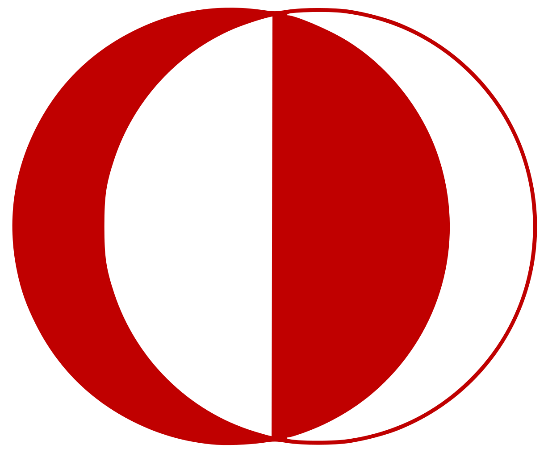
****

**MIDDLE EAST TECHNICAL UNIVERSITY**

***Department of Statistics***

**STAT 292**

**TERM PROJECT I**

**Bennur Kaya**

**Beste Karaçay**

**Merve Erşahin**

**Introduction**

* The problem to be investigated is whether there is a relation between CHD (Coronory Heart Disease) in countries and inactivity/sugar consumption/tobacco consumption/obesity and countries’ total health expenditure. This problem was chosen since it is an important disease of our age, 20-25. There is no exactly one reason for this disease and we want to check which one of the variables causes more risk.
* In our data, we have 10 variables and 3 of them are categorical. We collect data from different sources considering the causes of the coronary heart disease and some other factors that we think which might affect the CHD.
* In section 1, the variable (CHD) and our data are described. Then, in the section 2&3, the defined research questions below are answered with appropriate statistical methods. After the results of statistical tests, section 4 has some conclusions.

***Section 1***

**Description of the Problem**

* Coronary heart disease refers to a narrowing of the coronary arteries, the blood vessels that supply oxygen and blood to the heart. Research suggests that coronary heart disease (CHD) starts when certain factors damage the inner layers of the coronary arteries.

As it was mentioned earlier, we have 10 variables.

* Our dataset is consists of the countries where data are received, continents as in Africa, America, Asia, Europe, Ocenia, death rate per 100.000 people because of CHD, CHD Levels:

VH: Very High (If variable>=133.30)

H: High (If 133.30>variable>=89.76)

M: Medium (If 89.76>variable>=65.79)

L: Low (If variable<65.79),

life expectancy of people according to the countries, inactivity, sugar consumption per capita, obesity rate per country, number of cigarettes smoked per person and lastly total health expenditure per capita in U.S. dollars.

**Literature Review**

A study which is conducted by LeducMedia (2014) represents the leading causes of death in the world such as cancer, heart diseases, alcohol, smoking, fires etc. in all countries around the world. Their purpose is to stimulate meaningful research on this important subject through leading Academic Institutions worldwide, while displaying the data in ways the less informed visitor can understand and use. Our CHD data is collected from their studies.

Lancet (2008) found that, in the UK, 63.3% of adults (with higher rates in women than in men) do not meet recommended amounts of activity, such as walking briskly for 30 minutes or more, five times a week or taking more vigorous exercise for 20 minutes three times a week.  
• The US scores 41% and Canada 34%.  
• In Malta, 71.9% of adults are inactive and in Serbia the proportion is 68.3%.  
• The most active countries are Greece, where 16% are inactive, Estonia (17%) and the Netherlands (18%).

Inactivity causes between 6% and 10% of four major diseases – **coronary heart disease**, type 2 diabetes and breast and colon cancer, reported the Lancet. In 2008, it was responsible for about 5.3 million out of the 57 million deaths worldwide.

Helgi Analytics found that the world produced about 168 million tonnes of sugar in 2011. The average person consumes about 24 kilograms of sugar each year, equivalent to over 260 food calories per person per day. In modern times it has been questioned whether a diet high in sugars, especially refined sugars, is bad for health. Sugar has been linked to obesity and suspected of being implicated in diabetes, **cardiovascular disease**, dementia, macular degeneration and tooth decay. Historically, average sugar consumption per capita reached an all time high of 18.8 kg in 1987 and an all time low of 14.5 kg in 1961. The average annual growth amounted to 0.480 % since 1961.

In a research conserning obesity rate in the world Renew Bariatrics (2017) states that  there are nearly 650 million obese adults on the planet, as defined as a body mass index (BMI) over 30. There are also about 125 million obese children and adolescents in the entire world according to a [BMI](https://renewbariatrics.com/bmi-body-mass-index/) over 30. The majority of the obesity on the planet resides in a few countries, in fact, the top 10 countries contribute half the entire world’s obesity.

**Study Facts:**

* World Population: 7,505,257,673
* World Obesity Population : 774,000,000

**Top 8 Most Obese Countries (July 1st, 2017)**

1. United States of America – 109,342,839
2. China – 97,256,700
3. India – 65,619,826
4. Brazil – 41,857,656
5. Mexico – 36,294,881
6. Russia – 34,701,531
7. Egypt – 28,192,861
8. Turkey – 23,819,781

According to World Health Organization (2014) cigarettes are smoked by over 1 billion people, which is nearly 20% of the world population in 2014. About 800 million of these smokers are men. While smoking rates have leveled off or declined in [developed nations](http://www.wiki-zero.com/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvRGV2ZWxvcGVkX25hdGlvbg), especially among men, in [developing nations](http://www.wiki-zero.com/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvRGV2ZWxvcGluZ19uYXRpb24) tobacco consumption continues to rise. More than 80% of all smokers now live in countries with low or middle incomes, and 60% in just 10 countries, a list headed by China. Smokers are over half of adult males in Indonesia (57%, but mostly [kretek](http://www.wiki-zero.com/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvS3JldGVr), a local form of cigarette) and China (53% estimated), and nearly half in Bangladesh, though for women the figure is much lower.

It can be seen that there are several studies on causes of coronary heart disease. Inactivity, sugar consumption, obesity and tobacco consumption is highly effective when it comes to heart diseases in these studies. In this study, the results will be checked if the results are matching with previous researches or not.

***Section 2***

**Research Questions**

1. Is there a relationship between expenses for health of the continents and rate of disease?
2. The data set of Consumption of Sugar contains 94 observations with variance 11.6. We think that the mean rate of sugar consumption exceeds 25. Are we correct? Also, show that this claim is true or not true with an appropriate graph, diagram or table.
3. After examining the pie chart of Obesity Rate according to CHD levels, compare the obesity rates of low and high CHD levels. Is the obesity rates of low CHD level greater than high CHD level? If this is not sufficiently obvious, use the appropriate test.
4. Does number of people who died because of the disease changes according to continents?

***Section 3***

***Statistical Analysis***

Before we start to examine our research questions, we wanted to check our variables if they are related or not by using correlation.

CHD LifeEx Inactivity ConsOfSugar Obesity\_Rate

CHD 1.00000000 -0.04470635 -0.08082289 0.1256455 0.1844434

LifeEx -0.04470635 1.00000000 0.36936460 0.6050871 0.6013032

Inactivity -0.08082289 0.36936460 1.00000000 0.4394486 0.5818568

ConsOfSugar 0.12564552 0.60508705 0.43944861 1.0000000 0.6150723

Obesity\_Rate 0.18444339 0.60130322 0.58185680 0.6150723 1.0000000

Tobacco\_Consumption 0.24725178 0.59016726 0.19862919 0.2953404 0.4740833

THE -0.30317228 0.63086497 0.28015403 0.4227336 0.4486134

Tobacco\_Consumption THE

CHD 0.2472518 -0.3031723

LifeEx 0.5901673 0.6308650

Inactivity 0.1986292 0.2801540

ConsOfSugar 0.2953404 0.4227336

Obesity\_Rate 0.4740833 0.4486134

Tobacco\_Consumption 1.0000000 0.3574778

THE 0.3574778 1.0000000

**Question 1:** Is there a relationship between expenses for health of the continents and rate of disease?

H0: There is a relation between expenses for health of the continents and rate of disease.

H1: There is no relation between expenses for health of the continents and rate of disease.

First, we wanted to check that if there is a relation between expenses for health of the continents and rate of disease. As we can see from the linear regression table, since p-value (0.002975) is less than alpha value (0.05), we reject H0. That means there is no relation between expenses for health of the continents and rate of disease.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

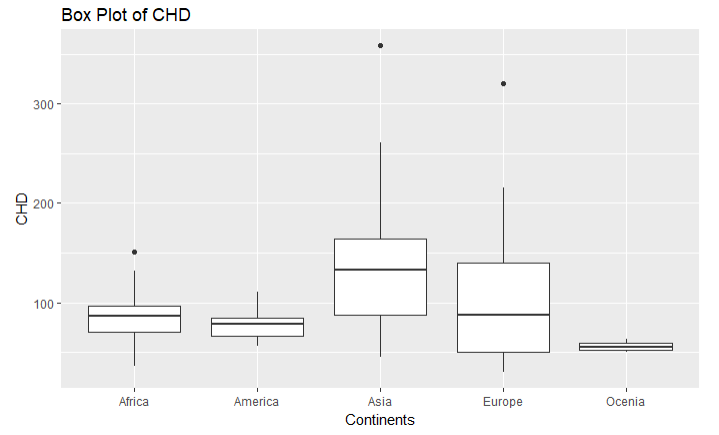
(Intercept) 116.488794 7.169141 16.249 < 2e-16 \*\*\*

THE -0.009884 0.003239 -3.052 0.00298 \*\*

F-statistic: 9.312 on 1 and 92 DF, p-value: **0.002975**

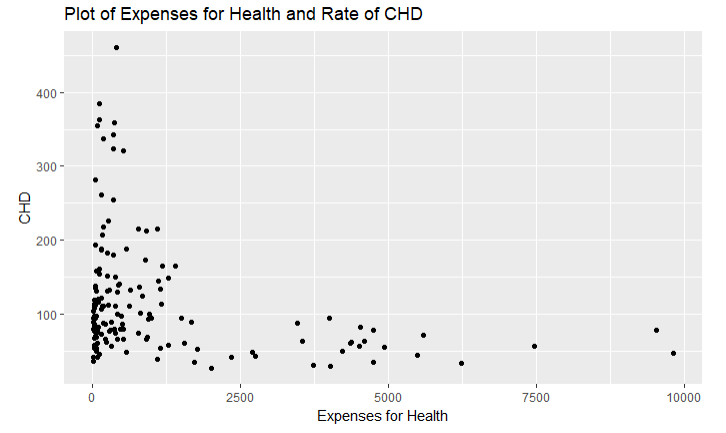
After these, we wanted to visualize it for you.

ekran görüntüsü içeren bir resim

Çok yüksek güvenilirlikle oluşturulmuş açıklama

For example, it can be understood from the box plots that although Europe has a larger health budget, it has also maximum number of people who die because of coronary heart disease, unlike Africa.

We can also see the relationship between CHD and expenses for health with scatter plot.

 It can be concluded that there is no linear relationship between these two variables.

**Question 2:** The data set of Consumption of Sugar contains 94 observations with variance 11.6. We think that the mean rate of sugar consumption exceeds 25. Are we correct? Also, show that this claim is true or not true with an appropriate graph, diagram or table.

H0: M<=25 (The mean rate of sugar consumption is less than or equal to 25)

H1: M>25 (The mean rate of sugar consumption is greater than 25)

Z-test is conducted in order to see whether the H0 is rejected or not.

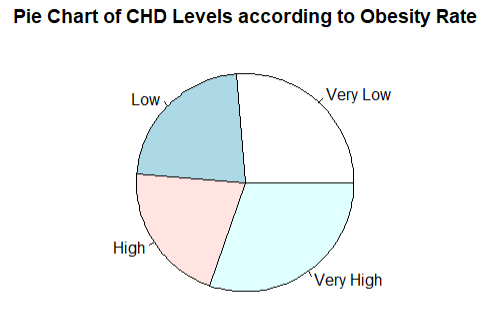
Zcalculated is -1.48 and our z-alpha(0.05) is -1.645. Since z-calculated is greater than z-alpha value, we fail to reject H0. That means, we have enough evidence to say that the mean rate of sugar consumption is greater than 25.

ekran görüntüsü içeren bir resim

Çok yüksek güvenilirlikle oluşturulmuş açıklama

Before we explain our graph, we would like you to know that if the data follows left skewed distribution, then the mean of data is less than median. Now, as it is seen from box plot, the data is left skewed and the median is less than 25. So, we conclude that the mean value of sugar consumption is less than 25.

**Question 3:** After examining the pie chart of Obesity Rate according to CHD levels, compare the obesity rates of low and high CHD levels. Is the obesity rates of low CHD level greater than high CHD level? If this is not sufficiently obvious, use the appropriate test.



Obviously, as it can be seen from the pie chart, average obesity rate for the “very high disease level” is the biggest slice, followed by “very low”. But we cannot see the difference between “low” and “high levels” clearly. So, z-test is used to clarify that.

H0: Mlow=Mhigh (Average obesity rate for high disease level is equal to average obesity rate for low disease level)

H1: Mlow>Mhigh (Average obesity rate for high disease level is greater than for low disease level)

Since the calculated z value which is 0.0307 is smaller than z alpha(0.05) value, 1.645, we reject H0. That is, we have sufficient evidence to conclude that average obesity rate for high disease level is greater than for low disease level.

**Question 4:** Does number of people who died because of the disease changes according to continents?

H0: CHD rate does not changes depending on the continent

H1: CHD rate changes depending on the continent

ANOVA table is created to investigate the hypothesis.

Df Sum Sq Mean Sq F value Pr(>F)

Continent 4 50689 12672 3.986 0.00507 \*\*

Residuals 89 282922 3179

Since our p value is less than alpha value, we reject H0. Thus, we have enough evidence to say that CHD rate changes depending on the continent. Now, we want to check which continents have significantly different. In order to see this, Tukey test is applied.

$Continent

diff lwr upr p adj

America-Africa -8.669371 -65.137362 47.79862 0.9929177

Asia-Africa 51.017138 7.041424 94.99285 0.0145830

Europe-Africa 20.944987 -21.456410 63.34638 0.6448523

Ocenia-Africa -30.085128 -125.812512 65.64226 0.9053036

Asia-America 59.686509 2.883755 116.48926 0.0344716

Europe-America 29.614357 -25.978519 85.20723 0.5759500

Ocenia-America -21.415758 -123.672402 80.84089 0.9772465

Europe-Asia -30.072152 -72.918357 12.77405 0.2969267

Ocenia-Asia -81.102267 -177.027502 14.82297 0.1378505

Ocenia-Europe -51.030115 -146.243905 44.18368 0.5701323

As it can be seen from the Tukey table, both for Asia-Africa and Asia-America, there is significant difference in number of people who died because of the disease.

Moreover, as we can see from the bar plot, the rate of death changes according to the continents. Asia has the highest rate and Europe follows Asia.

metin içeren bir resim

Çok yüksek güvenilirlikle oluşturulmuş açıklama

***Section 4***

**Conclusion**

At the end of this study, we saw that there was no relationship between expenses for health of the continents and the rate of disease. Then, by setting a hypothesis statement, if the mean of sugar consumption exceeds 25 or not was searched and the mean is less than 25 was founded by applying z-test. The other research question was obesity rates in different CHD levels. The highest rate was in the “Very high level” and then very low , respectively. Whether the low level has the greater percentage than the high level or not was investigated since there was no clear shape. At the end of this, we had sufficient evidence to conclude that average obesity rate for high disease level is greater than low disease level. The last research question showed that CHD rate changes depending on the continent.

**REFERENCES**

* Coronary Heart Disease Death Rate per 100,000 (2014). Retrieved from <http://www.worldlifeexpectancy.com/cause-of-death/coronary-heart-disease/by-country/>
* Life expectancy at birth (2015). Retrieved from <https://data.worldbank.org/indicator/SP.DYN.LE00.IN>
* List of countries by cigarette consumption per capita (2018). Retrieved from <http://www.wiki-zero.com/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvTGlzdF9vZl9jb3VudHJpZXNfYnlfY2lnYXJldHRlX2NvbnN1bXB0aW9uX3Blcl9jYXBpdGE>
* List of countries by total health expenditure per capita (2015). Retrieved from <http://www.wiki-zero.com/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvTGlzdF9vZl9jb3VudHJpZXNfYnlfdG90YWxfaGVhbHRoX2V4cGVuZGl0dXJlX3Blcl9jYXBpdGE>
* REPORT: OBESITY RATES BY COUNTRY (2017). Retrieved from <https://renewbariatrics.com/obesity-rank-by-countries/>
* Sugar Consumption Per Capita by Country (2011). Retrieved from <http://www.helgilibrary.com/indicators/sugar-consumption-per-capita/>
* Which are the laziest countries on earth? (2012). Retrieved from <https://www.theguardian.com/news/datablog/2012/jul/18/physical-inactivity-country-laziest>

**Appendix**

ch=na.omit(coronaryhd)

numeric=ch[,sapply(ch,is.numeric)]

cor(numeric)

attach(ch)

CHD LifeEx Inactivity ConsOfSugar Obesity\_Rate

CHD 1.00000000 -0.04470635 -0.08082289 0.1256455 0.1844434

LifeEx -0.04470635 1.00000000 0.36936460 0.6050871 0.6013032

Inactivity -0.08082289 0.36936460 1.00000000 0.4394486 0.5818568

ConsOfSugar 0.12564552 0.60508705 0.43944861 1.0000000 0.6150723

Obesity\_Rate 0.18444339 0.60130322 0.58185680 0.6150723 1.0000000

Tobacco\_Consumption 0.24725178 0.59016726 0.19862919 0.2953404 0.4740833

THE -0.30317228 0.63086497 0.28015403 0.4227336 0.4486134

Tobacco\_Consumption THE

CHD 0.2472518 -0.3031723

LifeEx 0.5901673 0.6308650

Inactivity 0.1986292 0.2801540

ConsOfSugar 0.2953404 0.4227336

Obesity\_Rate 0.4740833 0.4486134

Tobacco\_Consumption 1.0000000 0.3574778

THE 0.3574778 1.0000000

**#Q1**

#Ho= There is a relation between expenses for health of the continents and rate of disease.

#H1= There is no relation between expenses for health of the continents and rate of disease.

model=lm(CHD~THE)

summary(model)

# First, we wanted to check that if there is a relation between expenses for health of the continents and rate of disease. As we can see from the linear regression table, since p-value (0.002975) is less than alpha value (0.05), we reject Ho. That means there is no relation between expenses for health of the continents and rate of disease.

ggplot(coronaryhd, aes(x=THE,y=CHD))+geom\_point()+labs(title="Plot of Expenses for Health and Rate of CHD",x ="Expenses for Health", y = "CHD")

#Secondly, a scatter plot for THE and CHD to check if they are related.

ggplot(ch, aes(x=as.factor(Continent),y=THE))+geom\_boxplot()+labs(title="Box Plot of Expenses for health",x ="Continents", y = "Expenses for Health")

ggplot(ch, aes(x=as.factor(Continent),y=CHD))+geom\_boxplot()+labs(title="Box Plot of CHD",x ="Continents", y = "CHD")

a=tapply(ch$THE, ch$Continent, mean)

b=tapply(ch$CHD, ch$Continent, mean)

a

b

# For example, although Europe has a large health budget, it has also maximum number of people who die because of coronary heart disease, unlike Africa.

**#Q2**

#Ho: M<=25 (The mean rate of sugar consumption is less than or equal to 25)

#H1: M>25 (The mean rate of sugar consumption is greater than 25)

x=mean(ch$ConsOfSugar)

x

m=25

std=sd(ch$ConsOfSugar)

std

n=length(ch$ConsOfSugar)

n

ztest=function(x,m,std,n) {

zvalue=(x-m)/(std/sqrt(n))

return(zvalue)

}

ztest(23.23404,25,11.61742,94)

# z\_calculated is -1.48 and our z-alpha(0.05) is -1.645. Since z-calculated is greater than z-alpha value, we fail to rejet Ho.

ggplot(ch, aes(as.factor(0),y=ConsOfSugar))+geom\_boxplot(fill="indianred1",colour="indianred3")+labs(title="Box Plot of Consumption of Sugar",y="Sugar Consumption")

#Before we explain our graph, we would like you to know that if the data follows left skewed distribution, then the mean of data is less than median. Now, as it is seen from box plot, the data is left skewed and the median is less than 25. So, we conclude that the mean value of sugar consumption is less than 25.

**#Q3**

one=mean(ch$Obesity\_Rate[ch$CHD\_LEVEL=="Very Low"])

one

two=mean(ch$Obesity\_Rate[ch$CHD\_LEVEL=="Low"])

two

three=mean(ch$Obesity\_Rate[ch$CHD\_LEVEL=="High"])

three

four=mean(ch$Obesity\_Rate[ch$CHD\_LEVEL=="Very High"])

four

means=c(19.16,16.326,15.208,22.04)

chd\_levels=c("Very Low","Low","High","Very High")

pie(means, labels=chd\_levels, main="Pie Chart of CHD Levels according to Obesity Rate")

#Obviously, as it can be seen from the pie chart, average obesity rate for the very high disease level is the biggest slice, followed by very low. But we cannot see the difference between low and high levels clearly. So, z-test is used to clarify that.

#H0: Mlow=Mhigh (Average obesity rate for high disease level is equal to average obesity rate for low disease level)

#H1: Mlow>Mhigh (Average obesity rate for high disease level is greater than for low disease level)

xlow=16.326

xhigh=15.208

nlow=length(ch$Obesity\_Rate[ch$CHD\_LEVEL=="Low"])

nlow

nhigh=length(ch$Obesity\_Rate[ch$CHD\_LEVEL=="High"])

nhigh

sdlow=sd(ch$Obesity\_Rate[ch$CHD\_LEVEL=="Low"])

sdhigh=sd(ch$Obesity\_Rate[ch$CHD\_LEVEL=="High"])

sdlow

sdhigh

z\_test=function(xlow,xhigh,sdlow,sdhigh,nlow,nhigh) {

zvalue=(xlow-xhigh-0)/(sdlow^2/sqrt(nlow)+sdhigh^2/sqrt(nhigh))

return(zvalue)

}

z\_test(16.326, 15.208, 8.977, 9.776, 23, 24)

#Since the calculated z value which is 0.0307 is smaller than z alpha(0.05) value, 1.645, we reject H0. That is, we have sufficient evidence to conclude that average obesity rate for high disease level is greater than for low disease level.

**#Q4**

#H0: CHD rate does not changes depending on the continent

#H1: CHD rate changes depending on the continent

anovaCHD=aov(CHD~Continent, data=ch)

summary(anovaCHD)

#Since our p\_value is less than alpha value, we reject Ho. Thus we have enough evidence to say that CHD rate changes depending on the continent. Now, we want to check which continents have significant difference between them.

TukeyHSD(anovaCHD)

#As it can be seen from the ANOVA table, both for Asia-Africa and Asia-America, there is significant difference in number of people who died because of the disease.

ggplot(ch, aes(factor(Continent), CHD)) +

geom\_bar(stat="identity", position = "dodge") +

scale\_fill\_brewer(palette = "Set1")

# Moreover, as we can see from the bar plot, the rate of death changes according to the continents. Asia has the highest rate and Europe follows Asia.