Solutions 1

(1) What sort of variable is

(i) The type of cycle a person owns. **Categorical**
(ii) The height of a person. **(Continuous) Numerical**
(iii) The gender of a person. **Binary**
(iv) The number of a bus is **Categorical** (since the numbering has no numerical meaning or ordering).

(2) What is the mean and standard deviation of the following samples:

(a) 1, 1, 1, 1, 1.
   Mean is 1 and standard deviation is 0
(b) -99, -49, 1, 51, 101.
   Mean is 1 but the standard deviation is 79

(3) Sidney observes the data

\[ 1, 1.5, 2, 3, 3.5, 3.8, 6, 6.5, 7, 8. \]

The sample mean and standard deviation of this sample is 4.23 and 2.47 respectively.

What happens to the mean, standard deviation and quartiles of the sample if the last observation 8 is replaced with a much larger number?

(A) The mean and standard deviation stay about the same, but the quartiles and median change.
(B) Only the mean will change.
(C) The first quartile and median stay the same, but the third quartile, mean and standard deviation become larger.
(D) The quartiles and median stay the same, but the mean and most likely the standard deviation will become much larger.
(E) Unless we know the actual new number it is impossible to say what will happen to the values.

The answer is (D). This is an outlier. The quartiles stay about the same (either same if you use my definition or slightly different if you use the JMP quartiles) but the mean will becomes larger. Since the variation in the sample has increased the standard deviation is also likely to be larger.

(4) Input the M&M data into JMP. We want to see whether there is any difference in the distributions of number of Peanut, Peanote Butter and Milk chocolate M&Ms (later on we will use statistical tools to see if these difference are ‘real”).
(i) Make boxplots and histograms of the total number M&Ms for each type (Peanut, Peanut butter and Milk chocolate).

*Go to Analyze, Distribution and put Total into Y. column, then put Type into By.*

The individual means and standard deviations are

<table>
<thead>
<tr>
<th>Type</th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peanut</td>
<td>8.67</td>
<td>3.13</td>
</tr>
<tr>
<td>Peanut Butter</td>
<td>10.91</td>
<td>1.82</td>
</tr>
<tr>
<td>Milk Chocolate</td>
<td>17.29</td>
<td>2.87</td>
</tr>
</tbody>
</table>

(ii) Compare their distributions (through histograms and boxplots). Are the distributions roughly the same (in the sense of mean, median, IQR (the difference between the third and first quartile) and standard deviations)?

![Histograms](image)

Figure 1: A histograms of Milk chocolate, Peanut Butter and Peanut M&Ms.

We can see from Figure 2 that the sample means for all three type of chocolates are different. There also seems to be more spread for the Peanut flavour (as seen by the standard deviation and also the boxplot). An interest question that cannot be answered by just looking at the numbers and the histogram is whether the difference we see between the peanut and peanut butter M&Ms (compare mean of 8.67 with 10.91) is this just a ‘sampling’ difference or is such a difference see over all M&M packets? To answer this question we need to know the variability in the M&M bags (roughly 1.82 and 3.13) and the sample sizes (40 and 46). These numbers will help us to see whether it is easy to explain these difference by just sample/random variation, from which we will draw statistical conclusions on the differences.

(5) As we go through the course, one of the fundemental results will be that the sample mean (average based on the sample) is random with a distribution. This distribution is special, it’s spread (standard deviation) is less than the sample itself (original observations) and its shape is quite unique. The only problem for most real life problems we won’t be able to ‘see’ the distribution of the sample mean, since we have only one sample mean (the histogram of one number is not very informative). Therefore, I artificially make several small samples from a big sample (by grouping the 170 bags of M&Ms into bags of five).

(i) Calculate the standard deviation of the total number of M&Ms (you can do this in JMP).

**The mean is 13.54.**

(ii) In the fifth column each of the M&M bags have been put into one of 34 groups (there are 5 in each group). For each group of 5 calculate the average number of M&Ms in a bag.
The standard deviation is 4.65.

You can do this in JMP by Going to Tables → Summary selecting Group to go into Group (for Statistic you should choose mean) and highlighting Total. You should now get a new table containing the average in each group, it will have 34 rows. You will have to data sets, the original M&M data set and also the data sets containing all the means that you created.

(iii) Using the new table calculate the mean and standard deviation of the groups means (this column is denoted as Mean (Total)).

In the table below we give the means and standard deviations for the total number of M&Ms and also the group of size 5 and also size 10 (this one is additional and used for comparisons).

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (all the separate M&amp;M nags)</td>
<td>13.54</td>
<td>4.65</td>
</tr>
<tr>
<td>Group (averages based on samples of 5)</td>
<td>13.54</td>
<td>2.09</td>
</tr>
<tr>
<td>Group (averages based in samples of 10)</td>
<td>13.54</td>
<td>1.43</td>
</tr>
</tbody>
</table>

We observe that the means are the same for all the samples but the standard deviations are different.

(iv) Compare the standard deviation for the entire class and the standard deviation of the averages. Comment on their differences.

The standard deviation for the M&Ms which are grouped into samples of size five is 2.09. This can be predicted by the formula $2.09 = \frac{4.65}{\sqrt{5}}$. We will discuss why this is the case later in class. Indeed the standard deviation based on samples of size 10 (there are only 17 groups in this case, but this can be increased just by taking more groups of size 10) is 1.43 this can also be predicted by the formula $\frac{4.65}{\sqrt{10}}$.

(v) Make a histogram for the total number of M&Ms and also the averages. Compare the two shapes in the histograms, are they similar? Describe their shapes.

![Histograms](image)

Figure 2: A histogram of the total number of M&Ms in a bag, the average of five bags of M&Ms and the average of ten bags of M&Ms.

Besides the same centering for all three histograms and difference in spread, we also see something else rather interesting: the bi/tri-modality (two or three peaks) that is see in the original M&M data is being smoothened away in the
histograms of the averages. The distribution of the averages becomes more uni-modal (one peak), symmetric and may be (!!!) more bell shaped as we go from group sizes of 5 to 10. We are seeing the central limit theorem in action.