BMI Survey

02402 Introduction to statistics

Author:

M V A Suhas Kumar s191382

October 22, 2019

Contents

List of Figures

List of Tables

1 Question a

1.1 Question

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?

1.2 Solution

Description of Data Set and Variables

A person's BMI (Body Mass Index) score is a measure of the person's overweight which depends on various factors like gender, age ,education . In order to analyse BMI this dataset is collected which includes features like:

- **Gender** : which corresponds to gender of respondant
- **height** : which corresponds to height of respondant in cm
- **weight** : which corresponds to weight of respondant in kgs
- **Urbanity** : which corresponds to type of area where respondant resides
- **Fast Food level** : which corresponds to no of days respondant eats fast food in a year

Quantitative and/or Categorized variables

- **Categorized Varibales**: These are the variable where variable has finite discrete values and divide the observation into groups. Here categorized variables are **Gender** and **Urbanity** because Gender could take up only 2 values either male(1) or female(0) and Urbanity here is classified in 5 classes with numbers from 1 to 5 based on the no. of inhabitants.**Fastfood** intake is categorized in a like manner, with numbers from 1 to 8
- **Quantitative Variables**: These are the variable where variable can have any number of different values. Here **hieght** and **weight** where each can any value greater than or equal to zero.

No. of observations and No. of missing values

Running the command $Dim(D)$ in R yields the size of data. **No of observation** are the number of rows in the size which is **145**
To the no. of missing values we use the command sum(is.na(D)). Running this command yielded a reult of zero. **Hence there are no missing values**

2 Question b

2.1 Question

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?

2.2 Solution

Histogram of emperical densities of BMI

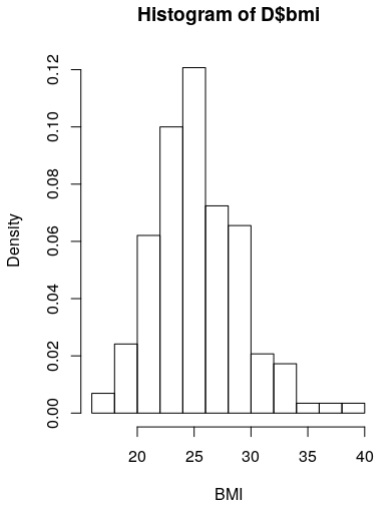

Figure 1: Emperical Density histogram plot of BMI

Can BMI score be negative?

No, BMI score can not be negative because all division of positive number($weight$) and positive number ($height^2$) can not yield a negative number

Emperical Distribution of BMI scores and Variance in Observation

The **mean,median,variance** of BMI scores that calculated from R script are **25.24795** $\frac{kg}{m^2}$**,** **24.69136** $\frac{kg}{m^2}$ **and 14.68608** $\frac{kg^2}{m^4}$ respectively.

The **min and max** values of BMI are **17.57707** $\frac{kg}{m^2}$ and **39.51974** $\frac{kg}{m^2}$

From the above calculations we could see that median is slightly left of mean by 2.2% of mean which implies the distribution **right skewed** i.e with slightly symmetric with longer right tail which could be seen from above min and max values with max value is fare from mean than min value
Also **standard deviation** of BMI is **3.83**$\frac{kg}{m^2}$ which is approximately 15% of the mean

3 Question c

3.1 Question

Make separate density histograms for the BMI scores of women and men, respec- tively. 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 gen- der difference in the distribution of the BMI scores (if so, describe the difference)?

3.2 Solution

Female BMI Histogram and analysis

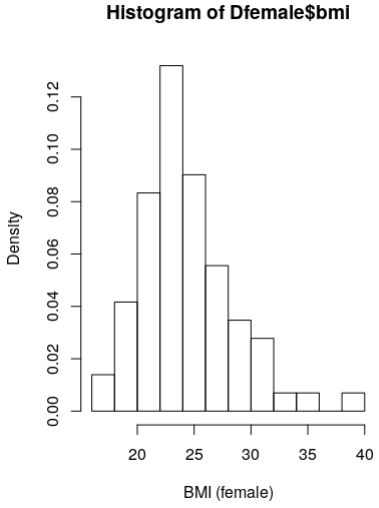

Figure 2: Emperical Density histogram plot of BMI of female

The **mean,median,variance** of female BMI scores that calculated from R script are **24.2164**$\frac{kg}{m^2}$**, 23.68911**$\frac{kg}{m^2}$ **and 16.41787** $\frac{kg^2}{m^4}$ respectively.
The **min and max** values of female BMI are **17.57707** $\frac{kg}{m^2}$ and **39.51974** $\frac{kg}{m^2}$

From the above calculations we could see that median is slightly left of mean by 2.1% of mean which implies the distribution **right skewed** i.e with slightly symmetric with longer right tail which could be seen from above min and max values with max value is fare from mean than min value
Also **Standard deviation** of female BMI is **4.05**$\frac{kg}{m^2}$ which is approximately 16.7%
On a concluding note mean of female BMI lies in normal region(<25 and >18) which implies an average female has normal weight.

Male BMI Histogram and analysis

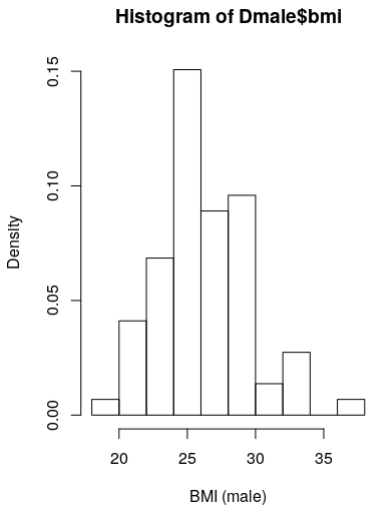

Figure 3: Emperical Density histogram plot of BMI of male

The **mean,median,variance** of male BMI scores that calculated from R script are **26.26**$\frac{kg}{m^2}$**, 25.72**$\frac{kg}{m^2}$ **and 11.068** $\frac{kg^2}{m^4}$ respectively.
The **min and max** values of male BMI are **19.75** $\frac{kg}{m^2}$ and **37.57** $\frac{kg}{m^2}$

From the above calculations we could see that median is slightly left of mean by 2.05% of mean which implies the distribution **right skewed** i.e with slightly symmetric with longer right tail which could be seen from above min and max values with max value is fare from mean than min value and also this male bmi distribution is less skewer than female.
Also **Standard deviation** of male BMI is **3.32**$\frac{kg}{m^2}$ which is approximately 12.66% which implies the distribution of male bmi is less varient than female.
On a concluding note mean of male BMI lies in over weight region(>25) which implies an average male has over weight.

4 Question D

4.1 Question

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

4.2 solution

BOX plot and observations

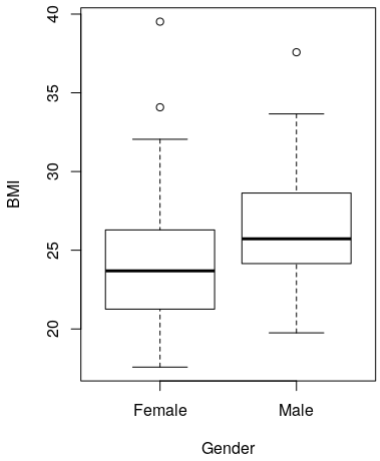

Figure 4: Box plot of BMI

We can clearly see from the box plot that both male and female distributions have slight longer tail because max is farther than min from median. From R script,IQR of **female(5.03**$\frac{kg}{m^2}$) is more than that of **male(4.48**$\frac{kg}{m^2}$) because of high variance and also female has more outliers than men (which are represented as dots in the above picture)

5 Question E

5.1 Question

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?

5.2 solution

Variable(BMI)	**No.of Obs**(n)	**Mean**	**Variance**(s^2)	**std. dev.** (s)	(Q1)	**Median**	(Q3)
Everyone	145	24.25	14.69	3.83	22.59	24.69	27.64
Women	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

Table 1: Descrpitive analysis

The above results are obtained from the R-script.
Unlike the observations from the graph we could get the precise values and variation of the data. Plots give the understanding and ables just give the values without any visualisation.

6 Question F

6.1 Question

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.

6.2 Solution

Statistical Model

Its been said that analysis of natural logarithm of bmi would yield a normal distribution. This could been seen from the qq plot below.

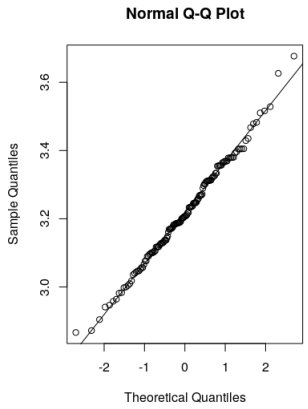

Figure 5: logarithm qq plot

From the R scripts we could see that **mean** and **variance** of log transformed data is $\mathbf{3.21}\frac{kg}{m^2}$ **and** $\mathbf{0.1489}^2\frac{kg^2}{m^4}$.Further assuming that variables are independent and identically distributed.The distribution of logarithm observation $X_1, X_2, ...X_{145}$ of BMI is identically to

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

From the Central limit theorem, we could assume that the mean and variance obtained from the data is an approximation to the real population mean and variance.

7 Question G

7.1 Question

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 the interval. Then, determine a 95% CI for the median BMI score of the population (see Section 3.1.9).

7.2 Solution

95% Confidence Interval

The formula for the 95% CI for the $(1-\alpha)$ quartile would be

$$CI = \bar{x} + t_{1-\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

Calculating 95% CI around median where $(\alpha = 0.5)$ and also obtaining median from R script

$$CI = 3.2176 + 1.9765 * \frac{0.1489}{\sqrt{145}} \sim (3.19, 3.24)$$

The 95% confidence interval of mean of the logarithmic data is 3.19-3.24,meaning true mean will lie in this class with 95% probability. This result could be transformed by using exp(x) and results in confidence that median will lie in 24.36-25.58 $\frac{kg}{m^2}$.

8 Question H

8.1 Question

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:

$$H_0 : \mu_{logBMI} = log(25)$$

,

$$H_1 : \mu_{logBMI} \neq log(25)$$

. Specify the significance level , the formula for the test statistic, as well as the distribution of the test statistic (remember to include the degrees of freedom). Insert relevant values and compute the test statistic and p-value. Write a conclusion in words. In particular, comment on whether it can be concluded that over half of the population is overweight.

8.2 Solution

Hypothesis

A null hypothesis set up :

$$H_0 : \mu_{logBMI} = log(25)$$

,

$$H_1 : \mu_{logBMI} \neq log(25)$$

. Firstly using t-statistics is used to calculate

$$t_{obs} = \frac{x - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$t_{obs} = \frac{3.2176 - log(25)}{\frac{0.1489}{\sqrt{145}}} = 0.09991274$$

Now calculating the P-value with knowning 144 degrees of freedom

$$p - value = 2 * pt(-abs(t_{obs}), df = n - 1) = 0.9205$$

The same results were obtained from t.test in built function of r scripts!! with ideal mean as 3.2176 and hypothesis μ_0 as log(25).

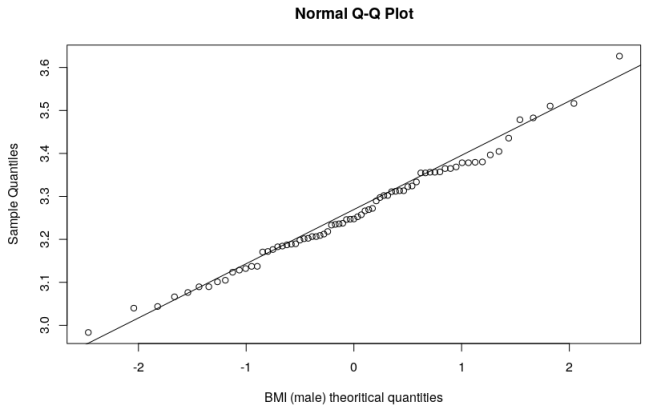

Figure 6: Men qq plot

Also significance level is set to $\alpha = 0.05$, clearly the P-value is greater than α , therefore our initial assumption of null hypothesis is true.
Hence acceptance of null hypothesis implies average BMI is 25 which is the lower limit for over weight which concludes that average person BMI is on the verge of Overweight.

9 Question I

9.1 Question

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).

9.2 Soution

For men.

From the above male qq plot we could accept that logarithmic male BMI distribution could be assumed as normal.
From the R scripts we could see that **mean** and **variance** of log transformed data is $\mathbf{3.260588}\frac{kg}{m^2}$ **and** $\mathbf{0.1239}^2\frac{kg^2}{m^4}$.Further assuming that variables are independent and identically distributed.The distribution of logarithm observation $X_1, X_2, ...X_{73}$ of BMI is identically to

$$X_i \sim N(3.260588, 0.1239^2)$$

For Female

From the above female qq plot we could accept that logarithmic male BMI distribution could be assumed as normal.

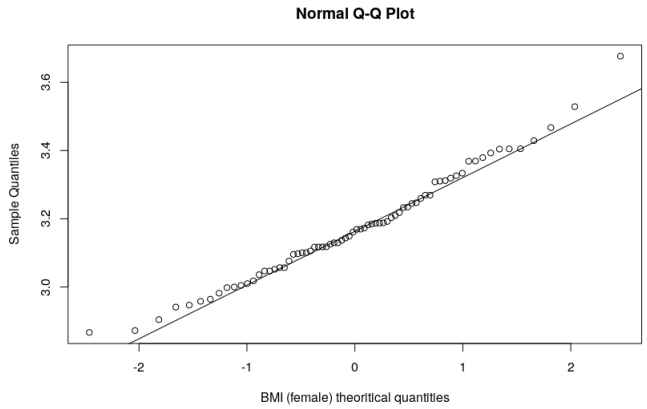

Figure 7: Female QQ plot

From the R scripts we could see that **mean** and **variance** of log transformed data is $\mathbf{3.174}\frac{kg}{m^2}$ **and** $\mathbf{0.1598}^2\frac{kg^2}{m^4}$.Further assuming that variables are independent and identically distributed.The distribution of logarithm observation $X_1, X_2, ...X_{72}$ of BMI is identically to

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

10 Question J

10.1 Question

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% con- fidence intervals for the median BMI score of women and men, respectively. Fill in the table below with the confidence intervals for the two medians.

10.2 Solution

For Men

The formula for the 95% CI for the $(1-\alpha)$ quartile would be

$$CI = \bar{x} \pm t_{1-\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

Calculating 95% CI around median where $(\alpha = 0.5)$ and also obtaining median from R script

$$CI = 3.260588 \pm 1.993464 * \frac{0.1239}{\sqrt{73}} \sim (3.231667, 3.289498)$$

The 95% confidence interval of mean of the logarithmic data is 3.231667 - 3.289498,meaning true mean will lie in this class with 95% probability. This result could be transformed by using exp(x) and results in confidence that median will lie in 25.32209 -26.8292 $\frac{kg}{m^2}$.

For Female

The formula for the 95% CI for the $(1-\alpha)$ quartile would be

$$CI = \bar{x} \pm t_{1-\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

Calculating 95% CI around median where ($\alpha = 0.5$) and also obtaining median from R script

$$CI = 3.174 \pm 1.993943 * \frac{0.1598}{\sqrt{72}} \sim (3.13652, 3.21169)$$

The 95% confidence interval of mean of the logarithmic data is 3.13652 - 3.21169,meaning true mean will lie in this class with 95% probability. This result could be transformed by using exp(x) and results in confidence that median will lie in 23.02372 -24.82048 $\frac{kg}{m^2}$.

Tables

-	**Lower bound of CI**	**Upper Bound of CI**
Men	3.231667	3.289498
Women	3.13652	3.21169

Table 2: In logarithmic Domain

-	**Lower bound of CI**	**Upper Bound of CI**
Men	25.32209	26.8292
Women	23.02372	24.82048

Table 3: In Real Domain

We could clearly see that from the results of R script that inbuilt function and calculated values are exactly the same!

11 Question K

11.1 Question

Perform a hypothesis test in order to investigate whether there is a difference be- tween the BMI of women and men. Specify the hypothesis as well as the significance level , 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.

11.2 Solution

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:

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

we get t_obs from R script as -3.6429 with degrees of freedom

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

From above we get degrees of freedom as 133.75
Calculating P-value = 0.000384
we could clearly see that P-value less than α ,Hence null hypothesis is rejected. This implies that BMI values of men and women are not same. Also we got the same result in-built R function t.test with two data inputs.

12 Question L

12.1 Question

Comment on whether it was necessary to carry out the hypothesis test in the pre- vious question, or if the same conclusion could have been drawn from the confi- dence intervals alone? (See Remark 3.59).

12.2 Solution

The above analysis is unnecasssary.The same conclusion cis obtained by looking at the CIs for men and women. They do not intersect, which concludes that the two groups are signicantly dierent, same as the result found in exercise k.

13 Question M

13.1 Question

State the formula for computing the correlation between BMI and weight. In- sert 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.

13.2 Solution

The sample correlation coefficient is

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

where s_x and s_y are standard deviations of X and Y variable respectively and s_{xy} is co variance of variables X and Y

Scatter plots and Correlation coefficient

Scatter plot of weight and BMI:

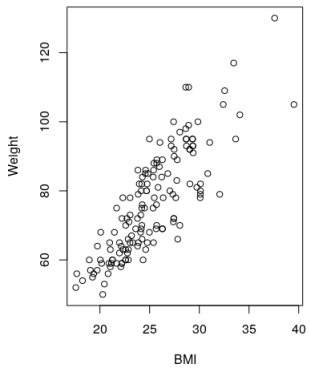

Figure 8: scatter plot weight and BMI

The correlation coefficient is $r_{xy} = 0.82826$.
The scatter plot and the correlation coefficient concludes that there is a positive and high correlation between weight and BMI.

Scatter plot of fast food and weight

The correlation coefficient is $r_{xy} = 0.2793$

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

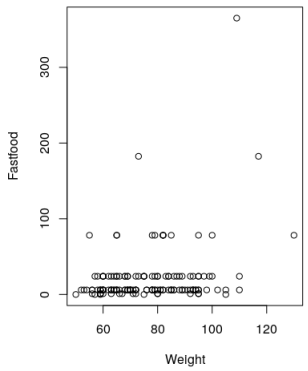

Figure 9: Scatter plot Fast food and weight

Scatter plot of fast food and BMI

The correlation coefficient is $r_{xy} = 0.1513$

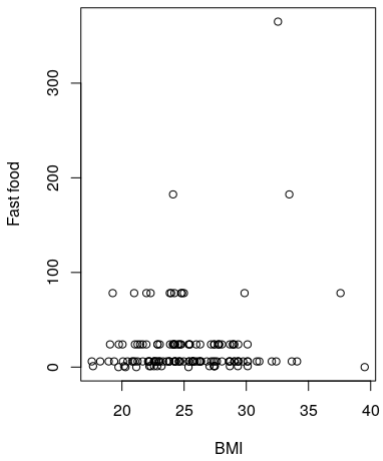

Figure 10: scatter plot fast food and BMI

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