Module 1: Fundamentals of Data Analysis

January 2006

Key Objectives
- Understand the difference between the descriptive and inferential aspects of statistics
- Understand the concept of a random sample, measurement, and levels of measurement
- Understand the use of basic summary statistics of measures of central tendency and measures of dispersion

In this module we will be
- Describing data using summary measures of Central Tendency and Dispersion
- Looking at graphical displays of data: box plots, stem and leaf, and time series smoothing
- Transforming data with logs, inversions, weighting schemes and dealing with outliers

Statistics
- Statistics is the science of data
- It refers to
  - Collecting data
  - Classifying, summarizing, and organizing data
  - Analysis of data
  - Interpretation of data

Statistics
- Statistics is both a field of study
  - ...and a set of tools used by many disciplines
    - Business and Economics
    - Social Sciences
    - Biological Sciences
    - Physical Sciences
We will focus on two types of statistical applications

- Descriptive
- Inferential

**Descriptive Statistics**

- Descriptive statistics uses summary measures, graphs, and measures of association to show relationships in data.
- The focus is on describing the data
- With an emphasis on **parsimony**

**Descriptive Statistics**

- We want to find summary measures which describe the data adequately and succinctly
- Be they a
  - Percentage
  - Mean or average
  - Standard deviation
  - Range from highest to lowest
  - Mode

**Inferential Statistics**

- Inferential statistics takes it a step further
- Now we use some of the same techniques to make estimates, decisions, predictions, or generalizations about a population from a smaller subset or sample

**Inferential Statistics**

- Inferential statistics are a powerful tool for research
- It enables us to make statements about a large group from a much smaller sample.
  - We can survey 1,000 people and make statements about 300 million people
  - And it is the same for 1,000 sample needed to make an inference for 1 million people!
HOW WOULD YOU EVALUATE OVERALL YOUR TRUST IN THE EUROPEAN UNION?

- Source: Prism Research, Public Opinion Poll
- Conducted June 2005.
- n=1550

Let’s look closer at this survey example

- It was based on a poll of 1,550 adult Bosnians taken by Prism Research in June 2005.
- The combined High and Very High is 34.3% or about one-third.
- The sampling error is ± 1.3%
- What does this mean?

Here’s my interpretation

- The survey is designed to represent adult Bosnians in June of 2005.
- Because we are taking a sample, we have some error associated with our estimate.
- Since the sample was taken randomly, we have a method to estimate the error of our estimate.
- In this case, we are reasonably sure (95% sure) that the true percentage is within ± 1.3% points of our estimate.
- Which means our interval for support (Very High and High) is 33.0% to 35.6%

A Population is the total number of units involved in the research question. The units are the members (or elements) of the population.

Populations could be:
- People
- Time periods
- Affinity groups
- Products on a manufacturing line
A POPULATION IS DEFINED BY

- Purpose of the study
- The units and elements involved
- Geographic coverage
- Time frame

Sampling

- When we collect data on all elements in a population, we take a census
- However, sometimes it is difficult to get information on the entire population
- So we take a sample of the population
- A sample is a subset of the units or elements of a population

Why Sample?

- It saves time
- Money
- Other resources (computation time)
- It may actually be impossible to collect information on everyone
  - Every person in the nation
  - Every transaction that takes place

More on Sample

- Samples are also defined in the terms we used for populations
  - purpose of the research,
  - the units and elements involved,
  - the geographic coverage, and
  - the time frame

More on Sampling

- A valuable property of a sample is that is representative of the population.
- The sample characteristics resemble those possessed by the population
- Inferential statistics require a sample to be representative of the population,
- And that can be done through a random process

More Terms

- A random sample is when each element or unit has the same chance of being selected
  - If we select a random sample of 1,000 from a population of 1,000,000,
  - Each unit has a 1,000/1,000,000 or 1/1,000th chance of being selected
More Terms

- **Measurement** is the process of assigning a number to variables of the individual units
- Some measurement seems relatively straight-forward
  - years of age, dollars of income, cholesterol counts, parts per million of a chemical

Measurement

- Other concepts are more difficult to measure
  - Attitudes
  - Emotions
  - Intelligence
  - LOVE

Measurement

- The process of measurement is often complex – don’t take it for granted
- It always comes with some error
- And perhaps Bias

Measurement

- With measurement we must also deal with
  - **Validity** – are we measuring what we think we are measuring
  - **Reliability** – is the measuring device consistent

Types of Data

- **Quantitative data** are measures that are recorded on a naturally occurring scale
- **Qualitative data** does not follow in natural numerical scale and thus are classified into categories

Types of Data

- We will use a more elaborate description of levels of measurement
  - Nominal
  - Ordinal
  - Continuous
Levels of Measurement

- **Nominal** (or categorical) - no implied order or superiority
  - Men and Women
  - Marital Status - Single, married, divorced
  - Geographic Location

- **Ordinal** - an implied order or rank, but the distance between units is not well specified
  - Ranking of a list of items
  - Time in a trend analysis
  - Attitude of Support: Strongly agree to Strongly disagree
  - Categories of Age - < 35, 35 to 55, 55+

Levels of Measurement

- **Continuous** (combination of interval and ratio) - data that is measured on a scale where we can say something about the magnitude between numbers
  - Age
  - Income
  - Sales

Graphing Data

- Histograms
- Stem and Leaf Plots
- Box Plots (a.k.a, Box and Whisker Plots)
- Scatter Plots
- Excel will help with some of these

Histograms

- Good visual depiction of the distribution of a variable, showing shape, modes, skew and outliers
- Can be graphed for a small sample large sample size
- Most software programs (including Excel) provide an easy means to construct
- Requires decisions on the number and width of intervals
- Choices made by a user can distort the graph

Box Plots or Box and Whisker Pots

- Box Plots are based a "Five Number Summary" based on position, including the median and quartiles;
- Good visual depiction of the distribution of a variable, showing shape, modes, skew and outliers
- Can be graphed for a small or large sample size
- There is a uniform approach to construct - no user decisions
- Excellent approach for comparing the distribution of two or more sub-groups
Stem and Leaf Plot

- Good visual depiction of the distribution of a variable, showing shape, modes, skew and outliers
- The plot actually uses the data itself to make the graph
- Limited to small and medium sized data sets - difficult to produce when the sample size is over 150
- The user (or program) must make some decisions that can influence the shape of the graph

<table>
<thead>
<tr>
<th>Stem-and-Leaf Display for MPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem unit: Whole number</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>31</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>33</td>
</tr>
<tr>
<td>34</td>
</tr>
<tr>
<td>35</td>
</tr>
<tr>
<td>36</td>
</tr>
<tr>
<td>37</td>
</tr>
<tr>
<td>38</td>
</tr>
<tr>
<td>39</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>42</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>44</td>
</tr>
</tbody>
</table>

Scatter Plots

- Scatter Plots are based on how two variables vary together, particularly over time.
- Most spreadsheet programs provide an easy mechanism to make a scatter plot (also called XY scatterplot).
- In scatter plots we tend to think of one of the variables as a dependent variable and label it Y.
- It forms the basis for correlation and regression

Excel Practices

- Create a Dummy Variable
- Create a Scatter Plot of two variables
- Open up Excel and open CATALOGS.xls

Creating Dummy Variables

- Any categorical variable with k categories can be represented with k-1 dummy variables
- A dummy variable takes on only two values, either zero or one. A value of one represents the presence of the attribute.
- The category that is not represented by a dummy variable (the k-1 category) is referred to as the Reference Category. It is included in regression analysis.
- When working with Dummy Variables, you must be clear on:
  - Which category is the reference category
  - What a one or zero represents

Dummy Variable Example

- In CATALOGS.xls, AGE is a categorical variable (ordinal) with three categories (30 or less, 31 to 55, and 56 and over.
- We will create two dummy variables, AGE1 and AGE2
- Persons 56 and Over will be the Reference Category
- To do this we will:
  - Insert two columns after AGE
  - Label the columns AGE1 and AGE2 (you can also apply comments)
  - Use IF Statements of the form:
    - For ages 30 and lower: =IF(B7=1,1,0)
    - For ages 31 to 55: =IF(B7=2,1,0)
  - Copy the formulas down the whole column
  - You can use Paste Special, Values to convert the formula into a value which no longer has a reference
Creating a XY Scatter Plot

- In CATALOGS.xls we will create a XY Scatter Plot of Sales (Y) versus Salary (X)
- Use the following sequence:
  - Insert
  - Chart
  - XY (Scatter)
- We will use the first type of scatter plot where the values are not connected by a line
- The Excel Wizard will help us create the plot, so click on Next.

Creating a XY Scatter Plot

- The next window allows you to select the variables
  - I have better luck with Series where I can specify exactly which variable is X and which is Y
  - Y is Sales, A7:A1006
  - X is Salary, G7:G1006 You can name it Salary
  - Click on Next.
- The next window allows you to add titles and labels
- The final window lets you specify where the graph is located

Creating a XY Scatter Plot

- After creating the XY Scatter Plot, we can still dress it up
  - Click on the graph
  - Notice now that Chart now appears in the menu bar
  - You can also right mouse click or click on specific areas of the chart, such as the title
  - I like to:
    - Change the color of the plot area to white
    - Delete the Legend
    - Change the fonts of the labels

Creating a XY Scatter Plot

- The last thing we will do is add a Trend Line to the graph.
  - Click on the chart
  - Click on Graph in the Menu Bar
  - Click on Trendline
  - In the is screen you can:
    - Select the type of trendline – we will use Linear
    - Select Options and add
      - Display equation on chart
      - Display R-squared value on chart
Measure of Central Tendency

- A useful concept when summarizing data is to find some way to measure the center of the data.
- The central tendency of a variable is the tendency of the data to cluster or center about certain numerical values.
- For central tendency we will focus on the mean, the mode, and the median.

The Mean

- The arithmetic mean or mean is the sum of the measurements divided by the number of measurements contained in the data set.
- For a sample we use \( x \) with a bar over it \( \bar{x} \).
- For a population, we use the Greek \( \mu \).

Two ways to express the mean

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{The sum of all the values, divided by the number of values}
\]

\[
\bar{x} = \frac{\sum_{i=1}^{n} (x_i / n)}{n} \quad \text{The sum of values weighted by the number of values}
\]

Properties of the Mean

- As a measure of central tendency the mean has several advantages:
  - The first is that the mean uses information of all the values in a variable.
  - We can make inferences from a sample to the population using the mean (we know its sampling distribution).
  - The mean is the basis for other statistics – product moment statistics.
  - It is sensitive to outliers and extreme values.

- The sum of the deviations about the mean equals zero.
- The sum of squared deviations about the mean is a minimum – Least Squares property:
  \[
  \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{There is no other value or constant we could substitute in the equation for the mean that would result in a lower sum of squares}
  \]
The Median

- The median is the middle value when the measurements are arranged in ascending order.
- It is a positional measure because it is based on the middle case in a variable. The median is also referred to as the 50th percentile.
- Other ordered measures include percentiles, deciles, and quintiles, and quartiles. Quartiles, used in box plots, represent the values at the 25th, 50th, and 75th percentiles.
- In order to find the median value, we first must sort the data, find the middle position (which observation is the middle), and then use that value to represent the median.

Properties of the Median

- The median is an intuitive measure of central tendency - the middle.
- The median has limited inferential properties
- But, it is not as sensitive to outliers and thus is used in data with extreme values
  - Income
  - Company size

The Mode

- The mode is the most frequent occurring value in a variable.
- In a qualitative variable, we refer to the Modal Class or Category.
- The mode may make more sense in reference to categorical data.
- In continuous level data, there may not be any single value that is the most frequent.

Mode

- We may also have the experience of multiple "modes" as in Bi-Modal or Tri-Modal.

Comparing the Mean, Median, and Mode

- In a symmetrical, bell shaped curve depicting the distribution of a variable, the mean, median, and mode would the same value
- The normal curve is a very special bell shaped curve where the mean, median, and mode are all equal.

Comparing the Mean, Median and Mode

- If the data are skewed to the right, the mean is greater than the median and it is being pulled by extreme values to the right
- Skew right
- If the data are skewed to the left, the mean is less than the median and it is being pulled by extreme values to the left
- Skew left
Variability
- Central tendency only tells part of the story of a variable, we also want summary measures of the spread of the data
- We will focus briefly on the
  - Range and Inter-quartile range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

Range and Inter-Quartile Range
- Let's start with the range - the difference between the largest measurement and the smallest measurement
- To calculate the range we need
  - Minimum Value
  - Maximum Value
- The range is an order statistic
- Note that the range depends upon the two most extreme values, and may be seriously influenced by outliers or unusual cases.

An alternative to the range
- The Inter-Quartile Range (IQR)
- Based on the difference between the Third Quartile (Q3 or the 75 Percentile) and the First Quartile (Q1 or the 25 Percentile)
- IQR is less sensitive to the extreme values in a data set than Range

What about using the mean to help measure variability?
- The concept of deviations around the mean can be intuitively appealing.
- If the mean is a good measure of central tendency, then it is reasonable to ask how different (or how far away) is a particular value of X from the mean of X.
- The mean deviation might be a summary measure

The Variance
- A second approach would be to square the differences from the mean
  - The square will always give positive values
  - This is called the variance
  \[ \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \]

NOTE: Population versus Sample
- When we are dealing with a population we use the Greek term \( \sigma^2 \) (sigma squared)
- When we are dealing with a sample we use \( s^2 \)
  - And, we use \( n-1 \) in the denominator
  - This has to do with degrees of freedom
  - It has to do with making inferences from a sample to the population.
  - Using \( n \) in the formula for \( s^2 \) tends to underestimate \( \sigma^2 \).
Sample Variance

\[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \]

A closer look at the Variance

- The numerator is called the Total Sum of Squares

\[ \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

And when we divide by n-1 we have the Mean Squared Deviation (Sample Variance)

\[ \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \]

Standard Deviation

- One problem with the variance is that it is expressed in squared units and can be difficult to interpret
- If you take it the square root of the variance we bring it back to original units
- This is called the standard deviation
  - \( s \) for a sample
  - \( \sigma \) for a population

A Note of Caution

- The variance and the standard deviation are very sensitive to outliers (extreme values)
- When you square large numbers you get much larger numbers
- Be careful when you have extreme values or outliers in the data

Coefficient of Variation

- A standard deviation takes some time to get used to and interpret
- Another way to express the standard deviation is in relation to the mean, called the Coefficient of Variation
  \[ CV = \frac{s}{|\bar{x}|} \times 100 \]
- It expresses the standard deviation as a percentage of the mean
Using Excel for Descriptive Statistics

- Excel has two main ways to help us with descriptive statistics
  - Functions ($fx$ button) which are Statistical (80 functions)
  - Data Analysis Tools
    - Tools Data Analysis Descriptive

- Return to CATALOGS.xls

---

Steps

1. Select the variables
2. Select if Labels in First Row (I suggest yes)
3. Identify Output Range - where the output goes (I suggest a new worksheet which you name)
4. Select Summary Statistics
5. Select Confidence Level for the Mean (95% is ok)

Dress up the output, here’s what I do:
- Shift the labels over one column with cut and paste and then delete the redundant columns of labels for the statistics names
- Bold the first column of descriptive statistics names and make the column width fit the labels
- Adjust the number of decimal places to an appropriate and consistent number

---

Interpreting the Standard Deviation

- We can use the standard deviation to express the proportion of cases that might fall within 1, 2, or 3 standard deviations from the mean.
- We can use Empirical Rule
  - Based on a symmetrical distribution where the mean, median, and the mode are similar

---

A Symmetrical Curve

- The standard deviation is a measure of dispersion or spread in a set of values. It gives us an idea of how much the values in a dataset deviate from the mean.
- The normal distribution, also known as the bell curve, is symmetrical around the mean. Approximately 68% of the values fall within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.

---

Using Excel for Descriptive Statistics

- We will generate Descriptive Statistics for all the variables in the data set
- Refer to the handout
- Select Tools
  - Data Analysis
    - Descriptive Statistics

---

Using Excel for Descriptive Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Mean</th>
<th>Standard Error</th>
<th>Median</th>
<th>Mode</th>
<th>Standard Deviation</th>
<th>Sample Variance</th>
<th>Kurtosis</th>
<th>Skewness</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Count</th>
<th>Confidence Level (95.0%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salary</td>
<td>12167.77</td>
<td>30.39</td>
<td>1918.00</td>
<td>494.00</td>
<td>516.00</td>
<td>502.00</td>
<td>710.00</td>
<td>1000</td>
<td>0.92</td>
<td>1.30</td>
<td>0.30</td>
<td>1000</td>
<td>49.59</td>
</tr>
<tr>
<td>Age</td>
<td>1.92</td>
<td>0.02</td>
<td>2.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1000</td>
<td>0.02</td>
<td>0.14</td>
<td>0.03</td>
<td>1000</td>
<td>0.04</td>
</tr>
<tr>
<td>Gender</td>
<td>0.49</td>
<td>0.02</td>
<td>2.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1000</td>
<td>0.00</td>
<td>0.21</td>
<td>0.03</td>
<td>1000</td>
<td>0.07</td>
</tr>
<tr>
<td>OwnHome</td>
<td>0.52</td>
<td>0.02</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1000</td>
<td>0.00</td>
<td>0.41</td>
<td>0.07</td>
<td>1000</td>
<td>0.07</td>
</tr>
<tr>
<td>Married</td>
<td>0.50</td>
<td>0.02</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1000</td>
<td>0.00</td>
<td>0.72</td>
<td>0.12</td>
<td>1000</td>
<td>0.00</td>
</tr>
<tr>
<td>Children</td>
<td>0.71</td>
<td>0.01</td>
<td>53700.00</td>
<td>1.00</td>
<td>12.00</td>
<td>1.00</td>
<td>12.00</td>
<td>1000</td>
<td>0.00</td>
<td>1.33</td>
<td>0.23</td>
<td>1000</td>
<td>0.00</td>
</tr>
</tbody>
</table>

---

Return to CATALOGS.xls
Empirical Rule

- Approximately 68% of the measurements will be $\pm 1$ standard deviation from the mean.
- Approximately 95% of the cases fall between $\pm 2$ standard deviations from the mean.
- Approximately 99.7% of the cases will fall within $\pm 3$ standard deviations from the mean.

This means it will be very rare to be more than 3 standard deviations from the mean when dealing with a symmetrical distribution.

Transformations of Data

- There will be times when we will need to transform our data during an analysis.
- Reasons include:
  - Adjusting
  - Taming high variability
  - Reducing the impact of extreme values
  - Make nonlinear relationships linear
- Transformations discussed
  - Z-scores
  - Log transformations
  - Weighting schemes

Ideas for Starting an Analysis

- Know your data
- Start with basic descriptives of individual variables
- Shift to looking at bi-variate relationships between variables
- Later move toward more sophisticated analysis
- Always document and record your decisions and observations, and try to verify using several methods

Ideas for Analysis

- Shift to looking at bi-variate relationships between variables
  - Can you think of one variable as the dependent variable - the one you wish to explain?
  - Graph and plot by groups
- Later move toward more sophisticated analysis
  - We will later look at regression
  - But there are more techniques

Ideas for Analysis

- Always document and record your decisions and observations, and try to verify using several methods
  - Be creative!
  - Expect the Unexpected!
- Seek to tell a story with your data, and let the story unfold