1. Summarizing Data

Homework #2 (Unit 1 – Summarizing Data)

Solutions

#1a.

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

---

#1b.

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

---

#1c.

\[
\sum_{i=1}^{4} \left( X_i - 1 \right)^2 = (3 - 1)^2 + (1 - 1)^2 + (4 - 1)^2 + (6 - 1)^2 \\
= 2^2 + 0^2 + 3^2 + 5^2 \\
= 4 + 0 + 9 + 25 \\
= 38.
\]
1. Summarizing Data

Note:
\[
\sum_{i=1}^{4} (X_i - 1)^2 = \sum_{i=1}^{4} \left[ X_i^2 - 2X_i + 1 \right] \\
= \sum_{i=1}^{4} X_i^2 - 2 \sum_{i=1}^{4} X_i + 4 \sum_{i=1}^{4} \\
= 62 - (2)(14) + (1)(4) \\
= 38.
\]

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

# 2a. A stem and leaf diagram might come in handy. Stems are shaded, leaves are not.

```
  3 | 68851865      →  3 | 1 5 5 6 6 8 8 8  
  4 | 50165165310          4 | 0 0 1 1 1 3 5 5 5 6 6  
  5 | 39113                          5 | 1 1 3 3 9  
  6 | 90                               6 | 0 9  
```

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

**MEDIAN**
First solve \( \left( \frac{n+1}{2} \right) = \left( \frac{26+1}{2} \right) = 13.5 \)

Median is midpoint of 13th and 14th observation.
\[
\bar{x} = \frac{1}{2} (41 + 43) \\
so \ \bar{x} = 42
\]
1. Summarizing Data

**MODE**  This sample is tri-modal  38, 41, 45

**RANGE**  Maximum - Minimum

\[
= 69 - 31
\]

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>
</tbody>
</table>

**TOTALS**  26  1998.50

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

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

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

75th Percentile

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

#2b.

\[
\begin{array}{cccccccccc}
2 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\
2 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \\
2 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 \\
3 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
3 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\
\end{array}
\]

\(\text{MEAN} \quad \bar{x} = \frac{1}{n} \sum X_i = \frac{1}{21} (568) = 27.04 \quad \text{So} \quad \bar{x} = 27.0\)

\(\text{MEDIAN} \quad \text{Solving} \quad \left(\frac{n+1}{2}\right) = \left(\frac{21+1}{2}\right) = 11\)

Median is the 11\(^{th}\) observation. \quad \text{So} \quad \bar{x} = 26

\(\text{MODE} \quad \text{mode} = 25\)

\(\text{RANGE} \quad \text{Maximum} = \text{Minimum} \quad = 34 - 25 \quad \text{So} \quad \text{Range} = 9\)
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><strong>TOTALS</strong></td>
<td><strong>21</strong></td>
<td><strong>80</strong></td>
<td><strong>167</strong></td>
<td></td>
</tr>
</tbody>
</table>

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

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

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.
#2c. Let’s produce a side-by-side box plot using lock5stat.com, which was introduced in “Show me #1”

__step 1. Launch www.lock5stat.com. The home page should appear__

__step 2. From the left navigation bar, click on StatKey__
__step 3. Click on **One Quantitative Variable and One Categorical Variable**

__step 4. Click on **Edit Data**
step 6. Delete the data that is shown, taking care to keep the header “label, value”

step 7. Enter the data for the panic disorder group, row by row. Then enter the data for the controls, row by row. Check your work before clicking the OK. Then click OK
step 8. StatKey will return a dot plot. Click on **BOX PLOT**

---

**StatKey**

Descriptive Statistics for one Quantitative and one Categorical Variable

Summary Statistics

<table>
<thead>
<tr>
<th>Statistics</th>
<th>panic</th>
<th>control</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>29</td>
<td>21</td>
<td>47</td>
</tr>
<tr>
<td>Mean</td>
<td>44.5</td>
<td>27.0</td>
<td>36.7</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>8.9</td>
<td>2.9</td>
<td>11.1</td>
</tr>
<tr>
<td>Minimum</td>
<td>31</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>Q1</td>
<td>38.00</td>
<td>25.00</td>
<td>26.00</td>
</tr>
<tr>
<td>Median</td>
<td>42.00</td>
<td>29.00</td>
<td>33.00</td>
</tr>
<tr>
<td>Q3</td>
<td>51.00</td>
<td>28.00</td>
<td>44.00</td>
</tr>
<tr>
<td>Maximum</td>
<td>69</td>
<td>34</td>
<td>69</td>
</tr>
</tbody>
</table>

---

#3a.

<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 |
#3b.

<table>
<thead>
<tr>
<th>Midpoint X&lt;sub&gt;j&lt;/sub&gt;</th>
<th>Frequency f&lt;sub&gt;j&lt;/sub&gt;</th>
<th>X&lt;sub&gt;j&lt;/sub&gt;f&lt;sub&gt;j&lt;/sub&gt;</th>
<th>(X&lt;sub&gt;j&lt;/sub&gt; - ̅X)</th>
<th>f&lt;sub&gt;j&lt;/sub&gt;(X&lt;sub&gt;j&lt;/sub&gt; - ̅X)&lt;sup&gt;2&lt;/sup&gt;</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><strong>12434.75</strong></td>
<td></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  
\[ \frac{n + 1}{2} = \frac{75 + 1}{2} = 38^{th} \text{ observation} \]

Examination of the table reveals that the 38<sup>th</sup> 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

\[ \hat{x} = l + \left( \frac{N/2 - R}{M} \right) (u - l) = 35 + \left( \frac{75/2 - 35}{22} \right) (44.99 - 35) = 36.135 \text{ or } 37 \]
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 \[ S = \sqrt{S^2} \text{ so } S = 13.0 \]