Unit 1 - Summarizing Data  
Practice Problems

Solutions

#1.

a. Qualitative - ordinal
b. Qualitative - nominal
c. Quantitative – continuous, ratio
d. Qualitative - nominal
e. Quantitative – continuous, ratio
f. Quantitative – continuous, interval

#2a. By hand, here is the stem and leaf diagram I constructed. Other groupings for the stem are okay.

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>0</td>
<td>2 3 3 3 3 3</td>
</tr>
<tr>
<td>0</td>
<td>4 4 4 4 5 5 5</td>
</tr>
<tr>
<td>0</td>
<td>7 7 7 7 7</td>
</tr>
<tr>
<td>0</td>
<td>8 8 8 8 8</td>
</tr>
<tr>
<td>1</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>1</td>
<td>2 2 3 3</td>
</tr>
<tr>
<td>1</td>
<td>7 7</td>
</tr>
</tbody>
</table>

#2b. By hand, this is what I produced. Other class intervals are okay.

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Frequency</th>
<th>Cumulative Rel. Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>4</td>
<td>.0889</td>
<td>4</td>
<td>.0889</td>
</tr>
<tr>
<td>2-3</td>
<td>7</td>
<td>.1556</td>
<td>11</td>
<td>.2444</td>
</tr>
<tr>
<td>4-5</td>
<td>10</td>
<td>.2222</td>
<td>21</td>
<td>.4667</td>
</tr>
<tr>
<td>6-7</td>
<td>7</td>
<td>.1556</td>
<td>28</td>
<td>.6222</td>
</tr>
<tr>
<td>8-9</td>
<td>6</td>
<td>.1333</td>
<td>34</td>
<td>.7556</td>
</tr>
<tr>
<td>10-11</td>
<td>5</td>
<td>.1111</td>
<td>39</td>
<td>.8667</td>
</tr>
<tr>
<td>12-13</td>
<td>4</td>
<td>.0889</td>
<td>43</td>
<td>.9556</td>
</tr>
<tr>
<td>14-15</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>.9556</td>
</tr>
<tr>
<td>16-17</td>
<td>2</td>
<td>.0444</td>
<td>45</td>
<td>1.0000</td>
</tr>
<tr>
<td>TOTAL</td>
<td>45</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
#2c. I did not construct a histogram by hand, I used the Shodor applet that can be found under http://www.shodor.org/interactivate/activities/Histogram/

At the dialogue box “select a data set”, I scrolled down to choose My Data. Next, I scrolled down until I found a data entry box. I entered my data, selected interval size = 1.00 and then clicked on Update Histogram. Here is what I got.

If you like, you can play with different choices of interval size. For example, interval size=2.00 yields the following.
If you would like to try generating a histogram in SAS or Stata or Minitab visit the Summarizing Data topic page in the course resource website. Go to

http://www-unix.oit.umass.edu/~biep540w/webpages/summarizing.htm

#2d. I did not do this by hand, either. A frequency polygon plot is similar to a histogram. The first step is to choose class intervals. Next, note for each class interval the frequency (or relative frequency of data values in the interval). Plotted on the x-axis is the midpoint of the interval. Plotted on the y-axis is the frequency (or relative frequency).

Thus, a frequency polygon can be appreciated as an overlay of the histogram (and therefore communicating the same summarization as the histogram). With a little bit of artistic license (MS word doesn’t allow lots of precision), the frequency polygon is the graphed line below in bold black.
#3a.

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Frequency</th>
<th>Cumulative Rel. Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>1</td>
<td>.05</td>
<td>1</td>
<td>.05</td>
</tr>
<tr>
<td>15</td>
<td>13</td>
<td>.65</td>
<td>14</td>
<td>.70</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>.30</td>
<td>20</td>
<td>1.00</td>
</tr>
<tr>
<td>TOTAL</td>
<td>20</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
#3b. 

![Histogram of Age](image)

#3c. Males tend to be taller than females

<table>
<thead>
<tr>
<th>Females</th>
<th>Stem</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 3 4</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>6 9</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>0 3</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>6 7</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>0 1</td>
<td>17</td>
<td>4 3 3</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>8 7</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>3 3</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>5</td>
</tr>
</tbody>
</table>
#3d.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>150-159</td>
<td>5</td>
<td>.45</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>160-169</td>
<td>4</td>
<td>.36</td>
<td>1</td>
<td>.11</td>
</tr>
<tr>
<td>170-179</td>
<td>2</td>
<td>.18</td>
<td>5</td>
<td>.56</td>
</tr>
<tr>
<td>180-189</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>.33</td>
</tr>
</tbody>
</table>

I then used STATA to produce the histogram.

```
. graph height, by(gender) bin(6) freq ytick(0,1,2,3,4,5,6) xlabel
```

Histograms by gender
#4a.

\[
\left( X_1 + X_2 + X_3 + X_4 \right)^2 = \left[ \sum_{i=1}^{4} X_i \right]^2 \\
= (3 + 1 + 4 + 6)^2 \\
= 14^2 \\
= 196.
\]

#4b.

\[
X_1^2 + X_2^2 + X_3^2 + X_4^2 = \sum_{i=1}^{4} X_i^2 \\
= 3^2 + 1^2 + 4^2 + 6^2 \\
= 9 + 1 + 16 + 36 \\
= 62.
\]

#4c.

\[
\sum_{i=1}^{4} (X_i - 1)^2 = (3-1)^2 + (1-1)^2 + (4-1)^2 + (6-1)^2 \\
= 2^2 + 0^2 + 3^2 + 5^2 \\
= 4 + 0 + 9 + 25 \\
= 38.
\]
Note:
\[
\sum_{i=1}^{4} (X_i - 1)^2 = \sum_{i=1}^{4} [X_i^2 - 2X_i + 1]
\]
\[
= \sum_{i=1}^{4} X_i^2 - 2 \sum_{i=1}^{4} X_i + 1 \sum_{i=1}^{4}
\]
\[
= 62 - (2)(14) + (1)(4)
\]
\[
= 38.
\]

#4d.
\[
\sum_{i=1}^{4} 3X_i = 3 \sum_{i=1}^{4} X_i
\]
\[
= 3(14)
\]
\[
= 42
\]

Sol_summarizing.doc
5A. A stem and leaf diagram might come in handy. Stems are shaded, leaves are not.

<table>
<thead>
<tr>
<th></th>
<th>3 68851865</th>
<th>3 1 5 6 8 8 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>50165165310</td>
<td>4 0 0 1 1 3 5 5 5 6 6</td>
</tr>
<tr>
<td>5</td>
<td>39113</td>
<td>5 1 1 3 3 9</td>
</tr>
<tr>
<td>6</td>
<td>90</td>
<td>6 0 9</td>
</tr>
</tbody>
</table>

\[ \text{MEAN } \bar{x} = \frac{1}{n} \sum_{i=1}^{26} X_i = \frac{1}{n} (1156) = 44.46 \text{ so } \bar{x} = 44.5 \]

\[ \text{MEDIAN } \text{First solve } \left( \frac{n+1}{2} \right) = \left( \frac{26+1}{2} \right) = 13.5 \]
Median is midpoint of 13\textsuperscript{th} and 14\textsuperscript{th} observation.

\[ \bar{x} = \frac{1}{2} (41 + 43) = 42 \text{ so } \bar{x} = 42 \]

\[ \text{MODE } \text{This sample is tri-modal} \]

\[ \text{RANGE } \text{Maximum - Minimum} = 69 - 31 \text{ so range} = 38 \]
VARIANCE  Let’s save ourselves the trouble of a very long brute force formula by using the formula for grouped data.

Let \( j \) index the unique values. There are 14 unique values.

<table>
<thead>
<tr>
<th>( j )</th>
<th>( X_j )</th>
<th>( f_j )</th>
<th>( (x_j - \bar{x})^2 )</th>
<th>( f_j (x_j - \bar{x})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>1</td>
<td>182.25</td>
<td>182.25</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>2</td>
<td>90.25</td>
<td>180.50</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>2</td>
<td>72.25</td>
<td>144.50</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>3</td>
<td>42.25</td>
<td>126.75</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>2</td>
<td>20.25</td>
<td>40.50</td>
</tr>
<tr>
<td>6</td>
<td>41</td>
<td>3</td>
<td>12.25</td>
<td>36.75</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>1</td>
<td>2.25</td>
<td>2.25</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>3</td>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>9</td>
<td>46</td>
<td>2</td>
<td>2.25</td>
<td>4.50</td>
</tr>
<tr>
<td>10</td>
<td>51</td>
<td>2</td>
<td>42.25</td>
<td>84.50</td>
</tr>
<tr>
<td>11</td>
<td>53</td>
<td>2</td>
<td>72.25</td>
<td>144.50</td>
</tr>
<tr>
<td>12</td>
<td>59</td>
<td>1</td>
<td>210.25</td>
<td>210.25</td>
</tr>
<tr>
<td>13</td>
<td>60</td>
<td>1</td>
<td>240.25</td>
<td>240.25</td>
</tr>
<tr>
<td>14</td>
<td>69</td>
<td>1</td>
<td>600.25</td>
<td>600.25</td>
</tr>
<tr>
<td><strong>TOTALS</strong></td>
<td></td>
<td>26</td>
<td></td>
<td><strong>1998.50</strong></td>
</tr>
</tbody>
</table>

\[
S^2 = \frac{\sum_{j=1}^{14} f_j (x_j - \bar{x})^2}{\left(\sum_{j=1}^{14} f_j\right) - 1} = \frac{1998.50}{25} \quad \text{So} \quad S^2 = 79.94
\]

Standard deviation \( S = \sqrt{S^2} \) \( \text{So} \quad S = 8.94 \).
25th Percentile

First solve \((.25) (n) = (.25) (26) = 6.5\)
So 25th percentile is the 7th observation \(P_{25} = 38\)

75th Percentile

First solve \((.75) (n) = (.75) (26) = 19.5\)
So 75th percentile is the 20th observation \(P_{75} = 51\)

\[\begin{array}{l}
\text{5B.} \\
\hline
2 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 2 \\
2 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \\
2 & 8 & 8 & 8 \\
3 & 0 & 1 \\
3 & 4 & 4 \\
\hline
\end{array}\]

\[\text{MEAN } \bar{x} = \frac{1}{n} \sum_{i=1}^{21} X_i = \frac{1}{21} (568) = 27.04 \quad \text{So } \bar{x} = 27.0\]

\[\text{MEDIAN} \quad \text{Solving } \left(\frac{n+1}{2}\right) = \left(\frac{21+1}{2}\right) = 11\]
Median is the 11th observation. \quad \text{So } \tilde{x} = 26

\[\text{MODE} \quad \text{mode } = 25\]

\[\text{RANGE} \quad \text{Maximum } - \text{ Minimum} \]
\[= 34 - 25. \quad \text{So } \text{Range } = 9\]
Variance There are 6 unique values.

<table>
<thead>
<tr>
<th>j</th>
<th>X_j</th>
<th>f_j</th>
<th>$(x_j - \bar{x})^2$</th>
<th>$f_j(x_j - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>9</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>1</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
<td>1</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>34</td>
<td>2</td>
<td>49</td>
<td>98</td>
</tr>
<tr>
<td>TOTALS</td>
<td>21</td>
<td>49</td>
<td>167</td>
<td></td>
</tr>
</tbody>
</table>

\[
S^2 = \frac{\sum_{j=1}^{6} f_j(x_j - \bar{x})^2}{\left(\sum_{j=1}^{6} f_j\right) - 1} = \frac{167}{20} = 8.35
\]

So \( S^2 = 8.35 \)

Standard deviation \( S = \sqrt{S^2} = \sqrt{8.35} = 2.89 \)

25th Percentile

Solving (.25) (n) = (.25) (21) = 5.25
So 25th percentile is 6th observation \( P_{25} = 25 \)

Note - I get this by noticing from the table above that the smallest value (=25) occurs with a frequency of 9 times in the sample.

75th Percentile

Solving (.75) (n) = (.75) (21) = 15.75
So 75th percentile is 16th observation \( P_{75} = 28 \)

Note – I get this by noticing in the table that the value = 28 occurs with a frequency of 3 times in the sample and comes after the first 9 observations all equal to 25 and after the next 5 observations all equal to 26, so that the value of 28 is the 15th, 16th and 17th observations in the ordered sample.
5C. **REMEMBER** - Use the *same* scale when comparing two groups.

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients</td>
<td>Controls</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>44.5</td>
<td>27.0</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>42</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>P25</td>
<td>38</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>P75</td>
<td>51</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range (IQR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P25 - (1.5)(IQR)</td>
<td>18.5</td>
<td>20.5</td>
<td></td>
</tr>
<tr>
<td>P75 + (1.5)(IQR)</td>
<td>70.5</td>
<td>32.5*</td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>31*</td>
<td>25*</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>69*</td>
<td>34</td>
<td></td>
</tr>
</tbody>
</table>

* = Whisker

**Notes on Whiskers**

1) IF P<sub>25</sub> - (1.5) (IQR) < minimum of the actual data, so use minimum of actual data instead
2) IF P<sub>75</sub> + (1.5) (IQR) > maximum of the actual data, so use maximum of actual data instead

---

Exercise #5C – Box and Whisker Plot

Healthy (n=21), Panic Disorder (n=26)

Sol_summarizing.doc
6A.

<table>
<thead>
<tr>
<th>Class Endpoints</th>
<th>Class Midpoint</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Frequency</th>
<th>Cumulative Relative Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-14.99</td>
<td>10</td>
<td>5</td>
<td>.067</td>
<td>5</td>
<td>.067</td>
</tr>
<tr>
<td>15-24.99</td>
<td>20</td>
<td>10</td>
<td>.133</td>
<td>15</td>
<td>.200</td>
</tr>
<tr>
<td>25-34.99</td>
<td>30</td>
<td>20</td>
<td>.267</td>
<td>35</td>
<td>.467</td>
</tr>
<tr>
<td>35-44.99</td>
<td>40</td>
<td>22</td>
<td>.293</td>
<td>57</td>
<td>.760</td>
</tr>
<tr>
<td>45-54.99</td>
<td>50</td>
<td>13</td>
<td>.173</td>
<td>70</td>
<td>.933</td>
</tr>
<tr>
<td>55-64.99</td>
<td>60</td>
<td>5</td>
<td>.067</td>
<td>75</td>
<td>1.000</td>
</tr>
</tbody>
</table>

TOTALS

1.000

6B.
A cumulative relative frequency polygon for grouped data is, unfortunately, not straightforward in SAS or Stata.

**Solution using Excel.**

**Step 1:** Enter your “x” and “y” points into your worksheet such that
“x” = Endpoint of class interval
“y” = Cumulative relative frequency for the interval
note – Be sure to include an (x,y) = (0,0)

<table>
<thead>
<tr>
<th>x=age</th>
<th>y=cumulative relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0.067</td>
</tr>
<tr>
<td>25</td>
<td>0.2</td>
</tr>
<tr>
<td>35</td>
<td>0.467</td>
</tr>
<tr>
<td>45</td>
<td>0.76</td>
</tr>
<tr>
<td>55</td>
<td>0.933</td>
</tr>
<tr>
<td>65</td>
<td>1</td>
</tr>
</tbody>
</table>
**Step 2:** Use the chart wizard in excel as follows.

<table>
<thead>
<tr>
<th>Highlight the data you want to plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Data Table Image]</td>
</tr>
</tbody>
</table>

Click on the chart wizard from the upper toolbar

<table>
<thead>
<tr>
<th>Under Chart Type:</th>
<th>Select XY (Scatter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Chart Wizard Image]</td>
<td></td>
</tr>
</tbody>
</table>

Under Chart sub-type
- Highlight the plot with the dots connected

Click Next

<table>
<thead>
<tr>
<th>You should see the following</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Graph Image]</td>
</tr>
</tbody>
</table>

Click Next

Sol_summarizing.doc
You will then see a menu that lets you add legends and titles, etc.

And, if you like, you can change such things as shading, tick marks, etc.

After some aesthetics on my part, this is what I got:

Estimates are $P_{10} = 17$  $P_{50} = 36$  $P_{25} = 26$  $P_{75} = 44.5$
6C.

<table>
<thead>
<tr>
<th>Midpoint $X_j$</th>
<th>Frequency $f_j$</th>
<th>$X_j f_j$</th>
<th>$(x_j - \bar{x})$</th>
<th>$f_j (x_j - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5</td>
<td>50</td>
<td>-25.7</td>
<td>3302.45</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>200</td>
<td>-15.7</td>
<td>2464.90</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>600</td>
<td>-5.7</td>
<td>649.80</td>
</tr>
<tr>
<td>40</td>
<td>22</td>
<td>880</td>
<td>4.3</td>
<td>406.78</td>
</tr>
<tr>
<td>50</td>
<td>13</td>
<td>650</td>
<td>14.3</td>
<td>2658.37</td>
</tr>
<tr>
<td>60</td>
<td>5</td>
<td>300</td>
<td>24.3</td>
<td>2952.45</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>75</strong></td>
<td><strong>2680</strong></td>
<td></td>
<td><strong>12434.75</strong></td>
</tr>
</tbody>
</table>

**MEAN** \( \bar{x} = \frac{\sum_{j=1}^{6} f_j x_j}{\sum_{j=1}^{6} f_j} = \frac{2680}{75} \) So \( \bar{x} = 35.7 \)

**MEDIAN**  
Note to reader – I’ve consulted a number of texts on this. There is no single correct answer. With interval data, whatever median you calculate is an approximation. Here is what is suggested in Think and Explain with Statistics (Lincoln E. Moses, page 64)

First solve \( \frac{n + 1}{2} = \frac{75 + 1}{2} = 38th \) observation

Examination of the table reveals that the 38th observation is in the interval 35 to 44.99

Set the following quantities:

- The letter \( l \) = lower limit of interval = 35
- The letter \( u \) = upper limit of interval = 44.99
- \( R \) = cumulative frequency up to the lower limit of interval = 35
- \( M \) = # observations contained in interval = 22
- \( N \) = total # observations = 75

An approximate solution for the median is calculated as

\[
\tilde{x} = l + \left[ \frac{N/2 - R}{M} \right] (u - l) = 35 + \left[ \frac{75/2 - 35}{22} \right] (44.99 - 35) = 36.135 \text{ or } 37
\]
VARIANCE

\[ S^2 = \frac{\sum_{j=1}^{6} f_j (x_j - \bar{x})^2}{\left(\sum_{j=1}^{6} f_j\right) - 1} = \frac{12434.75}{74} \text{ so } S^2 = 168.04 \]

Standard deviation \( S = \sqrt{S^2} \) so \( S = 13.0 \)