Dealing with Data

Students will:

- Form questions answered by data and collect data from a sample and from a population to answer the questions
- Display data in appropriate graphs, including box plots and scatter plots
- Analyze data using measures of center and measures of variation
- Describe how a change in one variable affects another variable
- Discuss how sample size affects sample statistics
- Compare data sets using graphs and statistics

Classroom Cases:

1. The following list shows the ages of cars (in years) in the faculty/staff parking lot: 4, 9, 9, 8, 7, 1, 5, 4, 4, 4, 7, 6, 7, 23. Find the 5-number summary and construct a box plot. Identify at least one interesting characteristic.

   **Case Closed - Evidence:**

<table>
<thead>
<tr>
<th>5-number summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
</tr>
<tr>
<td>Lower quartile</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Upper quartile</td>
</tr>
<tr>
<td>Maximum</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

   - Minimum: 1
   - Lower quartile: 4
   - Median: 6.5
   - Upper quartile: 8
   - Maximum: 23

   75% of the cars are 8 years or less in age. One car, probably a classic, is an outlier at 23 years old.

2. Who drinks more soft drinks: 7th graders or their parents? Conduct a survey to gather data to answer this question, display the data appropriately, and analyze it to answer the question.

   **Case Closed - Evidence:**

   I interviewed 24 7th graders and one of their parents. I found that 7th graders drank more soft drinks a day than their parents did. The mean number of sodas for 7th graders was 2 compared to 1.46 sodas per day for the parents. As the box plots at right show, the highest number of sodas for students was 5 but the highest for parents was only 4. The lowest for both groups was 0, but 29% of parents responded, “0” (that is, they drank no soft drinks). Also for parents, the median was only 1, meaning that 50% of my sample drank 1 or fewer soft drinks a day. For students, the median was 2, so 50% of the students I surveyed drank 2 or fewer sodas a day and 50% drank 2 or more.

   **Parents’ Responses**

   In the graph at left, I plotted parents’ responses on the horizontal axis and students’ responses on the vertical axis because I wanted to see if the number of sodas students drank depended on the number their parents drank. Since the points sort of make a line, I think that the more sodas parents drink, the more sodas their children drink. There seems to be a positive association.

3. Abby, Ben, and Carlos conducted the soda survey described above. Abby asked 12 people on the bus on the way home. Ben surveyed 35 people at the mall. And Carlos called 65 people on the telephone. Discuss how different sample sizes and survey techniques could affect their results.

   **Case Closed - Evidence:**

   Abby only surveyed 12 people. Her sample is small and it comes from a limited group. It may not accurately reflect the true characteristics of the population. Ben surveyed more people, but only those who could find rides to the mall. Ben’s sample may not represent people who shop online. Carlos surveyed a fairly large sample and should therefore be able to avoid an impact from unusual behaviors. Most people in the U.S. have telephones, so his sample would allow all members of the population to be represented. If he selected phone numbers randomly, there should be no sampling bias. I think Carlos’ sample will be an accurate representation of the population of 7th graders and their parents.

Further investigations:

With your student, find graphs and summary statistics in newspapers and magazines. Discuss the messages delivered by these displays and analyses of data.

Ask your student to record for 10 days the amount of minutes he spends each day on two non-school activities (such as watching TV and listening to music). Together, make a double box plot and compare how your student is spending his leisure time.

Use the data collected on leisure time activities to make a scatter plot that will answer the question, “Does time spent on activity 1 depend on time spent on activity 2?”

Terminology:

- **Box plot:** A diagram that summarizes data using five numbers: minimum, lower quartile, median, upper quartile, and maximum.
- **Census:** Collection of data from every member of a population.
- **Five number summary:** Minimum, lower quartile, median, upper quartile, and maximum of a data set.
- **Interquartile range:** The difference between the first and third quartiles.
- **Measures of center:** Numerical values used to describe the clustering of data in a set. Mean, median, and mode are common measures of center.
- **Measures of variation:** Numerical values used to describe the spread or dispersion of data in a set. Range and interquartile range are common measures.
- **Outlier:** A value that is very far away from most of the values in a data set.
- **Parameter:** A measured characteristic of a population.
- **Quartiles:** Numbers that divide data into quarters when data in a set are in order.
- **Sample:** A selected part of a population.
- **Scatter plot:** A graph of a set of ordered pairs.
- **Statistic:** A measured characteristic of a sample.

Related Files:

www.ceismc.gatech.edu/csi