**Diabetes analysis on BMI and Glucose**

**Submitted to: Lynn A. Agre, MPH, PhD**

**Statistics 390: Section MA**

**December 9, 2022**

**By: Mia (Mingya) Yang**

**RU Student ID: 194008295**

**I.  Abstract**

This study will investigate the relationship between BMI (x variable) and Glucose (y variable). This research will explore how BMI affects the Glucose of people with diabetes. This analysis will explore the association between BMI and Glucose. In this research, the sample size is 20, and they are randomly chosen from the 770 variables. SAS programing language is being used in this research to get accurate and visualized results. Several statistical methods, such as hypothesis tests, ANOVA, etc., are used in this research to provide professional results. The results can help diabetes patients better understand the relationship between BMI and glucose.

**II. Introduction/Background:**

Diabetes is one of the most well-known chronic diseases that have significant influences on people’s quality of life. Diabetes can be a cause of other diseases, such as blindness, stroke, kidney failure, or even heart attacks. World Health Organization points out that the world diabetes population has increased by almost 320 million over the past 30 years. According to the Centers for Disease Control and Prevention (CDC), there are 37.3 million people suffering from diabetes which is 11.3% of the US Population. Today many young teenagers (aged under 18 years old) are diagnosed to type 1 diabetes. Adults with high living pressures are adopting many unhealthy habits, such as having irregular working and resting schedules and eating fast food or food with too much sugar. These unhealthy living habits can be the causes of diabetes. Therefore, getting rid of those unhealthy habits and trying to have a healthy diet, regular physical exercise, reducing tobacco use, and maintaining body weight could be ways to prevent type one and type two diabetes. Hypothetically speaking, individuals with higher body weight tend to have higher glucose, and the risks of being diagnosed with diabetes are much more likely higher. To help diagnosed diabetes patients and prediabetes people improve their quality of life or even have longevity life and understand the relationship between their body weight and glucose, this research focuses on two variables: BMI (Body Mass Index) and Glucose.

Based on the study that investigated the relationship between BMI and Glucose in people with type 2 diabetes written by PoChung Cheng, Shangren Hsu, Shihte Tu, Yunchung Cheng, and Yuhsiu Liu: “BMI was significantly related to plasma glucose concentration during iatrogenic hypoglycemia in people with T2DM.” (Body mass index influences the plasma glucose concentration during iatrogenic hypoglycemia in people with type 2 diabetes mellitus: a cross-sectional study). In this case, in patients with type 2 diabetes mellitus, their BMI strongly correlates with glucose when they suffer clinically low blood sugar period. Another research discovered the correlations between BMI and Glucose provided more shreds of evidence for this research study. “As BMI increases, insulin resistance also increases which results in increased blood glucose level in body”, claims by Neelam Agrawall, M. Agrawall, Sunil Kumar in their article. (Correlation between Body Mass Index and Blood Glucose levels in Jharkhand Population). The article provides a clear direction for this research—whether BMI and glucose have significant correlations that can lead to diabetes. In this research, the analysis is based on a sample size of 20 randomly selected from 768 variables. The original owner of the data is the National Institute of Diabetes and Digestive and Kidney Diseases. The population of data was collected from Phoenix, Arizona, USA, and was collected on May 9th, 1990.

**III. Methods**

In this research, females aged above 21 of Pima Indian heritage are being studied. Plasma glucose concentration 2 hours in an oral glucose tolerance test. The formula for calculating the BMI is equal to weight in kilograms divided by the power of heights in feet. Four small studies using different SAS code arguments can help better analyze the data. The first part is (Homework No.1): **Proc means** argument summarizes the sample size and generates the mean, standard deviation, minimum value, and maximum value of BMI and Glucose. The other argument is **Proc Freq.** It creates frequency and cross-tabulation tables of BMI and Glucose. It can provide a clear relationship between 2 variables and help researchers to discover the association between BMI and Glucose. **Proc Univariate** can indicate data distribution and some other basic statistical measures. It can also test for location:mu0=0, an assessment of normality, extreme observations, and find the outliers for researchers. The second part is (Homework No.2): **Proc means t-test** conducted t-test for one sample, two samples, or paired observations. In this diabetes research, it helps conduct a one-sample t-test. **IF, Then** statement in this project used to cut the data into different ranges and categorize them into several types. **Proc Freq** in this part, provides the relationship between BMI and Glucose. **Proc corr** argument calculated the coefficients with original data and standardized data. The third part of this project is (Homework No.3). In this part, several charts and graphs are generated to visualize the relationship between BMI and Glucose. **Proc chart vbar makes some bar charts and assigns the values for each category. Proc univariate histogram creates histograms and can also help to draw parametric and nonparametric probability density curves. Proc freq to create Proc chart vbar; this command gives the value of the categorical variables, counts, and proportions. Proc tabulate argument offers descriptive statistics in tabular format. Proc boxplot creates side-by-side boxplots. The Proc plot procedure** plots the values of BMI and Glucose for each observation in the diabetes dataset that has already been imported into SAS. **Proc sgscatter** creates some scatter plots for multiple combinations of BMI and Glucose. The fourth part (Homework No.4) included hypothesis tests to help researchers generate accurate results. To determine whether there is an association between categorical data, we use **Proc freq** and **Proc chi square.** The hypothesis test is /\* H0:"[BMI] is dependent of [Glucose]" \*/

/\* H1:"[BMI] is independent of [Glucose]" \*/

/\* p-value is 0.25. Since 0.25 is much larger than 0.05 it fails to reject null\*/

/\*There is not sufficient evidence to say that there is a statistically significant difference between the means of BMI and Glucose. \*/ Another argument is **Proc glm** and **Proc anova.** It is used to conduct an analysis of variance for balanced data only. The hypothesis is /\* H0: mu1 = mu2 \*/

/\* H1: “At least one of the means is different from the others” \*/

/\*p-value is 0.5118 > 0.05, it fails to reject null hypothesis \*/

/\*There is not sufficient evidence to say that there is a statistically significant difference between the means of BMI and Glucose. \*/. For **Proc t-test** and paired t-test, the dependent sample t-test is used to compare to the population means. /\* H0: mu1 = mu2 (the paired population means are equal) \*/ /\* H1: mu1 is not equal to mu2 (the paired population means are not equal )\*/

/\*Since the p-value is very small (p< 0.0001) and it’s less than 0.05 which means we reject null \*/ /\*There is enough evidence to conclude that the mean difference between the paired observations is statistically significant. \*/. **Proc corr scatter** calculated correlation of coefficient of continuous variables. /\* H0: ro = 0 (The population correlation coefficient is not significantly differently from 0) \*/ /\*H1: ro not equal to 0 (The population correlation coefficient is significantly different from 0)\*/ /\*The p-value is 0.517. Since the p-value is 0.517 which is larger than 0.05. We fail to reject the null. \*/ /\* There is insufficient evidence to conclude that there is a significant linear relationship between BMI and Glucose because the correlation coefficient is NOT significantly different from zero\*/. The last part is (Homework No.5). **Proc reg** with outset will conduct the reg procedures, and it will fit the data into models. Proc glm is an important statement that can generate a linear regression model based on the least squares method.

**IV. Homework**

**Homework No.1:**

**Codes:**

proc import file = 'C:\Users\my397\Downloads\diabetes.csv'

out= work.Diabetes

dbms=csv;

run;

proc print data= work.diabetes (obs = 20) noobs;

run;

proc means data= work.diabetes (obs = 20);

Var BMI;

run;

proc freq data= work.diabetes (obs = 20);

table Glucose BMI;

run;

proc univariate data= work.diabetes normal plot;

var BMI;

ID Glucose;

run;

Results Output:

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table, Word

Description automatically generatedTable

Description automatically generated

Chart

Description automatically generated

**Homework No.2:**

**Codes:**

proc means N mean std t prt;

title 'one sample t-test';

var BMI Glucose ;

run;

data diabetes;

set diabetes;

if BMI <= 18.5 then mark = 'Underweight';

if 18.5 <= BMI <24.9 then mark ='Normal';

if 24.9 <= BMI < 29.9 then mark = 'Overweight';

if 29.9 <= BMI < 34.9 then mark = 'Obese';

if BMI > = 35 then mark = 'Extremely Obese';

if Glucose <= 80 then comment = 'Low';

if 80 <= Glucose < 100 then comment = 'Normal';

if 100 <= Glucose < 130 then comment = 'High';

if 130 <= Glucose < 150 then comment = 'Extremely High';

if Glucose > = 150 then comment = 'Dangerous';

run;

proc print data = diabetes;

run;

proc freq data = diabetes;

table mark\*comment;

run;

proc corr data = diabetes (obs = 20);

var Glucose;

with BMI;

run;

**Results Output:**

**Table

Description automatically generated**

**Graphical user interface, application, table

Description automatically generated**

**Homework No.3:**

**Codes:**

proc print data= diabetes;

run;

proc gchart data = diabetes;

vbar mark;

Title 'Bar Chart of marks from diabetes';

run;

proc gchart data = diabetes;

vbar comment;

Title 'Bar chart of comments from diabetes';

run;

proc univariate data = diabetes;

var BMI Glucose;

histogram / normal ;

title 'Histogram of marks from diabetes';

run;

proc freq data = diabetes;

tables mark comment / out= pct\_row;

title 'Frequency table';

run;

proc tabulate data = diabetes;

class mark comment;

var BMI Glucose;

table comment\*mark, BMI Glucose;

run;

data diabetes;

set diabetes;

group = 1;

run;

ods graphics off;

title 'Box plot for BMI';

axis1 label = none value = none major = none;

proc boxplot data = diabetes;

plot BMI\*group / haxis=axis1 boxwidth= 10;

run;

ods graphics;

ods graphics off;

title 'Box plot for Glucose';

axis1 label = none value = none major = none;

proc boxplot data = diabetes;

plot Glucose\*group / haxis = axis1 boxwidth = 10;

run;

ods graphics;

proc plot data = diabetes;

plot BMI \* Glucose;

title 'Scatter plot';

run;

proc sgscatter data = diabetes;

plot BMI \* Glucose

/ group = mark ;

title 'BMI vs. Glucose';

run;

Output Results:

Table

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

**Graphical user interface, table

Description automatically generated**

**Table

Description automatically generated**

Chart, histogram

Description automatically generated

Table

Description automatically generated

Graphical user interface, table

Description automatically generated

Table

Description automatically generated

Chart, histogram

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Homework No.4:**

**Codes:**

proc print data = diabetes;

run;

proc freq data = diabetes;

tables mark\*comment / chisq expected deviation norow nocol nopercent;

title1 'Frequency table hw#4';

run;

proc tabulate data = diabetes;

class mark comment;

var BMI Glucose;

table comment\*mark, BMI Glucose;

run;

/\* H0:"[BMI] is dependent of [Glucose]" \*/

/\* H1:"[BMI] is independent of [Glucose]" \*/

/\* p-value is 0.25. Since 0.25 is much larger than 0.05 it fails to reject null\*/

/\*There is not sufficient evidence to say that there is a statistically significant difference between the means of BMI and Glucose. \*/

proc glm data = diabetes;

class BMI Glucose;

model Glucose = BMI;

means BMI;

run;

proc anova data = diabetes;

class BMI Glucose;

model Glucose = BMI;

means BMI;

run;

/\* H0: mu1 = mu2 \*/

/\* H1: “At least one of the mean is different from the others” \*/

/\*p-value is 0.5118 > 0.05, it fails to reject null hypothesis \*/

/\*There is not sufficient evidence to say that there is a statistically significant difference between the means of BMI and Glucose. \*/

proc ttest data = diabetes alpha= 0.05;

paired BMI\*Glucose;

run;

/\* H0: mu1 = mu2 (the paired population means are equal) \*/

/\* H1: mu1 is not equal to mu2 (the paired population means are not equal )\*/

/\*Since the p-value is very small (p< 0.0001) and it’s less than 0.05 which means we reject null \*/

/\*There is enough evidence to conclude that the mean difference between the paired observations is statistically significant. \*/

proc corr data = diabetes plots=scatter(nvar = all);

var BMI Glucose;

run;

/\* H0: ro = 0 (The population correlation coefficient is not significantly differently from 0) \*/

/\*H1: ro not equal to 0 (The population correlation coefficient is significantly different from 0)\*/

/

/\*The p-value is 0.517. Since the p value is 0.517 which is larger than 0.05. We fail to reject the null. \*/

/\* There is insufficient evidence to conclude that there is a significant linear relationship between BMI and Glucose because the correlation coefficient is NOT significantly different from zero.\*/

Output Results:

Table

Description automatically generated

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

Chart, scatter chart, box and whisker chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

**Homework No.5:**

**Codes:**

proc reg data = diabetes

outest= work.param\_estimates;

model BMI = Glucose;

run;

ods output ParameterEstimates = work.param\_estimates;

proc glm data = diabetes;

model BMI = Glucose;

run;

Graphical user interface, table, Word

Description automatically generated

Chart, diagram, engineering drawing

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, application

Description automatically generated

Chart, scatter chart

Description automatically generated

**V. Results:**

From Homework 1:

In homework one, the SAS helps calculate the Sample size, mean, standard deviation, minimum value, and maximum value of BMI and Glucose. As this research considers BMI as an independent variable, SAS also gives significant results on Sample size, mean, standard deviation, skewness, uncorrected SS, and coefficient variance of BMI. The other significant result given by the research included basic statistical measures, such as quartiles. The distribution and probability plot for BMI is also significant as the output results provided. As we can see, the distribution of BMI is normally distributed. However, Coefficients in Cross-tabulations for BMI and Glucose seem not that significant.

From Homework 2:

The t-test statistics (t- values and p-values) and Pearson correlation coefficients show that BMI and Glucose significantly correlate. The IF-THEN statements provided a table with splitting points of both BMI and Glucose, and it is called mark and comment tables (recorded table) can generate good analysis. Still, as the result of homework, one shows, the Coefficients in Cross-tabulations for BMI and Glucose are not that significant.

From Homework3:

The Bar chart of marks and comments from diabetes data shows that each category’s distribution is significant. The histogram of marks and comments on diabetes data for BMI and Glucose are close to the normal distribution, which provided significant results. The box plot for Glucose data is a little bit skewed to the left, but since the mean and median are really close to each other, the distribution can be considered normal. BMI boxplot has the mean and median almost overlapping each other, meaning the sample data of BMI is normally distributed. The scatterplot for BMI and Glucose and the paneled graph of the scatter plot show certain patterns of the data, which can create some further research on the relationship between BMI and Glucose. The insignificant result can be the Coefficients in Cross-tabulations for BMI and Glucose.

From Homework 4:

In homework 4, we did some hypothesis tests to prove the relationship and correlations between BMI and Glucose. The significant results used to conduct the analysis can be the Sum of squares, F-value, p-value, and R-square of BMI and Glucose. The p-value can conclude the first two hypotheses, and the results are significant. Pearson correlation coefficients for BMI and Glucose are also useful and significant to draw the conclusion. Scatter plot for BMI and Glucose is also significant.

From Homework 5:

The fit diagnostic for BMI generated by Reg procedures based on model one analyzed the data in several plots and they showed a significant correlation between BMI and Glucose. The parameter estimations are also significant. The fit plot for BMI in 95% confidence interval indicates that most of the data fall into that range which is significant.

**VI. Discussions and Interpretation**

According to the SAS programming results, all the procedures used to conduct different tests or statistics calculations show that BMI and Glucose are highly correlated with each other. Individuals with higher BMI tend to have high Glucose, and the chance they get diabetes would be higher. Individuals with lower BMI may have lower Glucose. Investigating the relationship or the correlation between BMI and Glucose can help diabetes patients and healthy people better understand how their body systems work. For diabetes patients, understanding how BMI influences on Glucose can help them control and manage the use of insulin. For prediabetes patients, this research may give insight to them and prevent them from getting diabetes. Healthy individuals can learn how their BMI is calculated and live healthy life. The hypothesis test generated by the SAS programming system test the significance and the correlation between BMI and Glucose, which helped conclude that BMI and Glucose are highly correlated. The limitation of this research is that first, the formula to calculate the BMI provided by the World Health Organization doesn’t take muscles into account which means the individuals (have diabetes or not) with higher BMI may not be obese and be completely fit. In other words, since BMI does not take muscles into we cannot fully conclude that with a higher BMI, one will also have higher glucose. To better analyze the relationship between BMI and Glucose, we can either generate a new formula to calculate BMI that can take muscles into account or use A1C as a new variable. The A1C test measures the percentage of one’s red blood cells that have sugar-coated hemoglobin. To generate more accurate results, individuals’ A1C test results can be a good criterion. In general, this research can benefit individuals who are diagnosed with diabetes or prediabetes patients. The doctors may use the results of this research to help manage the patient’s insulin use. Furthermore, as there are more and more young teenagers are diagnosed with diabetes, this research can help them to improve their life quality. To expand and develop this research in the future, researchers or scientists can categorize diabetes into type one and type two and give specifically analyze of each type of diabetes to see if all diabetes follows the same relationship between BMI and Glucose. Also, researchers or scientists can use the BMI and Glucose relationship and correlations to further investigate the relationship between diabetes and its patient’s genetics. Even to figure out the cause of diabetes.

**VII. Citations:**

1. [Po Chung Cheng](https://pubmed.ncbi.nlm.nih.gov/?term=Cheng%20PC%5BAuthor%5D), [Shang Ren Hsu](https://pubmed.ncbi.nlm.nih.gov/?term=Hsu%20SR%5BAuthor%5D), [Shih Te Tu](https://pubmed.ncbi.nlm.nih.gov/?term=Tu%20ST%5BAuthor%5D), [Yun Chung Cheng](https://pubmed.ncbi.nlm.nih.gov/?term=Cheng%20YC%5BAuthor%5D), and [Yu Hsiu Liu](https://pubmed.ncbi.nlm.nih.gov/?term=Liu%20YH%5BAuthor%5D) (2018): “Body mass index influences the plasma glucose concentration during iatrogenic hypoglycemia in people with type 2 diabetes mellitus: a cross-sectional study.” PMCID: PMC5816962
2. [Neelam Agrawall](https://www.semanticscholar.org/author/Neelam-Agrawall/2103817264), [M. Agrawal](https://www.semanticscholar.org/author/M.-Agrawal/47280469), [Sunil Kumar](https://www.semanticscholar.org/author/Sunil-Kumar/2109862875) (2017): “Correlation between Body Mass Index and Blood Glucose levels in Jharkhand Population.” Medicine.
3. CDC website
4. WHO website

Appendix:

Table

Description automatically generated