R to calculate this chi-square value. The wdt.chi.sq() function from the {weights} package is typically used to perform a RW chi-square, but if you do not include a weight statement, and only specify the two columns which you wish to compare, it will perform a standard SRS chi-square test (assuming the weight for each observation is 1).

```
wtd.chi.sq(Political$Preference, Political$Sex)
```

2. What are the null and alternative hypotheses?

3. What p-value do you find using this test?

4. What do you conclude from this p-value?

5. Try switching the order of the two variables in the wtd.chi.sq function. Does it change the test output?

Raw Weights (RW) Method

Now that you have completed an SRS test, you can try the RW method. The main idea behind the RW method is to multiply each response by its respective weight as a way to adjust the sample to account for the weights. This provides a more representative estimate of the population proportion based on the sample design and other attributes addressed with the weighting variable. To create the two-way table for the RW analysis, you will need to first create a column of indicator variables for each combination of *Sex* and *Political Preference* (*PoliticalFD, PoliticalMD, PoliticalFR, PoliticalMR, PoliticalFAC*, and *PoliticalMAC*). Then, you will need to multiply each entry by its corresponding weight and sum those values to find the weighted frequencies.

```
#Create a new column where a 1 indicates that Sex = Female and Preference = Republ
ican and a 0 if both conditions aren't satisfied.
Political$FR <- ifelse(Political$Sex == "Female" &  Political$Preference == "Repub
lican", 1, 0)
```

Do this for every combination listed above (you should have a total of 6 new columns). Then, multiply each column by the weight column and sum those values by using the code below for each column:

```
#The code below first multiplies the FR column by the weight column, and then sums
to give the weighted frequency of Female Republicans.
sum(Political$FR*Political$Weight)
```

After computing each of the cell totals (and then adding them appropriately to get row and column totals), you should have an observed two-way table of weighted frequencies.

6. Using the R code above, fill in the table below.
7. Is the RW observed count of Female Democrats higher or lower than the SRS observed count?

Political Preference	Male	Female	Row Total
Alternative Choice	54.14	36.83	90.97
Republican	210.32	224.39	
Democrat		247.59	449.6
Column Total	466.5	508.8	975.3

Using this table of observed values, calculate the table of expected values using the same formula as before. Below is the resulting table.

Political Preference	Male	Female
Alternative Choice	43.51	47.46
Republican	207.9	226.8
Democrat	215.1	234.5

Once you have calcuated the observed and expected frequencies, you can continue with the chi-square test manually, using 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 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 *Political$Weight*. Use the code below to run a RW chi-square test:

```
wtd.chi.sq(Political$Sex, Political$Preference, weight=Political$Weight)
```

8. What p-value does this test yield?

9. Is it more or less extreme than the SRS p-value?

10. What conclusion would you make based on this p-value?

Rao-Scott Method

The first two tests used the general equation

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

For the Rao-Scott correction, use the weighted tables from the RW section and run a chi-square test using equation [2]. However, you also need to take into account weight variability by doing a few more calculations. First, multiply the chi-square test statistic from [2] by $\frac{n}{N}$, where n is the sample size (for example, 1087 in this data set) and N is the sum of the weights (for example, 975.3 in this data set). Then, divide by a design correction, D, found by taking into account the variability and uncertainty involved in creating the weights as well as taking into account the clustering and stratification involved in the collection of the sample. This gives 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, you can use 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 is necessary to help specify the design of the survey. It indicates that there is no clustering in this data set. If a data set does use clustering or stratification, you may need to consult the {survey} package for more information about what is the most appropriate code to use in your analysis.

```
#Specify the survey design (indicate the data frame and the weights column)
polidesign <- svydesign(~0, data=Political, weights=Political$Weight)
```

Create a table of the weighted frequencies of *Political Preference* by *Sex*. Notice that the code below specifies the survey design within the table command. This is to ensure that the table created is weighted. The weighted table involves multiplying each entry by their assigned weight and then using these new values to calculate the counts for each cell in the table.

```
#Create your two-way table (below, we call it "weightedtable")
weightedtable <- svytable(~Sex + Preference, polidesign)

#Get the chi-square value using the summary() function on the weightedtable
summary(weightedtable, statistic = "Chisq")
```

11. What p-value does this test give?

12. What conclusion would you make based on this p-value?

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

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 some cases, weights are designed to create population estimates. In these situations, the RW method over-inflates the test statistic and thus results in a superficially significant p-value **(see supplement for more detail)**. 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 D. Try it On Your Own

Go to the Political Data shiny app (https://shiny.grinnell.edu/Testing_Political_Data/).

Select *Political Preference* as your **X-Axis Variable** and *Sex* as your **Color By** Variable. Then select the **Test Output** tab. The results shown should match what you just found.

Now, select *Race* instead of *Sex* in the **Color By** variable to test the relationship between *Race* and *Political Preference*.

14. Compare the resulting chi-square test statistics and p-values. What do you notice?

15. How do these compare to the results you found when looking at *Sex*? Why might this be? More information on this data set available at:

https://www.icpsr.umich.edu/web/ICPSR/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.