3. Independent random samples of 500 households were taken from a large metropolitan area in the United States for the years 1950 and 2000. Histograms of household size (number of people in a household) for the years are shown below.

(a) Compare the distributions of household size in the metropolitan area for the years 1950 and 2000.

(b) A researcher wants to use these data to construct a confidence interval to estimate the change in mean household size in the metropolitan area from the year 1950 to the year 2000. State the conditions for using a two-sample $t$-procedure, and explain whether the conditions for inference are met.
Solution

Part (a):

Household size tended to be larger in 1950 than in 2000. The histograms reveal a much larger proportion of small (1-, 2-, and 3-person) households in 2000 than in 1950. Similarly, the histograms reveal a much smaller proportion of large (5-person and larger) households in 2000 than in 1950. Also, the median household sizes can be calculated to be 5 people per household in 1950 compared with 3 or 4 people per household in 2000. The year 1950 displayed slightly more variability in household sizes than the year 2000. Although the interquartile ranges for both years are the same (3 people), the standard deviation (1950: about 2.6 people; 2000: about 3.1 people) and the range (1950: 13 people; 2000: 11 people) are larger for 1950 than for 2000. Both distributions of household size are skewed to the right. In both years, there are a few households with very large families, as large as 14 people in 1950 and 12 people in 2000.

Part (b):

The conditions for applying a two-sample $t$-procedure are:

1. The data come from independent random samples or from random assignment to two groups;
2. The populations are normally distributed, or both sample sizes are large;
3. The population sizes are at least 10 (or 20) times the sample sizes.

The first condition is satisfied because independent random samples were selected for the years 1950 and 2000. The second condition is satisfied because the sample sizes (500 in each group) are quite large, despite the right skewness of the distributions of household sizes in the sample data. The third condition is satisfied because the number of households in the large metropolitan area in both 1950 and 2000 would easily exceed $10 \times 500 = 5,000$. 

1. Agricultural experts are trying to develop a bird deterrent to reduce costly damage to crops in the United States. An experiment is to be conducted using garlic oil to study its effectiveness as a nontoxic, environmentally safe bird repellant. The experiment will use European starlings, a bird species that causes considerable damage annually to the corn crop in the United States. Food granules made from corn are to be infused with garlic oil in each of five concentrations of garlic — 0 percent, 2 percent, 10 percent, 25 percent, and 50 percent. The researchers will determine the adverse reaction of the birds to the repellant by measuring the number of food granules consumed during a two-hour period following overnight food deprivation. There are forty birds available for the experiment, and the researchers will use eight birds for each concentration of garlic. Each bird will be kept in a separate cage and provided with the same number of food granules.

(a) For the experiment, identify
i. the treatments
ii. the experimental units
iii. the response that will be measured

(b) After performing the experiment, the researchers recorded the data shown in the table below.

<table>
<thead>
<tr>
<th>Garlic oil concentration</th>
<th>0%</th>
<th>2%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean number of food granules consumed</td>
<td>58</td>
<td>48</td>
<td>29</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>Number of birds</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

i. Construct a graph of the data that could be used to investigate the appropriateness of a linear regression model for analyzing the results of the experiment.

ii. Based on your graph, do you think a linear regression model is appropriate? Explain.
Solution

Part (a):

i. The treatments are the different concentrations of garlic in the food granules. Specifically, there are five treatments: 0 percent, 2 percent, 10 percent, 25 percent and 50 percent.
ii. The experimental units are the birds (starlings), each placed in an individual cage.
iii. The response is the number of food granules consumed by the bird.

Part (b):

i. The following scatterplot results from these data.

![ Scatterplot Image ]

ii. The curved pattern in this scatterplot reveals that a linear regression model would not be appropriate for modeling the relationship between these variables.
1. As a part of the United States Department of Agriculture’s Super Dump cleanup efforts in the early 1990s, various sites in the country were targeted for cleanup. Three of the targeted sites—River X, River Y, and River Z—had become contaminated with pesticides because they were located near abandoned pesticide dump sites. Measurements of the concentration of aldrin (a commonly used pesticide) were taken at twenty randomly selected locations in each river near the dump sites.

The boxplots shown below display the five-number summaries for the concentrations, in parts per million (ppm) of aldrin, for the twenty locations that were sampled in each of the three rivers.

(a) Compare the distributions of the concentration of aldrin among the three rivers.

(b) The twenty concentrations of aldrin for River X are given below.

3.4 4.0 5.6 3.7 8.0 5.5 5.3 4.2 4.3 7.3
8.6 5.1 8.7 4.6 7.5 5.3 8.2 4.7 4.8 4.6

Construct a stemplot that displays the concentrations of aldrin for River X.

(c) Describe a characteristic of the distribution of aldrin concentrations in River X that can be seen in the stemplot but cannot be seen in the boxplot.
Solution

Part (a):

Comparing the medians reveals that the concentration of aldrin tends to be highest for River X and lowest for River Z. About 50 percent of the concentrations of aldrin for Rivers X and Y are higher than all of the concentrations for River Z. River X also displays the most variability in aldrin concentrations, as seen by the largest range and largest IQR, and River Z has the least variability, as judged by both IQR and range. The shapes of the three distributions differ, in that the distribution appears to be skewed to the right for River X, roughly symmetric for River Y and slightly skewed to the left for River Z.

Part (b):

Aldrin concentrations (in ppm) for River X
Leaf unit = 0.1 (for example, 3 | 4 represents 3.4 ppm)

3  47
4  023678
5  13356
6
7  35
8  0267

Part (c):

The stemplot shows a gap in the distribution of aldrin concentrations for River X between the values of 5.6 and 7.3 ppm of aldrin. This gap is not apparent in the boxplot.
1. As gasoline prices have increased in recent years, many drivers have expressed concern about the taxes they pay on gasoline for their cars. In the United States, gasoline taxes are imposed by both the federal government and by individual states. The boxplot below shows the distribution of the state gasoline taxes, in cents per gallon, for all 50 states on January 1, 2006.

(a) Based on the boxplot, what are the approximate values of the median and the interquartile range of the distribution of state gasoline taxes, in cents per gallon? Mark and label the boxplot to indicate how you found the approximated values.

(b) The federal tax imposed on gasoline was 18.4 cents per gallon at the time the state taxes were in effect. The federal gasoline tax was added to the state gasoline tax for each state to create a new distribution of combined gasoline taxes. What are approximate values, in cents per gallon, of the median and interquartile range of the new distribution of combined gasoline taxes? Justify your answer.
Solution

Part (a):

The median and quartiles are marked and labeled on the boxplot above. The median is approximately 21 cents per gallon.

The first and third quartiles are approximately 18 cents per gallon and 25 cents per gallon, respectively. The IQR is Q3 − Q1, which is approximately 25 − 18 = 7 cents per gallon.

Part (b):

After adding 18.4 cents per gallon to each of the state taxes, the median of the combined gasoline taxes would be the median of the state tax plus the federal tax, which is approximately 21 + 18.4 = 39.4 cents per gallon.

Although the quartiles of the combined gasoline taxes will change (Q1 = 18 + 18.4 = 36.4 cents per gallon and Q3 = 25 + 18.4 = 43.4 cents per gallon), the IQR will remain the same as it was for the state taxes at 7 cents per gallon (43.4 − 36.4 = 7).