Indian Institute of Technology Hyderabad

MA4240: Applied Statistics

Report On

US Demographics & Cardiovascular Diseases: A Statistical Analysis

---

Authors:
Rajdeep Pathak (MA23MSCST11013)
Sonali Saha (MA23MSCST11022)
Deepak Yadav (MA23MSCST11005)
Sayani Mondal (MA23MSCST11019)
Rohit Kumar Das (MA23MSCST11016)

Course Instructor:
Dr. Sameen Naqvi

Introduction

Data-driven decision-making is crucial across industries like healthcare, insurance, fashion, and sports. Using the 2021 BRFSS dataset, this project involves statistical analysis on US demographics and various health factors that might lead to cardiovascular diseases.

The dataset provides a wealth of information on health parameters and demographics of US residents, making it ideal for exploring cardiovascular diseases. Through statistical analysis, the project aims to uncover insights transcending disciplinary boundaries. Methods like confidence interval estimation allow for robust inference on demographic parameters. Hypothesis testing helps to statistically verify claims on the demographic parameters, as well as identifies significant associations between population characteristics and heart diseases: aiding evidence-based strategies for public health and equitable outcomes.

Contents

1 About the Data

The Behavioral Risk Factor Surveillance System (BRFSS) is the U.S.' premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. It is a collaborative project between all the states in the United States and participating US territories and the Centers for Disease Control and Prevention (CDC). This dataset is the **2021 BRFSS Survey Data** [1] that contains health data about residents from 49 states in the US, the District of Columbia, Guam, Puerto Rico, and the US Virgin Islands.

Since the sample encompasses data from individuals across all 50 states of the US, it can be considered to be a good representative of the US population. With a sample size of 308,854, which is notably large, it will enhance the precision of population parameter estimates [2].

2 Data Wrangling

2.1 Transforming & Cleaning the Dataset

The original dataset comprises about 438,693 records and 304 unique features (or columns). The dataset was cleaned in the following procedure:

- The duplicate, redundant, and records with null values were dropped from the dataset. The new dataset comprises about 308,854 records.
- 19 features out of 304 were handpicked (based on domain knowledge) that relates to lifestyle factors of a person that can be contributed to being at risk with any form of Cardiovascular Diseases. [3]
- The units of measurement of fruit, vegetables, and fried potato consumption were converted to "average number of units consumed in a month." Similarly, the unit of alcohol consumption was converted to "average liters of alcohol consumed in a month".

Hereon, we might refer to each person of the dataset as 'patient' in a medical context.

2.2 The Transformed Dataset

The transformed dataset consists of 19 columns (features) in total: 7 numerical and 12 categorical.

Columns in the Dataset	
Column Name	**Explanation**
General_Health	In general, whether the health of the patient is poor, fair, good, very good, or excellent (categorical variable)
Checkup	How long it has been since the patient last visited a doctor for a routine checkup. The most common values are "within the past year" and "within the past 2 years"
Exercise	Whether the patient has participated in any kind of physical activities or exercises in the past month (categorical variable, Yes/No values)
Heart_Disease	Whether the patient has a coronary heart disease or has suffered a myocardial infarction (categorical variable, Yes/No values)
Skin_Cancer	Whether the patient has skin cancer (categorical variable, Yes/No values)
Other_Cancer	Whether the patient has any other type of cancer (categorical variable, Yes/No values)
Depression	Whether the patient has (has been) suffered (suffering) from depression of any kind (categorical variable, Yes/No values)
Diabetes	Whether the patient has diabetes, and if yes, of what type
Arthritis	Whether the patient has an Arthritis (categorical variable, Yes/No values)
Sex	Gender of the patient (categorical variable, Male/Female)
Age_Category	Age range of the patient (in years): the bins are 60-64, 65-69, etc.
Height_(cm)	Height of the patient in cm
Weight_(kg)	Weight of the patient in kg
BMI	Body Mass Index of the patient
Smoking_History	Whether the patient has a smoking history (categorical variable, Yes/No values)
Alcohol_Consumption	Average liters of alcohol consumed in a month (numerical variable)
Fruit_Consumption	Average number of fruits consumed in a month
Green_Vegetables	Average number of green vegetables consumed in a month
FriedPotato_Consumption	Average number of fried potatoes consumed in a month

3 Some Definitions and Theorems

Definition 1 (Likelihood Function): Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population having probability distribution (or mass) function $f(x;\theta)$, where θ is an unknown parameter. If $x_1, x_2, \ldots, x_n$ are the observed values of the random sample, then the likelihood function of the sample is

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Definition 2 (Maximum Likelihood Estimator): The maximum likelihood estimator (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta)$.

Theorem 1: Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ distribution, where $\mu \in (-\infty, \infty)$ and $\sigma^2 \in (0, \infty)$ are unknown. Then the MLE of μ is $\overline{X}$, and that of σ^2 is $\frac{n-1}{n}S^2$.

Definition 3 (Confidence Interval): A confidence interval is an interval which is expected to typically contain the parameter being estimated.

Theorem 2: Let $X_1, X_2, \ldots, X_n$ are normally distributed with unknown mean μ and variance σ^2, then a $(1-\alpha)100\%$ confidence interval for the population mean μ is:

$$\left(\overline{X} - t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right),\ \overline{X} + t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)\right)$$

Theorem 3: If $X_1, X_2, \ldots, X_n$ are normally distributed and a=$\chi^2_{1-\alpha/2,n-1}$ and $b = \chi^2_{\alpha/2,n-1}$, then a $(1-\alpha)100\%$ confidence interval for the population variance σ^2 is:

$$\left(\frac{(n-1)S^2}{b},\ \frac{(n-1)S^2}{a}\right)$$

and a $(1-\alpha)100\%$ confidence interval for the population standard deviation σ is:

$$\left(\frac{S\sqrt{n-1}}{\sqrt{b}},\ \frac{S\sqrt{n-1}}{\sqrt{a}}\right)$$

Theorem 4: For large random samples, a $(1-\alpha)100\%$ confidence interval for population proportion $\hat{p}$ is:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Theorem 5 (Two Sample Pooled t-interval): If $X_1, X_2, \ldots X_n \sim N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots Y_m \sim N(\mu_2, \sigma^2)$ are independent samples, then a $(1-\alpha)100\%$ confidence interval for the difference in population means $\mu_1 - \mu_2$ is:

$$(\overline{X} - \overline{Y}) \pm t_{\alpha/2,n+m-2}\ S_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

where S_p is the pooled standard deviation, given by $S_p = \sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}$

Theorem 6: If $X_1, X_2, \ldots X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \ldots Y_m \sim N(\mu_Y, \sigma_Y^2)$ are independent samples, then a $(1-\alpha)100\%$ confidence interval for the ratio of population variances $\frac{\sigma_X^2}{\sigma_Y^2}$ is:

$$\left(\frac{1}{F_{\alpha/2,\, n-1,\, m-1}} \frac{S_X^2}{S_Y^2},\ F_{\alpha/2,\, m-1,\, n-1} \frac{S_X^2}{S_Y^2} \right)$$

Theorem 7: For large random samples, an approximate $(1-\alpha)100\%$ confidence interval for the difference in two population proportions $p_1 p_2$ is:

$$(\hat{p_1} - \hat{p_2}) \pm z_{\alpha/2} \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

4 Analysis of Demographic Data of the US Residents

Demographic data of people such as age, gender, background, income, height, weight, BMI, etc. plays a crucial role in many fields. They provide valuable insights for entities like healthcare providers, health insurance companies, fashion industries, sports gear manufacturers, and so on.

In this section, we shall carry out inferential statistical analysis. We aim to draw some conclusion about the heights and weights of the underlying population through the sample (our dataset). In other words, we will estimate the population parameters such as the mean population height, standard deviation of the population height, and so on. Note that the population can be considered as the people living in US: the 49 states in the US, the District of Columbia, Guam, Puerto Rico, and the US Virgin Islands (from where the sample has been collected).

4.1 Age Range and Gender

Considering the age range of residents is crucial for studying health trends. Elderly people are more prone to certain diseases than the young. Figure 1 displays how many people are in different age groups in the dataset. Residents in the dataset are between 18 and 85 years old, with a median age of 45 years. This information helps healthcare providers and insurance companies design age-specific health plans and policies. Dividing the ages from 18 to 80+ into 13 parts is a useful way to organize the data for analysis.

It is important to consider gender as a factor in the healthcare system. Men, women, and others may have different health needs and risks. Our dataset consists of two genders viz. Male and Female, and the proportion is given in Figure 2.

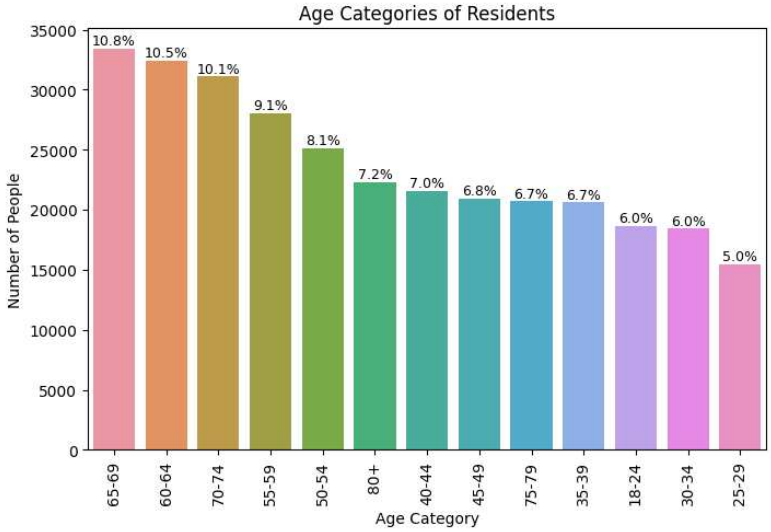


Figure 1: Age Categories of Residents

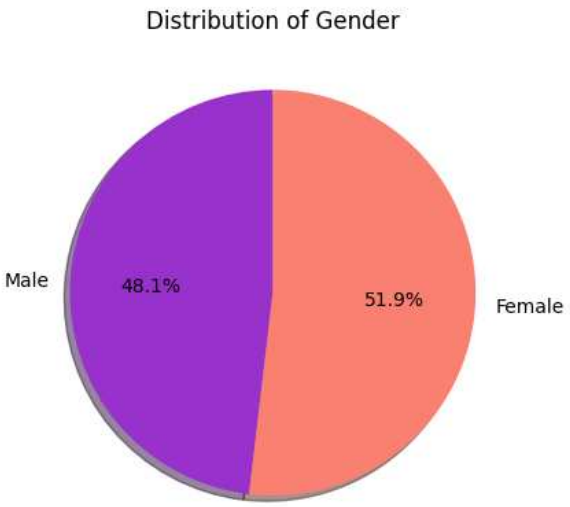


Figure 2: Distribution of Gender

4.2 Some Important Values

Note that the total sample size is $n = 308854$. The t-values are close to z-values for large values of n. All critical values were calculated using functions in the *stats* module of *SciPy*.
Value of $t_{\alpha/2,n-1}$, for 95% confidence interval ($\alpha = 0.05$) is $t_{\alpha/2,n-1} = 1.959$
Value of $t_{\alpha/2,n-1}$, for 99% confidence interval ($\alpha = 0.01$) is $t_{\alpha/2,n-1} = 2.5758$

For 95% confidence interval ($\alpha = 0.05$):
Value of $a = \chi^2_{1-\alpha/2,n-1} = 307314.4757$
Value of $b = \chi^2_{\alpha/2,n-1} = 310395.3129$

For 99% confidence interval ($\alpha = 0.01$):

Value of $a = \chi^2_{1-\alpha/2,n-1} = 306832.30237$
Value of $b = \chi^2_{\alpha/2,n-1} = 310881.21082$

4.3 Distribution of Height: Confidence Interval Estimation for Mean and Variance of Population Height

4.3.1 Distribution of Height of all Residents

Figure 3 shows that height of the US residents is approximately normally distributed with mean 170.62 cm and variance 113.59 cm^2.

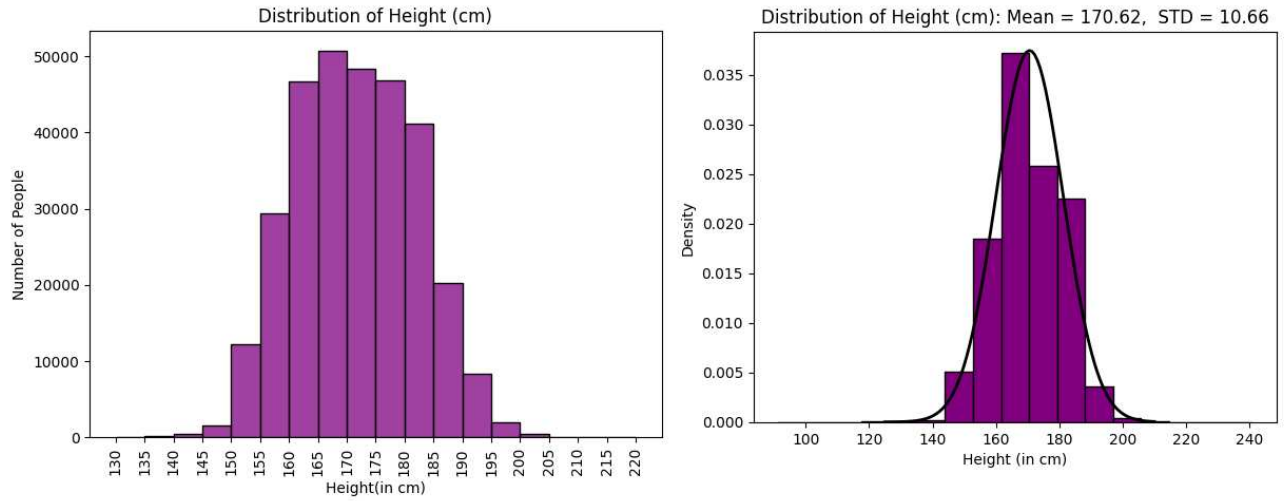


Figure 3: Distribution of Height of All Residents

Point Estimate for Mean Population Height of all Residents
As the height is normally distributed, we have by Theorem 1 that the maximum likelihood estimate for the population mean is the sample mean. That is,

$$\hat{\mu}_H = \overline{X}_H = 170.615\ cm$$

Hence, based on the dataset, we can say that the average height of the US residents is approximately 170.615 cm.

Confidence Interval for Mean Population Height of of all Residents
By Theorem 2, the confidence interval for mean height of the population is

$$\left(\overline{x}_H - t_{\alpha/2,n-1}\left(\frac{S_H}{\sqrt{n}}\right),\ \overline{x}_H + t_{\alpha/2,n-1}\left(\frac{S_H}{\sqrt{n}}\right)\right)$$

where the notations have their usual meaning.

Here, $\overline{x}_H = 170.615, S_H = 10.658, n = 308854$.

Values of $t_{\alpha/2,n-1}$ are mentioned in section 4.2.

The confidence intervals for the population mean of the height of all residents:
95% confidence interval: **(170.578, 170.653)**
99% confidence interval: **(170.566, 170.665)**

Hence, we can be 95% confident that the mean height of the US population lies between 170.578 *cm* and 170.653 *cm*. The interval length is considerably small, resulting in a more precise confidence interval. This is due to the large size of the dataset (large value of n).

Point Estimate for Population Standard Deviation and Variance of Height of all Residents
By Theorem 1, the maximum likelihood estimate for the population variance is $\frac{(n-1)S_H^2}{n}$.
Here, $S_H = 10.658$, n = 308854,

$$\hat{\sigma_H^2} = \frac{(n-1)S_H^2}{n} = 113.593\ cm^2$$

Hence, point estimate for population variance of height of all residents: 113.593 cm^2.
And, point estimate for population standard deviation of the same: 10.658 *cm*.

Confidence Interval for Population Variance and Standard Deviation of Height of all Residents

By Theorem 3,
Confidence Interval for the population variance is $\left(\frac{(n-1)S_H^2}{b},\ \frac{(n-1)S_H^2}{a}\right)$
where $a = \chi^2_{(1-\alpha/2),n-1}$, $b = \chi^2_{\alpha/2,n-1}$, whose values are mentioned in Section 4.2.

The confidence intervals for the population variance of height of all residents:
95% confidence interval: **(113.029, 114.162)**
99% confidence interval: **(112.852, 114.341)**

The confidence intervals for the population standard deviation of height of all residents:
95% confidence interval: **(10.631, 10.685)**
99% confidence interval: **(10.623, 10.693)**

4.3.2 Distribution of Height of Male Residents

Figure 4 shows that the height of the male population is approximately normally distributed with mean 178.34 *cm* and variance 60.996 cm^2.

Confidence Interval for Mean Height of Male Population
By Theorem 2, the confidence interval for the mean height of male population is:

$$\left(\overline{x}_{H_M} - t_{\alpha/2,n-1}\left(\frac{S_{H_M}}{\sqrt{n}}\right),\ \overline{x}_{H_M} + t_{\alpha/2,n-1}\left(\frac{S_{H_M}}{\sqrt{n}}\right)\right)$$

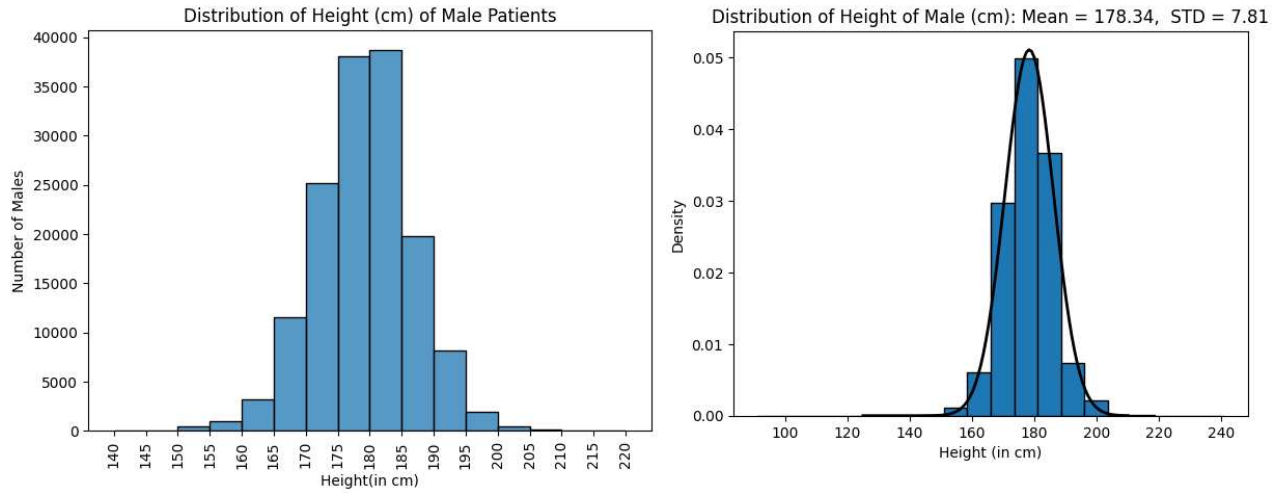


Figure 4: Distribution of Height of Male Residents

where the notations have their usual meaning.
Here, sample mean: $\overline{x}_{H_M} = 178.34,\ S_{H_M} = 7.808,\ n = 148658.$

Confidence intervals for (population) mean height of male residents:
95% confidence interval: **(178.3, 178.379)**
99% confidence interval: **(178.288, 178.392)**

Hence, we can be 95% confident that the mean height of the US male population lies between 178.3 cm and 178.379 cm.

Confidence Interval for Population Variance and Standard Deviation of Male Population Height

By Theorem 3, confidence interval for the population variance is $(\frac{(n-1)S^2_{H_M}}{b}, \frac{(n-1)S^2_{H_M}}{a})$, where $a = \chi^2_{(1-\alpha/2),n-1}$, $b = \chi^2_{\alpha/2,n-1}$, and S_{H_M} denotes the sample STD of male heights.

Point estimate for population variance of height of male residents: 60.965 cm^2.
The confidence intervals for the population variance of the height of male residents:
95% confidence interval: **(60.53, 61.407)**
99% confidence interval: **(60.394, 61.546)**

Point estimate for population standard deviation of male height: 7.808 cm. Confidence intervals for the same are:
95% confidence interval: **(7.78, 7.836)**
99% confidence interval: **(7.771, 7.845)**

4.3.3 Distribution of Height of Female Residents

Figure 5 shows that height of the female residents is approximately normally distributed with mean 163.45 cm and variance 55.65 cm^2.

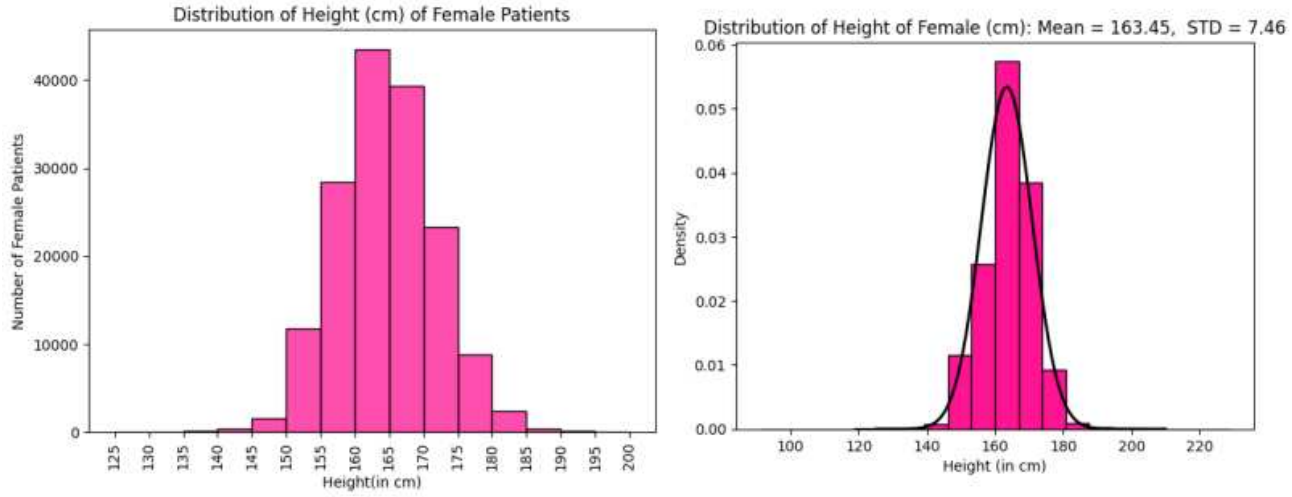


Figure 5: Distribution of Height of Female Residents

Point Estimate for Mean Height of Female Residents
We have

$$\hat{\mu_{H_F}} = \overline{X_{H_F}} = 163.447 \; cm$$

Hence, based on the dataset, we can say that the average height of the female residents is approximately 163.447 cm.

Confidence Interval for Mean Height of Female Population
By Theorem 2, using $\overline{x} = 163.447, S = 7.462$, and $n = 160196$,

The confidence intervals for mean height of the female population:
95% confidence interval: **(163.411, 163.484)**
99% confidence interval: **(163.399, 163.495)**

Hence, we can be 95% confident that the mean height of the female population lies between 163.411 cm and 163.484 cm.

Confidence Interval for Population Variance and Standard Deviation of Female Height
By Theorem 3,
Confidence intervals for population variance of height of female residents:
95% confidence interval: **(55.295, 56.066)**
99% confidence interval: **(55.175, 56.189)**

Confidence intervals for population standard deviation of height of female residents:
95% confidence interval: **(7.436, 7.488)**
99% confidence interval: **(7.428, 7.496)**

4.3.4 Confidence Intervals for Difference in Population Mean of Male and Female Heights

Figure 4 shows that the height of the male residents is approximately normally distributed with average height($\overline{X}_M$) = 178.34 cm and STD = 7.808 cm.
Figure 5 shows that the height of the female residents is approximately normally distributed with average height($\overline{X}_F$) = 163.45 cm and STD = 7.462 cm.
Number of male residents (n) = 148658 and number of female residents (m) = 160196.
The ratio of standard deviations of height (male/female): 1.046

As the ratio of sample standard deviation is 1.05, which is less than 2, we use the pooled t-interval for the confidence interval of difference of means.

Using theorem 5, confidence intervals for difference in population mean of male heights and female heights:
The 95% Confidence Interval: **(14.839, 14.946)**
The 99% Confidence Interval: **(14.822, 14.963)**

4.4 Distribution of Weight: Confidence Interval Estimation for Mean and Variance of Population Weight

4.4.1 Distribution of Weight of all Residents

Figure 6 shows that weight is approximately normally distributed with mean 83.59 kg and standard deviation 21.34 kg.

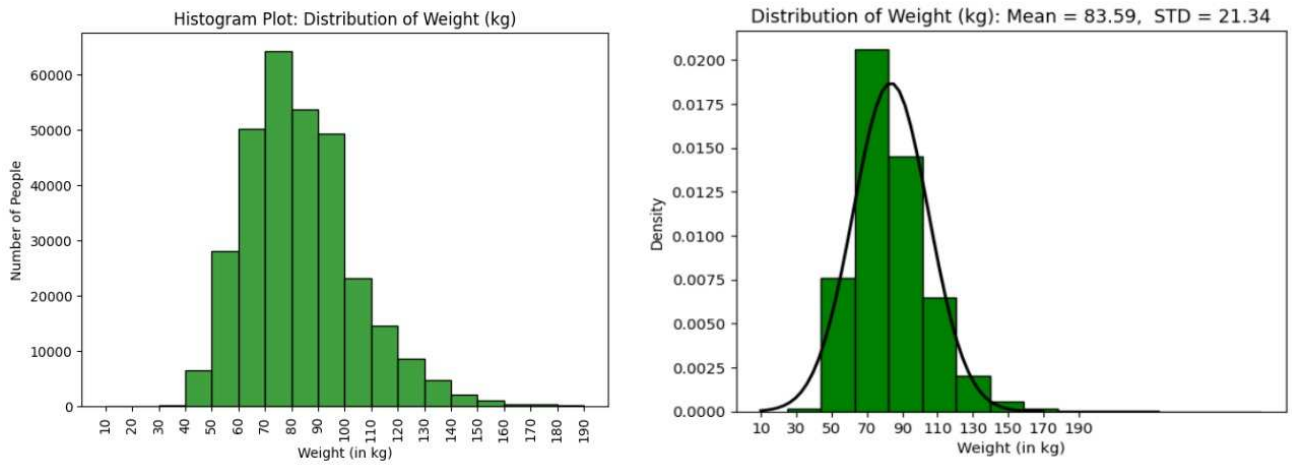


Figure 6: Distribution of Weight of all Residents

Point Estimate for the Mean of Weight of Population:

Point estimate for the mean weight of the US population is

$$\hat{\mu}_w = \overline{X}_w = 83.59\ kg$$

Confidence Interval for Mean Weight of Population:
As in the above sections, we use theorem 2 for computing the confidence interval.
Here, $\overline{x}_w = 83.59, S_w = 21.34, n = 308854$.

The 95% confidence interval for the mean weight of the population: **(83.513, 83.664)**
The 99% confidence interval for the mean weight of the population: **(83.491, 83.688)**

Hence, we can be 95% confident that the mean weight of the population lies between 83.513 kg and 83.664 kg.

Confidence Interval for Variance and Standard Deviation of Population Weight:

By Theorem 3,
Confidence Interval for the population variance is $(\frac{(n-1)S_w^2}{b}, \frac{(n-1)S_w^2}{a})$

where $a = \chi^2_{(1-\alpha/2),n-1}$, $b = \chi^2_{\alpha/2,n-1}$. The values of a and b are mentioned in section 4.2

The point estimate for the population variance of weight: 455.39
The confidence intervals for the population variance of weight are:
95% confidence interval: **(453.268, 457.812)**
99% confidence interval: **(452.559, 458.531)**

The point estimate for the population STD of weight is 21.34.
The confidence intervals for the population variance of weight are:
95% confidence interval: **(21.29, 21.397)**
99% confidence interval: **(21.273, 21.413)**

4.4.2 Distribution of Weight of Male Residents

Figure 7 shows that the weight of male residents is approximately normally distributed with mean 91.43 kg and STD 20.30 kg^2.

Point Estimate for Mean Weight of Male Residents:

$$\hat{\mu}_{W_M} = \overline{X}_{W_M} = 91.43\ kg$$

Hence, the average weight of the male residents is approximately 91.43 kg.

Confidence Interval for Mean Weight of Male Residents:
Using theorem 2 with $\overline{x}_{W_M} = 91.43,\ S_{W_M} = 20.30,\ n = 148658$ (number of male residents),
The confidence intervals for (population) mean weight of male residents:
95% confidence interval: **(91.329, 91.535)**
99% confidence interval: **(91.297, 91.568)**

Hence, we can be 95% confident that the mean weight of male residents in the US lies between 91.329 kg to 91.535 kg.

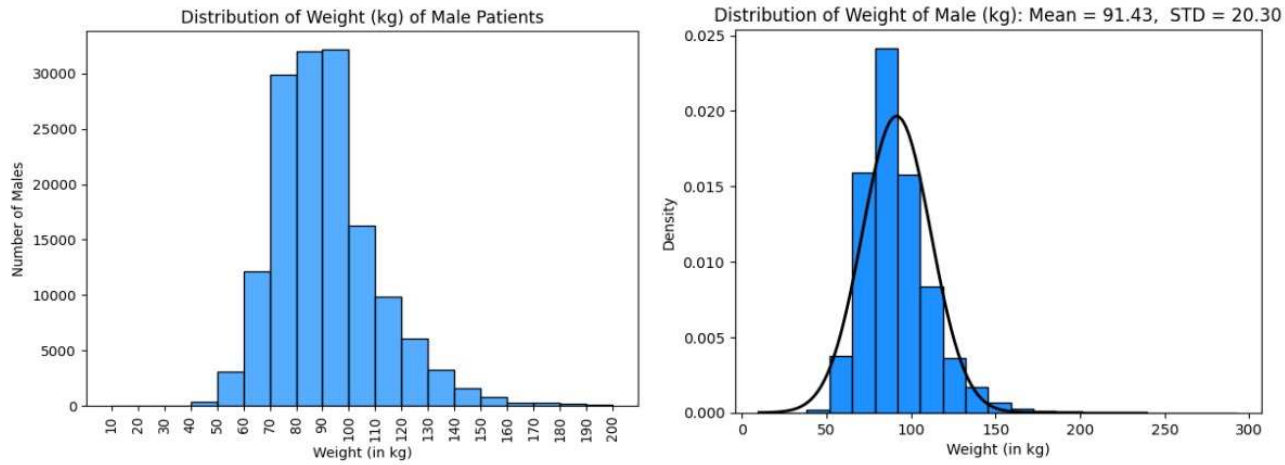


Figure 7: Distribution of Weight of Male Residents

Confidence Interval for Population Variance and Standard Deviation of Weight of Male Residents:
By theorem 3, confidence interval for the population variance is $(\frac{(n-1)S^2_{W_M}}{b}, \frac{(n-1)S^2_{W_M}}{a})$ where $a = \chi^2_{(1-\alpha/2),n-1}$, $b = \chi^2_{\alpha/2,n-1}$, and S_{W_M} denotes the sample standard deviation of male weight.

The confidence intervals for population variance of weight of male residents:
95% confidence interval: **(408.97, 414.896)**
99% confidence interval: **(408.08, 415.834)**

The confidence intervals for population standard deviation of weight of male residents:
95% confidence interval: **(20.223, 20.369)**
99% confidence interval: **(20.201, 20.392)**

4.4.3 Distribution of Weight of Female Residents

Figure 8 shows that the weight of female residents is approximately normally distributed with mean 76.31 kg and STD 19.64 kg.

Point Estimate for Mean Weight of Female Residents

$$\hat{\mu}_{W_F} = \overline{X}_{W_F} = 76.31 \ kg$$

Hence, based on the dataset, we can say that the average weight of the female residents in the US is approximately 76.31 kg.

Confidence Interval for Mean Weight of Female Population

Using theorem 2 with the values $\overline{x} = 76.31, S_{W_F} = 19.64$, and $n = 160196$,
The confidence intervals for the mean weight of female population:

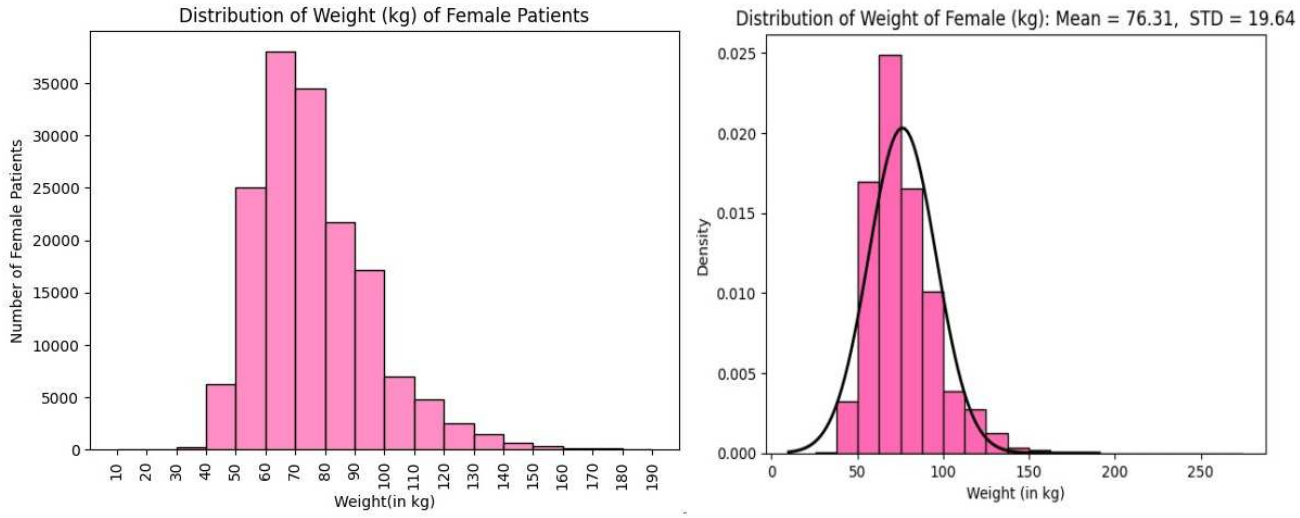


Figure 8: Distribution of Weight of Female Residents

95% confidence interval: **(76.214, 76.406)**
The 99% confidence interval: **(76.184, 76.436)**

Hence, we can be 95% confident that the mean weight of the female population lies between 76.214 kg and 76.406 kg.

Confidence Interval for Population Variance and Standard Deviation of Weight of Female Residents
By Theorem 3,
Confidence Interval for the population variance is $(\frac{(n-1)S^2_{W_F}}{b}, \frac{(n-1)S^2_{W_F}}{a})$

Point estimate for the population variance of weight of female residents: 385.918
The 95% confidence interval: **(383.262, 388.607)**
The 99% confidence interval: **(382.431, 389.456)**

Point estimate for the population standard deviation of weight of female residents: 19.645
The 95% confidence interval: **(19.577, 19.713)**
The 99% confidence interval: **(19.556, 19.735)**

4.5 Confidence Interval for Ratio of Variances: BMI of Women to Men

Body mass index (BMI) is a measure of body fat based on height and weight that applies to an adult human. It is an important measure to determine whether an individual is healthy, underweight, overweight, or obese. In this section, we calculate confidence intervals for the ratio of variance of BMI of female residents to that of male residents.

(Sample) Variance of BMI of women (S^2_X): 49.889

(Sample) Variance of BMI of men (S_Y^2): 34.611
Number of women (n) = 160196
Number of men (m) = 148658
F values: $F_{0.025,\, n-1,\, m-1} = 1.010033,\ F_{0.025,\, m-1,\, n-1} = 1.010031,\ F_{0.005,\, n-1,\, m-1} = 1.013206,$
$F_{0.005,\, m-1,\, n-1} = 1.013203$

Using theorem 6, the required confidence intervals are:
95% confidence interval: **(1.427, 1.456)**
99% confidence interval: **(1.423, 1.46)**

A ratio of variances less than 4 lets us conclude that the variability in BMI of men and women in US are not significantly different. However, there is slightly more variability in the BMI of women as compared to men. There could be several reasons for this small difference - it could be due to biological factors, such as hormonal differences between men and women, or it could be influenced by social and cultural factors related to body image and societal expectations regarding weight and appearance. It is important to note that this conclusion is based solely on the statistical analysis of variance of BMI.

5 Determination of Sample Size

Sample size plays an important role in statistical analysis, since a larger random sample better represents the population and increases the precision of population estimates. Further, if we have a huge dataset, we can choose a random sample out of it of a given size that fits the purpose, based on our goal.

Q1. **How many female residents should we survey to be 95% confident that their estimated average height will be within 3 cm of the mean height, given that the previous data indicates a normal distribution of heights ranging from 135 cm to 195 cm?**
Ans.
There are two methods to get the sample size.

Crude Method:
In this method, we simply replace the t-value that depends on n with a z-value that (does not because as n increases, the t-distribution approaches the standard normal distribution). Thus,

$$n \approx \frac{(z_{\alpha/2})^2 S^2}{E^2}$$

Here, E is the margin of error. Also, as the **empirical rule** states that approximately 95% of the measurements lie in the interval $\mu \pm 2\sigma$, we estimate σ by $S = \frac{Range}{4}$, where $Range = (\mu + 2\sigma) - (\mu - 2\sigma) = 4\sigma$ is the length of the interval.
Here, $\alpha = 1 - 0.95 = 0.05,\ E = 3,\ S = \frac{Range}{4} = \frac{195-135}{4} = 15$
So, $z_{\alpha/2} = 1.96$,

$$n = \frac{(z_{\alpha/2})^2 S^2}{E^2}$$

$$n = \frac{(1.96)^2 (15)^2}{3^2}$$

i.e., $n = 96.04$.
Through the crude method, we have that a sample size of approximately 96 or larger is recommended to obtain an estimated average female height that we are 95% confident is within 3 cm of the true mean female height.

Iterative Method: A more accurate method to estimate the sample size is to iteratively evaluate the formula since the t value also depends on n. We start with an initial guess for n and plugin the formula

$$n = \frac{(t_{\alpha/2,\, n1})^2 S^2}{E^2}$$

and iteratively solve for n.

The following Python code snippet calculates the sample size using the iterative method:

```
n = 30     # Initial guess
E = 3      # Margin of Error
S = (195-135)/4         # Sample STD estimate = Range/4
for i in range(8):
    t = stats.t.ppf(q=1-(0.05/2), df=n-1)
    new_n = ((t**2)*(S**2))/(E**2)
    print("Iteration {}: \t Assumed n = {} \t Calculated n = {}".format((i+1),
        round(n,3), round(new_n,3)))
    n = new_n

```

The following is the output:

```
Iteration 1:    Assumed n = 30          Calculated n = 104.574
Iteration 2:    Assumed n = 104.574     Calculated n = 98.32
Iteration 3:    Assumed n = 98.32       Calculated n = 98.47
Iteration 4:    Assumed n = 98.47       Calculated n = 98.466
Iteration 5:    Assumed n = 98.466      Calculated n = 98.466
Iteration 6:    Assumed n = 98.466      Calculated n = 98.466
Iteration 7:    Assumed n = 98.466      Calculated n = 98.466
Iteration 8:    Assumed n = 98.466      Calculated n = 98.466
```

In the first iteration, we assumed $n = 30$. Proceeding, we see that the value of n stabilizes to 98.466 by the 8^{th} iteration. Hence, a sample size of approximately 98 or larger is recommended to obtain an estimated average female height that we are 95% confident is within 3 cm of the true mean female height.

Q2. **We wish to determine the proportion of US residents who are involved in exercise or any kind of physical activities. How many residents should we survey in order to be 95% certain that our estimate will be correct to within 2%?**
Ans.

We know that the error term in the confidence interval for population proportion is given by $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$, where p is the sample proportion, α is the level of significance, and n is the number of samples. Here, we must have

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq 0.02$$

We know, $p(1-p) \leq 0.25$ (the maximum value is attained at $p = 0.5$ and this can be determined through Calculus). Therefore, we have

$$1.96\sqrt{1/4n} \leq 0.02$$

Solving which, we get $n = 2401.25$. Thus, a sample size of at least 2402 is required to achieve the desired goal.

6 Confidence Interval Estimation for Population Proportion & Difference of Proportions

By Theorem 4, as we have a large number of samples, a $100(1-\alpha)\%$ confidence interval for population proportion $\hat{p}$ is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

We now answer some questions regarding the US population through the dataset.

Q1. **What percent of the population has a smoking history?**

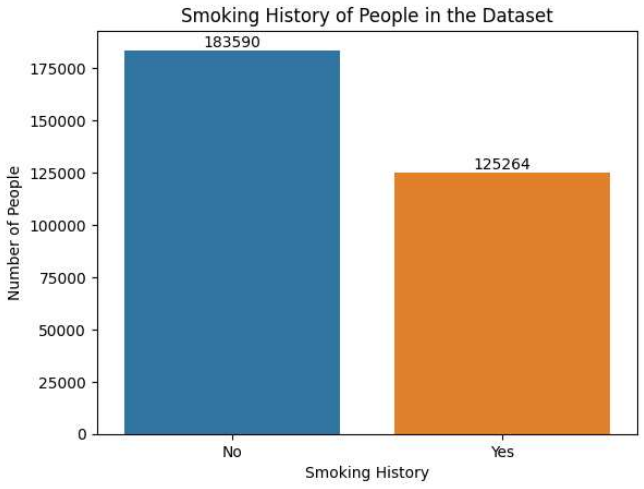


Figure 9: Smoking History of People in the Dataset

Figure 8 shows that out of 308854 people, 125264 have a smoking history in the dataset.

Here, $n = 308854$, $\hat{p} = \frac{125264}{308854} \approx 0.40558$, $\quad z_{\alpha/2} = 1.96$, where $\alpha = 1 - 0.95 = 0.05$
Thus,

$$\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.40731$$

and

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.40385$$

Conclusion: By theorem 4, we can be 95% confident that between **40.385% to 40.731%** of the US population smokes, or have a smoking history.

***Q2.* What percent of the population is diabetic?**

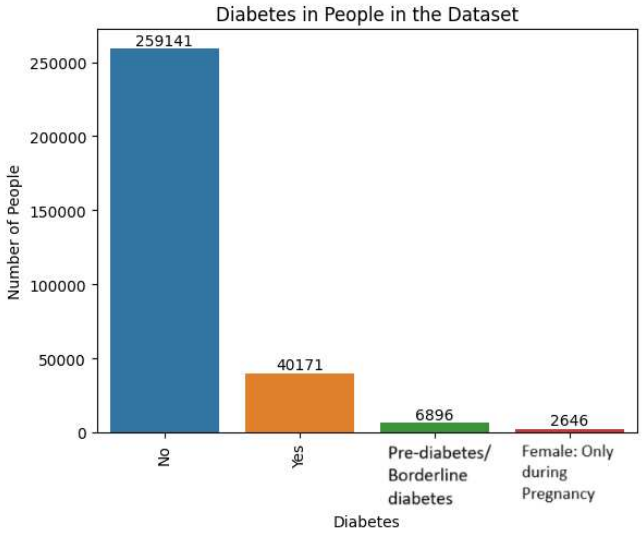


Figure 10: Diabetes in People in the Dataset

(For the sake of simplicity, let us treat pre-diabetic patients as diabetic, and women who reported diabetes during pregnancy only to be non-diabetic.)
From Figure 9, we can see that out of 308854 residents, 47067 are diabetic.

Here, we have $n = 308854$, $\hat{p} = \frac{47067}{308854} \approx 0.1524$, $\ z_{\alpha/2} = 1.96$, where $\alpha = 1 - 0.95 = 0.05$
Thus,

$$\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.15366$$

and

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.15112$$

Conclusion: By theorem 4, we can be 95% confident that between **15.112% to 15.366%** of

the US population is diabetic.

***Q3.* What percent of the population has depression?**

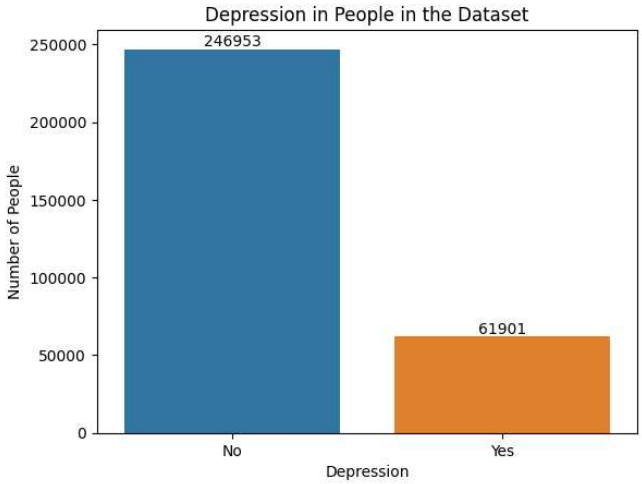


Figure 11: Depression in People in the Dataset

One can conclude from Figure 10 that out of 308854 people, 61901 are suffering from depression.
Proceeding as in the above questions, by theorem 4, we can be 95% confident that between **19.901% to 20.183%** of the population is suffering from depression.

***Q4.* What is the difference in proportion between male and female smokers?**

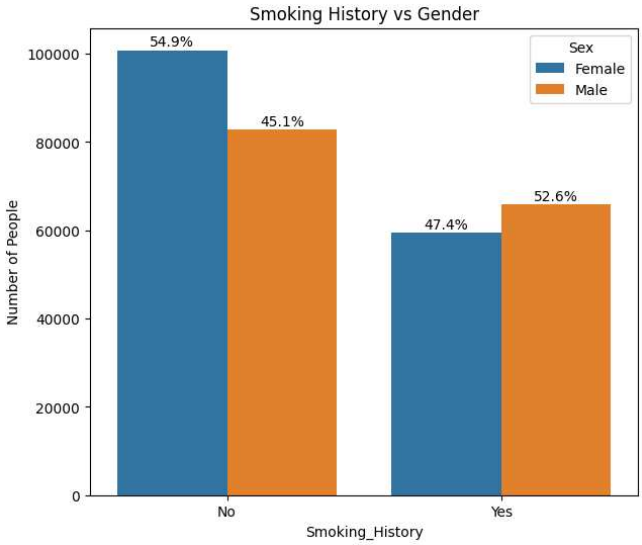


Figure 12: Smoking History vs Gender

From Figure 12, we can see that 52.6% of the smokers are men and the rest are women. In the

dataset, 65854 out of 148658 men are smokers, and so are 59410 out of 160196 women.
Proportion of men who smoke ($\hat{p_1}$) = 0.443
Proportion of women who smoke ($\hat{p_2}$) = 0.371

By theorem 7, the confidence intervals for difference in proportion of male and female smokers are as follows:
95% confidence interval: **(0.069, 0.076)**
99% confidence interval: **(0.068, 0.077)**

Thus, we can be 95% confident that 6.9% to 7.6% more men in the US are smokers as compared to women.

***Q5.* What is the difference in proportion of depressed women and men?**

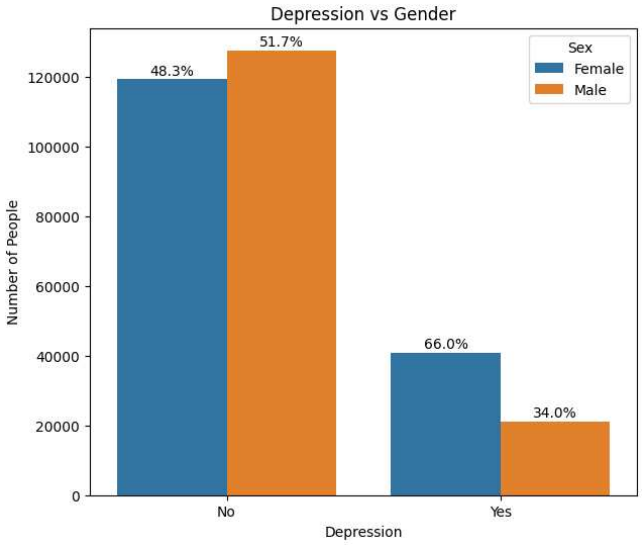


Figure 13: Depression vs Gender

From Figure 13, it is evident that 66% of the people suffering from depression in the dataset are women, and the rest are men.
Number of depressed men: 21056
Number of depressed women: 40845
(Sample) Proportion of depressed women (p_1): 0.255
(Sample) Proportion of depressed men (p_2): 0.142
Using theorem 7, the required confidence intervals for difference in proportion are:
95% confidence interval: **(0.111, 0.116)**
99% confidence interval: **(0.11, 0.117)**

We can be 95% confident that 11.1% to 11.6% more women in the US suffer from depression, as compared to men. As 0 does not lie in the interval, there is a statistically significant difference in the proportion of women suffering from depression compared to men, with women having a higher prevalence.

7 Hypothesis Testing on Demographic Data & Population Proportions

We divide the section of hypothesis testing into two. The current section deals with hypothesis testing on normally distributed numerical features like height of the US population. In section 9, we test several hypotheses related to the connection of factors like diabetes, arthritis, and age with heart disease. We use the original dataset for this section, and undersample it (section 8) for hypothesis testing in section 9.

7.1 Is the mean height of the US male population still 175.3 cm as of 2021?

According to Wikipedia, the average height of the male population in the US is 175.3 cm, based on 5,232 samples in the years 2015-18 [4]. Based on the sample we have, we want to check whether the average male height is still $\mu_0 = 175.3$ cm at 5% level of significance. Denoting H_0 as the null hypothesis, H_a as the alternative hypothesis, and μ as the average height of the male population in the US (as of 2021), we formulate our hypothesis as follows:

H_0: $\mu = 175.3$
H_a: $\mu \neq 175.3$

Level of Significance (α): 0.05
Number of samples: $n = 148658$
Mean height of the male residents in our sample: $\overline{X} = 178.34$
Sample standard deviation: $S = 7.808$
As the population variance is unknown, the test statistic is:

$$t^* = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \approx 150.1$$

Again, we get such a large value of t^* due to a considerably large number of samples. We conclude the hypothesis test through the following methods.

7.1.1 Rejection Region Approach

As this is a two-tailed test, the rejection region is $|t^*| \geq t_{\alpha/2,n-1}$, which means that we will reject the null hypothesis if $t^* \geq 1.96$ or $t^* \leq -1.96$.
We have critical value $t_{\alpha/2,n-1} \approx 1.959$ for $\alpha = 0.05$ and $n = 148658$.
Clearly, $t^* = 150.1 > 1.96$, so we will reject the null hypothesis H_0.

7.1.2 p-Value Approach

p-Value for two-tailed test is $p = 2P(t \geq |t^*|)$. We reject H_0 if $p < \alpha$.

We use the t.sf(t_test, df) function in the scipy.stats library to calculate the value of $2P(t \geq |t^*|)$.
To this end, we get that $p = 2P(t \geq |t^*|) \approx 0$. Clearly, $p < 0.05 = \alpha$, so we reject the null hypothesis H_0.

7.1.3 Conclusion

From the above two sections, we have sufficient statistical evidence to conclude that the average height of the male population in the US is not 175.3 cm, as of 2021. In fact, this result is consistent with the confidence interval estimation for mean male population height in section 4.3.2. We can say, with 95% confidence, that the mean population height of the male residents must lie in the range (178.3 cm, 178.379 cm). Clearly, 175.3 cm does not lie in this range, so it cannot be 175.3 cm.

7.2 Is the average female height in US greater than 161.3 cm?

The same Wikipedia article [4] states that the mean female height in the US (during 2015-18) is 161.3 cm. We want to test our claim that the value has increased as of 2021. Let μ denote the average female height in the US, then our hypothesis is:

$H_0 : \mu$ has not increased since 2018 ($\mu \leq 161.3$)
$H_a : \mu$ has increased since 2018 ($\mu > 161.3$)

Level of Significance (α): 0.05
Number of samples: $n = 160196$
Mean height of the female residents in our sample: $\overline{X} = 163.447$
Sample standard deviation: $S = 7.462$
Again, the population variance is unknown, so the test statistic is:

$$t^* = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \approx 115.2$$

7.2.1 Rejection Region Approach

As this is a right-tailed test, the rejection region is $t^* \geq t_{\alpha,n-1}$. So, we will reject the null hypothesis if $t^* \geq 1.6449$, since $t_{\alpha,n-1} \approx 1.6449$ for $\alpha = 0.05$ and $n = 160196$.
Clearly, $t^* = 115.2 > 1.6449$, so we will reject the null hypothesis H_0.

7.2.2 p-Value Approach

p-Value for right-tailed test is $p = P(t \geq t^*)$. We reject H_0 if $p < \alpha$.
We get that $p = P(t \geq t^*) \approx 0$. Clearly, $p < 0.05 = \alpha$, so we reject the null hypothesis H_0.

7.2.3 Conclusion

We have sufficient statistical evidence to conclude that the average female height in the US has indeed increased from 161.3 cm since 2018.

7.3 Does more than 40% of the US population smoke?

Based on our sample, we want to test the claim that more than 40% of the US population smokes or have a smoking history. Let us denote H_0 as the null hypothesis, H_a as the alternative hypothesis, and $\hat{p}$ as the (sample) proportion of the residents who smoke in the dataset.

Our formulated hypothesis is:

$H_0 : p \leq 0.4$ (less than or equal to 40% of the population smokes)
$H_a : p > 0.4$ (more than 40% of the population smokes)

Level of significance (α): 0.05
Number of samples: $n = 308854$
Number of smokers in the sample = 125264
Proportion of smokers in the dataset: $\hat{p} \approx 0.405577$
$p_0 = .040$

The test statistic is:

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx 6.32633$$

7.3.1 Rejection Region Approach

As this is a right-tailed test, the rejection region is $z^* > z_\alpha$
So, we reject the null hypothesis if $z^* > 1.64$ (since $z_\alpha = 1.64$ when $\alpha = 0.05$). It is clear that $z^* = 6.32633 > 1.64$, thus we reject null hypothesis H_0.

7.3.2 p-Value Approach

The p-value for the right-tailed test is $p = P(Z > z^*) = 1.25e - 10$, which is clearly $< \alpha = 0.05$. Hence, we reject the null hypothesis.

7.3.3 Conclusion

At 5% level of significance, we have sufficient statistical evidence to claim that more than 40% of the US population are smokers or have a smoking history.

7.4 Is the variance in BMI of the US population lesser than 40 kg/m^2?

Variability in BMI of a population gives us an idea about how different the body sizes and shapes of the people can be. Let σ^2 denote the population variance of BMI. To test the given hypothesis at 5% level of significance, we formulate it as follows:
H_0: $\sigma^2 \geq 40$
H_a: $\sigma^2 < 40$

Level of significance (α): 0.05
Number of samples: $n = 308854$
Sample variance in BMI: $S^2 = 42.54$, and $\sigma_0^2 = 40$.
This is a left-tailed test, where the test statistic is given by:

$$\chi^2_{test} = \frac{(n-1)S^2}{\sigma_0^2} = 328469.49$$

7.4.1 Rejection Region Approach

The rejection region is $\chi^2_{test} < \chi^2_L$, where χ^2_L is the lower tail value for $1-\alpha$ and $n-1$ degrees of freedom. We have $\chi^2_L = \chi^2_{1-\alpha,\, n-1} = 307561.37$ for $\alpha = 0.05$.

Hence, $\chi^2_{test} > \chi^2_L$, so we fail to reject the null hypothesis H_0.

7.4.2 p-Value Approach

The p-value is given by $p = P(\chi^2 \leq \chi^2_{test})$. We get that $p \approx 1$. As $p > \alpha$, we fail to reject the null hypothesis.

7.4.3 Conclusion

We do not have enough evidence to claim that the US population's BMI variability is less than 40 kg/m^2. A greater variance in BMI indicates that the individuals have a wide range of body sizes and shapes. This could include both underweight and severely obese individuals.

7.5 Are men in the US more than 14.8 cm (5.8 inches) taller than women on average?

Let μ_{men}, μ_{women} denote the mean population height of men and women respectively. The null and alternative hypotheses are defined as below:

H_0 : Men are not more than 14.8 cm taller than women on average ($\mu_{men} - \mu_{women} \leq 14.8$)
H_a : Men are more than 14.8 cm taller than women on average ($\mu_{men} - \mu_{women} > 14.8$)
So, this is a right-tailed test.
Average height of men in the US ($\overline{X_1}$)= 178.34 cm
Average height of women in the US ($\overline{X_2}$)= 163.447 cm
(Sample) STD of heights of men (S_1)= 7.808 cm
(Sample) STD of heights of women (S_2) = 7.462 cm
Number of male residents (n_1) = 148658 and number of female residents (n_2) = 160196
The ratio of standard deviations of height (male/female)= 1.046
We may assume that the population variances of heights of men and women are nearly equal (since the ratio of standard deviations is less than 2). Hence, we use the pooled standard deviation.

The pooled standard deviation is:

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = 7.63$$

The test statistic is:

$$t^* = \frac{(\overline{X_1} - \overline{X_2}) - 14.8}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = 3.368$$

7.5.1 Rejection Region Approach

The rejection region is $t^* \geq t_{\alpha,\, n_1+n_2-2}$. Now, $t_{\alpha,\, n_1+n_2-2} = t_{0.05,\, n_1+n_2-2} = 1.645$
So we will reject the null hypothesis, since $t^* \approx 3.368 \geq 1.645$

7.5.2 p-Value Approach

p-Value for the right-tailed test is $p = P(t \geq t^*) = 0.000379$. Clearly, it is less than $\alpha = 0.005$. Thus, we reject the null hypothesis.

7.5.3 Conclusion

At 5% significance level, we have enough evidence to claim that men in the US are more than 14.8 cm taller than women on average. The results are purely based on statistical inferences and do not imply inherent superiority or inferiority of either gender.

7.6 Do men have more variability in weights than women?

Let σ_m^2 and σ_w^2 denote the population variance of weight of men and women respectively.

H_0: $\sigma_m^2 \leq \sigma_w^2$
H_a: $\sigma_m^2 > \sigma_w^2$

Level of significance (α):0.05
Average weight of male residents: 91.432 kg
Average weight of female residents: 76.31 kg
(Sample) STD of weight of male residents (S_m): 20.296 kg
(Sample) STD of weight of female residents (S_w): 19.645 kg

The test statistic is given by:

$$F_{test} = \frac{S_m^2}{S_w^2} = 1.0674$$

7.6.1 Rejection Region Approach

As this is a right-tailed test, for level α with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$, we reject H_0 if $F_{test} \geq F_{\alpha,\, df_1,\, df_2}$.
Here, $F_{\alpha,\, df_1,\, df_2} = 1.0084$. Clearly, $F_{test} \geq F_{\alpha,\, df_1,\, df_2}$, so we reject H_0.

7.6.2 p-Value Approach

The p-value is given by $p = P(F > F_{test}) = P(F > 1.0674) \approx 1.11e-16$. As $p < \alpha$, we reject the null hypothesis H_0.

7.6.3 Conclusion

At 5% level of significance, we have sufficient statistical evidence to claim that men have more variability in weights than women. However, this result may not be practically significant, as

the ratio of the sample variances is much less than 4.

7.7 Is Diabetes More Common in Men than in Women?

Let p_1 denote the proportion of diabetic men and p_2 denote the proportion of diabetic women. We formulate the hypothesis as:
H_0 : Diabetes is not more common in men than in women. $(p_1 - p_2 \leq 0)$
H_a : Diabetes is more common in men than in women. $(p_1 - p_2 > 0)$
This is a right-tailed test.

Number of diabetic men: 23765, and number of men who are not diabetic: 124893 (where total number of male residents $(n_1) = 148658$)
Number of diabetic women: 23302, and number of women who are not diabetic: 136894.(where total number of female residents $(n_2) = 160196$)
(Sample) Proportion of diabetic men (p_1): 0.1599
(Sample) Proportion of diabetic women (p_2): 0.1454
The test statistic is:

$$z^* = \frac{(p_1 - p_2) - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = 11.115$$

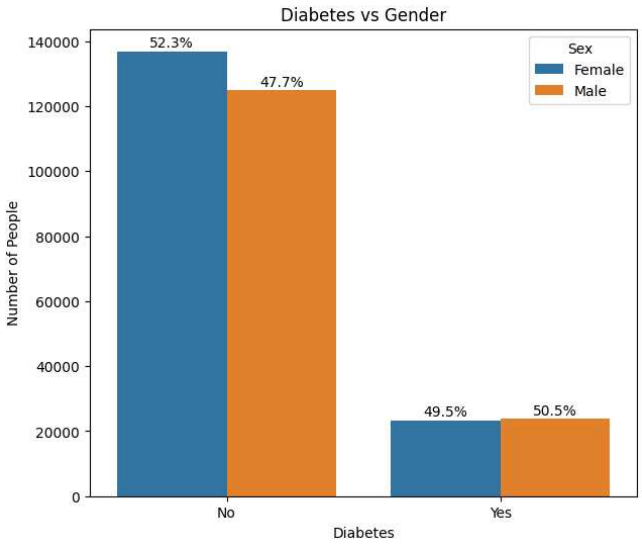


Figure 14: Diabetes Vs Gender

7.7.1 Rejection Region Approach

As this is a right-tailed test, the rejection region is $z^* > Z_\alpha$, where $\alpha = 0.05$. We reject the null hypothesis if $z^* > 1.64$ (since $z_\alpha = 1.64$). It is clear that $z^* = 11.115 > 1.64$, and thus we reject the null hypothesis H_0.

7.7.2 p-Value Approach

The p-value for the right-tailed test is $p = P(Z > z^*) \approx 0$, which is clearly $< \alpha = 0.05$. Hence, we reject the null hypothesis.

7.7.3 Conclusion

At 5% significance level, we have sufficient statistical evidence to claim that diabetes is more common in men than women.

8 Undersampling

8.1 What is Undersampling?

Most of the real-world data are imbalanced, which leads to unwanted biases in the results of analysis and predictive modeling. Undersampling is a technique of reducing the size of the dataset with imbalanced class distribution to balance the class distribution [5]. It involves selecting and deleting data points belonging to the majority class to reduce the skew in the class distribution.

8.2 The Need for Undersampling

As for the remaining part, we want to test several hypotheses related to heart disease based on lifestyle and other factors, our target class is "Heart Disease". It is evident from figure 15 that the target class is highly imbalanced, with 2,83,883 people not having a heart disease while only 24,971 people having a heart disease. This imbalance makes any data analysis or predictive modelling task highly biased towards the majority class. Thus, it is essential to balance the dataset first.

Figure 15: No. of People with Heart Disease in Original Dataset

8.3 Undersampling Technique

The simplest undersampling method is to randomly select and eliminate rows (data points) belonging to the majority class, while keeping examples of the minority class untouched. This method is known as *random undersampling* or *naive undersampling*, since no heuristics are used and nothing is assumed about the data. We use this technique to randomly select people (samples) with heart disease from the dataset, in order to balance the class distribution.

9 Hypothesis Testing on Categorical Features: Heart Diseases

The χ^2 test is used to test hypothesis on categorical variables. It tests whether two categorical variables are independent or dependent. The hypothesis is formulated as follows:
H_0 : The two categorical variables are independent ($\chi^2 = 0$)
H_a: The two categorical variables are dependent ($\chi^2 > 0$)

Then, a contingency (or crosstab) table is formed showing the **observed values or frequencies** in each category. Denote the observed frequency in cell i by f_{o_i}.

	Dependent Variable		
Independent Variable	Yes	No	Row Marginal (rm):
Yes	f_{o_1}	f_{o_2}	$f_{o_1} + f_{o_2}$
No	f_{o_3}	f_{o_4}	$f_{o_3} + f_{o_4}$
Column marginal (cm):	$f_{o_1} + f_{o_3}$	$f_{o_2} + f_{o_4}$	Total = $\sum f_{o_i}$

Contingency Table of Observed Frequencies

Next, a contingency table for **expected frequencies** is formed. The expected frequency is that value which would be seen if the null hypothesis was true (that is, if the two variables were indeed unrelated).

	Dependent Variable		
Independent Variable	Yes	No	Row Marginal (rm):
Yes	f_{e_1}	f_{e_2}	$f_{e_1} + f_{e_2}$
No	f_{e_3}	f_{e_4}	$f_{e_3} + f_{e_4}$
Column marginal (cm):	$f_{e_1} + f_{e_3}$	$f_{e_2} + f_{e_4}$	Total = $\sum f_{e_i}$

Contingency Table of Expected Frequencies

Denote the expected frequency in cell i by f_{e_i}. The formula for the expected frequency count is

$$f_{e_i} = \frac{rm_i \cdot cm_i}{N}$$

where f_{e_i} : Expected frequency for cell i (if H_0 was true)
$rm_i,\ cm_i$: The row and column marginals of cell i respectively
N : Total sample size

Next, the test statistic χ^2_{test} is obtained using the formula:

$$\chi^2_{test} = \sum_{i=1}^{N} \frac{(f_{o_i} - f_{e_i})^2}{f_{e_i}}$$

χ^2_{test} is substantially a measure of the difference between the observed frequencies and expected frequencies. The conclusion can be drawn in two ways:

Rejection region approach: The critical value is $\chi^2_{\alpha,\, df}$, where $df = (r-1)(c-1)$. Here, df denotes the degrees of freedom, and r and c denotes the number of rows and columns in the contingency table respectively. $\chi^2_{\alpha,\, df}$ is compared with χ^2_{test}. We reject the null hypothesis H_0 if $\chi^2_{test} > \chi^2_{\alpha,\, df}$.

p-Value approach: The p-value represents the area under the density curve of the χ^2 distribution to the right of the test statistic χ^2_{test}. In other words, $p = P(\chi^2 > \chi^2_{test})$. The null hypothesis is rejected if $p < \alpha$.

9.1 Is Diabetes Related to Heart Diseases?

Many medical studies have shown that diabetes is closely related to heart disease. In fact, diabetic people are more prone to having heart attacks and other complicated cardiovascular diseases [6]. We conduct a χ^2-test on our dataset to check whether we have enough statistical evidence to conclude that diabetes is indeed related to heart disease. Through the χ^2-test, we cannot determine whether diabetes increases the chances of heart diseases - since it tests only if the two variables are associated, and not the degree (positive or negative) of association.

Figure 16 shows that **71.8% of the people who have diabetes also have heart disease**, and **57.5% of the people who do not have diabetes do not have heart disease**. From this, we may say that diabetes may increase the chances of heart disease.
Now formulate the hypothesis at 5% level of significance.
H_0 : Diabetes has no effect on heart diseases.
H_a: Diabetes is associated with heart diseases.

Below figure 16 is the contingency table and heat-map for the observed frequencies.

Test statistic: $\chi^2_{test} = 3258.96$
Degrees of freedom: $df = (r-1)(c-1) = (2-1)(2-1) = 1$
Critical value: $\chi^2_{0.05,1} = 3.841$
p-Value: $p \approx 0$

Conclusion:

1. As $\chi^2_{test} > \chi^2_{critical}$, we reject the null hypothesis H_0.
2. Also, $p \approx 0 < 0.05 = \alpha$, so the p-value approach also leads us to rejecting H_0.

Hence, diabetes is indeed associated with heart diseases.

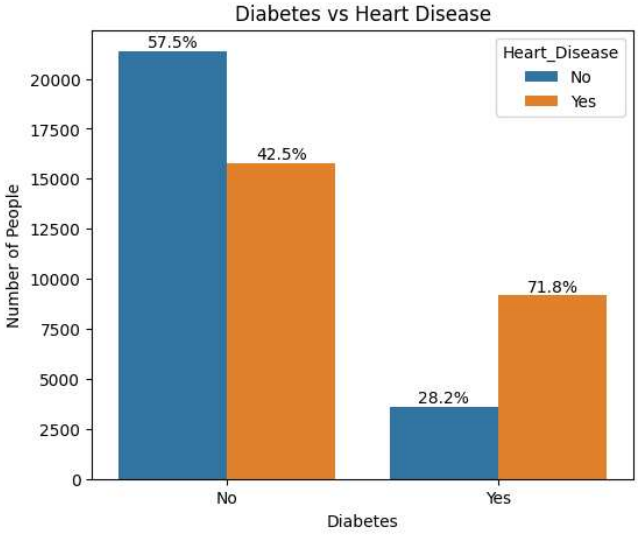
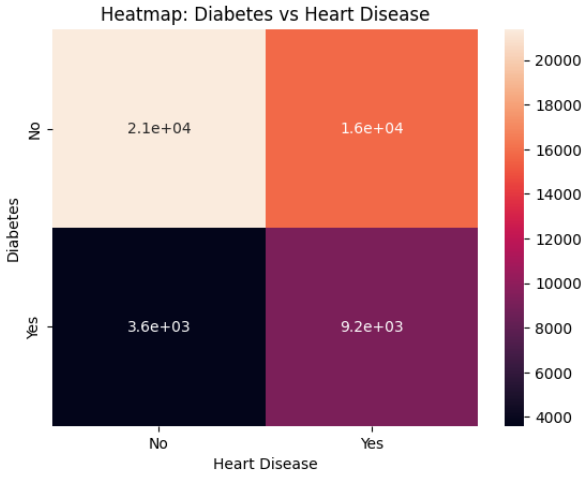


Figure 16: Diabetes vs Heart Disease

	Heart Disease	
Diabetes	No	Yes
No	21368	15801
Yes	3603	9170



9.2 Is Arthritis Related to Heart Diseases?

Arthritis is the inflammation of one or more joints that causes pain and stiffness. Different types of arthritis exist, each with different causes including wear and tear, infections and underlying diseases. Many studies have shown that people with arthritis have an increased risk of heart disease [7]. We conduct a χ^2-test on our dataset to check whether we have enough statistical evidence to conclude that arthritis is related to heart disease.

Figure 17 shows that **64.3% of the people who have arthritis also have heart disease**, and **61.4% of the people who do not have arthritis do not have heart disease**.

We formulate our hypothesis at 5% level of significance as follows:
H_0 : Arthritis has no effect on heart diseases.
H_a: Arthritis is associated with heart diseases.

Below figure 17 is the contingency table and heat-map for the observed frequencies.

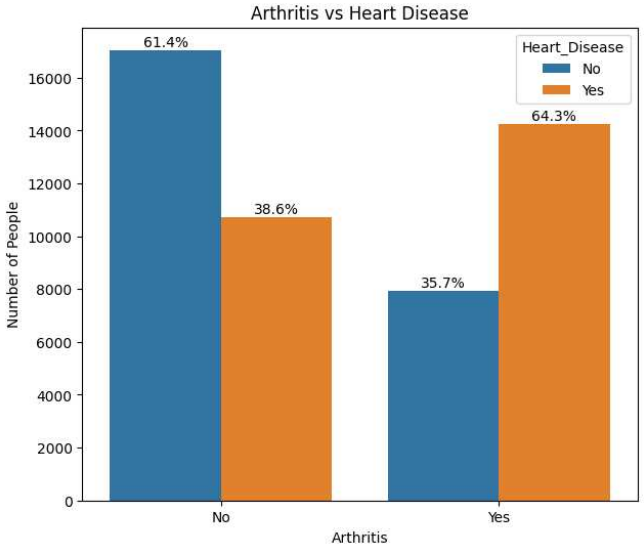
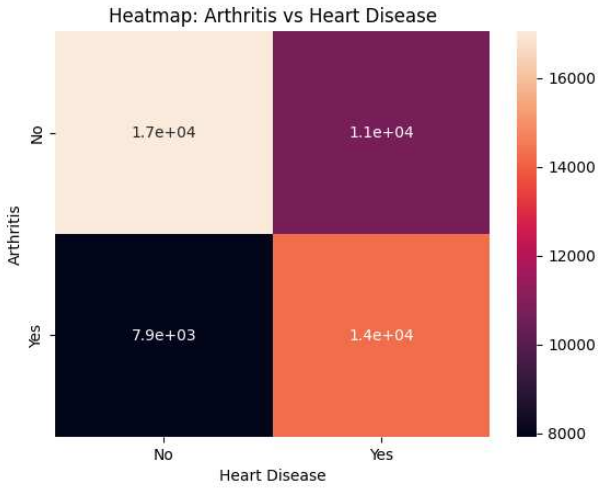


Figure 17: Arthritis vs Heart Disease

	Heart Disease	
Arthritis	No	Yes
No	17048	10719
Yes	7923	14252



Test statistic: $\chi^2_{test} = 3247.93$
Degrees of freedom: $df = (r-1)(c-1) = (2-1)(2-1) = 1$
Critical value: $\chi^2_{0.05,1} = 3.841$
p-Value: $p \approx 0$

Conclusion:

1. As $\chi^2_{test} > \chi^2_{critical}$, we reject the null hypothesis H_0.
2. $p \approx 0 < 0.05 = \alpha$, so we reject H_0 through the p-value approach as well.

Hence, we conclude that we have enough statistical evidence to claim that arthritis is associated with heart diseases.

9.3 Is Age Associated with Heart Diseases?

According to some studies, adults of age 65 and older are more likely than younger people to suffer from cardiovascular diseases. Aging can cause changes in the heart and blood vessels that may increase a person's risk of developing cardiovascular disease.
The following contingency table and heatmap (figure 18) of "Age category vs Heart Disease" clearly supports the assertion above, because a considerably large number of elderly people are having heart diseases as compared to the young.

Heart_Disease	No	Yes
Age_Category		
18-24	1414	94
25-29	1310	113
30-34	1550	201
35-39	1859	274
40-44	1895	435
45-49	1848	678
50-54	2246	1181
55-59	2415	1991
60-64	2653	3012
65-69	2653	3823
70-74	2292	4561
75-79	1429	3752
80+	1407	4856

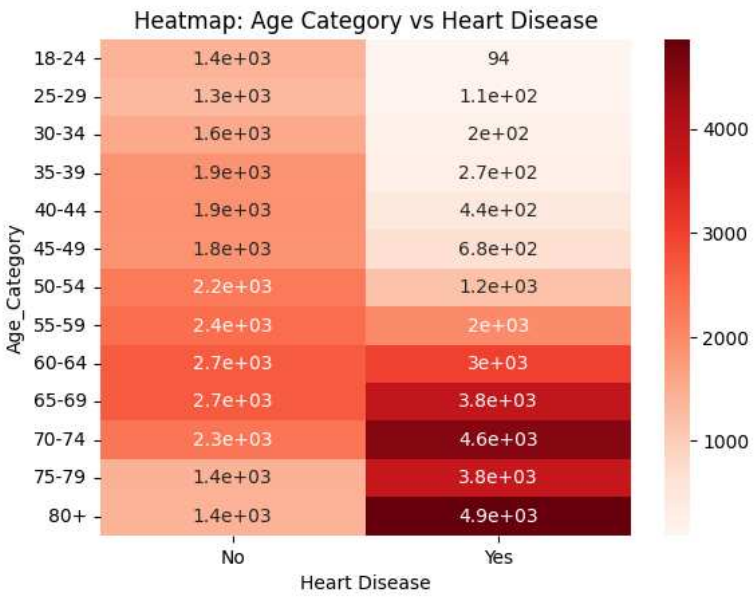


Figure 18: Age vs Heart Disease

We formulate our hypothesis at 5% level of significance as follows:
H_0 : Age does not have an effect on heart diseases.
H_a: Age is associated with heart diseases.

Test statistic: $\chi^2_{test} = 10134.2506$
Degrees of freedom: $df = (r-1)(c-1) = (13-1)(2-1) = 12$
Critical value: $\chi^2_{0.05,12} = 21.026$
p-Value: $p \approx 0$

Conclusion:

1. As $\chi^2_{test} > \chi^2_{critical}$, we reject the null hypothesis H_0.
2. $p \approx 0 < 0.05 = \alpha$, so we reject H_0 through the p-value approach as well.

Hence, we conclude that we have enough statistical evidence to claim that age is related to heart diseases.

10 Fruits & Green Vegetables Consumption vs Heart Disease

The **Fruit_Consumption** and **Green_Vegetables_Consumption** columns in the dataset show the average number of fruits and green vegetables consumed by a resident in a month. We make the following boxplots to study the effect of consumption of fruits and vegetables on heart diseases, if any.

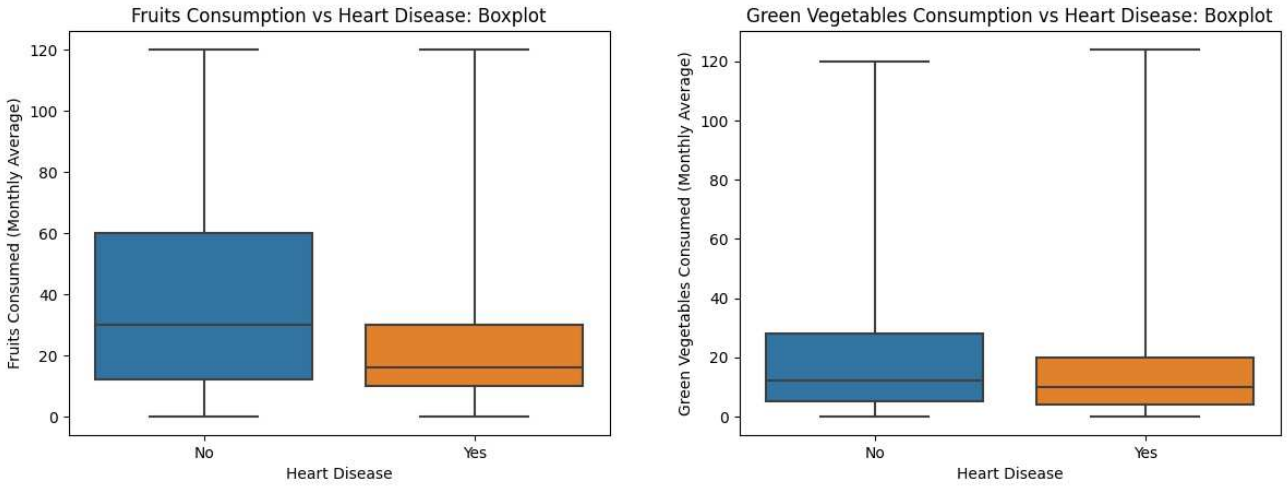


Figure 19: Fruits & Vegetables Consumption vs Heart Disease

From the plots, we can see that the "median number of fruits consumed in a month on average" is greater for people with no heart disease as compared to people with heart disease. The same goes for green vegetables. Hence, fruits and green vegetables form a healthy diet as they might decrease the risk of heart diseases. This is probably one of the reasons why Peter Parker said, "Eat your green vegetables".

References

[1] 2021 BRFSS Survey Data and Documentation: Centers for Disease Control and Prevention Available online: https://www.cdc.gov/brfss/annual_data/annual_2021.html

[2] Nestor Asiamah, Henry Kofi Mensah, and Eric Fosu Oteng-Abayie, "Do Larger Samples Really Lead to More Precise Estimates? A Simulation Study." American Journal of Educational Research, vol. 5, no. 1 (2017): 9-17. doi: 10.12691/education-5-1-2 Available online: https://www.researchgate.net/publication/312174643_Do_Larger_Samples_Really_Lead_to_More_Precise_Estimates_A_Simulation_Study

[3] Cleaned Dataset: Cardiovascular Diseases Risk Prediction Dataset (Kaggle) Available online: https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset

[4] Wikipedia Article: Average human height by country (Measured and self-reported figures) Available online: https://en.wikipedia.org/wiki/Average_human_height_by_country

[5] Fernández, Alberto et al. "Learning from Imbalanced Data Sets", Page 82 Available online: https://www.amazon.com/Learning-Imbalanced-Data-Alberto-Fern%C3%A1ndez/dp/3319980734/

[6] Leon BM, Maddox TM. Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research. World J Diabetes. 2015 Oct 10;6(13):1246-58. doi: 10.4239/wjd.v6.i13.1246. PMID: 26468341; PMCID: PMC4600176.

[7] Crowson CS, Liao KP, Davis JM 3rd, Solomon DH, Matteson EL, Knutson KL, Hlatky MA, Gabriel SE. Rheumatoid arthritis and cardiovascular disease. Am Heart J. 2013 Oct;166(4):622-628.e1. doi: 10.1016/j.ahj.2013.07.010. Epub 2013 Aug 29. PMID: 24093840; PMCID: PMC3890244.

Roles

- **Rajdeep Pathak (MA23MSCST11013):**
  - Data wrangling, calculations through Python codes, plots generation using Matplotlib
  - Hypothesis Testing
  - Report writing & Presentation
- **Sonali Saha (MA23MSCST11022):**
  - Hypothesis Testing
  - Report writing
- **Deepak Yadav (MA23MSCST11005):**
  - Hypothesis Testing
  - Report writing
- **Sayani Mondal (MA23MSCST11019):**
  - Report writing
- **Rohit Kumar Das (MA23MSCST11016):**
  - Report Writing