Organizing Data

Descriptive Statistics deals with collection, classification and analysis of DATA.

Types Of Variables

Definition: A variable is a characteristic that varies from one person or thing to another.

1. Qualitative Variable: Non-numeric valued variable
2. Quantitative Variable: Numeric valued variable
   a. Discrete Variable: A quantitative variable whose values form a countably finite set
   b. Continuous Variable: A quantitative variable whose possible values form some interval of numbers

Types Of Data -- Classification I

Definition: Information obtained by observing values of a variable.

1. Qualitative Data: Data obtained on a qualitative variable

   Examples:
   Eye color (Black, Blue, Brown, Green)
   Gender (Male, Female)
   Blood Type (A, B, AB, O)

2. Quantitative Data: Data obtained on a quantitative variable

   Examples:
   Height of all the students in a class
   Number of ice-creams purchased at an ice-cream store

Types of Quantitative Data

1. Discrete: Data obtained on discrete variable (countable set)
2. Continuous: Data obtained on a continuous variable
Types Of Data -- Classification II

1. Nominal Data
   Categories only. Data cannot be arranged in an ordering scheme.

   Example of Categories:
   East, West, North, South
   Voter categories

2. Ordinal, Ranked, or Ordered Data
   Data about order or rank on a scale. That is, categories are ordered
   but differences cannot be determined.

   Examples of Experiments:
   Taste testing
   Selection of candidates

   Examples of Scales:
   low, medium, high
   1, 2, 3,...
   A, B, C,...
   S. Agree, Agree, Indifferent, Disagree, S. Disagree

3. Interval Data
   Differences between values can be determined but no inherent starting point. Data generally obtained from the measurement of quantities such as, temperature.

4. Ratio Data
   Data contain an inherent starting point.

   Example: Height, Weight, etc.
Data Collection

Most common approach is to sample the population

Characteristics of a good sample

1. Large enough
2. Experiment values v/s questionnaire values
3. Cost of the survey
4. Represent the population

Types of Sampling (Review)

1. Random sampling -- equal chance approach
2. Stratified sampling -- subdividing the population into meaningful strata
3. Systematic sampling -- pick every kth item
4. Cluster sampling -- divide the population into clusters
5. Convenience sampling
Grouping Data

Frequency Data

Data on the number of individuals or items falling in various categories.

Examples:

The #people in each of the four blood type categories.
The #people that fall into each category for candidate selection.
The number of people over 21 that fall into one of the three height categories (<5', 5-6', 6'>).

Frequency Table

One of several methods available for organizing data. Terms needed are:

1. **Classes**: Categories for grouping data.
2. **Tally Marks**: Counting the number of items in each class.
3. **Frequency**: The number of data values (observations) in each class.
4. **Frequency Distribution**: A table listing all classes and their frequencies.
5. **Percentage**: \((\text{Frequency} / \text{total items}) \times 100\).
6. **Relative Frequency**: Percentage expressed as a decimal.
7. **Relative Frequency Distribution**: A table listing all classes and their relative frequencies.
8. **Lower Class Limit or Lower Cutpoint**: The smallest value that can go into a class.
9. **Upper Class Limit or Upper Cutpoint**: The largest value that can go into a class.
10. **Class Mark**: The Midpoint of a class.
11. **Class Width**: The difference between the lower class limit of the given class and the lower class limit of the next higher class.
12. **Grouped Data Table**: A table giving the classes, frequencies, relative-frequencies, and the class marks for a data set.
Constructing A Frequency Table or a Grouped Data Table

Steps:

1. Find the lowest and highest value in the data set.

2. Decide on the classes.
   
   * Try to have anywhere from 5 to 10 classes.
   * Formula you could use:
     
     \[
     \text{# of classes} = \text{Trunc}(1 + 3.3\log_{10}(\text{# of items})) + 1
     \]
   * Remember each data item must belong only to one class.
   * Whenever feasible all classes should have the same width.

3. Construct a Tally mark table to count the number of items in each class.

4. Count the tally marks to obtain the frequency in each class so as to build the frequency distribution table.

5. Calculate the Relative Frequency (Rf) of a class. That is,

   \[
   Rf = \frac{\text{Frequency}}{\text{Total Number of Items}}
   \]

   Recall: Percentage = Rf * 100.

   This would result in a relative frequency distribution table.

6. Calculate the class mark so as to compute the frequency table.

Note: Cumulative frequency can be calculated by keeping a running total of the frequency.
Example:

Days To Maturity For 40 Short-Term Investments

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>64</td>
<td>99</td>
<td>55</td>
<td>64</td>
<td>89</td>
<td>87</td>
<td>65</td>
</tr>
<tr>
<td>62</td>
<td>38</td>
<td>67</td>
<td>70</td>
<td>60</td>
<td>69</td>
<td>78</td>
<td>39</td>
</tr>
<tr>
<td>75</td>
<td>56</td>
<td>71</td>
<td>51</td>
<td>99</td>
<td>68</td>
<td>95</td>
<td>86</td>
</tr>
<tr>
<td>57</td>
<td>53</td>
<td>47</td>
<td>50</td>
<td>55</td>
<td>81</td>
<td>80</td>
<td>98</td>
</tr>
<tr>
<td>51</td>
<td>36</td>
<td>63</td>
<td>66</td>
<td>85</td>
<td>79</td>
<td>83</td>
<td>70</td>
</tr>
</tbody>
</table>

Grouped Data Table

<table>
<thead>
<tr>
<th>Days To Maturity</th>
<th>Tally</th>
<th>Freq.</th>
<th>Rel. Freq</th>
<th>Class Mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-49</td>
<td></td>
<td></td>
<td>1</td>
<td>1/40 --&gt;</td>
</tr>
<tr>
<td>50-59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70-79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80-89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90-99</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>40</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Checks:

- Relative Frequency total must equal 1.0.
- Frequency total must equal the sample size.
Special Cases Of Grouping

**Single-Value Grouping**

1. No class range.
2. The class mark is the class value itself.
3. Used with finite discrete data.

**Example:**

*Number of TVs Owned per Household*

<table>
<thead>
<tr>
<th>Number of TVs</th>
<th>Frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0.020</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>0.320</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>0.280</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>0.240</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.060</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.040</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td><strong>50</strong></td>
<td><strong>1.000</strong></td>
</tr>
</tbody>
</table>
Grouping For Continuous Data – CutPoint Grouping

1. Used with continuous or decimal data.
2. Either the upper- or lower-limit of the class must specify the exclusion value.
   
   Example: 120 - under 130 -or- over 120 - 130.

3. Class mark does not change. Example, for 120-under 130 the CM is:

   \[
   \frac{120+130}{2} = 125
   \]

Grouping For Qualitative Data

Similar to Single-Value Grouping.

Why Grouping?

1. Summarize
2. Enable to determine the shape of the data (probability model).
3. Important statistics
Graphs & Charts
http://onlinestatbook.com/2/

Histograms

1. Frequency histograms
   * Classes on the horizontal axis
   * Frequency on the vertical axis
   * Height of each bar is equal to the class frequency
   * Must have titles: overall & xy-axis

2. Relative-Frequency histograms
   * Classes on the horizontal axis
   * Relative-Frequency on the vertical axis
   * Height of each bar is equal to the class relative-frequency
   * Must have titles: overall & xy-axis

Features Of Histograms

* Effective summary of the sample (population)
* Shows the hypothetical probability distribution
* Useful only with numerical (metric) data

Short-Term Investments

![Histograms](image-url)
Dot Plots

* All possible data values on the x-axis

* For every occurrence of a specific data value put a DOT at the appropriate x-axis coordinate. If a DOT already exists, simply stack the new DOT on top of it.

Features:

* Shows the relative positions of the data in a data set.
* Allows easy comparisons of two or more data sets.
* Difficult to use with too much data with a lot of dispersion.

Example:

Oat Yields

<table>
<thead>
<tr>
<th>67</th>
<th>65</th>
<th>55</th>
<th>57</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>61</td>
<td>61</td>
<td>64</td>
<td>62</td>
</tr>
<tr>
<td>62</td>
<td>60</td>
<td>62</td>
<td>60</td>
<td>67</td>
</tr>
</tbody>
</table>

Oat Yields

Yield (bushels)
Pie Charts

- Useful for qualitative data
- Size of the pie-slice indicates relative-frequency.

<table>
<thead>
<tr>
<th>Party</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>13</td>
<td>0.325</td>
</tr>
<tr>
<td>Republican</td>
<td>18</td>
<td>0.450</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>0.225</td>
</tr>
<tr>
<td>TOTAL</td>
<td>40</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Political Party Affiliations

- 32.50% Democrat
- 45.00% Republican
- 22.50% Other

<< See example in Excel >>
Bar Charts

* Like a histogram, but bars are separated.
* Can be used in place of pie charts for non-qualitative data.
* Pie charts are preferred for qualitative data.

![Political Party Affiliations Chart]

Legend:
- Democratic
- Republican
- Other
Stem-&-Leaf Diagrams

Invented by John Tukey, 1960. Simultaneously, groups the data items and draws a horizontal histogram.

Steps:

1. Select the leading digits from the data. These will be denoted as STEMS.
2. List those leading digits on the left-hand side of a page (vertical)
3. Write the final digit (LEAVES) of each data item to the right of the appropriate leading digit.

Stem & Leaf Diagram for Days-To-Maturity Data.

<table>
<thead>
<tr>
<th>stems</th>
<th>---------------------------</th>
<th>Leaves</th>
<th>---------------------------</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>8 6 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>7 1 6 3 5 1 0 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2 4 7 3 6 4 0 9 8 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0 5 1 0 9 8 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5 9 1 7 0 3 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9 9 5 8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other Variations

1. Shaded stem-and-leaf-diagram --- leaves are shaded
2. Ordered stem-and-leaf diagram --- leaves are in ascending order.
Why Charts and Graphs?

To determine the distribution of a data set. A distribution of a data set is, a table, graph, or formula that provides the values of the observations and how often they occur.

Common Distribution Shapes:

Symetric

1. Bell-shaped
2. Triangular
3. Uniform

Skewed

4. Reverse J-shaped
5. J-shaped
6. Right-skewed
7. Left-skewed

Maybe Symetric or Skewed

8. Bimodal
9. Multimodal
DISTRIBUTION SHAPES

(a) Bell-shaped  (b) Triangular  (c) Uniform (or rectangular)

(d) Reverse J-shaped  (e) J-shaped  (f) Right skewed

(g) Left skewed  (h) Bimodal  (i) Multimodal
Population & Sample Distributions

The distribution of a population data is called the population distribution.

The distribution of sample data is called a sample distribution.

Example: A Look at 6 Sample Distributions
Misleading Graphs

1. Truncated graphs
2. Improper scaling

<< See Sec 2.5 in Textbook >>

NOTE: Click below for the Excel File:

http://www.nkd-group.com/sta308/notes/orgdata.xls