Estimating and Testing Supplement

Working With Weighted Survey Data

Types of Weights (in R)

The two most common types of sampling weights are **expansion weights**, which scale the sample up to the population, and **proportional weights**, which have a mean approximately equal to 1. **Frequency weights**, another type of weight used in survey data, are whole numbers (integers) that indicate how many cases each entry represents. Frequency weights do not take survey collection methods into account. Although many types of weights may be used in survey data analysis, this supplement will focus primarily on expansion and proportional weights.

To explore these weights, we will be using two data sets, the Political data set and the CAM data set.

Political Data Set

This data set can be found at:

Political GITHUB FILEPATH

In 2010, CBS and the New York Times conducted a national phone survey of 1,087 subjects as part of "a continuing series of monthly surveys that solicit[ed] public opinion on a range of political and social issues" (ICPSR 33183, 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 sampling weight based on age, sex, race, education, and region, where the average of the weights is .897. These weights are proportional weights, but the average is not exactly 1.

Below are the first 5 rows of this data set.

```
##   X.1 X Weight        Race    Sex Income   Age       Region Preference
## 1   1 1  2.702       Asian Female  50-75 26-35        South Republican
## 2   2 2  0.474       White   Male   > 75 46-55    Northeast   Democrat
## 3   3 3  0.616 African Am. Female   > 75 46-55        South   Democrat
## 4   4 4  0.686       White   Male  30-50 66-75 Northcentral Republican
## 5   5 5  0.458       White   Male   > 75   76+        South Republican
```

CAM Data Set

This data set can be found at:

Complementary and Alternative Medicine (CAM) Survey GITHUB FILEPATH

In 2012 the National Center for Health Statistics (NCHS) conducted a survey of 34,525 subjects. 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, where the individual weights assigned to each entry sum to 234,920,670. These weights are expansion

weights meant to scale the sample up to the population of people over 18 years of age in the United States. The data set provided contains a selection of variables (from the full survey) that we are interested in exploring.

Below are the first 5 rows of this data set:

```
##   weight strat psu cam_use use_type    sex    region          race
## 1   9428   109   1 NonUser  Nonuser   Male   Midwest    White Only
## 2   7609     9   2 NonUser  Nonuser Female Northeast    White Only
## 3   2550    58   1 NonUser  Nonuser   Male   Midwest    Asian only
## 4   3899   183   2 NonUser  Nonuser Female     South    White Only
## 5  17129   153   1 NonUser  Nonuser   Male     South Multiple Race
##   chiropractor physical_therapist
## 1            0                  0
## 2            0                  0
## 3            0                  0
## 4            0                  0
## 5            0                  0
```

Notice that within this data set we have a column called *psu* and another called *strat*. These columns indicate the clustering and stratification (respectively) that were involved in collecting this data and will be discussed in more detail later on.

Estimating with Weights

When using either estimation or proportion weights to estimate population parameters, we must take the weights into account. If we ignore the weights (and assume that the sample was collected using a SRS), we get an inaccurate estimation of the population, since weights are used to account for non-response, over-sampling and/or undersampling of a portion of the population. Thus, we must multiply each entry by its respective weight to accurately estimate the population.

Below is an example of the impact of using weights in population estimates, using the Political data, looking at Political Preference by Sex.

Unweighted Graph

This example uses R code to create bar plots using unweighted and weighted data.

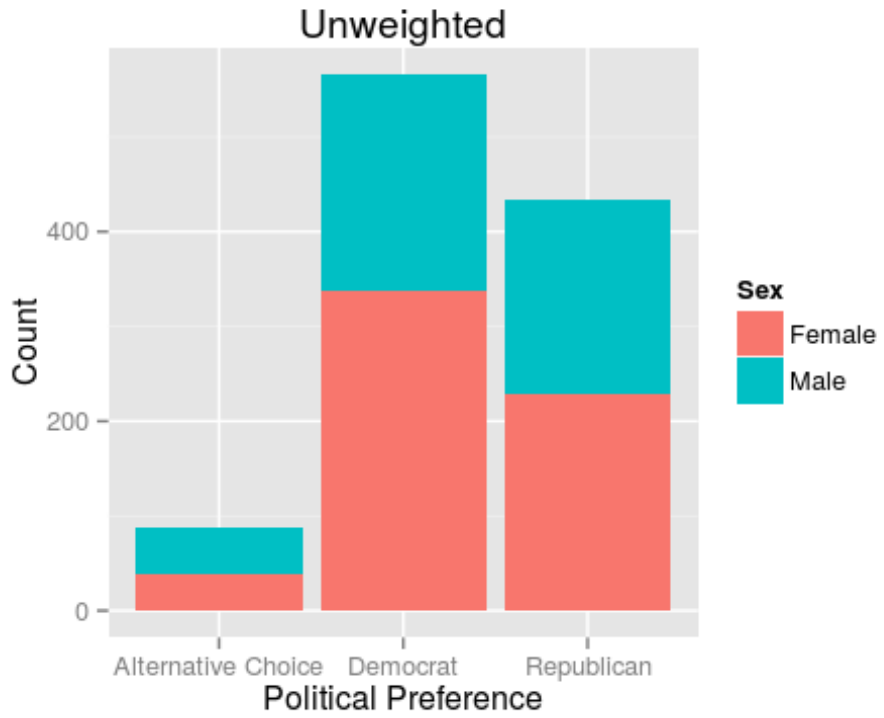
```
#Create a table of Sex by Preference
poliunweighted<-ftable(PoliticalData$Sex, PoliticalData$Preference)

#Convert unweighted table into data frame to be used with ggplot()
poliunweighteddata <- as.data.frame(poliunweighted)

#Assign the column names Race, Preference, and Counts to the data frame
colnames(poliunweighteddata) <- c("Sex", "Preference", "Count")

#Generate a barplot of the unweighted table created above
 ggplot(poliunweighteddata, aes(x = Preference, y = Count, fill = Sex)) +
     geom_bar(stat = "identity") +
```

```
  labs(x = "Political Preference", y = "Count", fill = "Sex") +
     ggtitle("Unweighted")
```



This graph shows the observed counts from the sample of Political Preference, colored by Sex.

Unweighted Proportion Table

Below is a table of the proportions of different Political Preference by Sex, as shown in the graph above.

```
##         Alternative Choice Democrat Republican
##
## Female                0.03     0.31       0.21
## Male                  0.05     0.21       0.19
```

Weighted Graph

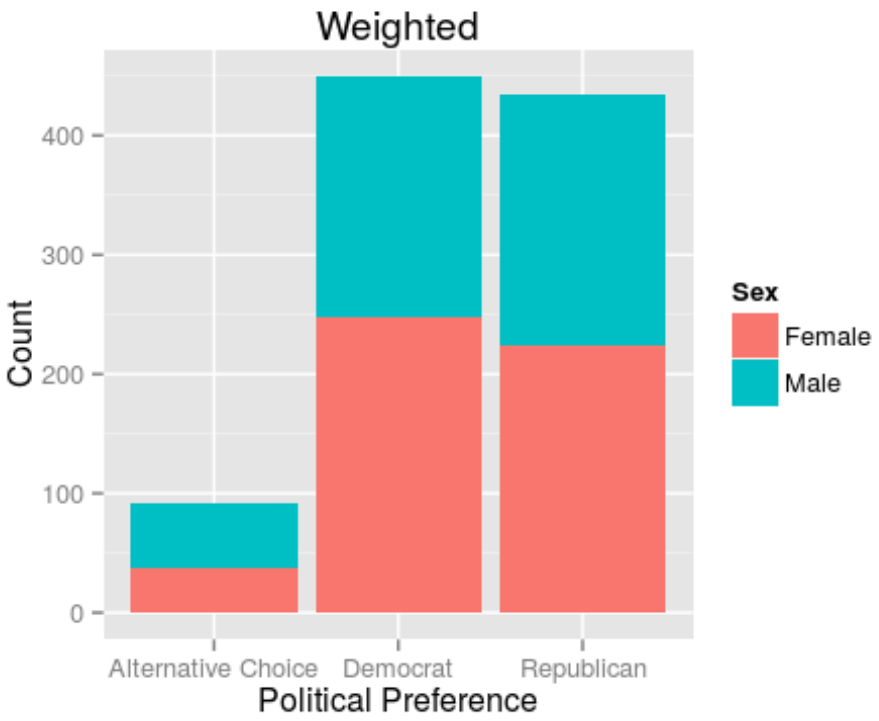
```
#Modify the data to be used in the {survey} package
polidesign <- svydesign(~0, data=PoliticalData, weights=PoliticalData$Weight)

#Create a table of the estimated weighted frequencies of the data
poliweighted <- svytable(~Sex + Preference, polidesign)

#Convert weighted table into a data frame to be used with ggplot()
poliweighteddata <- as.data.frame(poliweighted)

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

```r
#Generate a barplot of the weighted table created above.
ggplot(poliweighteddata, aes(x = Preference, y = Count, fill = Sex)) +
       geom_bar(stat = "identity") +
       labs(x = "Political Preference", y = "Count", fill = "Sex") +
       ggtitle("Weighted")
```



This graph shows the weighted observed frequencies (taking the weights into account) of Political Preference, again colored by Sex.

Weighted Proportion Table

Below is a table of the weighted proportions of different Political Preference by Sex, as shown in the graph above.

```
##         Preference
## Sex      Alternative Choice Democrat Republican
##   Female               0.04     0.25       0.23
##   Male                 0.06     0.21       0.22
```

As we can see, there is a difference between the weighted and unweighted graphs. Specifically, the number of Republican supporters increases when we take into account the weights. The weighted graph gives us a more accurate estimation of Political Preference by Sex in the population than the unweighted graph.

Subsetting

When interested in estimating parameters for a specific subset of the population it is ok to subset the data set as we would in a SRS. For example, if we wanted to examine differences in Political

Preference by Sex, looking only at subjects from the South, we could create a new data set containing only these entires:

```
PoliticalSouth <- PoliticalData[PoliticalData$Region == "South", ]
```

Using this subset, we could then create two graphs similar to those above. This method works fine for visualizating data, however when it comes to testing weighted data, subsetting becomes more complicated.

Note: See section on subsetting below for more detail

Testing with Weights

There are three common methods for conducting a chi-square test on survey 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. Below is the general equation used in calculating this chi-square test statistic

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

- The **Raw Weight (RW) Method** multiplies each entry by their respective weight and runs the analysis on this weighted sample. Below is the general equation used in calculating this chi-square test statistic

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

- The **Rao-Scott (RS) Method** takes into account both sampling variability and varibility among the assigned weights to adjust the chi-square from the RW method. This method also takes into account stratification and clustering elements of the survey design. Below is the general equation used in calculating this chi-square test statistic, where n is the sample size, N is the sum of the weights and D is a design correction

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

Note: For more information on this calculation, refer to the Rao-Scott supplement

Political Data

Let's look at the three methods for running a chi-square to test whether there is a significant difference in the proportion of males versus females that prefer various political parties.

SRS Two-Way Tables

Below are the tables of observed and expected counts used in the SRS chi-square test. The observed table corresponds to the unweighted graph we made earlier with the political data.

Observed Counts:

Political Preference	Male	Female	Row Total
Alternative Choice	50	38	88
Republican	205	228	433
Democrat	229	337	566
Column Total	484	603	1,087

Expected Counts:

Political Preference	Male	Female
Alternative Choice	39.18	48.82
Republican	192.8	240.2
Democrat	252	314

RW Two-Way Tables

Below are the tables of observed and expected weighted frequences used in the RW chi-square test. The weighted observed frequency table corresponds to the weighted graph we made earlier with the political data.

Weighted Observed Frequencies

Political Preference	Male	Female	Row Total
Alternative Choice	54.14	36.83	90.97
Republican	210.32	224.39	434.7
Democrat	202.01	247.59	449.6
Column Total	466.5	508.8	975.3

Weighted Expected Frequences

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

The Rao-Scott uses the same two tables as the RW method but, as mentioned earlier, the Rao-Scott test involves a few more calculations.

Test Output Comparison

Here is the R code to run each test, with a table of output comparisons below:

```
#-----------------------------SRS Method-----------------------
wtd.chi.sq(PoliticalData$Sex, PoltiicalData$Preference)

#-----------------------------RW Method-----------------------
wtd.chi.sq(PoliticalData$Sex, PoltiicalData$Preference,
weight=PoliticalData$Weight)
```

```r
#-----------------------------RS Method-----------------------

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

#Create a table of the estimated weighted frequencies of the data
poliweighted <- svytable(~Sex + Preference, polidesign)

summary(poliweighted, statistics="Chisq")
```

Test	χ^2	p-value	df	Comment
SRS	10.564	0.0051	2	Assumes that we are analyzing a simple random sample, and that this simple random sample is representative of the population
RW	6.543	0.0380	2	Multiplies each entry by their weight giving a slightly more representative sample. This method still assumes we are testing a SRS
RS	7.292	0.1673	2	Takes into account the variability and uncertainty involved in the sampling weights, giving us a more accurate result.

CAM Data

Let's look at the three methods for running a chi-square to test whether there is a significant difference in Chiropractor use by Region.

SRS Two-Way Tables

Below are the tables of observed and expected counts used in the SRS chi-square test:

Observed Counts:

Region	No	Yes	Row Total
Midwest	6,338	855	7,193
Northeast	5,355	419	5,774
South	11,846	690	12,536
West	8,201	821	9,022
Column Total	31,740	2,785	34,525

Expected Counts:

Region	No	Yes
Midwest	6,613	580.2
Northeast	5,308	465.8
South	11,525	1,011
West	8,294	727.8

RW Two-Way Tables

Below are the tables of observed and expected weighted frequences used in the RW and RS chi-square test:

Weighted Observed Frequences

Region	No	Yes	Row Total
Midwest	47,354,178	6,024,092	53,378,270
Northeast	39,545,300	3,214,503	42,759,803
South	80,569,430	50,08,310	85,577,740
West	48,083,035	5,121,822	53,204,857
Column Total	215,551,943	19,368,727	234,920,670

Weighted Expected Frequences

Region	No	Yes
Midwest	48,977,341	4,400,929
Northeast	39,234,345	35,25,458
South	78,522,031	7,055,709
West	48,818,226	4,386,631

Test Output Comparison

Here is the R code for running the three chi-square test on the CAM data, looking at region and chirpractor use. A table of test output comparisons is below:

```
#SRS Method
wtd.chi.sq(CAMData$chiropractor, CAMData$region)

#RW Method
wtd.chi.sq(CAMData$chiropractor, CAMData$region, weight=CAMData$weight)
```

For the RS method, first we must specify the survey design where we specify the weights, clustering and stratification used in data collection. See the clustering and stratification section for more detail.

```
#Specify survey design
camdesign <- svydesign(id=~psu, strata=~strat, nest=TRUE, data=CAMData,
weights=CAMData$weight)

#Create your two-way table
CAMweighted <- svytable(~region + chiropractor, camdesign)

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

Test	χ^2	p-value	df	Comment
SRS	270.629	<.0001	3	Assumes that we are analyzing a simple random sample, and that

				this simple random sample is representative of the population
RW	1464124	0	3	RW method multiplies each entry by their weight giving a slightly more representative sample. This method still assumes we are testing a SRS and greatly inflates the sample size used in the calculations
RS	215.17	<.0001	3	Takes into account the variability and uncertainty involved in the sampling weights as well as the clustering and stratification involved in the collection of the sample, giving us a more accurate result.

Subsetting

Even when working with subsetted data, it is essential that the full data set is taken into account when analyzing the data. Every entry in the data frame is given a weight that is dependent on the demographic make-up of the entire sample compared to the population, thus ignoring specific entries that aren't in the subpopulation of interest results in an incorrect analysis of the weights. We must specify the subpopulation we want to study in our analysis without subsetting the original data[1].

Below is an example of how to perform a Rao-Scott method chi-square test in R on a subset of the population, using the Political data set. We will be looking at only individuals who were from the West, comparing sex and political preference.

```
#-----------------------------Without taking the subset into account-------------
---------
#Subset the data to be just the individuals where use_type = Wellness
PoliticalSubset <- PoliticalData[PoliticalData$Region == "West", ]

#Specify survey design using the subsetted data set
inaccuratedesign <- svydesign(~0, data=PoliticalSubset,
weights=PoliticalSubset$Weight)

#Create your two-way table of region by chiropactor without subsetting being taken
into account
inaccuratetable <- svytable(~Sex + Preference, inaccuratedesign)

#-----------------------------Taking the subset into account--------------------
---------
#Specify survey design using the ORIGINAL data set
accuratedesign <- svydesign(~0, data=PoliticalData, weights=PoliticalData$Weight)

#Create your two-way table of region by chiropractor specifying your subset
accuratetable <- svytable(~Sex + Preference, subset(accuratedesign,
Region=="West"))
```

Outputs

```
#Run the Incorrect Test
summary(inaccuratetable, statistic = "Chisq")
```

```
##         Preference
## Sex      Alternative Choice Democrat Republican
##   Female                12       52         47
##   Male                   6       56         55
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~Sex + Preference, design = inaccuratedesign, statistic =
"Chisq")
## X-squared = 2.2114, df = 2, p-value = 0.4996
```

```
#Run the Correct Test
summary(accuratetable, statistic = "Chisq")
```

```
##         Preference
## Sex      Alternative Choice Democrat Republican
##   Female                12       52         47
##   Male                   6       56         55
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~Sex + Preference, design = subset(accuratedesign, Region ==
"West"), statistic = "Chisq")
## X-squared = 2.2114, df = 2, p-value = 0.4983
```

Note that there are slight differences between the p-values of the two chi-square tests. Although these differences don't impact our interpretation of these results, these differences can appear much larger and may have an impact on the significance of the result depending on your data set.

Strata & Clustering Issues

As we mentioned earlier, the CAM data set contains two columns, *psu* and *strat*. These columns indicate the clustering and stratification, respectively, used in the survey design. Both clustering and stratification affect the sampling error and variability within the sample. When working with a data set that specifies these two methods, it is important to factor them into our analysis to get accurate results.

Below is an example of how to take the clustering and stratification into account in a Rao-Scott chi-square test in R, comparing Physical Therapist use by Sex.

```
#-----------------------------Without Clustering and Stratification-------------
-----------
#Specify the survey design and weight column
nonclustereddesign <- svydesign(~0, data=CAMData, weights=CAMData$weight)

#Create the weighted table
nonclusteredtable <- svytable(~region + physical_therapist, nonclustereddesign)

#-----------------------------With Clustering and Stratification----------------
-----------
#Specify the survey design and weight column as well as the clustering and
stratification
```

```
clustereddesign <- svydesign(id=~psu, strata=~strat, nest=TRUE, data=CAMData,
weights=CAMData$weight)

#Create the weighted table
clusteredtable <- svytable(~region + physical_therapist, clustereddesign)
```

Test Outputs

```
#Evaluate both tables
summary(nonclusteredtable, statistic="Chisq")

##            physical_therapist
## region              0        1
##   Midwest   48246434  5131836
##   Northeast 38999272  3760531
##   South     79062876  6514864
##   West      48628210  4576647
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~region + physical_therapist, design = nonclustereddesign,
statistic = "Chisq")
## X-squared = 26.033, df = 3, p-value = 0.001639

summary(clusteredtable, statistic="Chisq")

##            physical_therapist
## region              0        1
##   Midwest   48246434  5131836
##   Northeast 38999272  3760531
##   South     79062876  6514864
##   West      48628210  4576647
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~region + physical_therapist, design = clustereddesign,
statistic = "Chisq")
## X-squared = 26.033, df = 3, p-value = 0.003282
```

Note that the clustering and stratification affect the values we get. Without the clustering and stratification specified, we geta p-value of .001639. When we take the clustering and stratification into account we get a p-value of .003282, a slightly different but much more appropriate value.

Endnotes

[1] The Carolina Population Center at UNC has more information on common erros when working with survey data. [http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial/sample_surveys/svy_errors]

References:

University of North Carolina Chapel Hill. Common Errors and How to Avoid Them. [http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial/sample_surveys/svy_errors]

Maletta, H. (12 March, 2007). Weighting. *Universidad del Salvador*.