The Effects of Department Popularity on Professors' Salaries

Meggie Huang	18716225	data analysis, references, data collection and summaries
Yuxi Liu	45149663	data analysis, data collection and summaries, appendix
Avril Yiheng Li	18780353	data analysis, data collection and summaries, appendix
Yiling Fan	81527616	introduction, data collection and summaries, discussion and conclusion
Sophia Zhang	65921389	introduction, data collection and summaries, discussion and conclusion

Department of Statistics, University of British Columbia

STAT 344: Sample Surveys

Dr. Lang Wu

November 13, 2020

1. Introduction

Objectives

The objective is to investigate whether the salaries of UBC professors in five popular Science departments (Mathematics, Computer Science, Statistics, Chemistry, and Physics) are higher than that of all UBC professors in the Faculty of Science.

We will apply sample surveys in simple random sampling method (SRS) and two different stratified sampling methods: stratified by positions and by departments. For stratification by position, the strata are: assistant professors, associate professors and professors. For stratification by department, we focus on the following departments: Mathematics, Statistics, Computer Science, Physics, and Chemistry.

Background

We discovered that there are GPA differences in admission when Science students enter different specializations in their second years. Competitive and popular specializations such as Computer Science, Statistics, Maths, Physics and Chemistry have higher GPA cutoffs compared to other specializations in the Faculty of Science. We wish to know whether professors in more popular departments receive more appealing salaries.

2. Data Collection and Summaries

Data Collection

Targeted Population

The targeted populations are the assistant, associate, and full professors in the Mathematics, Statistics, Computer Science, Physics, and Chemistry departments for the 2019 academic year. For the sake of simplicity, we did not make a distinction between whether a professor of any level of seniority was one of teaching or not. For instance, a professor of teaching was recorded as just a professor. Total population: N = 229.

Parameters of Interest

1) Average professors' salaries in five selected departments (Mathematics, Statistics, Computer Science, Physics, and Chemistry).

2) Proportion of salaries over $113,417.7, which is the mean salary in the Faculty of Science (Marmer and Sudmant, 2010, p. 4).

How and where the data is collected?

1) To avoid non-representative samples, we divide the entire population into homogeneous groups. We choose position and department for stratification, since they have a similar level of variance in each sub-group. For positions, we consider assistant, associate, and full professors. For departments, we consider Mathematics, Statistics, Computer Science, Physics, and Chemistry.

2) We check the lists of assistant, associate, and full professors from the web page of the following departments: Mathematics, Statistics, Computer Science, Physics, and Chemistry. The web pages are under the people directory in the official website of University of British Columbia.

3) Collect salary data of targeted professors from consolidated financial statements.

4) Build an excel spreadsheet consisting of professors' name, sex, position, salary, and department.

Sampling Methods

We set the sample size to be n = 50 for all methods. Sample size cannot be too small, otherwise it would limit the power of the study and increase the margin of error. And for our relatively small total population, sample size cannot be too large. Hence, 50 is a reasonable sample size. For each method, we arbitrarily chose to use set.seed(1) so that we would be performing different types of estimates on the same data. This allows for a more meaningful comparison of the different sampling methods.

Simple Random Sampling

1) Use the sample function in R to obtain a simple random sample

Stratified Sampling by Position

1) Assume within-strata variance of three assistant, associate, and full professors are the same. The proportion of assistant professors, associate professors, and professors are 19/229, 63/229, and 147/229 respectively. Hence, in the sample, we have 4 assistant, 14 associate, and 32 full professors.

Stratified Sampling by Department

1) We take separate random samples in each sub-population. Using proportional allocation, we obtain strata sample sizes of 4,12,13,11,10 for statistics, physics, mathematics, computer science and chemistry respectively.

Two different stratified sampling methods were used in order to find a more precise way to estimate the mean and proportion. The method that results in a larger between-strata variance will give a narrower confidence interval. After computing results for three different sampling methods in R, we compare the estimates with the average salary for the Faculty of Science found by Marmer and Sudmant (2010).

3. Data Analysis

Results and Formulas

Please refer to the appendix for computations

Simple Random Sample

1) Mean response

Estimate: $\bar{y_{\mathcal{S}}} = \frac{1}{n}\sum_{i \in \mathcal{S}} y_i = 164219.60$

Standard error: $SE(\bar{y_{\mathcal{S}}}) = \sqrt{\frac{s_{\mathcal{S}}^2}{n}\left(1 - \frac{n}{N}\right)} = 5807.08$ where $s_{\mathcal{S}} = \sqrt{\frac{1}{n-1}\sum_{i \in S}(y_i - \bar{y}_{\mathcal{S}})^2}$

95% CI: $\bar{y}_{\mathcal{S}} \pm 1.96 * SE(\bar{y_{\mathcal{S}}}) = (152837.7,\ 175601.4)$

2) Proportion

Estimate: $\hat{p} = 0.84$ Standard error: $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.04583776$

95% CI: $\hat{p} \pm 1.96 * SE(\hat{p}) = (0.750158,\ 0.929842)$

Stratification by Position

1) Mean response

Estimate:

$$\bar{y}_{str} = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\bar{y}_{\mathcal{S}_h} = 183022.94$$

Standard error:

$$SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1 - \frac{n_h}{N_h}\right)\frac{s_{\mathcal{S}_h}^2}{n_h}} = 4031.55$$

95% CI:

$$\bar{y}_{str} \pm 1.96 * SE(\bar{y}_{str}) = (175121.1, 190924.8)$$

2) Proportion

Formulas are the same as those for mean response. We denote the proportion of professors with a salary over \$113417.70 in stratum h by $\hat{p}_{\mathcal{S}_h}$. For the estimate, we replace the $\bar{y}_{\mathcal{S}_h}$ with $\hat{p}_{\mathcal{S}_h}$.

To compute the standard error, we replace $s_{\mathcal{S}_h}^2$ with $\hat{p}_{\mathcal{S}_h}(1 - \hat{p}_{\mathcal{S}_h})$.

Estimate: 0.8987445 Standard error: 0.03476802 95% CI: (0.8305992, 0.9668899)

Stratification by Department

Note that the formulas are the same as those for stratification by position

1) Mean response

Estimate: 179284.127 Standard error: 4317.851 95% CI: (170821.1, 187747.1)

2) Proportion

Estimate: 0.92794760 Standard error: 0.02329766 95% CI: (0.8822842, 0.9736110)

Interpretation and Analysis

In these three study designs, there are two assumptions made:
1) Samples from different strata are independent, so the covariance terms are equal to 0.
2) Proportional allocation is optimal because the within-strata variance of salaries is the same for each stratum and the cost of obtaining samples from each stratum is equal.

In SRS, the standard error of 5807.08 is the largest among the three sampling methods as there is the total variance. Hence, it may not be optimal for computing the estimated mean of professors' salaries. Despite having the largest standard error, SRS still has some advantages: it is easier to conduct and its results are more easily understood by those with limited knowledge of statistics. Standard errors of the mean response in stratification by position and department are about 44% and 34% smaller respectively compared to the standard error for SRS. This result is to be expected as stratification contains only the within-strata variance, whereas SRS is the total variance. Another advantage of stratification is that it avoids unrepresentative samples and ensures estimates of reasonable precision for small strata. Among the three designs, stratification by position is optimal for estimating the mean response as it has smallest standard error.

The purpose of performing stratification by position and department is to determine which one would give narrower confidence intervals. It was stated in class that when stratified sampling helps to greatly reduce within-strata variance, it will give a notably smaller standard error compared to SRS. The estimated mean response of salaries in three designs are all higher than the salaries in Science Faculty ($113,417.7). This demonstrates that professors in these five popular departments have higher average salaries compared to the average for all professors in Faculty of Science.

The estimated proportion of professors that have a salary higher than the average ($113417.70) is around 0.8 to 0.9 by these three methods. This result further confirms that the income for professors in these departments is generally higher than the salary of those in other departments, and assures us that our previous result is reasonable. The standard error of the proportion in stratification by department is the smallest (0.0233), indicating that stratification by department is optimal for estimating the proportion. An interesting result is that stratification by position is optimal for estimating the mean response, but stratification by department is optimal for estimating the proportion. A potential reason for this is that stratification by department produces smaller strata size, so we are more likely to have a high proportion of professors in each strata who make more than the mean salary of $113417.70. This will make $\hat{p}$ larger and the corresponding SE smaller.

4. Discussion and Conclusion

Discussion

Due to the simple nature of this project, our results have numerous limitations. Some of these limitations are:
(i) we did not account for the number of courses taught (professors who teach more courses will receive higher salaries)
(ii) the independence assumption for stratification by department may not be valid as salaries for certain departments may be correlated with each other. This is due to the fact that some faculty members are part of more than one department
(iii) we assumed that the strata have equal variance. This may not hold and it might be necessary to obtain larger samples from strata with higher variance. Note that cost is not an issue here, as the effort required to obtain the salary of one individual versus another is not notably different.
(iv) not all faculty members can be found in the report, as it only contains the salaries of those who have relatively higher salaries. The limitations above cause us to be unable to generalize our conclusions to larger or other populations.

We extracted the professor name list from departments web page in 2020, but the salary report was for the 2019 academic year. There may be an inconsistency in people, since some professors might have left or be new to UBC between 2019 and 2020. Also, due to the availability of data, we collect and estimate salaries data from the 2019 report, but this data is compared with the average salaries of professors in 2007. There is a potential increase in total salary for Faculty of Science between 2007 and 2019 (professors are usually promoted over time, there are new faculty members), which leads to an inaccurate comparison. Furthermore, due to inflation, the overall salaries for professors has risen from 2007 to 2019. This causes the proportions obtained for salaries higher than \$113,417.7 to be larger than if we had controlled for inflation.

Conclusion

The estimated mean response of salaries in three methods are all higher than the salaries in the Faculty of Science(\$113,417.7). The estimated proportion of professors that have a salary higher than the average(\$113417.7) is around 0.8 to 0.9 by these three methods. Therefore, on average, the salaries of UBC professors in the five selected departments are higher than the average of all UBC professors in the Faculty of Science. The popularity of departments can have an impact on professors' salaries, with professors in more popular departments receiving higher salaries on average.

Appendix

```
 Simple Random Sample
> cleaned = read.csv("˜/Downloads/cleaned.csv", sep = ",")
> View(cleaned)
> N = length(cleaned$Name)
> n = 50
> set.seed(1)
> mysample = sample(N,n, replace = FALSE)
> mysample.Name = cleaned$Name[mysample]
> sample = subset (cleaned, cleaned$Name %in% mysample.Name)
> View(sample)
> sample.salaries = as.numeric(gsub(",","", sample$Salary))
# sample mean
> sample.mean = mean(sample.salaries)
> sample.mean
[1] 164219.6
# sample variance
> sample.variance = var(sample.salaries)
# standard error
> se = sqrt((1-n/N)*sample.variance/n)
> se
[1] 5807.081
# 95% confidence level for population mean
> CI = c(sample.mean-1.96*se, sample.mean +1.96*se)
> CI
[1] 152837.7 175601.4
```

```
Simple Random Sample for Proportion
set.seed(1)
professor_sample=sample(cleaned$Salary,n,replace=FALSE)
professor_sample<- as.numeric(gsub(",","", professor_sample))
sum(professor_sample>113417.7)
[1] 42
#sample proportion
p_hat=42/n
[1] 0.84
#sample variance
se = sqrt((1-n/N)*(0.84*0.16)/n)
se
[1] 0.04583776
# 95% confidence level for population proportion
CI = c(0.84-1.96*se, 0.84 +1.96*se)
CI
[1]  0.750158 0.929842
```

```
 Stratification by position
> cleaned<-cleaned[,2:3]
> professor<-cleaned[cleaned$Position=="Professor",]
> Assistant_professor<-cleaned[cleaned$Position=="Assistant Professor",]
> Associate_professor<-cleaned[cleaned$Position=="Associate Professor",]
> sum(cleaned$Position=="Professor")=147
> sum(cleaned$Position=="Assistant Professor")=19
> sum(cleaned$Position=="Associate Professor")=63

# due to proportional allocation, the number of sampled units in these
three strata is 32,4,14
set.seed(1);
professor_sample=sample(professor$Salary,32,replace=FALSE)
Assistant_sample=sample(Assistant_professor$Salary,4,replace=FALSE)
Associate_sample=sample(Associate_professor$Salary,14,replace=FALSE)
professor_sample<- as.numeric(gsub(",","", professor_sample))
Assistant_sample<- as.numeric(gsub(",","", Assistant_sample))
Associate_sample<- as.numeric(gsub(",","", Associate_sample))

strata_mean=c(mean(professor_sample),mean(Assistant_sample),
mean(Associate_sample))
se_strata=c(sqrt(var(professor_sample)/32),sqrt(var(Assistant_sample)/4),
sqrt(var(Associate_sample)/14))
weight_strata=c(147/229,19/229,63/229)
estimate=sum(strata_mean*weight_strata)
se=sqrt((147/229)^2*(1-32/147)*var(professor_sample)/32+(19/229)^2*(1-4/19)
*var(Assistant_sample)/4+(63/229)^2*(1-14/63)*var(Associate_sample)/14)
print(c(estimate,se))= [183022.94,4031.55]
CI=c(estimate-1.96*se,estimate+1.96*se)
print(c(CI))=[175121.1 ,190924.8]
Assumptions: we assume the variance  within three strata(professor,assistant
professor and associate professor)are equal.
```

```
proportion estimate
> sum(professor_sample>113417.7)
[1] 30
> sum(Assistant_sample>113417.7)
[1] 2
> sum(Associate_sample>113417.7)
[1] 13
# the proportion estimated by strata sampling
strata_proportion=147/229*30/32+19/229*2/4+63/229*13/14
> strata_proportion
[1] 0.8987445
# the se estimated by strata sampling
> se.strata=sqrt((147/229)^2*(1-32/147)*(30/32)*(2/32)/32+(19/229)^2
*(1-4/19)*(2/4)*(2/4)/4+(63/229)^2*(1-14/63)*(13/14)*(1/14)/14)
> se.strata
[1] 0.03476802
# the confidence interval of proportion
> CI = c(strata_proportion-1.96*se.strata, strata_proportion +1.96*se.strata)
> CI
[1] 0.8305992 0.9668899
```

```
Stratification by department

prof = read.csv(’cleaned.csv’, header = T)
attach(prof)
prof$Salary = gsub(’,’, ’’, Salary) # remove commas in salaries
prof$Salary = as.numeric(prof$Salary)
N.h = tapply(Salary, Department, length)
depts = names(N.h)
N = sum(N.h)
n = 50

n.h <- round((N.h/N)*n)
STR.sample <- NULL
set.seed(1)
for (i in 1: length(depts)) {
  row.indices <- which(Department == depts[i])
  sample.indices <- sample(row.indices, n.h[i], replace = F)
  STR.sample <- rbind(STR.sample, prof[sample.indices, ])
}

ybar.h <- tapply(STR.sample$Salary, STR.sample$Department, mean)
var.h <- tapply(STR.sample$Salary, STR.sample$Department, var)
se.h <- sqrt((1 - n.h / N.h) * var.h / n.h)

ybar.str <- sum(N.h/N * ybar.h)
FPC = 1-(N.h/N)
se.str <- sqrt(sum((N.h/N)^2 * FPC * se.h^2))
str <- c(ybar.str, se.str)
> str
[1] 179284.127   4317.851
CI = c(ybar.str-1.96*se.str, ybar.str+1.96*se.str)
> CI
[1] 170821.1 187747.1
```

```
# estimate proportion of profs with salaries above the mean
for the faculty of science

STR.sample$Salary = as.numeric(STR.sample$Salary > 113417.7)

ybar.h <- tapply(STR.sample$Salary, STR.sample$Department, mean)
var.h <- tapply(STR.sample$Salary, STR.sample$Department, var)
se.h <- sqrt((1 - n.h / N.h) * var.h / n.h)
rbind(ybar.h, se.h)

ybar.str <- sum(N.h/N * ybar.h)
FPC = 1-(N.h/N)
se.str <- sqrt(sum((N.h/N)^2 * FPC * se.h^2))
str <- c(ybar.str, se.str)
> str
[1] 0.92794760 0.02329766
CI = c(ybar.str-1.96*se.str, ybar.str+1.96*se.str)
> CI
[1] 0.8822842 0.9736110
```

References

Marmer, O., Sudmant, W. (2010), *Statistical Analysis of UBC Faculty Salaries: Investigation of Differences Due to Sex or Visible Minority Status*
https://equity.ubc.ca/files/2010/06/salary_analysis.pdf

Professors' Salaries
https://finance.ubc.ca/sites/finserv.ubc.ca/files/FY19_Financial_Information_Act_Report.pdf