Section 2.4: Outliers, Boxplots, and Quantitative/Categorical Relationships

Objectives

1) Box plots
2) Using the IQR rule to detect outliers
   a. Lower and upper fences
3) Box plots and shapes of distributions
4) Back to back box plots
   a. One categorical and one quantitative variable
   b. Comparing box plots
**Section 2.4: Outliers, Boxplots, and Quantitative/Categorical Relationships**

Read section 2.4, page 96 to complete the following.

### Detection of Outliers – Using the IQR rule

- What formulas are used to detect outliers?

- We call a data value an outlier if it is

  Smaller than ________________________________ or larger than ____________________________.

- **What is a Box Plot**
  It’s a graphical display of the ________________________________

1) Use the data: “Distance between your home and MC” in miles.
   3, 13, 1, 7, 4, 1, 8, 13, 5, 3, 5, 25, 14, 16, 4, 5, 7, 14, 6, 8

   a) Sketch a box plot “by hand”
      - Write the numbers in ascending order.
      - Find the median.
      - Find Q1 = median of the lower 50% of the data
      - Find Q3 = median of the upper 50% of the data
      - Use the five number summary to construct a box plot.

   b) The five number summary is ________________________________

   c) Show also the dotplot. You can now appreciate that in each “portion” we have 25% of the data.

   d) Is 25 an outlier? Show calculations.

   e) Use the calculator to calculate the summary statistics using 1-Var Stats
      To see the graph: Press 2nd Y= ; Select the box plot with outliers; Press ZOOM and select 9:ZOOMStat
2) **Arsenic in Toenails**

Toenail clippings are used to measure the level of arsenic exposure of individuals in Great Britain. A similar study was conducted in the US. Table 2.23 gives toenail arsenic concentrations (in ppm) for 19 individuals with private wells in New Hampshire, and the data are also available in ToenailArsenic. Such concentrations prove to be an effective indicator of ingestion of arsenic-containing water.

**Table 2.23**

<table>
<thead>
<tr>
<th>Arsenic concentration in toenail clippings</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.119 0.118 0.099 0.118 0.275 0.358 0.080</td>
</tr>
<tr>
<td>0.158 0.310 0.105 0.073 0.832 0.517 0.851</td>
</tr>
<tr>
<td>0.269 0.433 0.141 0.135 0.175</td>
</tr>
</tbody>
</table>

**Figure 2.29 - Dotplot of arsenic concentration in toenails**

a) Use technology to find the mean, standard deviation and the five-number-summary.
b) Find the range and the IQR.
c) Use the formula to identify outliers.
d) Construct a box plot for the data. (show it above the dot plot)
e) Compute the z-score for the largest concentration and interpret it. Is it unusual?
f) What measures are more appropriate to describe the data, the mean and standard deviation or the five number summary? Why?
3) **Gross State Product**

The boxplot shows the Gross State Product (GSP) per capita, in dollars per resident, for the 50 US states.

a). Are there any outliers? Is so, approximately what are the values of the outliers?

b). Estimate the range of the data – show calculations:

c). Estimate the IQR for the data – show calculations.

d). Estimate the median of the data.

e). Does the data appear to be symmetric, skewed to the left, skewed to the right, or none of these?

f). Do you expect the mean to be greater than or less than the median? Why?

4) **Population of US States**

The five number summary for the populations of the 50 US states, in millions of people, is

\(0.506, 1.660, 4.170, 6.676, 35.842\).

The table shows all 50 populations.

<table>
<thead>
<tr>
<th>0.506</th>
<th>0.621</th>
<th>0.636</th>
<th>0.658</th>
<th>0.771</th>
<th>0.830</th>
<th>0.927</th>
<th>1.080</th>
<th>1.262</th>
<th>1.299</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.315</td>
<td>1.395</td>
<td>1.748</td>
<td>1.813</td>
<td>1.903</td>
<td>2.333</td>
<td>2.421</td>
<td>2.734</td>
<td>2.750</td>
<td>2.901</td>
</tr>
</tbody>
</table>

a. Find the IQR

b. Find the lower fence: \(Q_1 - 1.5 \times IQR\)

c. Find the upper fence: \(Q_3 + 1.5 \times IQR\)

d. List the outliers, if any.

e. Construct a boxplot.
Section 2.4 – Back to Back Box Plots
One Quantitative and One Categorical variable

Read this section of the book and complete the following:

To visualize the relationship between a quantitative variable and a categorical variable we use _______________________

5) Read example 2.29 based on the StudentSurvey dataset in the book and answer the questions given below the graphs.

![Boxplots and Dotplots](image.png)

**Figure 2.34 - Who watches more TV, males or females?**

*a) Who are the individuals? (Cases?)*

*b) List the variables and classify them as categorical or quantitative.*

*c) Use the graphs, compare the data and write five statements that describe the data.*
Section 2.4 – Back to Back Box Plots
One Quantitative and One Categorical variable

6) **StudentSurvey** dataset – continued. Consider the descriptive statistics output for the same data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StdDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>F</td>
<td>169</td>
<td>5.237</td>
<td>0.315</td>
<td>4.100</td>
<td>0.000</td>
<td>2.500</td>
<td>4.000</td>
<td>6.000</td>
<td>20.000</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>192</td>
<td>7.620</td>
<td>0.464</td>
<td>6.427</td>
<td>0.000</td>
<td>3.000</td>
<td>5.000</td>
<td>10.000</td>
<td>40.000</td>
</tr>
</tbody>
</table>

Figure 2.36
Output from Minitab comparing TV watching by gender

a) Compare the mean and standard deviation for number of hours spent watching television per week broken down by gender. Find the difference in means, using appropriate notation, and interpret it in terms of television viewing habits.

b) Determine any outliers for each of the data sets. The back to back box plots are on the prior page

7) **Example 3: Smokers by Region of the Country in the US**
The side-by-side boxplot shows the percent of adult residents who smoke in each state, categorized by region of the country (Midwest, Northeast, South, or West).

![Boxplot](image)

a). In general, which region seems to have the highest percent of smokers?

b). In which region is the state with the highest percent of smokers?

c). Which region has the greatest variability?

d). Do any of the regions have any outliers? If, so, which ones?
8) **Quantitative/Categorical Relationships**

In a recent study, participants were randomized to drink either tea or coffee every day for two weeks. After two weeks, blood samples were exposed to an antigen and an immune system response was measured, with higher values representing a stronger immune system.

![Boxplot](image)

a). Based on the boxplots, does there appear to be a relationship between the categorical variable (tea or coffee) and the quantitative variable (immune response)? Explain.

b). Which group appears to have the stronger immune response?

d). Determine if there are any outliers.

e). Can we conclude that tea *causes* an increase in this aspect of the immune response? Why or why not?

---

**Variable**     | **Drink** | **N** | **Mean** | **StDev** | **Min** | **Q1** | **Median** | **Q3** | **Max**
---|---|---|---|---|---|---|---|---|---
InterferonGamma | Coffee   | 10   | 17.70    | 16.69     | 0.00    | 2.25   | 15.50    | 25.25  | 52.00
              | Tea      | 11   | 34.82    | 21.08     | 5.00    | 13.00  | 47.00    | 55.00  | 58.00