BMI Survey

02323 Introduction to statistics

Mohamad Malaz Mohamed ALZarrad

s180424

October 20, 2020

Question a: Write a short description of the data. Which variables are included in the dataset?Are the variables quantitative and/or categorized? (Categorized variables are only

introduced in Chapter 8, but they are simply variables which divide the observations into categories/groups - e.g. three categories: low, medium, and high). How many observations are there? Are there any missing values? Remember to consult the extended description of the variables in Appendix 1 (p. 11).

Short Description of Data Set and Variables

BMI (Body Mass Index) appears to be as strongly correlated with various metabolic and disease, gender, education, fast food, age outcome
this dataset is collected which includes features like:

- **Gender** : gender of respondant man or woman.
- **height** : height of respondant in Centimeter.
- **weight** : weight of respondant in kilogram.
- **Urbanity** : where respondant lives.
- **Fast Food level** : number of days respondant eats fast food in a year

Quantitative or Categorized variables :

- **Quantitative Variables:** These are the numerical values that allow mathematical operations to be performed on them These are the variable where variable can have any number of different values and like Here (**hieght** and **weight)**where each can any value greater than or equal for zero but can't be minus.
- **Categorized Varibales:** (Categorical Data) consists of a set of elements or groups. These are discrete values and divide the observation into groups. In our case categorized variables are **Gender** and **Urbanity** and **Fast food** because Gender could be 2 values, male (1) or female (0) and **Fastfood** is categorized in 8 groups with numbers from 1 to 8 and **Urbanity** here is classified in 5 classes from 1 to 5

Number of observations and Number of missing values

When I run command *Dim*(*D*) in R Studio I get the size of data.

Number of observation are the number of rows in the size its is **145** and to get the number of missing values I use the command sum(is.na(D)). When I am Running this command yielded a result of zero. **So there are no missing values.**

Question b: Make a density histogram of the BMI scores. Use this histogram to describe the empirical distribution of the BMI scores. Is the empirical density symmetrical or skewed? Can a BMI score be negative? Is there much variation to be seen in the observations?

Emperical Distribution of BMI scores and Variance in Observation

The **mean25.247** **kg /m²**,**median24.691** **kg/m²** **,variance 14.686** **kg /m²** of BMI scores that calculated in R The **min** value of BMI is **17.577** **kg /m²** and The **max** value of BMI is **39.519** **kg /m²**.

A symmetric distribution is one where the left and right hand sides of the distribution are roughly equally balanced around the mean and in our histogram below shows that our distribution is not symmetric it is a **skewed right** distribution (also known as positively skewed) is shown below in our case. For a right skewed distribution, the mean is typically greater than the median and From the above calculations we could see that median is slightly left of mean by 2.2% of mean. Also notice that the tail of the distribution on the right hand (positive) side is longer than on the left hand side.

Also **standard deviation** of BMI is **3.83** **kg /m²** which is approximately 15% of the mean

density histogram of the BMI scores :

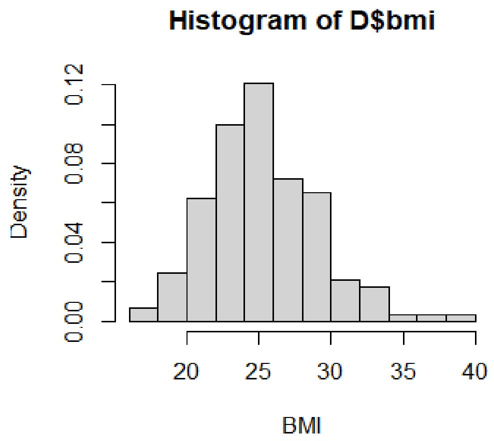


Figure 1: Empirical Density histogram

Can BMI score be negative?

No, BMI score can not be negative because all division of positive number(*weight*) andpositive number (*height*^2) can not give a negative numbers it will give only a positive numbers.

Question c : Make separate density histograms for the BMI scores of women and men, respectively.
Describe the empirical distributions of the BMI scores for men and women
using these histograms, like in the previous question. Does there seem to be a gender difference in the distribution of the BMI scores (if so, describe the difference)?

The **mean,median,variance** for female BMI scores that calculated by R script are **24.216 kg /m^2 , 23.689 kg /m^2 and 16.417 kg /m^2 .**
The **min and max** values of female BMI are **17.577 kg /m^2** and **39.519 kg /m^2**
As I explained in the previous question(b) about a **symmetric** distribution and **skewed right** distribution we have here the same situation it is a skewed right distribution is shown below in our histogram of BMI for Female, For a right skewed distribution, the mean is typically greater than the median and From the above calculations we could see that median is slightly left of mean by 2.1% of mean. Also notice that the tail of the distribution on the right hand (positive) side is longer than on the left hand side Also **standard deviation** of BMIfor female is **4.04 kg /m^2** which is approximately 16.7% of the mean.
The results give us a concluding note that mean of female BMI lies in normal region between 25 and 18(<25 and >18) which implies an average for female has a normal weight.
Histogram of emperical densities of BMI for female

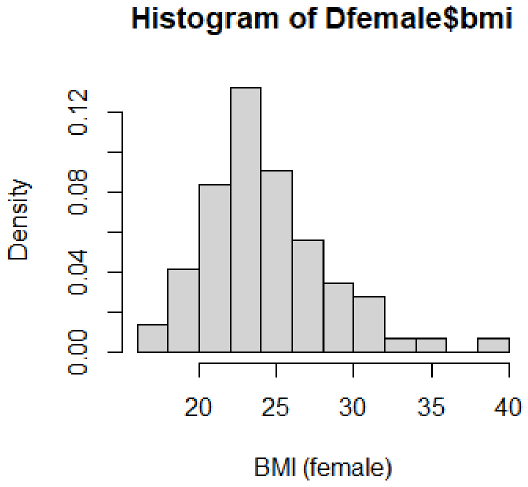


Figure 2 Empirical Density histogram for Female

The **mean,median,variance** for male BMI scores that calculated by R script are **26.25 kg /m^2 , 25.72 kg /m^2 and 11.06 kg /m^2** respectively.The **min and max** values of male BMI are **19.75 kg /m^2** and **37.57**

kg /m² .we have here also the same situation. it is a skewed right distribution is shown below in our histogram of BMI for Male, For a right skewed distribution, the mean is greater than the median and From the above calculations we could see that median is slightly left of mean by 2.06% of mean. Also notice that the tail of the distribution on the right hand (positive) side is longer than on the left hand side Also **standard deviation** of BMI for male is **3.32** **kg /m²** which is approximately 12.6 % of the female which implies the distribution of male bmi is less varient than female..

The results give us a concluding note that mean of male BMI lies in over weight region bigger than 25 (>25) which implies an average for male has a over weight.

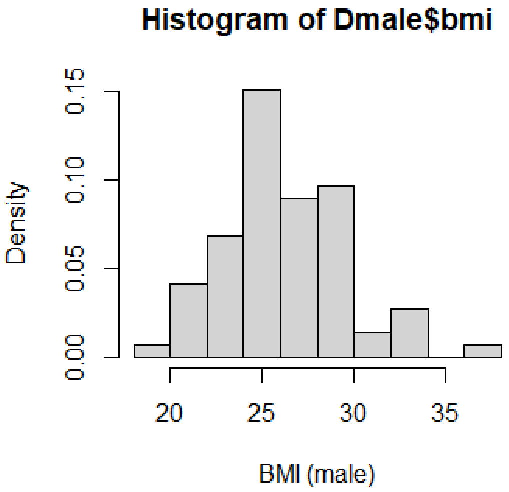


Figure 3 Empirical Density histogram for male

Question d : Make a box plot of the BMI scores by gender. Use this plot to describe the empirical distribution of the BMI scores for women and men. Are the distributions symmetrical or skewed? Does there seem to be a difference between the distributions (if so, describe the difference)? Are there extreme observations/outliers?

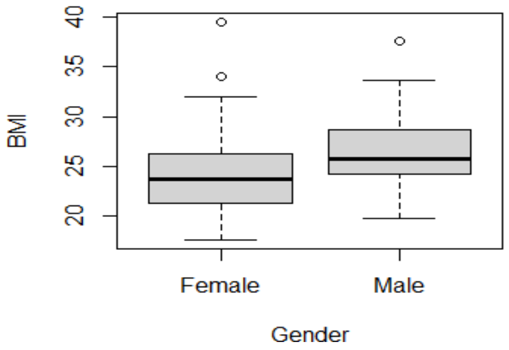


Figure 4 Box plot of BMI

It is seen clearly from our plot box that male and female distributions have long tail and that because if we look at max and min from the median we can see that max value is farther than min value from the median. we can also see that the median is closer to the first quartile than the third quartile in the meal distribution. By using R script, IQR of female will give **5.03 kg /m^2** and IQR of male will give **4.47 kg/m^2**And (**5.03 kg /m^2 > 4.48 kg/m^2)** It's because of high variance and female has more outliers.

Question e : Fill in the empty cells in the table above by computing the relevant summary statistics for BMI, first for the full sample (both genders combined), then separately for women and men. Which additional information may be gained from the table, compared to the box plot?

Table 1 summary statistics for BMI

Variable(BMI)	**No.of Obs (n)**	**Mean $(\bar{x})$**	**Variance (s^2)**	**std. dev. (s)**	**Lower quartile (Q1)**	**Median**	**Upper quartile (Q3)**
Every one	145	24.25	14.69	3.83	22.59	24.69	27.64
Woman	72	24.22	16.42	4.05	21.26	23.69	26.29
Men	73	26.26	11.07	3.32	24.15	23.72	28.63

Question f : Specify a statistical model for log-transformed BMI, making no distinction between men and women (see Remark 3.2). Estimate the parameters of the model (mean and standard deviation). Perform model validation (see Chapter 3 and Section 3.1.8). Since, in this case, confidence intervals and hypothesis tests involve the distribution of an average, it might also be useful to include the central limit theorem (Theorem 3.14) in the discussion.

By using R, scripts I could see that log of transformed data is **mean = 3.21 kg /m^2** and **variance = 0.1489^2 kg /m^2**. Assuming that variables are independent and identically distributed so the distribution of logarithm observation X_1,$X2$, ...X_{145} of BMI is identically to

$$X_i \sim N(3.21, 0.1489^2)$$

So we can assume that the mean and variance that we got from the data is an approximation to the real population, it means mean and variance (by using the Central limit theorem).

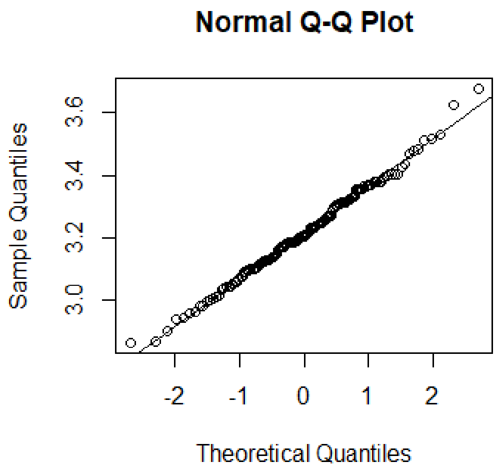


Figure 5: logarithm qq plot

Question g: *State the formula for a 95% confidence interval (CI) for the mean log-transformed BMI score of the population (see Section 3.1.2). Insert values and calculate theinterval. Then, determine a 95% CI for the median BMI score of the population(see Section 3.1.9).*

First I will start with write the formula for the 95% (CI) for (1 −α) quartile:

$$\text{(CI)} = \bar{x} + t \;\; 1 - \alpha * \frac{1}{2} * \frac{s}{\sqrt{n}}$$

I will calculate 95% CI for median when (α= 0.5) I got median by using R Script

$$\text{(CI)} = 3.21 + 1.97 * \frac{0.14}{\sqrt{145}} \sim (3.1 + 3.2)$$

Now we can see that The confidence interval of mean of the logarithmic data are (3.1 - 3.2), that mean, mean will take place in this class with 95% probability. By using exponential function exp(x) this result could be transformed and the results in confidence that median will be in (24.36-25.58)

Question h: Perform a hypothesis test in order to investigate whether the mean log-transformed

BMI score is different from log(25). This can be done by testing the following hypothesis, and corresponds to investigating whether the median BMI score is different from 25:

$$H0: \mu logBMI = \log(25)$$

$$H1: \mu logBMI \neq \log(25)$$

Using of t-statistics first I will do to calculate:

$$\boxed{t_{observation} = \frac{x-\mu 0}{\frac{s}{\sqrt{n}}}}$$

$$\boxed{t_{observation} = \frac{3.2-\log(25)}{\frac{0.1489}{\sqrt{145}}} = 0.099912}$$

By knowing Degrees of freedom 144 can I calculate the P-value:

$$\boxed{\text{P -value} = 2 * \text{pt}(-\text{abs}(t_{observation}), \text{df} = n-1) = 0.9205}$$

I got the same results from t.test in built function of r scripts, with mean as 3.2176 and hypothesis as log (25).

The significance level is α = 0.05

P-value is bigger than α → The initial assumption of null hypothesis is right.

So acceptance of null hypothesis implies average BMI is 25 which is the lower limit for over weight.

→ average person BMI is on the verge of Overweight.

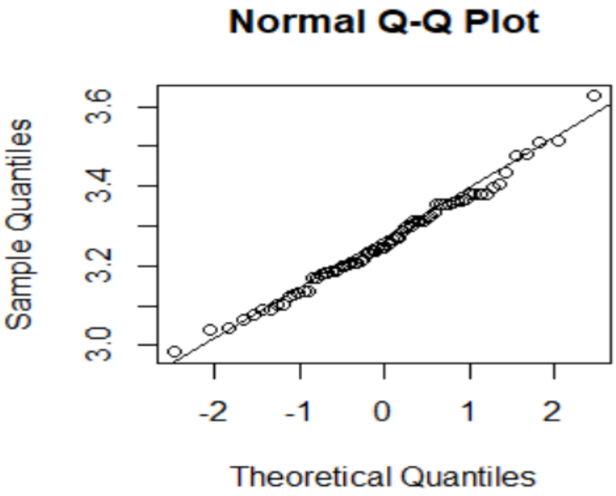


Figure 6 : Men qq plot

Question i: Specify separate statistical models for log-transformed BMI for men and women. Perform model validation for both models. Estimate the parameters of the models (mean and standard deviation for men and women, respectively).

***Start with men:** If we take a look at qq plot we can easily know that logarithmic male BMI distribution Can be assumed like normal.

And by using the R scripts we can see that **mean** and **variance** of log transformed data are **3.260588 kg /m^2 and 0.1239^2 kg^2 /m^4**. And variables are independent and identically distributed. distribution of logarithm observation X_1,X_2, ...X_{73} of BMI is identically to

$$X_i \sim N(3.26, 0.1329^2)$$

***End with women:** If we take also a look at qq plot we can easily know that logarithmic male BMI distribution Can be assumed like normal.

And by using the R scripts we can also see that **mean** and **variance** of log transformed data are **3.147 kg /m^2 and 0.1598^2 kg^2 /m^4**. And variables are independent and identically distributed. distribution of logarithm observation X_1,X_2, ...X_{72} of BMI is identically

$$X_i \sim N(3.174, 0.1598^2)$$

Question j: Calculate 95% confidence intervals for the mean log-transformed BMI score for women and men, respectively (se Section 3.1.2). Use these to determine 95% confidence intervals for the median BMI score of women and men, respectively. Fill
in the table below with the confidence intervals for the two medians.

***Start with men:**

First, I will start with write the formula for the 95% (CI) for (1 −α) quartile:

$$(\text{CI}) = \bar{x} \pm t \;\; 1 - \alpha * \frac{1}{2} * \frac{s}{\sqrt{n}}$$

I will calculate 95% CI for median when (α= 0.5) I got median by using R Script

$$(\text{CI}) = 3.2605 \pm 1.99346 * \frac{0.1239}{\sqrt{73}} \sim (3.23166 + 3.28949)$$

Now, we can see that the confidence interval of mean of the logarithmic data are (3.23166 - 3.28946), that’s mean, mean will take place in this class with 95% probability. By using exponential function exp(x) this result could be transformed and the results in confidence that median will be in
(24.36-25.58 kg /m^2).

***End with women:**

First, I will start with write the formula for the 95% (CI) for (1 −α) quartile:

$$(\text{CI}) = \bar{x} \pm t \;\; 1 - \alpha * \frac{1}{2} * \frac{s}{\sqrt{n}}$$

I will calculate 95% CI for median when (α= 0.5) I got median by using R Script

$$(\text{CI}) = 3.174 \pm 1.9939 * \frac{0.1598}{\sqrt{72}} \sim (3.13652 + 3.21169)$$

Now, we can see that the confidence interval of mean of the logarithmic data are (3.13652 - 3.21169), that's mean, mean will take place in this class with 95% probability. By using exponential function exp(x) this result could be transformed and the results in confidence that median will be in (23.02372-24.82048 **kg /m^2**)

	LOWER BOUND OF CL	UPPER BOUND OF CL
MEN	3.231667	3.289498
WOMEN	3.13652	3.21169

Table 2 Logarithmic Domain

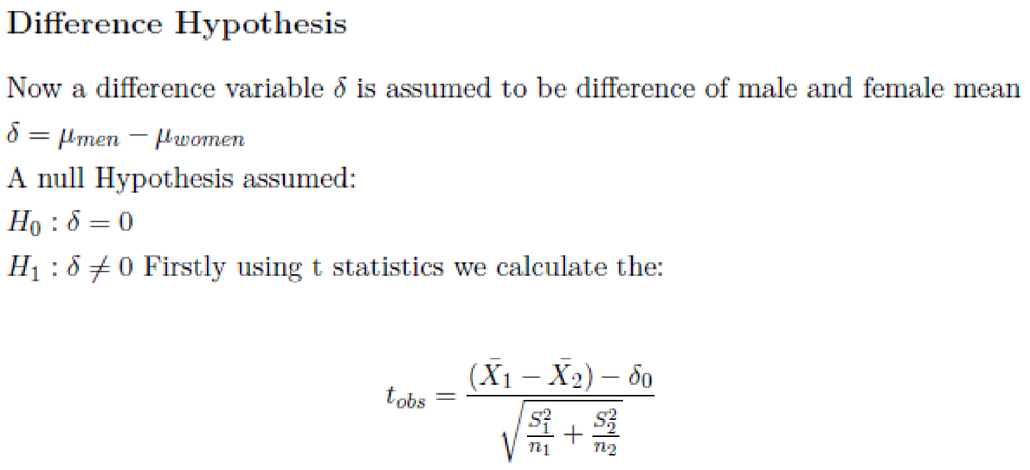
	LOWER BOUND OF CL	UPPER BOUND OF CL
MEN	25.32209	26.8292
WOMEN	23.02372	24.82048

Table 2 Real Domain

The results form R script values are the same!

Question k : Perform a hypothesis test in order to investigate whether there is a difference between

the BMI of women and men. Specify the hypothesis as well as the significance level *a*, the formula for the test statistic, and the distribution of the test statistic (remember the degrees of freedom). Insert relevant values and compute the test statistic and p-value. Write a conclusion in words.



Difference Hypothesis

Now a difference variable δ is assumed to be difference of male and female mean

$\delta = \mu_{men} - \mu_{women}$

A null Hypothesis assumed:

$H_0 : \delta = 0$

$H_1 : \delta \neq 0$ Firstly using t statistics we calculate the:

$$t_{obs} = \frac{(\bar{X_1} - \bar{X_2}) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

I get also t_{obs} from R script - 3.6429 degrees freedom

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

From it we get degrees of freedom 133.75

And by Calculating P-value = 0.000384

P-value less than α ,therefore null hypothesis is rejected. This mean that BMI values of men and women are different and not same . we got also the same result from R function t.test with two data.

Question l : Comment on whether it was necessary to carry out the hypothesis test in the previous

question, or if the same conclusion could have been drawn from the confidence

intervals alone? (See Remark 3.59).

our analysis is unnecasssary. if we take a look at the CIs for men and women we will get the same conclusion . They will not intersect, that’s mean the groups are different , the same result we got from exercise k.

Question m : State the formula for computing the correlation between BMI and weight. Insert values and calculate the correlation. Furthermore, compute the remaining

pairwise correlations involving BMI, weight and fast food. Make pairwise scatter

plots of these variables. Assess whether the relation between the plots and the

correlations is as you would expect.

sample correlation coefficient :

$$r = \frac{1}{n-1}\Sigma(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y}) = \frac{s_{xy}}{s_x * s_y}$$

Here we have S_x and S_y as standard deviations of X and Y variable and S_{xy} are variance of variables X and Y.

Scatter plots and Correlation coefficient

Scatter plot of weight and BMI:

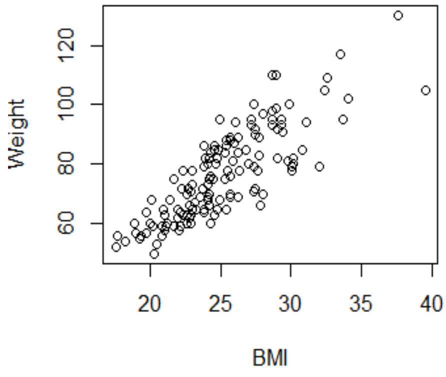


Figure 8: scatter plot weight and BMI

correlation coefficient r_{xy} = 0.82826.

correlation coefficient and scatter plot concludes that there is a positive and high correlation between BMI and weight.

Scatter plot of fast food and weight

correlation coefficient is r_{xy} = 0.2793

correlation coefficient and the scatter plot concludes that there is no correlation at all between fast food and weight.

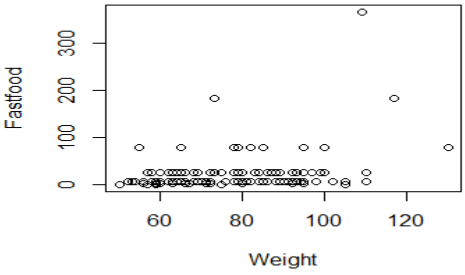


Figure 9: Scatter plot Fast food and weight

Scatter plot of fast food and BMI:

The correlation coefficient is r_{xy} = 0.1513

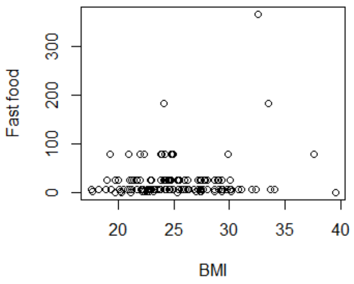


Figure 10: scatter plot fast food and BMI

The correlation coefficient and scatter plot and concludes that there is no correlation between Fast food and BMI.