Surveillance Research Program, NCI, Technical Report #2020-01

Fay's Variance Estimation Method for Combining Multiple TUS-CPS Data

Benmei Liu

Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute

Introduction

Let $\hat{\theta}$ denote the point estimate of an outcome of interest calculated from the Tobacco Use supplement to the Current Population Survey (TUS-CPS). The associated variance can be estimated using Fay's variance estimation method:

$$V_{Fay}(\hat{\theta}) = c\sum_{g=1}^{G}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)^2,$$

where $c = \frac{1}{G(1-K)^2}$, 1/(1-K)^2 is the Fay adjustment factor, G is the number of replicate weights.

The original Fay-adjustment factor is 4 (K=0.5) for each single cycle of TUS-CPS data. However, when multiple cycles of TUS-CPS data are to be combined, the weights in the combined data file need to be adjusted and the Fay-adjustment factor need to be changed so the variance for each data cycle before and after the combination stays the same. The following example illustrates how to adjust the weights and the Fay-adjustment factor to satisfy this criterion.

Combining 1992-1993, 2003 and 2006/2007 data illustration

Let's say we want to combine the 1992-93, 2003 and 2006/2007 data. The 1992-93 data has 48 replicates, 2003 TUS-CPS data has 80 replicate weight and the 2006-2007 data has 160 replicate weights. The total new replicate weights in the combined file would be: 48+80+160=288.

Using the 1992-93 data alone (3 months data), the variance for $\hat{\theta}$ is:

$$v_{1992/93}(\hat{\theta}) = \frac{4}{48}\sum_{g=1}^{48}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)^2. \qquad (1)$$

Using the 2003 data alone (3 months data), the variance for $\hat{\theta}$ is:

$$v_{2003}(\hat{\theta}) = \frac{4}{80}\sum_{g=1}^{80}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)^2. \qquad (2)$$

Using the 2006/07 data alone (3 months data), the variance for Let $\hat{\theta}$ is:

$$v_{2006/07}(\hat{\theta}) = \frac{4}{160}\sum_{g=1}^{160}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)^2. \qquad (3)$$

After combining the 1992/93, 2003 and the 2006/07 data, the variance of $\hat{\theta}$ for the subset of 1992/93 data should stay the same as (1), for the 2003 data should stay the same as (2) and for the 2006/07 data should stay the same as (3). To satisfies this, the new replicate weights for the first 48 replicate weights in the combined data should be $\hat{\theta}_0 + \frac{1}{2} \times \sqrt{\frac{288}{48}}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)$ so the variance stays the same as before, i.e.:

$$\frac{4 * 2^2}{288} \times \sum_{g=1}^{48}\left(\hat{\theta}_0 + \frac{1}{2} \times \sqrt{\frac{288}{48}}(\hat{\theta}_{(g)} - \hat{\theta}_0) - \hat{\theta}_0\right)^2 = \frac{4}{48}\sum_{g=1}^{48}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)^2;$$

The new replicate weights for replicates 49 to 128 in the combined data should be $\hat{\theta}_0 + \frac{1}{2} \times \sqrt{\frac{288}{80}}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)$ so the variance stays the same as before:

$$\frac{4 * 2^2}{240} \times \sum_{g=49}^{128}\left(\hat{\theta}_0 + \frac{1}{2} \times \sqrt{\frac{288}{80}}(\hat{\theta}_{(g)} - \hat{\theta}_0) - \hat{\theta}_0\right)^2 = \frac{4}{80}\sum_{g=49}^{128}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)^2;$$

The new replicate weights for the replicates 129 to 288 should be $\hat{\theta}_0 + \frac{1}{2} \times \sqrt{\frac{288}{160}}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)$ so the variance stays the same as before:

$$\frac{4 * 2^2}{288} \times \sum_{g=129}^{288}\left(\hat{\theta}_0 + \frac{1}{2} \times \sqrt{\frac{288}{160}}(\hat{\theta}_{(g)} - \hat{\theta}_0) - \hat{\theta}_0\right)^2 = \frac{4}{160}\sum_{g=129}^{288}\left(\hat{\theta}_{(g)} - \hat{\theta}_0\right)^2$$

The rest replicate weights from each data cycle will be equal to the full sample weights. The factor of $\frac{4*2^2}{240}$ corresponds to Fay-adjustment factor of 16.

When TUS-CPS data cycles with the same number of replicate weights are to be combined (e.g., 2006/07, 2010/11, 2014/15, 2018/19), the data can just be stacked together, there is no need to adjust the weights except dividing the weights by the number of data months to be combined. When multiple years of data with different number of replicate weights to be combined, the new weights need to be adjusted using similar approach as the above. The new Fay-adjustment factor will be 16 regardless when two or more years of data with different number of replicates weights are to be combined.

Sample SAS codes of combining 1992/93, 2003 and 2006/2007 data

```
Data CPS9207(Drop=I J RepWt001-RepWt288 SmplWgt);
  Set CPS9207;                                   /*Assuming CPS9207 is a data
set that combines 1992/93, 2003, and 2006/2007 data together*/
  Array OldR(160) RepWt001-RepWt160;
 Array NewR(288) NWgt001-NWgt288;
  NSmplWgt=SmplWgt/9;

If SurvGrp=1 Then Do;   /*survey year 1992-93*/
    Do I = 1 to 48;
      NewR(I)=(1/9) * (SmplWgt+(1.22475*(OldR(I)-SmplWgt)));  /* 1.22475 =
1/2 x (Sqrt(288/48)) */
      End;
    Do I = 49 to 288;
      NewR(I)=SmplWgt/9;
      End;
    End;

  If SurvGrp=2 Then Do;   /*survey year 2003*/
   Do I = 1 to 48;
      NewR(I)=SmplWgt/9;
      End;
 Do I = 49 to 128;
      J=I-48;
      NewR(I)=(1/9) * (SmplWgt+(0.948683*(OldR(J)-SmplWgt)));  /* 0.948683 =
1/2 x (Sqrt(288/80)) */
      End;
    Do I = 129 to 288;
      NewR(I)=SmplWgt/9;
      End;
    End;
Else if SurvGrp=3 Then Do;   /*survey year 2006/2007*/
Do I = 1 to 128;
     NewR(I)=SmplWgt/9;
      End;
Do I=129 to 288;
  J=I-128;
      NewR(I)=(1/9) * (SmplWgt+(0. 0.67082 *(OldR(J)-SmplWgt)));  /* 0.67082
= 1/2 x (Sqrt(288/160)) */
      End;
    End;
Run;
```