# STATISTICS ON HEIGHT AND WEIGHT

Vu Truong Tam Tran Prakruti Makwana Nikhil Vasudevan

# CONTENTS

# 1. Introduction

Did you know the maximum height you can have in order to be a part of Indian army is 6 feet and 8 inches? In a country like India where the average male height is 5 feet and 5 inches, the occurrence of 6 footers may be very rare. The average height in Vietnam is 5 feet 7 inches and in Netherlands its astoundingly 6 feet! And the female counterpart being at 5 feet and 7 inches.

So how does organizations and governments around the world collect the height of its citizens and members in order to facilitate some decisions that require the same. The public transport of Netherlands must be designed in such a way that it can carry loads of tall people while it is not the same case in India.

Collection of Height and Weight is also important in order to assess the health of citizens. More healthy citizens, the merrier.

# 2. Overview of Datasets

## 2.1 Dataset 1

We would like to name this as "*Dataset 1*". This dataset has more than 8000 rows of data which houses heights (centimeters) and weights (kilograms) of male and female students aged 18 to 24 from, where we assume is a province or state of a particular country.

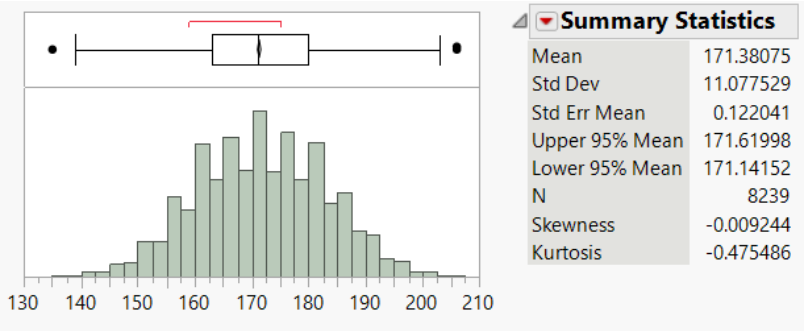



**Figure 2.1** *Height distribution and its summary statistics*

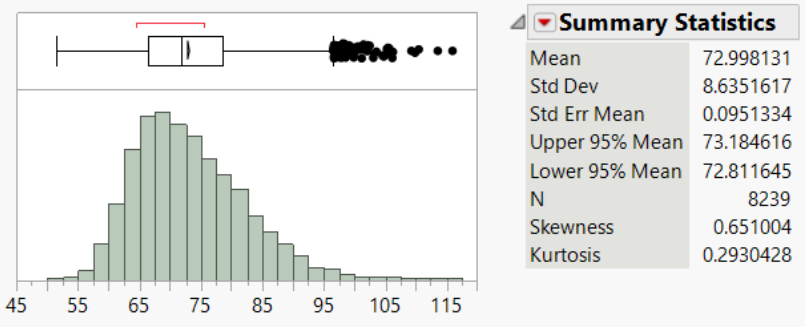



**Figure 2.2** *Weight distribution and its summary statistics*

The images attached above are the distributions of Height and Weight respectively and since the Height is not normally distributed, unfortunately, we would not be able to use it.

### 2.2 Dataset 2

Regarding to "*Dataset 2*", this contains 25,000 rows of height and weight of 18-year-old students from Hong Kong. Such a dataset is also normally distributed. We convert Height (inches) and Weight (pounds) into Height (meters) and Weight (kilograms) for our purpose of analyzing.

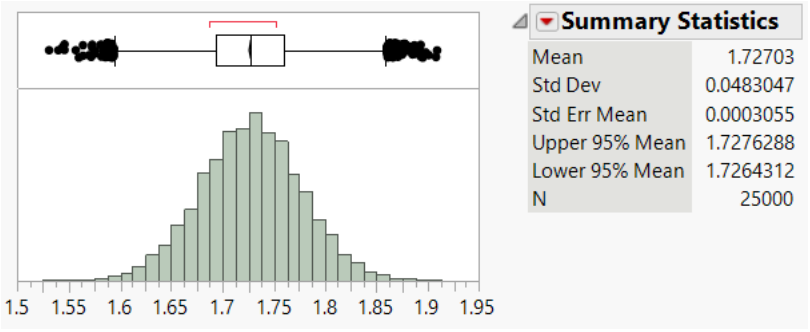



**Figure 2.3** *Distribution and summary statistics of students' Height in Hong Kong*

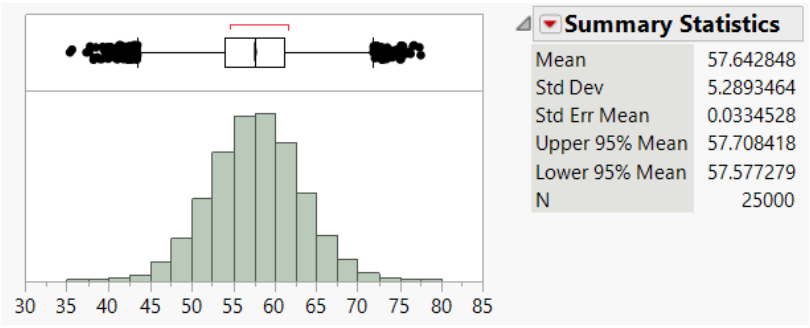



**Figure 2.4** *Distribution and summary statistics of students' Weight in Hong Kong*

The reason why we tend to use normally distributed data is that it is the most relevant distribution of probability since it fits many natural phenomena. It is also known as the bell curve and gaussian distribution.

As you can see in the above images, the graphs form a bell-shaped curve. Based on such diagrams, we would further analyze and give our outcomes.

## 3. Hypothesis testing on weight of male and female

### 3.1 Descriptive analysis of Population

We divide dataset 1 into two groups: male and female and consider them as two new populations. It is not difficult to use JMP to draw graphs and statistics values.

| ***male's weight*** | ***female's weight*** |
|---|---|
| $\sigma_1 = 7.67$ | $\sigma_2 = 5.23$ |

**Figure 3.1** *Standard deviation of two populations*

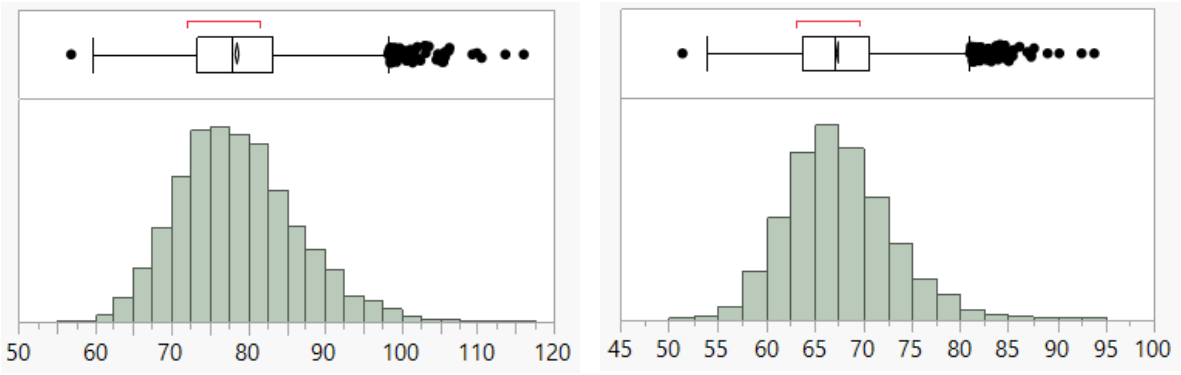



**Figure 3.2** *Normal distribution of two populations*

### 3.2 Descriptive analysis of Sample

400 samples of male and 400 samples of female were taken from the weight population (*figure 3.2*) and analysis was performed on that. The graphs and the results are as follows

#### 3.2.1 Sample male's weight of dataset 1

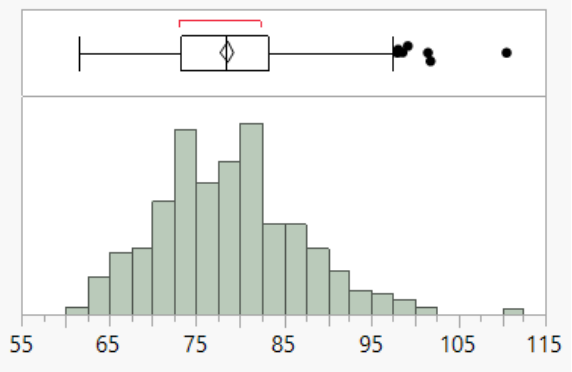



**Figure 3.3** *Histogram depicting weights of the male sample*

| | |
|---|---|
| ***Mean*** | 78.57 |
| ***Std deviation*** | 7.95 |
| ***Std Error Mean*** | 0.40 |
| ***Upper 95% Mean*** | 79.35 |
| ***Lower 95% Mean*** | 77.78 |
| ***N*** | 400 |
| ***Variance*** | 63.22 |
| ***Median*** | 78.45 |
| ***Mode*** | 73.8 |
| ***Skewness*** | 0.45 |
| ***Kurtosis*** | 0.36 |

**Figure 3.4** *Summary Statistics of male's weight sample*

| | |
|---|---|
| ***100%*** | 110.5 |
| ***75%*** | 83.07 |
| ***50%*** | 78.45 |
| ***25%*** | 73.2 |
| ***0%*** | 61.4 |

**Figure 3.5** *Quantiles of male's weight sample*

### 3.2.2 Sample female's weight of dataset 1

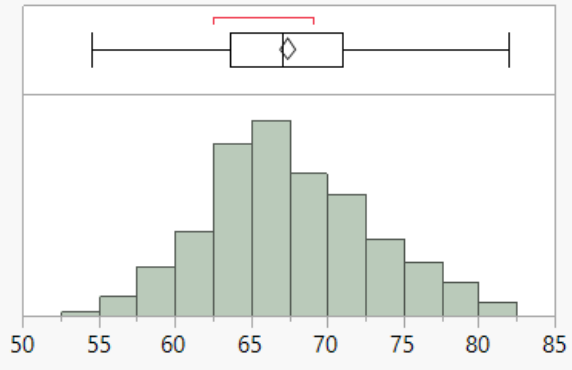



**Figure 3.6** *Histogram depicting weights of the female sample*

| | |
|---|---|
| ***Mean*** | 67.44 |
| ***Std deviation*** | 5.25 |
| ***Std Error Mean*** | 0.26 |
| ***Upper 95% Mean*** | 67.96 |
| ***Lower 95% Mean*** | 66.93 |
| ***N*** | 400 |
| ***Variance*** | 27.62 |
| ***Median*** | 67 |
| ***Mode*** | 67.3 |
| ***Skewness*** | 0.27 |
| ***Kurtosis*** | -0.26 |

**Figure 3.7** *Summary Statistics of female's weight sample*

| | |
|---|---|
| ***100%*** | 82 |
| ***75%*** | 70.98 |
| ***50%*** | 67 |
| ***25%*** | 63.6 |
| ***0%*** | 54.6 |

**Figure 3.8** *Quantiles of female's weight sample*

### 3.3 Hypothesis testing of weight of men and women

The weights in kilogram of males of a particular region is known to be normally distributed with a standard deviation of 7.67 ($\sigma_1 = 7.67$). The weights of female in the same region are also known to be normally distributed but with a standard deviation of 5.23 ($\sigma_2 = 5.23$). 400 samples are taken from each gender and the mean weight of male and female is found to be 78.57 and 67.44 respectively ($\bar{x}_1 = 78.57,\ \bar{x}_2 = 67.44$).

Investigate at the 5% level of significance, the claim that the mean weight of men is 12 kilograms $(\mu_1 - \mu_2 > 10)$ greater than the mean weight of women.

We use "Test on two mean" method for solving this problem:

- Null hypothesis: $H_0$: $\mu_1 - \mu_2 = d_0 = 10$
- Alternative hypothesis: $H_1$: $\mu_1 - \mu_2 > 10$ (one tail test)
- $\bar{x}_1 - \bar{x}_2 = 78.57 - 67.44 = 11.13$
- $n_1 = n_2 = 400$
- $\alpha = 0.05$
- Value of Test Statistic: $z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{11.13 - 10}{\sqrt{\frac{7.67^2}{400} + \frac{5.23^2}{400}}} = 2.43$
- $z_{0.05} = 1.64$

Because $z = 2.43 > z_{0.05} = 1.64$, so we will have to reject $H_0$ and accept $H_1$.

## 4. Analyze Height column from dataset 2

We are going to apply Hypothesis Testing on Dataset 2 by picking a sample data consisting of 51 random records from the dataset.

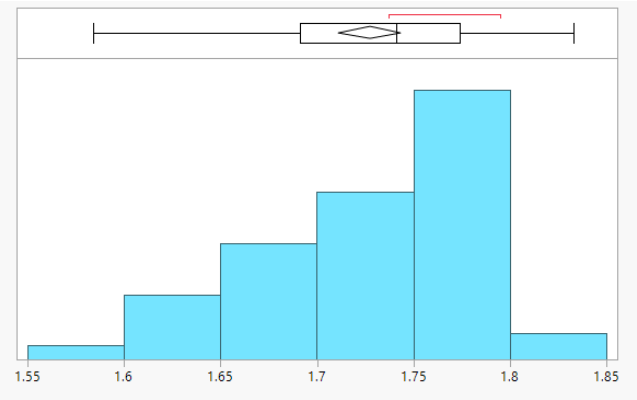



**Figure 4.1** *Height of the sample data*

Mean and Standard Deviation of the sample data is calculated using the following formulas:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

| | |
|---|---|
| ***Mean*** | 1.73 |
| ***Standard Deviation*** | 0.057 |
| ***N*** | 51 |

**Figure 4.2** *Summary Statistics*

Hypothesis testing will be used to determine whether sample variance is equal to population variance or not, at 5% level of significance.

- Null hypothesis: $H_0$: $\sigma^2 = \sigma_0^2$ ($\sigma_0^2$ is sample variance and $\sigma^2$ is population variance)
- Alternative hypothesis: $H_1$: $\sigma^2 \neq \sigma_0^2$ (2-tail Test)
- $\sigma = 0.048$
- $s = \sigma_0 = 0.057$
- $\alpha = 0.05 \quad \Rightarrow \quad \frac{\alpha}{2} = 0.025$
- $n - 1 = 51 - 1 = 50$
- Value of Test Statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{50 \cdot 0.057^2}{0.048^2} = 70.5$$

- $\chi^2_{0.025,50} = 71.42$ and $\chi^2_{1-0.025,50} = 32.36$

Because $\chi^2_{1-0.025,50} < \chi^2 < \chi^2_{0.025,50}$, so we do not have to reject $H_0$.

## 5. Analyze Weight column from dataset 2

We are going to apply Hypothesis Testing on Dataset 2 by picking a sample data consisting of 51 random records from the dataset.

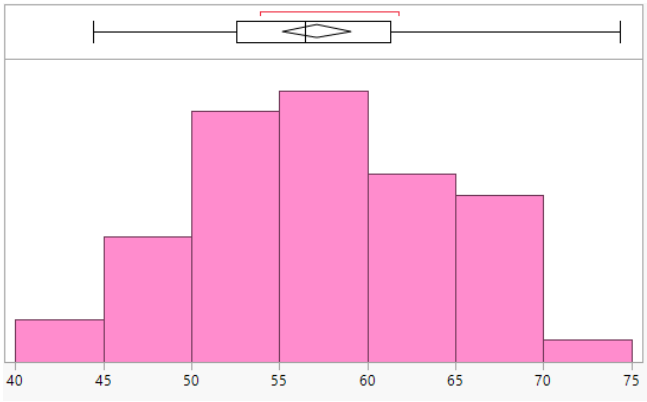



**Figure 5.1** *Weight of the sample data*

Mean and Standard Deviation of the sample data is calculated using the following formulas:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

| | |
|---|---|
| ***Mean*** | 57.13 |
| ***Standard Deviation*** | 6.9 |
| ***N*** | 51 |

**Figure 5.2** *Summary Statistics*

Hypothesis testing will be used to determine whether sample variance is equal to population variance or not, at 5% level of significance.

- Null hypothesis: $H_0$: $\sigma^2 = \sigma_0^2$ ($\sigma_0^2$ is sample variance and $\sigma^2$ is population variance)
- Alternative hypothesis: $H_1$: $\sigma^2 \geq \sigma_0^2$ (right-tail Test)
- $\sigma = 5.29$
- $s = \sigma_0 = 6.9$
- $\alpha = 0.05$
- $n - 1 = 51 - 1 = 50$
- Value of Test Statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{50 \cdot 6.9^2}{5.29^2} = 85.07$$

- $\chi^2_{0.05,50} = 67.50$

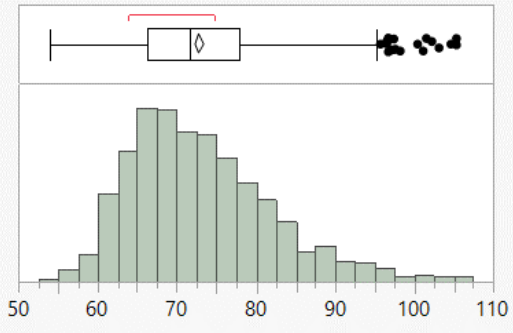
Because $\chi^2 > \chi^2_{0.05,50}$, therefore we have to reject $H_0$ and accept $H_1$.

## 6. Compare two datasets

In this section, we will analyze Weight columns from two datasets.

As mentioned in section 2.1, the average weight of 8239 students is 73 ($\mu_1 = 73$). From this data, we pick 1001 rows randomly (around 12% of population) as a sample. After calculating, the mean is 72.76 ($\bar{x}_1 = 72.76$) with the standard deviation of 8.78 ($s_1 = 8.78$).

Similarly, we also take 1001 records from dataset2 ($\mu_2 = 57.64$) and compute related values. This sample gives an average of 57.34 ($\bar{x}_2 = 57.34$) with the standard deviation of 5.22 ($s_2 = 5.22$).



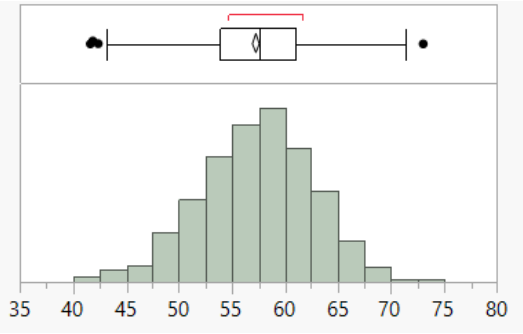




**Figure 6.1** *Histogram of sample of dataset 1 and dataset 2 respectively*

| | ***Dataset1*** | ***Dataset2*** |
|---|---|---|
| ***Population*** | $\mu_1 = 73$ | $\mu_2 = 57.64$ |
| ***Sample*** | $\bar{x}_1 = 72.76$ | $\bar{x}_2 = 57.34$ |
| | $s_1 = 8.78$ | $s_2 = 5.22$ |

**Figure 6.2** *Mean and standard deviation of two samples*

Thus, we conduct an experiment to compare the weight of two different datasets with 0.1 level of significant.

We use "Test concerning variance" for 2 samples:

- Null hypothesis: $H_0$: $\sigma_1^2 = \sigma_2^2$
- Alternative hypothesis: $H_1$: $\sigma_1^2 \neq \sigma_2^2$
- Degrees of two samples are the same: $v_1 = v_2 = 1000$
- Value of Test statistic: $f = \frac{s_1^2}{s_2^2} = \frac{8.78^2}{5.22^2} = 2.83 \sim f_{1000,1000}$
- We calculate:

$$f_{0.05}(v_1, v_2) = 1.11$$
$$f_{0.95}(v_1, v_2) = \frac{1}{f_{0.05}(v_2, v_1)} = \frac{1}{1.11} = 0.9$$

Because $f = 2.83 > f_{0.05}(v_1, v_2) = 1.11$, so we reject null hypothesis and accept alternative hypothesis, which means that $\sigma_1^2 \neq \sigma_2^2$.

# 7. Relation between Height and Weight

## 7.1 Scatter plot

The following graph shows the scatter plot between the height (y-axis) and the weight (x-axis) of the entire population that is 25,000 people.

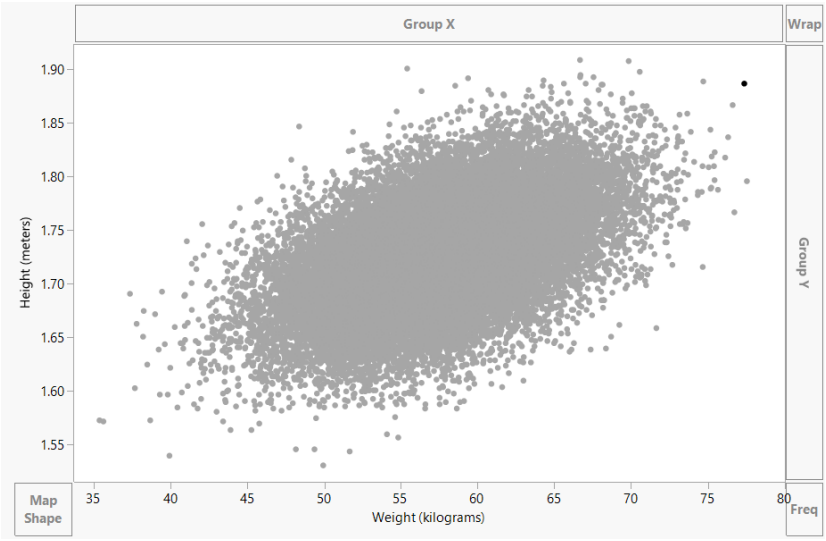



**Figure 7.1** *Scatter plot of Height and Weight*

The scatter plot does not indicate a linear relationship. The points do not follow a trend. In other words, there does not appear to be a relationship between the height and the weight of the population. Therefore, Height and weight are independent parameters.

## 7.2 BMI

### 7.2.1 What is BMI

Body Mass Index is a value determined from the mass and height of a person. The formula for BMI is weight in kilograms divided by height in meters squared. It is a way to measure whether the weight is in proportion to the height. It tells us if we are underweight, normal, overweight or obese according to the following table:

| | |
|---|---|
| **0 – 18.5** | Underweight |
| **18.5 – 24.9** | Normal |
| **25 – 29.9** | Overweight |
| **> 30** | Obese |

**Figure 7.2** *BMI table*

The formula of BMI is:

$$\text{BMI} = \frac{\boldsymbol{weight\ (in\ kg)}}{\boldsymbol{height^2\ (in\ m)}}$$

### 7.2.2 BMI of 18-year-old students

We use the Height and Weight of dataset 2 to compute BMI of 25,000 students.

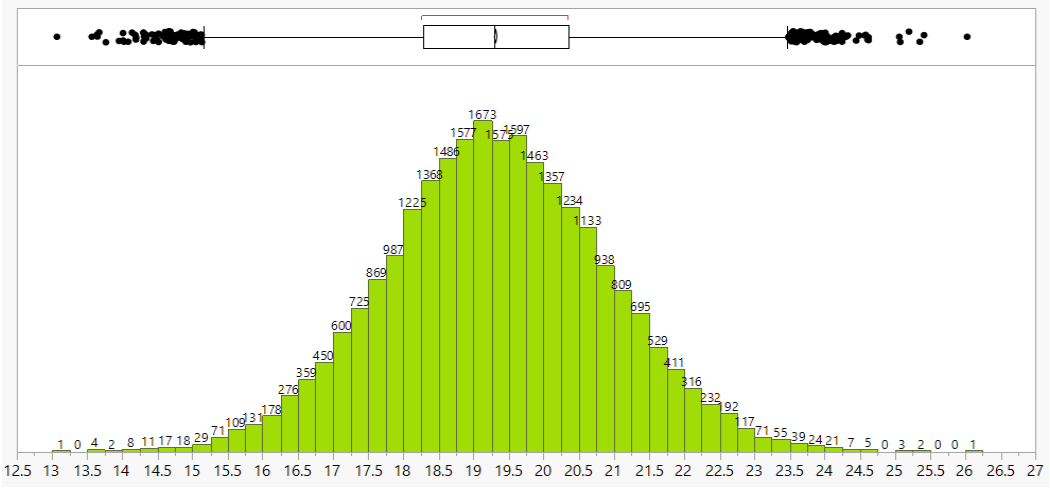



**Figure 7.3** *BMI of 25,000 students*

The histogram shows that BMI is normally distributed.

| | |
|---|---|
| ***Mean*** | 19.32 |
| ***Std deviation*** | 1.55 |
| ***Std Error Mean*** | 0.01 |
| ***Upper 95% Mean*** | 19.34 |
| ***Lower 95% Mean*** | 19.30 |
| ***N*** | 25000 |
| ***Variance*** | 2.40 |
| ***Median*** | 19.30 |
| ***Mode*** | 19.77 |
| ***Skewness*** | 0.02 |
| ***Kurtosis*** | 0.02 |

**Figure 7.4** *Summary Statistics*

| | |
|---|---|
| ***100%*** | 26.03 |
| ***75%*** | 20.40 |
| ***50%*** | 19.30 |
| ***25%*** | 18.28 |
| ***0%*** | 13.07 |

**Figure 7.5** *Quantiles*

**Observations with respect to the BMI values obtained:**

**Underweight:**

The striped part of the histogram shows the part of the population which is underweight i.e. their BMI lies between 0–18.5. There are 7438 people out of 25000 are underweight.

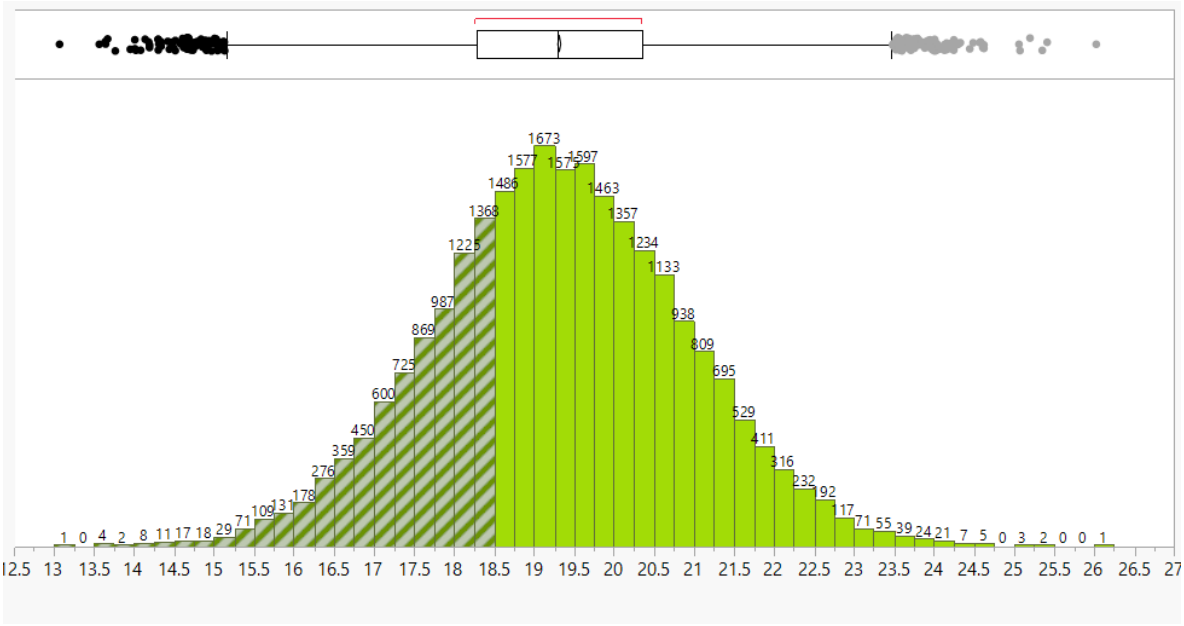



**Figure 7.6** *Histogram of underweight*

**Normal:**

The striped part of the histogram shows the part of the population which is normal i.e. their BMI lies between 18.5 – 24.9. There are 17556 people out of 25000 have normal weight.

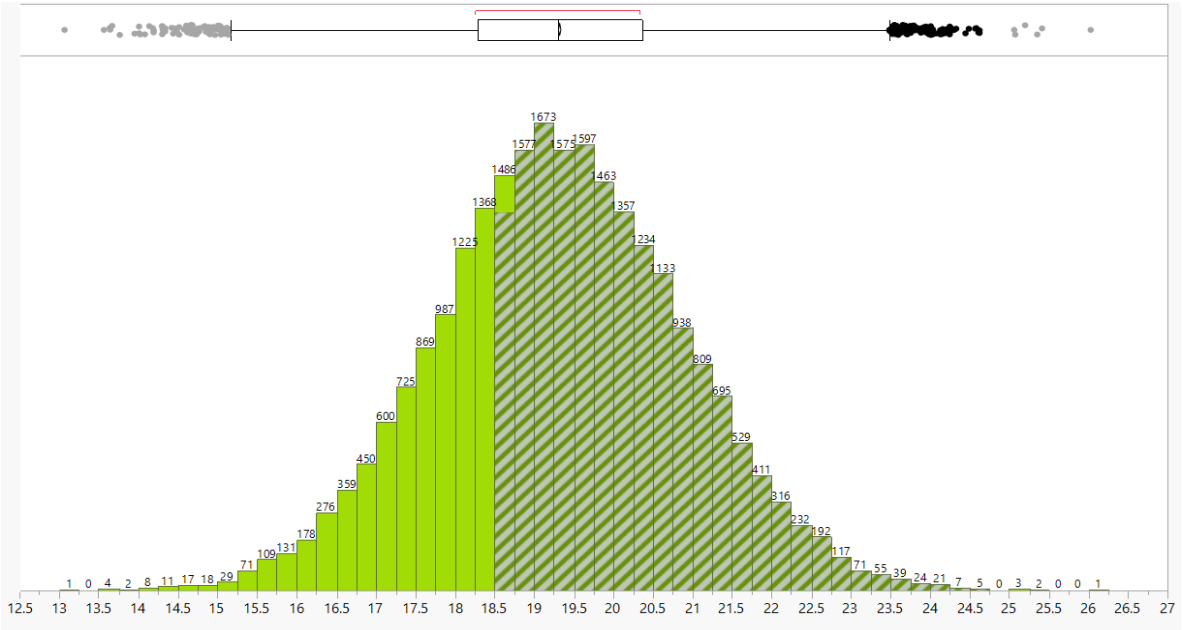



**Figure 7.7** *Histogram of normal weight*

**Overweight:**

The striped part of the histogram shows the part of the population which is overweight i.e. their BMI lies between 25 – 29.9. Only 6 people out of 25000 are overweight.

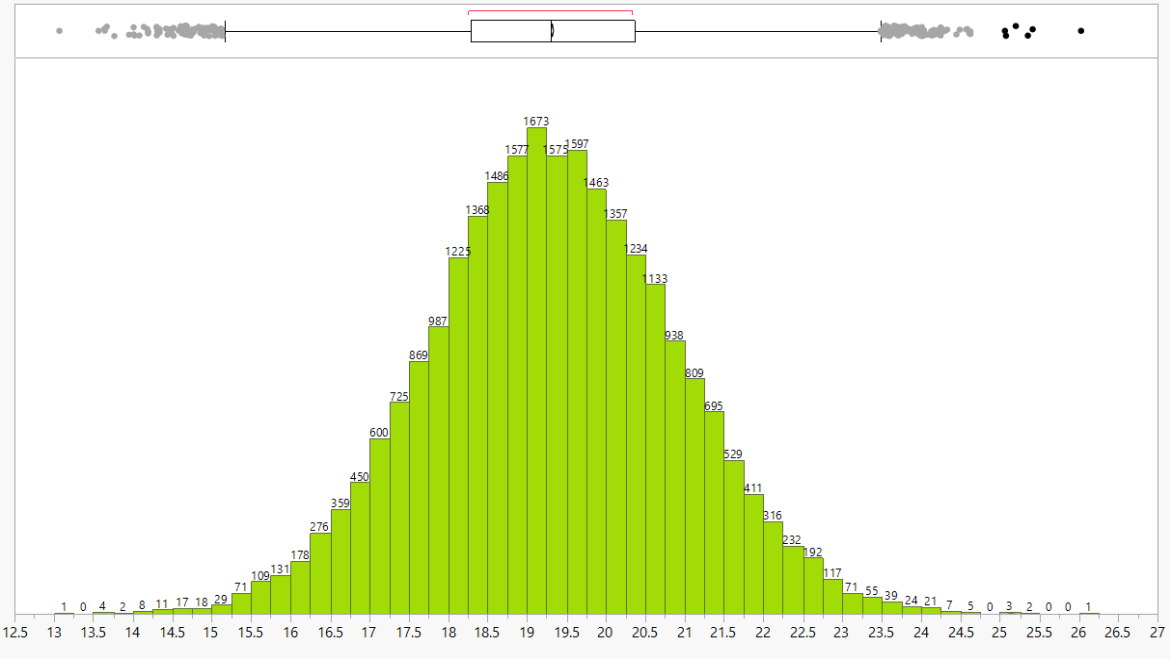



**Figure 7.8** *Histogram of overweight*

**Conclusion:**

From the 3 histograms given above, we conclude that the percentages of population falling into the categories of BMI are as follows:

(Using $Percentage = \frac{part}{total} \times 100$)

| | |
|---|---|
| Underweight | 29.75% |
| Normal | 70.22% |
| Overweight | 0.03% |

# REFERENCE

Geo.fu-berlin.de, (2021). Department of Earth Sciences - Freie Universität Berlin Website. [online] Available at https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Continous-Random-Variables/The-Standard-Normal-Distribution/The-Standard-Normal-Distribution-An-Example/index.html [Accessed 10 Jan. 2021]

Smit Patel (2020). *Height and Weight dataset*. [online] Available at https://www.kaggle.com/burnoutminer/heights-and-weights-dataset [Accessed 10 Jan. 2021]

wiki.stat.ucla.edu/socr, (no date). Main SORC Wiki Pages. [online] Available at http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights [Accessed 10 Jan. 2021]