STA304 Assignment 2

Yu-Chun Chien

3/11/2021

Question 1: Mainstreet Research Survey

A. Survey & Parameter of Interest

Chosen question

"The current Ontario sex-ed curriculum includes lessons that teach gender identity theory. That is, that there are many genders other than male and female. Do you agree or disagree with teaching gender identity theory to children in Ontario elementary schools?"

Parameter of Interest

Proportion of the respondents that selected strongly agree for the survey question

B. Estimation of Population Parameter

i. Weighted Frequency

Estimation of parameter using gender as the stratification variable:

$\hat{p_s} \pm 2\sqrt{\hat{Var}(\hat{p_s})}$: $0.3015917 \pm 0.01429863 = (0.2872949, 0.3158885)$

ii. Unweighted Frequency

Estimation of parameter using gender as the stratification variable:

$\hat{p_s} \pm 2\sqrt{\hat{Var}(\hat{p_s})}$: $0.2920089 \pm 0.01416435 = (0.2778445, 0.3061732)$

C. Compare Two Estimates

The value of the two estimates are similar. For unweighted frequency, the data is solely stratified based on the information that the respondents provided in the survey, which is the gender of the respondents. In constrast, for weighted frequency, the data is weighted based on additional information obtained by ways such as other survey or open resources available about the true gender proportion of the population. Since this additional adjustment is made after the observation, the weighted frequency is a post-stratified estimate. By making the adjustment, the results of the analysis might possibly become more accurate as the proportion of male and female is more similar to the entire population.

Question 2: Baseball Dataset

A. Stratified Random Sample

```
# get information about the teams and proportions
table(baseball_yc$team)
```

```
##
## ANA ARI ATL BAL BOS CHA CHN CIN CLE COL DET FLO HOU KCA LAN MIL MIN MON NYA NYN
##  26  28  28  25  27  26  29  27  28  27  26  26  25  27  24  25  25  28  29  26
## OAK PHI PIT SDN SEA SFN SLN TBA TEX TOR
##  27  25  27  26  27  28  26  26  27  26
```

```
team_size_yc <- c(26, 28, 28, 25, 27, 26, 29, 27, 28, 27, 26, 26, 25, 27, 24, 25, 25, 28, 29, 26, 27, 25

for(i in 1:30){
  team_size_yc[i] <- team_size_yc[i]/ 797 * 150
}
team_size_yc
```

```
##  [1] 4.893350 5.269762 5.269762 4.705144 5.081556 4.893350 5.457967 5.081556
##  [9] 5.269762 5.081556 4.893350 4.893350 4.705144 5.081556 4.516939 4.705144
## [17] 4.705144 5.269762 5.457967 4.893350 5.081556 4.705144 5.081556 4.893350
## [25] 5.081556 5.269762 4.893350 4.893350 5.081556 4.893350
```

```
# SRSWOR of 150 samples
set.seed(4380)
players_yc <- strata(baseball_yc, c("team"), size = team_size_yc, method = "srswor")
```

Steps for Selecting Stratified Sample

1. Use the table() function to count the number of players in each team.
2. After obtaining the number of players in each team, calculate n_i by using proportional allocation, $n_i = n \times \frac{Ni}{N}$.
3. using the strata() function in package "sampling", take stratified random sample without replacement, with sample size found in step (2).

B. Mean of Log Salary

Mean of ln(salary): 13.84192

95% CI: (13.74520, 13.93863), with margin of error = 2×0.09671471

C. Proportion of Pitchers

Proportion of ployers in the data set who are pitchers: 0.4577792

95% CI: (0.4561463, 0.4594121), with margin of error = 2×0.001632907

D. Proportion of Pitchers SRS

Proportion of ployers in the data set who are pitchers: 0.42

95% CI: (0.4186728, 0.4213272), with margin of error = 2×0.001327202

The estimate of proportion using simple random sampling is 0.04 smaller than the estimate of proportion using stratified random sampling. Using simple random sampling, the variance of sample proportion is much smaller than the variance of sample proportion using stratified random sampling.

E. Sample Variance of Log Salary

When using proportional allocation, we assume that the cost and variance of each stratum is similar. However, the variance between the stratums are quite different. Thus, instead of using proportional allocation, optimal

allocation would be better for this problem, since optimal allocation take into account the difference in variance between each stratum to determine the allocation.

F. Population Stratum Variances & Optimal Allocation

Sample Sizes for Optimal Allocation

By applying optimal allocation, the number of sample size in each strata is as follows: 9, 6, 9, 2, 12, 10, 6, 3, 6, 1, 2, 0, 0, 4, 5, 2, 5, 6, 4, 8, 5, 6, 1, 8, 2, 5 ,9, 1, 8, 3. The total sample size is 148 after rounding the decimals. There is two strata with sample size equal to zero after rounding, which would potentially make the two teams underrepresented. Thus, we fix this by sampling one unit from each team, making the total sample size up to 150.

The sample size of each strata by using proportional allocation is: 4, 5, 5, 4, 5, 4, 5, 5, 5, 5, 4, 4, 4, 5, 4, 4, 4, 5, 5, 4, 5, 4, 5, 4, 5, 4, 5, 5, 4, 4, 5, 4

The sample size of each strata by using optimal allocation is: 9, 6, 9, 2, 12, 10, 6, 3, 6, 1 , 2, 0, 0, 4, 5, 2, 5, 6, 4, 8, 5, 6, 1, 8, 2, 5, 9, 1, 8, 3.

By incorporating variance to decide allocation, optimal allocation differs significantly from proportional solution in this context. In our data set, each team might account for relatively similar proportion, but they differ quite a lot in variance of ln of salary. Thus, we could observe that optimal allocation suggested that we should sample more units from the teams with bigger variance in order to make a more accurate estimate of the entire data.

Appendix

Sample Variances of Logsal in Each Stratum

```
##          1          2          3          4          5          6          7
## 1.92700631 1.74564771 1.93946275 1.38054704 0.34230209 0.70000350 0.24741644
##          8          9         10         11         12         13         14
## 0.44572780 1.25214015 2.00310039 0.88227064 0.42057098 2.56063929 0.94450569
##         15         16         17         18         19         20         21
## 0.61371997 0.01494245 1.96558729 0.81140425 2.18979831 1.45228584 0.71762956
##         22         23         24         25         26         27         28
## 1.79626824 0.43117455 0.39714938 0.27987048 0.90346938 2.93335875 0.97548276
##         29         30
## 0.02760052 0.82654658
```

Population Stratum Variances of Logsal in Each Stratum

Using $\sigma^2 = s^2 \times \frac{N-1}{N}$, the population variance is estimated using the sample variance.

```
##  [1] 1.85289068 1.68330315 1.87019622 1.32532516 0.32962423 0.67308029
##  [7] 0.23888484 0.42921936 1.20742086 1.92891148 0.84833715 0.40439518
## [13] 2.45821372 0.90952399 0.58814830 0.01434476 1.88696380 0.78242553
## [19] 2.11428802 1.39642869 0.69105069 1.72441751 0.41520512 0.38187440
## [25] 0.26950490 0.87120262 2.82053726 0.93796419 0.02657828 0.79475632
```