A Weighted Analysis of the Senior Survey

Matt Brady

April, 25 2006

Introduction

The senior survey provides valuable information to the university administration, faculty, students, and prospective students. The administration uses the data from this survey to compare Trinity to its peer institutions and to study longitudinal trends. For example, the university community wants to know if Trinity students study as many hours per week as those at other schools. Furthermore, it would be beneficial to the university community to know if the amount of time Trinity students study is increasing, decreasing, or staying constant. Also, the community would like to know if students are getting a well rounded liberal arts education and if students are utilizing the resources available to them. Therefore, it is imperative to have accurate information to give these individuals so that Trinity is representing the true situation of students.

My project will focus on making the data analysis as accurate as possible. The answers students give to the questions on the survey vary dramatically across major; for example, English majors use the scientific method much less often than science majors. Therefore, if a disproportionate number of science majors are selected to take the survey, it will appear that students at Trinity use the scientific method more than they really do. Ideally, it would be best if the correct proportion of each major were selected before administering the survey. For example, if 30% of the graduating population are science majors, then it would be best if 30% of the students surveyed were science majors. No more and no less.

However, in administering the survey, upper division classes are selected and

seniors in that class fill out the survey. The classes are selected to generate approximately the correct proportion of students desired, but in selecting classes it is impossible to get correct percentages of students from each major. Additionally, some professors do not allow the survey to be administered in their class due to time constraints. Thus, the mean of each survey question could be biased because the correct proportion of majors is not selected.

Survey Sampling Theory

Proportionate sampling, in the case of the senior survey, would require that the proportion of each major sampled equal the proportion of students in that major that actually graduated. Since there is no efficient way to select a proportionate sample, post-stratification weighting is the best way to remedy the above problem. Each year we know how many students graduate and know what each student's major is. Therefore, we know the correct "weights" (the percentage of students graduating with a particular major) for each major. The survey population will be divided into different strata based on major. The mean and variance will then be calculated for each particular stratum (major or group of majors). If the survey is split into three strata (for example, science, business, and other majors) then there will be three means, one for each of these groups, and three variances. Post-stratification weighting is a linear combination of these statistics for each stratum, weighted with the true proportion of graduating seniors in that major, which will then yield one mean and one variance for a particular question on the survey. If you take the mean of all the responses of a particular question for a non-proportionate sample, the sample mean may be biased. Stratification eliminates this bias by weighting each stratum sample mean. This can increase the variance of the sample mean, but it eliminates the bias.

In order to provide a more theoretical understanding of the mean and variance for survey samples, we must first explore some basic statistical properties. Let Y be a random variable that models a student's response to a particular question. The population mean and variance of a set of N measurements is defined as follows.

Definition 1. *The population mean, $\bar{Y}$, is defined as the sum of all mea-*

surements in the population $(Y_1, Y_2, \ldots, Y_N)$ *divided by* N.

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

Definition 2. *The variance of the population,* S^2*, is a way to measure the variation of the measurements* $Y_1, Y_2, \ldots, Y_N$*. The variance of the population is essentially defined as the average of the squared deviations from the mean* $\bar{Y}$*.*

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

$\bar{Y}$ and S^2 are called population parameters and are unknown because the entire senior population was not polled. Hence, we will use estimators in place of the above population parameters. The estimator of the mean is the sample mean, $\bar{y}$.

Definition 3. *The sample mean,* $\bar{y}$*, is defined as the sum of all measurements of a simple random sample with replacement* $(y_1, y_2, \ldots, y_n)$ *divided by* n*.*

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

The estimator for the variance will be defined and used later. It should be noted that $E(y_i) = \bar{Y}$, because the expected value of any response is the population mean. Likewise, $Var(y_i) = S^2$. The expected value and variance of $\bar{y}$ that are found in any introductory statistics text assume a SRS (simple random sample) *with* replacement. Using basic properties of expected values we have that, for a SRS with replacement,

$$E(\bar{y}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{\sum_{i=1}^{n} E(y_i)}{n} = \frac{\sum_{i=1}^{n} \bar{Y}}{n} = \bar{Y}.$$

Therefore, $\bar{y}$ is an unbiased estimator of $\bar{Y}$. Similarly, because observations in a SRS are independent,

$$Var(\bar{y}) = Var\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{Var\left(\sum_{i=1}^{n} y_i\right)}{n^2} = \frac{\sum_{i=1}^{n} Var(y_i)}{n^2} = \frac{nS^2}{n^2} = \frac{S^2}{n}.$$

The mean and variance of $\bar{y}$ for survey sampling are similar to that of a SRS with replacement, but there are some twists that must be addressed. The properties of the mean and variance of $\bar{y}$ for a SRS shown above assume replacement. However, survey samples do not allow replacement: once a person is surveyed, theoretically, he or she cannot be surveyed again.

For a SRS with replacement, independence can be assumed because any observation has no effect on any other. However, for our survey results (SRS without replacement) every observation affects the probability of the next observation. For example, consider a box which contains 12 red balls and 12 white balls. If one ball is selected from the box, the probability of selecting a red ball is $\frac{1}{2}$. However, if a red ball is selected on that draw, then the probability of drawing a red ball on the next selection is $\frac{11}{23}$. If a red ball was not selected on the first draw, then the probability of drawing a red ball on the next selection is $\frac{12}{23}$. Either way, the probability of selecting a red ball changes when replacement is not permitted.

We can also calculate the population parameters for the example above. Let each red ball be assigned $y_i = 1$, and each white ball $y_i = 0$. The parameters are calculated as follows

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{\sum_{i=1}^{24} Y_i}{24} = \frac{12}{24}$$

$$S^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N-1} = \frac{\sum_{i=1}^{24} \left(Y_i - \frac{1}{2}\right)^2}{23} = \frac{12\left[\left(0 - \frac{1}{2}\right)^2 + \left(1 - \frac{1}{2}\right)^2\right]}{23} = \frac{6}{23}$$

Definition 4. *The sample mean of a simple random sample without replacement, $\bar{y}_{srvy}$, is defined as the sum of all measurements $(y_1, y_2, \ldots, y_n)$ divided by n.*

$$\bar{y}_{srvy} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

For survey results the expected value of the sample mean is still the same,

$$E(\bar{y}_{srvy}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \bar{Y}.$$

Again consider the example that we used above. Imagine two different scenarios: *Case 1*: 4 balls are selected from the box with replacement, and *Case 2*: 4 balls are selected from the box without replacement. $\sum_{i=1}^{n} y_i$ is the total number of red balls selected from the box. $\sum_{i=1}^{n} y_i$ in Case 1 is a binomial distribution and $\sum_{i=1}^{n} y_i$ in Case 2 is a hypergeometric distribution. Therefore, we can calculate the expected value of the sum in Case 1 by using the expected value formula for a binomial distribution with $p = \frac{1}{2}$ and $n = 4$,

$$E(\sum_{i=1}^{n} y_i) = np = 4 * \frac{1}{2} = 2.$$

Using the expected value of the sample mean,

$$E(\bar{y}) = \frac{E(\sum_{i=1}^{n} y_i)}{n} = \frac{2}{4} = \frac{1}{2} = \bar{Y}.$$

For Case 2 we have a hypergeometric distribution with $n = 4$, $N = 24$, and $M = 12$, which yields

$$E(\sum_{i=1}^{n} y_i) = n\left(\frac{M}{N}\right) = 4 * \frac{12}{24} = 2.$$

Similarly,

$$E(\bar{y}) = \frac{E(\sum_{i=1}^{n} y_i)}{n} = \frac{2}{4} = \frac{1}{2} = \bar{Y}.$$

Notice that in this example the expected value of the sample means are the same with and without replacement.

We can now calculate the variances for this example. The variance of the sum for Case 1 is a binomial random variable and can be calculated as follows,

$$Var(\sum_{i=1}^{n} y_i) = np(1-p) = 4 * \frac{1}{2} * \frac{1}{2} = 1.$$

The variance of the sample mean can be found as follows,

$$Var(\bar{y}) = \frac{Var(\sum_{i=1}^{n} y_i)}{n^2} = \frac{1}{4^2} = \frac{1}{16}.$$

Again, recall that Case 2 is a hypergeometric distribution with $n = 4$, $N = 24$, and $M = 12$. The variance of the sum for Case 2 is

$$Var(\sum_{i=1}^{n} y_i) = n\left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{N-1}\right) = 4 * \frac{12}{24} * \frac{12}{24} * (1 - \frac{12}{24}) = \frac{5}{16},$$

which yields the following variance of the sample mean for Case 2,

$$Var(\bar{y}) = \frac{Var(\sum_{i=1}^{n} y_i)}{n^2} = \frac{\frac{5}{6}}{4^2} = \frac{5}{96}.$$

The variance of the sample mean without replacement is smaller than the variance of the sample mean with replacement. The variance is smaller without replacement because we are removing people from the population; thus, we are getting a better estimate of the sample mean without replacement than with replacement.

Now can mathematically show how $Var(\bar{y}_{srvy})$ is different. The variance of the sum, in a sample without replacement, is not necessarily equal to the sum of the variances. To illustrate this concept we must investigate the variance formula. First note that,

$$E(\sum_{i=1}^{n} y_i) = n\bar{Y}.$$

$$
\begin{aligned}
Var(\sum_{i=1}^{n} y_i) &= E\left[\left(\sum_{i}^{n} y_i - n\bar{Y}\right)^2\right] && (1)\\
&= E\left[\left(\sum_{i}^{n} (y_i - \bar{Y})\right)^2\right] && (2)\\
&= E\left[\sum_{i,j}^{n} (y_i - \bar{Y})(y_j - \bar{Y})\right] && (3)\\
&= E\left[\sum_{i}^{n} (y_i - \bar{Y})^2 + \sum_{i \neq j}^{n} (y_i - \bar{Y})(y_j - \bar{Y})\right] && (4)\\
&= E\left[\sum_{i}^{n} (y_i - \bar{Y})^2\right] + E\left[\sum_{i \neq j}^{n} (y_i - \bar{Y})(y_j - \bar{Y})\right] && (5)\\
&= \sum_{i}^{n} E\left[(y_i - \bar{Y})^2\right] + \sum_{i \neq j}^{n} E\left[(y_i - \bar{Y})(y_j - \bar{Y})\right] && (6)
\end{aligned}
$$

A sample with replacement is independent; thus, the expected value of a product equals the product of the expected values for a sample with replacement. Furthermore, $E(y_i - \bar{Y}) = 0$. Thus for a sample with replacement

$$E\left[(y_i - \bar{Y})(y_j - \bar{Y})\right] = E\left[(y_i - \bar{Y})\right] E\left[(y_j - \bar{Y})\right] = 0.$$

However, independence does not hold under a sample without replacement. Therefore, the variance of the sum for a sample without replacement *is not* necessarily equal to the sum of the variances.

Definition 5. *The sampling fraction, f, is defined as the proportion of the population that was sampled.*

$$f = \frac{n}{N}$$

Theorem 1[1]**.** *The variance of a sum for a SRS* without *replacement is*

$$Var(\sum_{i=1}^{n} y_i) = (1 - f)nS^2$$

.

[1]Leslie Kish, *Survey Sampling*, New York: John Wiley and Sons, 1965, 63.

Proof. From (6) we have,

$$Var(\sum_{i=1}^{n} y_i) = \sum_{i}^{n} E\Big[(y_i - \bar{Y})^2\Big] + \sum_{i \neq j}^{n} E\Big[(y_i - \bar{Y})(y_j - \bar{Y})\Big].$$

Taking the expected values yields,

$$E\Big[(y_i - \bar{Y})^2\Big] = \frac{\sum_{i}^{N}(Y_i - \bar{Y})^2}{N}$$

and

$$E\Big[(y_i - \bar{Y})(y_j - \bar{Y})\Big] = \frac{\sum_{i \neq j}^{N}(Y_i - \bar{Y})(Y_j - \bar{Y})}{N(N-1)},$$

which yields,

$$Var(\sum_{i=1}^{n} y_i) = \frac{n}{N}\sum_{i}^{N}(Y_i - \bar{Y})^2 + \frac{n(n-1)}{N(N-1)}\Big[\sum_{i \neq j}^{N}(Y_i - \bar{Y})(Y_j - \bar{Y})\Big]$$

Recall from (2) and (4),

$$\Big(\sum_{i}^{N}(Y_i - \bar{Y})\Big)^2 = \sum_{i}^{N}(Y_i - \bar{Y})^2 + \sum_{i \neq j}^{N}(Y_i - \bar{Y})(Y_j - \bar{Y}).$$

Hence,

$$\sum_{i \neq j}^{N}(Y_i - \bar{Y})(Y_j - \bar{Y}) = \Big(\sum_{i}^{N}(Y_i - \bar{Y})\Big)^2 - \sum_{i}^{N}(Y_i - \bar{Y})^2.$$

Thus,

$$Var(\sum_{i=1}^{n} Y_i) = \frac{n}{N}\sum_{i}^{N}(Y_i - \bar{Y})^2 + \frac{n(n-1)}{N(N-1)}\Big[\Big(\sum_{i}^{N}(Y_i - \bar{Y})\Big)^2 - \sum_{i}^{N}(Y_i - \bar{Y})^2\Big].$$

Note that $\left(\sum_{i}^{N}(Y_i - \bar{Y})\right) = 0$. Therefore,

$$\begin{aligned} Var(\sum_{i=1}^{n} y_i) &= \frac{n}{N}\sum_{i}^{N}(Y_i - \bar{Y})^2 - \frac{n(n-1)}{N(N-1)}\sum_{i}^{N}(Y_i - \bar{Y})^2 \\ &= n\left(\frac{\sum_{i}^{N}(Y_i - \bar{Y})^2}{N-1}\right)\left[\frac{(N-1)-(n-1)}{N}\right] \\ &= nS^2\left(\frac{N-n}{N}\right) \\ &= (1-f)nS^2 \end{aligned}$$

□

Corollary 1. *The variance of a sample mean without replacement, $\bar{y}_{srvy}$, is*

$$Var(\bar{y}_{srvy}) = \frac{(1-f)S^2}{n}.$$

Proof.

$$Var(\bar{y}_{srvy}) = Var\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{Var(\sum_{i=1}^{n} y_i)}{n^2} = \frac{(1-f)nS^2}{n^2} = \frac{(1-f)S^2}{n}.$$

□

For a survey sample if only a small proportion of people are surveyed, then we would expect the variance of the sample mean to be large. However, if the proportion of people surveyed is large, then we would expect the variance of the sample mean to be much smaller. If the population size is large and the number of people surveyed is small then f is very small, and the variance resembles the variance of a SRS with replacement $Var(\bar{y}) = \frac{S^2}{n}$.

Proportionate Samples

We will begin by discussing stratification and proportionate samples. Definitions of the major variables pertaining to strata are given in Figure 1. First,

h		denotes a particular stratum
$\bar{y}_h$	$\bar{y}_h = \frac{\sum_{j=1}^{n_h} y_j}{n_h}$	the sample mean for the hth stratum
n_h		the number of people surveyed from the hth stratum
f_h	$f_h = \frac{n_h}{N_h}$	the sampling fraction for the hth stratum
W_h	$W_h = \frac{N_h}{N}$	the fraction of people in the population of the hth stratum over the entire population
$\bar{y}_{prop}$	$\bar{y}_{prop} = \sum_{h}^{H} W_h \bar{y}_h$	the sample mean that weights the individual sample means of each stratum
$\bar{y}_{ps}$	$\bar{y}_{ps} = \bar{y}_{prop}$	the sample mean for a post-stratified survey

Figure 1: Key Terms

we consider that the data has been partitioned into distinct strata (in our case all the students have been put into their major groups). We have H different strata. In a proportionate sample the proportion of people selected from each stratum is equivalent to the proportion of people in the entire stratum population. That is, if 30% of graduating seniors are science majors, then 30% of the survey sample will be science majors. The weight of each stratum is denoted by $W_h = \frac{N_h}{N}$. We have that

$$\frac{n_h}{n} = \frac{N_h}{N} = W_h.$$

Cross multiplying gives us

$$\frac{n_h}{N_h} = \frac{n}{N}.$$

Thus, the sampling fraction of each stratum equals the sampling fraction for the entire population: $f_h = f$ for $h = 1, 2, \ldots, H$.

The sample mean, $\bar{y}_h$, is found for each stratum h. The variance of the sample mean for the hth stratum is exactly the same as the variance without strata:

$$Var(\bar{y}_h) = \frac{(1 - f_h)S_h^2}{n_h}.$$

For this paper the weights signify the number of people with a given major in the population divided by the entire population. The weights are equal to both the proportion of people in the population in the stratum and the proportion of people in the sample in the stratum; hence, we can use these values to "weight" each stratum sample mean to get a mean that is representative of the students combined. Weighting the sample means for each stratum under a proportionate sample generates the exact same sample mean that would be computed if all the responses were added up and divided by the total number. However, introducing this notation now is necessary because it allows use to distinguish post-stratification from proportionate sampling later.

Definition 6. *The proportionate sample mean, $\bar{y}_{prop}$, is defined as a linear combination of each stratum's sample mean weighted by $W_1, \ldots, W_H$.*

$$\bar{y}_{prop} = \sum_{h}^{H} W_h \bar{y}_h.$$

Theorem 2.

$$Var(\bar{y}_{prop}) = \left(\frac{1-f}{n}\right) \sum_{h}^{H} W_h S_h^2$$

Proof.

$$Var(\bar{y}_{prop}) = Var(\sum_{h}^{H} W_h \bar{y}_h)$$

We will assume that the sample mean for each stratum is independent, because what one major does should not effect what other majors do. Therefore,

$$\begin{aligned} Var(\bar{y}_{prop}) &= \sum_{h}^{H} W_h^2 \frac{(1-f_h) S_h^2}{n_h} \\ &= \frac{1}{N^2} \sum_{h}^{H} \frac{(1-f_h) N_h^2 S_h^2}{n_h} \\ &= \frac{1}{N^2} \sum_{h}^{H} \frac{(1-f_h) N_h S_h^2}{f_h} \end{aligned}$$

Recall that $f_h = f$ for $h = 1, 2, \ldots, H$. Thus, we can pull the $\frac{(1-f)}{f}$ term out of the sum. Also Note that $(1-f) = \frac{N-n}{N}$. Hence,

$$\begin{aligned} Var(\bar{y}_{prop}) &= \frac{1}{N^2}\left(\frac{1-f}{f}\right)\sum_h^H N_h S_h^2 \\ &= \frac{1}{N^2}\left(\frac{\frac{N-n}{N}}{\frac{n}{N}}\right)\sum_h^H N_h S_h^2 \\ &= \left(\frac{N-n}{N}\right)\left(\frac{1}{n}\right)\sum_h^H \frac{N_h}{N} S_h^2 \\ &= \left(\frac{1-f}{n}\right)\sum_h^H W_h S_h^2. \end{aligned}$$

□

If the sample is perfectly proportionate, then the variance of the weighted stratum sample means would be the same as the variance of the sample mean with no weights, that is $Var(\bar{y}_{srvy}) = Var(\bar{y}_{prop})$. Thus, if the majors' weights for the senior survey had been selected in a perfectly proportionate manner then everything would be alright, but it was not.

Post-Stratification

Post-stratification can be used when creating a proportionate sample is hard, or costly, or when the data has already been collected. It is beneficial to use post-stratification analysis anytime that the sample mean for the survey may be biased because the sample is not proportionate. Post-stratification eliminates any bias that may be present in the sample mean. In order to do post-stratification analysis it is necessary to have information on the population as a whole, and it is also necessary to be able to fragment the survey sample into strata. The first step for post-stratification analysis is to determine the particular strata that will be used and then to group the data into these strata. Once that has been done, sample means should be found

for every strata and weighted by W_h to determine a sample mean for the entire survey that has *no* bias. As you might imagine the variance for post-stratification is larger than the variance for a proportionate sample because it is essentially trying to correct the data *after* it has been collected, which is much less precise than selecting a proportionate sample *before* the data is analyzed. The variance of the sample mean using post-stratification is given by[2]

$$Var(\bar{y}_{ps}) = \frac{1-f}{n}\sum_{h}^{H} W_h S_h^2 + \frac{1-f}{n}\sum_{h}^{H} W_h(1-W_h)\frac{S_h^2}{n_h^2}.$$

Note that the first term corresponds to $Var(\bar{y}_{prop})$, but the second term makes the variance larger due to weighting each stratum's sample mean after the survey has been completed. However, if n_h is relatively large for every value of h then the variance of a post-stratified sample is similar to that of a proportionate sample.

Theory Summary

The list below outlines the properties of the variance of the sample mean.

- The variance in a SRS with replacement is

$$Var(\bar{y}) = \frac{S^2}{n}.$$

- The variance in a SRS without replacement is

$$Var(\bar{y}_{srvy}) = \frac{(1-f)S^2}{n}.$$

- The variance for the hth stratum in a SRS without replacement is

$$Var(\bar{y}_h) = \frac{(1-f_h)S_h^2}{n_h}.$$

[2]Kish, 90. Hansen, Hurwitz, and Madow, *Sample Survey Methods and Theory*, New York: John Wiler and Sons, Vol. II, 1953, 5.13.

- The variance of the weighted proportionate sample mean is

$$Var(\bar{y}_{prop}) = \left(\frac{1-f}{n}\right)\sum_{h}^{H} W_h S_h^2. \tag{7}$$

- The variance using post-stratification is

$$Var(\bar{y}_{ps}) = \frac{1-f}{n}\sum_{h}^{H} W_h S_h^2 + \frac{1-f}{n}\sum_{h}^{H} W_h(1-W_h)\frac{S_h^2}{n_h^2}. \tag{8}$$

Figure 2 provides an excellent visual interpretation of the spread of the sample mean for a sample that is not proportionate and does not use the weights, a proportionate sample, and a sample that is not proportionate that utilizes the post-stratification weights. The sample mean for the non-proportionate sample without weights is biased, meaning that the middle of the estimator distribution lies to the left (or right) of the population parameter. A proportionate sample corrects the bias, the distribution is centered around the population parameter. The non-proportionate sample that uses the post-stratification weights is unbiased, it is centered about the population parameter, but the variance, or spread, of the distribution has increased due to using the weights after it has been collected, instead of having a proportionate sample. It is possible that the sample mean of a non-proportionate sample without weights is only slightly biased, which would mean that the distribution is centered just to the left or right of the population parameter. If this is the case, the benefits of post-stratification may be minimal or even harmful, because post-stratification would remove the bias but could increase the spread. Therefore, it may be possible that the sample mean using post-stratification weights lies in one of the outlying tails. In any case, Figure 2 illustrates the benefits and costs associated with post-stratification in a manner that is easy to comprehend.

In conclusion, I will be examining the weighted sample means for several questions on the survey. If a certain major was oversampled in the survey, then weighting the stratum sample means should correct the bias. I will also look at the variance of the sample mean for each question using both the post-stratification method and the non-weighted method to see if the variance is larger for post-stratification. If there is little difference in the two

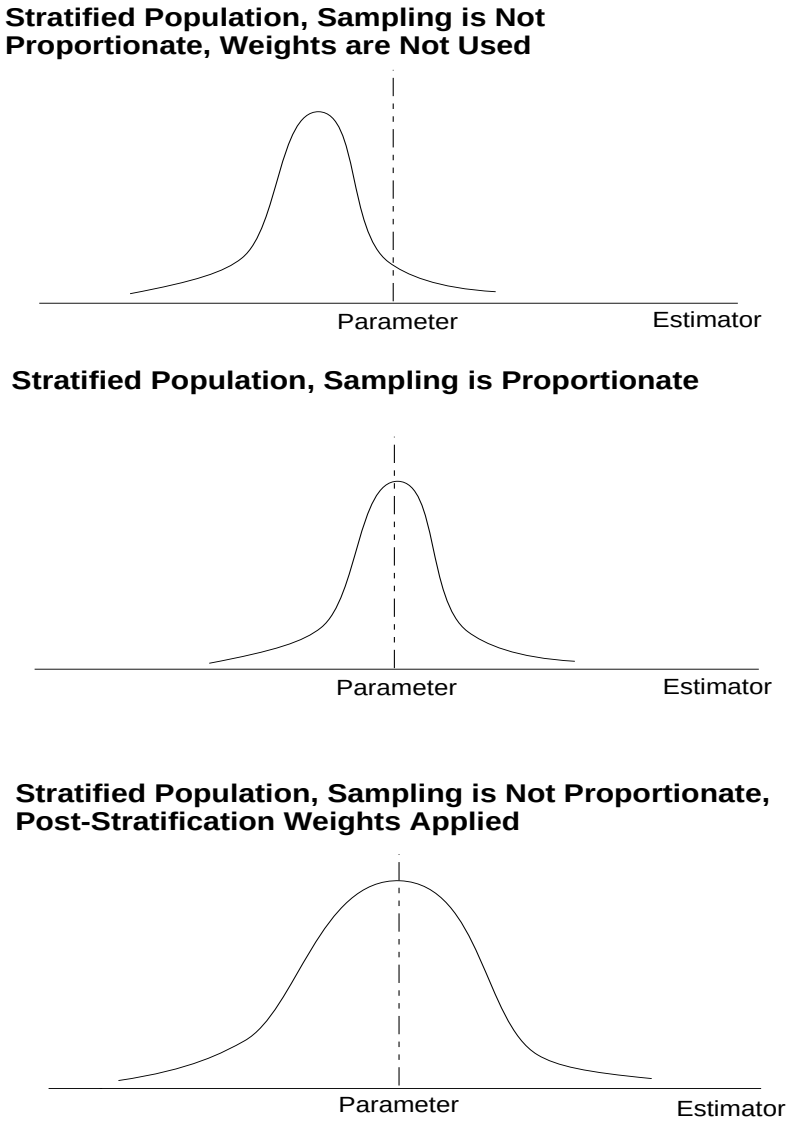

Figure 2: Sample Mean Distribution

variances, then post-stratification will be very beneficial: the bias will be corrected, with little change in the variance.

Since the variance of the population is not known, the sample variance, s^2, will be used in the calculations. The sample variance is an unbiased estimator[3] of S^2 and is defined as

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}.$$

The sample variance is an unbiased estimator of the population variance, so it can be substituted into the variance of the sample mean equations in place of the population variance, S^2. The sample mean for the non-weighted data will be calculated as follows

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

The weighted sample mean for the post-stratified data will be calculated as follows

$$\bar{y}_{ps} = \sum_{h}^{H} W_h \bar{y_h}.$$

I will compute the variance of the sample mean for all responses, before weights are introduced, using the formula

$$Var(\bar{y}) = \frac{(1-f)s_y^2}{n}.$$

I will then compute the variance of the weighted sample mean for post-stratification using the following formula

$$Var(\bar{y}_{ps}) = \frac{1-f}{n}\sum_{h}^{H} W_h s_h^2 + \frac{1-f}{n}\sum_{h}^{H} W_h(1-W_h)\frac{s_h^2}{n_h^2}.$$

[3]Kish, 36.

Survey Details

Now that the theory behind survey sampling has been covered in depth we can move on to the details of the survey and the data itself. In this paper I will examine the data corresponding to two different questions on the senior survey. One of the questions gathers quantitative data from the student and one of the questions gathers categorical data from the student. The analysis of the quantitative question just requires that the sample mean and the variance of the sample mean be calculated as documented in the previous section. The analysis of the categorial question is a bit more difficult, since we do not have numbers to work with. Hence, we will transform the categorical question into a binary event. That is, for the given question one answer to the question will be categorized as a success and will be assigned the value 1, and the other answer will be categorized as a failure and will be assigned the value 0. After this adjustment is made the sample mean and variance of the sample mean can easily be calculated because the responses are quantitative. The two questions on the survey that I will examine are presented below.

Question 1. *During the time school is in session, about how many hours a week do you usually spend outside of class on activities related to your academic program, such as studying, writing, reading, lab work, rehearsing, etc.?*
(A) 5 or fewer hours a week
(B) 6-10 hours a week
(C) 11-15 hours a week
(D) 16-20 hours a week
(E) 21-25 hours a week
(F) 26-30 hours a week
(G) more than 30 hours a week

The responses to the question are assigned a single numerical value according to the average of the high and low numbers in the range. Thus, if a student responds that he studies 11-15 hours a week, then we will denote that as 13 hours of study. 35 hours will be assigned for the 'more than 30 hours' response.

Question 2. *How often have you used a computer to produce visual displays of information (charts, graphs, spreadsheets, etc.)?*

(A) Very Often
(B) Often
(C) Occasionally
(D) Never

Question 2 involves categorical data. For this question I have combined responses (A) and (B) and will from now on call them "frequent" responses, and I have combined responses (C) and (D) and will refer to them as "infrequent" responses. This was done to create a binomial distribution instead of a multinomial distribution. A binomial distribution works well for this question because it is simple and it gives us a pretty good idea of the percentage of students who use visual displays on a regular basis. Frequent responses are assigned a value of 1 and infrequent responses are assigned a value of 0. The sample mean for this question is a percentage; for example, a sample mean of .6 would indicate that 60% of those surveyed produce visual displays *frequently*.

Now before we can begin to analyze the data for each question, I must explain how strata were introduced to the data. To refresh, we are using strata to correct for possible oversampling. It would seem likely that science majors tend to study more than all other majors. Therefore, if too many science majors were selected to take part in the survey, then it would appear that the entire student body studies more than it really does. Students will be split into three different strata: *science, business*, and *other*. If a student is a double major, then he or she will be assigned to the science group if either major is science. If the student is a double major and one of the majors is business, then he or she will be assigned to the business group unless the other major is science. Finally, if a student is double majoring in two other disciplines, then he or she will remain in the other category. Figure 3 provides the number of students in each stratum in both the population and in the sample. The "Prop" row provides the proportion of students in that stratum that were sampled. If the Prop value is greater than W_h, then that stratum has been oversampled. For example, Figure 3 shows that science majors were oversampled every year. It is possible that the sample mean of hours studied without stratification would be biased in those years due to science majors studying more (as we will see later). Likewise, business and science majors tend to use a lot of graphs and charts, so if too many of them are selected

Student Information									
	Business			Science			Other		
	2001	2003	2005	2001	2003	2005	2001	2003	2005
N_h	166	147	138	111	117	127	220	261	271
n_h	109	62	46	62	71	60	88	47	54
W_h	0.33	0.28	0.26	0.22	0.22	0.24	0.44	0.50	0.51
Prop	0.42	0.34	0.29	0.24	0.39	0.38	0.34	0.26	0.34

Figure 3: Student Demographics by Major

Hours Studied			
	2001	2003	2005
$\bar{y}_{srvy}$	14.31	14.55	17.14
$\text{Var}(\bar{y}_{srvy})$	.1125	.2403	.3365
std dev$(\bar{y}_{srvy})$	.3354	.4902	.5801
n	259	180	160
f	0.52	0.34	0.30

Figure 4: Question 1, No Strata

then it will appear that more students utilize these visual aids than they do in reality. Figure 3 also shows that Business majors were oversampled in all years. Couple that with the oversampling of science majors in those years as well, and there could be significant bias in the sample mean without post-stratification. These questions should provide fruitful results for post-stratification.

Data Analysis: Question 1

The statistics for Question 1 without stratification are presented in Figure 4 for the three years in which this survey was given. Figure 4 illustrates a clear upward trend in the hours studied over these three years, in the absence of post-stratification. However, we must be careful not to assume that this is the actual trend without first exploring how the introduction of strata affects the results. Figure 5 provides all the valuable information for the three separate strata and it also provides the key statistics for each stratum.

Hours Studied by Major									
	Business			Science			Other		
	2001	2003	2005	2001	2003	2005	2001	2003	2005
$\bar{y}_h$	12.39	13.10	15.04	16.98	17.27	20.90	14.81	12.36	14.74
std dev($\bar{y}_h$)	0.414	0.777	0.884	0.716	0.664	0.911	0.598	0.724	0.894
s_h^2	54.43	64.78	53.91	72.05	79.54	94.29	52.36	30.02	53.86
n_h	109	62	46	62	71	60	88	47	54
W_h	0.33	0.28	0.26	0.22	0.22	0.24	0.44	0.50	0.51
Prop	0.42	0.34	0.29	0.24	0.39	0.38	0.34	0.26	0.34

Figure 5: Question 1, Stratified

Comparing Hours Studied			
	2001	2003	2005
$\bar{y}_{ps}$	14.49	13.66	16.28
$\bar{y}_{srvy}$	14.31	14.55	17.14
Var($\bar{y}_{ps}$)	.1062	.1855	.2783
Var($\bar{y}_{srvy}$)	.1125	.2403	.3365
std dev($\bar{y}_{ps}$)	.3259	.4306	.5275
std dev($\bar{y}_{srvy}$)	.3354	.4902	.5801

Figure 6: Question 1, Post-Stratified and Non-Stratified

Figure 5 shows that there is a defined upward trend in hours studied for business and science majors, but the results do not translate to the other majors. Hours studied for other majors declined in 2003 and rose in 2005 to reach roughly the 2001 mark. The "Prop" row provides data on the proportion of each type of major sampled and the W_h row provides data on the proportion of each type of major in the population. Looking at how the two values differ for a given year and major will tell us if a major was oversampled for a particular year. For example, it appears that science majors were oversampled in 2003, while other majors were undersampled. The ramifications of this are addressed in Figure 6, which compares the statistics using post-stratification and no stratification.

After weighting the stratified means we see that the average number of hours for 2003 has decreased due to the oversampling mentioned earlier. We see the similar thing happening in 2005, science majors were again oversampled and other majors were undersampled. Hence, the average hours studied decreases after stratification because science majors, on average, study more than other majors.

One interesting thing to note is that the variance of the sample mean actually decreases after stratification. From equations (7) and (8), we see that the variance for a post-stratified sample mean (non-proportionate sample) is larger than the variance of sample mean with proportionate sampling because of the extra term on the end. However, the extra term is divided by n_h^2 and the number of students sampled from every major is quite high, so that number should be very large, making the extra term basically zero. Therefore, the variance for post-stratification for this question would be roughly the same as the variance of a proportionate sample. The variance for post-stratification is lower than the variance without weights because science majors were oversampled all three years and science majors have the highest variance.

Hence, for Question 1 we can conclude that post-stratification provides a more accurate sample mean, because it takes into account oversampling and undersampling by correcting the bias in the sample mean, it provides a variance that is roughly the same as would be the variance of a proportionate sample, and it provides a variance that is better than the case without weights.

Furthermore, we can conclude for the entire senior population that studying time rose to its highest level of the three years in 2005. We can do a z-test on the difference between the two means to determine if the difference in hours studied in 2001 and 2005 is statistically significant. The z value can be determined as follows:

$$z = \frac{\bar{y}_{ps2005} - \bar{y}_{ps2001}}{\sqrt{Var(\bar{y}_{ps2005}) + Var(\bar{y}_{ps2001})}}.$$

Gathering the variables from Figure 6,

$$z = \frac{16.28 - 14.49}{\sqrt{.1062 + .2783}} = 2.887.$$

Proportion of Students Producing Visual Displays			
	2001	2003	2005
$\bar{y}_{srvy}$	0.680	0.730	0.720
$\text{Var}(\bar{y}_{srvy})$	0.00040	0.00074	0.00087
std dev$(\bar{y}_{srvy})$	0.020	0.027	0.029
n	259	178	164
f	0.52	0.34	0.31

Figure 7: Question 2, No Strata

The z-value can then be compared with a normal distribution to determine the rejection region for the two tailed test. We can reject the hypothesis that the two means are the same at the 1% level. We can conclude that we are reasonably certain that the average hours studied in 2005 is higher than in 2001.

Data Analysis: Question 2

The statistics for Question 2 without stratification are outlined in Figure 7. A brief glance at the data reveals that apparently students in 2003 and 2005 produced and used visual displays more often than did students in 2001. This seems logical. More computers are available on campus now than in 2001. Computers are in most classrooms now enabling Power Point presentations for in class presentations. Furthermore, students have probably begun to use visual displays much more often in response to the increased demand by employers of science and business majors for students to have knowledge of these skills. However, we must be very careful not to interpret the results from Figure 7 as proof that students are using visual displays more now than in 2001. After all, business and science students are going to use visual displays much more than other majors. Business majors routinely give power point presentations and many science majors work will graphs on a daily basis and use spreadsheets routinely. Thus, this is an excellent question to see how, or if, post-stratification alters the results. Figure 8 illustrates the statistics for Question 2 by stratum. Figure 8 provides some revealing information. The proportion of science majors who produced visual displays in 2005 is roughly the same proportion as 2001. Similarly, the proportion of

Visual Displays Produced by Major									
	Business			Science			Other		
	2001	2003	2005	2001	2003	2005	2001	2003	2005
$\bar{y}_h$	0.76	0.75	0.83	0.85	0.80	0.85	0.45	0.59	0.46
std dev($\bar{y}_h$)	0.024	0.042	0.044	0.030	0.030	0.033	0.041	0.066	0.061
s_h^2	0.18	0.19	0.14	0.13	0.16	0.13	0.25	0.25	0.25
n_h	109	61	48	62	71	62	88	46	54
W_h	0.33	0.28	0.26	0.22	0.22	0.24	0.44	0.50	0.51
Prop	0.42	0.34	0.29	0.24	0.40	0.38	0.34	0.26	0.33

Figure 8: Question 2, Stratified

Comparing Hours Studied			
	2001	2003	2005
$\bar{y}_{ps}$	0.646	0.682	0.651
$\bar{y}_{srvy}$	0.680	0.730	0.720
Var($\bar{y}_{ps}$)	0.00037	0.00079	0.00082
Var($\bar{y}_{srvy}$)	0.00040	0.00074	0.00087
std dev($\bar{y}_{ps}$)	0.019	0.028	0.0287
std dev($\bar{y}_{srvy}$)	0.020	0.027	0.0295

Figure 9: Question 1, Post-Stratified and Non-Stratified

other majors utilizing visuals in 2005 is roughly the same as the proportion in 2001. It appears that increase from 2001 to 2005 is due to a sharp rise in business students who say that they produce visual displays more often. The rise in $\bar{y}_{srvy}$ in 2003 appears to be due to other majors producing visual displays more frequently. Figure 9 compares the results of post-stratification to the original statistics without stratification. Figure 9 runs counter to the earlier theory that students produced visual displays more often in 2005 than they did in 2001. Looking at the weights and proportions in Figure 8 shows that the decrease in $\bar{y}_{ps}$ in 2005 can be attributed to the fact that science majors were oversampled in 2005 and other majors were undersampled.

Once again, the variance for post-stratification is below or near the variance without stratification. Thus, post-stratification provides insightful analysis

for Question 2 as well. Post-stratification provides a better estimate of the mean of the population. We will use the values in Figure 9 to formulate a z-value to determine if the proportion in 2001 and 2005 are different.

$$z = \frac{\bar{y}_{ps2005} - \bar{y}_{ps2001}}{\sqrt{Var(\bar{y}_{ps2005}) + Var(\bar{y}_{ps2001})}} = \frac{.651 - .646}{\sqrt{.00082 + .00037}} = 0.145$$

We cannot reject the hypothesis that the proportions in 2001 and 2005 are different even at the 10% level. Finally, we will use a z-test one last time to determine if the proportions in 2003 and 2005 might be different.

$$z = \frac{\bar{y}_{ps2005} - \bar{y}_{ps2003}}{\sqrt{Var(\bar{y}_{ps2005}) + Var(\bar{y}_{ps2003})}} = \frac{.651 - .682}{\sqrt{.00082 + .00079}} = -0.773$$

Again, we cannot reject the hypothesis that the proportions in 2003 and 2005 are different even at the 10% level. Thus, using post-stratification we are not able to determine if Trinity students are producing more visual images or not. However, post-stratification still has its benefit because it provides an unbiased estimate of the mean. Post-stratification provides better estimates of the population, even though we could not detect if the proportion was rising or falling.

Conclusion

Post-stratification can be a very valuable tool to provide better estimates of a population if creating a proportionate sample is costly. As we have seen the benefits of post-stratification are maximized when a large sample from each stratum is collected. The large sample reduces, or even eliminates, the extra term in the variance for a post-stratified sample. Post-stratification analysis provided valuable insight for the senior survey because the number of students in each stratum was sufficiently large.

Further research can investigate the benefits of post-stratification for the senior survey with more refined strata, perhaps individual majors. Since there are some majors with few students it is plausible that the variance for the post-stratified sample mean will be much larger than the sample mean without stratification.

References

Kish, Leslie. *Survey Sampling.* New York: John Wiley and Sons, 1965.