Chapter 05: Understanding and Comparing Distributions

Boxplots

First up, as promised in the last chapter…a new graphic. Also called a "box and whisker" plot.

Standard Boxplot

First, find the five number summary for the data. Draw an axis horizontally (for the variable). Make small vertical marks at each value of the five number summary. Connect the lines for Q_1 and Q_3, making a divided rectangle. Now draw lines from the edges of the box to the remaining vertical lines. Here is an example:

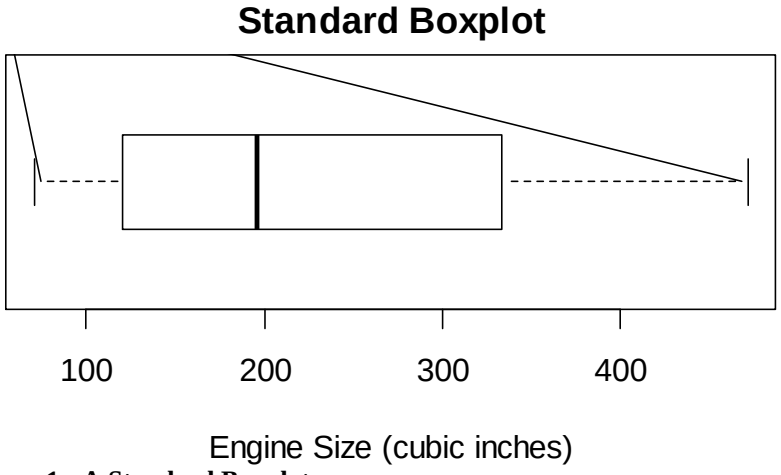

Figure 1 - A Standard Boxplot

Modified Boxplot

…but wait; there's more!

We need to talk for a moment about outliers (I promised that I would in the last chapter). Remember that an outlier is a datum that doesn't seem to fit with the others (one of these things is not like the others, one of these things just doesn't belong…). Deciding whether or not a datum is an outlier is another one of those things that probably bothers you at first since not everyone will agree…and, like the earlier topics, there is a numeric way to handle it!

Tukey's Rule of Thumb for Outliers: any datum lower than $Q_1 - 1.5(IQR)$, or higher than $Q_3 + 1.5(IQR)$ is possibly an outlier.

This rule is used to construct a modified boxplot (which is actually closer to the way that Tukey made them all the time…but whatever). Find the five number summary as before. Mark the quartiles and median, but not the extrema (minimum and maximum). Put a line at the lowest value that isn't an outlier. Also put a line at the highest non-outlier. Connect the lines to the box, then mark the outliers with separate symbols. Here is an example:

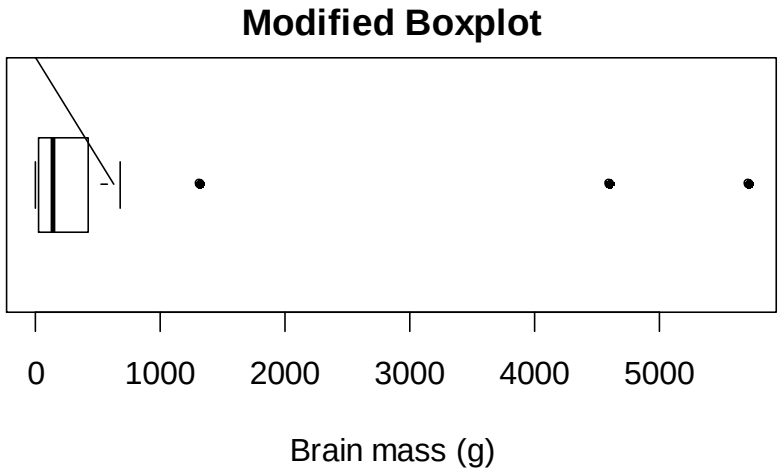

Figure 2 - A Modified Boxplot

By the way…your calculator will make both of these plots. Use it!

Examples

[1.] The masses of 11 chicks (grams) were recorded six weeks after hatching. The data are as follows:

Table 1 - Chicken Masses for Example 1

325	257	303	315	380	153	263	242	206	344	258

Construct a boxplot of these data.

Drop the data in your calculator and press the buttons! Here's what I get:

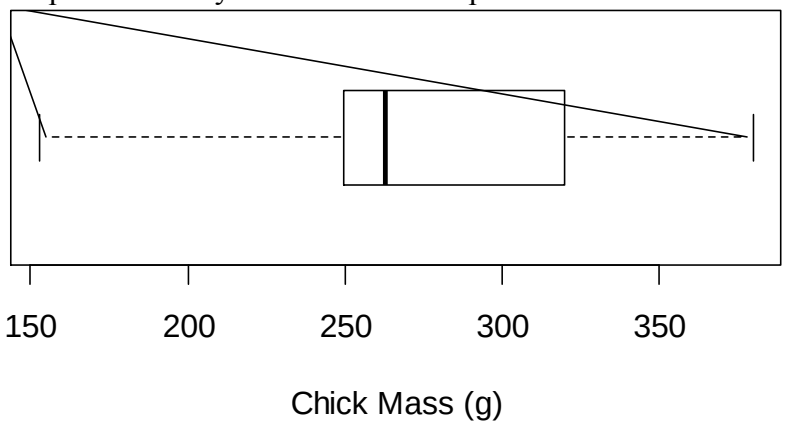

Figure 3 - Chicken Mass Boxplot for Example 1

[2.] The sepal width of 50 specimens of iris setosa were measured (cm). The data are as follows:

Table 2 - Sepal Widths for Example 2

3.5	3.0	3.2	3.1	3.6	3.9	3.4	3.4	2.9	3.1	3.7	3.4	3.0
3.8	3.4	3.7	3.6	3.3	3.4	3.0	3.4	3.5	3.4	3.2	3.1	3.0
3.0	3.4	3.5	2.3	3.2	3.5	3.8	3.0	3.8	3.2	3.7	3.3	
3.6	3.8	4.2	3.1	3.5	4.0	4.4	3.9	3.2	3.5	3.4	4.1	

Construct a boxplot of these data.

Once again…let the calculator do it for you. Here's what I got:

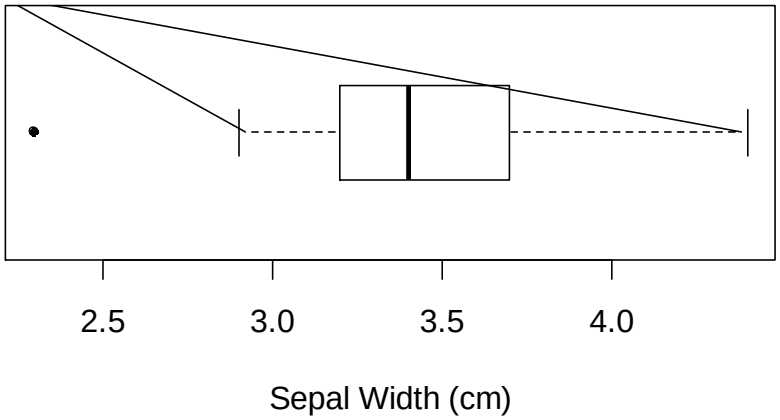

Figure 4 - Sepal Width Boxplot for Example 2

Comparing Distributions

When comparing two distributions, be sure to explicitly compare center and spread…that means that you should use a comparison word, like "larger" or "smaller." You should mention the shape of both distributions, but there isn't really any way to compare them.

When graphing the two distributions, using the same scale on both graphs is important. Your calculator does this most easily with boxplots. You'll have to do the work yourself if you want to graph two histograms…and if you're doing the work yourself, why not make a back-to-back stemplot (two stemplots that share a set of stems)?

Examples

[3.] 24 newly hatched chicks were randomly divided into two groups. Each group received a different type of feed for six weeks. Each chick was massed (grams) at the end of that time. Here are the masses for Feed Type 1:

Table 3 - Masses with Feed Type 1 for Example 3

309	229	181	141	260	203	148	169	213	257	244	271

Here are the masses for Feed Type 2:

Table 4 - Masses with Feed Type 2 for Example 3

368	390	379	260	404	318	352	359	216	222	283	332

Compare the effects of the two feed types.

Let's start with a graph:

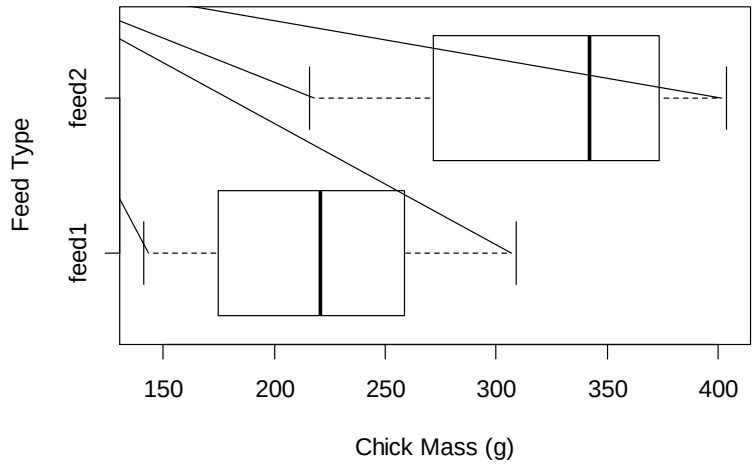

Figure 5 - Chicken Mass Boxplot for Example 3

Both distributions are close to symmetric—feed type 1 is closer to symmetric than feed type 2 (which might be ever so slightly skew left). The center of feed type 2 is much larger than that of feed type 1 (indicating that feed type 2 may cause more growth in chicks than feed type 1). The spread of masses is nearly equal in both distributions (IQR of approximately 100 grams).

[4.] 60 guinea pigs were randomly assigned to two groups. One group received a daily dose of Orange Juice, and the other group received a daily dose of ascorbic acid (Vitamin C). At the end of the experiment, the lengths of the teeth (mm) of the guinea pigs were measured.

Here are the data for the guinea pigs receiving Orange Juice:

Table 5 - Tooth Length with Orange Juice for Example 4

15.2	21.5	17.6	23.3	9.7	14.5	10.0	19.7
25.2	25.8	21.2	14.5	27.3	25.5	26.4	22.4
9.4	16.5	23.0	26.4	23.6	9.7	8.2	
30.9	26.4	27.3	29.4	20.0	24.8	24.5	

Here are the data for the guinea pigs receiving Vitamin C:

Table 6 - Tooth Length with Vitamin C for Example 4

4.2	11.5	5.8	7.3	5.2	7.0	16.5	6.4
7.3	13.6	14.5	18.8	15.5	23.6	18.5	33.9
11.2	16.5	15.2	17.3	22.5	10.0	11.2	
32.5	26.7	21.5	23.3	29.5	25.5	26.4	

Compare the growth of teeth for these two groups.

Again, start with a graph:

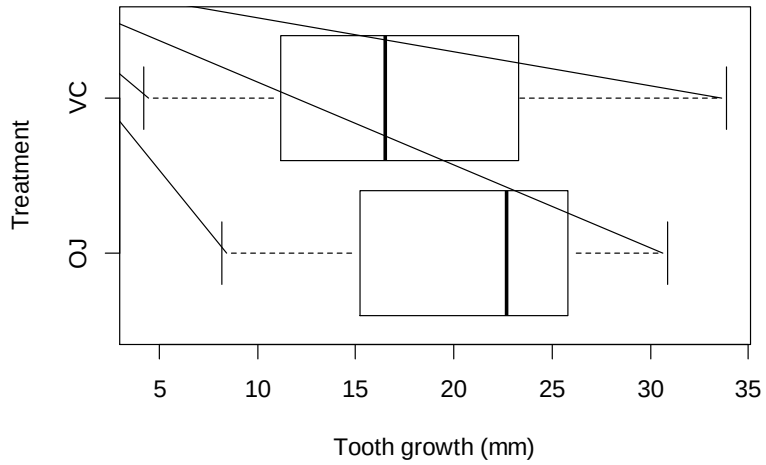

Figure 6 - Tooth Length Boxplot for Example 4

The distribution of tooth growth for those receiving Vitamin C appears symmetric; the graph for those receiving Orange Juice appears less symmetric (probably skew left). The center for the Orange Juice group is higher than the center for the Vitamin C group. The spread of tooth sizes for the Vitamin C group appears larger than that of the Orange Juice group.

It appears that Orange Juice provides larger tooth group with less variation than Vitamin C.

Re-expressing Data

If your data a strongly skew right—maybe even with a few high outliers—it becomes hard to read the graph and make useful statements. In this case, transforming the data can help. Of primary interest is taking the logarithm of the data…not that logarithms are the only way to transform data, but they are probably the most useful (especially in this course).

Example

[5.] Here are the masses (grams) of the brains of 28 animals.

Table 7 - Brain Masses for Example 5

8.1	423.0	70.0	119.5	115.0	50.0	4603.0	419.0	115.0	175.0
680.0	406.0	1320.0	5712.0	154.5	56.0	1.0	655.0	25.6	
157.0	440.0	1.9	179.0	5.5	3.0	180.0	0.4	12.1	

When graphing these data, the graph is very difficult to read and interpret because of the extreme skew and outliers.

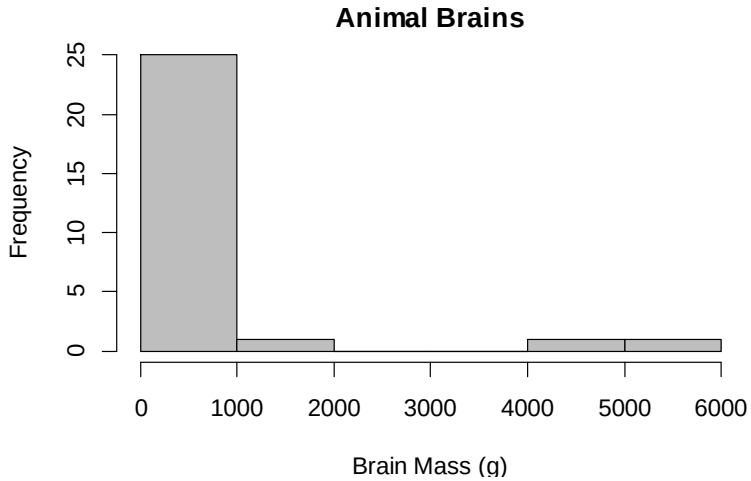

Figure 7 - Brain Mass Histogram for Example 5

This is a perfect case where transforming the data can be helpful. I'll take the natural logarithm of the data and re-graph.

Table 8 - Natural Log of Brain Masses for Example 5

2.092	6.047	4.248	4.783	4.745	3.912	8.434	6.038	4.745	5.165
6.522	6.006	7.185	8.650	5.040	4.025	0.000	6.485	3.243	
5.056	6.087	0.642	5.187	1.705	1.099	5.193	-0.916	2.493	

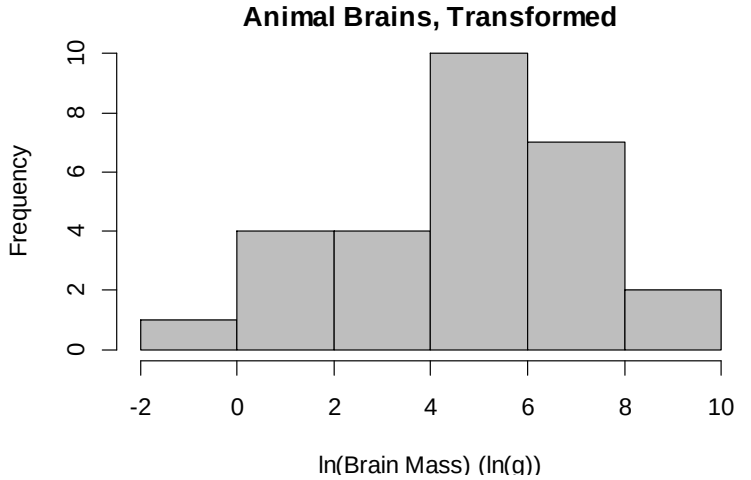

Figure 8 - Transformed Brain Mass Histogram for Example 5

That is much better!