Angela Boakye Danquah
Mike Wood

Statistical Analysis of Average BMI in the Americas

Introduction

Worldwide trends in weight and overall health has become a primary focus of public health scientists globally. Historically, the variable Body Mass Index (BMI) is used in medicine to determine and classify weight categories. BMI is a person's weight in kilograms divided by the square of their height in meters. According to the CDC, a high BMI can be an indicator of high body fatness. BMI can be used to screen for weight categories that may lead to health problems. Average BMI differs across countries, and disparities may be explained by regional differences in exercise and eating habits. To determine this, we focus on the BMIs of the Americas. Our analysis aims to determine if the mean BMI for countries is different across four geographic regions, North America, Central America, the Caribbean, and South America. Specifically, we aim to determine if the average BMI of a country increases with latitude, meaning that the South American countries have the lowest mean BMIs, followed by Central America, the Caribbean, and North America Secondly, we aim to determine between which regions these differences exist, if they exist. Lastly, we seek to discover whether the incidence of overweight mean country BMI differs across regions.

Data Summary and Discussion

The World Health Organization (WHO) publishes estimations of the mean BMI of adults, individuals aged 18 years or older. We obtained the mean BMI data for all countries in the Americas from the most recent WHO estimates published in 2014. We then divided the estimates by region, North America, Central America, the Caribbean, and South America. We believe our explanatory variable to be region, defined as the four regions mentioned, and our explanatory variable to be average BMI, the average BMI of adults within a country as reported by the World Health Organization. The data for South America has a sample size of 12, for Central America the same size is 7, for the Caribbean the sample size is 13, and for North America the sample size is 3. The overall sample size is small; therefore, caution must be taken in the selection of our tests and the conclusions we make about the data. A histogram of the variable BMI (see appendix,

Histogram 1) shows that the data is heavy tailed and slightly skewed left, meaning that there are fewer observations for smaller BMIs and that there is a relatively higher concentration of large BMIs as compared with moderate BMIs.

Methodology

To begin, we asked the question: Is there a difference in the mean country BMI across regions? To answer this question, we applied the Jonckheere-Terpstra test, which is a test used to determine whether there is a significant difference in mean between samples when the difference is believed to be ordered, as in the mean of sample one is believed to be less than the mean of sample 2 and the mean of sample 2 is believed to be less than the mean of sample 3 etc. Compared to other tests that determine if there is a significant difference in mean for multiple samples, which includes the ANOVA F test, the Permutation F test, and the Kruskal-Wallace test, the Jonckheere-Terpstra test has proven to be more powerful than these tests when there exists an ordered hypothesis. Because our hypothesis is ordered (we believe the mean of country BMIs in North America is greater than the mean of country BMIs in Central America etc.), the Jonckheere-Terpstra test is the best method to test this hypothesis.

The Jonckheere-Terpstra test is performed under the assumption that the samples have the same mean, meaning that the difference between the location parameters, in this case the mean of the average BMI of each region, is 0. We label the data for South America as sample 1, the data for Central America as sample 2, the data for the Caribbean as sample 3, and the data for North America sample as 4. Therefore, the null hypothesis is: $H_0: F_1(x) = F_2(x) = F_3(x) = F_4(x)$. Because we believe the average BMI is higher for regions in the north as compared with regions in the south we have the alternative hypothesis: $H_a: F_1(x) \geq F_2(x) \geq F_3(x) \geq F_4(x)$, with at least one strict inequality.

Next, we analyzed the data to identify between which regions these differences in mean exist. The Jonckheere-Terpstra test allowed us to determine if there exists a difference in mean, but it does not determine where the differences occur between the samples. To determine this, we used the Wilcoxon Rank Sum test. The Wilcoxon rank sum test must be applied to each pair of samples

in our data; therefore, we will repeat the test 6 times. Compared with other tests for differences in mean including the T-test, the Permutation Test for Difference in Mean, and the Mann Whitney test, the Wilcoxon Rank Sum test has greater power for distributions that are heavy tailed and skewed. The histograms created for each pair of samples (see appendix) shows that between most groups the data is skewed or heavy tailed. Therefore, the Wilcoxon rank sum test is the best test to use in this case.

The Wilcoxon Rank Sum test is performed assuming that the difference between the means of two samples is 0. Therefore our null hypotheses are $H_0: F_1(x) = F_2(x), F_1(x) = F_3(x), F_1(x) = F_4(x), F_2(x) = F_3(x), F_2(x) = F_4(x), F_3(x) = F_4(x)$. Because we believe average BMI is larger for regions in the north. Our alternative hypotheses are $H_a: F_1(x) \geq F_2(x), F_1(x) \geq F_3(x), F_1(x) \geq F_4(x), F_2(x) \geq F_3(x), F_2(x) \geq F_4(x), F_3(x) \geq F_4(x)$. The Wilcoxon Rank Sum tests assigns ranks to each of the observations within the two samples. The ranks of the first sample are then summed to determine the value of the test statistic, w. The P-value is determined by permuting the observations between the two samples and determining the probability that the test statistic will be as extreme as the test statistic our data produced.

Lastly, we analyzed the provided data to determine whether the incidence of overweight mean country BMI differs across regions. From the data, we find that the countries with an overweight average BMI are Argentina, Chile, Ecuador, Suriname, and Venezuela in South America, Belize El Salvador, and Panama in Central America, Antigua and Barbuda, the Bahamas, Barbados, Dominica, Grenada, Jamaica, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, and Trinidad and Tobago in the Caribbean and Canada, the U.S. and Mexico in North America. We used the Permutation Chi-Squared test, which is a test used to determine if there is a significant association between two categorical variables. In this case, the two variables are overweight BMI with categories “yes” and “no” and region with categories North America, South America, the Caribbean, and Central America (See Table 1 in appendix). Compared with other tests for two way tables, which includes the parametric Chi-Squared test and Fisher’s exact test among others, the Permutation Chi-Squared test is most appropriate in the cases in which there are

more than 2 categories for one variable and when the sample size is small. Therefore, the Permutation Chi-Squared test is the best method for our data.

The null hypothesis for this test is H_0: $p_{i|j} = p_{i|j*}$(where *i* refers to a given row and *j* refers to a given column). In other words, there is no association between the incidence of overweight mean country BMI and region. The alternate hypothesis is H_a: $p_{i|j} \neq p_{i|j*}$, meaning there is an association between the incidence of overweight mean country BMI and region. The Permutation Chi-Square Test is completed by performing the following steps. First, calculate the Chi-Square test statistic of the observed data in the table. The Chi-Square test statistic is calculated using this formula: : $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}$. This means that for each cell in the table, calculate the value in the table (n_{ij}) minus the expected value for that cell ($E_{ij} = (n_i * n_j)/n$), all squared, divided by the expected value (E_{ij}). The sum of these numbers is the Chi-squared statistic. In order to obtain the p-value for the test we found all possible permutations of the responses across the categories of the explanatory variable (in this case, mean BMIs across regions), calculated the Chi-Square test statistic for each permutation, and then calculate the fraction of test statistics that are more extreme than the observed one. This fraction is the p-value, or the probability that a value more extreme than observed will occur. Since there are so many permutations for this test, we took a random sample of 2000 permutations which produces an almost identical p-value.

Results

- Jonckheere-Terpstra test
  - Using R, we found that the test statistic for our Jonckheere-Terpstra test is JT = 302. The p-value associated with the test is p = 0.0028.
- Wilcoxon Rank Sum tests
  - South America v. Central America: The test statistic is w = 28.5 with an associated p-value of p = 0.1352.
  - South America v. Caribbean: The test statistic is w = 40.5 with associated p-value p = 0.02186

- South America v. North America: The test statistic is w = 3.5 with associated p-value p = 0.021.
- Central America v Caribbean: The test statistic is w = 32.5 with associated p-value p = 0.1606.
- Caribbean V North America: The test statistic is w = 16 with associated p-value p = 0.3428.
- Central America V North America: w = 4 with associated p-value p = 0.08508.

- Permutation Chi-Squared Test
  - Using R, we found that the test statistic for our Permutation Chi-Square Test is χ^2 = 6.089 with associated p-value of p = 0.1074.

Discussion

This data analysis process began by considering the questions posed to us. We inspected the data and noted the total sample size, the sample sizes by region, and the possibility of some observations being outliers. We then went about critically choosing appropriate tests based on our questions and the characteristics of our data. The p-value for our Jonckheere-Terpstra test is p = .0028. At significance level α = .05, we reject the null hypothesis. The test concludes that there is evidence of a significant difference in mean in the increasing order we predicted. For our Wilcoxon Rank Sum tests, we found p-values of p = 0.1352. for the test between South America and Central America, p = 0.1606 for the test between Central America and the Caribbean, p = 0.3428 for the test between the Caribbean and North America, and p = 0.08508. This means that for each of these tests we fail to reject the null hypotheses and that the tests do not conclude that there is a significant difference in the means of average BMI between each of these pairs. We also found p-values of p = 0.02186 for the test between South America and the Caribbean, and p = 0.021 for the test between South America and North America. For these tests, at significance level α =.05, we reject the null hypotheses and conclude that there is a significant difference in the means of average BMI between these pairs. The Chi-Squared test produced a p-value of p = 0.1074, which means, at significance level α = .05, we fail to reject the null hypothesis and the test

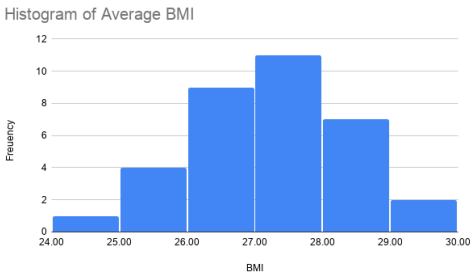
does not conclude that there is a significant association between incidence of overweight BMI and region.

Conclusions

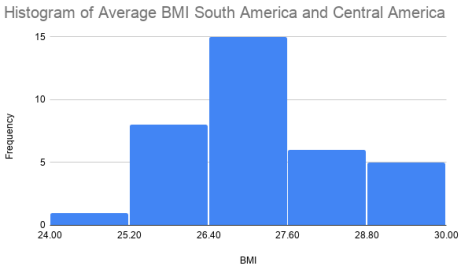
Using various statistical analysis approaches we have reached the following conclusions. Firstly, we have determined that there is a significant difference in the means of average BMI between the four regions of the Americas. We have also found that this difference increases with latitude, meaning that southern regions have lower average BMIs than northern regions. Secondly, we have found that the means of average BMI in North America and the Caribbean are significantly larger than the mean of average BMI in South America, but between the other pairs of regions (North America v. Central America, etc.) we cannot conclude that there is a significant difference. Lastly, we found that we cannot conclude that there is a significant association between the incidence of overweight average BMIs and region.

Appendix

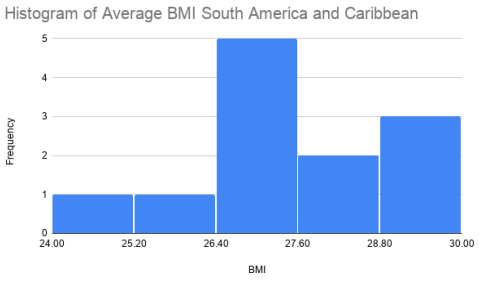
Histogram 1 (Histogram of all samples)



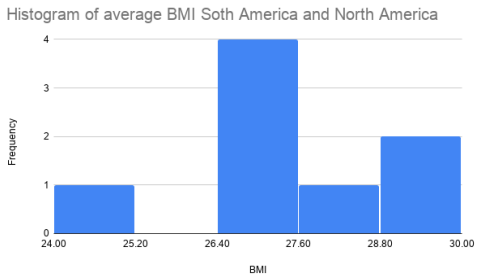
Histogram 2 (South America and Central America)



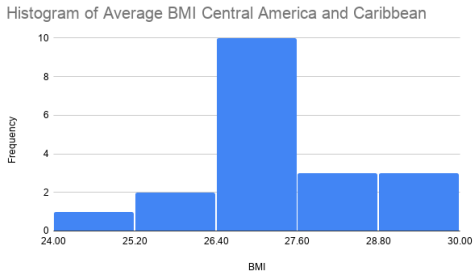
Histogram 3 (South America and Caribbean)



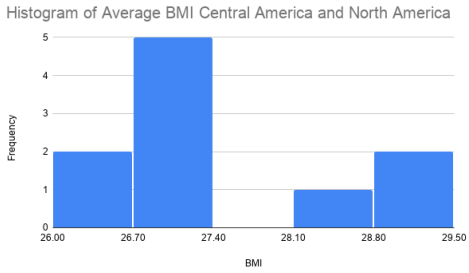
Histogram 4 (South America and North America)



Histogram 5 (Central America and Caribbean)



Histogram 6 (Central America and North America)



Histogram 7 (Caribbean and North America)

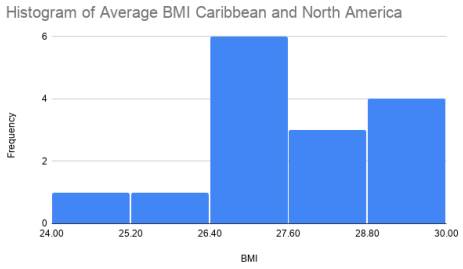


Table 1

		South America	Central America	Caribbean	North America
Overweight?	Yes	5	3	10	3
	No	7	4	3	0

R Code

Question 1:

```
PDATA<-read.csv("STAT 3480 Project Data - Sheet1.csv")

samp1SA<-PDATA$BMI[1:12]

n1<-length(samp1SA)

samp2CA<-PDATA$BMI[13:19]

n2<-length(samp2CA)

samp3C<-PDATA$BMI[20:32]

n3<-length(samp3C)

samp4NA<-PDATA$BMI[33:35]

n4<-length(samp4NA)

hist(PDATA$BMI)

data<-c(samp1SA,samp2CA,samp3C,samp4NA)

groups<-c(rep(1, n1), rep(2, n2), rep(3, n3),rep(4,n4))

### JONCKHEERE-TERPSTRA TEST

library("clinfun")

jonckheere.test(data, groups, alternative="increasing", nperm=5000)
```

Question 2:

```
#SOUTH AMERICA V CENTRAL AMERICA
```

```
wilcox.test(samp1SA, samp2CA, alternative="less")

#South America V Caribbean

wilcox.test(samp1SA, samp3C, alternative="less")

#South America V North America

wilcox.test(samp1SA, samp4NA, alternative="less")

#Caribbean V Cental America

wilcox.test(samp2CA, samp3C, alternative="less")

#Caribbean V North America

wilcox.test(samp3C, samp4NA, alternative="less")

#Central America V North America

wilcox.test(samp2CA, samp4NA, alternative="less")
```

Question 3:

```
source("http://www4.stat.ncsu.edu/~lu/ST505/Rcode/functions-Ch5.R")

regions <- c(rep("South America", 12), rep("Central America", 7), rep("Caribbean", 13), rep("North America", 3))

overweight <- c(rep("yes", 5), rep("no", 7), rep("yes", 3), rep("no", 4), rep("yes", 10), rep("no", 3), rep("yes", 3))

table(overweight, regions) #list the explanatory variable (x) second so that it becomes columns

# PERMUTATION CHI-SQUARE TEST #

chisq.test(overweight, regions)

permX2<-permapproxX2(overweight, regions, R=2000) #R sets the number of randomly sampled permutations

mean(permX2 >= chisq.test(overweight, regions)$statistic)
```

Works Cited

"Body Mass Index (BMI)." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 15 May 2015. https://www.cdc.gov/healthyweight/assessing/bmi/index.html. Web. 12 Dec, 2019.