IE7280: Statistical Methods in Engineering

Project Report

Topic 1: Diabetes in Indian Women: Analyzing using ANOVA Data

Nidhi Saraf

Shruthi Gadkari

Rasagnya Reddy

Northeastern University
College *of* Engineering

Topic1

Diabetes in Indian Women: Analyzing using ANOVA Data

Description

The raw dataset consists of 10 variables and 769 records. The variables used in this study are:

1. Glucose
2. Insulin
3. Pregnancies
4. Skin Thickness
5. Diet
6. Outcome
7. DiabetesPedigreefunction
8. BMI
9. Age
10. BloodPressure

Data Source: https://www.kaggle.com/rushikeshjoshi/biostatastics

Objective Statement

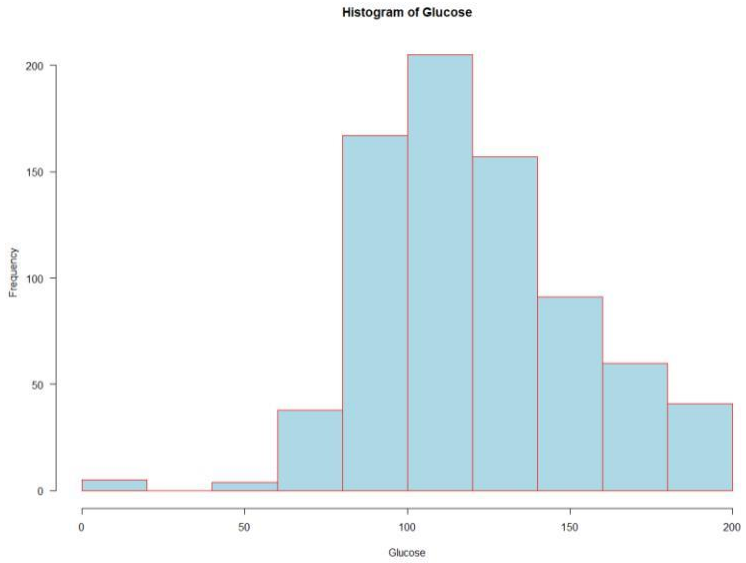
In this study, we want to analyze outcome of diabetes in women according to their Glucose levels, Insulin, BMI, Age, BloodPressure and DiabetesPedigreefunction. We would like to address the following questions.

1. Whether each of the independent variable would affect the outcome?
2. Are there any interactions between the independent variables?
3. What is the most significant variable contributing to the outcome?

Statistical Procedure

This study applies ANOVA on analyzing the outcome of diabetes. We will perform ANOVA to gather the required statistics. Test for Hypothesis and check if there's any significant difference between the variables. Also, perform Duncan's multiple range test to indicate the significant variable

Distributions of independent variables 1. Distribution of Glucose



Summary:

Min. 1st Qu. Median Mean 3rd Qu. Max.

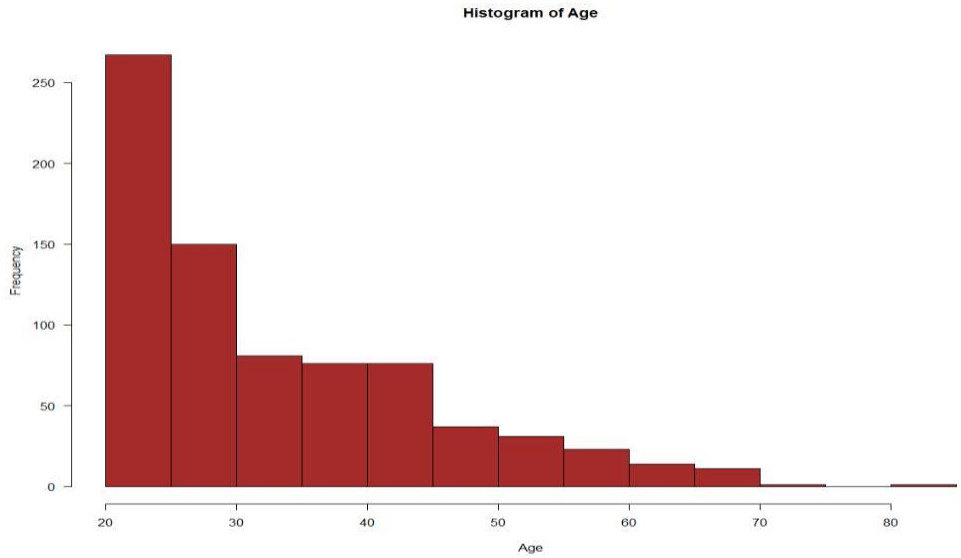
0.0 99.0 117.0 120.9 140.2 199.0

From the histogram it can be observed that the concentration of glucose levels is more in th e range of 50-150.

2. Distribution of Age

Min. 1st Qu. Median Mean 3rd Qu. Max.

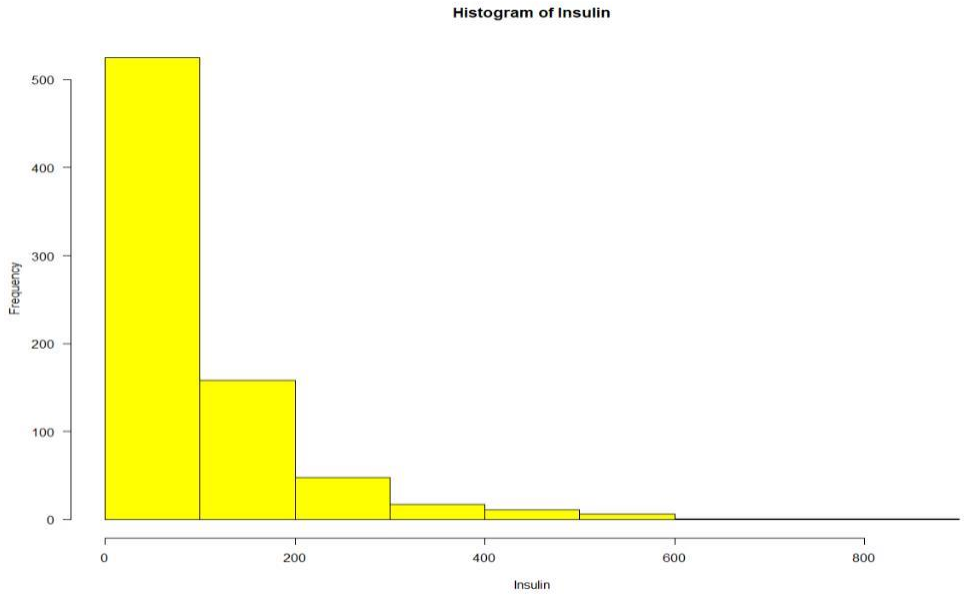
21.00 24.00 29.00 33.24 41.00 81.00



Majority of the women belong in the age group of 20-30 and there's equal distribution from 30-45

3. Distribution of Insulin

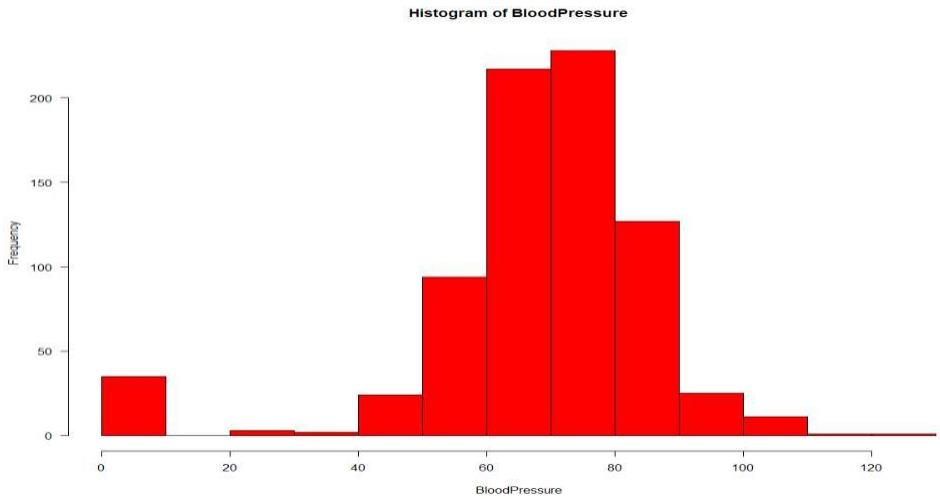
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0 0.0 30.5 79.8 127.2 846.0



The Insulin levels are mostly at ranging from 0-100 and a subset is ranging from 100-200

4. Distribution of BloodPressure

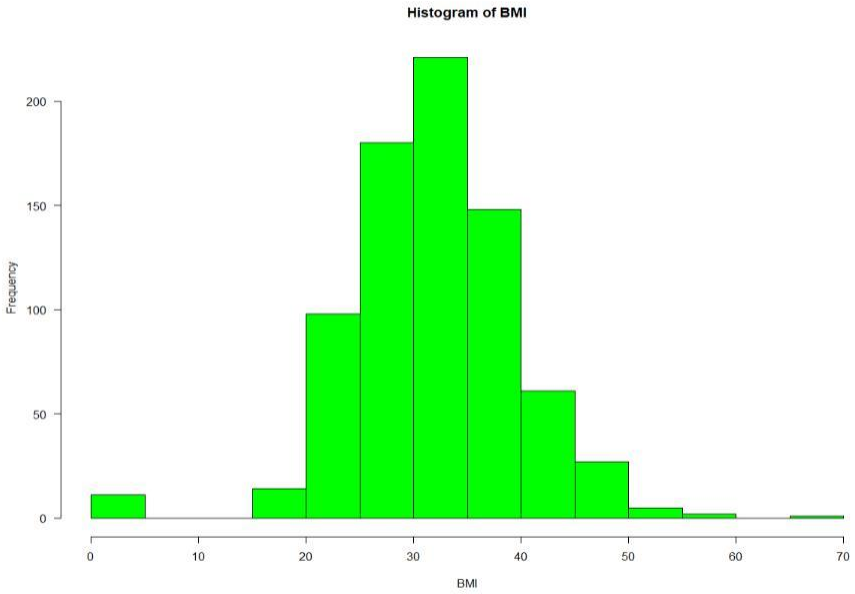
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 62.00 72.00 69.11 80.00 122.00



BloodPressure has values ranging in between 60-110 and a few values fall in the level of 0-10 which is low when compared with avg BP 120/80.

5. Distribution of BMI

Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 62.00 72.00 69.11 80.00 122.00



BodyMassIndex has majority of the values fall in to the buckets of 20-50

Fitting Linear Model to the dataset

The aim of linear regression is to model a continuous variable *Y* as a mathematical function of one or more *X* variable(s), so that we can use this regression model to predict the *Y* when only the *X* is known.

Y here is the Outcome(whether diabetic or not) and X are the independent variables contributing to the outcome.

The coefficients of the model are:

Coefficients:

(Intercept)	Pregnancies	Glucose	BloodPressure
-0.8396891	0.0207249	0.0059085	-0.0023364
SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
0.0001150	-0.0001777	0.0132543	0.1478877
Age	DietVegan	DietVegetarian	
0.0025810	-0.0244909	-0.0108945	

The Linear model equation looks like:

-0.8396891 + 0.0207249*Pregnancies + 0.0059085*Glucose – 0.0023364*BloodPressure + 0.0001150*SkinThickness - 0.0001777*Insulin + 0.0132543*BMI + 0.1478877*DiabetesPedigreeFunction + 0.0025810*Age – 0.0244909*DietVegan – 0.0108945*DietVegtarian

Looking at the coefficients from the equation, variables Pregnancies, Glucose, BMI, DiabetesPedigreeFucntion and Age have positive and larger coefficients, so it can be concluded that their contribution will be significant in way.

Summary of Model

Residuals:

Min	1Q	Median	3Q	Max
-1.01588	-0.29578	-0.09749	0.32298	1.24010

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.8396891	0.0897752	-9.353	< 2e-16	***
Pregnancies	0.0207249	0.0051394	4.033	6.08e-05	***
Glucose	0.0059085	0.0005159	11.452	< 2e-16	***
BloodPressure	-0.0023364	0.0008149	-2.867	0.00426	**
SkinThickness	0.0001150	0.0011149	0.103	0.91790	
Insulin	-0.0001777	0.0001502	-1.183	0.23715	
BMI	0.0132543	0.0020909	6.339	3.97e-10	***
DiabetesPedigreeFunction	0.1478877	0.0451090	3.278	0.00109	**
Age	0.0025810	0.0015515	1.664	0.09661	.
DietVegan	-0.0244909	0.0354270	-0.691	0.48958	
DietVegetarian	-0.0108945	0.0356048	-0.306	0.75970	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4006 on 757 degrees of freedom
Multiple R-squared: 0.3037, Adjusted R-squared: 0.2945
F-statistic: 33.02 on 10 and 757 DF, p-value: < 2.2e-16

Insights: Looking at the summary and p-values we can say that the variables Pregnancies, Glucose, BMI, BloodPressure and DiabetesPEdigreeFunction are significant. The p-values for the above-mentioned variables are < 0.05(alpha) which is why we conclude that they are significant.

ANOVA of model

```
Analysis of Variance Table

Response: Outcome
                          Df  Sum Sq Mean Sq  F value    Pr(>F)
Pregnancies                1   8.591   8.591  53.5307 6.503e-13 ***
Glucose                    1  34.021  34.021 211.9807 < 2.2e-16 ***
BloodPressure              1   0.123   0.123   0.7694 0.3806913
SkinThickness              1   0.864   0.864   5.3822 0.0206083 *
Insulin                    1   0.255   0.255   1.5911 0.2075637
BMI                        1   6.780   6.780  42.2466 1.459e-10 ***
DiabetesPedigreeFunction   1   1.818   1.818  11.3263 0.0008027 ***
Age                        1   0.459   0.459   2.8595 0.0912464 .
Diet                       2   0.077   0.038   0.2398 0.7868380
Residuals                757 121.491   0.160
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insights: From the ANOVA table we can deduce that the variables Pregnancies and SkinThickness are also significant at different confidence levels.

Test of Hypothesis for these variables

1. Glucose

0: 1 = 2 = 3 = 4 = 5…= un

1: ’ h . = 1 n

P_value = 2.2e-16 < 0.05

∴ 0 There is enough evidence to prove that the effect of Glucose is significant to determine the diabetic outcome in women.

2. BloodPressure

0: 1 = 2 = 3 = 4 = 5…= un

1: ’ h . = 1 n P_value = 0.00426 < 0.05

∴ 0 There is enough evidence to prove that the effect of BloodPressure is significant to determine the diabetic outcome in women.

3. BMI

0:1=2=3=4=5
1: ′ *h* . = 1 n
P_value = 3.97e-10 < 0.05

∴ 0 There is enough evidence to prove that the effect of BMI is significant to determine the diabetic outcome in women.

4. DiabetesPedigreeFunction

0:1=2=3=4=5
1: ′ *h* . = 1 n
P_value = 0.00109 < 0.05

∴ 0 There is enough evidence to prove that the effect of DiabetesPedigreeFunction is significant to determine the diabetic outcome in women.

5. Pregnancies

0:1=2=3=4=5
1: ′ *h* . = 1 n

P_value = 6.08e-5 < 0.05 ∴ 0 There is enough evidence to prove that the effect of Pregnancies is significant.

GLM model for data

Generalized Linear Model to give a better picture with respect to interactions between the independent variables

Summary:

Call:
glm(formula = Outcome ~ .^2, data = DB)

Deviance Residuals:
Min	1Q	Median	3Q	Max
-0.99482	-0.27129	-0.06944	0.25594	1.37114

Coefficients:

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

(Intercept)	-9.429e-01	3.949e-01	-2.388	0.01721 *
Pregnancies	7.482e-02	3.510e-02	2.132	0.03338 *
Glucose	5.929e-03	3.458e-03	1.715	0.08682 .
BloodPressure	-2.555e-03	4.877e-03	-0.524	0.60048
SkinThickness	1.059e-03	7.985e-03	0.133	0.89450
Insulin	-1.920e-03	1.285e-03	-1.494	0.13566
BMI	3.419e-03	1.183e-02	0.289	0.77265
DiabetesPedigreeFunction	6.951e-01	2.841e-01	2.447	0.01465 *
Age	1.827e-03	1.052e-02	0.174	0.86212
DietVegan	2.462e-01	2.300e-01	1.070	0.28494
DietVegetarian	8.953e-02	2.199e-01	0.407	0.68404
Pregnancies:Glucose	-2.932e-05	1.859e-04	-0.158	0.87473
Pregnancies:BloodPressure	-1.705e-04	2.485e-04	-0.686	0.49276
Pregnancies:SkinThickness	2.134e-04	3.623e-04	0.589	0.55601
Pregnancies:Insulin	2.293e-05	5.627e-05	0.408	0.68376
Pregnancies:BMI	-5.920e-04	7.458e-04	-0.794	0.42764
Pregnancies:DiabetesPedigreeFunction	3.947e-02	1.665e-02	2.370	0.01803 *
Pregnancies:Age	-1.314e-03	4.709e-04	-2.789	0.00542 **
Pregnancies:DietVegan	1.452e-03	1.265e-02	0.115	0.90868
Pregnancies:DietVegetarian	1.196e-02	1.349e-02	0.887	0.37556
Glucose:BloodPressure	-3.700e-05	3.323e-05	-1.113	0.26591
Glucose:SkinThickness	-8.491e-05	3.742e-05	-2.269	0.02356 *
Glucose:Insulin	7.761e-06	5.302e-06	1.464	0.14372
Glucose:BMI	1.691e-04	8.495e-05	1.990	0.04693 *
Glucose:DiabetesPedigreeFunction	-8.116e-04	1.611e-03	-0.504	0.61463
Glucose:Age	-4.523e-05	5.164e-05	-0.876	0.38136
Glucose:DietVegan	-1.083e-03	1.265e-03	-0.856	0.39238
Glucose:DietVegetarian	2.104e-03	1.336e-03	1.576	0.11558
BloodPressure:SkinThickness	8.037e-06	7.284e-05	0.110	0.91217
BloodPressure:Insulin	6.976e-06	1.326e-05	0.526	0.59901
BloodPressure:BMI	-7.360e-06	8.013e-05	-0.092	0.92684
BloodPressure:DiabetesPedigreeFunction	-2.351e-03	2.990e-03	-0.786	0.43197
BloodPressure:Age	2.231e-04	8.992e-05	2.481	0.01333 *
BloodPressure:DietVegan	8.047e-05	2.114e-03	0.038	0.96965
BloodPressure:DietVegetarian	-1.039e-03	2.107e-03	-0.493	0.62208
SkinThickness:Insulin	1.745e-05	1.532e-05	1.139	0.25507
SkinThickness:BMI	4.588e-05	1.565e-04	0.293	0.76949
SkinThickness:DiabetesPedigreeFunction	9.171e-03	3.563e-03	2.574	0.01026 *
SkinThickness:Age	6.188e-05	1.055e-04	0.587	0.55756
SkinThickness:DietVegan	-3.645e-03	2.897e-03	-1.258	0.20865
SkinThickness:DietVegetarian	2.452e-03	2.843e-03	0.862	0.38873
Insulin:BMI	-3.264e-05	2.805e-05	-1.164	0.24500

```
Insulin:DiabetesPedigreeFunction          -9.378e-04 3.492e-04 -2.685 0.00741 **
Insulin:Age                            2.563e-05 1.472e-05  1.741 0.08206 .
Insulin:DietVegan                         9.596e-04 4.300e-04  2.232 0.02595 *
Insulin:DietVegetarian                   -2.951e-05 4.112e-04 -0.072 0.94281
BMI:DiabetesPedigreeFunction             -6.990e-03 6.041e-03 -1.157 0.24758
BMI:Age                                -6.948e-05 2.246e-04 -0.309 0.75715
BMI:DietVegan                             2.373e-03 5.545e-03  0.428 0.66880
BMI:DietVegetarian                       -6.658e-03 5.226e-03 -1.274 0.20306
DiabetesPedigreeFunction:Age             -6.289e-03 5.440e-03 -1.156 0.24803
DiabetesPedigreeFunction:DietVegan       -1.908e-01 1.137e-01 -1.679 0.09365 .
DiabetesPedigreeFunction:DietVegetarian -3.691e-02 1.174e-01 -0.314 0.75326
Age:DietVegan                          -3.941e-03 3.802e-03 -1.036 0.30036
Age:DietVegetarian                       -4.066e-03 4.017e-03 -1.012 0.31175
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1546964)
Null deviance: 174.48 on 767 degrees of freedom
Residual deviance: 110.30 on 713 degrees of freedom
AIC: 801.11

Number of Fisher Scoring iterations: 2
```

The significant interactions from glm model are: Pregnancies:Age , Insulin:DiabetesPedigreeFunction. Insulin and Age have not been recognized as efficient contributors towards determining the results hence these interactions are ignored.

Results of 2-way ANOVA

Furthermore, to determine the effect of individual variables when in contact with other significant contributors 2-way ANOVA is used

a. 2-way ANOVA for Pregnancies and Glucose

```
                Df Sum Sq Mean Sq F value   Pr(>F)
DB$Pregnancies   1   8.59    8.59   49.84 3.75e-12 ***
DB$Glucose       1  34.02   34.02  197.36  < 2e-16 ***
Residuals      765 131.87    0.17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insights: Both the variables are significant when interacted upon.

b. 2-way ANOVA for Glucose and BloodPressure

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

```
DB$Glucose          1   37.98   37.98 212.894 <2e-16 ***
DB$BloodPressure    1    0.01    0.01   0.038  0.846
Residuals         765  136.49    0.18
---
Signif. codes:   0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insights: Only Glucose is significant when it encounters BloodPressure to determine the diabetic outcome

c. 2-way ANOVA for Glucose and BMI

```
             Df Sum Sq Mean Sq F value   Pr(>F)
DB$Glucose    1   37.98   37.98  223.68 < 2e-16 ***
DB$BMI        1    6.59    6.59   38.81 7.7e-10 ***
Residuals   765  129.90    0.17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insights: Both BMI and Glucose are significant enough when interacted with each other. P-value corroborate for the same.

d. 2-way ANOVA for Pregnancies and BMI

```
                Df Sum Sq Mean Sq F value   Pr(>F)
DB$Pregnancies   1    8.59   8.591   43.43 8.18e-11 ***
DB$BMI           1   14.55  14.554   73.57  < 2e-16 ***
Residuals      765  151.33   0.198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insights: Both Pregnancies and BMI are equally significant to determine outcome of diabetes

e. 2-way ANOVA for DiabetesPedigreeFunction and BMI

```
                            Df Sum Sq Mean Sq F value    Pr(>F)
DB$DiabetesPedigreeFunction  1   5.27   5.273  25.79  4.78e-07 ***
DB$BMI                       1  12.81  12.808  62.65  8.67e-15 ***
Residuals                  765 156.40   0.204
---
Signif. codes:  0 ‘***’ 0.001  ‘**’ 0.01 ‘*’ 0.05 ‘.’0.1‘’1
```

Insights: Both DiabetesPedigreeFunction and BMI are significant to determine outcome of diabetes in India women

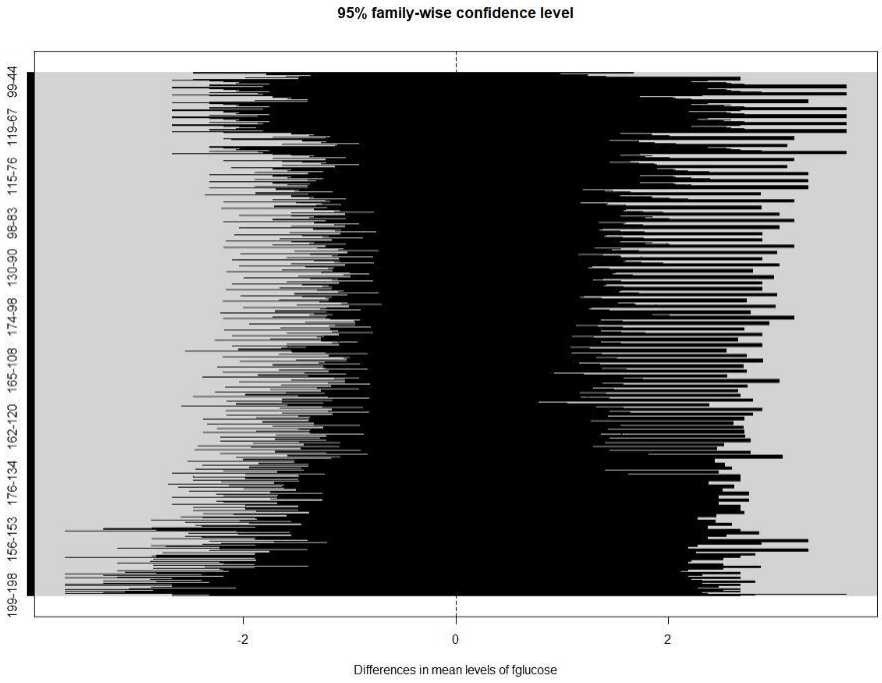
Tukey's Test

The Tukey Test (or Tukey *procedure*), also called Tukey's Honest Significant Difference test, is a post-hoc test based on the studentized range distribution. An ANOVA test can tell you if your results are significant overall, but it won't tell you exactly where those differences lie. After you have run an ANOVA and found significant results, then you can run Tukey's HSD to find out which specific groups means (compared with each other) are different. The test compares all possible pairs of means.

Assumptions for the test

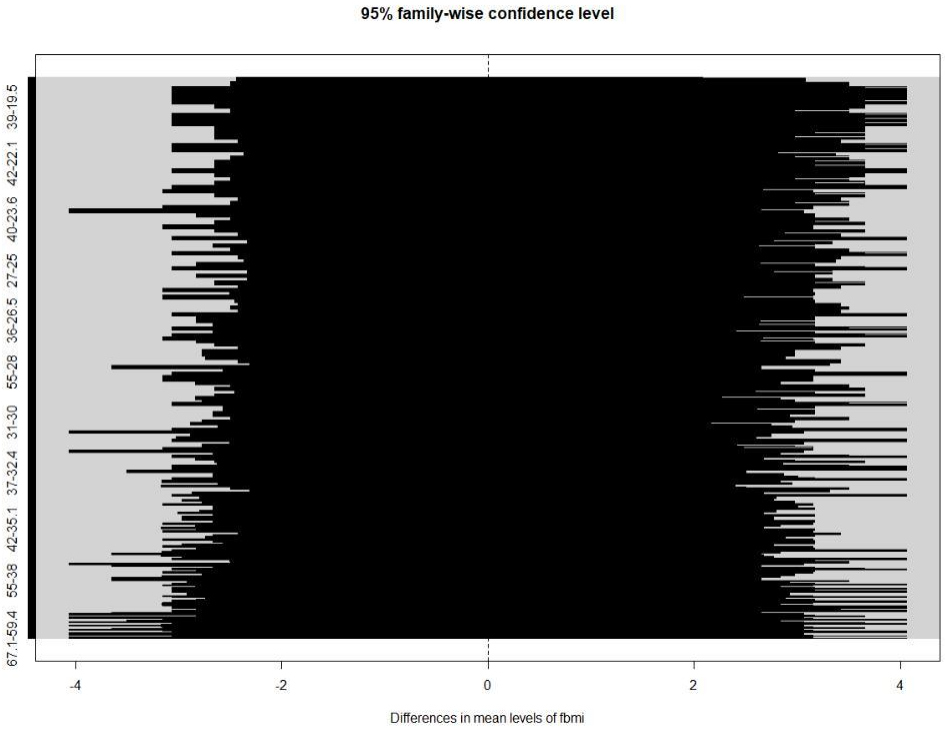
- Observations are independent within and among groups.
- The groups for each mean in the test are normally distributed.
- There is equal within-group variance across the groups associated with each mean in the test (homogeneity of variance).

1. Glucose



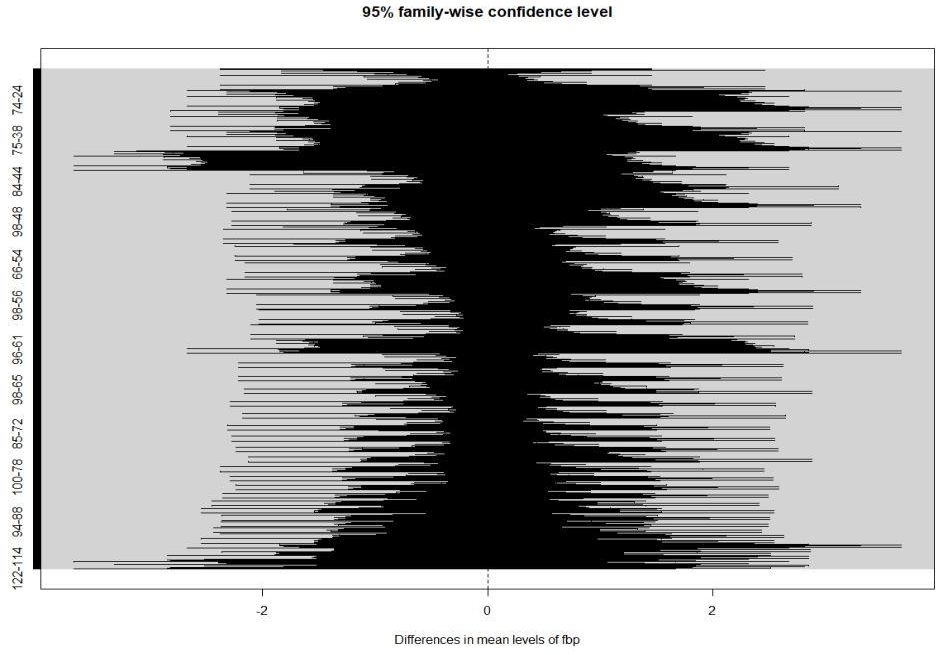
The plot tells us that there's difference in means of Glucose at levels 198-199 and 153-156. The rest of the levels have pretty much no significant mean difference. TukeyHSD() divides the Glucose range of values into different intervals and plots the intervals to show the difference in means of the levels Women with Glucose levels at 198-199 and 153-15 have higher chances of being diabetic.

2. BMI



TukeyHSD plot tells us that there is significant difference in the means from levels 19.5-39 and 23.6-40 and at 28-55,30-31. Women with BMI in the range of 19.5-39, 23.6-40, 28-55, 30-31 are more likely to be diagnosed as diabetic.

3. BloodPressure



The difference in means is constant within the intervals 44-88,114-122 and there's a difference between the intervals 24-74 and 38-75. This tells us that people with BP in the range are prone to be diabetic.

From the above observations, looking at Tukey's test and ANOVA values the strongest contributor to determining the diabetic outcome is Glucose.