Heart Disease

Final Report

Math 3430 - Sample Survey Design

York University

Ammar Mughal , Aayushi Garg , Jasveen Kaur Khanuja

April 4 , 2025

Abstract

The purpose of the report is to analyze the variables in heart disease data that can affect the heart disease.
The data was collected from the Kaggle site. Telephone survey was conducted to collect data on the health status of US residents. The data was collected in all 50 states , the District of Columbia and three U.S territories by conducting approx 400,000 adult interviews. However the data was shortlisted to 319795 rows ie. individuals in the dataset and 18 columns ie. Variables affecting heart disease for analysis. This study focuses on two factors - BMI and sleep time and preliminary findings suggest that high BMI may have a significant impact , while sleep time shows no clear effect in the stratified samples examined. Further investigation will explore broader trends and relationships within the dataset.

Introduction

According to the CDC(US Centre for Disease Control and Prevention) heart disease is the leading cause of death in the United States. In the year 2022 alone, 702,880 people died from heart disease. That is the equivalent of 1 in every 5 deaths. So we decided to analyze factors associated with heart disease among individuals. The dataset, sourced from Kaggle, was collected through a telephone survey by the CDC on the health status of U.S. residents across the country. It includes approximately 400,000 adult interviews, later narrowed down to 319,795 rows i.e. the total number of individuals and 18 columns i.e. total number of factors affecting heart disease. Among the 18 factors, we decided to focus on BMI and Sleep time for our analysis.

Survey problem: What factors are associated with the prevalence of heart disease among individuals?

Objective : To identify the relationship between heart disease and variables such as BMI, smoking, alcohol drinking, physical health, sleep time etc.

Collect real data
The data was collected from Kaggle. A telephone survey was conducted by the CDC to collect data on the health status of US residents. The data was collected in all 50 states by conducting approx 400,000 adult interviews. However the data was shortlisted to 319795 rows and 18 columns (i.e variables) affecting heart disease for analysis. Variables like BMI, smoking, alcohol drinking, physical health, sleep time etc were considered.

Population (units, characteristics, mean, total, proportion, etc.);
Units: Individual units of the telephone survey population. An object on which measurement was taken
Characteristics: Variables like BMI, smoking, alcohol drinking, physical health, sleep time etc.
Parameters:

- Population Size N = 319795 (number of rows in excel file)
- Mean BMI: Average BMI of all individuals in the population.
- Mean Sleep Time: Average hours of sleep per night in the population
- Mean PhysicalHealth: Avg # of days with poor physical health in the past 30 days

- Total number of individuals with the heart disease: Total count of individuals in the population who have heart disease
- Total number of smokers: Total count of individuals in the population who smoke

Proportions (for categorical variables):

- Proportion of individuals with heart disease: Number of individuals with heart disease /Total population size
- Proportion of smokers: Number of smokersTotal population size / Total population size
- Proportion of individuals with good or very good general health:
- Number of individuals with good or very good health/Total population size

Sample factors (units and frame, questionnaire, precision, etc.):

Units: Number of samples collected from the data was 5000 samples (sample frame randomly selected) from the population of the size N = 319795 i.e. the sample units are the individuals or the adults of the telephone survey population that are selected as a sample. **Frame :** Sampling frames are the list of all the adults selected in the 5000 samples from the regions of 50 states of the District of Columbia and three US territories.

Method

1) **SRSWOR**

   SRSWOR is a method of selecting samples randomly from a population. Once a member is selected, it cannot be used again-hence the name simple random sampling without replacement. It is often used to ensure fairness and randomness in the selection process. We used SRSWOR in our analysis to ensure each unit in the population has an equal probability of selection ,with sampled units excluded from subsequent draws to prevent duplication.

   Process:

   In our analysis, from the population of approximately 319,000 individuals, **5,000 random samples** were selected. The **unbiased mean estimator** and **unbiased variance estimator** were then calculated using standard statistical formulas to derive population inferences.

2) **Stratified Random Sampling**

   Stratified Sampling is a method to divide our population into subgroups also called strata. To account for subgroup heterogeneity, the population was partitioned into strata based on shared characteristics, and random samples were drawn from each stratum. Stratified sampling is useful to get more precise estimates.

   Process:

   In our analysis, we stratified the data based on gender, with further division into four subgroups based on heart disease status:

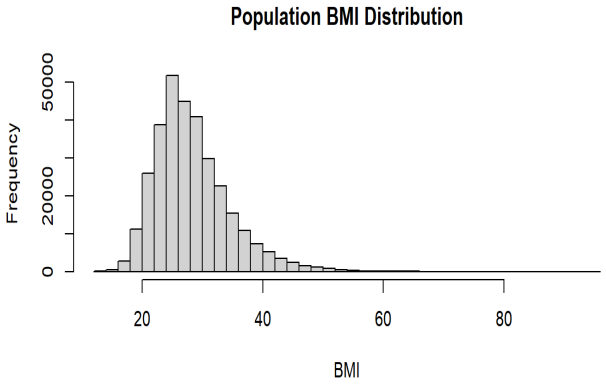
   - Female with no heart disease

- Female with heart disease
- Male with no heart disease
- Male with heart disease

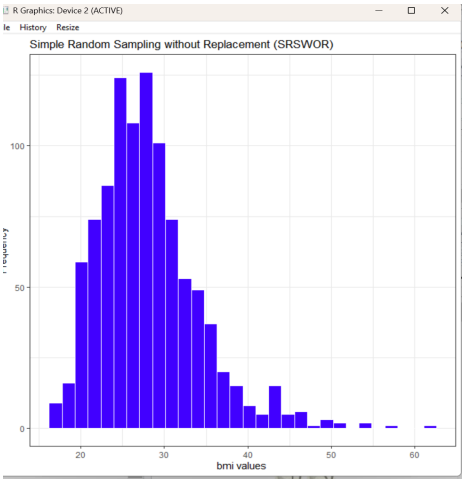
This approach ensured proportional representation of each subgroup, enhancing the precision and reliability of estimates compared to SRSWOR. We will now use both of these methods to analyse our two factors, **BMI and Sleep Time.**

Analysis on BMI

Population Parameters for BMI:
Count: 319,795 (N)
Mean: 7.097
Variance: 2.0621
Standard Deviation: 1.4360



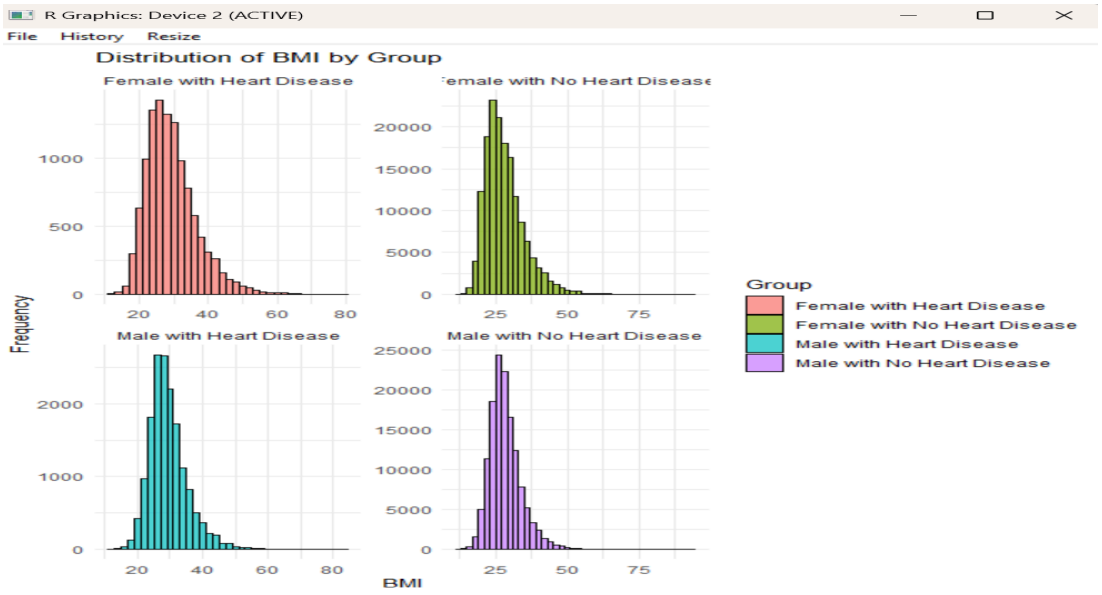
SRSWOR for BMI:
SRSWOR is chosen as a baseline sampling design
We selected 5000 samples from the population. We are interested in calculating the average BMI of the 5000 selected samples of the column BMI.
Count (n): 5000
Mean: 28.4143
Variance: 40.7259
Standard Deviation: 6.3817
Unbiased Variance Estimator: 0.008018



Stratified Sampling for BMI and General Neyman Allocation:
Stratified sampling with four groups:
Female with Heart Disease
Male with Heart Disease
Female with no Heart Disease
Male with no Heart Disease

Row Labels	Count of HeartDisease	Sum of BMI	Average of BMI2
No	**292422**	**8253511.04**	**28.22465834**
Female	156571	4395787.72	28.07536338
Male	135851	3857723.32	28.39672376
Yes	**27373**	**804809.78**	**29.40159208**
Female	11234	330009.9	29.37599252
Male	16139	474799.88	29.41941136
Grand Total	**319795**	**9058320.82**	**28.32539852**

Histogram of a sampled BMI



Results for Stratified Sampling for BMI:

Group Sample Size Sample Mean BMI Sample Variance Sample Stnd. Deviation

---

Neyman Allocation

- $N = 319,795 \mid n = 5,000$
- $N_1 = 156,571 \quad N_3 = 11,234$
  $N_2 = 135,851 \quad N_4 = 16,139$
- $\sigma_{y_1} = 6.7832 \quad \sigma_{y_3} = 6.2559$
  $\sigma_{y_2} = 5.7897 \quad \sigma_{y_4} = 5.5883$
- $n_h = n \dfrac{N_h \sigma_{y_h}}{\sum_{k=1}^{H} N_k \sigma_{y_k}} \mid \sum N_k \sigma_{y_k} = (156571)(6.7832) + (135851)(5.7897) + (11234)(6.2559) + (16139)(5.5883)$
  $\sum N_k \sigma_{y_k} = 2,009,057.296$

$\rightarrow n_1 = (5000)\dfrac{(156571)(6.7832)}{2,009,057.296} \approx 2,643.16$

$n_2 = (5000)\dfrac{(135851)(5.7897)}{2,009,057.296} \approx 1957.48$

$n_3 = (5000)\dfrac{(11,234)(6.2559)}{2,009,057.296} \approx 174.91$

$n_4 = (5000)\dfrac{(16,139)(5.5883)}{2,009,057.296} \approx 224.46$

$\sum n_h \approx 5,000.01 \approx 5,000 \Rightarrow$ Similar to what we chose for stratified sampling

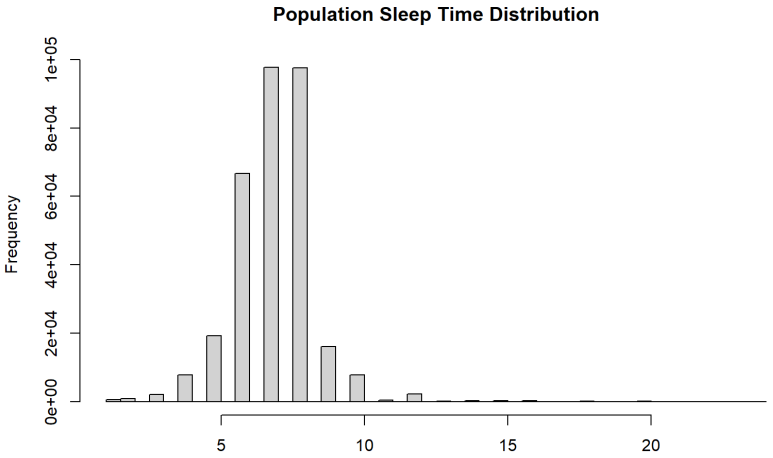
Female with No Heart Disease	2448	27.9965	46.0112	6.7832
Male with No Heart Disease	2124	28.3523	33.5208	5.7897
Female with Heart Disease	176	29.3993	39.1359	6.2559
Male with Heart Disease	252	29.3796	31.2291	5.5883

```
> # unbiased mean estimator
> sum(result$Wh*result$Average)
[1] 28.26673
> # unbiased variance estimator for \bar{Yst}
> sum((result$Wh)^2*(1/result$nh-1/result$Nh)*result$Variance)
[1] 0.007819296
>
```

Analysis on Sleep Time

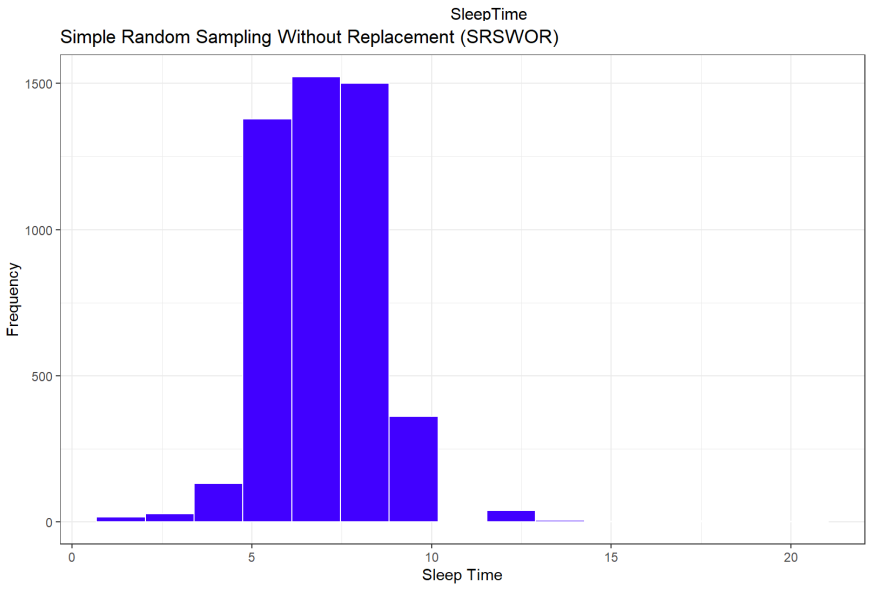
Population Parameters for Sleep Time:

Count: 319,795 (N)
Mean: 7.097
Variance: 2.0621
Standard Deviation: 1.4360

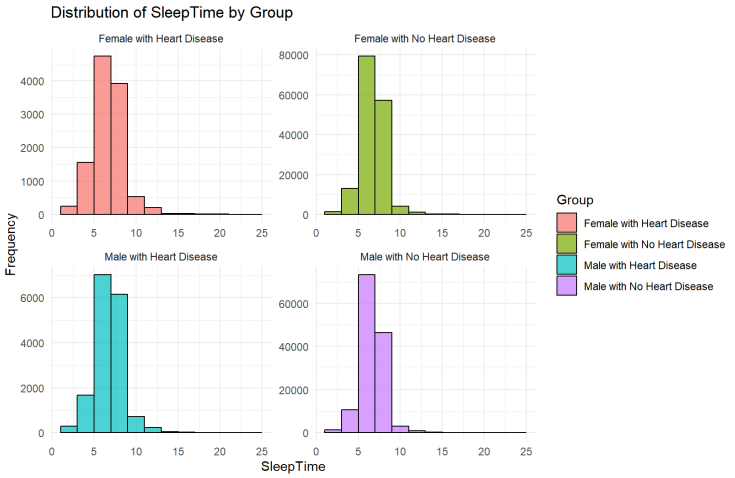


SRS for Sleep Time:

Count: 5000(n)
Mean: 7.0794
Variance: 2.0555
Standard Deviation: 1.4337
Unbiased Variance Estimator: 0.000415



Stratified Sampling for Sleep Time:




```
>
> result = data.frame(Group = Group, SampleSize = strata.n, Averag
+                     Nh = strata.N, nh = strata.n, Missing = str
> result
                          Group SampleSize  Average
1 Female with No Heart Disease       2448 7.144608
2   Male with No Heart Disease       2124 7.060264
3    Female with Heart Disease        176 7.261364
4      Male with Heart Disease        252 7.063492
  Variance         Wh     Nh   nh Missing
1 1.845856 0.48959802 156571 2448       0
2 1.665702 0.42480652 135851 2124       0
3 3.074156 0.03512875  11234  176       0
4 3.422248 0.05046671  16139  252       0
>
>
> # unbiased mean estimator
> sum(result$Wh*result$Average)
[1] 7.108786
>
> # unbiased variance estimator for \bar{Yst}
> sum((result$Wh)^2*(1/result$nh-1/result$Nh)*result$Variance)
[1] 0.0003724929
>
```

Interpretation of Results for BMI

1. Stratum Means (Average BMI per Group)

Each stratum (group) has a calculated sample mean (Average), which represents the **average BMI** for individuals in that particular category:

- **Observations:**
  - Individuals **with heart disease** tend to have a slightly **higher average BMI** than those without heart disease
  - Among **females**, the average BMI is **higher for those with heart disease** than those without (**29.40 vs. 28.00**)
  - A similar pattern is seen among **males**, where those with heart disease have a slightly higher BMI than those without (**29.38 vs. 28.35**)

2. Stratum Variances

The Variance column represents the **variability of BMI within each group**:

- **Observations:**
    - The variance indicates **how spread out the BMI values are** within each group.
    - **Females without heart disease** have the **highest BMI variance (46.01)**, meaning their BMI values are more widely distributed.
    - **Males with heart disease** have the **lowest BMI variance (31.23)**, indicating that their BMI values are more concentrated around the mean.

3. Unbiased Mean Estimator (Y¯st\bar)

- The estimated **overall population mean BMI** (across all groups, weighted by their proportion in the dataset) is **28.27**.
- This value suggests that, when considering all individuals in the dataset, the expected BMI is around **28.27**.

4. Unbiased Variance Estimator for Y¯st\bar{Y}_{st}

- The unbiased estimate of the **variance of the overall mean BMI** is **0.0078**.
- A **lower variance** suggests that **Stratified Sampling** is slightly **more precise** than **SRSWOR**

Interpretation of Results for Sleep Time

Interpretation between SRSWOR and Stratified:

- Individuals selected for the variable Sleep Time have a slightly higher mean in **Stratified Sampling than SRSWOR i.e. 7.10878 v/s 7.079**; this indicates more accuracy in stratified
- The unbiased variance estimator turned out to be lower for **Stratified Sampling than SRSWOR i.e. 0.000372 v/s 0.000415**; again indicating higher accuracy in stratified
- In general the average sleep time of the female with heart disease and no heart disease is higher than the average as compared to men; but that doesn't really say much for our required results
- The means for all stratum are very similar; meaning there is **not a significant difference in the sleep time of those with and without heart disease**
- Comparing the means for all three cases (Population vs SRSWOR vs Stratified): **7.097 vs 7.0794 vs 7.108 respectively**; we see that stratified is slightly closer to the population mean and therefore a better method/estimator than SRSWOR

Conclusion

- BMI is slightly higher for individuals with heart disease across both genders.
- Females without heart disease have the most BMI variability, while males with heart disease have the least.
- The small variance in the mean BMI estimate (0.0078) suggests better reliability in the estimation of Stratified Sampling versus SRSWOR.

- Our stratified sampling approach helped reduce bias and improve precision when estimating BMI trends across different groups
- We conclude that there is no effect of sleep time on heart disease as all stratum means were very close in value
- Stratified sampling gave slightly more accurate results than SRSWOR

References

1. Data Set: Indicators of Heart Disease (2022 UPDATE)
2. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html
3. Lecture notes
4. Sampling: design and analysis by Lohr, Sharon L. Chapman and Hall/CRC, 2019.