Stat 321 - Day 30
Descriptive Statistics: Comparisons, Transformations

Today we continue our study of descriptive statistics by considering a new graphical display (boxplots), a test for outliers, the idea of transforming data, and some cautions about measures of center and spread.

Example: Cloud Seeding
Scientists have long investigated whether humans can help nature to produce more rainfall. In one study researchers in southern Florida explored whether injecting silver iodide into cumulus clouds would lead to increased rainfall. On each of 52 days that were judged to be suitable for cloud seeding, a target cloud was identified and a plane flew through the target cloud in order to seed it. Randomization was used to determine whether or not to load a seeding mechanism and seed the target cloud with silver iodide on that day. Radar was used to measure the volume of rainfall from the selected cloud during the next 24 hours. The results (measured in volume units of acre-feet, "height" of rain across one acre) are presented below and also in the Minitab worksheet `CloudSeeding.mtw`:

Unseeded:

1.0	4.9	4.9	11.5	17.3	21.7	24.4	26.1	26.3	28.6	29.0	36.6	41.1
47.3	68.5	81.2	87.0	95.0	147.8	163.0	244.3	321.2	345.5	372.4	830.1	1202.6

Seeded:

4.1	7.7	17.5	31.4	32.7	40.6	92.4	115.3	118.3	119.0	129.6	198.6	200.7
242.5	255.0	274.7	274.7	302.8	334.1	430.0	489.1	703.4	978.0	1656.0	1697.8	2745.6

(a) Use Minitab to produce dotplots for comparing the distributions of rainfall between seeded and unseeded clouds (`Graph> Dotplot; One Y, With Groups`). Compare and contrast the distributions of rainfall amounts between these two groups of seeded and unseeded clouds. In particular, based on these data, would you say there was evidence that the seeded clouds tended to have more rainfall than the unseeded clouds?

An important idea when comparing distributions between two or more groups is that of a *statistical tendency*. A tendency is not a hard-and-fast rule, but *in general*, or *on average*, the seeded clouds *tended* to produce higher rainfall amounts than the unseeded clouds.
A graphical display called a *boxplot* (or box-and-whisker plot) is based on the *five-number summary*, which consists of the minimum, lower quartile, median, upper quartile, and maximum. Boxplots are particularly useful for comparing distributions between two or more groups. A boxplot is constructed as follows:

- A box is formed around the middle 50% of the data, with boundaries at the quartiles.
- A line is marked within the box at the median.
- Lines (known as "whiskers") are extended from the lower quartile down to the minimum, and from the upper quartile up to the maximum.
- Outliers are marked separately with an asterisk. Outliers are values that fall more than 1.5 IQR away from their nearer quartile. In other words, the cut-off values for outliers are Q1 - 1.5IQR and Q3 + 1.5IQR.
- When outliers are present, the "whiskers" are pulled back to the most extreme non-outlier.

(b) Use Minitab to calculate the five-number summaries of rainfall amounts for the two groups (`Stat> Basic Statistics> Display Descriptive Statistics`). Record these below:

	Minimum	Lower quartile	Median	Upper quartile	Maximum
Unseeded					
Seeded					

(c) Use Minitab to produce boxplots of the rainfall amounts between the two groups (`Graph> Boxplot; One Y, with groups`). Comment on what the boxplots reveal about comparing the rainfall amounts between the two groups. [*Hint*: Comment on center, spread, shape, and unusual observations.]

(d) Use the "outlier rule" described above to confirm Minitab's identification of outliers, giving the cut-off values (on both the low and high ends) for observations to be considered as outliers in each group.

Unseeded: Low cut-off: High cut-off:

Seeded: Low cut-off: High cut-off:

When a distribution is skewed, we often apply a transformation in an effort to make the distribution more symmetric. Typical choices for transformations are log functions and power functions, such as square root. When the distribution is skewed to the right, try a log transformation or a power less than one; when the skew is to the left, try a power greater than one.

(e) Take the log (base 10) of the rainfall amounts (`MTB> let c5=logt(c3)`). Compare the seeded and unseeded groups by producing boxplots of the log of the rainfall amounts. Comment on whether the distributions appear to be more symmetric. Then comment on what they reveal about the center and spread of log(rainfall) between the two groups.

Example: Roller Coasters
The Minitab worksheet coasters.mtw contains data on roller coasters in the United States.

(f) Determine five-number summaries, and produce comparative boxplots, of the speeds of these roller coasters, comparing wooden and steel coasters. Write a paragraph comparing and contrasting the distributions of speed between these two types.

(g) Between what two values do 50% of the speeds of the wooden coasters fall? How about for the steel coasters?

(h) Based solely on the five-number summaries, what can you say about the percentage of coasters of each type that travel faster than 60 miles per hour? Explain how you know.

(i) Based solely on the five-number summaries, what can you say about the percentage of coasters of each type that travel faster than 50 miles per hour? Explain how you know.