Part I

STAT344 Course Project

Estimating the Average Medical Insurance Charges and the Proportion of Smokers Who Have Children

Lin, Li-Kun 79656682 (Introduction & Background, Conclusion),

Hao Jiang 58301110 (Simple Random Sampling Coding),

Yucheng Zhang 84960715 (Simple Random Sampling Explanation),

Sandro Xu 29145364 (Stratified Sampling Coding),

Henry Zhang 47708805 (Stratified Sampling Explanation)

Introduction

Medical costs and lifestyle choices are core aspects when studying medical insurance charges. According to the Arabian Journal of Business, "Accurate cost estimates can help health insurers and, increasingly, healthcare delivery organizations to plan for the future and prioritize the allocation of limited care management resources (Milovic et al., 2012)." The accurate prediction of the medical insurance charges ensures the insurance policyholder gains some insight into the expected costs for the potential people in need. Moreover, it also helps the insurance policy sector to allocate limited funds to stay in financial liquidity for unpredictable adverse shocks in public health, such as COVID-19.

The study of insurance charges not only benefits the insurance policyholder to allocate the finances wisely, but it also benefits the customer to make wiser decisions when choosing the insurers. For instance, an accurate estimate of the medical insurance charge furnishes a general idea of the potential cost for customers, so customers are able to take advantage to tailor their insurance coverage based on their own needs and budget constraints to attain individual utility maximization. Additionally, the accurate estimation of medical costs also safeguards the customers from potential scams, such as overpriced charges and misleading information.

In addition to the medical insurance charges, we are also aware of the long-run public health planning. Therefore, we decided to investigate the population proportion of parenthood smokers who own insurance. According to European Addiction Research, "Children's exposure to secondhand smoking has been associated with sudden infant death syndrome, reduced lung function, and lower respiratory illnesses (Schaycka et al., 2020)." In other words, parenthood smoking is negatively correlated to the children's health. To allocate the limited medical insurance funds and improve healthcare accountability; intuitively, we might assume that if the smoking population proportion increases, it would lead to a corresponding increase in the expectation of health insurance costs in the future.

In this study, we will be using a simple random sample and stratified sample to estimate the average insurance costs per individual and the smoking proportion of the parenthood. We assume the observation in the dataset to be randomly selected and independent. The goal of these estimates is to find which of the sampling provides a better estimation of both parameters.

Introduction

Dataset background

The "Medical Cost Personal Datasets" was published by Packt Publishing in 2013. It includes 1338 participants enrolling in an insurance plan. According to the author of the dataset, "these data were created using demographic statistics from the U.S. Census Bureau, and thus approximately reflect real-world conditions (Lantz, 2013)." Therefore, we can assume that the chosen dataset is a valid statistical inference for simulating the real-life situation, which facilitates understanding both average insurance costs and the smoking population proportion of those who have children.

We intend to enhance our analysis by utilizing age (beneficiary's age) as an independent variable and charges (reflecting individual medical insurance charges in U.S. Dollars) as a response variable to estimate the population average charges in dollars in the United States. Then, we will also employ $Child_j$ and $Smoker_i$ (see image 1) to estimate the population proportion of smoking parenthood.

$$Smoker_i = \begin{cases} 0 \ \ if\ observation\ doesn't\ smoke \\ 1\ if\ observation\ smokes \end{cases}$$

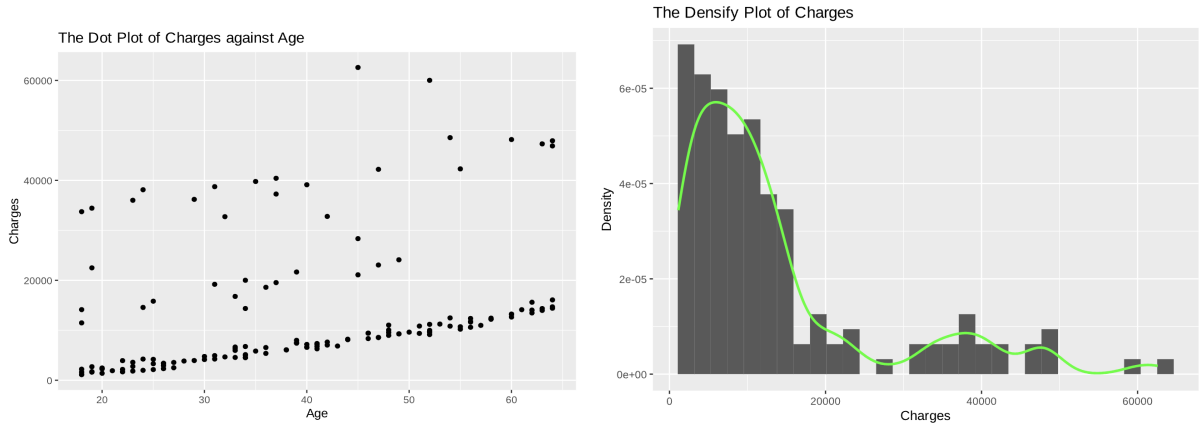
$$Child_j = \begin{cases} 0 \ \ if\ observation\ doesn't\ have\ child \\ 1\ if\ observation\ \ has\ at\ least\ a\ child \end{cases}$$

Image 1: Binary variable to determine whether the observation is a smoker or not, and determine whether the observation has at least a child or not.

Continuous variable (The average medical insurance charges)

In the simple random sample, our initial step involved selecting 150 observations from the population to constitute our sample. Therefore, the ratio of n/N stands at 11.2%, meeting the requirements for using the Finite Population Correction (FPC). By examining correlation values, we identified 'age' as the variable most strongly correlated with 'charges,' with a correlation coefficient of 0.299. Subsequently, a dot plot was employed to visually confirm the positive correlation and the linear relationship between the two variables.

Introduction



Additionally, a histogram plot was generated to explore the distribution of 'charges,' revealing that the majority falls within the range of 0 to 20,000, with few instances exceeding 40,000. Consequently, we assert that the number of outliers is within manageable limits. We need to assume that each observation is independently collected.

Estimates

- **Vanilla**: Obtaining the vanilla estimation involved calculating the mean value directly from the charges within the sample. Subsequently, after applying Finite Population Correction (FPC) to the variance of the sample charges, a z-score of 1.96 was used for constructing a 95% two-sided confidence interval with the Central Limit Theorem valid in the sample size of 150. This process yielded a vanilla estimated mean value of 13,434.85 dollars, with a 95% confidence interval ranging from 11,461.02 dollars to 15,408.69 dollars.

$$\bar{y}_{\text{srs}} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad s_{y_{\text{srs}}} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{\sum_{i=1}^{n}(y_i - \bar{y}_{\text{srs}})^2}{n-1}}$$

- **Ratio**: Even though the correlation between them is only 0.3, we still believe that elder people have a higher medical insurance charge. As Jessie Li (2011) reported in The Effects of Age and Income on Individual Health Insurance Premiums “Elderly people are more prone to illnesses so the insurance company charges them higher premiums in order to compensate for the risk that it is bearing”, we decided to use the 'age' as an auxiliary variable for the ratio estimation. By calculating the respective sample mean age (Xsbar) and the overall population mean age (Xpbar), we utilized these values in a formula to derive the estimated charges. Similarly, applying Finite

Population Correction (FPC) to the variance of the sample charges and using a z-score of 1.96 for a two-sided 95% confidence interval, we obtained an estimated ratio mean of charges as 13,301.53 dollars, with a 95% confidence interval ranging from 11,423.69 dollars to 15,179.37 dollars.

$$\hat{R}_{\text{srs}} = \frac{\bar{y}_{\text{srs}}}{\bar{x}_{\text{srs}}} \times \bar{x}_{\text{pop}} \quad \text{SE}^2_{\text{srs}} = \frac{\sum_{i=1}^{n} \left(y_i - \frac{\bar{y}_{\text{srs}}}{\bar{x}_{\text{srs}}} \times x_i\right)^2}{n-1} \quad s_{\hat{R}_{y_{\text{srs}}}} = \sqrt{\left(1 - \frac{n}{N}\right) \frac{\text{SE}^2_{\text{srs}}}{n}}$$

- **Regression**: Building on the evident linear relationship observed in the previous dot plot between age and charges, we also try to use linear regression. The obtained regression parameters were an intercept hat of 1872 and a beta age hat of 291.99. Consequently, the estimated charges for the population, calculated using the mean age of the population, amounted to 13,320.06 dollars. The associated 95% two-sided confidence interval ranged from 11,444.87 dollars to 15,195.25 dollars.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{SE}^2_{r_{\text{srs}}} = \frac{\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2}{n-1}$$

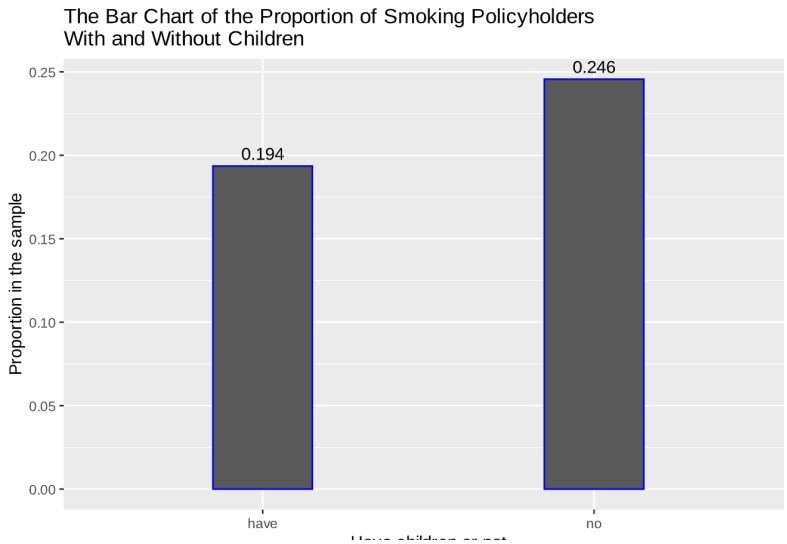
Summary

method	y_bar	standard_error	CI_low	CI_high
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
vanilla estimation	13434.85	1007.0579	11461.02	15408.69
ratio estimation	13301.53	958.0825	11423.69	15179.37
regression estimation	13320.06	956.7296	11444.87	15195.25

From the summarized table, it's noticeable that the standard error obtained through ratio estimation with the weak correlation between age and charges is significantly smaller than that of vanilla estimation. Moreover, the values obtained through ratio estimation closely align with those derived from regression. Consequently, at an equivalent confidence level, the confidence intervals generated by ratio estimation and regression are substantially narrower than those produced by vanilla estimation.

Binary Variable (The proportion of smokers who have children)

The different sample proportions of smoking policyholders with children and without children would be different. Thus, we assume that we randomly and independently get another sample with 150 observations that all of them have children to achieve our study aims.



The estimated proportion of smokers who are parents is 0.208, with a standard error of 0.031. Consequently, the 95% two-sided confidence interval spans from 0.147 to 0.269. The confidence interval suggests that the proportion of the smoking population who are parents is likely to be between 0.13 and 0.2. This means that we are 95% confident (or 19 times out of 20) that the true proportion of smokers who are parents in the entire population falls within this range.

	estimate	p_hat	standard_error	CI_low	CI_high
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
proportion of smokers who have children		0.208	0.031	0.147	0.269

Stratified Sampling

Continuous variable (The average medical insurance charges)

In addition to simple random sampling, we also decided to perform stratified sampling on our data. The parameter of interest is the population mean medical costs billed by health insurance of each individual in the U.S. We chose to use regions to split our strata because we believe that there should be some level of variations between the medical costs for people living in different regions. There are four regions in our data set, so we chose to use proportional allocation to decide the sample sizes for each strata.

Proportional allocation

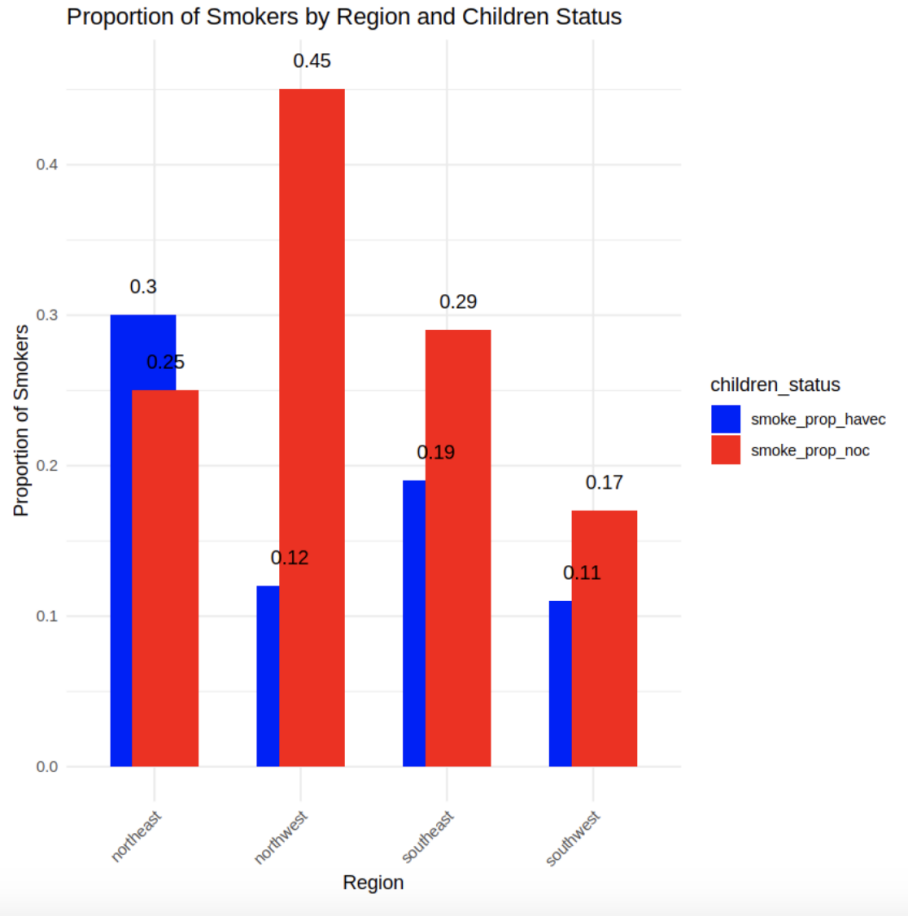
We decided to obtain a total sample size of 150, which is consistent with the sample size of the simple random sample in the previous section. The proportional allocation assigns sub-sample sizes to each strata based on the proportion of population sizes of each strata. By examining our data, we decided to sample 36 observations from the regions Northeast, Northwest, and Southwest, and sample 41 observations from the region Southeast.

Estimate of population mean

After deciding the sample sizes for each strata, we used R functions to obtain the stratified samples for each region and calculated the mean and standard error (with Finite Population Correction) for each strata. To obtain a stratified sample estimate of the population mean, we multiplied the sample mean of each strata by their relative population size and summed these results up. In the end, we obtained an estimate of 12945.80, which means that the population mean medical costs of individuals in the U.S. is estimated to be 12945.80 dollars. We also get the standard error of this estimate by combining the standard errors of each stratum following the formula: $SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N})^2 SE^2(\bar{y}_{S_h})}$. This gave us a result of 872.32 dollars. Since we had an estimate of the population mean and its standard error, we were also able to construct a 95% confidence interval of the population mean. The lower and upper bounds of this confidence interval are given by our estimate of less or plus 1.96 times the standard error. The resulting boundaries of the confidence interval are from 11236.046 to 14655.548 dollars. An interpretation of the 95% confidence interval under the context of our parameter of interest would be that over repeated samplings, approximately 95% of our confidence intervals will capture the true population mean of medical costs of individuals in the U.S.

Binary Variable (The proportion of smokers who have children)

Then we wanted to apply the same idea of stratified sampling to our second parameter of interest, which is the population proportion of individuals in the U.S. who have children and are smokers. We kept those individuals who have children in our data set and obtained stratified samples from the four strata with the same sub-sample sizes used in the previous section. Then we found out the sample proportions of smokers in each stratum. The following bar chart illustrates our results, along with sample proportions of smokers who do not have children:



Incorporating with Finite Population Correction (FPC), the standard error of sample proportion for each strata can be found using the formula $SE(\hat{p}) = \sqrt{\sum_{i=1}^{h}(1 - \frac{n_h}{N_h}) * \frac{\hat{p}*(1-\hat{p})}{n_h}}$. The following table

summarizes our final result of the sample proportion and its standard error in each region:

region <chr>	smoke_prop_havec <dbl>	se_smoke_prop_havec <dbl[1d]>
northeast	0.1111111	0.04938272
northwest	0.2222222	0.06533975
southeast	0.1951220	0.05830110
southwest	0.2222222	0.06533975

Estimate of population proportion

Knowing these statistics about our stratified sample, we combined these results to obtain a stratified estimate of the population proportion of people who are parenthood smokers. We summed up the products of the sample proportions of each strata and their relative population sizes, which gave us a result of around 0.188. This means that the population proportion of smokers in the U.S. who have children is estimated to be 0.188 by our stratified estimator. The standard error of this stratified estimate is given by summing up the product of the square of each stratum's relative population size and the square of standard errors of each stratum, then taking the square root of the result. This gave us a resulting standard error of our estimated population proportion of around 0.030. We again obtained a 95% confidence interval for the population proportion of smokers who have children and found the resulting boundaries to be from 0.129 to 0.247. Therefore, over repeated samplings, approximately 95% of our confidence intervals will capture the true population proportion of smokers who have children in the U.S.

Conclusion

In the analysis of the continuous variable in a simple random sample, we utilized vanilla, ratio, and regression estimations to find the average insurance charge. The regression estimation tended to have the smallest standard error with 956.7 dollars and a mean of 13320 dollars, which has the narrowest confidence interval out of all three estimators. However, when we utilized the stratified sample strata by region, we observed that the stratified sample's mean and standard error are 12945.80 dollars and 872.3 dollars, reflecting an approximate 9% reduction in standard error compared to the SRS.

On the other hand, in the analysis of the binary variable in a simple random sample on population proportion in parenthood smokers, we observed that population proportion and standard deviation are 20.8% and 0.031. Whereas, in the stratified random sample, we have 18.8% and 0.030 respectively.

In conclusion, our analysis reveals that stratified sampling provides higher consistency in the measurement of both continuous and binary variables than SRS. The reduction in standard error denotes the accuracy and precision of the population parameters. Thus, we choose the results from the stratified sampling as our estimation of population parameters in the US: the average insurance costs per policyholder are 12945.80 dollars, and the population proportion of parenthood smokers is 18.8%.

Discussion (Limitation)

According to Lantz (the author of the dataset), samples were collected between 1999~2012. Therefore, it is notable that the standard error of the estimation might increase due to exogenous factors such as nominal inflation, and real price (CPI). According to Macrotrends, the average annual inflation rate in the US is approximately 2.5%, which added up to 32.57% from 1999 to 2012 (Macrotrends, 2023). Thus, if we are calculating with the real price index, our confidence interval might be narrower than the presented values above. Unfortunately, we are unable to access the specific year in which the individual is being selected. Thus, the insurance charges are calculated in the nominal price index.

Part II

The paper “The Emperor’s New Tests” talks about the issues regarding hypothesis testing in the area of statistics. It reveals that there exist critics nowadays who argue that traditional likelihood ratio tests (LRT) can be inferior, biased, or flawed, and people come up with other tests such as “superior” tests. However, that is not a correct interpretation. The key message is that LRTs are often undervalued despite being less biased, more powerful, and generally more suitable for a wide range of statistical problems. He uses a very famous allegory called “The Emperor’s New Clothes “ to implicitly indicate that although new methods look better they are not necessarily better, if these methods are used without proper scrutiny it will be equivalent to the emperor being naked but thinking that he wore the most fancy and good looking dress. The authors advocate for a return to traditional methods and a reconsideration of what constitutes statistical practice and think that a fundamental reassessment of the mission of mathematical statistics is urgently needed. At the same time, we view with skepticism new method holdings that are guaranteed to produce simple answers without explicitly testing or considering the underlying assumptions.

Citation

Hassan, A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh M.A.A., & Ullah, S.S. (2021). "A Computational Intelligence Approach for Predicting Medical Insurance Cost." Hindawi. Volume 2021. Doi: https://doi.org/10.1155/2021/1162553

Lantz, Brett. (2013). Machine Learning with R. Page[173]. ISBN: 1782162143

Li, J. (2011). Undergraduate economic review - Effects of Age and Income on Individual Health Insurance Premiums. https://digitalcommons.iwu.edu/cgi/viewcontent.cgi?article=1114&context=uer

Macrotrends. (2023). "U.S. Inflation Rate 1960-2023." https://www.macrotrends.net/countries/USA/united-states/inflation-rate-cpi

Milovic, B., & Milovic, M. (2012). "PREDICTION AND DECISION MAKING IN HEALTH CARE USING DATA MINING. Kuwait Chapter of the Arabian Journal of Business and Management Review", 1(12), 126-136. Retrieved from https://www.proquest.com/scholarly-journals/prediction-decision-making-health-care-using-data/docview/1221952501/se-2

Schaycka, S.T., Mujcicc A., Ottene R., Engels R., Kleinjan, M. (2020). "The Effectiveness of Smoking Cessation Interventions Tailored to Smoking Parents of Children Aged 0–18 Years: A Meta-Analysis." European Addition Research. 27:278–293. DOI: 10.1159/000511145

Wyszynski et al. (2011). "Parental Smoking Cessation and Child Daily Smoking: A NineYear Longitudinal Study of Mediation by Child Cognitions About Smoking." Health Psychol.; 30(2): 171–176. doi:10.1037/a0022024.

Appendix

```
library(tidyverse)
library(dplyr)
library(ggplot2)
insurance_df <- read.csv("insurance.csv")
head(insurance_df)
N <- nrow(insurance_df)
N # populatin size
x_p <- insurance_df['bmi']
cor(insurance_df$charges, insurance_df$age)
ybar_pop <- mean(insurance_df$charges)
xbar_pop <- mean(insurance_df$age)
n <- 150
n/N
set.seed(1)
SRS.index <- sample.int(N, n, replace = FALSE)
insurance_sam_srs <- insurance_df[SRS.index, ]
head(insurance_sam_srs)
options(repr.plot.width = 7, repr.plot.height = 5)
y_sam <- insurance_sam_srs$charges
x_sam <- insurance_sam_srs$age
sam_df <- data.frame(x_sam, y_sam)
y_s_plot <- ggplot(sam_df, aes(x_sam,y_sam))+
```

```
        geom_point()+

        labs(x="Age", y="Charges", title="The Dot Plot of Charges
        against Age")

y_s_plot

y_dist <- ggplot(sam_df, aes(y_sam,y = after_stat(density)))+

        geom_histogram() +

        geom_density(color = "green", linewidth = 1)+

        labs(x="Charges", y="Density", title="The Densify Plot of
        Charges")

y_dist

# vanilla

ybar_sam_srs <- mean(y_sam)

xbar_sam_srs <- mean(x_sam)

ybar_sam_srs

sd_val_y_srs <- sqrt((1-n/N) * var(insurance_sam_srs$charges)/n)

sd_val_y_srs

CI_val_y_srs <- c(ybar_sam_srs - 1.96*sd_val_y_srs, ybar_sam_srs
        + 1.96*sd_val_y_srs)

CI_val_y_srs

# ratio

ratio_est_srs <- ybar_sam_srs/xbar_sam_srs * xbar_pop

ratio_est_srs

Se2_srs <- sum((y_sam - ybar_sam_srs/xbar_sam_srs *
        x_sam)^2)/(n-1)
```

Appendix

```
sd_rat_y_srs <- sqrt((1-n/N) * Se2_srs/n)

sd_rat_y_srs

CI_rat_y_srs <- c(ratio_est_srs - 1.96*sd_rat_y_srs, ratio_est_srs +
        1.96*sd_rat_y_srs)

CI_rat_y_srs

# regression

summary(lm(charges~age, data=insurance_sam_srs))

regr_est_srs <- 1872+291.99 * xbar_pop

regr_est_srs

Se2_r_srs <- sum((y_sam - (1872+291.99 * x_sam))^2)/(n-1)

sd_rgr_y_srs <- sqrt((1-n/N) * Se2_r_srs/n)

sd_rgr_y_srs

CI_rgr_y_srs <- c(regr_est_srs - 1.96*sd_rgr_y_srs, regr_est_srs +
        1.96*sd_rgr_y_srs)

CI_rgr_y_srs

data.frame(method = c("vanilla estimation",

                      "ratio estimation",

                      "regression estimation"),

           y_bar = c(ybar_sam_srs,

                     ratio_est_srs,

                     regr_est_srs),

          standard_error = c(sd_val_y_srs,

                             sd_rat_y_srs,
```

```
                sd_rgr_y_srs),

        CI_low = c(CI_val_y_srs[1],

             CI_rat_y_srs[1],

             CI_rgr_y_srs[1]),

        CI_high = c(CI_val_y_srs[2],

             CI_rat_y_srs[2],

             CI_rgr_y_srs[2]))

insurance_sam_srs <- insurance_sam_srs %>%

          mutate(

            children = case_when(

            grepl(0, children) ~ "no",

            grepl("1", children) | grepl("2", children) | grepl("3",
          children) | grepl("4", children) | grepl("5", children) ~
          "have",

            TRUE ~ as.character(children)))


insurance_sam_srs_havec <- insurance_sam_srs %>%

                    filter(children == "have")

smoker_prop_havec <- mean(insurance_sam_srs_havec$smoker ==
          "yes")


insurance_sam_srs_noc <- insurance_sam_srs %>%

                    filter(children == "no")
```

```
smoker_prop_noc <- mean(insurance_sam_srs_noc$smoker ==
        "yes")

smoker_prop_havec

smoker_prop_noc


noc_havec_df <- data.frame(children=c("have", "no"),

            prop=c(smoker_prop_havec, smoker_prop_noc))


noc_havec_plot <- ggplot(noc_havec_df, aes(children,prop))+

          geom_col(width = 0.3, color="blue")+

          geom_text(aes(label = round(prop, 3)),

                    vjust = -0.5,

                    color = "black")+

          labs(x="Have children or not",

               y="Proportion in the sample",

              title="The Bar Chart of the Proportion of Smoking
         Policyholders \nWith and Without Children")

noc_havec_plot


set.seed(1)

insurance_df <- insurance_df %>%

          mutate(
```

```
        children = case_when(

        grepl(0, children) ~ "no",

        grepl("1", children) | grepl("2", children) | grepl("3",
      children) | grepl("4", children) | grepl("5", children) ~
      "have",

        TRUE ~ as.character(children)))



SRS.index_havec <- sample.int(N, n, replace = FALSE)

insurance_df_havec <- insurance_df %>%

              filter(children == "have")

insurance_sam_srs_havec <- insurance_df_havec[SRS.index, ]

head(insurance_df_havec)

# vanilla

smoker_prop_havec <- mean(insurance_df_havec$smoker ==
      "yes")

smoker_prop_havec

sd_val_prop_srs <- sqrt((1-n/N) *
      (smoker_prop_havec*(1-smoker_prop_havec))/n)

sd_val_prop_srs

CI_val_prop_srs <- c(smoker_prop_havec - 1.96*sd_val_prop_srs,
      smoker_prop_havec + 1.96*sd_val_prop_srs)

CI_val_prop_srs

data.frame(estimate = "proportion of smokers who have children",

      p_hat = round(smoker_prop_havec,3),

      standard_error = round(sd_val_prop_srs,3),
```

```
        CI_low = round(CI_val_prop_srs[1],3),
        CI_high = round(CI_val_prop_srs[2],3))
```

```
#STRATIFIED SECTION CODE
library(tidyverse)
library(dplyr)

insurance <- read.csv("insurance.csv")

insurance <- na.omit(insurance) #removing NA observations
head(insurance) #check the dataset loaded
N <- nrow(insurance) #population size
N

N.h <- tapply(insurance$charges, insurance$region, length)
        #population size for different regions
regions <- names(N.h) # name of the regions
regions
N.h
```

Appendix

```
n <- 150 #sample size total

n.h.prop <- round( (N.h/N) * n) # WE ARE USING
        PROPORTIONAL ALLOCATION TO GET THE SAMPLE
        SIZE FOR EACH OF THEM

n.h.prop

set.seed(0)  # Set a seed for reproducibility

stratified_sample <- NULL  # Initialize an empty data frame for the
        stratified sample



# Loop over each region to create a stratified sample

for (i in 1:length(regions)) {

  # Get row indices for the current region starting from northeast

  row_indices <- which(insurance$region == regions[i])



  # Sample the indices without replacement

  sample_indices <- sample(row_indices, n.h.prop[i], replace =
        FALSE)



  # Extract the rows for the sampled indices and select only the
        charges and age columns

  stratified_sample <- rbind(stratified_sample,
        insurance[sample_indices, ])

}
```

```
ybar.h.prop <- tapply(stratified_sample$charges,
        stratified_sample$region, mean) #mean of charges for each
        strata sampled

var.h.prop <- tapply(stratified_sample$charges,
        stratified_sample$region, var) #variance of charges for each
        strata sampled

se.h.prop <- sqrt((1 - n.h.prop / N.h) * var.h.prop / n.h.prop)
        #standard error for each strata sampled



ybar.str.prop <- sum(N.h / N * ybar.h.prop) #Estimated population
        mean using Stratified Sampling

se.str.prop <- sqrt(sum((N.h / N)^2 * se.h.prop^2)) #Standard error
        of the estimated population mean

str.prop <- c(ybar.str.prop, se.str.prop) #Combined both the
        Estimated population mean and the Standard error




ybar.h.prop

var.h.prop

se.h.prop

rbind(ybar.h.prop, se.h.prop) #binding the mean and standard error
        for each strata

ybar.str.prop

se.str.prop

Str.prop
```

Appendix

```
# Obtain a 95% C.I. for our estimate of population mean

lower = ybar.str.prop - 1.96*se.str.prop

upper = ybar.str.prop + 1.96*se.str.prop

c(lower, upper)
```

```
# Turning the children column into having or not having children
        with binary output have and no

insurance_stratified <- stratified_sample %>%

          mutate(

             children = case_when(

             grepl(0, children) ~ "no",

             grepl("1", children) | grepl("2", children) | grepl("3",
          children) | grepl("4", children) | grepl("5", children) ~
          "have",

             TRUE ~ as.character(children)))
```

```
# Calculate proportion of smokers among those who have children,
        for each region

insurance_stratified_havec <- insurance_stratified %>%

                        group_by(region) %>%

                        filter(children == "have") %>%

                        summarise(smoke_prop_havec = mean(smoker ==
        "yes"))
```

```
# Calculate proportion of smokers among those who have no
      children, for each region

insurance_stratified_noc <- insurance_stratified %>%

                  group_by(region) %>%

                  filter(children == "no") %>%

                  summarise(smoke_prop_noc = mean(smoker ==
      "yes"))


insurance_stratified_havec

insurance_stratified_noc


library(ggplot2)

library(dplyr)


# Sample data

insurance_stratified_havec <- data.frame(

  region = c('northeast', 'northwest', 'southeast', 'southwest'),

  smoke_prop_havec = c(0.30, 0.12, 0.19, 0.11)  # sample proportions
      of smokers who have children

)
```

Appendix

```
insurance_stratified_noc <- data.frame(

  region = c('northeast', 'northwest', 'southeast', 'southwest'),

  smoke_prop_noc = c(0.25, 0.45, 0.29, 0.17)  # sample proportions of
        smokers who do not have children

)



# Merge the two data frames by 'region'

merged_df <- merge(insurance_stratified_havec,
        insurance_stratified_noc, by = "region")



# Convert to long format for ggplot

long_df <- gather(merged_df, key = "children_status", value =
        "proportion", -region)



# Plot

ggplot(long_df, aes(x = region, y = proportion, fill =
        children_status)) +

  geom_bar(stat = "identity", position = position_dodge(width =
        0.3)) +

  geom_text(aes(label = round(proportion, 3),

              y = proportion + 0.01,

              group = children_status),

          position = position_dodge(width = 0.3),

          vjust = -0.5) +

  labs(x = "Region",
```

```
    y = "Proportion of Smokers",

    title = "Proportion of Smokers by Region and Children Status")
      +

 theme_minimal() +

 scale_fill_manual(values = c("blue", "red")) +

 theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x
      labels


# We are interested in proportion of smokers that have children so
      we do the analysis


#reproducibility

set.seed(0)




# Turning the children column into having or not having children
      with binary output have and no

# Then this outputs the data frame after filtering for only the ones
      having children

insurance_havec <- insurance %>%

          mutate(

          children = case_when(

          grepl(0, children) ~ "no",
```

```
        grepl("1", children) | grepl("2", children) | grepl("3",
      children) | grepl("4", children) | grepl("5", children) ~
      "have",

        TRUE ~ as.character(children))) %>%

        filter(children == "have")


dim(insurance_havec)


#We sample 150 again using the same proportional allocation

stratified_sample_bin <- NULL  # Initialize an empty data frame
      for the stratified sample for binary data


# Loop over each region to create a stratified sample

for (i in 1:length(regions)) {

  # Get row indices for the current region starting from northwest

  row_indices <- which(insurance_havec$region == regions[i])


  # Sample the indices without replacement

  sample_indices <- sample(row_indices, n.h.prop[i], replace =
      FALSE)


  # Extract the rows for the sampled indices and select only the
      charges and age columns

  stratified_sample_bin <- rbind(stratified_sample_bin,
      insurance_havec[sample_indices, ])
```

```
}


#Sample with only children column being true

head(stratified_sample_bin)




#Grouping by region and finding the proportion of smokers that
      have children

stratified_sample_bin_havec <- stratified_sample_bin %>%

                  group_by(region) %>%

                  filter(children == "have") %>%

                  summarise(smoke_prop_havec = mean(smoker ==
      "yes"))

stratified_sample_bin_havec


# Calculate the standard error for proportion of smokers that have
      children

stratified_sample_bin_havec <- stratified_sample_bin_havec %>%

  mutate(se_smoke_prop_havec = sqrt((1 - n.h.prop / N.h) *
      (smoke_prop_havec * (1 - smoke_prop_havec) / n.h.prop)))




# The dataframe now has the standard error for each region and its
      proportion
```

```
stratified_sample_bin_havec


# Estimated population proportion using Stratified Sampling

bin.str.prop <- sum(N.h / N *
        stratified_sample_bin_havec$smoke_prop_havec) #Estimated
        population proportion using Stratified Sampling

bin.str.prop




# Standard error of the estimated population proportion

bin.se.str.prop <- sqrt(sum((N.h / N)^2 *
        stratified_sample_bin_havec$se_smoke_prop_havec^2))

bin.se.str.prop




#Combined both the Estimated population proportion and its
        Standard error

pse.str.prop <- c(bin.str.prop, bin.se.str.prop)

pse.str.prop

# Obtain a 95% C.I. for our estimate of population proportion

lower.prop = bin.str.prop - 1.96*bin.se.str.prop

upper.prop = bin.str.prop + 1.96*bin.se.str.prop

c(lower.prop, upper.prop)
```