36-303: Sampling, Surveys and Society

Statistics of Surveys III
Brian W. Junker
132E Baker Hall
brian@stat.cmu.edu

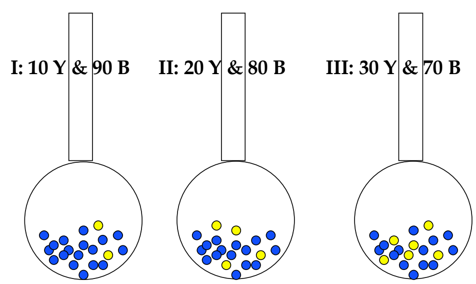
Handouts, Etc.

- Homework
- Lecture Notes
- Turn in Today:
  - HW01 in class on paper
  - I.2 in email (pdf or msword) by Midnight tonight
- FOR NEXT WEEK: Groves, Chapter 7
  - I'll start on survey questionnaires next Tues
- Turn in next week:
  - Tue: I.3 (more about this coming in email)
  - Thu: HW02

Outline

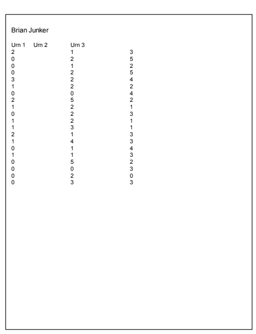
- Continuing our Survey Sampling Experiment
- Central Limit Theorem?
- Finite Population Correction

Continuing Survey Sampling Experiment



Continuing Survey Sampling Experiment

- Circulate all three urns
- Each student should mix the balls; then draw a sample and record # of yellows out of 10
  - Turn in a piece of paper with your name, and 3 neat columns of 20 results each (20 for each urn!)



Brian Junker

Urn 1	Urn 2	Urn 3
2	1	3
0	2	5
0	1	2
0	2	5
3	2	4
1	2	2
0	0	4
2	5	2
1	2	1
0	2	3
1	2	1
1	3	1
2	1	3
1	4	3
0	1	4
1	1	3
0	5	2
0	0	3
0	2	0
0	3	3

Central Limit Theorem for Surveys?

- For *simple random sampling (SRS) with replacement*,

$$E[\overline{X}] = \mu\,, \quad Var(\overline{X}) = \frac{\sigma^2}{n}$$

- The Central Limit Theorem then tells us

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

- σ is the SD of X_i; $\sigma/\sqrt{n}$ is the SE of $\overline{X}$
- *But in survey sampling we sample w/o replacement!*

Central Limit Theorem for Surveys?

- We will look at 500 draws from urn 3, at different sample sizes: [overprinted: 420 $\hat{p}$'s]
  - n=1, 2, 5, 10, 20, …, 98, 99
  - N=100 always
- Compare histogram of $\hat{p}$'s with a normal curve with the same center and spread as the $\hat{p}$'s
- *If CLT holds, histogram & curve will agree*

CLT Exploration in R here…

Conclusions from the CLT Exploration

- Small samples – CLT hasn't kicked in yet
- For "moderate" samples, CLT seems to work
- Moderate means … important to have n > 20 (or whatever rule of thumb), but also n/N has to be not close to 1
- CLT breaks when sample size is nearly whole population – then we are more certain about p, than CLT would have us believe

Finite Population Correction

- The goal is to figure out what the right SE is
- Requires us to "think differently" about sampling
- Involves a little bit of summation notation tedium
  - Statistics is sometimes like that: we "pay for" good insights with the need for tedious calculation…

Sampling from a Finite Population

- N = size of our fixed, finite population
- We want to measure Y. Y might be
  - income,
  - 'did you cheat'
  - number of "free" PAT bus rides taken…
- For each person in the population, Y *is not random*, it is a fixed value: $y_1, y_2, \ldots, y_N$
- *What is random* is whether the person gets in our sample or not:

$$Z_i = \begin{cases} 1, & \textit{if } i \textit{ is in our sample} \\ 0, & \textit{if } i \textit{ is not in our sample} \end{cases}$$

for i=1, 2, …, N

The Z's are a "trick" for thinking about how sampling works…

- Population size N = 10
- Sample size n = 3
- y's are respondents' ages

Nonrandom Population y_i's	44	35	21	62	27	19	23	56	28	45
Random sampling indicators Z_i's	0	0	1	0	0	1	1	0	0	0
Random sample of Y_i's			21			19	23			

Example: Drawing Balls from an Urn

- The colors of the 100 balls were not random. We could say

$$y_i = \begin{cases} 1, & \text{if ball is yellow} \\ 0, & \text{else} \end{cases}$$

- What was random was which 10 balls were drawn:
  - For 10 balls, $Z_i = 1$, for the rest, $Z_i=0$
  - We could write the fraction of yellows in the sample as

$$\hat{p} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1}{10}\sum_{i=1}^{100} Z_i y_i$$

Sampling Without Replacement

- Population size N
- Sample of size n without replacement.
- What is $P[Z_i=1]$?

$$\begin{aligned} P[Z_i = 1] &= \frac{\#(\text{samples of size } n \text{ including } i)}{\#(\text{all possible samples of size } n)} \\ &= \frac{\#(\text{put } i \text{ in sample}) \times \#(\text{samples of size } n-1 \text{ from the remaining } N-1)}{\#(\text{samples of size } n)} \\ &= \frac{1 \times \binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \quad \textit{(special case of hypergeometric distribution!)} \end{aligned}$$

Sampling Without Replacement

- The Z_i's are Bernoulli's with

$$E[Z_i] = \frac{n}{N}, \quad Var(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

- Therefore

$$\begin{aligned} E[\overline{Y}_{sample}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = E\left[\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{N} y_i E[Z_i] = \frac{1}{n}\sum_{i=1}^{N} y_i \frac{n}{N} \\ &= \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{y}_{pop} \end{aligned}$$

Sampling Without Replacement

- But the Z_i's are not independent,

$$\begin{aligned} E[Z_i Z_j] &= P[Z_i = 1 \cap Z_j = 1] \\ &= P[Z_j = 1 | Z_i = 1] P[Z_i = 1] \\ &= \left(\frac{n-1}{N-1}\right)\left(\frac{n}{N}\right) \end{aligned}$$

- We can calculate the covariance

$$\begin{aligned} Cov(Z_i, Z_j) &= E[Z_j Z_j] - E[Z_i]E[Z_j] \\ &= \left(\frac{n-1}{N-1}\right)\left(\frac{n}{N}\right) - \left(\frac{n}{N}\right)^2 \\ &= -\frac{1}{N-1}\left(1 - \frac{n}{N}\right)\left(\frac{n}{N}\right) \end{aligned}$$

- So having i "in" makes j a little less likely…

Sampling Without Replacement

$$\begin{aligned} Var(\overline{Y}_{sample}) &= Var\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i\Big) = Var\Big(\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\Big) \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 Var(Z_i) + \sum\sum\nolimits_{i \neq j} y_i y_j Cov(Z_i, Z_j)\right] \\ &= \frac{1}{n^2}\left[\left(\frac{n}{N}\right)\left(1 - \frac{n}{N}\right)\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\left(1 - \frac{n}{N}\right)\left(\frac{n}{N}\right)\sum\sum\nolimits_{i \neq j} y_i y_j\right] \\ &= \frac{1}{n^2}\left(\frac{n}{N}\right)\left(1 - \frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum\sum\nolimits_{i \neq j} y_i y_j\right] \\ &= \cdots = \left(1 - \frac{n}{N}\right)\frac{S_{pop}^2}{n} \end{aligned}$$

where $S_{pop}^2 = \sum_1^N (y_i - \overline{y}_{pop})^2/(N-1)$, the population variance.

The Finite Population Correction (FPC)

- We have seen that for SRS without replacement

$$\begin{aligned} E[\overline{Y}_{samp}] &= \overline{y}_{pop} & &(\overline{Y}_{samp} \textit{ is unbiased}) \\ Var(\overline{Y}_{samp}) &= (1-f)S_{pop}^2/n, & &f = n/N \end{aligned}$$

- The quantity (1-f) is called the *finite population correction (fpc).*
  - *When n/N $\approx$ 0, (1-f) $\approx$ 1, so "With or without replacement doesn't matter for small SRS's!"*
  - *As n/N -> 1, (1-f) -> 0 and SE($\overline{y}_{samp}$) ->0. "We don't need statistical estimates for a true census!"*

FPC, continued

- In practice we replace S^2_{pop} with s^2_{samp}

$$Var(\overline{Y}_{samp}) \approx (1-f)s^2/n,$$
$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y}_{samp})^2$$

- When y_i = 0 (blue ball) or 1 (yellow ball), one can show, since $\overline{y}_{samp} = \hat{p}$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \hat{p})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

and so

$$Var(\hat{p}) \approx (1-f)\frac{1}{n-1}\hat{p}(1-\hat{p})$$

Returning to our Sampling Experiment…

- The SE under SRS ***w/o*** replacement should have been

$$SE(\hat{p}) = (1-f)\hat{p}(1-\hat{p})/(n-1)$$

rather than

$$SE(\hat{p}) = \hat{p}(1-\hat{p})/(n-1)$$

- So in our survey experiment we hope to see that estimated SE's from SRS ***with*** replacement were *too large*.

Review

- Continuing our Survey Sampling Experiment
- Central Limit Theorem?
- Finite Population Correction

- FOR NEXT WEEK: Groves, Chapter 7
  - I'll report on our Survey Sampling Experiment
  - I'll start on survey questionnaires next Tues
- Turn in next week:
  - Tue: I.3 (more about this coming in email)
  - Thu: HW02