Social contact data analysis: participant weights

Lander Willem[1], Andrea Torneri[1], Niel Hens[1,2]

[1]University of Antwerp & [2]Hasselt University

December 21th, 2020

In the context of (social contact) surveys, **participant weights** have been commonly used to align sample and population characteristics. For example, weights have been used to define the contribution of participants if the age-distribution in the survey sample differs from the age-distribution in the population. These **age-specific weights** can be calculated as:

$$w_{age} = \frac{P_a/P}{N_a/N},$$

or, if we remove the constant values:

$$\dot{w}_{age} = \frac{P_a}{N_a}$$

with P the population size, P_a the population fraction of age a, N the survey sample size and N_a the survey fraction of age a. Both weighting methods result in the same standardized weights ($\tilde{w}$, see below), though the advantage of the first method is that we can limit the influence of single participants by a general truncation of the relative differences. If we for example **truncate** w at 3, we account for relative differences in the population and survey proportions up to 3 in the participant weights. With the second approach ($\dot{w}$) the range of the weights depends on P_a and N_a, hence a generic cutoff is not feasible. Therefore, we continue here with the first approach.

Temporal effects such as day of the week, have also been reported as a driving factor for social contact behavior. As such, this can also be included in participant weights to account for differences in the survey sample and the weekly 2/5 distribution of weekend/week days. This is represented as

$$w_{day.of.week} = \frac{5/7}{N_{weekday}/N} \text{ OR } \frac{2/7}{N_{weekend}/N}$$

The **combination** of age-specific and temporal weights for participant i of age a can be constructed as:

$$w_i = w_{age} * w_{day.of.week}$$

To use these weights, **standardization** is performed as follows:

$$\tilde{w}_i = \frac{w_i}{\sum w} * N$$

Participant weights within (age) groups

Social contact analyses are commonly based on (age) stratification, which works by splitting the population into non-overlapping groups (strata), with the purpose of drawing samples independently between the (age) groups. **Post-stratification weights** have to be standardized so that the weighted totals within mutually exclusive cells equal the known population totals (Kolenikov 2016). Post-stratification relies on data obtained in the survey itself that were not available before sampling, and adjusts the weights so that the totals in each group are equal to the known population totals. It needs the post-stratification cells to be mutually exclusive and cover the whole population. The **post-stratified (PS) weight** for participant i of group g is:

$$\tilde{w}_i^{PS} = \frac{w_i}{\sum_{\text{j}}^{\text{group g}} w_j} * N_g$$

All weights within a group are re-scaled proportionately so that the sum of post-stratified weights equals the known population total.

Reference: Kolenikov. 2016. Post-stratification or non-response adjustment. Survey Practice. 9(3) p112

Numerical example

The goal is to calculate a weighted average number of contacts by accounting for age and day of week with respect to a uniform population. We will apply the weights by day of week and age separately, though the combination is straightforward via multiplication. With this numeric example, we show the importance of post-stratification weights in contrast to using the crude weights directly within age-groups.

Survey data

We start from a survey including 6 participants of 1, 2 and 3 years of age. The ages are not equally represented in the sample, though we assume they are equally present in the reference population. We examine the weighted average number of contacts by age and by age group, using [1,2] and [3] years of age.

```
##   age day_type age_group cnt_total
## 1   1  weekend         A         3
## 2   1  weekend         A         2
## 3   2  weekend         A         9
## 4   2     week         A        10
## 5   2     week         A         8
## 6   3     week         B        15
```

Summary statistics for the sample (N) and reference population (P):

```
N             = 6
N_age         = c(2,3,1)
N_age.group   = c( 5 ,1)
N_day.of.week = c(3,3)

P             = 3000
P_age         = c(1000,1000,1000)
P_age.group   = c( 2000 ,1000)

P_day.of.week = c(5/7,2/7)*3000
```

Unweighted average number of contacts

```
## [1] "unweighted population average: 7.83"
```

The age-specific unweighted average number of contacts:

```
##   age age_group cnt_total
## 1   1         A       2.5
## 2   2         A       9.0
## 3   3         B      15.0
```

Weight by day of week

If we calculate the weights by including the population and sample size ("w") or excluding the constants ("w_dot"), we obtain the same standardized weights ("w_tilde" and "w_dot_tilde"):

```
##   age day_type age_group cnt_total    w w_tilde  w_dot w_dot_tilde
## 1   1  weekend         A         3 0.57    0.57 285.71        0.57
## 2   1  weekend         A         2 0.57    0.57 285.71        0.57
## 3   2  weekend         A         9 0.57    0.57 285.71        0.57
## 4   2     week         A        10 1.43    1.43 714.29        1.43
## 5   2     week         A         8 1.43    1.43 714.29        1.43
## 6   3     week         B        15 1.43    1.43 714.29        1.43
```

```
## [1] "weighted population average: 9.2"
```

Note the different scale of w and $\dot{w}$, and the more straightforward interpretation of the numerical value of w in terms of relative differences to apply truncation.

Using age groups (or age strata)

If the population-based weights are directly used for single-year age groups, the contact behavior of the 3 year-old participant, which participated during week day, is inflated to due the under-representation of week days in the survey sample. In addition, the number of contacts for participants of 1 year of age is decreased because of the over representation of weekend days in the survey. Using the population-weights within the two predefined age groups, we obtain a more intuitive weighting for age group A, but it is still affected by individuals in age group B. In addition, the weighted average for age group B has no meaning in terms of social contact behavior:

```
##   age cnt_total * w_tilde
## 1   1               1.425
## 2   2              10.290
## 3   3              21.450
```

```
##   age_group cnt_total * w_tilde
## 1         A               6.744
## 2         B              21.450
```

As such, we need to use post-stratification weights ("w_PS") in which the weighted totals within mutually exclusive cells equal the sample size.

```
##   age day_type age_group cnt_total    w w_tilde w_PS
## 1   1  weekend         A         3 0.57    0.57 0.62
## 2   1  weekend         A         2 0.57    0.57 0.62
## 3   2  weekend         A         9 0.57    0.57 0.62
## 4   2     week         A        10 1.43    1.43 1.56
## 5   2     week         A         8 1.43    1.43 1.56
## 6   3     week         B        15 1.43    1.43 1.00
```

The weighted means equal:

```
##   age_group cnt_total * w_PS
## 1         A            7.352
## 2         B           15.000
```

Age-specific weights

We repeated the example by calculating age-specific participant weights on the population and age-group level:

```
##   age day_type age_group cnt_total    w w_tilde w_PS
## 1   1  weekend         A         3 1.00    1.00 1.25
## 2   1  weekend         A         2 1.00    1.00 1.25
## 3   2  weekend         A         9 0.67    0.67 0.83
## 4   2     week         A        10 0.67    0.67 0.83
## 5   2     week         A         8 0.67    0.67 0.83
## 6   3     week         B        15 2.00    2.00 1.00
```

```
## [1] "weighted population average: 8.85"
```

Using age-groups

If the age-specific weights are directly used for the age groups, the contact behavior of people of 3 years of age is inflated to unrealistic levels and the number of contacts for age group 2 is reduced, although we use single year age groups:

```
##   age cnt_total * w_tilde
## 1   1               2.50
## 2   2               6.03
## 3   3              30.00
```

```
##   age_group cnt_total * w_tilde
## 1         A               4.618
## 2         B              30.000
```

Using the post-stratification weights, we end up with:

```
##   age_group cnt_total * w_PS
## 1         A            5.732
## 2         B           15.000
```

Truncation

Less balanced survey data

If the survey data contains more participants of 1 year of age, the difference in relative proportions for participants of 2 and 3 years of age increased. The age-specific participant weights on the population and age-group level are calculated as follows:

```
N             = nrow(survey_data)
N_age         = table(survey_data$age)
N_age.group   = table(survey_data$age_group)
N_day.of.week = table(survey_data$day_type)
```

```
##    age day_type age_group cnt_total    w w_tilde   w_dot w_dot_tilde
## 1    1  weekend         A         3 0.47    0.47  100.00        0.47
## 2    1  weekend         A         2 0.47    0.47  100.00        0.47
## 3    1  weekend         A         3 0.47    0.47  100.00        0.47
## 4    1  weekend         A         2 0.47    0.47  100.00        0.47
## 5    1  weekend         A         3 0.47    0.47  100.00        0.47
## 6    1  weekend         A         2 0.47    0.47  100.00        0.47
## 7    1  weekend         A         3 0.47    0.47  100.00        0.47
## 8    1  weekend         A         2 0.47    0.47  100.00        0.47
## 9    1  weekend         A         3 0.47    0.47  100.00        0.47
## 10   1  weekend         A         2 0.47    0.47  100.00        0.47
## 11   2  weekend         A         9 1.56    1.56  333.33        1.56
## 12   2     week         A        10 1.56    1.56  333.33        1.56
## 13   2     week         A         8 1.56    1.56  333.33        1.56
## 14   3     week         B        30 4.67    4.67 1000.00        4.67
```

```
## [1] "unweighted population average: 5.86"
```

```
## [1] "weighted population average: 13.86"
```

Truncation

The single participant of 3 years of age has a very large influence on the weighted population average. As such, we propose to truncate the relative age-specific weights w at 3. Defining a general cutoff of the $\dot{w}$ is not straightforward, so we continue with w. As such, the weighted population average equals:

```
survey_data$w_tilde[survey_data$w_tilde>3] = 3
```

```
## [1] "weighted population average after truncation: 10.28"
```

Acknowledgements

The authors acknowledge support from the Research Foundation – Flanders (FWO) (postdoctoral fellowship 1234620N) and the European Union's Horizon 2020 research and innovation program from European Research Council (grant 682540 – TransMID and 101003688 – EpiPose).