# Should It Pass?: An Introduction to Weighted Data

Many statistics classes talk about data that comes from a **simple random sample** (SRS), used to accurately estimate the population. An SRS minimizes any potential bias by ensuring that each individual (and each combination of individuals) in a population has an equal chance of being selected.

When a researcher is interested in examining distinct subgroups within a population, it is often best to use a **stratified random sample** to better represent the entire population. A **stratified random sample** involves dividing the population of interest into several smaller groups, called "strata" and then taking a simple random sample from each of these smaller groups. This method is commonly used when we want to guarantee a large enough sample from each subgroup. When this type of sampling method is used, it is important to use **weights** to take the relative size of each subgroup into account.

Weights are values assigned to individuals in a sample. These assigned values take into account how representative the individual is of the total population. When subjects are multiplied by their weights, the adjusted sample is more reflective of the actual population.

## Part A. Estimating Population Proportions Using SRS and Stratified Samples:

Suppose you want to know whether the students at a local university would like to pass a resolution to increase tuition costs in order to hire two statistical consultants to help with graduate research. The graduate program has a total of 1,000 students and the undergraduate program has a total of 20,000 students. For this resolution to pass, at least 50% of the entire student body must support it.

1. The school conducted a survey of 200 students asking them to indicate whether or not they support this resolution. One hundred and ten of the students surveyed said yes, and ninety students said no. When data is collected through a simple random sample, we expect that our sample statistic (55% = 110/200) is our best estimate of the population percentage. If these 200 observations were based upon a SRS, would you expect the resolution to pass?

2. You would most likely conclude that this resolution would have enough support to pass given the survey counts in question 1. However, that conclusion is only appropriate if the students sampled were an accurate representation of the total student body.

   **Stratification** is the process of dividing members of the population into similar subgroups. Then a SRS is taken within each subgroup. Stratified samples are commonly used in survey analysis to account for differences in subgroups within a population (Winship, et al). This stratification is used to better represent the population by reducing sampling variability. For example, if a SRS of 200 students were collected, how many graduate students would you expect to have been in the sample? [Hint: First calculate the percentage of graduate students in the population.] Would this be a large enough sample size to be confident in the results?

3. Suppose you now learn that the university did collect a stratified sample. Exactly 100 undergraduates and 100 graduates were randomly sampled. Below is a table representing the results of the survey. Why might you expect graduate students to have a different response to the resolution than undergraduates?

| Student Status | Yes | No | Total |
|---|---|---|---|
| Undergraduate | 30 | 70 | 100 |
| Graduate | 80 | 20 | 100 |
| Total | 110 | 90 | 200 |

4. 30% of the undergraduate students responded with a "Yes" to the survey. Thus, you can estimate that $20{,}000 * .30 = 6{,}000$ of the undergraduate student population would support the resolution. Complete the table below with estimates of the total population response to the resolution, using weights.

| Student Status | Yes | No | Total |
|---|---|---|---|
| Undergraduate | 6,000 | | 20,000 |
| Graduate | | | 1,000 |
| Total | | | 21,000 |

   Based upon this stratified sample, estimate the percentage of students in this university that would support the resolution. Does this change your decision in question 1) on whether or not the sample data provides evidence that the resolution will pass?

## Part B. Using Weights to Estimate Population

When a stratified random sample is implemented, **weighted data analysis** is used to estimate population parameters. In the previous example, you used the size of each subgroup (graduate and undergraduate students) as *weights* to better estimate the true percentage of students who support the resolution.

Survey data is usually available with individual entries as opposed to a summary table of counts. Below are the first 6 rows of the data set containing the results of the university survey:

```
ID Weight    Status Vote Weighted.Vote
1    200 Undergrad    0             0
2     10  Graduate    1            10
3    200 Undergrad    1           200
4    200 Undergrad    1           200
5     10  Graduate    0             0
6     10  Graduate    1            10
```

Each individual entry is given a *Weight* that can be used to scale the sample up to the population size. The sum of the values in the *Vote* column is 110, meaning that 110 students voted yes to the resolution (and thus 90 voted no). Note that 1 represents a yes vote and 0 represents a vote against the resolution.

However, this stratified random sample needs to be adjusted to reflect a more accurate population estimate. We multiply each student's vote by their assigned weight (stored in the *Weighted.Vote* column). This new column is summed to show the estimated number yes votes is 6,800 (and thus 14,200 weighted no votes).

## Part C. Understanding the Influence of Weighted Data on Population Estimates:

Part A and B demonstrate the importance of accounting for weights is survey data. When using a stratified sample, population estimates can be dramatically biased if weights are ignored (i.e. data is incorrectly treated as if it came from a SRS).

To better understand how to estimate population proportions using weighted data, go to the Weighted Data Introduction shiny app, http://rstudio.grinnell.edu/Weighted_Data_Introduction/. In this app you will see two charts. The first chart shows the information given by the sample data and the second chart demonstrates how using weights would influence the population estimates. Tabs on the left side of the app allow you to adjust the sample size of each subgroup and the sample proportion of positive responses.

5. Adjust the proportion of undergraduates who support the resolution to .6 and the proportion of graduates who are in favor of the resolution to .2 (keep both samples at 100). What proportion of the sample says yes? What is the best estimate of the proportion of all 21,000 students that would say yes?

6. Using the proportions given above (.6 undergraduate support and .2 graduate support), increase the sample size of undergraduates to 2,000. Notice that the proportion of graduates in the sample is now equal to the proportion of graduates in the population. In this case, what is the relationship between the proportion of students saying yes in the sample and the population?

7. Using sample sizes of 100 for both undergraduates and graduates, describe both graphs when the proportion of support from each group is exactly 30%. When both subgroups (graduates and undergraduates) have identical proportions saying yes in the sample, why do both the sample data graphs and the weighted population estimate graphs appear to come to the same conclusion?

8. Continue exploring this shiny app to create general statements about the influence of weights. For example, does increasing the graduate sample size reduce the differences between weighted and unweighted estimates? How about increasing the undergraduate sample size? Under what conditions are weighted and unweighed estimates the most different? Under what conditions are weighted and unweighted estimates the same?

## A Final Note

When studying a weighted survey it is important to note that weights affect not only visualizations of the data but also impact test statistics. Therefore it is essential to use caution when testing weighted data to be sure to account for the weight adjustments.

---

### *References:*

Winship, Christopher and Larry Radbill (1994). Sampling Weights and Regression Analysis. *Sociological Methods and Research.* 23(2), 230-257. Web. 9 June 2015.