STAT344 Group Project

Pulse Pressure Exploration in Cardiovascular Health

1 Introduction

Cardiovascular diseases (CVDs) are a health concern causing a considerable amount of illness and mortality worldwide. These diseases include a variety of disorders that impact the heart and blood arteries, such as hypertension, coronary artery disease, stroke and heart failure (Cardiovascular diseases (CVDs), 2021) C. Among these diseases, pulse pressure emerges as an important metric that provides a detailed look at cardiovascular health. Pulse pressure is defined as the difference between diastolic blood pressure and systolic blood pressure that provides insights into stiffness and cardiovascular risk (Homan et al., 2023) C. Its close relationship with CVDs makes it a focal point for exploration, motivating our endeavour to investigate it.

This study is grounded in the analysis of the Cardiovascular Disease Dataset sourced from the Kaggle website (Ulianova, S., 2018) C. The data is a collection of 70,000 patient records which is obtained at the moment of a medical examination. The targeted population comprises all individuals (patients) undergoing the medical examination.

Our primary objectives pivot toward pulse pressure parameters. Firstly, we aim to estimate the average pulse pressure of all patients in the medical examination. Secondly, we aim to estimate the proportion of all patients having high pulse pressure in the medical examination. To distill meaningful insights, we will employ both simple random sampling (SRS) and stratified sampling methods in our study. The two different sampling approaches recognize the significance of capturing the natural diversity within the dataset, which may enhance the robustness and representativeness of our findings.

By concentrating on these two pulse pressure parameters, our goal is to provide a nuanced insight into pulse pressure levels across a diverse population. It enables public health planners to finely tailor interventions and educational campaigns, fostering the development of targeted strategies that promote heart health and proactively prevent cardiovascular diseases.

Objectives:

The objectives of our study centred around two facets:

(1) We aim to estimate the average pulse pressure of all patients in the medical examination, providing a quantitative overview of cardiovascular health;

(2) We aim to estimate the proportion of all patients in the medical examination who exhibit high pulse pressure, providing us with a different view of cardiovascular health.

2 Data Collection and Data Summary

In this study, our focus is on the entire population of 70,000 individuals undergoing the medical examination based on the Cardiovascular Disease dataset. The parameters in this study are the average pulse pressure among all the individuals and the proportion of individuals with over 60mmHg pulse pressure.

Data Cleansing and Data Wrangling

The initial dataset contains 12 columns that capture information about age, height, weight, gender, blood pressure (both systolic and diastolic), cholesterol and glucose levels, smoking habits, alcohol consumption, physical activity level, and the presence or absence of cardiovascular disease.

Given that our study focuses on blood pulse pressure, we performed data cleansing to eliminate outliers characterized by systolic and diastolic blood pressure values that are implausibly extreme for human physiology. Specifically, we identified and removed individual data with systolic and diastolic blood pressure values below 40 or above 300. (see appendix [B1]) B

To enhance our estimation capabilities, we introduced three new columns derived from existing data: BMI, pulse pressure, and high pulse pressure (see appendix [B2]) B. The first two are continuous variables, and the last one is binary.

The "BMI" column is computed as weight divided by the square of height. Its inclusion is motivated by previous research indicating a positive correlation between blood pressure and Body Mass Index (BMI). Elevated BMI is associated with increased body fat, raising the risk of elevated overall blood pressure (Landi et al., 2018) C. Considering the research, BMI may exhibit a strong positive correlation with pulse pressure, making it a potential auxiliary variable for pulse pressure in ratio estimation.

The values of the column "pulse pressure" are the difference between systolic and diastolic blood pressure. This new column allows us to consolidate the status levels of systolic and diastolic blood pressure, providing an overview of the overall blood pressure status across all individuals in the medical examination.

In the "high pulse pressure" column, individuals with a pulse pressure surpassing 60mmHg are classified as having high pulse pressure, represented by the value 1 in the dataset. Conversely, a value of 0 signifies that an individual's pulse pressure falls below 60 mmHg.

After completing the processes of data cleansing and wrangling, the dataset now encompasses 15 variables and 68,759 observations, laying the foundation for our targeted study.

Data Sampling

To gather the sample, we employ both simple random sampling (SRS) and stratified sampling methods. For both sampling approaches, we have chosen a sample size of $n = 200$. This decision aims to avoid decreased variability that occurs when sampling without replacement from a finite population to ensure that our sample provides more precise and unbiased estimates of the population variance.

Furthermore, we set the sample size to 200 to address the condition $np >= 10$. Since we lack information about the population proportion of individuals with high pulse pressure, we make a conservative guess by setting $p = 0.5$. This guess is made in the idea that, when $p = 0.5$, the guessed population variance $p * (1 - p)$ is maximized. By choosing $p = 0.5$, $n = 200$, we ensure that $np = 100$, which is surpassing the threshold of 10.

Simple random sampling

To implement SRS, we randomly collect 200 individuals' data from the population without replacement. (see appendix [B3]) B

Stratified sampling

We define strata based on the presence or absence of cardiovascular disease in individuals, guided by previous research suggesting a potential association between elevated pulse pressure and an elevated risk of developing cardiovascular disease. To implement stratified sampling, we partition the population into two distinct sub-populations. The first stratum (labeled ”0” in the ”cardio” column) includes those without cardiovascular diseases, while the second stratum (labeled ”1” in the ”cardio” column) consists of individuals with cardiovascular diseases. (see appendix [B4]) B

Due to insufficient information regarding the population variance within each stratum, we make the assumption that the variance is consistent across both strata. Consequently, we employ proportion allocation to determine the sample size for each stratum in achieving the optimal estimation results.

With proportional allocation, we apply the simple random sampling method to sample 101 patients from the sub-population of people who do not have cardiovascular diseases and 99 patients from the sub-population of people who have cardiovascular diseases.

3 Data Analysis

Recalling from the introduction, in this study, we focus on the pulse pressure for all patients in the medical examination. We will study pulse pressure from two perspectives: the average pulse pressure of all patients in the medical examination and the proportion of patients with high pulse pressure in the medical examination. To study both parameters, we consider three different estimations: vanilla estimation, ratio estimation, and regression estimation.

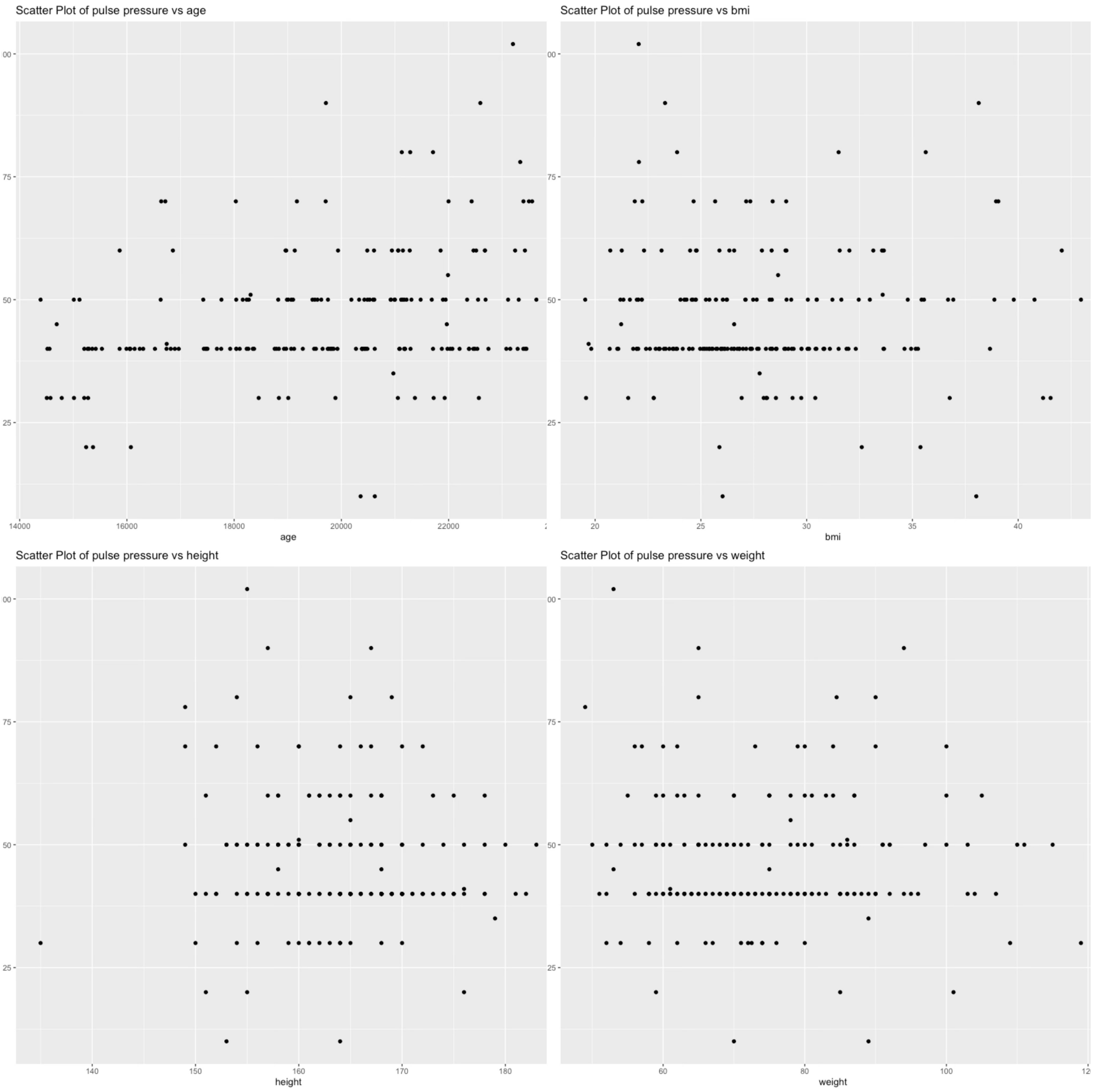
Before performing ratio estimation, we obtain the correlation between pulse pressure and other variables. We compute the correlation using cor() function in R. (see appendix [B5]) B

```
          id    age  gender   height   weight   ap_hi   ap_lo cholesterol
[1,] -0.04882 0.3178 0.05935 -0.06515 -0.03577 0.82833 0.11671     0.20396

        gluc   smoke     alco  active  cardio      bmi ap_pulse high_ap_pulse
[1,] 0.14509 -0.0094 -0.04024 0.05794 0.37974 -0.00548        1       0.77853
```

Based on the result, we see that except for “systolic blood pressure” (ap_hi) and “high pulse pressure” (high_ap_pulse), none of the variables have a strong positive correlation with pulse pressure. (Since the variable “pulse pressure” is derived from systolic blood pressure and the variable “high pulse pressure” is unknown, we do not use these as the auxiliary variable.) Thus using any of the variables as an auxiliary variable may not provide us with a good estimate. Therefore, we will abandon the ratio estimation method in this study.

We are also interested in whether regression estimation can provide us with a good estimate. We first explore the relationships between pulse pressure and other continuous variables via plots. The visualizations are plotted using the ggplot 2 package in R (see appendix [B5]) B.



Since none of the continuous variables exhibit a strong linear relationship with pulse pressure, we will not apply regression estimation in this study.

In vanilla estimation, where sample statistics serve as direct estimators for population parameters, it is sensible to apply the vanilla estimation method for estimating the average pulse pressure and the proportion of individuals with high pulse pressure across the entire population in our study.

Estimate Average Pulse Pressure with SRS

According to the findings above, in this study, we will only use vanilla estimation to estimate the average pulse pressure for all patients in the medical examination with the sampled data. We denote the sample as S and the estimated average pulse pressure as

$\bar{y}_s$, which estimates the average pulse pressure $\bar{y}_p$. The average pulse pressure is estimated with the following formula (see appendix [A1]) A:

$$\bar{y}_s = \frac{\sum_{i=1}^{n} yi}{n}$$

The standard error (SE) for the estimate $\bar{y}_s$ is computed by (see appendix [A2]) A:

$$SE(\bar{y}_s) = \sqrt{\frac{(n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y}_s)^2}{n}}$$

The 95% confidence interval for $\bar{y}_s$ is computed by (see appendix [A3]) A:

$$\bar{y}_s \pm 1.96 * SE(\bar{y}_s)$$

Based on the results (see appendix [B6]) B, the estimated mean pulse pressure of all patients in the medical examination is 46.31 mmHg. The standard error of the estimated mean pulse pressure is 0.922. The 95% confidence interval is [44.501, 48.117], thus we are 95% confident that the true mean pulse pressure falls in this interval.

Estimate Average Pulse Pressure with Stratified Sampling

As mentioned earlier in the section, when implementing sampling we chose a sample of 101 individuals from the subgroup without cardiovascular diseases and 99 individuals from the subgroup with cardiovascular diseases.

We first calculate the estimated average pulse pressure for each stratified sample, and we denote this as $\bar{y}_{str}$. For the non-cardiovascular diseases group, the estimated average pulse pressure is 39.11 mmHg. For the cardiovascular diseases group, the estimated average pulse pressure is 49.17 mmHg. The standard errors are 1.0513 and 1.0670 respectively. Then, we combine the estimates for each stratified sample into one using the formula (see appendix [A4]) A:

$$\bar{y}_{str} = \sum_{h=1}^{H} (\frac{N_h}{N}) \bar{y}_{s_h}$$

We obtain the standard error of the estimate $bary_{str}$ by (see appendix [A5]) A:

$$SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N})^2 SE^2(\bar{y}_{str})}$$

The 95% confidence interval for the estimate is computed by (see appendix [A6]) A:

$$\bar{y}_{str} \pm 1.96 * SE(\bar{y}_{str})$$

Based on the results (see appendix [B7]) B, the estimated average pulse pressure for all patients in the medical examination is 44.09 mmHg and the standard error for the estimated average pulse pressure is 0.7489. The 95% confidence interval is [42.621,45.557], so we are 95% confident that the true mean pulse pressure falls in this interval.

Recall the estimation of SRS, the estimated mean pulse pressure is 46.31 and the standard error is 0.9222. Compared to the estimation of stratified sampling, the estimated mean pulse pressure is 44.09 and its standard error is 0.7489, which is smaller than 0.9222. Based on the obtained results, we conclude that stratified sampling provides a better estimate of the mean pulse pressure of all patients in the medical examination because of its smaller standard error.

Estimate the Proportion of Patients with High Pulse Pressure with SRS

To comprehensively investigate the pulse pressure among all patients in the medical study, we want to study from a different perspective - the proportion of patients with high pulse pressure in the medical examination. We denote the proportion of patients with high pulse pressure as p, and the estimate of the proportion with high pulse pressure as $\hat{p}$. With the sampled data, the estimated proportion of high pulse pressure is 0.185 (see appendix [A7]) A. The standard error of the estimated proportion is (see appendix [A8]) A:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}*(1-\hat{p})}{n})}$$

The 95% confidence interval of $\hat{p}$ is computed by (see appendix [A9]) A:

$$\hat{p} \pm 1.96 * SE(\hat{p})$$

From the results (see appendix [B8]) B, the estimated proportion of people with high pulse pressure in the medical examination is 0.185 and its standard error is 0.02746. The 95% confidence interval is [0.1312,0.2388], thus we are 95% confident that the true proportion of people with high pulse pressure falls in this interval.

Estimate the Proportion of Patients with High Pulse Pressure with Stratified Sampling

Next, we estimate the proportion of patients with high pulse pressure using our stratified sample. We denote this estimate as $\hat{p}_{str}$. For the non-cardiovascular diseases group, the estimated proportion of patients with high pulse pressure is 0.02970. For the cardiovascular diseases group, the estimated proportion of patients with high pulse pressure is 0.2929. The standard errors are 0.01689 and 0.04574 respectively. Then, we combine our estimate for each stratified sample as one (see appendix [A10]) A:

$$\hat{p}_{str} = \sum_{h=1}^{H}(\frac{N_h}{N})\hat{p}_{S_h}$$

The standard error of the estimated proportion is computed by (see appendix [A11]) A:

$$SE(\hat{p}_{str}) = \sqrt{\sum_{h=1}^{H}(\frac{N_h}{N})^2 SE^2(\bar{p}_{str})}$$

The 95% confidence interval is (see appendix [A12]) A:

$$\hat{p}_{str} \pm 1.96 * SE(\hat{p}_{str})$$

Based on the above results (see appendix [B9]) B, the estimated proportion of people with high pulse pressure in the medical examination is 0.160 and its standard error is 0.02419. The 95% confidence interval is [0.1126, 0.2074], thus we are 95% confident that the true proportion of people with high pulse pressure falls in this interval.

In summary, with SRS, the estimated proportion is 0.185 with SE of 0.02746; with stratified sampling, the estimated proportion is 0.160 with SE of 0.02419. In this case, stratified sampling gives a relatively smaller standard error than SRS. Therefore, we conclude that stratified sampling provides a better estimate on estimating the proportion of people with high pulse pressure in the medical examination.

Advantages and Limitations of Each Sampling Method

SRS ensures that each individual in the population has an equal chance of being selected in the sample, which helps in creating a representative sample of the entire population, so the sample is less likely to be biased and gives more unbiased results. Also, when the population size is reasonably large, it's straightforward and easy to implement the sampling procedure as each individual is selected independently from the other.

However, as the population size grows, it will be hard to sample using SRS as each member in the population needs to be listed and randomly selected, which is time-consuming, expensive and very inefficient. Additionally, SRS does not guarantee that certain subpopulations within the population can be adequately represented by the sample, which may lead to potential bias.

Stratified sampling can appropriately solve the above concern. By using stratified sampling, the population is divided into different stratum and we sample from each stratum which ensures that each stratum can be adequately represented. Moreover, when there are variations between each stratum, stratified sampling allows for better comparison between strata. This leads to increased precision and reliability of the results, and it also provides more accurate statistical inferences and detailed insights.

Although stratified sampling usually provides us with smaller standard errors and better estimation, it also has limitations. As we lack information about the population data, we made an assumption that they share the same variance within each stratum. However, in reality, this may not hold true. Consequently, the chosen sample sizes for each stratum might not yield the most optimal estimation. Additionally, when there are numerous strata, it might be complicated and hard to implement the sampling procedure as we need to sample from each stratum which can be time-consuming. Also, to effectively implement stratified sampling, good knowledge about the population and the characteristics of different strata might be needed, which is not always available. Thus, if the population is misclassified to not well-defined strata, it's possible to receive biased results.

Both SRS and stratified sampling can provide us with sensible, reliable and unbiased results, but each does have some limitations. In this study, the stratified sampling method gives better estimates so we chose stratified sampling as our sampling method. This does not mean stratified sampling is always a better choice than SRS, the choice of sampling methods really depends on the research objectives, the nature of the population being studied, and the available resources in the study.

4 Conclusion

In summarising our findings, the estimated mean pulse pressure for individuals in the medical examinations using the simple random sampling (SRS) method is 46.31, with a standard error of 0.922. Employing the stratified sampling method yields an estimated mean pulse pressure of 44.09, accompanied by a smaller standard error of 0.749. The reduced standard error in the stratified sampling approach suggests its superior performance in estimating mean pulse pressure.

Similarly, when estimating the proportion of individuals with high pulse pressure, SRS produces an estimated proportion of 0.185 with a standard error of 0.0275. In contrast, the stratified sampling method yields a lower standard error of 0.0242, resulting in a more precise estimated proportion of 0.160. This consistent pattern reinforces the conclusion that, across both parameters, the stratified sampling method consistently outperforms SRS.

It is essential to recognize that while stratified sampling demonstrates superior estimation precision, each sampling method presents its unique set of advantages and limitations. SRS, being straightforward and easily calculable, may lack representativeness for all subgroups in the population, leading to increased variation and standard error. On the other hand, the meticulous nature of stratified sampling, ensures representation from distinct strata but introduces potential bias if the strata are not well-defined or if sample sizes within each stratum are not optimally allocated.

The conclusion drawn from this study, particularly regarding the estimated mean pulse pressure and the proportion of individuals with high pulse pressure, can provide sensible results for our targeted population but should be cautiously generalized to larger or other populations. The study focuses on individuals aged between 30 and 64, extrapolating these findings to populations beyond this age range may introduce bias as age significantly influences cardiovascular health and blood pulse pressure. In addition, since the dataset is sourced from Kaggle and only captures a finite set of patients during a medical examination, it may not fully represent the demographics and characteristics of broader populations. Furthermore, the conclusion emphasizes the superior performance of stratified sampling over simple random sampling in estimating pulse pressure parameters. While this finding holds within the confines of the dataset and the defined age range, the efficacy of sampling methods may vary based on the unique characteristics of different populations.

Thus, while the conclusion offers valuable insights for the studied population, its broader applicability requires careful consideration of age-related factors and the specific characteristics of the target population. Future research endeavours including diverse age ranges and demographic profiles are essential for extending and validating these findings across broader populations. Taking this study as a stepping stone, future investigations can refine and broaden the understanding of pulse pressure estimation using different sampling methods.

5 Appendix

A Formula and Calculation

1. $\bar{y}_s = \frac{\sum_{i=1}^{n} yi}{n} = \frac{9262}{200} = 46.31$

2. $SE(\bar{y}_s) = \sqrt{\frac{(n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y}_s)^2}{n}} = \sqrt{\frac{(200-1)^{-1}\sum_{i=1}^{200}(y_i - 46.31)^2}{200}} = \sqrt{\frac{170.0843}{200}} = 0.92218$

3. $\bar{y}_s \pm 1.96 * SE(\bar{y}_s) = [46.31 - 1.96 * 0.92218, 46.31 + 1.96 * 0.92218] = [44.50252, 48.11748]$

4. $\bar{y}_{str} = \sum_{h=1}^{H}(\frac{N_h}{N})\bar{y}_{s_h} = 0.505 * 39.1089 + 0.495 * 49.1717 = 44.08887$

5. $SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H}(\frac{N_h}{N})^2 SE^2(\bar{y}_{str})} = \sqrt{0.505^2 * 1.0513^2 + 0.495^2 * 1.0670^2} = 0.74887$

6. $\bar{y}_{str} \pm 1.96 * SE(\bar{y}_{str}) = [44.09 - 1.96 * 0.74887, 44.09 + 1.96 * 0.74887] = [42.62109, 45.55665]$

7. $\hat{p} = \frac{number\, of\, people\, with\, high\, pulse\, pressure}{number\, of\, people\, in\, the\, sample} = 0.185$

8. $SE(\hat{p}) = \sqrt{\frac{\hat{p}*(1-\hat{p})}{n})} = \sqrt{\frac{0.185*(1-0.185)}{200}} = 0.02746$

9. $\hat{p} \pm 1.96 * SE(\hat{p}) = [0.185 - 1.96 * 0.02746, 0.185 + 1.96 * 0.02746] = [0.13118, 0.23882]$

10. $\hat{p}_{str} = \sum_{h=1}^{H}(\frac{N_h}{N})\hat{p}_{S_h} = 0.505 * 0.02970 + 0.495 * 0.2929 = 0.160$

11. $SE(\hat{p}_{str}) = \sqrt{\sum_{h=1}^{H}(\frac{N_h}{N})^2 SE^2(\bar{p}_{str})} = \sqrt{0.505^2 * 0.01689^2 + 0.495^2 * 0.04574^2} = 0.02419$

12. $\hat{p}_{str} \pm 1.96 * SE(\hat{p}_{str}) = [0.160 - 1.96 * 0.02419, 0.160 + 1.96 * 0.02419] = [0.11256, 0.20738]$

B Code

```
### Set up
set.seed(1)
options(digits=12)
data <- read.csv("cardio_train.csv", header = TRUE, sep=";")
head(data)
  id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active
1  0 18393      2    168     62   110    80           1    1     0    0      1
2  1 20228      1    156     85   140    90           3    1     0    0      1
3  2 18857      1    165     64   130    70           3    1     0    0      0
4  3 17623      2    169     82   150   100           1    1     0    0      1
5  4 17474      1    156     56   100    60           1    1     0    0      0
6  8 21914      1    151     67   120    80           2    2     0    0      0

cardio
     0
     1
     1
     1
     0
     0


[1]
# filter the outliers out
data <- data[data$ap_hi < 300 &
data$ap_hi > 40 & data$ap_lo < 300 & data$ap_lo > 40, ]

[2]
# add columns for BMI, pulse pressure, high pulse pressure
data$bmi <- data$weight / (data$height/100)^2
data$ap_pulse <- data$ap_hi - data$ap_lo
data$high_ap_pulse <- ifelse(data$ap_pulse >= 60, 1, 0)

head(data)
  id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active
1  0 18393      2    168     62   110    80           1    1     0    0      1
2  1 20228      1    156     85   140    90           3    1     0    0      1
3  2 18857      1    165     64   130    70           3    1     0    0      0
4  3 17623      2    169     82   150   100           1    1     0    0      1
5  4 17474      1    156     56   100    60           1    1     0    0      0
6  8 21914      1    151     67   120    80           2    2     0    0      0
  cardio           bmi  ap_pulse high_ap_pulse
1      0 21.9671201814        30             0
2      1 34.9276791584        50             0
3      1 23.5078053260        60             1
4      1 28.7104793250        50             0
5      0 23.0111768573        40             0
6      0 29.3846761107        40             0

# sample size
n = 200
N = 68759

[3]
### Simple Random Sampling
sample_data = data[sample(N,n,replace=F),]

[4]
### Stratified sampling based on whether the person has this disease
N.h <- tapply(data$ap_pulse, data$cardio, length)
n.h.prop <- round((N.h/N) * n)
cardios <- names(N.h)
STR.sample.prop <- NULL

for (i in 1: length(cardios))
{
row.indices <- which(data$cardio == cardios[i])
sample.indices <- sample(row.indices, n.h.prop[i], replace = F)
STR.sample.prop <- rbind(STR.sample.prop, data[sample.indices, ])
}

attach(sample_data)

[5]
# Check correlation via cor() and plots
cor(ap_pulse, sample_data)
## id age gender height
## [1,] -0.0488195953548 0.317796800495 0.0593461452453 -0.0651542530957
## weight ap_hi ap_lo cholesterol
## [1,] -0.0357734738085 0.82832667518 0.116709553965 0.203957217743
## gluc smoke alco active
## [1,] 0.145092444772 -0.00940115366163 -0.0402403067763 0.0579401942691
## cardio bmi ap_pulse high_ap_pulse
## [1,] 0.37973730815 -0.00547532084204 1 0.77853285773

# Plots
library(ggplot2)

for (col in colnames(sample_data)[-1]) {
  p <- ggplot(sample_data, aes_string(x = col, y = "ap_pulse")) +
    geom_point() +
    labs(title = paste("Scatter Plot of pulse pressure vs", col),
         x = col,
         y = "pulse pressure")

  print(p)
}


[6]
### Parameter 1: pulse pressure as a continuous variable with SRS
# Compute estimate
vanilla.est.p1 <- mean(ap_pulse)
vanilla.est.p1
[1] 46.31

# Compute SE
se.function <- function(sample.value, estimated.value)
{
res = sample.value - estimated.value # residual
temp = sum(res^2)/(n-1)
se = sqrt(temp/n)
return (se)
}
vanilla.se <- se.function(ap_pulse, vanilla.est.p1)
vanilla.se
[1] 0.922183066446
```

```
# Compute 95% CI
vanilla.lower_bound.p1 <- round(vanilla.est.p1 - 1.96 * vanilla.se, 5)
vanilla.upper_bound.p1 <- round(vanilla.est.p1 + 1.96 * vanilla.se, 5)
cat("confidence interval:[",vanilla.lower_bound.p1,',',
                            vanilla.upper_bound.p1,"]")
## confidence interval: [ 44.50252, 48.11748 ]

detach(sample_data)

[7]
### Parameter 1: pulse pressure as continuous variable with stratified sampling

ybar.h.prop <- tapply(STR.sample.prop$ap_pulse, STR.sample.prop$cardio, mean)
var.h.prop <- tapply(STR.sample.prop$ap_pulse, STR.sample.prop$cardio, var)
se.h.prop <- sqrt(var.h.prop / n.h.prop)
rbind(ybar.h.prop, se.h.prop)
##                           0              1
## ybar.h.prop 39.10891089109 49.17171717172
## se.h.prop    1.05125111602 1.06702224323

# Compute estimate
ybar.str.prop.p1 <- sum(N.h / N * ybar.h.prop)
ybar.str.prop.p1
[1] 44.0888723815

# Compute SE
se.str.prop.p1 <- sqrt(sum((N.h / N)^2 * se.h.prop^2))
se.str.prop.p1
[1] 0.748868611143

#Compute 95% CI
str.prop.lower_bound.p1 <- round(ybar.str.prop.p1 - 1.96 * se.str.prop.p1, 5)
str.prop.upper_bound.p1 <- round(ybar.str.prop.p1 + 1.96 * se.str.prop.p1, 5)
cat("confidence interval: [",str.prop.lower_bound.p1,',',
                             str.prop.upper_bound.p1,"]")
## confidence interval: [ 42.62109 , 45.55665 ]


[8]
### Parameter 2: pulse pressure as binary variable with SRS

# Compute estimate
p_high_ap_pulse_sample <- mean(sample_data$high_ap_pulse)
vanilla.est.p2 <- p_high_ap_pulse_sample
vanilla.est.p2
[1] 0.185

# Compute SE
vanilla.se.p2 <- sqrt(p_high_ap_pulse_sample*(1-p_high_ap_pulse_sample)/n)
vanilla.se.p2
[1] 0.0274567842254

# Compute 95% CI
vanilla.lower_bound.p2 <- round(vanilla.est.p2 - 1.96 * vanilla.se.p2, 5)
vanilla.upper_bound.p2 <- round(vanilla.est.p2 + 1.96 * vanilla.se.p2, 5)
cat("confidence interval: [", vanilla.lower_bound.p2, ',',
                              vanilla.upper_bound.p2, "]")
## confidence interval: [ 0.13118 , 0.23882 ]

[9]
### Parameter 2: pulse pressure as binary variable with stratified sampling

ybar.h.prop_p2 <- tapply(STR.sample.prop$high_ap_pulse,
                         STR.sample.prop$cardio, mean)
se.h.prop_p2 <- sqrt((ybar.h.prop_p2 * (1 - ybar.h.prop_p2)) / n.h.prop)
rbind(ybar.h.prop_p2, se.h.prop_p2)
##                              0               1
## ybar.h.prop_p2 0.0297029702970 0.2929292929293
## se.h.prop_p2   0.0168924096413 0.0457399017142

# Compute estimate
ybar.str.prop.p2 <- sum(N.h / N * ybar.h.prop_p2)
ybar.str.prop.p2
[1] 0.15997050337

# Compute SE
se.str.prop.p2 <- sqrt(sum((N.h / N)^2 * se.h.prop_p2^2))
se.str.prop.p2
[1] 0.0241908819403

# Compute 95% CI
str.prop.lower_bound.p2 <- round(ybar.str.prop.p2 - 1.96 * se.str.prop.p2, 5)
str.prop.upper_bound.p2 <- round(ybar.str.prop.p2 + 1.96 * se.str.prop.p2, 5)
cat("confidence interval: [", str.prop.lower_bound.p2, ',',
                              str.prop.upper_bound.p2, "]")
## confidence interval: [ 0.11256 , 0.20738 ]
```

C References

1. Cardiovascular diseases (CVDs). (2021, June 11). *World Health Organization.* Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2. Homan, T., Bordes, S., Cichowski, E. (2023, July 10). Physiology, Pulse Pressure. National Library of Medicine. Retrieved from https://www.ncbi.nlm.nih.gov/books /NBK482408/

3. Ulianova, S. (2018). Cardiovascular Disease dataset. Retrieved from https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/

4. Landi, F., Calvani, R., Picca, A., Tosato, M., Martone, A. M., Ortolani, E., Sisto, A., D'Angelo, E., Serafini, E., Desideri, G., Fuga, M. T., Marzetti, E. (2018). Body Mass Index is Strongly Associated with Hypertension: Results from the Longevity Check-up 7+ Study. *Nutrients, 10(12), 1976.* https://doi.org/10.3390/nu10121976

5. Tang, KS, Medeiros ED, Shah, AD. (2020). Wide pulse pressure: A clinical review. *J Clin Hypertens, 22(11):1960-1967.*https://doi.org/10.1111/jch.14051.

The paper “The Emperor’s New Tests” conveys and emphasizes the importance of common sense in Statistics (e.g. making statistical decisions, creating new tests for multiparameter hypothesis testing, etc.). The paper starts with a statistical allegory of an Emperor who relies on statistical tests (especially the likelihood ratio test, abbreviated as LRT) to make wise decisions for his domain. A young statistician, who creates new tests that are theoretically superior to LRT but often lead to counter-intuitive results that lack practicality, was first appointed as the better test creator in the Imperial Court but later dismissed because of the defectiveness (violation of statistical intuition) in his “better tests”.

The key message in both the tale and the paper is that when someone considers a statistical test, he/she should not blindly follow and believe it without thinking about real-world relevance. One needs to always keep in mind not only the statistical accuracy but also the intuitiveness of the tests. In other words, even though a statistical test may be unbiased and more powerful than any other test, if it intuitively doesn’t make sense, one needs to consider carefully whether it’s appropriate to use it. People should also remember that statistics serves common sense and intuition, but does not replace them (if we have intuitive thoughts, we find a statistical way to confirm and prove our thoughts). In summary, mathematical statistics is coupled with common sense and intuition to ensure the statistical results are meaningful and have practical value.

In addition, we also want to highlight from the abstract that, the authors contended that although the LR criterion is not the most powerful test, it remains a generally reasonable first option for non-Bayesian parametric hypothesis-testing problems.