Introduction to Statistics
Spring 2016

Project 1: BMI Survey

Rahul Sharma

March 12th, 2016

Student ID s155316
Instructor .... Per Bruun Brockhoff

1 Introduction

The incidence of the overweight issue in Denmark is a gradually increasing problem and can result in a whole list of consequences for the Danish population. Increasing hypertension, heart disease, and general unhealthiness can significantly impact society economically, militarily, or even in national productivity. Being American myself, I can relate to these issues as the United States has consistently ranked in the top echelon for percentage of people who are overweight. In order to better understand the causes and effects of this epidemic, we can study the trends of body mass index (BMI) in a statistical sense. BMI is the measure of the level of body fat as it relates to an individual's weight and height and for our purposes, is the focus of our study.

In order to measure these effects we will take a look at BMI as it pertains to the ulterior factors of age, gender, urbanity, education, number of children and so on with data that are "a small part of a larger survey on these issues." By carrying out a variety of statistical studies, we hope to narrow down several of the correlations between BMI and the factors while simultaneously analyzing several of them for statistical significance to begin to understand this ever increasing issue in Denmark. Of course our methods are reliant on our data and we will assume for the purpose of this study that the data are a simple random sample taken out of the Danish population at large.

2 Descriptive Analysis

a) Our data-set has a total of 847 observations, or people who have responded to the survey with the following variables measured: Gender, Age, Height, Weight, Education, Urbanity, FastFood, HouseholdIncome, Alone, Children, and Labeling. On top of this we contrived two extra variables, namely BMI with formula

$$\frac{Weight}{Height^2}$$

and FastFood_k which was a recode with more appropriate values for quantitative analysis.

The majority of our data is categorical - e.g. Gender is either a 0 for female or 1 for male which leads to a lack of meaningful statistical analysis on these specific variables unless paired with a quantitative variable such as BMI. Hence, the quantitative data, namely, BMI, Weight, Height, and the recoded Fastfood_k, will be the meat of our data-set in analysis.

The mean of our BMI is 0.002557 and will not be reclassified to the standard measures of BMI which are on a scale of 10^1. The variance is 1.7788×10^{-7}.

Lastly, as far as data quality is concerned, our data-set is largely complete as any factors that could be considered are accounted for although not tested in this particular study. Even though there could be another variable such as amount of exercise, the statistical effect for our purposes would be fairly obvious.

b) Quantitative data with associated boxplots are as follows:

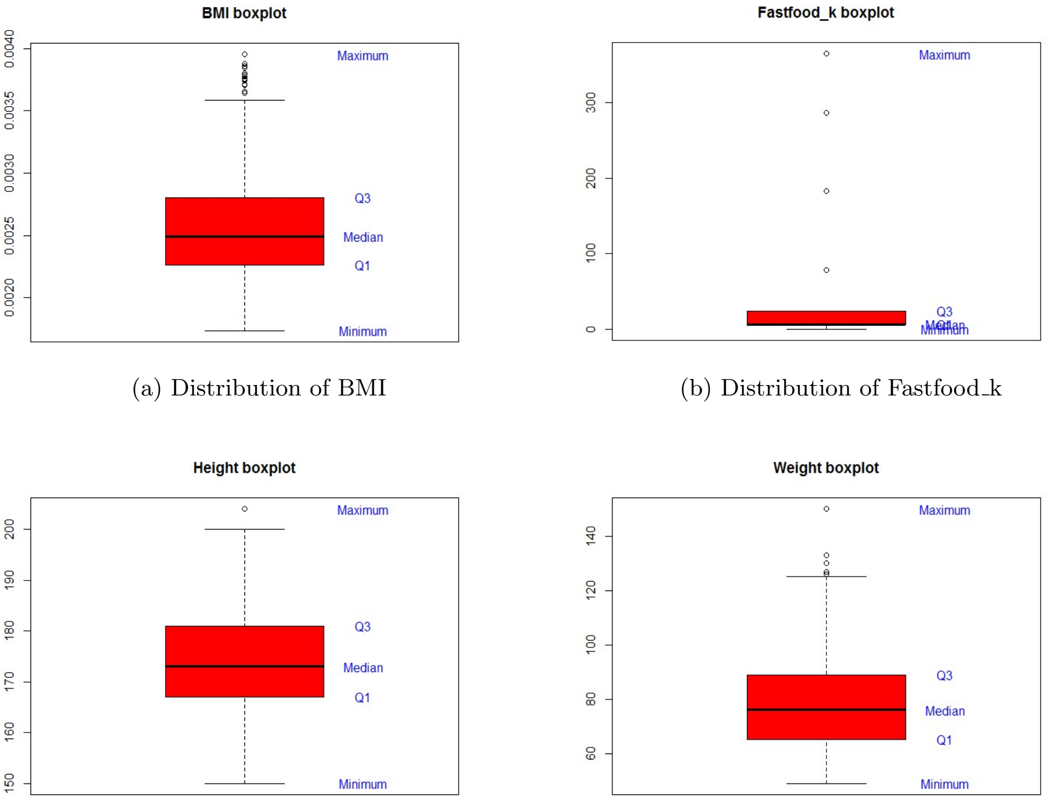
Label	n	mean	var	sd
BMI	847	2.557e-03	1.778e-07	4.217e-04
Weight	847	7.791e+01	2.568e+02	1.602e+01
Height	847	1.741e+02	8.732e+01	9.345e+00
Fastfood_k	847	1.904e+01	1.066e+03	3.265e+01

Table 1: Summary of selected variables

Label	0%	25%	50%	75%	100%
BMI	1.736e-03	2.263e-03	2.493e-03	2.804e-03	3.952e-03
Weight	4.900e+01	6.500e+01	7.600e+01	8.900e+01	1.330e+02
Height	1.500e+02	1.670e+02	1.730e+02	1.810e+02	2.040e+02
Fastfood_k	0.000e+00	6.000e+00	6.000e+00	2.400e+01	3.650e+02

Table 2: Percentile values of selected variables

c) For BMI, the mean is 2.557×10^{-3} and the standard deviation is 4.217×10^{-4}. We can see from the plot that the data are relatively evenly spread out and the interquartile range (IQR) is reasonably sized. This is a positive indication for the use of the assumption of the normal distribution however more tests will be required. There are several outliers beyond .0035 which might slightly skew the data to the right but aren't of too much consequence considering that there are 847 samples. Our values fall between $[0.001736111, 0.003951974]$.



(a) Distribution of BMI

(b) Distribution of Fastfood_k

(c) Distribution of Height

(d) Distribution of Weight

Figure 1: Boxplots for respective values shown in Table 1

With the Fastfood_k boxplot, we can see a strong skew towards the right with the IQR being located in the extreme left of the diagram. This indicates that if necessary, a log normal distribution might be required to model the data. Our values fall between $[0, 365]$.

The Height boxplot is the most evenly distributed with no significant skew and very few outliers or in other words, symmetrical. The IQR is clearly around the middle of the data set with a mean of 1.741×10^2 and variance of 8.732×10^1 which confirms our observations that the data are evenly spread out. Our values fall between $[150, 204]$.

Our last distribution is that of weight boxplot shown in figure D and this is largely similar to the distribution of the Height as shown, symmetrical. There is a large IQR representative of the values being relatively close to the mean of 7.791×10^1. The variance for this variable is 2.568×10^2. This is slightly skewed right by several outliers but it isn't a strong skew by any means. Our values fall between $[49, 133]$.

All of our plots show reasonable data and distributions, according to all the averages, the average person in Denmark has a height of 174 cm, a weight of approximately 78 kg, and a resulting BMI of .00255 or BMI index of 25 indicating that the person is moderately overweight.

3 Problem 1

d) Our model for the approximation of the data set for BMI is the normal distribution with

$$X_i \sim N\Big(2.557\,302 \times 10^{-3}, \big(4.217\,671 \times 10^{-4}\big)^2\Big)$$

based on the results of the QQ plot of BMI shown in figure 2. In this graph we can see that the majority of the data, approximately 99%, falls within 3 standard deviations of the mean, which suggests that a normal distribution can be used. In figure 2, the straight line represents the theoretical, default QQ plot and we can observe that the data from our 'qqnorm' command is reasonably close to this line which confirms the use of the normal distribution.

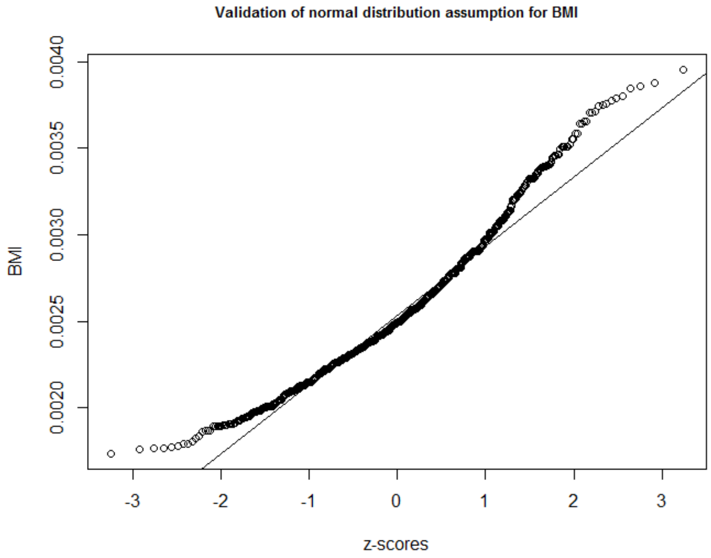


Figure 2: QQ plot of BMI data with Normal line

e) 95% confidence interval for mean of BMI

1. The population of interest is the values of the BMI index reported in the survey with a sample size of n = 847, sample mean of .0025, and sample standard deviation of .00042 and we are testing for true mean of the BMI of the population.
2. The sample is randomly selected as provided from our problem, the samples are independent of each other since they are all distinct people replying to a survey, and as already proven on the QQ plot, our sample can be reasonably modeled by a normal distribution. Hence, we will use a 95% confidence z-interval test with formula :

$$\mu \in \left[\bar{x} - t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]$$

3. We will be using 846 degrees of freedom since there are 847 samples
Plugging in values : $t_{.975} = 1.96$ using 'qt' command in R

   95% CI $= .0025 \pm 1.96 \cdot \frac{0.000421}{\sqrt{847}} = \big(0.00252, 0.00258\big)$

4. We are 95% confident that the true mean of the BMI of the sampled population is somewhere between .00252 and .00258.

95% confidence interval for variance of BMI

1. The population of interest is the values of the BMI index reported in the survey with a sample size of n = 847, sample mean of .0025, and sample standard deviation of .00042 and we are testing for true variance of the BMI of the population.
2. The sample is randomly selected as provided from our problem, the samples are independent of each other since they are all distinct people replying to a survey, and as already proven on the QQ plot, our sample can be reasonably modeled by a normal distribution. We will use the 95% confidence interval formula for variances :

$$\sigma^2 \in \left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \quad \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \right]$$

3. We will be using 846 degrees of freedom since there are 847 samples
Plugging in values: $\chi^2_{.025} = 767.2897$ and $\chi^2_{.975} = 928.4983$

$$\begin{aligned} 95\% \text{ CI} &= \left[\frac{846 \cdot 0.000421^2}{767.2897}, \frac{(846 \cdot 0.000421^2}{928.4983} \right] \\ &= [1.620 \times 10^{-7}, 1.961 \times 10^{-7}] \end{aligned}$$

4. We are 95% confident that the true variance the BMI is somewhere between 1.620×10^{-7} and 1.961×10^{-7}.

f) Testing whether the mean BMI is greater than .0025

1. We are testing the mean BMI for whether or not it is greater than .0025, using an α-level of .05 and 846 degrees of freedom. Our relevant hypotheses are as follows:

$$\begin{aligned} H_0 &: \mu = .0025 \\ H_a &: \mu > .0025 \end{aligned}$$

2. The appropriate procedure is a one-tailed t-test for means and the conditions for this test, namely, normal, independent sampling, and random are all met as outlined previously. However for thoroughness the justifications are T̈he sample is randomly selected as provided from our problem, the samples are independent of each other since they are all distinct people replying to a survey, and as already proven on the QQ plot, our sample can be reasonably modeled by a normal distribution."
3. Formula for computing t_{obs}:

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

Plugging in values:

$$\begin{aligned} t_{obs} &= \frac{0.002557302 - .0025}{0.0004217671/\sqrt{847}} \\ &= 3.954 \end{aligned}$$

Computing p-value:

$$\begin{aligned} p - value &= 2 \cdot P(T > |t_{obs}|) \\ &= 2 \cdot P(T > 3.954) \\ &= 4.163 \times 10^{-5} \end{aligned}$$

4. Because our p-value of 4.163×10^{-5} is less than our stated α-level of .05, we reject the null hypothesis. The data provide convincing statistical evidence that the mean of BMI for our population is greater than .0025.

4 Problem 2

g) Testing gender based differences in BMI

1. We are testing the mean BMIs of the female and male responses against each other for whether or not they are equal, using an α-level of .05 and 844.99 degrees of freedom (calculated below). Our relevant hypotheses are as follows:

$$\begin{aligned} H_0 &: \mu_{male} = \mu_{female} \\ H_a &: \mu_{male} \neq \mu_{female} \end{aligned}$$

2. The appropriate procedure is a two-sample t-test for means and the conditions for this test, namely, normal, independent sampling, and random are all met as outlined previously. However for thoroughness the justifications are "The sample is randomly selected as provided from our problem, the samples are independent of each other since they are all distinct people replying to a survey, and as already proven on the QQ plot, our sample can be reasonably modeled by a normal distribution." Since both the male and female responses are subsets of our larger data-set, the same conditions still apply.

3. Relevant values specified below:

   Gender 1 = female, Gender 2 = male
   $\mu_1 = 0.002488228$, $s_1 = 0.0004359721$, $n_1 = 447$
   $\mu_2 = 0.002634493$, $s_2 = 0.0003916573$, $n_2 = 400$

   Computing degrees of freedom ν :

$$\begin{aligned} \nu &= \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} \\ &= \frac{(0.000435972^2/447 + 0.000391657^2/400)^2}{(0.000435972^2/447)^2/(447-1) + (0.000391657^2/400)^2/(400-1)} \\ &= 844.99 \end{aligned}$$

   Computing t_{obs} :

$$\begin{aligned} t_{obs} &= \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \\ &= \frac{.002488228 - .0004359721}{\sqrt{\frac{0.0004359721^2}{447} + \frac{0.0003916573^2}{400}}} \\ &= -5.1433 \end{aligned}$$

Computing p-value:

$$\begin{aligned} p - value &= 2 \cdot P(T > |t_{obs}|) \\ &= 2 \cdot P(T > -5.1433) \\ &= 3.357 \times 10^{-7} \end{aligned}$$

4. Because our p-value of 3.357×10^{-7} is less than our stated α-level of .05, we reject the null hypothesis. The data provide convincing statistical evidence that the true difference between mean BMI in gender is not equal to 0.

h) Testing urbanity based differences in BMI

1. We are testing the mean BMIs of the highest level in a particular urbanity against the lowest level for whether or not they are equal, using an α-level of .05 and 95.599 degrees of freedom (calculated below). Our relevant hypotheses are as follows:

$$\begin{aligned} H_0 &: \mu_{urban_{low}} = \mu_{urban_{high}} \\ H_a &: \mu_{urban_{low}} \neq \mu_{urban_{high}} \end{aligned}$$

2. The appropriate procedure is a two-sample t-test for means and the conditions for this test, namely, normal, independent sampling, and random are all met as outlined previously. However for thoroughness the justifications are ”The sample is randomly selected as provided from our problem, the samples are independent of each other since they are all distinct people replying to a survey, and as already proven on the QQ plot, our sample can be reasonably modeled by a normal distribution.” Since the two urbanities are subsets of our larger data-set, the same conditions still apply.

3. The highest mean BMI is in urbanity 1 and is referred to by subscript 1 and the lowest mean BMI is in urbanity 5 and is referred to by subscript 2.

   Relevant values specified below:
   $\mu_1 = 0.002645397$, $s_1 = 0.0004333438$, $n_1 = 72$
   $\mu_2 = 0.002491761$, $s_2 = 0.0004063971$, $n_2 = 388$

   Computing degrees of freedom ν :

$$\begin{aligned} \nu &= \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} \\ &= \frac{(0.0004333438^2/72 + 0.0004063971^2/388)^2}{(0.0004333438^2/72)^2/(72-1) + (0.0004063971^2/388)^2/(388-1)} \\ &= 95.599 \end{aligned}$$

   Computing t_{obs} :

$$\begin{aligned} t_{obs} &= \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \\ &= \frac{0.002645397 - 0.002491761}{\sqrt{\frac{0.0004333438^2}{72} + \frac{0.0004063971^2}{388}}} \\ &= 2.7893 \end{aligned}$$

Computing p-value :

$$\begin{aligned} p - value &= 2 \cdot P(T > |t_{obs}|) \\ &= 2 \cdot P(T > 2.8793) \\ &= 0.006375 \end{aligned}$$

4. Because our p-value of 0.006375 is less than our stated α-level of .05, we reject the null hypothesis H_0. The data provide convincing statistical evidence that the true difference between highest mean BMI based on urbanity and the lowest is not equal to 0.

5 Problem 3

Label	Age	Height	Weight	Fastfood_k	BMI
Age	1.00000000	-0.03154333	0.1161790	-0.28788580	0.16835262
Height	-0.03154333	1.00000000	0.6026076	0.15725390	0.09141162
Weight	0.11617903	0.60260764	1.0000000	0.12882564	0.84509068
Fastfood_k	-0.28788580	0.15725390	0.1288256	1.00000000	0.05810985
BMI	0.16835262	0.09141162	0.8450907	0.05810985	1.00000000

Table 3: Correlations between all relevant variables

i) The formula for correlation is:

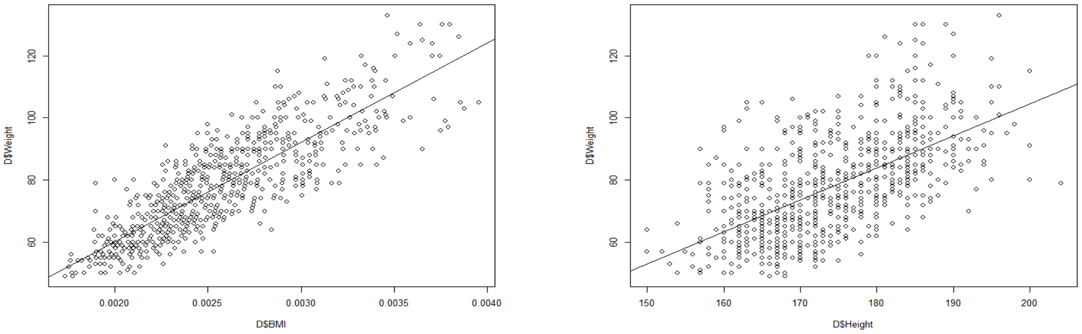
$$\hat{\rho} = \frac{1}{n-1} \sum_{i=1}^{N} \Big(\frac{x_i - \bar{x}}{s_x}\Big)\Big(\frac{y_i - \bar{y}}{s_y}\Big)$$

Since we have 847 data points, showing a calculation for correlation is difficult and a sample calculation using BMI as 'x' and Weight as 'y' is shown below (also shown in Figure 3a) :

$$\begin{aligned} \hat{\rho} = \frac{1}{847-1}\Big(\frac{0.001736111 - 0.002557302}{0.0004217671}\Big)\Big(\frac{49 - 77.91499}{16.02487}\Big) + \dots \\ + \frac{1}{847-1}\Big(\frac{0.003951974 - 0.002557302}{0.0004217671}\Big)\Big(\frac{133 - 77.91499}{16.02487}\Big) \end{aligned}$$

This results in the correlation value .8409068 shown in table 3 for BMI vs Weight which as shown in Figure 3a, is reasonably correlated. The majority of the correlations in our table were not very high as expected as factors such as Age and Weight or Fastfood_k and Height had little to no correlations which makes sense in the real world considering that anyone can really be any weight regardless of age. Similarly, frequency of eating fast food has little bearing on an individuals height. The two that did show a reasonably high level of correlation are also results that we expected. BMI is highly correlated with weight because as the two have a more or less linear relationship in our previously defined equation. As for Weight and Height, the taller a person is, the heavier they generally are as a result of that extra height.

Essentially, there are no meaningful correlations or surprising values that jump out at us based on the results of table 3.



(a) BMI vs Weight

(b) Height vs Weight

Figure 3: Scatterplots with added regression line

6 Conclusion

Understanding the factors and effects of the gradual increase of overweight people is a difficult task even with the use of statistics. Of surveyed individuals, one of the most meaningful conclusions that we drew is that the mean BMI is greater than .0025 which according to the supplied chart, means that the majority of people are more than ’moderately overweight.’ As far as the factors are concerned, the two tested variables were urbanity and gender.

For urbanity, we found that the difference between the location of highest surveyed mean BMI and lowest was statistically significant. The highest mean BMI was located in urbanity 1 which represents localities outside urban areas while the lowest was located in urbanity 5 which represents large cities with over 100,000 inhabitants. Ultimately, this significance can be interpreted to mean that rural areas will likely have more overweight individuals than cities, however because our sample size was significantly smaller in the rural areas (72 responses), the results have to be taken with a grain of salt.

As far as gender differences the results were as we expected, there is a statistically significant difference between the mean BMI of the two genders. Due to biological factors such as body composition and other physical differences, this is a reasonable conclusion to draw. Since both sides saw a large sample size, the effects here are likely reliable and conclusive.

Our last test was looking through the correlations between our four most used variables - BMI, Weight, Height, and Fastfood_k. As stated earlier, no meaningful conclusions could be drawn beyond the obvious such as BMI vs Weight or Height vs Weight which are intuitive correlations.

Although not many conclusions were drawn overall, understanding the fundamental aspects of the issue was the goal of this study, and all in all we achieved this aim. Moving forward, testing multiple factors beyond the aforementioned four will likely lead to interesting statistical effects that weren’t tested for in this particular study.