Homework 6 Solutions
303 Spring 2003
Steve Fienberg

Thursday, April 3

Part I

1. Suppose you want to estimate the percentage of persons who have been immunized against polio in each city and can take a SRS from each. What should you sample size be for each of the five cities so that each can be estimated within a margin of error of 4%? For which cities does the finite population correction make a difference?

Recall that for a margin of error of 4%, that means that

$$z_{\alpha} \sqrt{\frac{N-n}{N-1}} \frac{p(1-p)}{n} = 0.04$$

With 95% confidence, $$z_{\alpha} = 1.96$$. Using the worst case scenario, $$p = 0.5$$. Using the population values given in the table, we will then solve for $$n$$.

$$1.96 \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}} = 0.04$$

$$\frac{N-n}{N-1} \frac{p(1-p)}{n} = \left(\frac{0.04}{1.96}\right)^2$$

$$\frac{N}{n} - 1 = \frac{N-1}{p(1-p)} \left(\frac{0.04}{1.96}\right)^2$$

$$\frac{N}{n} = \frac{N-1}{p(1-p)} \left(\frac{0.04}{1.96}\right)^2 + 1$$

So ...

$$n = \frac{N}{\frac{N-1}{p(1-p)} \left(\frac{0.04}{1.96}\right)^2 + 1}$$
<table>
<thead>
<tr>
<th>City</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buckeye</td>
<td>535</td>
</tr>
<tr>
<td>Gilbert</td>
<td>595</td>
</tr>
<tr>
<td>Gila Bend</td>
<td>446</td>
</tr>
<tr>
<td>Phoenix</td>
<td>600</td>
</tr>
<tr>
<td>Tempe</td>
<td>598</td>
</tr>
<tr>
<td>Total</td>
<td>2774</td>
</tr>
</tbody>
</table>

Now, let’s see what the difference is if we don’t use the finite population correction. Again, we get:

\[
1.96 \times \sqrt{\frac{p(1-p)}{n}} = 0.04
\]

\[
\frac{p(1-p)}{n} = \left(\frac{0.04}{1.96}\right)^2
\]

\[
n = \left(\frac{1.96}{0.04}\right)^2 p(1-p)
\]

Using the worst case where \( p = 0.5 \), we get \( n = 601 \). This is the same for all of the cities (i.e. there is no dependence on \( N \)). It appears that the largest differences are with Guckeye (difference of 65) and Gila Bend (difference of 154). Gilbert and Tempe have only small differences (5 and 2 respectively).

2. Sum the sample sized for each of the five cities. Describe an alternative sampling scheme with different sample sizes for the five cities that would get better overall precision for the percentage of persons who have been immunized against polio for the population as a whole.

To do this, we first need to determine the precision of the case above. Recall that the variance of the stratified random sampling is

\[
Var(\bar{p}_{strat}) = \sum_{i=1}^{k} \left( \frac{N_i}{\sum_j N_j} \right) \frac{p_i(1-p_i)}{n_i} \cdot \frac{N - n_i}{N - 1}
\]

So plugging in the numbers, the variance of the above estimator is 0.0000329 so the standard error is 0.005.

Let’s see what proportional allocation can do here. First, assume that the total sample size is the same as the total sample size in part A, 2774. We will allocate the sample sizes proportional to the sizes of the population (i.e. \( \frac{4857}{4857+59338+1724+1149417+153821} = 0.0035 \) percent from Buckeye – 2774*0.0035=9.8. So 10 people will be sampled from Buckeye). Doing this, we get the following scheme.
Computing the variance as we did for the previous problem, we find that the variance is 0.0000000037, for a standard error of 0.0006. This is much better than the above sampling scheme.

**Part II**

1. Find the probability of selection $\pi_i$ for each unit $i$.

To find the probability of unit 1 (i.e. $y_1$), note that 1 is in sample 1 and 3 with probabilities $1/8$ and $1/8$ respectively. Thus, the probability of unit 1 is $1/8+1/8=2/8=1/4$. Continuing in this manner, we have

<table>
<thead>
<tr>
<th>Unit</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1/8+1/8=2/8$</td>
</tr>
<tr>
<td>2</td>
<td>$1/4+3/8=5/8$</td>
</tr>
<tr>
<td>3</td>
<td>$1/8+1/4=3/8$</td>
</tr>
<tr>
<td>4</td>
<td>$1/8+3/8+1/8=5/8$</td>
</tr>
<tr>
<td>5</td>
<td>$1/8+1/8=2/8$</td>
</tr>
<tr>
<td>6</td>
<td>$1/8+1/8+3/8=5/8$</td>
</tr>
<tr>
<td>7</td>
<td>$1/4+1/8=3/8$</td>
</tr>
<tr>
<td>8</td>
<td>$1/4+1/8+3/8+1/8=7/8$</td>
</tr>
</tbody>
</table>

2. What is the sampling distribution for the estimate of the population total $\hat{t} = 8\bar{y}$?

First determine the possible values of $\hat{t}$

<table>
<thead>
<tr>
<th>Sample</th>
<th>Probability</th>
<th>$\bar{y}$</th>
<th>$t = 8\bar{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1/8$</td>
<td>($1+4+7+8)/4=19/4$</td>
<td>$19/4\times8=38$</td>
</tr>
<tr>
<td>2</td>
<td>$1/4$</td>
<td>($2+4+7+8)/4=21/4$</td>
<td>$21/4\times8=42$</td>
</tr>
<tr>
<td>3</td>
<td>$1/8$</td>
<td>($1+4+7+8)/4=20/4$</td>
<td>$20/4\times8=40$</td>
</tr>
<tr>
<td>4</td>
<td>$3/8$</td>
<td>($2+4+7+8)/4=21/4$</td>
<td>$21/4\times8=42$</td>
</tr>
<tr>
<td>5</td>
<td>$1/8$</td>
<td>($4+7+7+8)/4=36/4$</td>
<td>$26/4\times8=52$</td>
</tr>
</tbody>
</table>

Thus, I find that the sampling distribution is:

<table>
<thead>
<tr>
<th>$t = 8\bar{y}$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>$1/8$</td>
</tr>
<tr>
<td>40</td>
<td>$1/8$</td>
</tr>
<tr>
<td>42</td>
<td>$1/4+3/8=5/8$</td>
</tr>
<tr>
<td>52</td>
<td>$1/8$</td>
</tr>
</tbody>
</table>
3. Construct the sampling distribution for \( \hat{t} \) under SRS without replacement for samples of size \( n = 4 \).

See attached file

4. Compare the sampling scheme in part B with the SRS scheme in part C, in terms of expectations, variances, and MSE.

From the attached file, we see that the mean of the SRS is 40 and the variance is 54.8. We can see that the total of the population is \((1+2+4+4+7+7+7+8)=40\). Thus the bias of the SRS is 0 and the MSE is 54.8.

From part B, we can find that the mean is 
\[
\frac{38}{8} + \frac{42}{4} + \frac{40}{8} + 42 \cdot \frac{3}{8} + \frac{52}{8} = 42.5.
\]

So the bias is 2.5. The variance is 
\[
\left(\frac{38}{8} - 42.5\right)^2/8 + \left(\frac{42}{4} - 42.5\right)^2/4 + \left(\frac{40}{8} - 42.5\right)^2/8 + \left(\frac{42}{4} - 42.5\right)^2/4 = 14.75.
\]
Thus, the MSE is \(2.5^2 + 14.75 = 21\). We can see that the sampling scheme in part B is better than the SRS. Even though there is slightly more bias, the reduction in the variance makes up for it, providing a lower MSE.

Part III

1. Compute an estimate of the proportion of students in the population who have used illegal drugs.

Recall that

\[
\hat{\pi} = \frac{\hat{\lambda} - \Pr(\text{Yes}|\text{Innocuous Question}) \cdot (1 - P)}{P}
\]

In this case \(\hat{\lambda} = 10/21\) which is the fraction of students who answered yes. The probability of yes given the social security question is 0.5. \(P\) which is the probability the person is asked the sensitive question has probability 4/5. Thus

\[
\hat{\pi} = \frac{(10/21) - 0.5 \cdot 1/5}{4/5} = 0.470
\]

2. Give a formula for the variance of your estimate in part A.

The formula is

\[
V(\hat{\pi}) = \frac{\hat{\lambda}(1 - \hat{\lambda})(N - n)}{P^2 \cdot n \cdot (N - 1)}
\]

Plugging in the numbers we get

\[
V(\hat{\pi}) = \frac{10/21(11/21)(200 - 21)}{16/25 \cdot 21 \cdot (200 - 1)} = 0.0166
\]

3. How would your answers to A and B change if there were only 5 red balls in the urn?

In this case

\[
\hat{\pi} = \frac{(10/21) - 0.5 \cdot 5/50}{45/50} = 0.473
\]

and

\[
V(\hat{\pi}) = \frac{10/21(11/21)(200 - 21)}{45^2/50^2 \cdot 21 \cdot (200 - 1)} = 0.013
\]
So the higher the percentage that the person answers the sensitive question, the lower the variance of the estimate.

4. Using the answers in part A and B, construct a 95% confidence interval for the population proportion who have used illegal drugs.

   The confidence interval is $\hat{p} \pm 1.96 \times \sigma_{\hat{p}}$. This is $0.470 \pm 1.96 \times \sqrt{0.0166} = (0.217, 0.722)$. 