Chapter 1. Data Presentation
Statistics and Data

J.C. Wang

Goal and Objectives

1. To understand the difference between types of data — categorical and numerical
2. To learn how to analyze each type
   ▶ To explore the categorical levels of data
   ▶ To explore the numerical levels of data
Outline

Statistics and Data
Statistics and Data
Types of Data
Levels of Measurement
Example of Data Types and Levels of Measurement

Summarizing Categorical Data
Relative Frequency Table
Bar Chart
Pie Chart

Summarizing Numerical Data
Stem-and-Leaf Plot
Relative Frequency Table and Histogram
Five-Number Summary and Boxplots
Dotplot
Data Shape
Statistics and Data

- Statistics
  - (broad sense) collection of methods/procedures for collecting data and analyzing data
  - (narrower sense) numbers arise out of data analysis, e.g., average

- Data
  - collection of measurements made on a number of subjects
  - example: CLASS DATA set (page 2 of textbook) has six different measurements on ten students

Two Types of Data

- Categorical such as color of student eyes
- Numerical such as temperature of this room in Fahrenheit
Levels of Measurement

Categorical Data

1. Nominal: ordering does not exist, eg., gender, SSN, eye color
2. Ordinal: ordering does exist, eg., military rank, class levels, rating scales

Levels of Measurement

Numerical Data

- Interval:
  - distance exists but no ratios
  - zero is arbitrary and not an indication of absence of the measurement; eg.'s, temperature scale, IQ scores, GPA
- Ratio: ratios exists and zero indicates an absence of the measurement
Interval Measurement
Example: Temperature

▶ Subtraction of two interval values = distance which makes sense
  
  \[ 60^\circ F - 30^\circ F = 30^\circ F \]

▶ However, ratio does not make sense  
  
  \[ 60^\circ F \text{ is not twice as warm as } 30^\circ F \]

▶ \(0^\circ F\): not absence of temperature

Ratio Measurement
Two types

▶ Discrete: result of a counting process, whole numbers (i.e., integers), eg., result of question “how many children in your family?”

▶ Continuous: result of a measuring process, eg.’s, age, money, time, mph, height, weight
Example

CLASS DATA

- Variables Sex and Transport are nominal
- Variable (Class) Level is ordinal
- Variable GPA is an interval measurement
- Variables Hours taken and Sleep hours last night are ratio measurements.
- Note: The column Student is used as data label. It’s not used in data analysis. However, it could be used to label observations in graphical presentation or tabular outcomes. Other data labels include students’ names, SSNs, PINs, etc.

Types of Data: a summary

- Types of Data: **Categorical** and **Numerical**
- Levels of Categorical Data: **Ordinal** and **Nominal**
- Levels of Numerical Data: **Interval** and **Ratio**
Tools to Summarize Categorical Data

- Relative Frequency Table
- Bar Chart
- Pie Chart

Relative Frequency Table
for categorical data

- Frequency table lists categories of a variable and their respective frequencies
- Frequency is a count of subjects who fall into a particular category
- Relative frequency is the frequency divided by total count. Independent of size of data set. Used to compare datasets of different sizes.
Example
Graduation Rate Data

<table>
<thead>
<tr>
<th>Major</th>
<th>Frequency</th>
<th>Relative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>16</td>
<td>28.6%</td>
</tr>
<tr>
<td>MGT</td>
<td>8</td>
<td>14.3%</td>
</tr>
<tr>
<td>MKT</td>
<td>4</td>
<td>7.1%</td>
</tr>
<tr>
<td>GBS</td>
<td>28</td>
<td>50.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>56</strong></td>
<td><strong>100.0%</strong></td>
</tr>
</tbody>
</table>

Data on page 3 of textbook

Relative Frequency Table
for nominal data

Arrange the rows of relative frequency table in decreasing order of frequencies of the categories. Redo previous example:

<table>
<thead>
<tr>
<th>Major</th>
<th>Frequency</th>
<th>Relative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBS</td>
<td>28</td>
<td>50.0%</td>
</tr>
<tr>
<td>ACT</td>
<td>16</td>
<td>28.6%</td>
</tr>
<tr>
<td>MGT</td>
<td>8</td>
<td>14.3%</td>
</tr>
<tr>
<td>MKT</td>
<td>4</td>
<td>7.1%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>56</strong></td>
<td><strong>100.0%</strong></td>
</tr>
</tbody>
</table>

Note: it is recommended not to apply this to ordinal data — in such case, the rows should be kept in the order of categories.
Bar Chart

is used to present frequencies or relative frequencies graphically — draw bars (at the same base) with heights representing frequencies or relative frequencies.

Note: use only relative frequency for comparison of multiple data sets having the same variable of same categories.

Pareto Chart

for nominal data

A Pareto chart is a bar chart used for nominal data (only, recommended) in which bars are arranged in decreasing order (left to right) of heights. Redo previous example:
Pie Chart

- A pie chart is a widely used graphical display to express information from frequency summary table of categorical data.
- Picture illustrating the distribution of data
- Circle divided into slices — number of slices corresponds to the number of categories
- Size of slice is proportional
- Relative frequency percents make it easier to create a proportional pie chart

Example
Summer II Quiz Data

![Pie Chart Diagram]

- ACT: 28.6%
- MGT: 14.3%
- MKT: 7.1%
- GBS: 50%
Drawback of Pie Chart

1. Human eyes not good at comparing areas (of slices)
2. Poor choice of colors (or filled patterns) makes judging area even more difficult
3. Human eyes are good at judging positions (of bar tops) hence a bar chart works better than a pie chart
4. Pie chart with proportions of categories shown is no good, bar chart with reference grid lines works better or even a relative frequency table is enough to convey the comparison

iClicker Question 1.2
Graphs and Tabular Tools

- Stem-and-Leaf Plot
- Relative Frequency Table
- Histogram
- Box-and-Whisker Plot
- Dotplot (see page 13)

Interpret data shape, symmetry and skewness, based on graphical tools.
Use of Stem-and-Leaf Plot

to identify the following

▶ typical value
▶ spread about the typical value
▶ data gaps
▶ shape (distribution) of data
▶ number & location of peaks
▶ outlying values

Construction of Stem-and-Leaf Plot

1. sort data (recommended)
2. select leading digits for stems
3. list all possible stems
4. record leaf for each observation beside the corresponding stem, round off if necessary
5. indicate units for stems & leaves
6. note: if isolated extreme values (low or high) exist, write them out instead
Example of Stem-and-Leaf Plot
Summer II Quiz Data (sorted)

8 11 13 19 21 23 25
25 25 28 31 35 39 47

Stemplot of Summer Quiz Data
stem width = 10
0 | 8
1 | 1 3 9
2 | 1 3 5 5 5 8
3 | 1 5 9
4 | 7

Stemplot of Quiz Data
different set of stems

Each stem in the previous example split in two

stem width = 5
0 | 8
1 | 1 3
1 | 9
2 | 1 3
2 | 5 5 5 8
3 | 1
3 | 5 9
4 | 7
Increased sample sizes required data to be condensed

- **Advantage:** summary table of data arranged into numerically ordered class groupings
- **Disadvantage:** information lost — individual values lost through grouping

**Calculations of Rel.Freq.**

\[
\text{Class RF} = \frac{\text{Class freq.}}{\text{sample size}}
\]

- Expressed as a proportion (fraction) or percent
- Independent of sample size
- Used when comparing multiple data sets of different size (a right way to do)
Construction of Rel. Freq. Table

1. Determine number of classes, \( k \) (5 to 15 recommended)
2. Determine class interval width = range \( \div (k - 1) \)
3. Establish classes’ boundaries and use proper inclusion rule (left- or right-) so that classes not overlap
4. Must include entire range of data

5. Choose classes to facilitate interpretation — shape, etc.
6. Tally data into appropriate classes
7. Total frequency of each class
8. Calculate rel. freq.’s
9. Calculate cumulative (rel.) freq.’s if needed
Example of Rel.Freq. Table
Summer II Quiz Data (sorted)

8 11 13 19 21 23 25 25 25 28 31 35 39 47

range = 47 − 8 = 39 and set k = 4
then width = 39 ÷ (4 − 1) = 13
right-inclusion ⇒ (−5)–8, 8–21, 21–34, 34–47

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>−5–8</td>
<td>1</td>
<td>0.071</td>
<td>0.071</td>
</tr>
<tr>
<td>8–21</td>
<td>4</td>
<td>0.286</td>
<td>0.357</td>
</tr>
<tr>
<td>21–34</td>
<td>6</td>
<td>0.429</td>
<td>0.786</td>
</tr>
<tr>
<td>34–47</td>
<td>3</td>
<td>0.214</td>
<td>1.000</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

class interval includes right boundary

Construction of Histogram

- Vertical bar chart constructed at boundaries of each class
- Bars of same width are joined at boundaries
- Height represents class freq. or rel.freq.
- Classes listed and evenly spaced along horizontal axis
- (Rel.) Freq. listed along vertical axis, scaled to account for highest (rel.) freq.
- Comparison histogram: use rel. freq. only
Freq. Histogram Example
Summer II Quiz Data, 4 classes

Frequency Distribution
Summer II Quiz Data

quiz scores
class interval includes right boundary

Freq. Histogram Example
Summer II Quiz Data, 5 classes

Frequency Histogram
Summer II Quiz Data

Quiz Scores
range ÷ (k − 1) = 39 ÷ 4 = 10
Five-number Summary

- $MIN = \text{Minimum data value}$
- $Q_1 = \text{first quartile} = \text{upper boundary of first quarter of data values}$
- $MED = \text{middle value of data} = \text{second quartile}$
- $Q_3 = \text{third quartile} = \text{upper boundary of third quarter of data values}$
- $MAX = \text{Maximum data value}$
Five-number Summary
Summer II Quiz Example

data  8 11 13 19 21 23 25 25 25 28 31 35 39 47
order(1 2 3 4 5 6 7 8 9 10 11 12 13 14)
n=14; \quad .25(n+1)=3.75; \quad .5(n+1)=7.5; \quad .75(n+1)=11.25

Use textbook formula (pages 12 & 13)

\begin{align*}
MIN &= 8 \\
Q_1 &= 16 = \frac{(3rd \, obs. \, + \, 4th \, obs.)}{2} = \frac{(13+19)}{2} \\
MED &= 25 = \frac{(7th \, obs. \, + \, 8th \, obs.)}{2} = \frac{(25+25)}{2} \\
Q_3 &= 33 = \frac{(11th \, obs. \, + \, 12th \, obs.)}{2} = \frac{(31+35)}{2} \\
MAX &= 47
\end{align*}

Construction of Boxplot
Skeletel Boxplot

1. Draw 5-lines that includes range of data and quartiles
2. Draw box extending from $Q_1$ to $Q_3$
3. Draw line inside box at median
4. Extend horizontal lines (whiskers) from box to minimum and maximum
Skeletal Boxplot
Summer II Quiz Example

Note: different computation methods used for quartiles

Schematic Boxplot

- Box and median line remained
- Outliers (outlying values) displayed with symbol such as circle
- Whiskers extended to smallest and largest values that are ‘normal’
- IQR = Inter-Quartile Range = $Q_3 - Q_1$
  (values $< Q_1 - 1.5IQR$) or (values $> Q_3 + 1.5IQR$) are outliers (Note: these values are lower fence and upper fence, respectively)
- Example: as seen in the comparison boxplots of HS GPA & GPA2 (page 12 of textbook)
Draw Boxplot Using TI Calculator

Enter data into $L_1$ by choosing STAT $\rightarrow$ EDIT $\rightarrow$ choose the list $L_1$.

After entering the data, do 2$^{nd}$STATPLOT $\rightarrow$ choose PLOT$_1$ $\rightarrow$ choose ON $\rightarrow$ Select boxplot icon (one of two types) $\rightarrow$ xLIST: $L_1$ $\rightarrow$ select ZOOM 9 shows the boxplot $\rightarrow$ TRACE $\rightarrow$ select RIGHT ARROW shows the 5-numbers

Construction of Dotplot

- Simple graph that displays along an axis line
- Each observation is shown as a dot above the line
- Illustrates the pattern of variation in data
- Provides information similar to that found in a stem-and-leaf plot
- Example: as seen in the comparison dotplots of HS GPA & GPA2 (page 13 of textbook)
Symmetry and Skewness
based on histogram, stemplot and dotplot

- **Histogram**: draw a smooth curve over the top of the bars
- **Stemplot**: first turn 90 degrees counterclockwise, draw a smooth curve over the ends of the leaves
- **Dotplot**: draw a smooth curve over the top of the stacked dots
- **Examine the tails of the curve** (examples on page 14 of textbook)
  - Left tail longer than right ⇒ left-skewed data
  - Right tail longer than left ⇒ right-skewed data
  - Both tails approximately the same ⇒ symmetric data

Symmetry and Skewness
based on boxplot

- Use only skeleton boxplot in horizontal position to examine data shape
- **Examine the length of the whiskers** (example on page 14 of textbook)
  - Left whisker longer than right ⇒ left-skewed data
  - Right whisker longer than left ⇒ right-skewed data
  - Both whiskers approximately the same ⇒ symmetric data
### Stemplot of Tomato Soup Sold

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>1 6 9</td>
</tr>
<tr>
<td>7</td>
<td>1 1 6 7</td>
</tr>
<tr>
<td>8</td>
<td>1 1 1 4 4 6 6 6 8 9</td>
</tr>
<tr>
<td>9</td>
<td>1 1 6 9</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>1 1 8</td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1 3</td>
</tr>
</tbody>
</table>

- What are the stem values: 5, 6, 7, 8, 9, 10, 11, 12, 13
- There are 9 stem values
- More stems? no
- Fewer stems? no
- 86 cases of tomato soup would most likely be sold
- There are nine classes
5-number summary:

\[ MIN = 56, \quad Q_1 = 76.5, \quad MED = 86, \quad Q_3 = 97.5, \quad MAX = 133. \]