STAT 344 Group Project: Blood Glucose Level

Group C1

Leader: Sicily Xie 54385315

- lead the discussion with open-ended questions and process data

Editor: Tingyu Chen 85784080

- integrate different pieces of writing from group members to create consistency

Brainstomer: Jia Jun Cheng Xian 55325294

- critique the project based on requirements to ensure success

Organizer: Junan Yao 95298444

- keep the project on track

Researcher: Yutong Li 89878672

- provide the group members with sources and information

Introduction

Blood glucose level can be a major indicator of many serious health issues, including heart disease, vision loss, and kidney disease. Without consideration of keeping a healthy glucose level, individuals are at higher risk of suffering from many conditions. In this study, our objective is to analyze the blood glucose levels of the public. To further investigate the population, this study analyzes patient data collected by medical clinics in Bangladesh.

The Glucose level of the body is considered healthy if lower or equal to 100 mg/dl (Danaei G., 2011). This study is interested in two parameters of interest in this population in order to better understand the level of healthy patients in this population: the population means of the blood glucose level of the patients and the proportion of the patients with mean blood glucose levels lower or equal to 100mg/dl, from medical clinics in Bangladesh. Using retrieved sample data, we use two sampling methods to predict it and analyze the results from each sampling method in this population.

Source of the data:

The dataset used within this study is retrieved from Kaggle. The original source of the data is collected from many hospitals/clinics in Bangladesh, containing 5110 patient anonymized information. The variables of interest within this dataset are as described below.

Description of variable:

avg_glucose_1eve1: This attribute describes the average level of the patient's blood glucose level. It's a numerical variable with unit mg/dL.

bmi: This attribute describes the body mass index (BMI) of the patients. It's a numerical variable with a unit kg/m^2.

smoking_status: This categorical attribute represents the smoking condition of an individual.

Research Methods

1.Method I (SRS)

To estimate the *average blood glucose level* of all patients in our dataset using simple random sample (SRS) sampling, both a vanilla estimator and a ratio estimator is implemented in this study. For the ratio estimator, this study uses variables correlated with the variable of interest to improve the precision of the estimation of the population mean; previous studies revealed that BMI are correlated with glucose level (Nogueira-de-Almeida, 2017), BMI is chosen as our auxiliary variable in this study as we hypothesized that there might exist a positive correlation with glucose levels. We denote N as *population size*; n as our *sample size*. The sample obtained in this study used sample.int() function in R with arguments (N, n, replace = FALSE), to extract a random sample of size n without replacement.

Vanilla estimator and ratio estimator are both used as population estimators for the average blood glucose level. The following are the variables denoted in this part of study to conduct the method of estimation. Denote $\overline{y_p}$ as the population average glucose level, $\overline{y_{vanilla}}$ as a vanilla estimator, $\overline{y_s}$ as the sample mean of average blood glucose level. According to lectures, $\overline{y_{vanilla}} = \overline{y_s}$ and is an unbiased estimator for population parameter $\overline{y_p}$.

In ratio estimation, denote $\overline{y_{ratio}}$ as our ratio estimator, $\overline{x_s}$ as the sample mean value of the auxiliary variable, BMI. Ratio estimator makes further assumption that the population mean of patients' BMI is known and denoted as $\overline{x_p}$, and $\overline{y_{ratio}} = \frac{\overline{y_s}}{\overline{x_s}} \times \overline{x_p}$ would be the ratio estimate of population average blood glucose level, given population mean BMI.

For the parameter of interest on proportion, the vanilla estimator is chosen as the sample proportion of patients' blood glucose levels less than 100 mg/dL, the population

parameter is denoted as p. The patient data is categorized as a binary variable, *1* as less than the threshold and *0* as higher than threshold. Denote $\widehat{p}$ as the vanilla estimator, representing the sample proportion of individuals with blood glucose level less than 100 mg/dL. From lecture, $\widehat{p}$ is the unbiased estimator for our population proportion p.

2.Method II (STR)

For the stratified sample (STR), the same *population size* and *sample size* is the same as the SRS. Previous studies have indicated that smoking status is correlated to glucose level (Lebech C.,2020), therefore in this study the population is stratified for sampling by their smoking_status parameter. We divide population P into H strata (sub-populations), where $H = 3$ since we have three categories of smoking status, "formerly smoked", "never smoked", "smokes".

This study take the assumption of equal variance between categories, it would be optimal to use proportional allocation to allocate our sample size for each stratum, denote each sample size as $n_{h1}, n_{h2}\ and\ n_{h3}$ respectively. Assume the population size for each stratum is known, denote them as $N_{h1}, N_{h2}\ and\ N_{h3}$, sample size $n_{hi} = \frac{N_{hi}}{N} * n$ for i = 1, 2, 3. Then $\overline{y_{h1}}$, $\overline{y_{h2}}$ and $\overline{y_{h3}}$ are the population mean glucose level of the three stratum and P_{h2}, P_{h2}, P_{h3} are the population proportion of people with glucose level less than 100 mg/dL of the three stratum, respectively .The sample is obtained using sample.int() function in R with arguments N_{hi}, n_{hi}, replace = FALSE, for i = 1,2,3, corresponding to the three stratum respectively, and the total sample size is still 346.

The *mean blood glucose level* based on the patient's smoking status is the parameter of interest here. The *targeted population* is the same as discussed in the previous (Method I SRS) part. The blood glucose level is estimated using STR with proportional allocation.

Denote $\overline{y_{str}}$ as a stratify estimator, representing a sample mean of average blood glucose level and estimates $\overline{y_p}$ based on a sample of size 346, $\widehat{\overline{y_{h1}}}$ as the sample mean of the stratum patients with "formerly smoked" and estimate $\overline{y_{h1}}$, $\widehat{\overline{y_{h2}}}$ as the sample mean of the stratum patients with "never smoked" and estimate $\overline{y_{h2}}$ and $\widehat{\overline{y_{h3}}}$ as the sample mean of the stratum patients with "smokes" and estimate $\overline{y_{h3}}$, then, the formula would be

$$\overline{y_{str}} = \frac{n_{h1}}{N} * \widehat{\overline{y_{h1}}} + \frac{n_{h2}}{N} * \widehat{\overline{y_{h2}}} + \frac{n_{h3}}{N} * \widehat{\overline{y_{h3}}}$$

We choose sample proportion P_{str} to estimate the population proportion p of patients' blood glucose levels which are less than 100 . Denote P_{normal} as the population proportion that has glucose level less than 100 ml/dl, P_{str} as a stratify estimator that estimates population parameter P_{normal} based on a sample of size 346. Let $\widehat{P_{h1}}$ as the sample mean of the stratum patients with "formerly smoked" and estimate P_{h1}, $\widehat{P_{h2t}}$ as the sample mean of the stratum patients with "never smoked" and estimate P_{h2}, and $\widehat{P_{h3}}$ as the sample mean of the stratum patients with "smokes" and estimate P_{h3}. Then we let our

$P_{str} = \frac{n_{h1}}{N} * \widehat{P_{h1}} + \frac{n_{h2}}{N} * \widehat{P_{h2t}} + \frac{n_{h3}}{N} * \widehat{P_{h3}}$ as our stratify estimate of P_{normal}.

Data Analysis

1.Data Cleaning

This study's data sampling and data analysis use three variables, average glucose level, BMI, and smoking status. During the data cleaning process, the dataset was found to contain undefined BMI and smoking status for some patients. Data from those patients were removed from our dataset for clarity. After cleaning, the total data size shrinks from the original 5110 to 3426. Our population is then approximated by the entries within this dataset after cleaning.

2.Study Planning

We set the margin of error (m.o.e) for the proportion of patients with blood glucose levels below the threshold to be less than 0.05, or 19 out of 20 times. Due to the lack of prior information regarding the proportion of patients with normal glucose levels, we would choose the conservative proportion of 0.5 to obtain the largest confidence interval. Due to the FPC effect, the minimum sample size required to achieve this m.o.e. is derived to be 346. This sample size would subsequently be applied for both simple random sampling and stratified sampling methods.

Denote $\alpha = 0.95$ as our 95 percent confidence interval. Denote δ as the margin of error, δ = 0.05. Denote S^2_{guess} as our conservative variance, S^2_{guess} = 0.5^2 = 0.025. Denote $Z_{\alpha/2}$ = 1.96 as the $(1 - \alpha/2)$ quantile of standard normal. The sample size n is then calculated as below.

$$n_0 = \frac{(Z_{\alpha/2})^2 * S^2_{guess}}{\delta^2} = \frac{(1.96)^2 * 0.25}{0.05^2}$$

$$n = ceiling(n) = ceiling(\frac{n_0}{(1+n_0/N)}) = 346$$

3.Discussion and analysis of data

The vanilla estimate of the average glucose level among this dataset obtained by a simple random sample is 109.69 and has a standard error of 2.53 while the ratio estimate of the average glucose level is 106.94 and has a standard error of 2.70. Compared to the population true value of the average glucose level at 108.32, the vanilla estimate has little difference of the gap to the true value of the parameter but the vanilla estimate gives a smaller standard error as well as a tighter 95% confidence level, which makes it a better estimator than ratio estimate in this case. The reason why the ratio estimator did not perform better than the vanilla estimator may be resulted from our assumption on method I, where we only assumed the auxiliary variable BMI has positive correlation with the average glucose level. However, the ratio estimator will have a better performance than vanilla estimator only if there is a strong positive correlation between the auxiliary variable and the response variable. But the correlation of the average glucose level and BMI from population is 0.157, which is positive but not strong enough to support the performance of the ratio estimator.

Table: SRS_Continuous_Data

	estimates	standard erros	lower confidence intervals	upper confidence intervals
Vanilla	109.6894	2.532423	104.7258	114.6529
Ratio	106.9390	2.624594	101.7948	112.0832

Figure 1: SRS (continuous)

Table: SRS_Binary_Data

	estimates	standard erros	lower confidence intervals	upper confidence intervals
SRS (binary)	0.6184971	0.02611438	0.5673129	0.6696813

Figure 2: SRS (binary)

In terms of stratified estimator for the average glucose level, it has an estimate of 106.55 and a standard error of 2.32, where it has the lowest standard error among all the 3 estimators of the average glucose level. Though this estimator gives the most desirable result among these 3 estimators, there is a potential way to improve the performance of the estimator: one can consider that the strategy of the stratification might not be very suitable for this dataset, as the assumptions to use proportional allocation require the within-stratum variance of the response variable is the same across each stratum, but this is not the case in terms of ours, as we have 2 within-stratum variance that has a similar value, at about 2100, and the other within-stratum variance has the value about 3000, which will make the proportional allocation not the optimal choice. In general, if an educated guess of population variance of each of the stratum is given, one could assume the cost to stratify every group is identical and make use of the optimized allocation strategy to allocate the sample size for each subpopulation to achieve a potential better estimator.

As for the estimator for the proportion of individuals with blood glucose level less than 100 mg/dL, both the vanilla estimator from SRS and the stratified estimator covered the true value of the population parameter, at a value of 0.63. The stratified estimator does have a smaller standard error, which makes it a better estimator in terms of accuracy, though the difference of standard error between these two variables are not very significant, at an absolute difference of 0.0014. Considering the insignificant difference on the standard error and close estimates on the parameter of these 2 estimators, this result suggests that the stratified sampling method stratified on the population's smoking status does not have too much impact, the SRS estimator for the proportion parameter in this population is good enough due to its simplicity and potentially lower cost without the need to stratify and find stratums.

Table: STR_Continuous_Data

	estimates	standard erros	lower confidence intervals	upper confidence intervals
STR (continuous)	106.5451	2.319689	101.9985	111.0917

Figure 3: STR (continuous)

Table: STR_Binary_Data

	estimates	standard erros	lower confidence intervals	upper confidence intervals
STR (binary)	0.6156036	0.02478154	0.5670318	0.6641754

Figure 4: STR (binary)

Conclusion

After comparing the vanilla estimate, ratio estimate and stratified sample estimate (smoke status as stratum) of the average glucose level, we conclude that the stratified sample estimate is a more suitable estimate for the dataset collected from the clinics of Bangladesh, as it produces the smallest standard error among these three methods. As for the estimation on the proportion of the patients with mean blood glucose level lower or equal to 100mg/dl, the vanilla estimate is good enough for this dataset since the stratified sampling method provide insignificant improvement on the result, if there is no consideration on swapping the variable to stratify; Further analysis could be invited to investigate whether changing stratum helps to improve the performance of the stratified sampling on this parameter. In addition, the suggesting estimation methods on both of the parameters of interest might not be suitable to generalize to dataset beside those collected in Bangladesh due to some limitations on our design.

This study conducted a series of investigations on the results obtained using different sampling methods. Through statistical calculation and discussion, we have shown that stratified sampling outstrips other sampling methods in providing an accurate population estimate. This nonetheless should not diminish other methods' significance in sampling as the SRS method is great in estimating the population parameter while maintaining a considerably reasonable cost. Specifically, future studies are encouraged to optimize their sampling methods in reference to this study's conclusion and recommendations.

Limitation: In our data sampling and data analysis, certain limitations and concerns of both our data and our analysis. The first limitation is about how the data is collected, the information provided that the data comes from the clinic of Bangladesh, but we don't know how it is collected. Specifically, there are several ways of retrieving the glucose levels of patients, the resulting glucose level would be interpreted differently with different tests, but

there is no mention of how the glucose level is obtained in the dataset, the validity of the dataset and thus our analysis assumption of this dataset is questionable. In this study, the initial assumption was made that the data obtained represented the population of the patients in medical clinics across Bangladesh. This assumption could be questioned due to insufficient amount of data collected. Additional studies need to be conducted in the future to better represent the population.

References

The source of the data being used for the project:

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Cichosz, Simon L., Morten H. Jensen, and Ole Hejlesen. "Associations between Smoking, Glucose Metabolism and Lipid Levels: A Cross-Sectional Study." *Journal of Diabetes and its Complications*, vol. 34, no. 10, 2020, pp. 107649-107649.

Danaei, Goodarz, MD, et al. "National, Regional, and Global Trends in Fasting Plasma Glucose and Diabetes Prevalence since 1980: Systematic Analysis of Health Examination Surveys and Epidemiological Studies with 370 Country-Years and 2·7 Million Participants." *The Lancet (British Edition)*, vol. 378, no. 9785, 2011, pp. 31-40.

Nogueira-de-Almeida, Carlos A., and Elza D. d. Mello. "Correlation of Body Mass Index Z-Scores with Glucose and Lipid Profiles among Overweight and Obese Children and Adolescents." *Jornal De Pediatria*, vol. 94, no. 3, 2018, pp. 308-312.

Appendix

R code:

```r
library(tidyr)
library(dplyr)
library(formattable)
library(knitr)

# data <- healthcare.dataset.stroke.data
data <- read.csv("/Users/yuqianxie/Desktop/healthcare-dataset-stroke-data.csv",header=T)
# data cleansing (remove row with NA value)
data[data == "N/A"] <- NA
data[data == "Unknown"] <- NA
data <- data |> drop_na()
data <- data[-c(1:3)]
data <- data[-c(3:5)]

# Set up
set.seed(20)
data$bmi <- as.numeric(data$bmi)
N <- length(data$avg_glucose_level) # population size = 3426

n <- 346 # sample size = 346 (based on study planning)

# SRS
# Continuous (Vanilla estimator)
ybar.pop <- mean(data$avg_glucose_level) # true value
SRS.indices <- sample.int(N,n,replace = F)
SRS.sample <- data[SRS.indices, ]
# attach(SRS.sample)
vanilla.estimator <- mean(SRS.sample$avg_glucose_level)

se.function  <-  function(sample.value, estimated.value) {
  res <- sample.value - estimated.value  # residual
  temp <- sum(res^2)/(n-1)
  se <- sqrt((1-n/N) *(temp/n))
  return (se)
}

vanilla.se  <- se.function(SRS.sample$avg_glucose_level, vanilla.estimator)
# vanilla.se <- sqrt((1 - n / N) * var(SRS.sample$avg_glucose_level) / n)
vanilla.srs <- c(vanilla.estimator, vanilla.se)
vanilla.lower.limit.CI <- vanilla.estimator - 1.96*vanilla.se
vanilla.upper.limit.CI <- vanilla.estimator + 1.96*vanilla.se
```

```r
# Continuous (Ratio estimator)
# Note: we use bmi as our auxiliary variable
xbar.pop <- mean(data$bmi)

ratio.estimator <- (mean(SRS.sample$avg_glucose_level) /mean(SRS.sample$bmi))*xbar.pop
ratio.se  <- se.function(SRS.sample$avg_glucose_level, (mean(SRS.sample$avg_glucose_level) /mean(SRS.sample$bmi))*SRS.sample$bmi)
ratio.srs <- c(ratio.estimator,ratio.se)
ratio.lower.limit.CI <- ratio.estimator - 1.96*ratio.se
ratio.upper.limit.CI <- ratio.estimator + 1.96*ratio.se
cor(bmi,avg_glucose_level)


# Table for SRS Continuous Data
tab <- matrix(c(vanilla.estimator,ratio.estimator,vanilla.se,ratio.se,
vanilla.lower.limit.CI,ratio.lower.limit.CI,vanilla.upper.limit.CI, ratio.upper.limit.CI)
,nrow=2,ncol=4)
colnames(tab) <- c('estimates','standard erros','lower confidence intervals','upper confidence intervals')
rownames(tab) <- c('Vanilla','Ratio')
tab <- as.table(tab)
formattable(tab) %>% kable(caption = "SRS_Continuous_Data")

# Some comments:
# We know that the standard error for the vanilla estimator is 1.836228;
# the standard error for the ratio estimator is 2.026795.
# Check the correlation value is 0.1618729; we learned that the ratio beats the vanilla only if
# there is a strong correlation between X and Y, so our output is reasonable.

# SRS Binary
# Note: we are interested in patients' glucose_level whose below 100
count <- 0
SRS.sample$avg_glucose_level <- as.array(SRS.sample$avg_glucose_level)
for (i in 1:n) {
  if (SRS.sample$avg_glucose_level[i] < 100)
    count <- count + 1
}
sample.value <- count
prop.estimator <-  sample.value / n # sample proportion (estimator)
prop.se <- sqrt((prop.estimator*(1-prop.estimator))/n)
prop.lower.limit.CI <- prop.estimator - 1.96*prop.se
prop.upper.limit.CI <- prop.estimator + 1.96*prop.se
tab3 <- matrix(c(prop.estimator,prop.se,prop.lower.limit.CI,prop.upper.limit.CI)
               ,nrow=1,ncol=4)
colnames(tab3) <- c('estimates','standard erros','lower confidence intervals','upper confidence intervals')
rownames(tab3) <- c('SRS (binary)')
tab3 <- as.table(tab3)
formattable(tab3) %>% kable(caption = "SRS_Binary_Data")
# detach(SRS.sample)
```

```r
# STR continuous
# Note: we are interested in estimating the average glucose level based on smoking status.
N.h <- tapply(data$avg_glucose_level, data$smoking_status, length)   # population size for peoples' smoking status
smoke <- names(N.h) # smoking status

# Estimate the population mean using STR with proportional allocation
STR.sample.prop <- NULL
n.h.prop <- round( (N.h/N) * n)
for (i in 1: length(smoke)) {
row.indices <- which(data$smoking_status == smoke[i])
sample.indices <- sample(row.indices, n.h.prop[i], replace = F)
STR.sample.prop <- rbind(STR.sample.prop, data[sample.indices, ])
}
ybar.h.prop <- tapply(STR.sample.prop$avg_glucose_level, STR.sample.prop$smoking_status, mean)
var.h.prop <- tapply(STR.sample.prop$avg_glucose_level, STR.sample.prop$smoking_status, var)
se.h.prop <- sqrt((1 - n.h.prop / N.h) * var.h.prop / n.h.prop)
# rbind(ybar.h.prop, se.h.prop)
ybar.str.prop <- sum(N.h / N * ybar.h.prop)
se.str.prop <- sqrt(sum((N.h / N)^2 * se.h.prop^2))
str.prop <- c(ybar.str.prop, se.str.prop)
STR.CI <- c(ybar.str.prop - 1.96*se.str.prop , ybar.str.prop + 1.96*se.str.prop)
STR.lower.limit.CI <- ybar.str.prop - 1.96*se.str.prop
STR.upper.limit.CI <- ybar.str.prop + 1.96*se.str.prop

# Table for STR Continuous Data
tab2 <- matrix(c(ybar.str.prop,se.str.prop,STR.lower.limit.CI,STR.upper.limit.CI)
               ,nrow=1,ncol=4)
colnames(tab2) <- c('estimates','standard erros','lower confidence intervals','upper confidence intervals')
rownames(tab2) <- c('STR (continuous)')
tab2 <- as.table(tab2)
formattable(tab2) %>% kable(caption = "STR_Continuous_Data")

# STR binary
ybar.prop <- tapply(as.numeric(STR.sample.prop$avg_glucose_level < 100), STR.sample.prop$smoking_status, mean)
var.prop <- ybar.prop * ( 1- ybar.prop)
se.prop <- sqrt((1 - n.h.prop / N.h) * var.prop / n.h.prop)
ybar.prop <- sum(N.h / N * ybar.prop)
se.new.prop <- sqrt(sum((N.h / N)^2 * se.prop^2))
str.new.prop <- c(ybar.prop, se.new.prop)
STR.new.lower.limit.CI <- ybar.prop - 1.96*se.new.prop
STR.new.upper.limit.CI <- ybar.prop + 1.96*se.new.prop

# Table for STR Binary Data
tab4 <- matrix(c(ybar.prop,se.new.prop,STR.new.lower.limit.CI,STR.new.upper.limit.CI)
               ,nrow=1,ncol=4)
colnames(tab4) <- c('estimates','standard erros','lower confidence intervals','upper confidence intervals')
rownames(tab4) <- c('STR (binary)')
tab4 <- as.table(tab4)
formattable(tab4) %>% kable(caption = "STR_Binary_Data")
```

PART II:
Paper Summary

In the field of statistics, there exists many ways to test for the true value of some statistics that we want. Recently, many test procedures built upon a commonly-accepted best testing method, Likelihood Ratio Tests (LRTs) have been found to produce worse results than newly developed testing methods. This has led many people to believe that the LRTs are in fact not the best testing method, and that the method of LRTs is flawed. This paper argues that many of the new tests claimed to be superior to LRTs disregard intuition and use loose hypotheses in the optimization process, and are often of no practical value. From that the author further shows that LRTs is still the most practical method for some types of statistical situations, namely non-Bayesian parametric hypothesis testing. In conclusion, the paper claims that the statistical community should not casually discard LRTs as an important statistical inference tool.