BST5420 Sampling Theory and Survey Design Homework 1

Miao Cai[*]

2019-02-11

1. (2 Points) Suppose that a population consists of 50 individuals, numbered 1 through 50. The following three samples represent (a) a simple random sample, (b) a stratified random sample, and (c) a cluster sample (but not necessarily in that order). Note that each is a sample of size 15.

1. Sample 1 = 6, 7, 8, 9, 10, 31, 32, 33, 34, 35, 41, 42, 43, 44, 45
2. Sample 2 = 3, 4, 9, 11, 17, 19, 23, 25, 26, 31, 33, 38, 44, 45, 50
3. Sample 3 = 1, 3, 8, 9, 10, 18, 23, 26, 27, 29, 31, 32, 34, 38, 48

Which is which? (Hint: Find the ones that could be stratified and cluster samples, and the one left over is the SRS.

Answers:

- Sample 1 seems to be cluster samples since there are three clusters in the sample: {6, 7, 8, 9, 10}, {31, 32, 33, 34, 35}, {41, 42, 43, 44, 45}.
- Sample 2 seems to be stratified samples since there are samples from all 5 stratas: { 3, 4, 9 } from strata 0 - 10, { 11, 17, 19 } from strata 11 - 20, { 23, 25, 26 } from strata 21 - 30, { 31, 33, 38 } from strata 31 - 40, { 44, 45, 50 } from strata 41 - 50.
- Sample 3 is more like simple random sample since I don't see any significant pattern in the sample.

2. (2 Points) Do problem 10 of Chapter 2. "Which of the following SRS designs will give the most precision . . . "
Which of the following SRS designs will give the most precision for estimating a population mean? Assume that each population has the same value of the population variance S^2.

1. An SRS of size 400 from a population of size 4000
2. An SRS of size 30 from a population of size 300
3. An SRS of size 3000 from a population of size 300,000,000

Answers:

The variance of sample mean estimate is $\frac{s^2}{n}\frac{N-n}{N} = s^2(\frac{1}{n} - \frac{1}{N})$. Therefore, below are the variances of each of the samples:

- $\frac{s^2}{n}(1 - \frac{n}{N}) = s^2(\frac{1}{400} - \frac{1}{4000}) = \frac{9}{4000}s^2$
- $\frac{s^2}{n}(1 - \frac{n}{N}) = s^2(\frac{1}{30} - \frac{1}{4000}) = \frac{9}{300}s^2$
- $\frac{s^2}{n}(1 - \frac{n}{N}) = s^2(\frac{1}{3000} - \frac{1}{300000000})$

```
variance = c(9/4000, 9/300, (1/3000 - 1/300000000))
which(variance == min(variance))
```

[*] PhD student, Department of Epidemiology and Biostatistics, College for Public Health and Social Justice, Saint Louis University. Email address miao.cai@slu.edu

```
## [1] 3
```

Therefore, the last sample "An SRS of size 3000 from a population of size 300,000,000" has the most precise estimate of the population mean.

3. (6 Points) Do problem 33 in Chapter 2 of the textbook. "Online bookstore. The website amazon.com can be used to obtain populations of books, CDs, and other wares." Use a sample size of 10, rather than the 50 that the problem asks for in part c. Be creative in your choice of genre, but don't be too general.

1. Do part (a). In the books search window, type in a genre you like, such as mystery or sports; you may want to narrow your search by selecting a subcategory since an upper bound is placed on the number of books that can be displayed. Choose a genre with at least 20 pages of listings. The list of books forms your population.
2. Do part (b). What is your target population? What is the population size, N?
3. Do part (c) but don't worry about the amount of time you spent collecting data. Take an SRS of 10 books from your population. Describe how you selected the SRS, and record the amount of time you spent taking the sample and collecting the data.
4. Record just the price. Record the following information for each book in your SRS: price
5. Do part (e). Give a point estimate and a 95
6. Skip part (f)
7. Do part (g). Explain, to a person who knows no statistics, what your estimates and CIs mean.

Answers:

1. I choose the key words "Bayesian data statistics"[1]
2. The target population is all the book results returned by Amazon. The population size is 506.
3. I used a random number generated in R to get 10 unrepeated samples from a population `1:506`, and I get the information of the 10 books from Amazon.

The book names are:

x
Bayesian Methods for Data Analysis
Bayesian Statistics: An Introduction, 4th Edition: An Introduction, 4th Edition
Bayesian Analysis of Gene Expression Data (Statistics in Practice)
Machine Learning for Decision Makers: Cognitive Computing Fundamentals for Better Decision Making
Interval-Censored Time-to-Event Data: Methods and Applications (Chapman & Hall/CRC Biostatistics Series)
Analysis of Metal-loss Corrosion on Energy Pipelines Based on ILI Data: A Bayesian Approach
Advanced Techniques in Knowledge Discovery and Data Mining (Advanced Information and Knowledge Processing)
Multiscale Modeling: A Bayesian Perspective (Springer Series in Statistics)
Dynamical Search: Applications of Dynamical Systems in Search and Optimization (Chapman & Hall/CRC Interdisciplinary Statistics)
Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS (Wiley and SAS Business Series)

[1] https://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Dstripbooks&field-keywords=bayesian+data+analysis&rh=n%3A283155%2Ck%3Abayesian+data+analysis

4. The price of the 10 SRS books are: 80.28, 52.11, 77.13, 29.99, 77.99, 84, 68.5, 44.99, 68.23, 23.14. The sample mean is 60.64, while sample standard deviation is 21.83.
5. The formula to estimate confidence interval for the population mean is $\bar{x} \pm t * \sqrt{\frac{s^2}{n}\frac{N-n}{N}}$. The degrees of freedom is $n - 1 = 9$, so the t statistic is 2.26. The confidence interval can be calculated as follows in R:

```r
N = 506
n = 10
t = qt(0.975, n - 1)
s = sd(dat$price)
xbar = mean(dat$price)
(cl_left = round(xbar - t*sqrt(s^2/n*(N-n)/N), 2))
```

```
## [1] 45.18
```

```r
(cl_right = round(xbar + t*sqrt(s^2/n*(N-n)/N), 2))
```

```
## [1] 76.09
```

6. The mean price of book with keywords “Bayesian data analysis” on Amazon is about 60.636 US dollars. If we take samples for a lot of times, 95% times the confidence intervals will capture the population mean.