STAT 344

Group Project Report: Average of UBC Professors' Salaries

With Appendices

Presented to:

Professor Lang Wu

Date Submitted: November 11, 2022

Section 101

Han Cho (Data collection and summary and group leader) – 13852249

Amy Perna (Data analysis writing and Part 2) - 34956771

Joel Vandervalk (Data analysis R portion) - 98820277

Yifan Wu (Part 2 and proofreading) - 99677320

Part 1

1.1 Introduction (background, objectives)

This report aims to estimate two parameters: the average salary of all professors at the University of British Columbia and the proportion of professors earning $135,000 or more. This paper uses two sampling methods, simple random sampling, and stratified sampling, to estimate the parameters. As full-time students studying statistics and members of the UBC community, we are interested in taking advantage of the information provided by the institution regarding the compensation of the professors providing our education.

1.2 Data collection and Data summary

The name and salaries of full-time UBC professors are obtained from the Schedule of Remuneration and Expenses Paid to Employees dataset in the March 2021 Statement of Financial Information [1]. The target population is 7019 full-time UBC professors.
For the continuous variable, the parameter of interest is the salary of each UBC professor. For the binary variable, the parameter of interest is the proportion of professors with a salary over $135,000 CAD. We chose a sample size of 400 to ensure the sample size is a good portion of the target population.
The two sampling methods used are:

1) Simple Random Sampling

   For simple random sampling, 400 random data are selected from the population by sample function in R. One advantage of simple random sampling is that it is easier to obtain samples than stratified sampling as it does not require any extra information other than professors' names and salaries. One disadvantage of simple random sampling is that it might not lead to the most accurate estimate of the population characteristics.

2) Stratified Sampling

   Before Stratification, we first made assumptions that professors doing research supervising would make more money than the lecturers. Therefore, we used the research supervisor directory to stratify the population into research supervisors and non-research supervisors. Based on our data, 27% of UBC professors are research supervisors, and

73% are non-research supervisors (Table 1) [2]. We used optimal allocation for stratified sampling using $s_{h,guess}$ from a similar study conducted previously.

$\frac{n_h}{n} = \frac{N_h * \left(\frac{S_{h,guess}}{\sqrt{c_h}}\right)}{\sum_{k=1}^{H} \frac{N_k\left(S_{k,guess}\right)}{\sqrt{c_k}}}$ since c is equal, it reduces to $\frac{n_h}{n} = \frac{N_h * s_{h,guess}}{\sum_{k=1}^{H} n_k\left(s_{k,guess}\right)}$.

The resulting sample includes 154 Research Supervisors and 246 Non-Research Supervisors, with a total sample size of 400. One advantage of stratified sampling is that its sample is more likely to be representative than simple random sampling. One disadvantage of stratified sampling is that it requires a cost to stratify the population.

Table 1: Research supervisor number and proportion

	Number	Proportion
Research Supervisors	1907	0.27
Non-Research Supervisors	5112	0.73
Total	7019	1.00

1.3 Data analysis

1.3.1 Analysis on Continuous Variable

We estimate the average salary of the UBC professors' population by calculating the average salary of the samples ($\overline{y_s}$).

1.3.1.1 Analysis on Continuous Variable by Simple Random Sampling

We first took a simple random sample (SRS) of size 400 from the overall population. We used the 'sample' function in R to take a sample of the size 400 from the total population without replacement.

We observed:

$$\overline{y_s} = 134424.3 \qquad \text{and} \qquad SE(\overline{y_s}) = \sqrt{\left(1 - \frac{n}{N}\right)\left(\frac{var(\overline{y_s})}{n}\right)} = 2802.022$$

Our estimation of the average salary of a UBC professor by SRS is $134,424.30 CAD with a standard error of $2,803.02 CAD.

We used these estimates to calculate a 95% confidence interval for the average salary of a UBC professor:

$$\overline{y_s} \pm 1.96 \times SE(\overline{y_s}) = (128932.3,\ 139916.256)$$

Based on our SRS, we are 95% certain that the true average salary of professors at UBC is between $128,932.30 CAD and $139,916.26 CAD. While our SRS is representative of the whole population, it is possible that this sampling method has not led to the most accurate estimate of average salary. We will also take a stratified sample to see if we are able to obtain a better estimate.

1.3.1.2 Analysis on Continuous Variable by Stratified Sampling

We estimated the average salary from a stratified sample ($\overline{y_{str}}$), and the sample was stratified on whether a professor is a Research Supervisor or a Non-Research Supervisor. The resulting size of the sample containing the Research Supervisor was 154, and the sample size of the Non-Research Supervisor was 246 based on optimal allocation determined by the guessed variance from the previous study. From these samples, we observed that:

$$\overline{y_{str}} = 133430.0 \qquad \text{and} \qquad SE(\overline{y_{str}}) = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \left(\frac{s_{s_h}^2}{n_h}\right)} = 2249.881$$

Based on stratified sampling, we estimate the average salary of a UBC professor to be $133,430.00 CAD with a standard error of $2,249.88 CAD.

We observed that a stratified sample gives a lower standard error, leading us to believe that dividing the data based on whether a professor is a research supervisor or not and combining the results from both strata gives us a better estimate of the average salary.

We used these estimates to calculate a 95% confidence interval for the average salary of a UBC professor:

$$\overline{y_s} \pm 1.96 \times SE(\overline{y_{str}}) = (129020.3,\ 137839.809)$$

Based on our stratified samples, we are 95% certain that the true average salary for professors at UBC lies between $129,020.30 CAD and $137,839.81.

1.3.2 Analysis on binary Variable

We estimate the proportion of professors earning a salary of more than $135,000 CAD ($p$) by calculating the sample proportion.

1.3.2.1 Analysis on Binary Variable by Simple Random Sampling

We used the same simple random sampled population taken in section 1.3.1.1 to begin our analysis of our binary variable to estimate the proportion of professors making more than $135,000 CAD per year.

We observed that:

$$\hat{p} = 0.390 \qquad \text{and} \qquad SE(\hat{p}) = \sqrt{\left(1 - \frac{n}{N}\right)\left(\frac{\hat{p}(1-\hat{p})}{n}\right)} = 0.0244$$

Therefore, we estimated the proportion of professors making over $135,000 CAD to be 0.390 (or 39%) with a standard error of 0.0244.

Using these estimates to construct a 95% confidence interval for p:

$$\hat{p} \pm 1.96 \cdot SE(\hat{p}) = (0.342,\ 0.438)$$

Based on our estimation, the true proportion of professors making over $135,000 CAD per year lies between 0.342 and 0.438 95% of the time.

1.3.2.2 Analysis on Binary Variable by Stratified Sampling

We used the same stratified sampled population taken in section 1.3.1.2 to estimate the proportion of professors earning salaries of more than $135,000 CAD.

We observed that:

$$\widehat{p_{str}} = 0.338 \text{ and } \mathrm{SE}(\hat{p}_{\mathrm{str}}) = \sqrt{\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1 - \frac{n_h}{N_h}\right)\left(\frac{\widehat{p_{sh}}(1-\widehat{p_{sh}})}{n_h}\right)} = 0.0236$$

Therefore, we estimated the proportion of professors earning over $135,000 CAD to be 0.338 (or 33.8%) with a standard error of 0.0236.

Then, we calculated the 95% of the confidence interval for p:

$$\hat{p} \pm 1.96 \cdot SE(\hat{p}) = (0.292,\ 0.384)$$

Based on our estimation, the true proportion of professors making over $135,000 CAD per year lies between 0.292 and 0.0.384 95% of the time.

1.4 Conclusions and Discussion

We used simple random sampling and stratified sampling for both parameters, continuous and binary. Stratified sampling better estimated both continuous and binary variables of interest with smaller standard errors.

The limitation of our conclusion is that if our sample size is small, the conclusions will not still hold for both the simple random and stratified methods as the salary data collected is highly variable.

The results we obtained cannot be generalized to the larger population because professors' salaries in each institution vary a lot. For example, suppose our objective was to obtain the mean salary of university professors in Canada. In that case, we cannot guarantee that the mean salary or proportion of professors earning $135,000 CAD or more at UBC is representative of all universities in Canada.

Part 2

Over the past two decades, there have been many attempts to find a test superior to the likelihood ratio tests in terms of the resulting bias and overall power of the tests. In multiparameter testing, it has been observed that the LRT has shortcomings. As a result, it has been asserted that tests of the same size α resulting in a less biased or unbiased result, are superior to the LRT. However, these tests go directly against statistical intuition and make an inappropriate assertion. This then begs whether a test's overall power or statistical intuition should be considered more when constructing tests. If we, as statisticians, begin to blatantly go against our intuition in favour of more "powerful tests, "we risk sacrificing our credibility within the scientific community. Therefore, it can be concluded that while we can acknowledge the shortcomings of the LRT, if it comes to a direct conflict between a new test with seemingly "better" results and our common sense as statisticians, we should always trust our intuition.

Reference

[1] University of British Columbia, “Statement of Financial Information.” March 31, 2021.

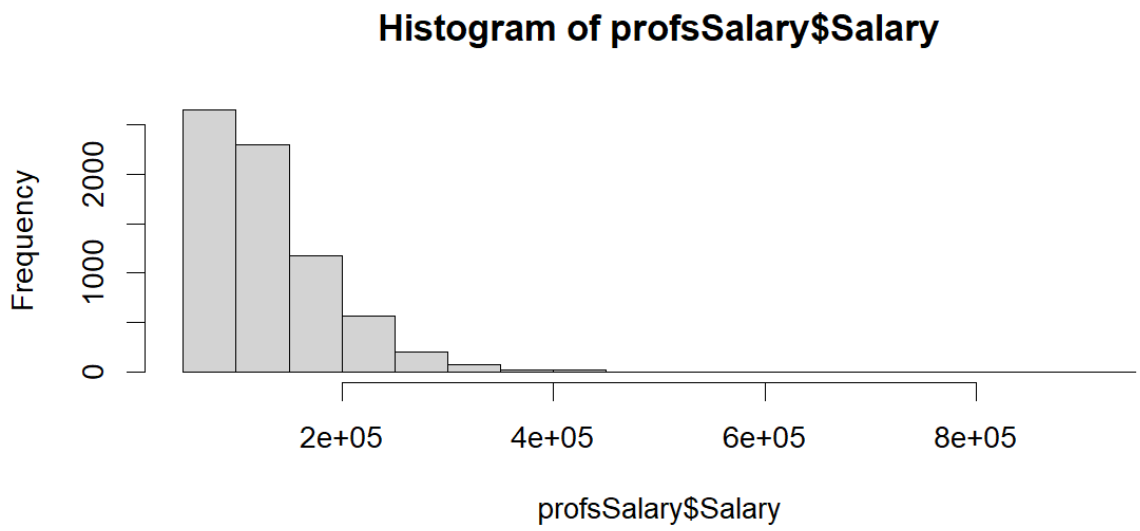
[2] UBC Graduate and Postdoctoral Studies, “Research Supervisors Directory.” Graduate School.

Appendix: Data, R code

The data includes three columns: name, salary, and the status as a research supervisor. The first ten rows are as follows:

Name	Salary	Research Supervisor
Aamodt, Tor	213,861	TRUE
Abanto Salguero, Arleni Karina	88,964	FALSE
Abbassi, Arash	104,162	FALSE
Abdalkhani, Arman	75,388	FALSE
Abdelaziz, Morad	111,360	FALSE
Abdellatif, Waleed	77,082	FALSE
Abdi, Ali	234,350	TRUE
Abdul-Mageed, Muhammad	156,351	TRUE
Abe, Masumi	85,040	FALSE

This is the histogram of professors' salaries:



Following is the R code used:

```
# You need to import that dataset from excel I named it data1
colnames(data1)[1] ="Name"
colnames(data1)[2] ="Salary"
colnames(data1)[3] = "true" # true is the column for the status as research supervisor
data1 <- data1[!is.na(data1$Salary),]
data1 <- data1[!is.na(data1$true),]
```

```
#These are the proportions of the data split into the true and false categories
N.h <- tapply(data1$Salary, data1$true, length)
#This is the population size
N <- length(data1$Salary)
#This is the sample size
n <- 400
#SRS for continuous population variable
SRS.indices <- sample.int(N,n,replace = F)
SRS.sample <- data1[SRS.indices,]
ybar.continuous.srs <- mean(SRS.sample$Salary)
se.continuous.srs <- sqrt((1 - n/N) * var(SRS.sample$Salary) / n)
srs.continuous <- c(ybar.continuous.srs, se.continuous.srs)
set <- c(FALSE,TRUE)

#Stratifying where the samples sizes are optimal for continuous population variable
sd1 <- sd(data1[data1$true == FALSE,]$Salary)
sd2 <- sd(data1[data1$true == TRUE,]$Salary)
w1 <- (sd1 * (N.h[1]/N)) / ((sd1 * (N.h[1]/N)) + (sd2 * (N.h[2]/N)))
w2 <- (sd2 * (N.h[2]/N)) / ((sd1 * (N.h[1]/N)) + (sd2 * (N.h[2]/N)))
weights <- c(w1,w2)
n.h.optl <- round(weights * n)
STR.sample.optl <- NULL
for (i in 1: 2)
{
 row.indices <- which(data1$true == set[i])
 sample.indices <- sample(row.indices, n.h.optl[i], replace = F)
 STR.sample.optl <- rbind(STR.sample.optl, data1[sample.indices, ])
}
ybar.continuous.h.optl <- tapply(STR.sample.optl$Salary, STR.sample.optl$true, mean)
var.continuous.h.optl <- tapply(STR.sample.optl$Salary, STR.sample.optl$true, var)
str.continuous.ybar.optl <- sum((N.h/N) * ybar.continuous.h.optl)
str.continuous.se.optl <- sqrt(sum((N.h/N)^2 * (((1 - n.h.optl / N.h) * var.continuous.h.optl)/n.h.optl)))
str.continuous.optl <- c(str.continuous.ybar.optl, str.continuous.se.optl)
```

```
#These outputs are the final result

# 95% confidence interval for SRS continuous population variable

srs.continuous.ci <- c(ybar.continuous.srs - 1.96 * se.continuous.srs, ybar.continuous.srs + 1.96 *
se.continuous.srs)

# 95% confidence interval for stratified sampling with optimal weights for continuous population variable

str.continuous.optl.ci <- c(str.continuous.ybar.optl - 1.96 * str.continuous.se.optl, str.continuous.ybar.optl +
1.96 * str.continuous.se.optl)

rbind(srs.continuous,srs.continuous.ci)

rbind(str.continuous.optl,str.continuous.optl.ci)


#Using a binary population variable

srs.binary.ybar <- sum(SRS.sample$Salary > 135000) / n

# SRS sampling showed that 0.395 were above 135000

srs.binary.se <- sqrt((srs.binary.ybar)*(1 - srs.binary.ybar)/n)

# SE was 0.02444 for the SRS

srs.binary.ci <- c(srs.binary.ybar - 1.96*srs.binary.se, srs.binary.ybar + 1.96*srs.binary.se)


#Using stratified sampling with optimal allocation for binary population variable

researchAssistant <- STR.sample.optl[STR.sample.optl$true == TRUE,]

nonResearchAssistant <- STR.sample.optl[STR.sample.optl$true == FALSE,]

rA.greaterThan <- sum(researchAssistant$Salary > 135000)

nRA.greaterThan <- sum(nonResearchAssistant$Salary > 135000)

beforeGreaterThan <- c(nRA.greaterThan, rA.greaterThan)

str.binary.ybar <- sum((N.h/N) * (beforeGreaterThan / n.h.optl))

str.binary.se.optl <- sqrt(sum((N.h/N)^2 * (1 - n.h.optl / N.h) * ((str.binary.ybar) * (1 -
str.binary.ybar)/n.h.optl)))

str.binary.optl.ci <- c(str.binary.ybar - 1.96 * str.binary.se.optl, str.binary.ybar + 1.96 * str.binary.se.optl)

rbind(srs.binary.ybar,srs.binary.se)

srs.binary.ci

rbind(str.binary.ybar,str.binary.se.optl)

str.binary.optl.ci
```