STAT301 Solutions 1

(1) What sort of variable (either numerical or categorical) is

(i) The type of cycle a person owns. 
   **Categorical**

(ii) The height of a person. 
   **Numerical**

(iii) The gender of a person. 
   **Categorical**

(iv) A questionnaire where the participant gives an answer between 1 and 5, where 5 is strongly agree and 1 is strongly disagree. The values inbetween are between strongly agree and disagree. 
   **Numerical**

(v) Five football teams are arbitrarily given the names Team 1, Team 2, Team 3, Team 4, Team 5. 
   The favourite football team of a student, a choice of Team 1, Team 2, Team 3, Team 4 or Team 5. 
   **There is no real meaning to the ordering of the teams, thus it is CATEGORICAL**

(2) In this question use the peanut and plain M&Ms together.
   Our aim is to see whether the distribution of blue and yellow M&Ms are the same or not.

(i) By using the **Graphics** and **Histogram** (choosing relative frequency rather than frequency in the drop menu) options make relative frequency histograms of the number of yellow and blue M&Ms in a bag (over the entire class, ignoring the difference between peanut and plain).
   **To make a fair comparison use the same binwidths for both plots**

(ii) By using the **Summary Statistics** option evaluate the mean, median, IQR and standard deviations for the number of blue and yellow M&Ms.
   **Table:**
### Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>median</th>
<th>standard deviation</th>
<th>First Quartile</th>
<th>Third Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>2.71</td>
<td>3</td>
<td>1.80</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Yellow</td>
<td>2.01</td>
<td>2</td>
<td>1.43</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

(iii) Based on the plots and summary statistics do you think the distribution of blue and yellow M&Ms are roughly the same?

From this sample it is extremely hard to say. Comparing both the plots and the summary statistics there are differences but it is difficult with any formal methods to say whether we can explain away these differences by variation between the samples. Or if there is systematic difference between the distribution of blue and yellow M&Ms. This example illustrates the need for us to develop a more formal method which ‘tests’ whether it is possible these differences can easily be explained by sample variation.

(3) In this question we want to investigate whether there is an relationship between the number of blue and yellow M&Ms.

(i) By reasoning that the number of plain M&Ms in a bag and the number of Peanut M&Ms in a bag is about 7 and 17 respectively. Do you think that there may be a relationship between the number of blue and yellow M&Ms in a bag?

Possibly: on average there are 17 M&Ms in a milk chocolate bag, this suggests that a bag of M&Ms with a large number of blues will have to be compensated with a smaller number of yellows. However, there are also other M&M colours in the bag and the number of blue and yellows tend to be a lot less than the total number of M&Ms. So any relationship that may exist would be very weak.

(ii) We check the above assertion by making some plots (later we can do a statistical test to see whether there is any evidence from the data that there is a relationship).

By using the Graphics and Scatterplot options (does not matter which you choose as the x or y variable). In the Group by option choose the variable type, this will allow us to distinguish between the Peanut and Plain M&Ms in the scatterplot by the different colours.

From the plots, does there appear to be a relationship?

From the plots there does not appear to be an obvious relationship. However, it is possible the sample is too small to see a very ‘weak’ relationship.
This question will help prepare ourselves in understanding that a sample mean is random, has a distribution too and the spread of the sample mean is less than the original population.

We will focus on the total number of M&Ms in this question.

(i) By using the **summary statistics** option calculate the standard deviation for total number of M&Ms.

The sample mean is 13.54 standard deviation is 4.65 (last time when I did the same example the sample mean and standard deviation for that sample was 12.12 and 5.26. This illustrates that from sample to sample the sample mean and standard deviation vary randomly)

(ii) You will have noticed that I put each M&M in one of 34 groups. There are roughly 5 bags in each of the groups.

By using the **Summary Statistics, Columns** putting *total* into the Select column option and *Group* in the Group by option you should be able to calculate the mean number of M&Ms for each of the 34 groups.

Make a Table with the Group number and the mean number of M&Ms for that group.

**The data can be found on my website**

(ii) Type these 34 mean number of M&Ms as one column in Statcrunch (if you don’t want to do this, you can do following calculation by hand). Call this new column TotalGroup.

(iii) Either by using the **Summary Statistics** option in Statcrunch or by hand evaluate the mean and standard deviation of TotalGroup (the average and standard deviation based on 34 groups).

**The mean for Total group is 13.54 the standard deviation is 2.09**
(iv) Compare the standard deviation for the total number of M&Ms (this number is given in part 4(i)) with the standard deviation over the groups (standard deviation of Total-Group this number is given in 4(iii)). What is the main difference that you see between these standard deviations?

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>13.54</td>
<td>4.65</td>
</tr>
<tr>
<td>TotalGroup</td>
<td>13.54</td>
<td>2.09</td>
</tr>
</tbody>
</table>

We observe that the means for both sets of data are exactly the same (so both data sets have the same center), however, the spreads are completely different. The spread for Total (individual bags of M&Ms) is a lot more (4.65) than the spread for the group averages (which is 2.09). This difference in spreads CANNOT be explained by the sample size of Group being 34 and the sample size of Total being 170. The reason for this difference is that data set of Group are the averages of five M&M bags. Remember an average tends to smooth out extremes numbers and has a lot less spread. For example one sample of 5 from the data was 6, 12, 13, 17, 18. The average of this sample is 13.2. See how the average has, averaged out the extreme values of 6 and 18.

We will show in a future class that the standard deviation of the groups averages can be predicted by the formula \[ \frac{4.65}{\sqrt{5}} = 2.07. \]

(v) Make relative frequency histograms for both the total number of M&Ms and the means in each group (ie. histogram of TotalGroup).

Is there are a difference between the spreads of these two variables?

Figure 2: For Queston 4(iv) and 4(v). A histogram of both the total number of M&Ms in a bag and the sample mean number in a bag over 34 groups (with 5 bags per group).

It is clear from the plots histogram that the spread is a lot less for the mean over the groups than the actual data itself. In addition, whereas the histogram of the total M&Ms has is bimodal (or possibly even tri-modal), the histogram of the groups means is closer to unimodal with a possibly a more bell-shaped type curve. In future class we will show that this close bell-shaped curve is due to the central limit theorem.