Exploring Data with Graphs and Tables

Dr. Tom Ilvento
FREC 408

Describing Data
- Statistics provides us tools to describe a set of data
- The tools involve numerical and graphical summaries
- The first distinction is the nature of the data
  - Qualitative (Nominal and Ordinal)
  - Quantitative (Continuous)

Work in your groups
- Describe the data using any tools you now have
- Use graphs if you wish

Look at Data Set I
- 5.8 6.1 6.5 6.6 6.7
  6.9 7.0 7.1 7.1 7.1
  7.1 7.4 7.6 7.6 7.7
  7.8 7.9 7.9 8.0 8.0
  8.1 8.2 8.2 8.3 8.4
  8.4 8.4 8.4 8.5 8.6
  8.8 9.0 9.0 9.0 9.1
  9.4 9.4 9.8 10.1 10.2
  10.4 10.9 11.0 11.1 11.5
  12.6 14.5 15.5 16.4 88.2

Data Set II
- 2 5 4 3 3 6 2 5
  1 3 6 2 2 3 4 5
  3 3 2 2 4 3 1 5
  2 4 2 3 5 2 4 3
  1 1 3 2 6 2 3 3
  2 2 3 1 2 2 3 1
  2 2

Data Example
- Data Set One was Marriage data for the U.S., 50 states
  - Quantitative data
  - Leads itself to things like **mean**, **mode**, **median** (We'll discuss more about them next class)
Data Example

- Data set 2 was the pattern of responses to a Likert scale of satisfaction with the poultry business from a survey of poultry growers on Delmarva
  - More Qualitative (or ordinal)
  - Leads itself to frequencies, relative frequencies, and graphing

Data Set 2: Poultry Grower Survey

I am satisfied with the Poultry growing business

Qualitative Data – Coding Strategy

- In qualitative data we had both nominal groups (no order or size implied) and ordinal (implied order or ranking)
- We often represent this data with numbers or categories
  - 0, 1 dichotomy can represent the presence of an attribute
  - 1, 2, 3, 4, 5…. Can represent categories or rankings

Qualitative Data

- In qualitative data the categories are also referred to as CLASSES (Def 2.1 p59)
- The number of observations in a class is called the CLASS FREQUENCY (Def 2.2 p59)
- If we calculate a percentage for each class (of the total number of observations) we have the CLASS RELATIVE FREQUENCY (Def 2.3 p59)

Example: Poultry Grower Survey

I would recommend the poultry growing business to someone who is interested

Class
- Strongly Agree
- Agree
- Somewhat Agree
- Somewhat Disagree
- Disagree
- Strongly Disagree

Example: Poultry Grower Survey

I would recommend the poultry growing business to someone who is interested

Class        Code
- Strongly Agree  1
- Agree          2
- Somewhat Agree  3
- Somewhat Disagree  4
- Disagree       5
- Strongly Disagree  6
Example: Poultry Grower Survey

- I would recommend the poultry growing business to someone who is interested

<table>
<thead>
<tr>
<th>Class</th>
<th>Code</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Agree</td>
<td>1</td>
<td>79</td>
</tr>
<tr>
<td>Agree</td>
<td>2</td>
<td>273</td>
</tr>
<tr>
<td>Somewhat Agree</td>
<td>3</td>
<td>308</td>
</tr>
<tr>
<td>Somewhat Disagree</td>
<td>4</td>
<td>163</td>
</tr>
<tr>
<td>Disagree</td>
<td>5</td>
<td>196</td>
</tr>
<tr>
<td>Strongly Disagree</td>
<td>6</td>
<td>327</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>1343</td>
</tr>
</tbody>
</table>

Example: Poultry Grower Survey

- I would recommend the poultry growing business to someone who is interested

<table>
<thead>
<tr>
<th>Class</th>
<th>Code</th>
<th>Freq</th>
<th>Rel Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Agree</td>
<td>1</td>
<td>79</td>
<td>5.9%</td>
</tr>
<tr>
<td>Agree</td>
<td>2</td>
<td>273</td>
<td>20.3%</td>
</tr>
<tr>
<td>Somewhat Agree</td>
<td>3</td>
<td>308</td>
<td>22.9%</td>
</tr>
<tr>
<td>Somewhat Disagree</td>
<td>4</td>
<td>163</td>
<td>12.1%</td>
</tr>
<tr>
<td>Disagree</td>
<td>5</td>
<td>196</td>
<td>14.6%</td>
</tr>
<tr>
<td>Strongly Disagree</td>
<td>6</td>
<td>326</td>
<td>24.2%</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>1344</td>
<td></td>
</tr>
</tbody>
</table>

Proportions and Percentages

- We generally make a distinction between absolute and relative numbers when looking at data, particularly when comparing data.
- Each tells us a different aspect of reality, and rarely is either one sufficient by themselves.
For example, in the 1996 Presidential election over 96,000,000 people voted. However, that was only about 49 percent of the voting age population (194 million).

For comparative purposes we generally want some sort of relative numbers — ratios such as proportions, and percentages.

A ratio is simply one number expressed in relation to another. We do this by dividing the one number by the other.

PROPORTIONS are special kinds of ratios where the denominator is the total population, and the numerator is a class in the population. The interpretation of this measure is what part of the total is the numerator.

Proportions are probabilities.

Examples of Proportions
- If we had 202.6 million eligible voters in 2000, and 105.4 million votes cast
  - The proportion is $105.4/202.6 = .520$
  - The probability of voting, given you are a eligible voter, is $.520$
  - The given statement is a condition, resulting in a conditional probability, which you will see more in Chapter 4.

A Step Further
- We can use conditional probabilities to compare different groups
  - Probability of voting given only a high school education
    \[ p = 32.75/66.3 = .494 \]
  - 66.3 million eligible voters in 2000 had only a high school education and 32.75 million indicated they voted

Conditional Probabilities
- 48.0 million eligible voters in 2000 had 4 or more years of college and 34.56 million voted
  - Probability of voting given a college degree \[ p = 34.56/48.0 = .720 \]
  - If I had just said that 33 million high school graduates voted and 35 million College graduates voted, it would tell a different story
**Percentages**

- Percentages are simply proportions multiplied by 100
- Percent means “per 100”
- It provides a convenient way to express the relative frequency

**A Few Ideas on Graphing Qualitative Data**

- Many software programs can help you graph data
  - Excel or a spreadsheet
  - Word or word processing
- But, there is not substitute for common sense, and good taste!

**Features of a Well Designed Graph**

- The data should stand out clearly from the background
- There should be clear labeling that indicates
  - The title or purpose
  - What each axes, bars, pie segments and so on, denote
  - The scale of each axis including starting point

**Features of a Well Designed Graph**

- A source should be given for the data
- There should be as little “chart junk” - that is, extraneous material - in the picture as possible

*Seeing Through Statistics, Jessica M. Utts, 1999 Brooks/Cole Publishing Company*

**A few ideas on graphing**

- A graph should tell a story – if there is no story, don't graph it!

**A few ideas on graphing**

- Too many pie slices render it useless

---

**Sales**

($1,000,000)

<table>
<thead>
<tr>
<th>1st Qtr</th>
<th>2nd Qtr</th>
<th>3rd Qtr</th>
<th>4th Qtr</th>
</tr>
</thead>
<tbody>
<tr>
<td>$40</td>
<td>$30</td>
<td>$20</td>
<td>$10</td>
</tr>
</tbody>
</table>

**Market Share of Computer Sales in the North East**

- DE
- MD
- NY
- PA
- NC
- ME
- NH
- MA
- CT
- VT
- RI
A few ideas on graphing

- Watch your use of colors!
- Keep in mind the method of printing when using colors.

Distorting data using the axis scale

Checklist for Statistical Pictures

- Does the message of interest stand out clearly?
- Is the purpose or title of the picture evident?
- Is a source given for the data, either with the picture or in an accompanying article?
- Did the information in the picture come from a reliable, believable source?

Checklist for Statistical Pictures

- Is everything clearly labeled, leaving no ambiguity?
- Do the axes start at zero or not?
- Do the axes maintain a constant scale?
- Are there any breaks in the numbers on the axes that may be easy to miss?

Checklist for Statistical Pictures

- For financial data, have the numbers been adjusted for inflation?
- Is there information cluttering the picture or misleading the eye?

Other ideas on graphing

- Appropriate use of graph types – Excel gives you 14 different types
  - Pie chart is good for parts of a whole when the # categories are small
  - Bar Charts allow for comparisons of something by different groups
  - Scatter-plots or XY graphs can show the relationship between two continuous variables

Graphing Quantitative Data

- There are several ways to graph quantitative data
  - Relative Frequency Histograms
  - Stem and Leaf Plots
- Relative frequency histograms require you to divide the data into intervals
  - By doing this you always lose information
  - Choosing the intervals can be problematic

Stem and Leaf Plots

This approach graphs the data using the data itself
- It provides a graphical picture of a variable
  - Distribution
  - Range
  - Skewness
  - Outliers

Look at Data Set I

| 5.8 | 6.1 | 6.5 | 6.6 | 6.7 |
| 6.9 | 7.0 | 7.1 | 7.1 | 7.1 |
| 7.1 | 7.4 | 7.6 | 7.6 | 7.7 |
| 7.8 | 7.9 | 7.9 | 8.0 | 8.0 |
| 8.1 | 8.2 | 8.2 | 8.3 | 8.4 |
| 8.4 | 8.4 | 8.4 | 8.5 | 8.6 |
| 8.8 | 9.0 | 9.0 | 9.0 | 9.1 |
| 9.4 | 9.4 | 9.8 | 10.1 | 10.2 |
| 10.4 | 10.9 | 11.0 | 11.1 | 11.5 |
| 12.6 | 14.5 | 15.5 | 16.4 | 88.2 |

Constructing A Stem-And-Leaf Plot

- Sort the data
- Choose the Stems.
- Add the *Leaves*
Constructing A Stem-And-Leaf Plot

- **Sort the data** from lowest to highest
- It makes it easier to make decisions about the stems and leaves
- Reduces the number of errors – forgetting an observation or two

Steps In Constructing A Stem-And-Leaf

- **ADD THE LEAVES**, or the following digits. In most cases it will be one digit, but it is possible to use more than one digit
- Make sure the distance between digits are
  - Uniform
  - Large enough to show the separate observations

SAS Stem and Leaf of EPA Data of MPG of 100 Cars

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>0799</td>
<td>4</td>
</tr>
<tr>
<td>33</td>
<td>124589</td>
<td>6</td>
</tr>
<tr>
<td>34</td>
<td>024589</td>
<td>6</td>
</tr>
<tr>
<td>35</td>
<td>012345678999</td>
<td>11</td>
</tr>
<tr>
<td>36</td>
<td>012345678999</td>
<td>20</td>
</tr>
<tr>
<td>37</td>
<td>012345678999</td>
<td>21</td>
</tr>
<tr>
<td>38</td>
<td>012345678999</td>
<td>21</td>
</tr>
<tr>
<td>39</td>
<td>003457899999</td>
<td>8</td>
</tr>
<tr>
<td>40</td>
<td>012356789999</td>
<td>7</td>
</tr>
<tr>
<td>41</td>
<td>002222222222</td>
<td>3</td>
</tr>
<tr>
<td>42</td>
<td>002222222222</td>
<td>1</td>
</tr>
<tr>
<td>43</td>
<td>002222222222</td>
<td>1</td>
</tr>
<tr>
<td>44</td>
<td>002222222222</td>
<td>1</td>
</tr>
</tbody>
</table>

Constructing A Stem-And-Leaf Plot

- **CHOOSE THE STEMS.** The **Stems** are the initial digit in the plot, such as
  - 1 in the number 10;
  - 10 in the number 10.6;
  - 2 in the number 215.
- It is helpful to look at the range of the variable to decide the appropriate stems.
- Stems can be 1, 2 or more digits – for 215 it could be 2 or 21

Stem and Leaf Plots

- Some statistical packages will create stem and leaf plots for us (SAS, Minitab, SPSS)
- Each program has unique aspects on how they do it and how it is labeled
- Stem and Leaf plots work best with moderate sized data sets (100 or less). Larger data sets may get unwieldy.

Scatter Plots

- When you want to look at the relationship between two continuous variables, use a scatter plot
- The book on page 90 provides some guidance
- You will use this in a homework assignment
Scatterplot Example

Average State SAT scores by Percent Taking the Test, 2001

\[ y = -2.133x + 1145.3 \]

\[ R^2 = 0.7657 \]