

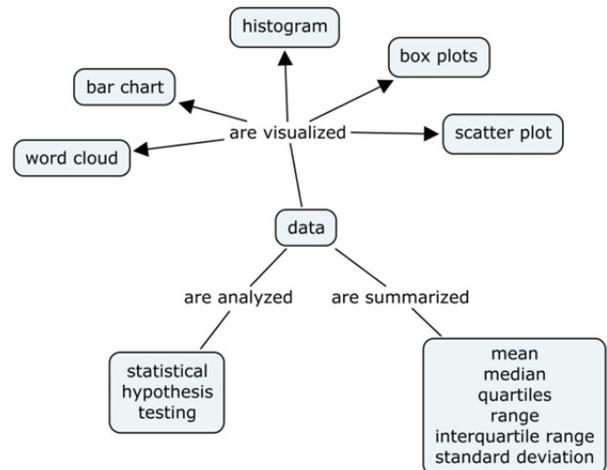
Exploratory Data Analysis

Learning Objectives:

1. After completion of this module, the student will be able to explore data graphically in Excel using

- histogram
- boxplot
- bar chart
- scatter plot

2. After completion of this module, the student will be able to employ web-based tools for statistical hypothesis testing and determine significance levels under multiple testing using the Bonferroni correction.



Knowledge and Skills

- Graphing in Excel
- Functions in Excel
- Pivot tables in Excel
- Concepts: mean, standard deviation, bar chart, histogram, boxplot, quantiles, error bar, line graph, scatter plot

Prerequisites

- Statistical hypothesis testing
- Some basic familiarity with graphing in Excel
- Familiarity with
 - Cardiac events
 - Risk factors for cardiac events
 - Dobutamine stress echocardiography

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 1

Materials and Pedagogical Approaches

We will use a data set that is analyzed in Garfinkel, Alan, et. al. "Prognostic Value of Dobutamine Stress Echocardiography in Predicting Cardiac Events in Patients With Known or Suspected Coronary Artery Disease." *Journal of the American College of Cardiology* 33(3) (1999) 708-16. You can download the data set from the NUMB3R5 COUNT web site (<http://bioquest.org/numberscount/data/>). The data set is called "Garfinkel Cardiac Data."

The data set is a compilation of medical records of patients who underwent dobutamine stress echocardiography and were followed for twelve months afterwards for occurrences of cardiac events. The data set thus includes data on patient characteristics, physiological data under rest and dobutamine conditions, the history of medical conditions relevant to cardiac events, and outcome data during the subsequent twelve months. As preparation for the data set and the accompanying article, students should become familiar with coronary events (myocardial infarction, cardiac death, percutaneous transluminal coronary angioplasty, and coronary artery bypass surgery) and risk factors (high blood pressure, diabetes, and smoking), at the level of being able to read a newspaper article that uses these terms. Students should review patient information on dobutamine stress echocardiography prior to this module (see, for instance, <http://stanfordhospital.org/healthLib/greystone/heartCenter/heartProcedures/dobutamineStressEchocardiogram.html>).

The data set and the accompanying article can be deployed in numerous ways in the classroom, depending on the background of the students. (1) If students are already familiar with statistical methods, they can start with the data set and explore the data on their own or with some guidance. Both a hypothesis-based or a discovery-based approach are appropriate: based on prior knowledge, students can formulate hypotheses and test them subsequently; or students can explore the data and discover patterns. (2) Students can reanalyze the data based on results from the accompanying paper to deepen their grasp of statistical analysis. (3) Students can "unpack" the paper to learn how a scientific paper in this discipline is written and how results can be succinctly communicated. (4) The paper relies almost exclusively on tables to convey results. A student could focus on a specific aspect of the paper and develop a poster where the results are visually displayed.

We will focus in the following on visual displays of the data and their statistical analysis.

Exploratory Data Analysis (EDA)

Graphical techniques are important for exploratory data analysis. The NIST Engineering Statistics Handbook describes many of these techniques: <http://www.itl.nist.gov/div898/handbook/eda/eda.htm>.

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 2

In addition, we will divide patients into groups based on the outcome of categorical data (yes vs. no), and then investigate differences among groups.

The following table contains a list of the variables and their data type in the spreadsheet:

Column	Variable	Data Type
A	Basal Heart Rate	numerical
B	Basal Blood Pressure	numerical
C	Basal Double Product BHR*BBP	numerical
D	Peak Heart Rate	numerical
E	Systolic Blood Pressure	numerical
F	Double Product PKR*SBP	numerical
G	Dobutamine Dose	numerical
H	Maximum Heart Rate	numerical
I	% Maximum Predicted Heart Rate Achieved by Patient	numerical
J	Maximum Blood Pressure	numerical
K	Double Product of Maximum on Dobutamine	numerical
L	Double Product Maximum on this Dobutamine Dose	numerical
M	Age	numerical
N	Gender (male=0)	categorical
O	Baseline Cardiac Ejection Fraction	numerical
P	Baseline Cardiac Ejection Fraction on Dobutamine	numerical
Q	Chest Pain (yes=0)	categorical
R	Sign of Heart Attack on ECG (yes=0)	categorical
S	Equivocal ECG (yes=0)	categorical
T	Heart Wall Motion Anomaly Observed (yes=0)	categorical
U	Stress ECG Positive (yes=0)	categorical
V	New Heart Attack (yes=0)	categorical
W	New Angioplasty (yes=0)	categorical
X	New Bypass Surgery (yes=0)	categorical
Y	Death (yes=0)	categorical
Z	History of Hypertension (yes=0)	categorical
AA	History of Diabetes (yes=0)	categorical
AB	History of Smoking (yes=0)	categorical
AC	History of Heart Attack (yes=0)	categorical
AD	History of Angioplasty (yes=0)	categorical
AE	History of Bypass Surgery (yes=0)	categorical
AF	Any New Cardiac Event (yes=0)	categorical

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 4

Guidelines for Exploratory Data Analysis

After you have become familiar with the different variables in the data set, determine which variables are independent, i.e., input variables, and which ones are dependent, i.e., outcome variables. Next, get a sense for the range of the numerical variables by finding the minimum and maximum value, the mean, median, and standard deviation, and the first and third quartile (i.e., the 25th and 75th percentile). This information can be visualized with box plots. A histogram is helpful to visualize the range and the shape of the distribution. It is generated by first grouping the data and counting the number of occurrences in each group or category. The counts are then displayed as vertical bars in a histogram where the area of each bar is proportional to the frequency in that category. Pairs of numerical data can be displayed as scatter plots to reveal relations among the pairs. Pivot tables are helpful to find counts of categorical data, which can then be displayed as bar charts where the height of each bar is proportional to the frequency in the respective category. To determine statistical significance of any relationships, the chi-squared test and the *t*-test are frequently employed.

Statistical hypothesis testing determines whether there are statistically significant relationships among variables. With a large data set like the one considered in this module, we will perform a large number of statistical tests. To set the significance level in multiple comparisons, we need to be careful. A significance value of 5% means that there is a 5% chance of rejecting the null hypothesis even if it is true. If we do multiple comparisons and stick with the same significance level of say 5%, then just by chance it becomes increasingly likely that we will reject a null hypothesis even if it is true as the number of tests increases. You can think of it in this way: every time you perform a statistical hypothesis test with significance level 5%, there is a 5% chance that you reject the null hypothesis even though it is true, in other words, you are wrong 5% of the time. If you have a coin that comes up tails 5% of the time, then if you toss the coin often enough you will eventually see tails. In fact, we expect to see tails about every twenty coin tosses since 5% is 1 in 20. One solution to the multiple comparison problem is provided by the Bonferroni correction, which says that you need to divide the desired significance level across all tests by the number of tests and apply this new level to each of the individual tests. For instance, if you perform ten tests and want an overall level of significance of 5%, each individual test should be conducted at the 0.5% level since $0.5 = 5/10$. You can read about the Bonferroni correction on http://en.wikipedia.org/wiki/Bonferroni_correction.

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 5

Histograms

Histograms are a common form to summarize univariate numerical data. You can find a description on the NIST pages: <http://www.itl.nist.gov/div898/handbook/eda/section3/histogra.htm>

We will create a histogram for the frequency distribution of the basal heart rate. The data are listed in the spreadsheet under the “Basal Heart Rate” tab. The range of the data is “A2:A559.” We begin with arranging the data into bins of equal width and counting the number of patients in each bin.

To determine the bin size, we first calculate the range of the data, that is, the difference between the largest and the smallest value. The minimum of the data can be found using the Excel function “MIN(*range*)” and the maximum of the data can be found using the Excel function “MAX(*range*).” Confirm that the minimum is 42 and the maximum is 210. The range of the data is the difference between the maximum and the minimum, that is, 168. Let’s start with somewhere close to ten bins. The number of bins may have to be revised to obtain a meaningful histogram. If each bin is of width 15, we end up with 12 bins. To capture all the data, we start the first bin at 41 so that it ends at 55, and the last bin at 206, which ends at 220. The choice of values is somewhat arbitrary: we need to capture all the data and may want to choose “nice” numbers, such as 55, 70,... for the right end points. The right end points are then 55, 70, 85,..., 205, and 220. Enter these numbers into the cells E2:E13.

To count the number of patients whose basal heart rate is greater than 70 and less than or equal to 85, we count the number of patients whose heart rate is less than or equal to 85 and subtract from this the number of patients whose basal heart rate is less than or equal to 70. The “COUNTIF” function in Excel can accomplish this task. For instance, to count the number of patients whose basal heart rate is less than or equal to 70, we would enter into a cell “=COUNTIF(*range*, “<=70””, where “*range*” is the data range. If “70” is in the cell E3, say, then the formula becomes “=COUNTIF(*range*, “<=”&E3).”

To count the number of patients whose basal heart rate is less than or equal to the right boundary of each bin, enter the function “=COUNTIF(\$A\$2:\$A\$559,“<=”&E2)” into the cell F2. Drag this down to F13. These are the cumulative counts. To find the counts in each bin, we take differences, except for the first bin. The count of the first bin is in Cell F2. We thus enter “=F2” into Cell G2. To calculate the remaining counts, we enter F3-F2 into cell G3 and drag this down to G13.

We can now create a table with the counts. The table should list the ranges of each of the bins, the midpoints of each of the bins and the counts for each of the bins. To find the midpoint of a bin, average the left and the right boundary values of each bin. That is, to determine the midpoint of the bin that counts the number of patients whose basal heart rate is between 41 and 55, find

$$\frac{41 + 55}{2} = 48$$

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 6

The midpoint of the first bin is thus 48. Enter the midpoints into the Cells H2-H13. The following table lists the ranges of each bin, their midpoints, and the counts.

Range	Midpoint	Count
41...55	48	36
56...70	63	192
71...85	78	202
86...100	93	95
101...115	108	29
116...130	123	3
131...145	138	0
146...160	153	0
161...175	168	0
176...190	183	0
191...205	198	0
206...220	213	1

Excel has bar charts that display data as bars where the *height* is proportional to the count in the respective category. Histograms display counts as bars where the *area* is proportional to the count in each bin. If all bins are of equal width, both the height and the area of each bar are proportional to the respective counts. Note that in our example, all bins are of equal width. Figure 2 shows the histogram and we will proceed with how to create this figure.

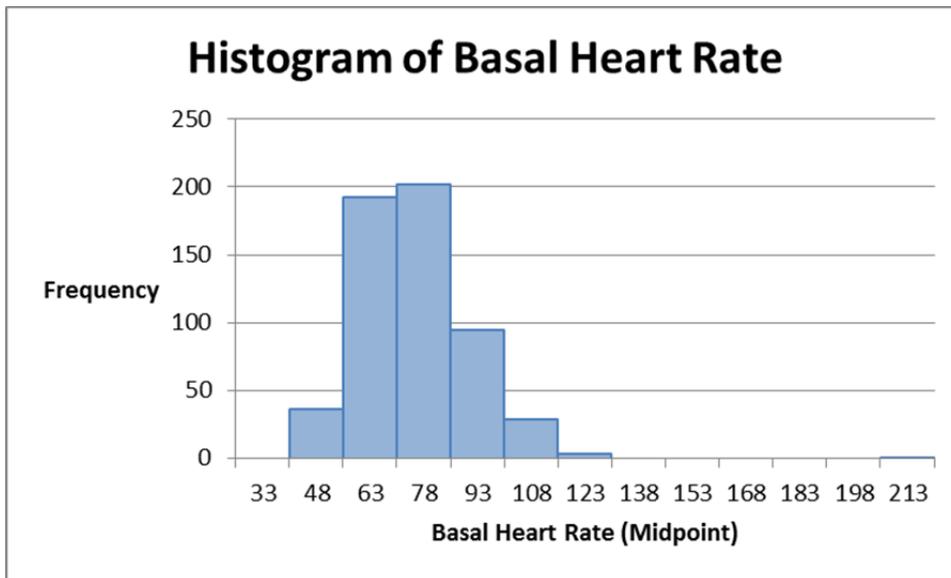


Figure 2: Histogram of the basal heart rate.

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 7

The basal heart rate data of the 558 patients is divided into equal-sized bins and the vertical axis counts the number of patients in each bin. For instance, the bin with midpoint 63 counts the number of patients whose basal heart rate is between 56 and 70 (both end points are included).

To graph the histogram, click on the **Insert** tab. Under the **Charts** group, you find the **Column** charts. We will use the 2D Clustered Column to create the histogram. Highlight the counts in the range G2:G13 and click on the 2D Clustered Column chart. This creates the graph.

To modify the figure, click on the **Design** tab in the **Chart Tools** group. Under the **Data** group, click on **Select Data**. A window appears. Click on **Edit** in the Horizontal (Category) Axis Label. A window for the Axis label range appears. Select the range and enter the range of the midpoints, H2:H13. Click OK.

We can now add a title and axes titles, and increase the width of the bars so that they touch (a standard feature of histograms). We can also remove the “Series 1” label by clicking on it and then hitting the Delete button on the key board. To add a title and axes title, click on the **Layout** tab in the **Chart Tools** group and click on **Chart Title** for the title and on **Axis Title** for the axes titles. To change the width of the bars, double click on any of the bars. A window appears where you can change the gap width: set it to No Gap and the gaps will disappear. To change the border color of the bars, click on **Border Color** and select a color that is different from the fill color. To change the color of the bars, click on the **Fill** option. Close the window. Your histogram should now look like the one in Figure 2.

A histogram is a graphical way to show you the location of the data, how much they are spread out, whether the data are skewed, whether there are outliers, and what the shape of the frequency distribution is. We see that the shape is unimodal, that is, the distribution has only one hump. The distribution is not too skewed and the tails are not too heavy. In other words, the distribution looks near normal. We find that there is one outlier, namely the maximum value of 210. Without any further knowledge, it is difficult to determine whether this value is accurate or a measurement error.

Boxplots

Boxplots are very useful in summarizing univariate numerical data. The NIST Engineering Statistics Handbook introduces boxplots as an “excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.” (<http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm>)

Excel does not have a built-in tool to graph boxplots. However, we can use stacked columns and error bars to build boxplots. We consider the Baseline Cardiac Ejection Fraction as an example. The data are in the spreadsheet under the “BCEF” tab. Click on the tab. We divide the patients into two groups: in one

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 8

group, heart wall motion anomaly was observed (Group 0), in the other group, no heart wall motion anomaly was observed (Group 1). To group the data, use the **Sort & Filter** option in the **Editing** group on the **Home** tab. Highlight the range of values in both columns, click on Sort & Filter, select Custom Sort... and sort according to heart wall motion anomaly. We then create boxplots for each group, following the NIST handbook: the vertical axis lists the response variable, which is the Baseline Cardiac Ejection Fraction, and the horizontal axis lists the factor of interest, which is "Heart Wall Motion Anomaly Observed" (yes=Group 0; no=Group1). Figure 3 shows the final result.

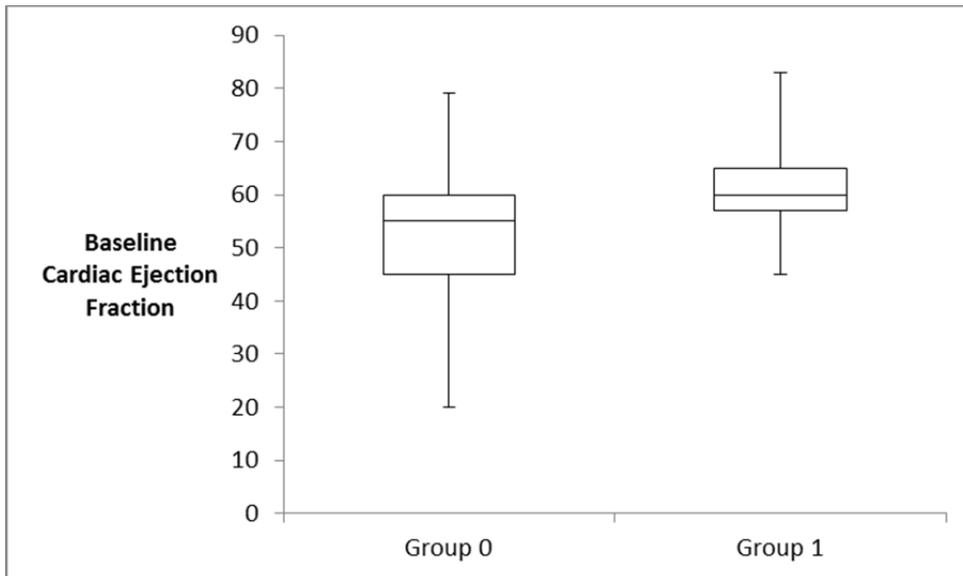


Figure 3: Boxplot for Baseline Cardiac Ejection Fraction for two groups: Group 0 has heart wall motion anomaly observed, whereas Group 1 does not have heart wall motion anomaly observed.

We will describe in the following how to create this figure. A boxplot identifies the median (the middle 50%), the lower quartile (the 25th percentile), the upper quartile (the 75th percentile), and the extreme points. It consists of a box with the bottom of the box indicating the location of the lower quartile, the top of the box indicating the upper quartile, and band in the box indicating the median. Two whiskers on either end of the box indicate the extremes of the data, for instance, the minimum and the maximum of the data. Alternatively¹, the ends of the whiskers may indicate the value that is within 1.5 times the interquartile range (upper minus lower quartile) of the lower and the higher quartile, respectively. Sometimes, the 9th and the 91st or the 2nd and the 98th percentiles are used. Since the whiskers are not plotted uniformly, we need to explicitly state their meaning. In the following, we will use the minimum and the maximum of the data, respectively, to indicate the ends of the whiskers.

¹ http://en.wikipedia.org/wiki/Box_plot

We begin with calculating the lower quartile, the median, and the upper quartile, together with the minimum and the maximum data values in each of the two groups. Excel has functions for these values: To calculate quartiles, use “=PERCENTILE(*range*, *p*)” where *p* denotes the percentile, such as 0.25 for the lower quartile. To calculate the minimum and the maximum, respectively, use “=MIN(*range*)” and “=MAX(*range*)”, respectively. We find

	Group 0	Group 1
Q1	45	57
Median	55	60
Q3	60	65
Min	20	45
Max	79	83

For the box, we will use stacked bar graphs. We therefore need to know the height of the boxes. For the whiskers, we will use error bars. We therefore need to know how far below the bottom of the box and above the top of the box the whiskers extend. We find

	Group 0	Group 1
Q1	45	57
Median-Q1	10	3
Q3-Median	5	5
Q1-Min	25	12
Max-Q3	19	18

We can now draw the boxes for the two groups. Highlight the values for Q1, Median-Q1, and Q3-Median for both groups simultaneously, and select the “Stacked Columns” option in the **Charts** group under **Column**. Click on one of the bars. This opens the **Chart Tools** group with three tabs. Click on the **Design** tab. In the **Data** group, you find the **Switch Row/Column** option. Click it to have the boxes stacked according to their groups. Double click on either of the bottom boxes of the stacked boxes. A window opens that allows you to format the data series. Go to “Fill” and select no color for the boxes and no color for the shape outline. Close the window and double click on one of the middle boxes. Again, select no fill but now choose a black shape outline for the two middle boxes. Repeat this for the top boxes. To add one-sided error bars to the bottom boxes, click on either of the (now invisible) bottom boxes. In the **Analysis** group under the **Layout** tab, you will find the **Error Bars** option. Click on “More Error Bar Options.” Click on the Minus direction for the display and use the Custom option for the error amount. The value you specify is the Q1-Min value. Repeat this for the top whisker, except that now the direction of the whisker is Plus and its length is given by Max-Q3. Add the names of the groups and a vertical axis title. If you double click on the horizontal lines in your figure, that is, the Vertical Axis Major Gridlines,

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 10

you can delete them by clicking on “Line Color” and then “No Line.” Your figure should now look like the one in Figure 3.

To test whether the two groups are statistically significant, perform a *t*-test. There are online calculators for *t*-tests

- <http://studentsttest.com/>
- <http://www.dimensionresearch.com/resources/calculators/ttest.html>

We need to calculate the mean, the standard deviation, and count the number of patients in each group. For the mean, use the Excel function “=AVERAGE(*range*)” and for the standard deviation, use the Excel function “=STDEV(*range*).” To count the number of patients, sum the 0s and 1s. The sum is equal to the number of patients in Group 1. If you subtract the value from the total number of patients, 558, you obtain the number of patients in Group 0. We find

	Group 0	Group 1
Mean	51.21595	60.74319
S.D.	11.58346	5.040093
Sample Size	301	257

Go to either web site for the *t*-test calculator and enter the data. You will find that the two groups are significantly different from each other.

Pivot Tables and Bar Charts

The article by Garfinkel states that “[p]atients with a positive SE had a 34% cardiac event rate within the ensuing 12 months (Table 3) versus an event rate of only 10% in patients with a negative SE ($p < 0.001$).” We will use pivot tables to confirm these figures and bar charts to illustrate the difference.

Creating a pivot table:

1. Highlight all the data that is in the spreadsheet under the “Data” tab (but only the data, not the headers), go to Insert, and click on PivotTable. This opens a pivot table in a new tab.
2. There are four areas where we can drag fields from the Pivot Table Field List. We drag the “Stress ECG Positive (yes=0)” field to the Row Labels and the “Any New Cardiac Event (yes=0)” to the Column Labels. To count the values in each of the four categories in the pivot table, we can drag, for instance, “Gender” into the Σ Values area. Make sure that the Value Field setting is “Count.”

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 11

3. To create a bar graph, click on Options in the PivotTables Tools and click on PivotChart. Select the bar graph.
4. To delete a field from the areas, either unclick the box in the Pivot Table Field List or left-click the label and delete the label.

We find

Count of Gender	Any New Cardiac Event (yes=0)		
Stress ECG Positive (yes=0)	0	1	Grand Total
0	46	90	136
1	43	379	422
Grand Total	89	469	558

and the corresponding bar chart

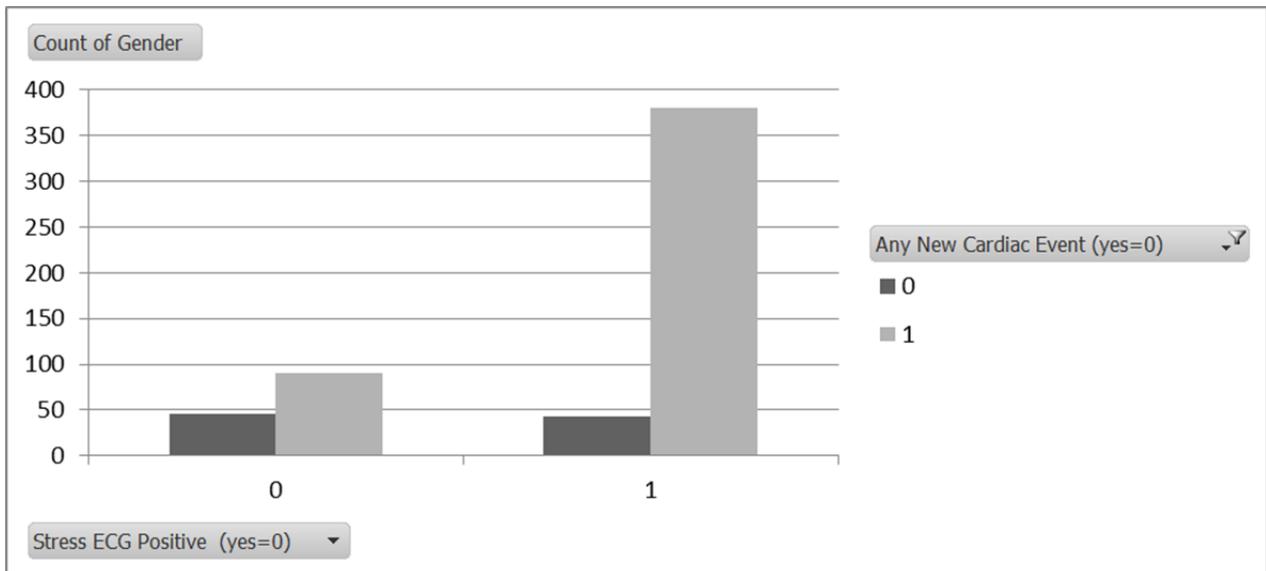


Figure 4: Bar chart of new cardiac events as a function of whether the stress ECG was positive.

To confirm the statement in Garfinkel et al., we calculate the percentage of patients with a positive SE who experienced a cardiac event. 136 patients had a positive SE and 46 of them experienced a cardiac event. Therefore

$$\frac{46}{136} = 0.338 \approx 34\%$$

There were 422 patients who had a negative SE and 43 of them experienced a cardiac event. Therefore

$$\frac{43}{422} = 0.102 \approx 10\%$$

The statistical significance of this difference can be ascertained with a chi-square test, which tests whether the two variables (positive SE and cardiac event) are associated. Online chi-square calculators are available:

- <http://people.ku.edu/~preacher/chisq/chisq.htm>
- <http://faculty.vassar.edu/lowry/tab2x2.html>
- <http://statpages.org/ctab2x2.html>

We use the first link. The online calculator asked you to enter the total number of patients in each of the four categories into cells. Clicking the button to calculate the chi-square value, which is 42.854 (Figure 5), confirms that

$$p = 0 < 0.001$$

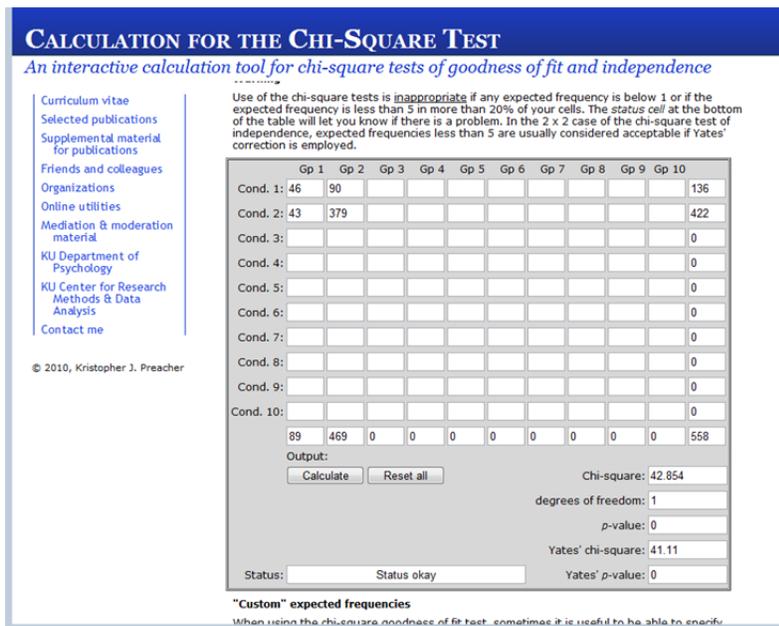


Figure 5: Screenshot of Chi-square calculator (<http://people.ku.edu/~preacher/chisq/chisq.htm>)

The Report Filter can be used to understand the effect of for instance History of Diabetes. The next table and graph looks at the number of patients with Heart Wall Anomaly and Stress ECG Positive filtered according to diabetes. Set the filter so that the data set includes only those patients with diabetes. Below is the table that you should get:

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 13

History of Diabetes (yes=0)	0				
Count of Gender		Any New Cardiac Event (yes=0)			
Heart Wall Motion Anomaly Observed	Stress ECG Positive (yes=0)	0	1	Grand Total	
0	0	22	28	50	
	1	16	59	75	
0 Total		38	87	125	
	1	0	3	6	9
		1	3	69	72
1 Total		6	75	81	
Grand Total		44	162	206	

The bar charts looks then as follows:

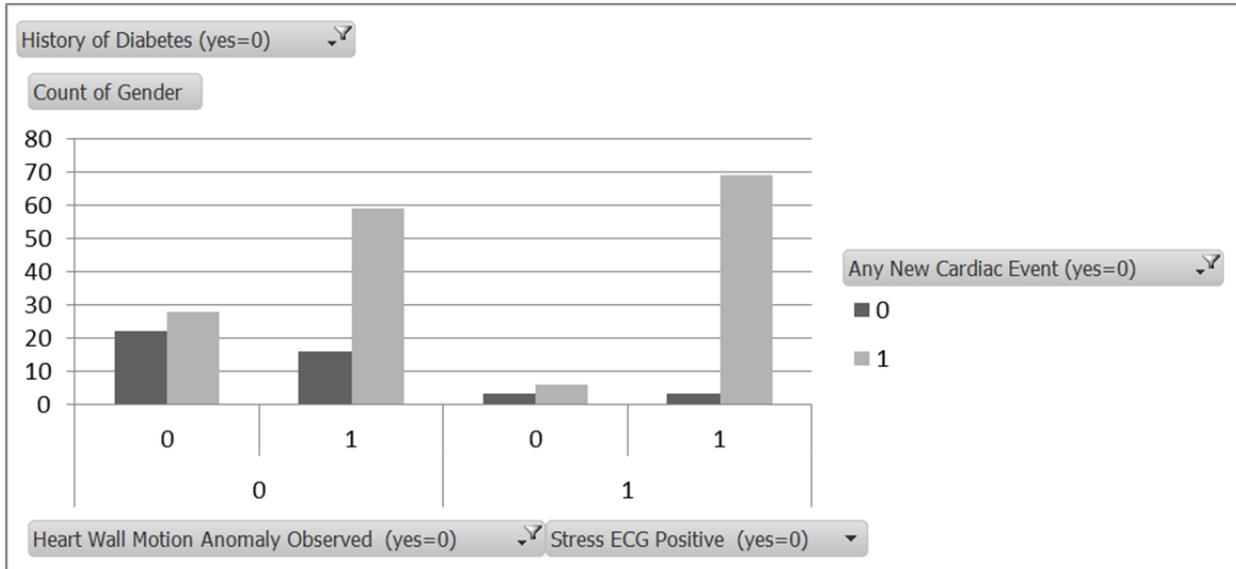


Figure 5: Cardiac event rate for patients with a history of diabetes as a function of WMA and positive stress ECG.

Scatterplots

You Tube has videos on many of the features of Excel. Here is one on scatter plots: http://youtu.be/-SeCPLC30_g

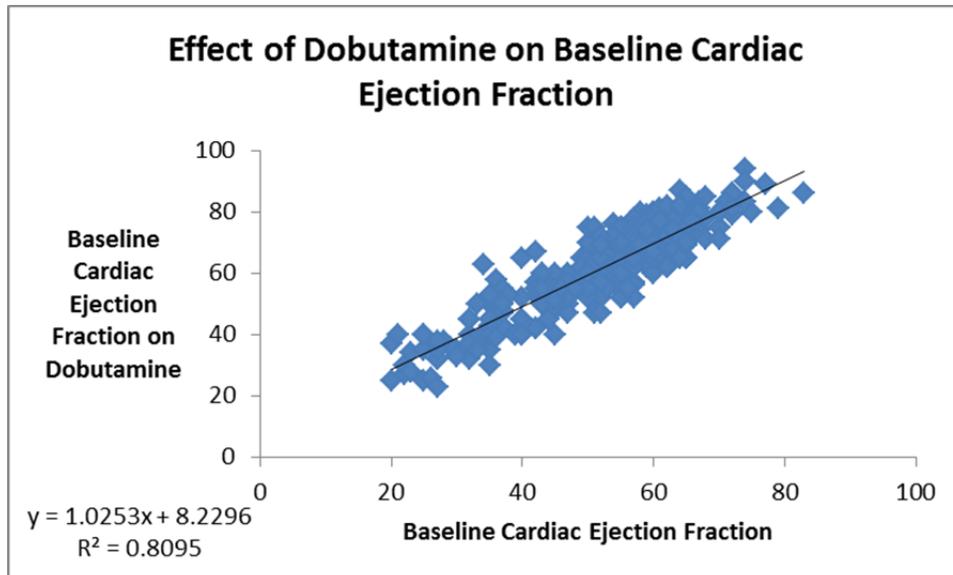


Figure 6: Scatterplot

A scatterplot can convey visually whether two numeric variables are related. We added a regression line and the R^2 value, both are readily available in Excel.

Effective Graphing

1. Take the Graph Design IQ test: <http://www.perceptualedge.com/files/GraphDesignIQ.html> and record your score.
2. Go to <http://www.perceptualedge.com/examples.php>
 - a. A graph and a table: <http://www.perceptualedge.com/example2.php>
 - b. Pie chart versus bar graph : <http://www.perceptualedge.com/example12.php>
 - c. Using line markers: <http://www.perceptualedge.com/example14.php>
 - d. Multiple solutions: <http://www.perceptualedge.com/example10.php>

Citation: Neuhauser, C. Exploratory Data Analysis.

Created: June 3, 2011

Copyright: © 2011 Neuhauser. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial Share Alike License, which permits unrestricted use, distribution, and reproduction in any medium, and allows others to translate, make remixes, and produce new stories based on this work, provided the original author and source are credited and the new work will carry the same license.

Funding: This work was partially supported by a HHMI Professors grant from the Howard Hughes Medical Institute. Page 15

For each of the four examples, explain in your own words (250 words or less) what the problem of the original visualization was and how the visualization was improved.

3. Retake the Graph Design IQ test: <http://www.perceptualedge.com/files/GraphDesignIQ.html> and record your score.

4. Read pages 1-4 in the White Paper *Effectively Communicating Numbers* by S. Few.

Web 2.0 Visualization Tools

<http://www-958.ibm.com/software/data/cognos/manyeyes/>

<http://xtimeline.com/index.aspx>

<http://circos.ca/>

<http://www.analytictech.com/Netdraw/netdraw.htm>

References

Few, S. 2005. *Effectively Communicating Numbers*. Principal Perceptual Edge. White Paper. Downloaded from <http://www.perceptualedge.com/library.php#Whitepapers>

Garfinkel, A., et. al. 1999. Prognostic Value of Dobutamine Stress Echocardiography in Predicting Cardiac Events in Patients With Known or Suspected Coronary Artery Disease. *Journal of the American College of Cardiology* 33(3): 708-16.