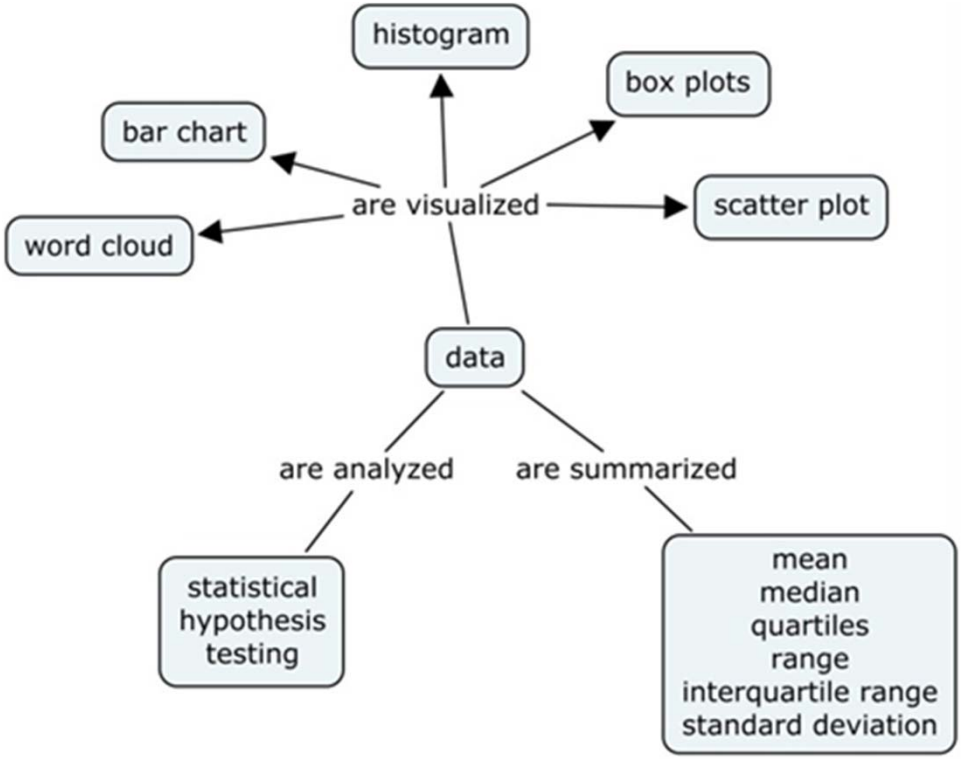
Exploratory Data Analysis

Claudia Neuhauser
University of Minnesota Rochester
June 11, 2011

Concept Map



Created with Cmap: http://cmap.ihmc.us/

Learning Objectives

- After completion of this module, the student will be able to explore data graphically in Excel using
  - histogram
  - boxplot
  - bar chart
  - scatter plot
- 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

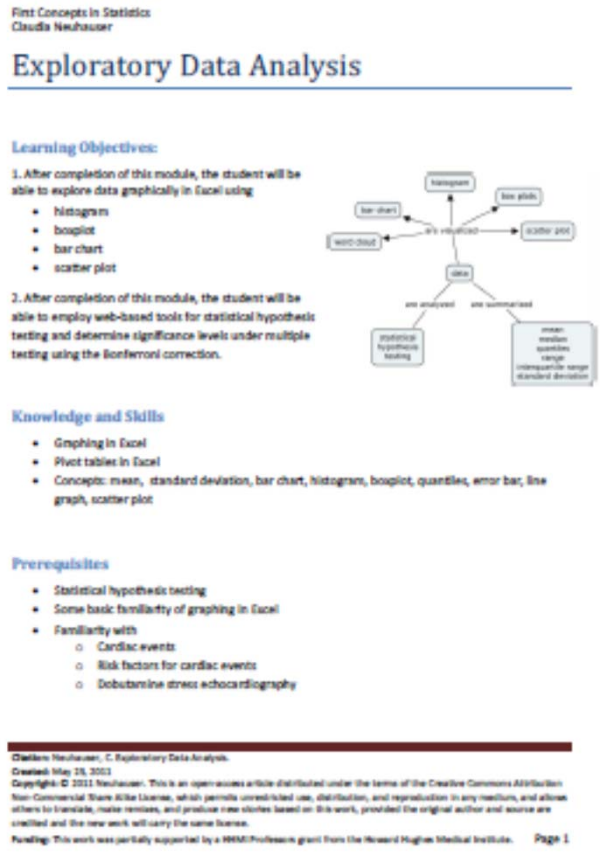
- Graphing in Excel
- Functions in Excel
- Pivot tables in Excel
- Concepts: mean, standard deviation, bar chart, histogram, boxplot, quantiles, error bar, line graph, scatter plot

Prerequisites

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

Worksheet

- **Citation:** Neuhauser, C. Exploratory Data Analysis.
- **Created:** May 23, 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.



First Concepts in Statistics
Claudia Neuhauser

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
- 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 of graphing in Excel
- Familiarity with
  - Cardiac events
  - Risk factors for cardiac events
  - Dobutamine stress echocardiography

Citation: Neuhauser, C. Exploratory Data Analysis.
Created: May 23, 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

The Data

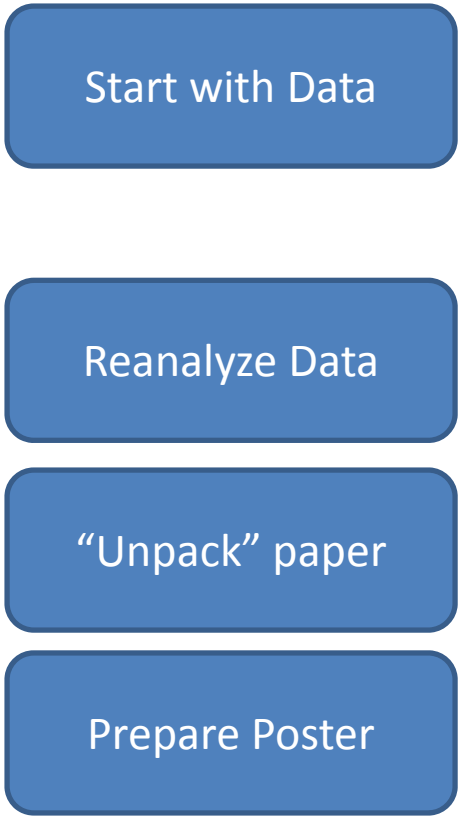
- Garfinkel Cardiac Data: http://bioquest.org/numberscount/data/
- Citation: Garfinkel, Alan, et. al. 1999. "Prognostic Value of Dobutamine Stress Echocardiography in Predicting Cardiac Events in Patients With Known or Suspected Coronary Artery Disease." Journal of the American College of Cardiology 33(3) (1999) 708-16

Background and Preparation

- 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 includes data on
  - patient characteristics
  - physiological data under rest and dobutamine conditions
  - the history of medical conditions relevant to cardiac events
  - outcome data during the subsequent twelve months.
- Preparation: review patient information on dobutamine stress echocardiography
  - http://stanfordhospital.org/healthLib/greystone/heartCenter/heartProcedures/dobutamineStressEchocardiogram.html

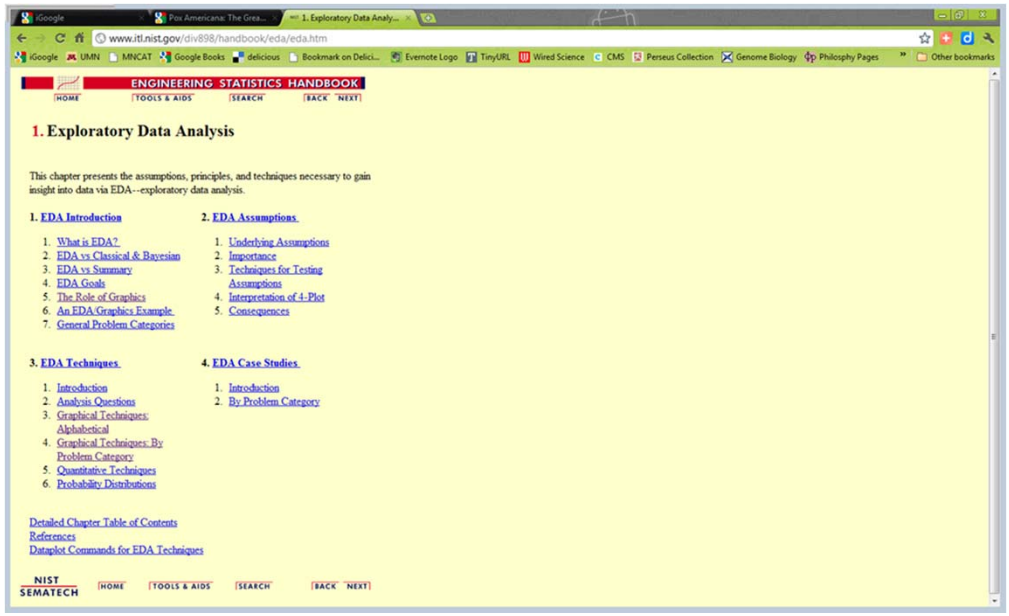
Classroom Use

- 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.
- Students can reanalyze the data based on results from the accompanying paper to deepen their grasp of statistical analysis.
- Students can "unpack" the paper to learn how a scientific paper in this discipline is written and how results can be succinctly communicated.
- 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.



Exploratory Data Analysis (EDA)

- NIST Engineering Statistics Handbook
  - http://www.itl.nist.gov/div898/handbook/eda/eda.htm

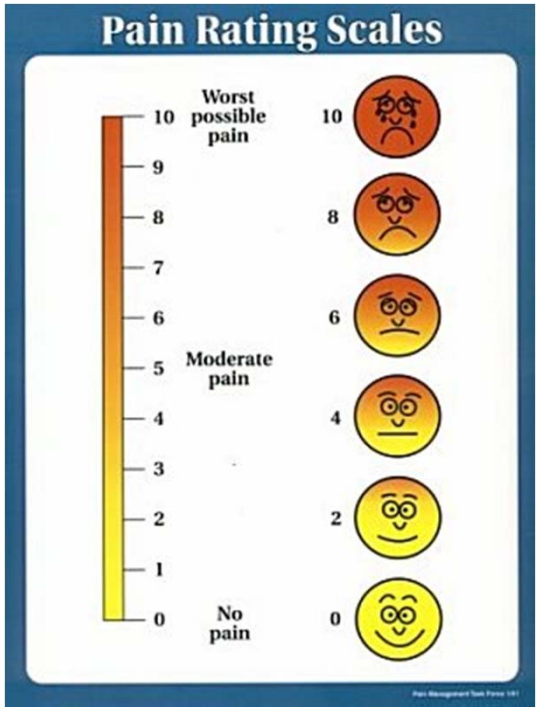


Data I

- Free text
  - Word clouds: http://www.tagxedo.com/

Data II

- Categorical data
  - Nominal (unranked)
    - Yes/No
  - Ordinal (ranked)
- Numerical data
  - Discrete
    - Heart rate
  - Continuous
    - Glucose or cholesterol levels in blood



Pain rating scales are an example of ranked categorical data

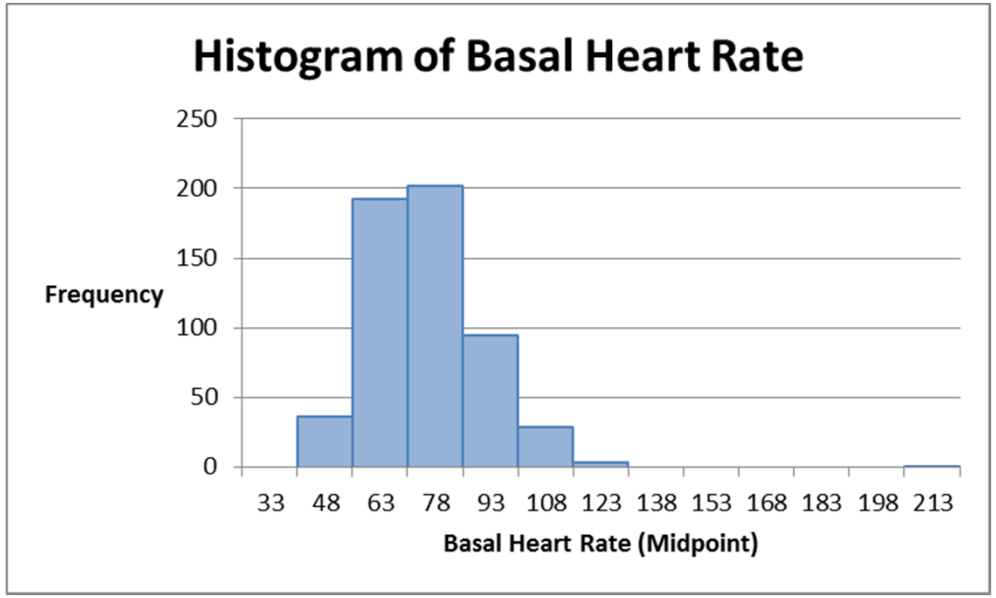
Garfinkel's Cardiac Data

Column	Variable
A	Basal Heart Rate
B	Basal Blood Pressure
C	Basal Double Product BHR*BBP
D	Peak Heart Rate
E	Systolic Blood Pressure
F	Double Product PKR*SBP
G	Dobutamine Dose
H	Maximum Heart Rate
I	% Maximun Predicted Heart Rate Achieved by Patient
J	Maximum Blood Pressure
K	Double Product of Maximum on Dobutamine
L	Double Product Maximum on this Dobutamine Dose
M	Age
N	Gender (male=0)
O	Baseline Cardiac Ejection Fraction
P	Baseline Cardiac Ejection Fraction on Dobutamine

Column	Variable
O	Baseline Cardiac Ejection Fraction
P	Baseline Cardiac Ejection Fraction on Dobutamine
Q	Chest Pain (yes=0)
R	Sign of Heart Attack on ECG (yes=0)
S	Equivocal ECG (yes=0)
T	Heart Wall Motion Anomaly Observed (yes=0)
U	Stress ECG Positive (yes=0)
V	New Heart Attack (yes=0)
W	New Angioplasty (yes=0)
X	New Bypass Surgery (yes=0)
Y	Death (yes=0)
Z	History of Hypertension (yes=0)
AA	History of Diabetes (yes=0)
AB	History of Smoking (yes=0)
AC	History of Heart Attack (yes=0)
AD	History of Angioplasty (yes=0)
AE	History of Bypass Surgery (yes=0)
AF	Any New Cardiac Event (yes=0)

Histograms

- http://www.itl.nist.gov/div898/handbook/eda/section3/histogra.htm



The histogram graphically shows the following:

- center (i.e., the location) of the data;
- spread (i.e., the scale) of the data;
- skewness of the data;
- presence of outliers; and
- presence of multiple modes in the data.

Instructions

- **Basal Heart Rate** tab
- Excel functions
  - MIN(*range*)
  - MAX(*range*)
  - COUNTIF(*range*, “<=“&E4)
- Determine bins and their midpoints
- Correct use of bar charts to create histograms
  - Area is proportional to counts
- Detailed instructions are on pp. 6-8

Boxplots

- http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm

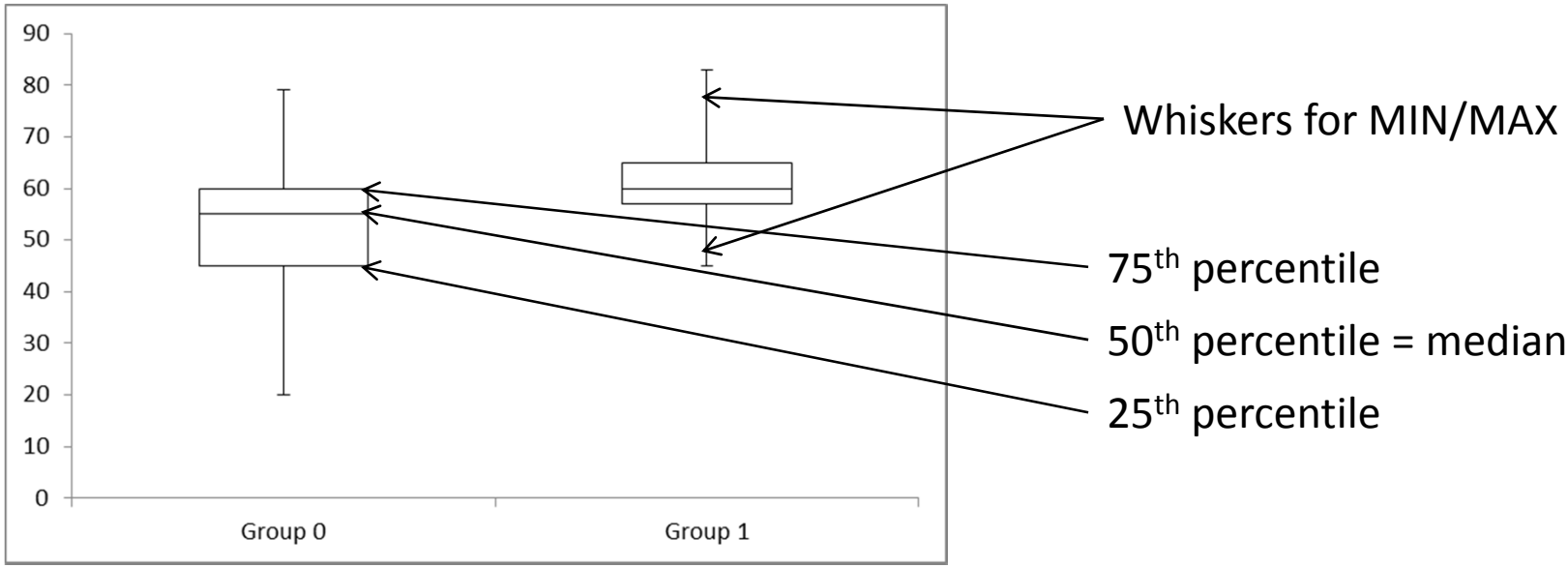


Figure 3: Boxplot for Baseline Cardiac Ejection Fraction for two groups: Group 0 has heart wall motion anomaly observed, whereas Group 1 does not have heart wall motion anomaly observed.

Instructions

- **BCEF** tab
- Excel functions
  - PERCENTILE(*range*, *p*)
  - MIN(*range*)
  - MAX(*range*)
- Determine 25^{th}, 50^{th}, and 75^{th} percentiles and the minimum and maximum of the data
- Use of “Stacked Columns” and “Error Bars” to create box plots
- Detailed instructions are on pp. 8-11

Online Tools for Hypothesis Testing

- To compare the two groups, we perform a *t*-test
  - http://studentsttest.com/
  - http://www.dimensionresearch.com/resources/calculators/ttest.html
- Excel functions
  - AVERAGE(*range*)
  - STDEV(*range*)
- Bonferroni correction (see p. 5)

Pivot Tables and Bar Charts

- “Patients 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 ($p<0.001$).”
  - Use a pivot table to confirm these figures and bar charts to illustrate the difference.
- Detailed instructions on pp. 11-12

Online Tools for Hypothesis Testing

- To test for associations, we use a chi-square test
  - http://people.ku.edu/~preacher/chisq/chisq.htm
  - http://faculty.vassar.edu/lowry/tab2x2.html
  - http://statpages.org/ctab2x2.html

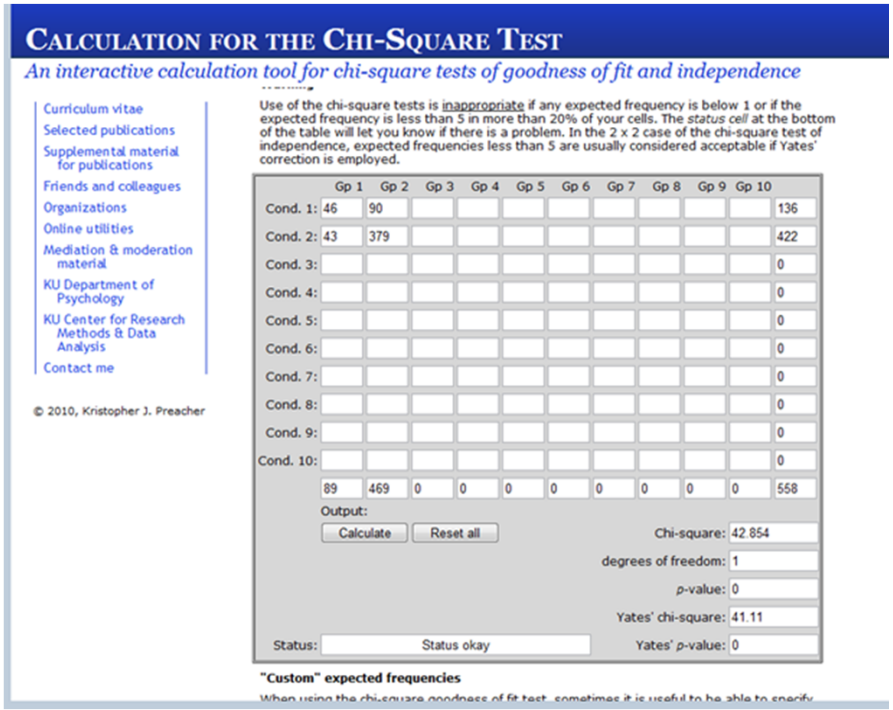


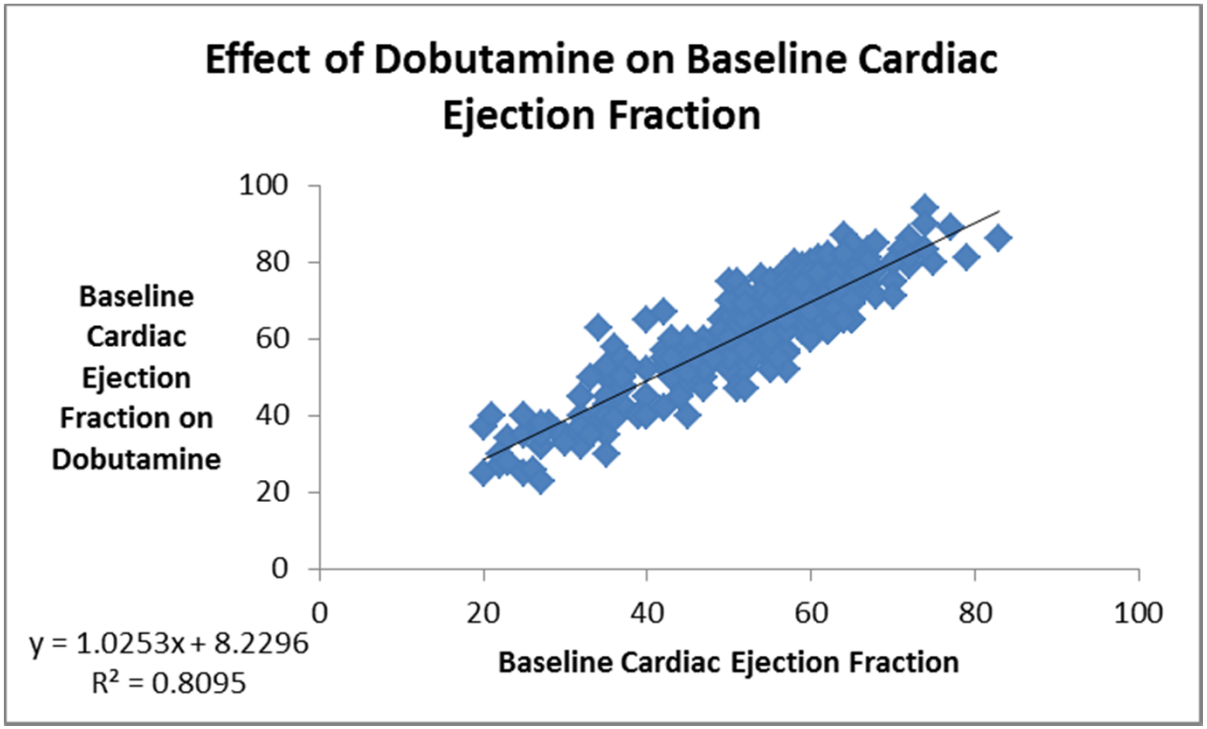
Figure 5: Screenshot of Chi-square calculator (http://people.ku.edu/~preacher/chisq/chisq.htm)

Multiple Testing

- Bonferroni Correction
  - http://en.wikipedia.org/wiki/Bonferroni_correction
- 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.
  - 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**.

Scatterplots

- You Tube videos
  - http://youtu.be/-SeCPLC30_g



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

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