Rao-Scott Supplement

The Rao-Scott correction for weighted chi-square tests can be used to obtain a more accurate chi-square value than traditional SRS or RW analysis. To begin the Rao-Scott correction calculation, we must calculate the raw weight Pearson chi-square test statistic, χ^2_{RW}. We then calculate the design correction used to account for sampling and weight variability, D. Additionally, we need to multiply the raw weight test statistic by the sample-population ratio of $\frac{n}{\hat{N}}$, where n is the sample size and $\hat{N}$ is the sum of the individual weights assigned to each entry. Then, the Rao-Scott chi-square test statistic, χ^2_{RS}, can be expressed as

$$\chi^2_{RS} = \frac{n}{\hat{N}D} \cdot \chi^2_{RW}$$

As in any chi-square test, we will use summary tables of observed and expected values to perform our analysis. Using the **Political Preferences** dataset[1], we will test whether there is a significant difference in the proportion of males versus females that prefer various political parties. The observed summary table of the sample data is shown below:

```
##
##          Alternative Democrat Republican Total
##   Female          38      337        228   603
##   Male            50      229        205   484
##   Total           88      566        433  1087
```

To calculate the raw weight chi-square, we must first find the expected values of our summary table. To do this, we need to determine:

1. the indicator columns

2. the observed table of weights associated with the data set, and

3. the corresponding expected values

Pearson Chi-Square Calculation for Raw Weights

Step 1: Indicator Columns

An indicator variable is simply a column of 1's and 0's used to indicate whether or not a particular condition is true.

Once the data has been read into the console, use the code below to create the indicator variables for each level of the column variable (Preference) $\delta_i(\cdot\ c)$ where c=1,2,3 and each level of the row variables Sex, $\delta_i(r\ \cdot)$ where r= 1,2.

```
delta_r<-model.matrix(~Sex +0, data=Political)
delta_c<-model.matrix(~Preference +0, data=Political)
delta_rc<-model.matrix(~Sex:Preference +0, data=Political)
Political <- cbind(Political, delta_r, delta_c, delta_rc)

#Note that delta_.1 is "Alternative", delta_.2 is "Democrat", and delta_.3 is
```

```
"Republican"
#Additionally, delta_1. is "Female" and delta_2. is "Male"
colnames(Political)[10:20] <- c("delta_1.", "delta_2.", "delta_.1", "delta_.2",
"delta_.3", "delta_11", "delta_21", "delta_12", "delta_22", "delta_13",
"delta_23")
```

For r = 1, 2 and c = 1,2,3

```
Political$delta_rc <- Political$delta_r. * Political$delta_.c
```

Once we have completed all of the delta code we should have a total of 11 indicator columns.

```
head(Political)
```

```
##   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
## 6   6 6  0.392 African Am. Female   > 75 56-65        South   Democrat
##   delta_1. delta_2. delta_.1 delta_.2 delta_.3 delta_11 delta_21 delta_12
## 1        1        0        0        0        1        0        0        0
## 2        0        1        0        1        0        0        0        0
## 3        1        0        0        1        0        0        0        1
## 4        0        1        0        0        1        0        0        0
## 5        0        1        0        0        1        0        0        0
## 6        1        0        0        1        0        0        0        1
##   delta_22 delta_13 delta_23
## 1        0        1        0
## 2        1        0        0
## 3        0        0        0
## 4        0        0        1
## 5        0        0        1
## 6        0        0        0
```

Notice that for r=1 (Sex=Female), $\delta_i(1\cdot) = 1$ when Sex=Female and $\delta_i(1\cdot) = 0$ when Sex=Male. Similarly, when c=2 (Preference=Democrat), $\delta_i(\cdot 2) = 1$ when Preference=Democrat. Finally, for r=1 and c=2 (Sex=Female and Preference=Democrat), $\delta_i(rc) = 1$ when Sex=Female and Preference=Democrat, and $\delta_i(rc) = 0$ otherwise.

Step 2: Weighted Frequency Sums

Next, we need to calculate the $\hat{N}$ values, the weighted frequency values for each row, column and row-column combination. This is calculated by multiplying the δ_i entry for each individual subject i by that entry's assigned weight W_i, as shown in the equations below.

$$\hat{N}_{rc} = \sum_{i=1}^{n} \delta_i\,(r, c) \cdot W_i$$

Sex	Alternative (c=1)	Democrat (c=2)	Republican (c=3)	Total
Female (r=1)	$\hat{N}_{11}$	$\hat{N}_{12}$	$\hat{N}_{13}$	$\hat{N}_{1\cdot}$

Male (r=2)	$\hat{N}_{21}$	$\hat{N}_{22}$	$\hat{N}_{23}$	$\hat{N}_{2\cdot}$
Total	$\hat{N}_{\cdot 1}$	$\hat{N}_{\cdot 2}$	$\hat{N}_{\cdot 3}$	$\hat{N}$

So, for Male Republicans (r=2 and c=3)

$$\hat{N}_{23} = \sum_{i=1}^{n} \delta_i\,(2,3) \cdot W_i$$

In R:

```
Nhat_23 <- sum(Political$delta_23 * Political$Weight)
Nhat_23
```

```
## [1] 210.322
```

So for our political data set, the table of observed raw weights looks like:

```
PoliTable <- wtd.table(Political$Sex, Political$Preference,
weights=Political$Weight)

PoliticalTable <- addmargins(PoliTable, FUN = list(Total = sum), quiet = TRUE)

PoliticalTable
```

```
##        Alternative Democrat Republican   Total
## Female      36.834  247.589    224.387 508.810
## Male        54.140  202.008    210.322 466.470
## Total       90.974  449.597    434.709 975.280
```

Notice that in the table, the Male row, Republican column weighted observed value is 210.32, the same value we found above.

Step 3: Expected Weighted Frequencies

Now that we have a table of observed weighted frequencies, we can calculate the expected weighted frequencies for each cell. Just as in the traditional Pearson chi-square test, expected values are calculated as the row total times the column total divided by the overall total.

The E values can be obtained using the equation below.

$$E_{rc} = \frac{\hat{N}_{r\cdot} \cdot \hat{N}_{\cdot c}}{\hat{N}}$$

Notice that this equation is the same equation we would use in calculating the expected values in a standard chi-square test.

Raw Weight Test Statistic

The raw-weights chi-square test statistic is calculated by[2]:

$$\chi^2_{RW} = \sum r \sum c \frac{(\hat{N}_{rc} - E_{rc})^2}{E_{rc}}$$

The expected table and chi-square output from R are printed below:

```
##        Alternative Democrat Republican
## Female    47.46174 234.5577   226.7905
## Male      43.51226 215.0393   207.9185

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

##      Chisq         df    p.value
## 6.54250166 2.00000000 0.03795892
```

Calculating the Design Effect

After calculating χ^2_{RW}, the next step in determining the Rao-Scott correction is to calculate a value for D, the design correction used in the Rao-Scott chi-square equation. To find D, we need to

1. Find the estimated proportions $\hat{P}$ for each row and column combination
2. Calculate the estimated variance $\hat{Var}$ of the $\hat{P}$ values
3. Find the variance values for each row and column combination assuming an SRS, $\hat{Var}_{SRS}$
4. Calculate the design effects, DEFF, for each row and column combination

Step 1: Estimated Proportions

First, we need to calculate $\hat{P}$, the estimated proportions for each row and column combination. This is calculated by

$$\hat{P}_{rc} = \frac{\hat{N}_{rc}}{\hat{N}}$$

Sex	Alternative (c=1)	Democrat (c=2)	Republican (c=3)	Total
Female (r=1)	$\hat{P}_{11}$	$\hat{P}_{12}$	$\hat{P}_{13}$	$\hat{P}_{1\cdot}$
Male (r=2)	$\hat{P}_{21}$	$\hat{P}_{22}$	$\hat{P}_{23}$	$\hat{P}_{2\cdot}$
Total	$\hat{P}_{\cdot 1}$	$\hat{P}_{\cdot 2}$	$\hat{P}_{\cdot 3}$	

So, for Male Republicans (r=2 and c=3)

$$\hat{P}_{23} = \frac{\hat{N}_{23}}{\hat{N}}$$

In R:

```
Phat_23 <- Nhat_23/975.3
Phat_23

## [1] 0.2156485
```

So for our political data set, the table of estimated proportions looks like:

```
##        Alternative Democrat Republican  Total
## Female      0.0378   0.2539     0.2301 0.5217
## Male        0.0555   0.2071     0.2156 0.4783
## Total       0.0933   0.4610     0.4457 0.0000
```

Notice that in the table, the Male row, Republican column estimated proportion is 0.2156, the same value we found above.

Step 2: Estimated Variances

Once we have calculated the $\hat{P}$s, the estimated individual variances, the e^is and the estimated group variances, the $\bar{e}$s must be approximated in order to find the total estimated variance ($\hat{Var}$) of the $\hat{P}$s for each row and column combination.

To calculate the e^i columns, we use the formula:

$$e^i_{rc} = \frac{(\delta_i(r,c) - \hat{P}_{rc}) \cdot W_i}{\hat{N}}$$

To create the Male Republican (r=2, c=3) e^i column in R:

```
Political$e_23 <- ((Political$delta_23-Phat_23)*Political$Weight)/975.3

head(Political)
```

```
##   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
## 6   6 6  0.392 African Am. Female    > 75 56-65        South   Democrat
##   delta_1. delta_2. delta_.1 delta_.2 delta_.3 delta_11 delta_21 delta_12
## 1        1        0        0        0        1        0        0        0
## 2        0        1        0        1        0        0        0        0
## 3        1        0        0        1        0        0        0        1
## 4        0        1        0        0        1        0        0        0
## 5        0        1        0        0        1        0        0        0
## 6        1        0        0        1        0        0        0        1
##   delta_22 delta_13 delta_23          e_1.          e_2.          e_.1
## 1        0        1        0  0.0013251079 -0.0013250511 -2.584200e-04
## 2        1        0        0 -0.0002535464  0.0002535564 -4.533349e-05
## 3        0        0        0  0.0003020971 -0.0003020842 -5.891441e-05
## 4        0        0        1 -0.0003669470  0.0003669614 -6.560923e-05
## 5        0        0        1 -0.0002449879  0.0002449975 -4.380325e-05
## 6        0        0        0  0.0001922436 -0.0001922354 -3.749099e-05
##            e_.2          e_.3          e_11          e_12          e_13
## 1 -0.0012771217  0.0015355986 -1.046304e-04 -0.0007032994  2.133038e-03
## 2  0.0002619644 -0.0002166210 -1.835485e-05 -0.0001233767 -1.118149e-04
## 3  0.0003404432 -0.0002815159 -2.385356e-05  0.0004712628 -1.453122e-04
## 4 -0.0003242433  0.0003898670 -2.656419e-05 -0.0001785579 -1.618249e-04
## 5 -0.0002164773  0.0002602902 -1.773527e-05 -0.0001192121 -1.080405e-04
## 6  0.0002166457 -0.0001791465 -1.517954e-05  0.0002998945 -9.247137e-05
```

```
##            e_21          e_22          e_23
## 1 -1.537897e-04 -5.738224e-04 -5.974390e-04
## 2 -2.697865e-05  3.853412e-04 -1.048061e-04
## 3 -3.506086e-05 -1.308196e-04 -1.362037e-04
## 4 -3.904504e-05 -1.456855e-04  5.516919e-04
## 5 -2.606797e-05 -9.726523e-05  3.683307e-04
## 6 -2.231145e-05 -8.324884e-05 -8.667509e-05
```

Next, we need to calculate the $\bar{e}$s for every row-column combination by using the equation:

$$\bar{e}_{rc} = \sum_{i=1}^{n} \frac{e_{rc}^{i}}{n}$$

For the Male Republican $\bar{e}$ value

$$\bar{e}_{23} = \sum_{i=1}^{n} \frac{e_{23}^{i}}{n}$$

In R:

```
ebar_23 <- sum(Political$e_23)/1087
ebar_23
```

```
## [1] 4.06826e-09
```

To view all of the $\bar{e}$ values:

```
##             Alternative     Democrat   Republican          Sex
## Female     7.124803e-10 4.789116e-09 4.340319e-09 9.841916e-09
## Male       1.047230e-09 3.907442e-09 4.068260e-09 9.022933e-09
## Preference 1.759711e-09 8.696558e-09 8.408580e-09 0.000000e+00
```

After we have calculated all of the $\bar{e}$ values, we can calculate the estimated variances using:

$$\hat{Var}(\hat{P}_{rc}) = \frac{n}{n-1} \cdot \sum_{i=1}^{n} (e_{rc}^{i} - \bar{e}_{rc})^2$$

For the Male Republican value:

$$\hat{Var}(\hat{P}_{23}) = \frac{n}{n-1} \cdot \sum_{i=1}^{n} (e_{23}^{i} - \bar{e}_{23})^2$$

Or, in R:

```
Var_23 <- 1087/1086 * sum((Political$e_23 - ebar_23)^2)
Var_23
```

```
## [1] 0.0003040566
```

To see all of the $\hat{Var}$ values:

```
##             Alternative     Democrat   Republican          Sex
## Female     6.087287e-05 0.0003306836 0.0003037697 0.0004464314
## Male       1.326370e-04 0.0002877114 0.0003040566 0.0004464331
## Preference 1.831738e-04 0.0004410630 0.0004383805 0.0000000000
```

Step 3: Simple Random Sample Variances

After calculating the $\hat{\text{Var}}$ values, we next need to calculate the estimated variance of each of the row-column combinations assuming the sample is an SRS. These values are referred to as $\hat{\text{Var}}_{\text{SRS}}$ and the equation for finding these values is:

$$\hat{\text{Var}}_{\text{SRS}}(\hat{P}_{rc}) = \frac{\hat{P}_{rc} \cdot (1 - \hat{P}_{rc})}{n - 1}$$

For example, the Male Republican $\hat{\text{Var}}_{\text{SRS}}(\hat{P}_{23})$ can be calculated as:

$$\hat{\text{Var}}_{\text{SRS}}(\hat{P}_{23}) = \frac{\hat{P}_{23} \cdot (1 - \hat{P}_{23})}{n - 1}$$

In R:

```
Varsrs_23 <- (Phat_23 * (1 - Phat_23))/1086
Varsrs_23
```

```
## [1] 0.0001557498
```

To see all of the $\hat{\text{Var}}_{\text{SRS}}$ values:

```
##             Alternative     Democrat   Republican          Sex
## Female     3.346271e-05 0.0001744151 0.0001631102 0.0002297691
## Male       4.827775e-05 0.0001512188 0.0001557498 0.0002297683
## Preference 7.787955e-05 0.0002288008 0.0002274894 0.0000000000
```

Step 4: Design Effect for Each Row-Column Combination

The second to last step in calculating D is to determine the design effect (DEFF) for each row-column combination. The equation below demonstrates how to find these values.

$$\text{DEFF}(\hat{P}_{rc}) = \frac{\hat{\text{Var}}(\hat{P}_{rc})}{\hat{\text{Var}}_{\text{SRS}}(\hat{P}_{rc})}$$

For the Male Republican value

$$\text{DEFF}(\hat{P}_{23}) = \frac{\hat{\text{Var}}(\hat{P}_{23})}{\hat{\text{Var}}_{\text{SRS}}(\hat{P}_{23})}$$

```
deff_23 <- Var_23/Varsrs_23
deff_23
```

```
## [1] 1.952212
```

To see all of the DEFF values:

```
##            Alternative Democrat Republican      Sex
## Female        1.819125 1.895958   1.862359 1.942956
## Male          2.747374 1.902616   1.952212 1.942970
## Preference    2.352015 1.927716   1.927037 0.000000
```

Design Effect

The final step in calculating the design effect D is to use the equation:

$$D = \frac{\sum_{r}\sum_{c}(1-\hat{P}_{rc})DEFF(\hat{P}_{rc}) - \sum_{r}(1-\hat{P}_{r\cdot})DEFF(\hat{P}_{r\cdot}) - \sum_{c}(1-\hat{P}_{\cdot c})DEFF(\hat{P}_{\cdot c})}{(R-1)(C-1)}$$

In R:

```
D <-((((1-Phat_11)*deff_11)+((1-Phat_12)*deff_12)+((1-Phat_13)*deff_13)+((1-
Phat_21)*deff_21)+((1-Phat_22)*deff_22)+((1-Phat_23)*deff_23))-(((1-
Phat_1.)*deff_1.)+((1-Phat_2.)*deff_2.))-(((1-Phat_.1)*deff_.1)+((1-
Phat_.2)*deff_.2)+((1-Phat_.3)*deff_.3)))/2

D
```

```
## [1] 2.025382
```

Rao-Scott Chi-Square Value

The last step in calculating the Rao-Scott correction is to divide the χ^2_{RW} calculated earlier by the D we just calculated. Below, we restate the original equation given at the beginning of the supplement.

$$\chi^2_{RS} = \frac{n}{\hat{N}D} \cdot \chi^2_{RW}$$

This equation gives us:

```
Q_RW <- 6.54250166
Q_RS<- (1087/975.3)*Q_RW/D
Q_RS
```

```
## [1] 3.600213
```

Additional Information

Now that we have calculated the first-order Rao-Scott correction, it is important to note that there are several other key concepts to keep in mind when testing weighted data [2].

To begin, although the first-order Rao-Scott correction provides a more accurate test statistic than other types of survey data analysis, the second-order correction improves this test statistic even more. Although the calculations involved in determining this correction will not be included in this supplement, it may be worth continued investigation to examine the details behind how this correction is calculated. Fortunately, many statistics packages are now capable of computing the second order correction, so calculations by hand are not absolutely necessary.

In addition, the Rao-Scott correction explained in this supplement does not include adjustments for stratification and clustering. If a data set relies on either of these survey methods, they must be taken into account when calculating the Rao-Scott. Again, this is a complicated process to complete by hand, but many statistical packages available today are able to effectively address this issue.

End Notes

[1] The Political Preference data set that we will use is a 2010 national survey conducted by CBS and the New York Times. The survey (conducted via phone) involved 1189 subjects (Note: due to non-response, the data set that we will be analyzing has 1087 subjects) who were asked for their opinion on "a range of political and social issues" such as their political party preference. In addition, information about race, sex, age, and region of residence was also collected. There was no stratification or clustering used, and individuals were assigned weights based on age, sex, race, education and region.

[2] Due to additional nonresponse and oversampling error, the {survey} package makes additional adjustments to the p-value after calculating the Chi-Square test statistic.

[2] Some packages use this equation and some use a variation of it.

References:

Scott, A. (2007). Rao-Scott corrections and their impact. In Proceedings of the 2007 joint statistical meetings, Salt Lake City, Utah.

SAS (2015) User's Guide. Rao-Scott Chi-Square Test.