Read: Sections 1 through 5 pp 92-149

Work the following text examples:
Section 3.2, 3-1 through 3-17
Section 3.3, 3-22 through 3.28, 3-42 through 3.82
Section 3.4, 29, 30, 32-46
Section 3.5, 38-42

Examples to be worked in class:
3.2 Exercise 6, 14, 28 p. 107-109
3.3 Exercise 52, 14, 74
3.4 Exercise 102, 110, and 111
3.5 Exercise 122, 130

Homework:
Exercises 1-21 odd, 42 - 67 odd, 89-105 odd, 107, 109, 118, 119, 120, 123 - 129 odd
3.4 Measures of Position

- Measures of position are used to locate relative position within a dataset.
- Example, the 50th percentile (the median) indicates that half of the observations fall below the median and half fall above the median.
- We can make comparisons of position between datasets using a z-score.

The z-score

- (defined) a measure obtained by subtracting the mean from a value and dividing the result by the standard deviation.
- General formula:
  \[ z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \]
- z-score for samples:
  \[ z = \frac{X - \bar{X}}{s} \]
- z-score for populations:
  \[ z = \frac{X - \mu}{\sigma} \]
- interpretation: the number of standard deviations a data value falls above or below the mean.
- If all the data are transformed into z-scores, the resulting distribution will have a mean of zero and a standard deviation of 1.
- Example - Exercise 102 on p. 136 Math \( z \)-score = \( (60 - 54)/3 = 2.0 \) and History \( z \)-score = \( (80 - 75)/2 = 2.5 \). She did better on the History exam relative to the rest of the class. What if the standard deviation of the history scores was the same as for History? Then she would have had a higher math score relative to the class.

Percentiles

- Percentiles divide the data into 100 equal groups.
- Percentiles are used to compare scores.
- The percentile represents the relative position of the score - not the score itself.
- We can find the percentile rank of a certain data point with the following formula (where \( X \) is the particular data value in question).
  \[ \text{percentile} = \frac{\text{number of values below } X + .5}{\text{total number of values}} \times 100\% \]
Observation | Percentile  
---|---
78 | \((0 + \frac{.5}{6}) \times 100\% = 8\text{th}\)
82 | \((1 + \frac{.5}{6}) \times 100\% = 25\text{th}\)
86 | \((2 + \frac{.5}{6}) \times 100\% = 42\text{nd}\)
88 | \((3 + \frac{.5}{6}) \times 100\% = 58\text{th}\)
92 | \((4 + \frac{.5}{6}) \times 100\% = 75\text{th}\)
97 | \((5 + \frac{.5}{6}) \times 100\% = 92\text{nd}\)

What if we want to find a data value corresponding to a given percentile? Note, the solution may not be part of the observed values.

See Procedure Table on p. 133

Example - Exercise 111 p. 137 What value corresponds to the 30th percentile?

- Step 1: Sort the data  
  
  78, 82, 86, 88, 92, 97

- Step 2: Let \(n\) = total number of values in the data set and let \(p\) be the percentile. Now, calculate \(c\) as follows:
  
  \[ c = \frac{n \times p}{100} = \frac{6 \times 30}{100} = 1.8 \]

- Step 3A: If \(c\) is not a whole number then round up to the next whole number. In our case, we round up to 2. Now, using the sorted dataset count over 2 positions. In our case, 82 represents the 30th percentile.

What if \(c\) is a whole number? Then we use the value halfway between the \(c\) and \(c + 1\) when counting from the minimum.

Example - Hypothetical dataset - find the 60th percentile.

- Step 1: Sort the data  
  
  33, 34, 37, 39, 41, 42, 48, 50, 51, 55

- Step 2: Let \(n\) = total number of values in the data set and let \(p\) be the percentile. Now, calculate \(c\) as follows:
  
  \[ c = \frac{n \times p}{100} = \frac{10 \times 60}{100} = 6 \]

- Step 3B: Since \(c\) is a whole number we will take the middle value between the 6th and 7th values in the ordered dataset. The 6th value is 42 and the 7th is 48. The middle value is 45 which represents the 60th percentile.
Quartiles

- Quartiles divide the distribution of the data values into 4 groups.
- The second quartile is equal to the median, the 50th percentile.
- Our procedure for finding quartiles involves finding medians.

**Procedure**
- Find the median. This is Q2.
- Find the median of the values less than Q2. This is Q1.
- Find the median of the values greater than Q2. This is Q3.

Example - Hypothetical dataset.

- Step 1: Sort the data
  33, 34, 37, 39, 41, 42, 48, 50, 51, 55
- Step 2: Find the median. This is Q2. Averaging 41 and 42, Q2 = 41.5.
- Step 3: Find the median of the values less than 41.5. In our case, Q1 = 37, which is the middle value of the data points less than 41.5.
- Step 4: Find the median of the values above 41.5. In our case, Q3 = 50, which is the middle value of the data points greater than 41.5.

**Interquartile Range (IQR)**

The calculation for IQR is given by $IQR = Q_3 - Q_1$. The IQR is another measure of dispersion (variability).
Exploratory Data Analysis (EDA)

- Traditional data analysis uses frequency distributions to make inference about populations.
- EDA uses stem-and-leaf plots along with box-plots to assess what information can be obtained from the data.
- In EDA we are looking for patterns.
- Data mining is a current application of EDA to enormous datasets.

The Stem-and-Leaf Plot

- This graphic produces a histogram-type picture while maintaining the actual data values.
- A data value is split into two parts. The leading digit(s) are called the stem. The trailing digit(s) are called the leaves.

Example - Exercise 122, p. 150. We have $n = 29$ observations.

- Step 1: Sort the data (makes it easier)
  
  3, 8, 9, 9, 10, 12, 12, 12, 14, 14, 16, 18, 19, 21, 22, 22, 25, 28, 29, 31, 33, 36, 37, 41, 49, 52, 54, 58

- Step 2: Decide on a partition of the data value. We will use the 10’s position as the stem and the one’s position as leaves. (Perhaps mark the dataset above by the respective groupings).

- Step 3: Construct the graph with the appropriate stem and leaf components as shown below:

```
  0 | 3 8 9 9
  1 | 0 2 2 2 4 4 6 8 9
  2 | 1 2 2 5 8 9
  3 | 1 3 6 7
  4 | 1 9
  5 | 2 4 8
```
Boxplots

- A boxplot is another EDA graphic that displays the *five-number summary*

- The five number summary is composed of the minimum, Q1, Q2, Q3, and the maximum.

- Boxplots can be used to compare multiple datasets (See Figure 3-10, p. 149).

- Review the green box on p 148 regarding the information obtained from a boxplot.

**Example - Exercise 130, p. 151.**

- **Step 1:** Calculate the five-number summaries for USA and South America

<table>
<thead>
<tr>
<th>Statistic</th>
<th>USA</th>
<th>S.Am.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>50,000</td>
<td>46,563</td>
</tr>
<tr>
<td>Q1</td>
<td>57,642.5</td>
<td>56,242</td>
</tr>
<tr>
<td>Med</td>
<td>72,100</td>
<td>103,979</td>
</tr>
<tr>
<td>Q3</td>
<td>85,004</td>
<td>274,026</td>
</tr>
<tr>
<td>Max</td>
<td>125,628</td>
<td>311,539</td>
</tr>
</tbody>
</table>

- **Step 2:** Draw the box-plots side by side using a common scale