Biostatistics - Dr. Patrick
BMEN 350; Section 201
Saurabh Dhole 626 002 135
September 15th, 2021

Project 1

A. Read and Reflect

Upon reading the esteemed article called "Error Bars" by Dr. Krzywinski, I learned about some of the technical aspects involved in uncertainty when performing statistical analysis. I was deeply intrigued by a few of Dr. Krzywinski's points on uncertainty and I would like to reflect on them here.

Firstly, I was educated on the three types of error bars in statistical analysis. These three types are as follows: standard deviation (s.d.), standard error of the mean (s.e.m.), and confidence interval (CI). I learned that these three types of error bars serve different purposes. For example, error bars given in s.d. inform about the spread of the population and are useful when looking at the range of a sample. I also learned that the s.e.m. informs about the uncertainty in the mean and the dependance on sample size. I observed that as the sample size increases, the s.e.m. error bars shrink in size. As a Biomedical Engineer, this information on these three types of error bars is very crucial. I say this because there are going to be many times in internships and research projects when I will have to include error bars on collected data. I must know the differences between the three types of error bars shown in this article so that I can use them wisely in creating plots. This information on these three types of error bars was very useful for me as I initially did not know about the s.e.m. error bar, even though I had heard of the other two types of error bars before. Reading about s.e.m. in this article will serve as an introduction, and I shall study more into the topic as we will be applying the concept of s.e.m. in the BMEN 350 course soon. I plan to use the error bar of s.e.m. to show how much dependency there is on the sample size in my undergraduate research projects as well as in course work projects.

Secondly, I was enlightened on misinterpretation upon reading Dr. Krzywinski's article. Dr. Krzywinski has astutely stated that one can misinterpret the data that one is analyzing if he or she does not know what type of error bar is being shown. I learned that in journals such as Nature Methods, approximately 5% of the displayed error bars are not specified, meaning one cannot tell if it is a standard deviation, a standard error of the mean, or if it's a confidence interval. As a Biomedical Engineer, this information is indeed valuable. I say this because I personally know the importance of being clear, concise, and forthcoming when writing a publication. There will be many times when I will have to publish data along with error bars, and it will be important for me to let the readers know which kind of error bar is being displayed. This information about potentially misinterpreting data due to not knowing what type of error bar is being displayed is indeed important for me. I say this because the next time that I read a publication or a scholarly peer reviewed article, I will be sure to inquire about which type of error bar is

being used instead of just assuming that the error bar is standard deviation. I plan to use this information throughout my time in scholarly research and in the BMEN 350 course. I will be sure to improve clarity in my statistical analyses by indicating which type of error bar I used.

Finally, I was warned about the limited utility of error bars in small sample sizes. Dr. Krzywinski states that error bars involving small sample sizes are not robust. Dr. Krzywinski also states that it is often times better to just show the individual data values than to show error bars involving small sample sizes. As a Biomedical Engineer, this information has great significance. I say this because I now understand that even the error bars are not safe from small sample sizes. I had initially known that having a small sample size does not yield any favors in terms of statistical significance, but I now know that even error bars can be affected by small sample sizes. This information is indeed valuable for me as I will be sure to be more vigilant of sample size when performing undergraduate research so that I do not render my error bars useless. I plan to use this information in the BMEN 350 course by testing the effects of small sample size on error bars and tests for statistical significance. I also plan on using this information in undergraduate research when selecting the number of samples to run in experiments so that the error bars in the data can actually tell a story.

Upon reading the knowledgeable article called “Ten Simple Rules for Better Figures” by Dr. Rougier, I was happy to learn about some straightforward techniques and guidelines that I can follow to make better figures. I was enlightened upon a few very thought-provoking techniques that yield better figures and I would like to reflect on them here.

Firstly, I learned about the importance of knowing one’s audience. Dr. Rougier states in the article that knowing the audience goes hand in hand with the intent of the conveyor. I observed that the statistician presenting the figures must be forthcoming in regards to what he or she wants the audience to see. The statistician must keep in mind different professional bodies and critics who may see his or her figures. I found it interesting that the statistician must cater to the different needs of different audiences when presenting figures. As a Biomedical Engineer, this information about the importance of knowing one’s audience can go a long way. I can see myself simplifying complex figures to display information to fellow interns, and I can also see myself including details and professional standards for when I present figures to principal investigators. This information of knowing one’s audience is very useful in a practical sense. I may perhaps, before giving a presentation, read up on the qualifications and experiences of the audience members so as to ensure my communication of figures is clear. I plan on using this information in internships and undergraduate research as it will help me communicate my thoughts more appropriately with faculty, and supervisors.

Secondly, I was educated on the importance of adapting the figure to support the medium. Dr. Rougier states in his article that the figures that one presents may have to change depending on what medium is being used to present said figures. For example, the colors and textures on a bar graph may need alteration depending on if the bar graph is being presented on a poster or on a screen via power point. Additionally, the figure labels may need to be resized according to the medium being used. For example, figure labels may need to have a larger font size if they will be displayed on a poster. As a Biomedical Engineer, this information is very important to me. I say this because there are so many instances where I have to give presentations in which I present multiple figures. I will remember to take into account the medium of presentation before presenting figures in my next research presentation. The information presented here is very useful on a daily basis as it serves as a reminder to always ask oneself if one's figures are well suited for the medium of presentation. This can be in any setting, private, academic, or personal. I plan on using this information in the BMEN 350 course when it is time to submit projects that include figures.

Finally, I was enlightened on the importance of being forthcoming and not misleading one's audience members. Dr. Rougier states very clearly in his article that one must be as objective as possible when presenting figures. I observed that the figures must have appropriate scales and sufficient number of ticks on the axes. Dr. Rougier states that it's easy to inadvertently mislead your audience, this is not ideal. As a Biomedical Engineer, this information is crucial. I say this because when presenting figures on the performance of a gel for example, it is important for me to use an appropriate scale, and to include a sufficient number of tick marks so that the performance parameters in the figure are easily interpreted. There should be no funny business such as zooming in on the scale to show a large difference (like in the study from the UK shown in class). This information is very valuable in an academic sense as well. I say this because when submitting a presentation or project, one must not try to force the audience to see something that isn't there. The presenter should simply be forthcoming in providing figures, and the audience should interpret for themselves. I plan to use this information in the BMEN 350 course when submitting figures, and explanations in the appendix.

After reading the informative article called "Reveal, Don't Conceal" by Dr. Weissgerber, I was delighted to learn that there are some helpful guidelines to follow when displaying figures on publications and presentations. I was exposed to a few wonderful techniques to improve the quality of figures when giving a presentation of submitting a publication. I reflect on these techniques and guidelines here.

Firstly, I learned about the importance of including flow charts and study design diagrams as figures in publications or presentations. Dr. Weissgerber states that figures such as flow charts and design diagrams make it easy for audience members to understand the study design and they also make it easy for the audience members to follow the participants or animals (if they were used) through the study. As a Biomedical Engineer, this information is very important to me. I say this because I can imagine myself using flow charts and design diagrams to show researchers and supervisors my thought process of how I went about experimenting about polymers during my quality control internship. On a practical level, this information is indeed very useful. I say this because many students such as myself are involved in laboratory courses in which we have to follow a certain protocol. It will certainly be helpful for students such as myself to draw flow diagrams before we go into lab so that we have a better idea of what to do in lab. I plan on using this information to clearly explain and display my thought process when designing experiment sin undergraduate research.

Something that was very interesting for me to learn in this article was that there are established zones of invisibility and zones of irrelevance when plotting a graph for a figure. Further, I learned that these established zones of invisibility and zones of irrelevance are backed by research in perceptual learning. Dr. Weissgerber explains that there are certain spaces or components that you can eliminate from the graph to make the figure more concise. The unneeded space at the top comprises of the irrelevance and can be removed from the figure. The space towards the bottom of the graph that shows no difference between bars on a bar graph is the invisible and can also be removed from the figure. As a Biomedical Engineer I imagine that I will have to display quite a few plots and graphs during internships and research projects. The information about irrelevance and invisibility is important to me as I will incorporate it into my figures to make them more concise. In a practical sense, the information about zones of irrelevance and zones of invisibility can apply to any figure or graphic that one may want to display (not just bar graphs). For example, if one is taking an image of fluorescently labelled cells, he or she must ensure that there is no irrelevant space in the image. I plan to use this information about zones of irrelevance and zones of invisibility in the BMEN 350 course when it comes time to submit module projects.

Finally, in this article I learned to be accommodative of the needs and visual abilities of all audience members. Dr. Weissgerber indicates that individuals who are red-green color blind may have difficulty visualizing figures that predominantly use the colors red and green. Dr. Weissgerber also mentions that in some cases, it may be best to use textures instead of colors. An example would be using textured shapes and lines on a bar graph instead of colors. This way, audience members who have visual difficulties will easily be able to decipher the information being conveyed. As a Biomedical Engineer this information is very important to me. I say this because I will be encountering many professionals through

internships and research projects, I must be able to clearly communicate data to all of my team members. This information serves a very practical purpose as well, if an individual in my team does have visual difficulties, I must be diligent enough to accommodate that individual. I plan on incorporating colors and textures that are not difficult to visualize for all audience members, especially those with visual difficulties. From this piece of information in Dr. Weissgerber's article, I learned the importance of inclusivity and accommodation in academic, clinical, and professional environments.

B. Problem 1

The mean pain score for the patients suffering from diabetic neuropathy was determined to be 40.2 $\pm$ 29.8 (mean $\pm$ std.dev.). The 25^{th}, 50^{th}, and 75^{th} percentile pain scores for these patients were determined to be 13.0, 29.5, and 70.0 respectively. The median is the 50^{th} percentile. Figure 1 displays a box plot and truncated violin plot of the pain scores for these patients. The key difference between the box plot and the truncated violin plot is that the truncated violin plot allows us to visualize the frequencies of pain scores along with the measures of central tendency. This is because we can see that the mean, 25^{th}, 50^{th}, and 75^{th} percentiles are displayed in both plots, but only the truncated violin plot displays the frequencies of the pain scores. If a pain score has a high frequency of occurrence among the patients with diabetic neuropathy, it shows as a wider portion on the violin plot. It can be observed from the truncated violin plot that the bottom portion of the plot is wider than the top potion of the plot. This indicates that the frequency of high pain scores (pain scores above the 75^{th} percentile) is less than the frequency of lower pain scores (pain scores below the 50^{th} percentile). It can also be observed from the violin plot that the mean pain score is greater than the median pain score. This is because the mean is affected by outliers on the high end. In addition, the shape of the violin plot indicates that the frequencies of lower pain scores is higher than the frequencies of higher pain scores. Hence, the data appears that it was not drawn from a normally distributed population as it is right skewed (mean > median). Due to such detailed information being conveyed by the truncated violin plot in such as concise manner, the truncated violin plot is indeed my preferred choice. Unlike the violin plot, the box plot does not allow us the ability to visualize the skewing of the data. This is why I would prefer the violin plot over the box plot.

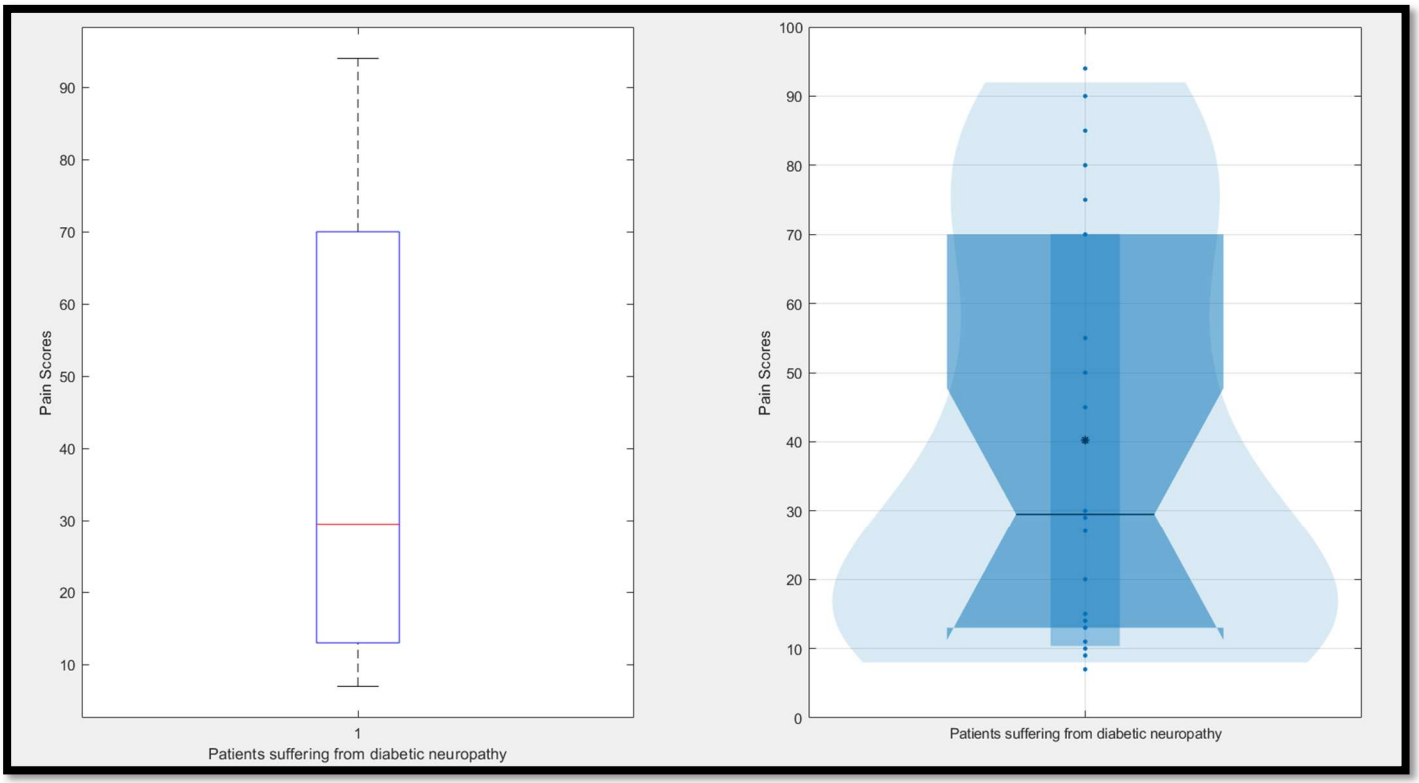


Figure 1: On the left, the box plot of the pain scores of the diabetic neuropathy patients can be observed. The mean pain score is 40.2 ± 29.8, and the 25th, 50th, and 75th percentile pain scores are 13.0, 29.5, and 70.0 respectively. On the right, the violin plot of the pain scores of the diabetic neuropathy patients can be observed. This violin plot has the same measures of central tendency as the box plot. The Matlab code for these two plots is provided in the Appendix.

C. Problem 2

The mean of the duration of hospital stay for the 25 patients was 8.6 days ± 5.7 days (mean ± std.dev.). The median of the duration of hospital stay for the 25 patients was 8.0 days. The 25th, 50th, and 75th percentiles of the duration of hospital stay for the 25 patients was 5.0 days, 8.0 days, and 11.0 days respectively. The range of the duration of hospital stay for the 25 patients was 27 days. Comparing the mean and median of the data set can allow us to guess which way the distribution will skew. If the mean is greater than the median, the distribution will be right skewed. If the mean is less than the median, the distribution will be left skewed. Figure 2 displays a scatter plot, histogram, and a box plot of the duration hospital stay for the 25 patients. The data set was split according to which patients received the antibiotic. The number of days in hospital for the patients who took the antibiotic was put into one group, and the number of days in hospital for the patients who did not take the antibiotic was put into another group. The mean duration of days in hospital for the antibiotic group was 11.5 days ± 8.8 days (mean ± std.dev.). The mean duration of days in hospital for the non-antibiotic group was 7.4 days ± 3.7 days (mean ± std.dev.). Comparing the mean duration of hospital stay of the antibiotic and non-antibiotic groups, it can

be observed that the antibiotic group had a longer hospital stay, and thus it looks like patients who took the antibiotic stayed longer in the hospital than the patients who didn't take the antibiotic. However, the mean does not tell the whole story! It can be observed that the distribution for duration of days in hospital is closer together for the non-antibiotic group, as the standard deviation is less than that of the antibiotic group. It was observed that the greater standard deviation for the antibiotic group was caused by an outlier. This outlier is patient 7 and this patient's stay in the hospital lasted for 30 days. Once patient 7's duration was removed from the antibiotic group, the new mean duration of hospital stay for the antibiotic group was 8.5 days $\pm$ 3.7 days (mean $\pm$ std.dev.). Now, if the new antibiotic group's mean and standard deviation are compared with the mean and standard deviation of the non-antibiotic group, it can be observed that the two distributions resemble each other. Thus, no comment can be made about whether or not duration of hospitalization is affected by whether a patient has received antibiotics.

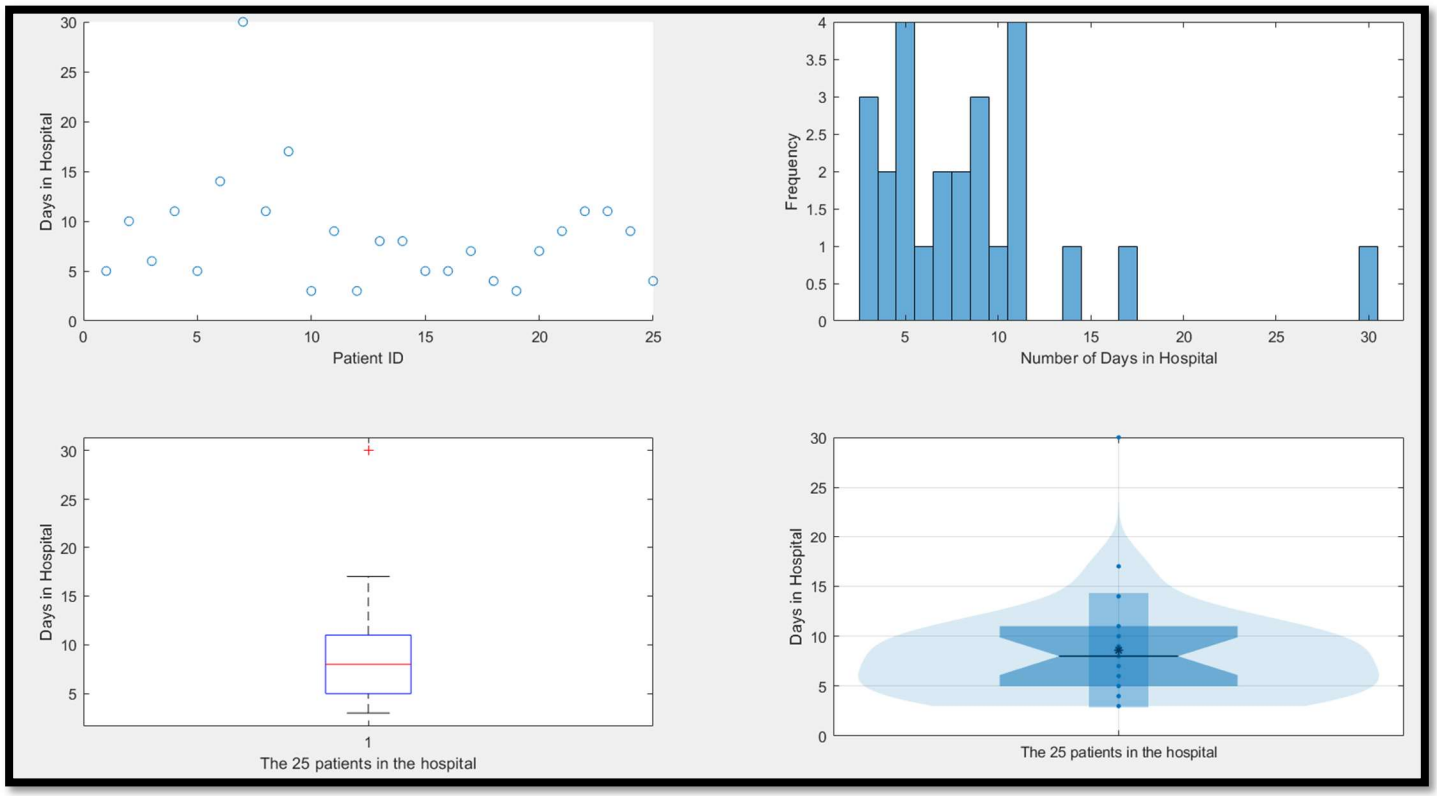


Figure 2: This is figure is an amalgamation of scatter plot, histogram, box plot, and violin plot of the duration of hospital stay for the 25 patients. On the top left is the scatter plot, on the top right is the histogram, on the bottom left is the box plot, and on the bottom right is the truncated violin plot. The mean of the duration of hospital stay for the 25 patients was 8.6 days $\pm$ 5.7 days (mean $\pm$ std.dev.). The median of the duration of hospital stay for the 25 patients was 8.0 days. The 25th, 50th, and 75th percentiles of the duration of hospital stay for the 25 patients was 5.0 days, 8.0 days, and 11.0 days respectively. Matlab code for these plots and measures of central tendency are provided in the Appendix.

D. Problem 3

The mean change in serum cholesterol for the 24 hospital employees was 22.1 mg/dL $\pm$ 13.1 mg/dL (mean $\pm$ std.dev.). The median serum cholesterol level for the baseline was determined to be 179 mg/dL. Individuals with baseline serum cholesterol levels above this median were said to have high cholesterol, while individuals with baseline serum cholesterol levels below this median were said to have low cholesterol. This is how the data was split. The mean change in serum cholesterol levels for the high cholesterol level group was determined to be 28.2 mg/dL $\pm$ 12.5 mg/dL (mean $\pm$ std.dev.). The mean change in serum cholesterol levels for the low cholesterol group was determined to be 16.1 mg/dL $\pm$ 10.9 mg/dL (mean $\pm$ std.dev.). Just by comparing the mean changes in serum cholesterol levels of the high cholesterol and low cholesterol groups, one may think that the mean change in serum cholesterol levels for the high cholesterol group was indeed much higher than that of the low cholesterol group. So, one may think that the effects of diet on cholesterol are indeed more evident in individuals with higher rather than lower cholesterol. However, the mean does not tell the whole story! It can be observed that the standard deviations of the high and low cholesterol groups, 12.5 mg/dL and 10.9 mg/dL respectively, do very much overlap. Since there is so much overlap in standard deviation of the high and low cholesterol groups, the difference between mean change in serum cholesterol levels of the high and low cholesterol groups is not statistically significant. Hence, we cannot comment descriptively on whether or not the effects of diet on cholesterol are more evident in individuals with high rather than low cholesterol levels. A more comprehensive statistical test will have to be performed in order to be able to comment on this issue. The calculations for this problem are displayed in the Appendix.

E. Appendix

Problem 1 Matlab code:

Project_1_Workings.m

```
%% Project 1, Problem 1

Imported_File = uigetfile('*.xlsx');
T = readtable(Imported_File);
Pain_Scores_Table = T(:,2);
Pain_Scores = table2array(Pain_Scores_Table);

Pain_Scores_Mean = mean(Pain_Scores);
Pain_Scores_STD = std(Pain_Scores);
twentyFifth = quantile(Pain_Scores,0.25);
fiftieth = quantile(Pain_Scores,0.50);
seventyfifth = quantile(Pain_Scores,0.75);

MEAN = sprintf('The mean of pain scores is: % .2f\n',Pain_Scores_Mean);
STD = sprintf('The standard deviation of pain scores is: % .2f\n',Pain_Scores_STD);
TwoFive = sprintf('The 25th percentile of pain scores is: % .2f\n',twentyFifth);
FiveZero = sprintf('The 50th percentile of pain scores is: % .2f\n',fiftieth);
SevenFive = sprintf('The 75th percentile of pain scores is: % .2f\n',seventyfifth);

disp(MEAN)
disp(STD)
disp(TwoFive)
disp(FiveZero)
disp(SevenFive)

subplot(1,2,1)
boxplot(Pain_Scores)
xlabel('Patients suffering from diabetic neuropathy')
ylabel('Pain Scores')

subplot(1,2,2)
al_goodplot(Pain_Scores);
xticks([1])
xticklabels('Patients suffering from diabetic neuropathy')
ylabel('Pain Scores')

```

Command Window

```
>> clear
>> Project_1_Workings
Warning: Column headers from the file were modified to make them valid MATLAB identifiers before creating variable names for the table. The
original column headers are saved in the VariableDescriptions property.
Set 'VariableNamingRule' to 'preserve' to use the original column headers as table variable names.
The mean of pain scores is:  40.21

The standard deviation of pain scores is:  29.84

The 25th percentile of pain scores is:  13.00

The 50th percentile of pain scores is:  29.50

The 75th percentile of pain scores is:  70.00

>>
```

UTF-8 | script | Ln 83 | Col 36

Problem 2 Matlab code and excel calculations:

Project_1_Workings.m

```
%% Project 1, Problem 2

Import_File = uigetfile('*.xlsx');
t = readtable(Import_File);
DaysInHospital_Table = t(1:25,2);
PatientID = t(1:25,1);
PatientNumber = table2array(PatientID);
DaysInHospital = table2array(DaysInHospital_Table);

Hospital_Days_Mean = mean(DaysInHospital);
Hospital_Days_StandardDeviation = std(DaysInHospital);
Hospital_Days_Median = median(DaysInHospital);
Hospital_Days_Range = range(DaysInHospital);

twentyfifthptile = quantile(DaysInHospital,0.25);
fiftythptile = quantile(DaysInHospital,0.50);
seventyfifthptile = quantile(DaysInHospital,0.75);

MEAn = sprintf('The mean days in hospital is: % .2f\n',Hospital_Days_Mean);
STd = sprintf('The standard deviation of days in hospital is: % .2f\n',Hospital_Days_StandardDeviation);
twofive = sprintf('The 25th percentile of days in hospital is: % .2f\n',twentyfifthptile);
fivezero = sprintf('The 50th percentile of days in hospital is: % .2f\n',fiftythptile);
sevenfive = sprintf('The 75th percentile of days in hospital is: % .2f\n',seventyfifthptile);
Ranges = sprintf('The range of days in hospital is: % .2f\n', Hospital_Days_Range);
Medians = sprintf('The median of days in hospital is: % .2f\n',Hospital_Days_Median );

disp(MEAn)
disp(STd)
disp(twofive)
disp(fivezero)
disp(sevenfive)
disp(Ranges)
disp(Medians)

subplot(2,2,1)
scatter(PatientNumber, DaysInHospital)
```

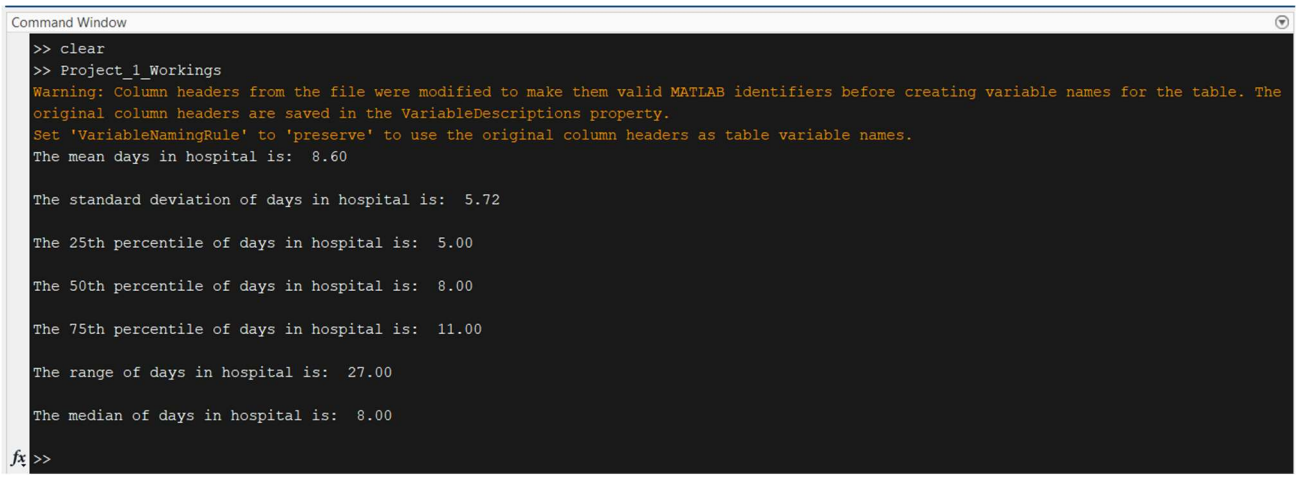
Project_1_Workings.m

```
subplot(2,2,1)
scatter(PatientNumber, DaysInHospital)
xlabel('Patient ID')
ylabel('Days in Hospital')

subplot(2,2,2)
histogram(DaysInHospital)
xlabel('Number of Days in Hospital')
ylabel('Frequency')

subplot(2,2,3)
boxplot(DaysInHospital)
xlabel('The 25 patients in the hospital')
ylabel('Days in Hospital')

subplot(2,2,4)
al_goodplot(DaysInHospital);
xticks([1])
xticklabels('The 25 patients in the hospital')
ylabel('Days in Hospital')
```



```
Command Window
>> clear
>> Project_1_Workings
Warning: Column headers from the file were modified to make them valid MATLAB identifiers before creating variable names for the table. The
original column headers are saved in the VariableDescriptions property.
Set 'VariableNamingRule' to 'preserve' to use the original column headers as table variable names.
The mean days in hospital is:  8.60

The standard deviation of days in hospital is:  5.72

The 25th percentile of days in hospital is:  5.00

The 50th percentile of days in hospital is:  8.00

The 75th percentile of days in hospital is:  11.00

The range of days in hospital is:  27.00

The median of days in hospital is:  8.00

fx >>
```

Duration of stay for AntiBioticReceived	Duration of stay forAntiBioticNotReceived		Duration of stay for AntiBioticReceived	Duration of stay forAntiBioticNotReceived
14	5		14	5
30	10			10
8	6		8	6
8	11		8	11
7	5		7	5
3	11		3	11
11	17		11	17
	3			3
	9			9
	3			3
	5			5
mean	5		mean	5
11.57142857	4		8.5	4
std	7		std	7
8.810167287	9		3.728270376	9
	11			11
	9			9
	4			4
	mean			mean
	7.444444444			7.444444444
	std			std
	3.697729386			3.697729386

Problem 3 Matlab code and excel calculations:

```
%% Project 1, Problem 3
Imported_Cholesteral = uigetfile('*.xlsx');
Imported = readtable(Imported_Cholesteral);
Difftable = Imported(1:24,4);
BaseLinesT = Imported(1:24,1);
DifferencesBtwAftBef = table2array(Difftable);
BaseLines_Cholesterals = table2array(BaseLinesT);

AllPositives = abs(DifferencesBtwAftBef);
Mean_Change_in_Cholesteral_levels = mean(AllPositives);
Stdev_of_change_in_Cholesteral = std(AllPositives);
MedianOfBaseLines = median(BaseLines_Cholesterals);
```

Window

	BaseLineBelowMedian	DiffCorresponding	BaseLineAboveMedian	DiffCossesponding
	145	10	195	49
	159	13	205	27
	166	19	244	36
	169	13	250	48
	158	31	236	21
	151	2	192	8
	168	8	224	16
	168	23	238	32
	167	13	197	28
	161	8	180	19
	178	41	222	35
	137	12	197	19
			std:	12.52512626
std:	10.91669558		mean	28.16666667
mean	16.08333333			

Details

Workspace

Name	Value
AllPositives	*24x1 double*
BaseLines_Cholesterals	*24x1 double*
BaseLinesT	*24x1 table*
DifferencesBtwAftBef	*24x1 double*
Difftable	*24x1 table*
Imported	*30x10 table*
Imported_Cholesteral	'Cholesteral.xlsx'
Mean_Change_in_Chol...	22.1250
MedianOfBaseLines	12.5000
Stdev_of_change_in_Ch...	13.0428