1. Suppose that the observational units in a study are the 21 purchases that I made on amazon.com in the year 2011. Identify each of the following as a categorical variable, a quantitative variable, or not a variable. (Circle your answer for each; do not bother to explain.)

   a) How much did I spend on the purchase?  
      Categorical  Quantitative  Not a variable
   b) Was the purchase shipped to me or to someone else?  
      Categorical  Quantitative  Not a variable
   c) Do I tend to spend more on purchases sent to others than on purchases sent to me?  
      Categorical  Quantitative  Not a variable
   d) Did the purchase include free shipping?  
      Categorical  Quantitative  Not a variable
   e) What was the average price of these purchases?  
      Categorical  Quantitative  Not a variable

2. A few years ago, I asked students in my classes whether they drink coffee every day, sometimes, or (almost) never. I also asked each student to report his/her gender. Results are summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Every day</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>Sometimes</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>(Almost) never</td>
<td>19</td>
<td>7</td>
</tr>
</tbody>
</table>

   a) What kind of graphical display would be appropriate to display these data? (Circle your choice. Do not bother to explain. Do not bother to create the graph.)
      
      histogram  boxplot  segmented bar graph
   b) What proportion of the female students drink coffee every day?
      
      \[
      \frac{15}{32} \approx 0.469
      \]
   c) What proportion of the students who drink coffee every day are female?
      
      \[
      \frac{15}{22} \approx 0.682
      \]

3. The news website CNN.com regularly posts a poll question that people who view the website can respond to. The following results were posted on January 10, 2012:
Do you suspect that the sampling method is biased? If so, identify which direction, and explain why you think this. If not, explain why not.

This sampling method is very likely to be biased in favor of people who surf the web often while at work, because a respondent had to be surfing the web in the first place in order to see the poll.

4. The National Retail Federation conducted a national survey of 8526 consumers on September 1-9, 2009. Among the findings reported were that:
   - 29.6% of those surveyed said that the state of the U.S. economy would affect their Halloween spending plans;
   - the average amount that the respondents said they expect to spend on Halloween is $56.31.

a) Identify the observational units in this study.

The observational units are the 8526 consumers who were surveyed.

b) Identify a quantitative variable in this study.

A quantitative variable is the amount that the person expects to spend on Halloween.

c) Are the numbers 29.6% and $56.31 parameters or statistics? (Circle your answer; do not bother to explain.)

   parameters statistics

d) The information above does not say whether randomness was used in this study. Which kind of randomness would have been appropriate? (Circle your answer; do not bother to explain.)

   random sampling random assignment

5. Create a hypothetical example of 10 exam scores (integers between 0 and 100, inclusive, with repeats allowed) with the property that only one of the exam scores is greater than the mean.

One example that works is: 80, 80, 80, 80, 80, 80, 80, 80, 80, 100

6. In a study published in a 2007 issue of the journal *Preventive Medicine*, researchers found that smokers were more likely to have used candy cigarettes as children than non-smokers were.
a) What type of study is this? (Circle your answer; do not bother to explain.)

**observational**  
**experimental**

b) When hearing about this study, John responded: “But isn’t the smoking status of the person’s parents a confounding variable here?” When Karen asked what he meant, John said: “Children whose parents smoke are more likely to become smokers themselves when they become adults.” What else does John need to say in order to explain how the parents’ smoking status can be a confounding variable in this study?

In order for the parents’ smoking status to be a confounding variable, it must be the case that children of parents who smoke are more likely to use candy cigarettes than children whose parents do not smoke.

7. Consider the following four distributions of quiz scores:

Arrange these in order from smallest standard deviation to largest standard deviation. (Just write in the letters A, B, C, D. Do not bother to calculate any standard deviations. Do not bother to explain.)

Smallest: B  
Next-to-smallest: C  
Next-to-largest: D  
Largest: A

8. The book *Day Hikes in San Luis Obispo County* by Robert Stone gives information on 72 different hikes that one can take in the county. The following stemplot displays the distribution of the lengths (in miles) of these hikes (for instance, the shortest hike is 0.6 miles and the longest is 9.5 miles):

```
0 | 68
1 | 00000555555588
2 | 00000001225555556668
3 | 000002224458
4 | 00000566
5 | 0055668
6 | 0000
7 | 004
8 |
9 | 5
```

a) Determine the five-number summary of this distribution. Also indicate how you calculate these values.
The minimum hike length is 0.6 miles. The maximum hike length is 9.5 miles.

Because \((72 + 1)/2 = 36.5\), the median is the average of the 36th and 37th values, which is \((2.8 + 3.0) / 2 = 2.9\) miles.

Because \((36 + 1)/2 = 18.5\), the quartiles are the average of the 18th and 19th values. The lower quartile is \((2.0 + 2.0) / 2 = 2.0\) miles, and the upper quartile is \((4.5 + 4.6) / 2 = 4.55\) miles.

b) Is the 9.5-mile hike an outlier according to the 1.5IQR rule? Justify your answer with appropriate calculations.

\[ \text{IQR} = (4.55 - 2.0) = 2.55 \text{ miles.}\]

\[ 1.5 \times \text{IQR} = 1.5 \times 2.55 = 3.825. \]

The mileage needed to be an outlier on the high end is \(4.55 + 3.825 = 8.375\), so the 9.5 mile hike is an outlier.

9. In a recent study, researchers purchased 40 food items in New York City and determined the actual calorie content of each through a laboratory analysis. They then calculated the percentage difference between the actual calorie content and the calorie count listed on the item’s label. (A positive percentage difference corresponds to a food item whose actual calorie content was higher than what the label claimed.) Each food item was also classified according to whether it was marketed locally, nationally, or regionally. The boxplots below were constructed to compare the distributions:

Write a paragraph summarizing what these boxplots reveal about the percentage differences between actual and advertised calorie content among the three marketing groups of food items.

The locally marketed items have the highest discrepancy between actual and advertised calorie counts. All of these items had a higher actual than advertised calorie count, and the median is about a 70% discrepancy. Moreover, there is large variability in these discrepancy amounts, from about 5% to about 250%. The regionally marketed items also tended to underestimate their calorie amounts, but much less so and with much less variability. The median for the regional items is only about a 30% difference. The nationally marketed items list the calorie amounts very close to the actual amounts; the median is about 0 and there is very little variability. The national group even has a low outlier at about -20%, indicating that one item had about 20% fewer calories than listed. The distribution of percentage differences is skewed to the right for the local items and fairly symmetric for the regional and national items.