Catalin Cris
Birchacherstrasse 34
3184 Wünnewil

Data Science Project

Project Diabetes

Conceptual Design Report

17 September 2019

ABSTRACT

The current project aims at analyzing the Diabetes data set provided by the UIC Machine Learning Repository.

This data set provides measured outpatient management data from 70 diabetic patients. The data are organized as one file per patient, each file including a data table.

The data were pooled in a unique data file and analyzed using descriptive and inferential statistical methods.

Our results illustrate an important interpatient variability of blood glucose control and insulin regimes. The results also suggest that two artificially generated sub-samples originate from different yet similar patient populations.

TABLE OF CONTENTS

OBJECTIVE

Diabetes mellitus is a chronical disease, which shows as a deficiency of insulin action due to one of the following causes:

a) Low or absent production of insulin by the beta islet cells of the pancreas subsequent to an auto-immune attack;
b) Insulin-resistance typically associated with older age and obesity, which leads to a relative insulin-deficiency even though the insulin levels might be normal.

The goal of the diabetes mellitus therapy is to maintain blood glucose ranges within a pre-defined range. This is achieved by delivery of exogenous insulin, combined with diet and exercise. This being said, it becomes clear that the diabetes mellitus therapy depends on various factors and influences patients' daily life.

The goal of the current project is to provide a statistical analysis of the "Diabetes" data set provided by the UIC Machine Learning Repository at Diabetes.

This data set contains data collected from 70 diabetic patients, i.e. suffering from insulin insufficiency, during outpatient management, i.e. therapy provided outside clinical facilities, in particular at home. This therapy relies on three interventions: diet, exercise and exogenous insulin. The entries in the data set document events are related to these interventions, either related to measurements as e.g. for insulin doses or plain entries as e.g. meals or hypoglycemic symptoms. As outpatient management is very demanding from the patients in terms of self-discipline and many environment influences on the glucose metabolism are unknown or difficult to quantify, e.g. stress, the control of blood glucose, i.e. the blood glucose values, may strongly vary among patients as well as for the same patient within different time intervals. On top of the mentioned variation factors, genetic factors play a part as well further influencing blood glucose control.

Our first objective is the illustration of the differences in blood glucose control and insulin therapy among patients, the so-called intrapatient variability. A second objective is to uncover possible correlations between the variables included in the data set and a third objective is to compare patient groups with respect to their origin from the same or different populations.

METHODS

Which infrastructure, tools, software libraries, statistical methods etc will be used

The machine used for the project is an **HP EliteBook 840 G1** using an **Intel(R) Core(TM) i7-4600U CPU @ 2.10GHz, 2701 MHz, 2**.

The operating system is **Microsoft Windows 10 Enterprise**.

The **Anaconda3-2019.07-Windows-x86_64** software distribution is used to perform the data analysis.

The development environment is **Jupyter Notebook 6.0.0**.

The code is written in **Python 3** and uses the following libraries: **Pandas**, **Numpy**, **SciPy** and **Matplotlib**.

We use descriptive statistical measures to assess the variability of the blood glucose values. Therefore, mean and median values as well as spread measures are calculated. The spread measures include variance, standard deviation, skewness and kurtosis. We will represent the calculated values in a tabular format in this report. Besides the use of such measures, we provide graphical illustrations of the data, which are meant to allow a visual variability assessment. Plots combining data measured for the same patient at different time points give an impression of intrapatient variability while plots combining data measured for different patients represent interpatient variability qualitatively.

We use inferential statistical methods in order to achieve the second objective as described in the previous chapter. We calculate the covariance matrix and the correlation coefficients between blood glucose values and the interventions influencing them in order to assess their interdependence. The interdependence between the administered insulin doses and the resulting blood glucose values is of particular interest in this context.

We also quantify the effect of the different intervention types, e.g. meal ingestion, on the blood glucose values using statistical tests. Our data set contains paired as well as independent samples. The paired samples include measurements related to same patients within different time intervals while measurements originating from different patient groups build the independent samples. Therefore, statistical test for paired as well as independent samples are used to assess the influence of the different interventions.

DATA

As already mentioned in the previous chapters, the “Diabetes” data set provided by the UIC Machine Learning Repository at Diabetes is used for this project.

This data set contains data collected from 70 diabetic patients. Data collected from one patient is contained in the file **data-<patient number>** where patient number is an integer running from 1 to 70. This means that the data set is composed by 70 such files. We have merged these files into a unique raw data file called **TotalData**. Merging the data allows, on one hand, less complex coding leading to higher coding efficiency, and simplifies, on the other hand, data management. However, we have saved all data files in the GitHub repository cris954 in order to allow reproducibility of our results.

The data are provided as a 29329 rows x 4 columns array. Each row represents data from an event at a specific time point. As explained in chapter “Objective”, events might be related to measurements, e.g. blood glucose measurements, or to interventions without connection to measured data, e.g. meals. The columns include the following values: event date, measurement time, code specifying the event type and value. Whenever an event relates to a measurement, the value field displays the measured value. If an

event is not connected to any measurement, the value field displays a zero value. The code values, formatted as integers, are explained in the metadata file **Data-Codes**.

Figure 1 below shows the first five rows of the data set:

Patient	Date	Time	Code	Value
1	04-21-1991	9:09	58	100
1	04-21-1991	9:09	33	9
1	04-21-1991	9:09	34	13
1	04-21-1991	17:08	62	119
1	04-21-1991	17:08	33	7

Figure 1

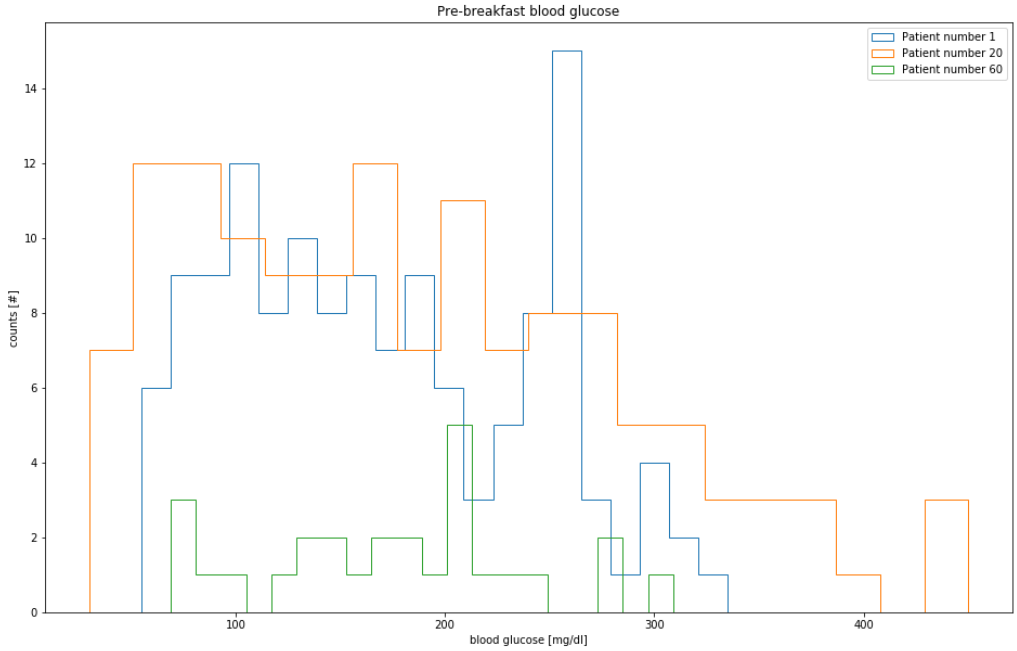


Figure 2

The histograms in Figure 2 illustrate the variability of measured blood glucose between patients. We chose patients 1, 20 and 60, as their data sets contain enough measured values corresponding to this event code as to allow a good illustration of this variability. This variability originates from various factors, e.g. gene pool, therapy quality, i.e. prescription as well as compliance, accurate assessment of meal composition and quantity, etc. However, differences between these histograms might also be explained by the data collection, as the number of data points is clearly lower for patient number 60 then for the other two patients.

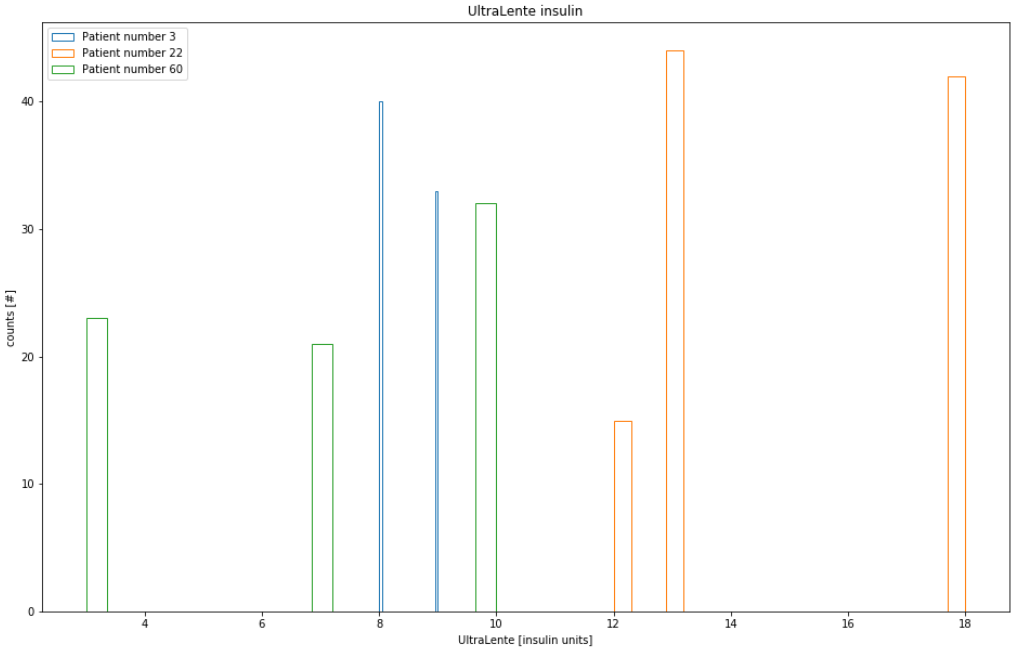


Figure 3

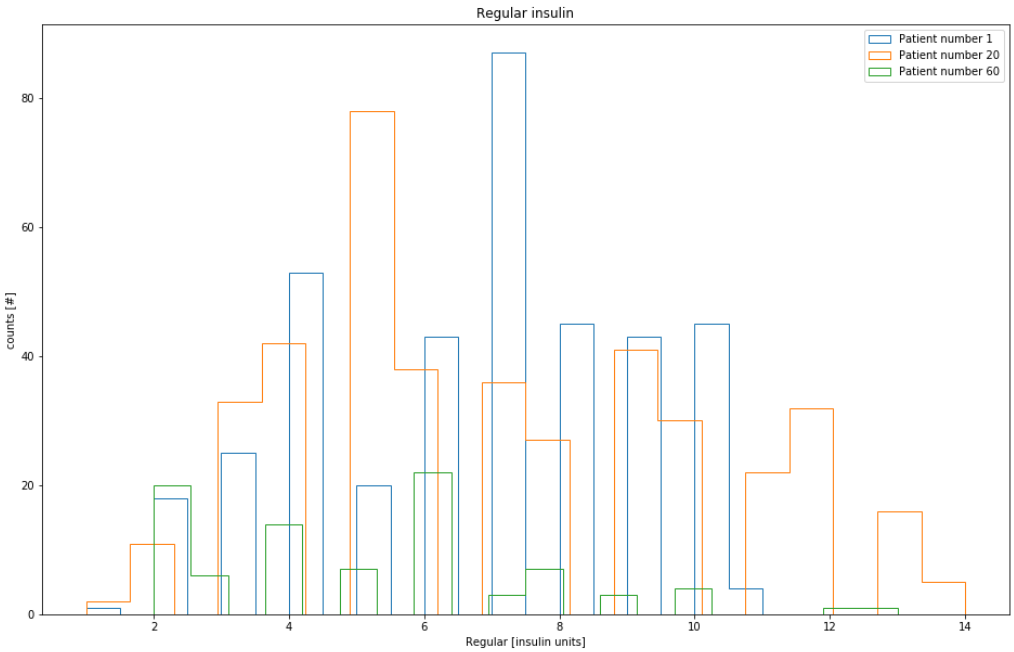


Figure 4

Figures 3 and 4 show the distributions of the administration of two different insulin types for patients 3, 22 and 60 (s. Figure 3) and patients 1, 20 and 60 (s. Figure 4). We chose these particular patients as their data sets offer enough values for satisfactory plots.

As can be seen from both figures, insulin dose distributions are characterized by several discrete values. However, the regular insulin distributions comprises more values then the ultralente insulin distributions. This difference originates from the different use of these insulin types: while ultralente insulin, due to its long-lasting action ([2]) as the name points out, is used for control over typically 24 hours ([2]) and, therefore, is administered once a day, regular insulin ([3]) is used for control of blood glucose surges after meals being, consequently, administered more several times over the day.

METADATA

What metadata is required for reproducing your analysis.

Where do you store the metadata, how can people access it.

The following files contain the metadata related to the "Diabetes" data set: **Data-Codes**, **Domain-Description**, **README** and **README-DIABETES**. These files are saved in the GitHub repository, together with the Jupyter notebook and the current report.

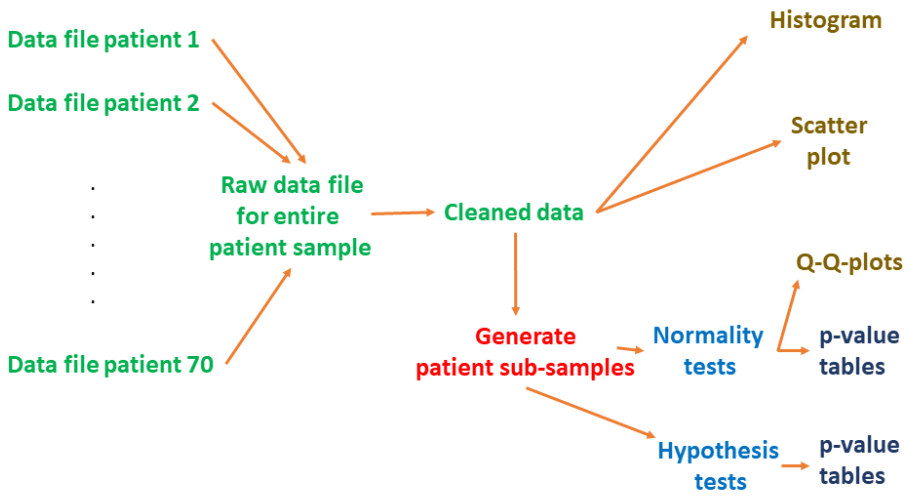
DATA QUALITY

The data set should provide enough data for reaching a significant statistic accuracy, allow calculate various statistics, and apply statistical tests.

Upon data inspection, it became clear that the number of missing values was small when compared to the total number of data points. In fact, 67 out of 23329 rows included missing values, corresponding to approximately 0.23% of the entries. Merely replacing the missing values by e.g. row or column average values might not be reasonable, as data points of various categories, e.g. measurement date and measured blood glucose value build the rows while the columns either include discrete values as event codes or measured blood glucose values corresponding to different event types, e.g. per-breakfast or pre-lunch blood glucose, or to different patients. Therefore, we preferred simply deleting the rows including missing entries given that we do not expect such a small number of entries to have a significant impact on our results. We stored the clean data in a file called **DiabetesData_NaN_removed** which can be found in the Github repository mentioned in chapter **DATA**.

More serious a problem is the reduced number of entries for same patients or some particular events, e.g. post meal measurements impeding the calculation of sound statistics in these cases. We tried to circumvent this problem by selecting events, e.g. pre meal measurements, and patients for our analysis, which provide enough data for our purposes. Additionally, we pooled data for all patients for specific events and analyzed these pooled data. Of course, this approach does not allow to draw conclusions for about blood glucose control of specific patients but leads to conclusions for the entire patient sample.

DATA FLOW



DATA MODELS

Conceptual

We analyze data from a set of 70 diabetic patients. The subset for each patient includes information about events concerning his diabetes therapy outside clinical facilities, i.e. during daily life, in particular at home. There is no mutual influence between data collected from different patients.

Logical (with dataframes and with databases)

Data are organized in 70 files; each of them included data collected from one patient. We have merged these data into one single file in order to simplify the coding of the data analysis application and data management. Data are given as a 29329 rows x 4 columns array. They contain text data.

Metadata is comprised in four text files listed in the corresponding chapter above.

We generate a dataframe containing the entire data set which we use for our analysis.

Physical (infrastructure needs)

The analysis was carried out using the infrastructure listed in chapter **METHODS**.

The required storage is divided as follows:

- Data set: 803 KB;

- Metadata files: 13 KB
- Code file: 253 KB
- Figure files: 248 KB

RISKS

One risk is the loss of the data set meaning the impossibility of carrying out our analysis. However, the data set availability at Diabetes mitigates this risk as the data set can be readily downloaded from this site.

Other risks are the loss of the code at some time point during the development or the mix up of versions, which does not allow further coding. In order to mitigate the former we save the code on a memory stick as well as on a desktop after each coding session. The latter is provided for by using the version management control offered by the checkpoints mechanism within Jupyter Notebooks.

Another risk is the loss of the current, which would impede documenting the project properly. We mitigate this risk as explained for the code loss above.

As to the quality of the data set, we listed the risks for our analysis related to this factor in the chapter **DATA QUALITY**. We explained in this chapter how we think to reduce these risks and what measures we used to improved data quality.

PRELIMINARY STUDIES

As already described in chapter “Objective” our analysis had a threefold objective:

1) Illustrate interpatient variability;
2) Discover correlations between the measured variables included in the data set;
3) Compare patient samples in order to assess whether they are drawn from the same or different patient populations.

Intrapatient variability is illustrated in Figures 2 to 4 in chapter “Data”. As commented in this chapter, intrapatient variability shows as well in the measured blood glucose values as in the administered insulin regime.

In order to reach our second objective, we identified measured variables, which are suitable for exploring correlations. Such correlations might appear between pre- and post-meal blood glucose values as one could expect that e.g. high pre-meal blood glucose values be followed by similar post-meal values. In order to test this hypothesis visually, we generated scatter plots showing pre-meal blood glucose values along the x-axis and post-meal blood glucose values along the y-axis summarized pooled together over all patients. Such plots were generated for the three meal types included in the data set: breakfast, lunch and supper.

The plots are shown in the following figures.

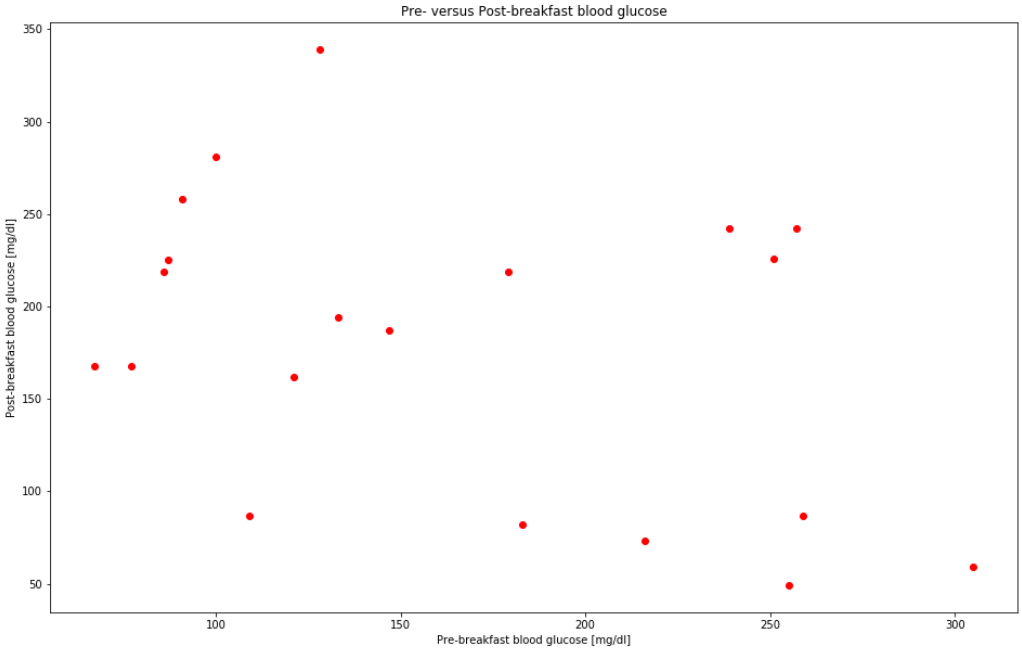


Figure 5

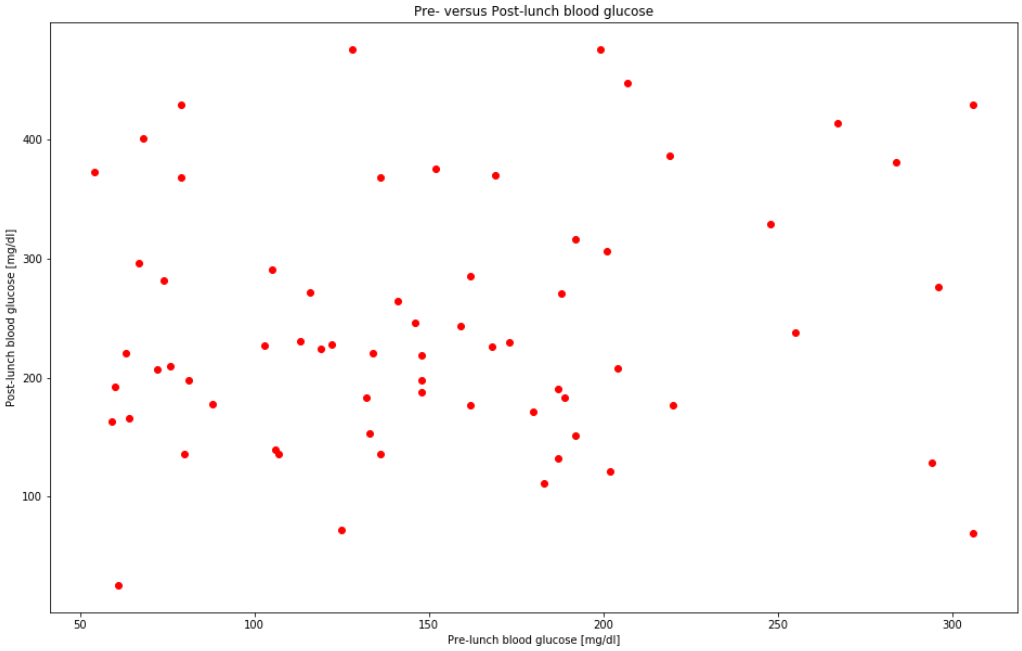


Figure 6

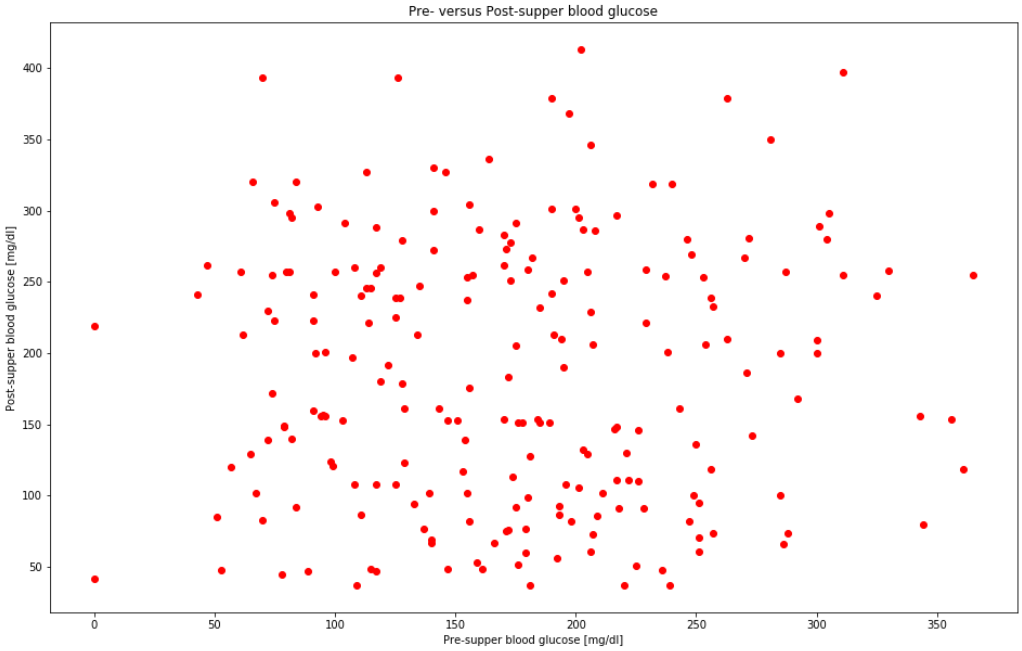


Figure 7

As can be seen from Figures 5 to 7, pre- and post-meal blood glucose values show no correlation, at least when pooled together over the entire patient sample. This effect might be due to the pooling and such correlations might appear if we analyzed the values for individual patients. However, the data might not be rich enough for some patients in order to show statistically significant correlations.

We focused our efforts on the second objective subsequently.

As the patients included in our data set build a unique sample, we split this sample in two sub-samples artificially. In order to reduce the subjectivity involved in such a procedure, we chose to determine the first sub-sample at random and define the second sample as being the complement to the first one with respect to the patient sample included in our data set.

Both patient sub-samples are stored in the current version of our code. Due to a coding error, namely selecting random integers starting from zero as lower limit, the first sub-sample includes twice the integer zero. As no patient index in the data set corresponds to zero, this value was ignored during the step described below.

For the sake of illustration, we pooled together the pre-breakfast blood glucose values per sub-sample for all patients. This step led to two blood glucose distributions, represented as histograms in Figure 8:

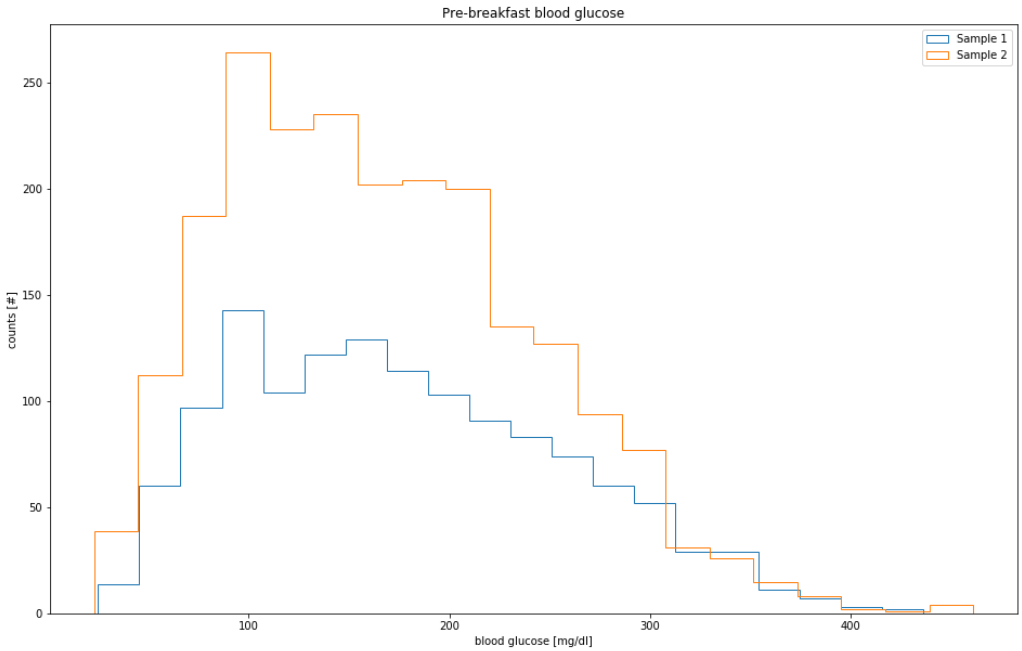


Figure 8

The legend in Figure 8 mentions sample 1 and 2. They correspond to the first and second sub-samples described above. The histograms show a large overlap and similar distribution shapes. However, the sub-sample 2 exhibits a larger number of values. The histograms suggest skewed distributions with longer tails towards larger blood glucose values. This observation seem to cope well with the fact that diabetic patients spend more time in the hyperglycemic (blood glucose values higher that the target) regime than in the hypoglycemic (blood glucose values lower that the target) regime.

Our next goal was to verify our visual impressions by against results obtained from statistical hypothesis tests.

Therefore, we first checked the normality of the sub-sample distributions. We first chose a visual approach using Q-Q-plots. The results are shown in Figures 9 and 10:

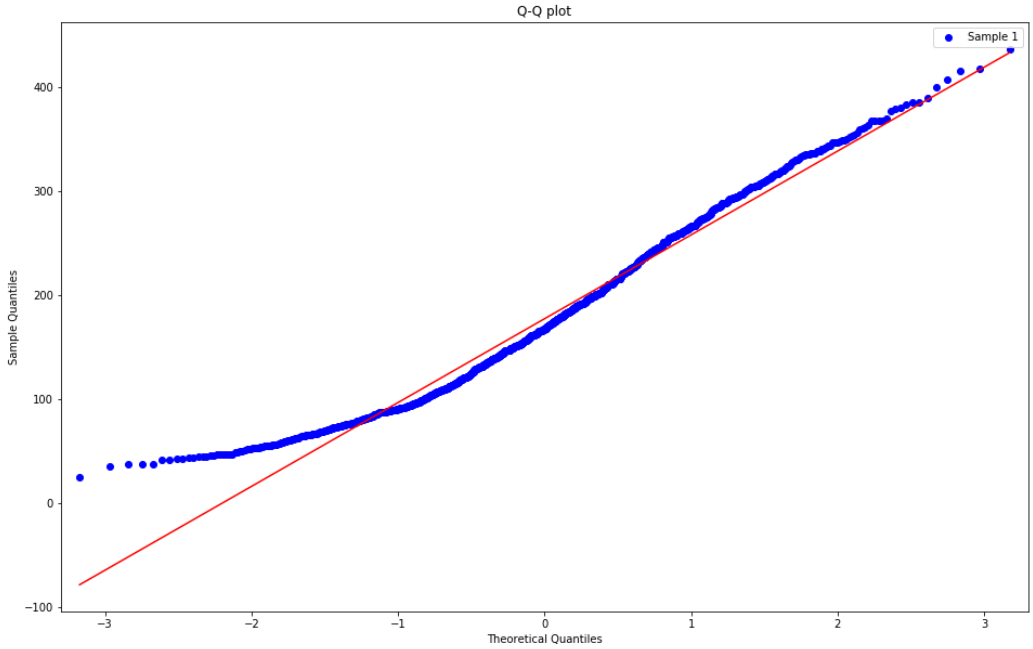


Figure 9

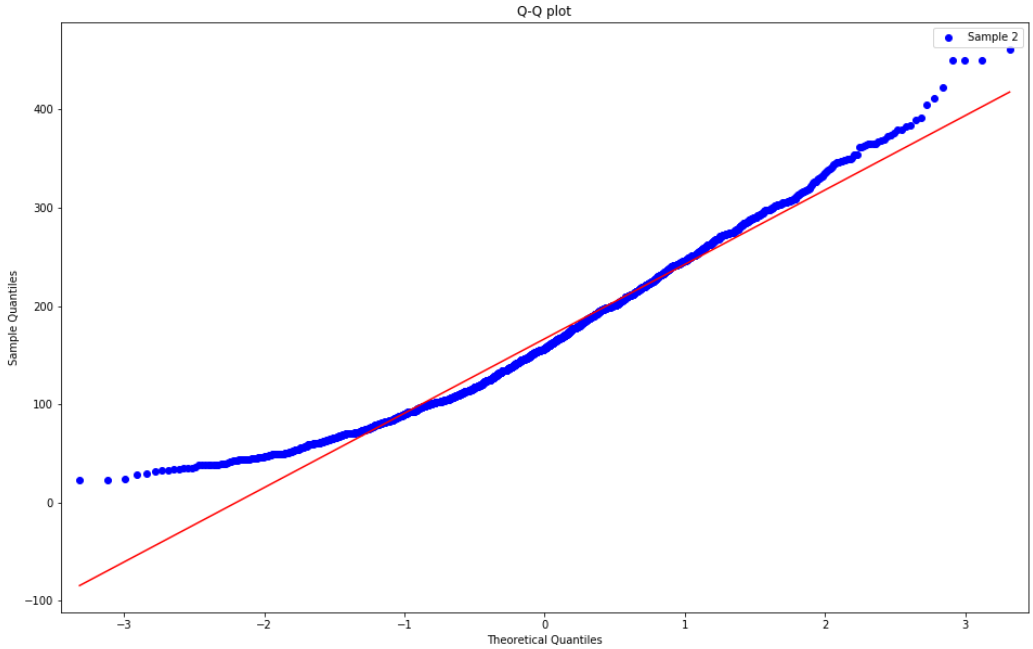


Figure 10

Figures 9 and 10 show deviations from normality for both blood glucose distributions. While the distribution of sub-sample 1 mainly deviates in at low values, the distribution of sub-sample 2 deviates as well for low as high values but with the largest deviations still exhibited at low values. This observation correlates with the hard cut observed at low blood glucose values in the histograms in Figure 8.

We performed normality tests during next step in order to check the visual assessment gained from the histograms and Q-Q-plots. We chose the D'Agostino and Pearson's (AP) and Shapiro-Wilk (SW) tests performed at a confidence level of 0.05. The corresponding p-values are shown in the table in Figure 11.

sample [#]	p-value (AP)	p-value (SW)
1	8.696e-16	1.420e-15
2	7.254e-22	8.343e-20

Figure 11

As can be seen from this figure, the p-values for both sub-samples resulting from both tests are much smaller than the confidence level. Therefore, the H0 hypothesis can be rejected in all cases, i.e. both sub-samples do not originate from normal distributions.

Therefore, we applied non-parametric tests to determine whether the blood glucose distributions of originate from the same populations. We performed the Mann-Whitney and Kolgomorov-Smirnov tests at a confidence level of 0.05. The corresponding p-values are shown in the table in Figure 12,

Test	p-Value
Mann-Whitney	2.873e-04
Kolgomorov-Smirnov	3.450e-03

Figure 12

The H0 hypothesis can be rejected for both tests as the p-values lie below the confidence level. Therefore, the results from both tests confirm the H1 hypothesis, i.e. the sub-samples originate from different populations.

Finally, we looked into the correlation of the blood glucose values corresponding to the sub-samples. We did so by generating a scatter plot of the blood glucose value distribution of the first sample versus the blood glucose value distribution of the second sample. Despite the fact that no correlation emerges from this scatter plot we overlaid a regression line on top of it in order to illustrate this method two. This regression line is close to horizontal confirming the lack of correlation between the blood glucose values as visually suggested by the scatter plot. This lack of correlation is as well confirmed by the correlation coefficients between the blood glucose distributions corresponding to both sub-samples.

The scatter plot and regression line are illustrated in Figure 13 while the correlation coefficients table is shown in Figure 14.

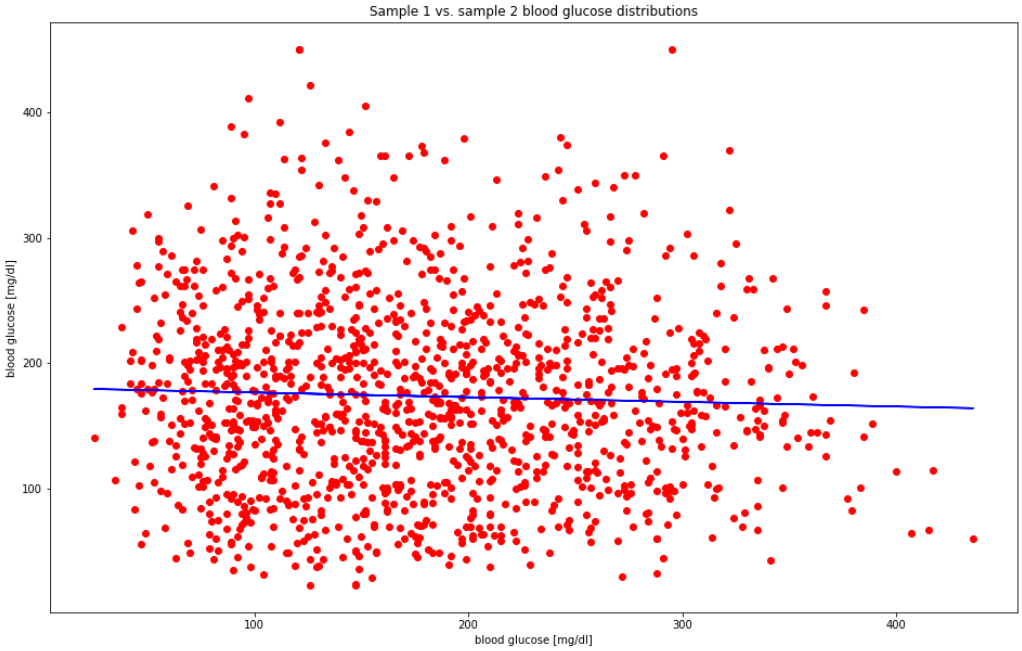


Figure 13

sample [#]	1	1
1	1.00	-0.04
2	-0.04	1.00

Figure 14

CONCLUSIONS

The data set Diabetes includes enough data points in order to allow a coherent exploration of the data. The data are not heavily flawed, i.e. data cleaning was not a tedious process and the cleaned data set still provided enough data for the exploration.

The methods chosen for the current analysis allowed us to reach our objectives.

Intrapatient variability was illustrated in chapter “DATA” using methods of descriptive statistics. This variability is important and can be explained by e.g. genetic factors or compliance to the prescribed therapy as explained in chapter “OBJECTIVE”.

We combined methods of descriptive and inferential statistics to obtain the results included in chapter “PRELIMINARY STUDIES”. The scatter plots presented in this chapter suggest that no correlations exist between pre- and post-meal blood glucose values. Therefore, this issue was not further investigated. In order to reach our third objective defined in chapter “OBJECTIVE” we constructed two artificial sub-samples of the patient sample represented in the data. We compared overall blood glucose values measured within these sub-samples, again using graphical methods as well as hypothesis testing. The results show that the sub-samples originate from different populations. The blood glucose values in both sub-samples are not correlated as shown by the scatter plot of the first versus the second distribution and emphasized by the gradient of the regression line.

We notice that the results obtained from the comparison of both sub-samples clearly depend on the random generation of the first sample. It would be interesting to explore this issue in future by generating a large set of sub-samples and investigating the distribution of the results presented in the previous chapter over this set.

REFERENCES

[1] Diabetes : the data set site at the UIC Machine Learning Repository

[2] Article “Ultralente” in Wikipedia Ultralente

[3] Article “Regular Insulin” in Wikipedia Regular_insulin