4.1: Ways to Summarize Data

In Chapter 3 we used data collected from 30 bags of M&Ms to explore different ways to visualize data. In this chapter we consider several ways to summarize data using the net weights of the same bags of M&Ms. Here is the raw data.

Table 4.1.1: Net Weights for 30 Bags of M&Ms.

49.287	48.870	51.250	48.692	48.777	46.405
49.693	49.391	48.196	47.326	50.974	50.081
47.841	48.377	47.004	50.037	48.599	48.625
48.395	51.730	50.405	47.305	49.477	48.027
48.212	51.682	50.802	49.055	46.577	48.317

Without completing any calculations, what conclusions can we make by just looking at this data? Here are a few:

- All net weights are greater than 46 g and less than 52 g.
- As we see in Figure 4.1.1, a box-and-whisker plot (overlaid with a stripchart) and a histogram suggest that the distribution of the net weights is reasonably symmetric.
- The absence of any points beyond the whiskers of the box-and-whisker plot suggests that there are no unusually large or unsually small net weights.

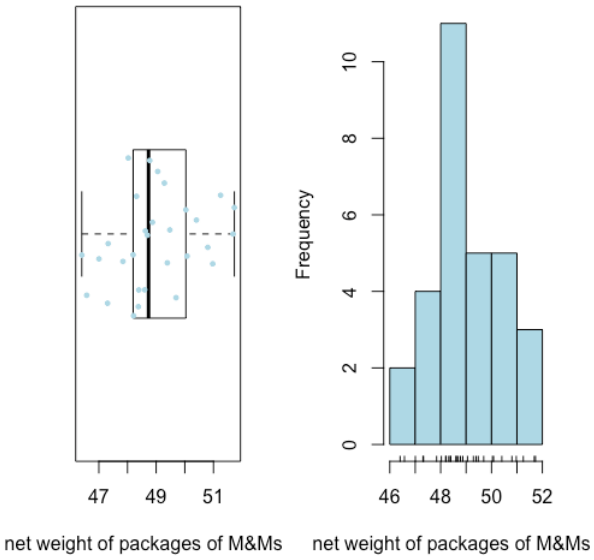

Figure 4.1.1: Two visualizations of the net weights of packages of M&Ms.

Both visualizations provide a good qualitative picture of the data, suggesting that the individual results are scattered around some central value with more results closer to that central value that at distance from it. Neither visualization, however, describes the data quantitatively. What we need is a convenient way to summarize the data by reporting where the data is centered and how varied the individual results are around that center.

Where is the Center?

There are two common ways to report the center of a data set: the mean and the median.

The mean, $\overline{Y}$, is the numerical average obtained by adding together the results for all *n* observations and dividing by the number of observations

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{49.287 + 48.870 + \cdots + 48.317}{30} = 48.980 \text{ g}$$

The median, $\widetilde{Y}$, is the middle value after we order our observations from smallest-to-largest, as we show here for our data.

Table 4.1.2: The data from Table 4.1.1 Sorted From Smallest-to-Largest in Value.

46.405	46.577	47.004	47.305	47.326	47.841
48.027	48.196	48.212	48.317	48.377	48.395
48.599	48.625	48.692	48.777	48.870	49.055
49.287	49.391	49.477	49.693	50.037	50.081
50.405	50.802	50.974	51.250	51.682	51.730

If we have an odd number of samples, then the median is simply the middle value, or

$$\widetilde{Y} = Y_{\frac{n+1}{2}}$$

where *n* is the number of samples. If, as is the case here, *n* is even, then

$$\widetilde{Y} = \frac{Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1}}{2} = \frac{48.692 + 48.777}{2} = 48.734 \text{ g}$$

When our data has a symmetrical distribution, as we believe is the case here, then the mean and the median will have similar values.

What is the Variation of the Data About the Center?

There are five common measures of the variation of data about its center: the variance, the standard deviation, the range, the interquartile range, and the median average difference.

The variance, s^2, is an average squared deviation of the individual observations relative to the mean

$$s^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}{n-1} = \frac{\left(49.287 - 48.980\right)^2 + \cdots + \left(48.317 - 48.980\right)^2}{30-1} = 2.052$$

and the standard deviation, *s*, is the square root of the variance, which gives it the same units as the mean.

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}{n-1}} = \sqrt{\frac{\left(49.287 - 48.980\right)^2 + \cdots + \left(48.317 - 48.980\right)^2}{30-1}} = 1.432$$

The range, *w*, is the difference between the largest and the smallest value in our data set.

$$w = 51.730 \text{ g} - 46.405 \text{ g} = 5.325 \text{ g}$$

The interquartile range, *IQR*, is the difference between the median of the bottom 25% of observations and the median of the top 25% of observations; that is, it provides a measure of the range of values that spans the middle 50% of observations. There is no single, standard formula for calculating the *IQR*, and different algorithms yield slightly different results. We will adopt the algorithm described here:

1. Divide the sorted data set in half; if there is an odd number of values, then remove the median for the complete data set. For our data, the lower half is

Table 4.1.3: The Lower Half of the Data in Table 4.1.2.

46.405	46.577	47.004	47.305	47.326
47.841	48.027	48.196	48.212	48.317
48.377	48.395	48.599	48.625	48.692

and the upper half is

Table 4.1.4: The Upper Half of the Data in Table 4.1.2.

48.777	48.870	49.055	49.287	49.391
49.477	49.693	50.037	50.081	50.405
50.802	50.974	51.250	51.682	51.730

2. Find F_L, the median for the lower half of the data, which for our data is 48.196 g.

3. Find F_U , the median for the upper half of the data, which for our data is 50.037 g.

4. The *IQR* is the difference between F_U and F_L.

$$F_U - F_L = 50.037 \text{ g} - 48.196 \text{ g} = 1.841 \text{ g}$$

The median absolute deviation, *MAD*, is the median of the absolute deviations of each observation from the median of all observations. To find the *MAD* for our set of 30 net weights, we first subtract the median from each sample in Table 4.1.1.

Table 4.1.5: The Results of Subtracting the Median From Each Value in Table 4.1.1.

0.5525	0.1355	2.5155	-0.0425	0.0425	-2.3295
0.9585	0.6565	-0.5385	-1.4085	2.2395	1.3465
-0.8935	-0.3575	-1.7305	1.3025	-0.1355	-0.1095
-0.3395	2.9955	1.6705	-1.4295	0.7425	-0.7075
-0.5225	2.9475	2.0675	0.3205	-2.1575	-0.4175

Next we take the absolute value of each difference and sort them from smallest-to-largest.

Table 4.1.6: The Data in Table 4.1.5 After Taking the Absolute Value.

0.0425	0.0425	0.1095	0.1355	0.1355	0.3205
0.3395	0.3575	0.4175	0.5225	0.5385	0.5525
0.6565	0.7075	0.7425	0.8935	0.9585	1.3025
1.3465	1.4085	1.4295	1.6705	1.7305	2.0675
2.1575	2.2395	2.3295	2.5155	2.9475	2.9955

Finally, we report the median for these sorted values as

$$\frac{0.7425 + 0.8935}{2} = 0.818$$

Robust vs. Non-Robust Measures of The Center and Variation About the Center

A good question to ask is why we might desire more than one way to report the center of our data and the variation in our data about the center. Suppose that the result for the last of our 30 samples was reported as 483.17 instead of 48.317. Whether this is an accidental shifting of the decimal point or a true result is not relevant to us here; what matters is its effect on what we report. Here is a summary of the effect of this one value on each of our ways of summarizing our data.

Table 4.1.7: Effect on Summary Statistics of Changing Last Value in Table 4.1.1 From 48.317 g to 483.17 g.

statistic	original data	new data
mean	48.980	63.475
median	48.734	48.824
variance	2.052	6285.938

statistic	original data	new data
standard deviation	1.433	79.280
range	5.325	436.765
IQR	1.841	1.885
MAD	0.818	0.926

Note that the mean, the variance, the standard deviation, and the range are very sensitive to the change in the last result, but the median, the *IQR*, and the *MAD* are not. The median, the *IQR*, and the *MAD* are considered robust statistics because they are less sensitive to an unusual result; the others are, of course, non-robust statistics. Both types of statistics have value to us, a point we will return to from time-to-time.
