Introduction to Statistics for the Life Sciences

Spring 2015
Chapter 2, continued.
Example of a heavy right-tailed distribution

Formally, we say skewed to the right.
Example of a heavy left-tailed distribution

Formally, we say skewed to the left.
Example of a distribution with similar left and right tails.

Formally, we say **symmetric**.
Unimodal

**Figure**: These are all examples of **unimodal** distributions: they have one peak.
A **bimodal** distribution has two peaks.
A statistic is a numerical measure calculated from a set of data. For example, the mean.
Suppose we observe a sample of size $n$ of a numeric variable. Denote the observed values by $y_1, \ldots, y_n$.

**Sample Mean**

The sample mean, or ordinary average, is defined as

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$
Suppose we observe a sample of size $n$ of a numeric variable. Denote the observed values by $y_1, \ldots, y_n$.

**Sample Mean**

The sample mean, or ordinary average, is defined as

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$ 

For example, if $n = 3$, $y_1 = 2$, $y_2 = 3$, and $y_3 = 4$, then

$$\bar{y} = \frac{1}{3} (2 + 3 + 4) = \frac{1}{3} (9) = 3$$

**Measure of Center**

A statistic is a measure of center if it attempts to offer a typical value for a data set.
Definition

Suppose we observe a sample of size \( n \) of a numeric variable. Denote the observed values by \( y_1, \ldots, y_n \).

Deviation

We call the value of each observation minus the \( i \)th observation, or \( y_i - \bar{y} \), a deviation.

The sum of deviations is always 0:

\[
\sum_{i=1}^{n} (y_i - \bar{y}) = y_1 + \ldots + y_n - n\bar{y} = n\bar{y} - n\bar{y} = 0
\]
Suppose we observe a sample of size $n$ of a numeric variable. Denote the observed values by $y_1, \ldots, y_n$.

**Order Statistics**

For $n$ observed values, the observed order statistics are simply a re-ordering of the observed values in ascending order. They are denoted by

$$y(1), y(2), \ldots, y(n)$$

where

$$y(1) \leq y(2) \leq \cdots \leq y(n).$$
Example

Suppose the data are given by

\[ 4 \quad 2 \quad 6 \quad 4 \quad 7. \]

We could write

\[
\begin{align*}
  y_1 &= 4 \\
  y_2 &= 2 \\
  y_3 &= 6 \\
  y_4 &= 4 \\
  y_5 &= 7
\end{align*}
\]
Example

Suppose the data are given by

\[ 4 \quad 2 \quad 6 \quad 4 \quad 7. \]

Or alternatively,

\[
\begin{array}{c|c}
  y_1 &= 4 \\
  y_2 &= 2 \\
  y_3 &= 6 \\
  y_4 &= 4 \\
  y_5 &= 7 \\
\hline
  y(1) &= 2 \\
  y(2) &= 4 \\
  y(3) &= 4 \\
  y(4) &= 6 \\
  y(5) &= 7 \\
\end{array}
\]
Definition

Minimum
The minimum of $n$ numeric observations is given by $y_{(1)}$.

Maximum
The maximum of $n$ numeric observations is given by $y_{(n)}$. 

Another Measure of Center

Suppose we observe a sample of size $n$ of a numeric variable. Denote the observed values by $y_1, \ldots, y_n$.

Median

The median, denoted by $\tilde{y}$, is defined by two cases. Case 1:

$n$ is odd. \quad \Rightarrow \quad \tilde{y} = y\left(\frac{n+1}{2}\right).

Case 2:

$n$ is even. \quad \Rightarrow \quad \tilde{y} = \frac{1}{2} \left(y\left(\frac{n}{2}\right) + y\left(\frac{n}{2}+1\right)\right).
Example

Consider the previous data:

4 2 6 4 7.

Calculate the median: Since \( n = 5 \), we can write

\[ \tilde{y} = y_{\left(\frac{n+1}{2}\right)} = y_{(6/2)} = y_{(3)}, \]

or the third order statistic (after the data have been arranged in ascending order). From the previous table, we know \( \tilde{y} = 4 \).
Example

Suppose instead the data are given by

\[ 4 \ 2 \ 6 \ 4 \ 7 \ 9 \]

Calculate the median: Since \( n = 6 \), we can write the ordered data \( y(i) \):

\[ 2 \ 4 \ 4 \ 6 \ 7 \ 9. \]

Then

\[
\tilde{y} = \frac{1}{2} \left( y\left( \frac{n}{2} \right) + y\left( \frac{n}{2} + 1 \right) \right) = \frac{1}{2} \left( y(3) + y(4) \right) = \frac{1}{2} (4 + 6) = 5.
\]
Suppose the following ordered data:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.39</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
<td>3.83</td>
</tr>
</tbody>
</table>
Suppose the following ordered data:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1.39</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1.39</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
</tr>
</tbody>
</table>

Then

\[ \bar{y} = 2.65 \quad \hat{y} = 2.69 \]
Median Versus Mean

Suppose the following ordered data:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1.39</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1.39</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
</tr>
</tbody>
</table>

If $y_{(10)} = 4$ instead of 3.83, then

$$\bar{y} = 2.67, \quad \tilde{y} = 2.69$$
Suppose the following ordered data:

\[
\begin{array}{cccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
1 & 1.39 & 1.50 & 1.55 & 2.40 & 2.65 & 2.72 & 3.11 & 3.50 & 3.83 & 3.83 \\
\end{array}
\]

If \( y_{(10)} = 5 \) instead of 3.83, then

\[
\bar{y} = 2.77 \quad \tilde{y} = 2.69
\]
Suppose the following ordered data:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1.39</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
</tr>
</tbody>
</table>

If \( y_{(10)} = 6 \) instead of 3.83, then

\[
\bar{y} = 2.86 \quad \hat{y} = 2.69
\]
Suppose the following ordered data:

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.39</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
</tr>
</tbody>
</table>

If \( y_{(10)} = 7 \) instead of 3.83, then

\[
\bar{y} = 2.96 \quad \hat{y} = 2.69
\]
Median Versus Mean

Suppose the following ordered data:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
<td>3.83</td>
</tr>
</tbody>
</table>

If \( y_{(10)} = 8 \) instead of 3.83, then

\[
\bar{y} = 3.06 \quad \tilde{y} = 2.69
\]
Suppose the following ordered data:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1.39</td>
<td>1.50</td>
<td>1.55</td>
<td>2.40</td>
<td>2.65</td>
<td>2.72</td>
<td>3.11</td>
<td>3.50</td>
<td>3.83</td>
</tr>
</tbody>
</table>

If \( y_{(10)} = 100 \) instead of 3.83, then

\[
\bar{y} = 12.27 \quad \tilde{y} = 2.69
\]
What happened? Why isn’t the median moving? Amazing?!? Or obvious?

Robustness

A statistic is robust if it is relatively unaffected by changes in a small portion of the data.
What happened? Why isn’t the median moving? Amazing?!? Or obvious?

**Robustness**

A statistic is robust if it is relatively unaffected by changes in a small portion of the data.

Conclusion: $\bar{y}$ is not robust. $\tilde{y}$ is robust.
What happened? Why isn’t the median moving? Amazing?!? Or obvious?

**Robustness**

A statistic is robust if it is relatively unaffected by changes in a small portion of the data.

Conclusion: $\bar{y}$ is not robust. $\tilde{y}$ is robust.

Question: Is $\tilde{y}$ always better than $\bar{y}$?
Definition

**Q₁ – The First Quartile**

In a data set of \( n \) observations, \( Q₁ \) is defined as the median of those observations in the lower half of the data set.

**Q₃ – The Third Quartile**

In a data set of \( n \) observations, \( Q₃ \) is defined as the median of those observations in the upper half of the data set.
The second quartile = the median

\[ Q_2 = \tilde{y} \]
Example

Find $Q_1$ and $Q_3$ for the following data:

8  2  3  5  1  3  6.

Ordered:

1  2  3  3  5  6  8

Since $n = 7$ is odd, we know that $\tilde{y} = y_{(4)} = 3$. To find $Q_1$, take the median of 1, 2, and 3, so that $Q_1 = 2$. Similarly, $Q_3$ is the median of 5, 6, and 8, giving $Q_3 = 6$. 
Example

Find $Q_1$ and $Q_3$ for the following data:

\[ 8 \ 2 \ 3 \ 5 \ 1 \ 3 \ 6. \]

Ordered:

\[ 1 \ 2 \ 3 \ 3 \ 5 \ 6 \ 8 \]

Since $n = 7$ is odd, we know that $\tilde{y} = y_{(4)} = 3$. To find $Q_1$, take the median of 1, 2, and 3, so that $Q_1 = 2$. Similarly, $Q_3$ is the median of 5, 6, and 8, giving $Q_3 = 6$. 
Example

Repeat the previous question with a slightly modified dataset:

8  2  3  5  1  3  6  9

Ordered:

1  2  3  3  5  6  8  9

Since $n = 8$ is even, we know that

$\tilde{y} = (y_{(4)} + y_{(5)})/2 = (3 + 5)/2 = 4$. To find $Q_1$, take the median of 1, 2, 3, and 3, so that $Q_1 = 2.5$. Similarly, $Q_3$ is the median of 5, 6, 8 and 9, giving $Q_3 = 7$. 
Example

Repeat the previous question with a slightly modified dataset:

\[
\begin{array}{cccccccc}
8 & 2 & 3 & 5 & 1 & 3 & 6 & 9 \\
\end{array}
\]

Ordered:

\[
\begin{array}{cccccccc}
1 & 2 & 3 & 3 & 5 & 6 & 8 & 9 \\
\end{array}
\]

Since \( n = 8 \) is even, we know that
\[
\tilde{y} = \frac{y_{(4)} + y_{(5)}}{2} = \frac{3 + 5}{2} = 4.
\]
To find \( Q_1 \), take the median of 1, 2, 3, and 3, so that \( Q_1 = 2.5 \). Similarly, \( Q_3 \) is the median of 5, 6, 8 and 9, giving \( Q_3 = 7 \).
Repeat the previous question with an ordered dataset:

1 1 3 3 5 5

Since \( n = 6 \) is even, we know that 
\[
\bar{y} = \frac{y_{(3)} + y_{(4)}}{2} = \frac{3 + 3}{2} = 3.
\]
To find \( Q_1 \), take the median of 1, 1 and 3, so that \( Q_1 = 1 \). Similarly, \( Q_3 \) is the median of 3, 5 and 5, giving \( Q_3 = 5 \).
Interquartile Range

The interquartile range is the difference between the third and first quartile, or

\[ IQR = Q_3 - Q_1 \]

In the previous slide, \( Q_3 = 7 \) and \( Q_1 = 2.5 \); so

\[ IQR = 7 - 2.5 = 4.5 \]
Definition

5 Number Summary

The 5 number summary for a data set is given by

\[y(1), Q_1, \bar{y}, Q_3, y(n),\]

or in plain language, the minimum, the first quartile, the median, the third quartile, and the maximum.

5 number summary from previous example

The ordered data were

\[1 2 3 3 5 6 8 9,\]

giving a 5 number summary of

\[y(1) = 1, Q_1 = 2.5, \bar{y} = 4, Q_3 = 7, y(n) = 9.\]
Definition

5 Number Summary

The 5 number summary for a data set is given by

\[ y(1), Q_1, \tilde{y}, Q_3, y(n), \]

or in plain language, the minimum, the first quartile, the median, the third quartile, and the maximum.

5 number summary from previous example

The ordered data were

\[ 1 \ 2 \ 3 \ 3 \ 5 \ 6 \ 8 \ 9, \]

giving a 5 number summary of

\[ y(1) = 1, Q_1 = 2.5, \tilde{y} = 4, Q_3 = 7, y(n) = 9. \]
Definition

Loosely speaking,

**Outlier**

A data point that differs from the rest of the data so much that it seems not to belong is called an outlier.

Formally,

**Outlier**

If in data points $y_1, ..., y_n$ there exists a data point, say $y_j$, such that either

$$y_j < Q_1 - 1.5 \times IQR = \text{Lower Fence}$$

or

$$y_j > Q_3 + 1.5 \times IQR = \text{Upper Fence}$$

then $y_j$ is considered an outlier.
Definition

Loosely speaking,

**Outlier**

A data point that differs from the rest of the data so much that it seems not to belong is called an outlier.

Formally,

**Outlier**

If in data points $y_1, ..., y_n$ there exists a data point, say $y_j$, such that either

$$y_j < Q_1 - 1.5 \times IQR = \text{Lower Fence}$$

or

$$y_j > Q_3 + 1.5 \times IQR = \text{Upper Fence}$$

then $y_j$ is considered an outlier.
Outlier Example

Dotplot of Observations

Value of Observations

Relative Frequency
Outlier Example

5 number summary:

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.21</td>
<td>-0.74</td>
<td>0.03</td>
<td>0.25</td>
<td>4.00</td>
</tr>
</tbody>
</table>
Outlier Example

\[ IQR = .25 - (-.74) = 0.99 \]

So lower fence = \(-.74 - 1.5 \times .99 = -2.24\); upper fence = \(.25 + 1.5 \times .99 = 1.75\). Conclusion? There is one outlier, the observation equal to 4.
The natural graphical way to encode the quantile information is a boxplot. For the dot chart on the last page, the boxplot follows:

Note the upper fence value of 1.75 is NOT where the right whisker ends.
The natural graphical way to encode the quantile information is a boxplot. For the dot chart on the last page, the boxplot follows:

Note the upper fence value of 1.75 is NOT where the right whisker ends.
Boxplot of Observations

Dotplot of Observations
Boxplot

Boxplot of Observations

- 1st line: Minimum
- 2nd line: First Quartile
- 3rd line: Median
- 4th line: Third Quartile
- 5th line: Maximum
Boxplot

- Minimum
- 25th percentile
- Median
- 75th percentile
- IQR
- 1.5 x IQR

To minimum
The textbook calls the type of boxplot just presented a **modified boxplot**. We won’t make a distinction between the two types of boxplots in the textbook. For us, all boxplots will be so-called “modified boxplots.”
Diamonds Data Set

- 'data.frame': 53940 obs. of 10 variables:
  - carat - numeric
  - cut - categorical (ordinal)
  - color - categorical
  - clarity - categorical (ordinal?)
  - depth - numeric
  - table - numeric
  - price - numeric
  - x - numeric
  - y - numeric
  - z - numeric
Boxplots as a Basis for Comparison

Boxplot of Price versus Cut Quality

Cut Quality

Fair  Good  Very Good  Premium  Ideal

Price

0  5000  10000  15000
Boxplot of Carat versus Cut Quality
Looking deeper into the Diamond Data

Bar Chart of Diamonds by Cut

- Fair
- Good
- Very Good
- Premium
- Ideal

Frequency
Looking deeper into the Diamond Data

Histogram of Carat

Frequency vs Carat
Looking deeper into the Diamond Data

Scatter Plot of Price versus Carat