Lesson Plan

- Box Plots
- Linear Relationships
- Correlation
The boxplot is a graphic that summarizes a dataset by plotting 5 numbers:

- the least
- the largest
- the median
- the 25th percentile
- the 75th percentile

The last two are the medians of the numbers less than the median and the numbers greater than the median, respectively (according to the convention for this class).
Boxplots are often made for different groups and plotted side by side to highlight differences.

A boxplot looks like this:
Suppose one made a side-by-side boxplot of the IQ scores for Duke students and the IQ scores for students at the Sam Houston Institute of Technology. What do you think they would look like?
A modified boxplot is slightly different. Instead of having the "whiskers" extend from the box-sides to the least and largest observations, the whiskers only extend to the least and largest observations that are not outliers. Outliers are plotted separately as points, to facilitate inspection.

In the context of a modified boxplot, we define an outlier to be any observation that falls more than 1.5*IQR below the 25th quartile or less than 1.5*IQR above the 75th quartile. For further modification, JMP-IN puts a diamond on the plot. The centerline of the diamond is at the mean of the data, and the width of the diamond is one standard deviation.

The modified boxplot is slightly different. Instead of having the "whiskers" extend from the box-sides to the least and largest observations, the whiskers only extend to the least and largest observations that are not outliers. Outliers are plotted separately as points, to facilitate inspection.
Consider the following (ordered) data:

-10, -2, -2, -2, -1, 0, 0, 1, 2
The algebraic equation for a line is:

\[ q + X^a = \lambda \]

The use of coordinate axes to show functional relationships was invented by Rene Descartes (1596-1650). He was an artillery officer, and probably got the idea from pictures that showed the trajectories of cannonballs.
5.4 Correlation

Sir Francis Galton explored Africa, invented eugenics, studied whether ships that carried missionaries were less likely to be lost at sea, pioneered birth-and-death models and meteorology, and was Charles Darwin’s cousin. He also was the first to conceive of linear regression (although he did not have the mathematical skill to develop the formulae, and got a friend of his at Cambridge to do the derivations).
Typically, correlation is a measure of the strength of the linear association between two continuous variables. The book discusses the relationship between the height of fathers and the height of sons. Typically, correlation is a measure of the strength of the linear association between two continuous variables. The term comes from regression analysis. A line to the points in a scatterplot. The term comes from scatterplot.
The correlation tells you how much information one variable provides about the other. If the absolute value of the correlation is near 1, then knowing one tells the other almost perfectly (if the relationship is linear). If the correlation is near zero, then knowing one provides no information about the other. All correlations are between -1 and 1, inclusive.

The square of the correlation is called the coefficient of determination. It is the proportion of the variation in $Y$ that is explained by knowledge of $X$. If this is 0, then $X$ provides no insight about $Y$ (assuming linearity).

The correlation tells you how much information one variable provides about the other.
The assumption of linearity is important. Specifically, it asserts that:

\[ Y_i = aX_i + b + e_i \]

where the \( Y_i \) and \( X_i \) are the observed values for the \( i \)th case (e.g., a father-son pair) and the \( e_i \) is random error (due to genetics, measurement error, etc.).

If the assumption does not hold, then one can get strange and even misleading behavior.

To calculate the correlation coefficient, just take the average of the products of the \( z \)-transforms of the \( X \) values with the \( z \)-transforms of the \( Y \) values.

\[ \bar{\epsilon} + q + \bar{X}q = \bar{Y} \]

The assumption of linearity is important.
Example: Suppose your data are (1,0), (2,1), and (3,5). Then the mean of the \( X \) values is 2, the mean of the \( Y \) values is 2, and the \( s_d \)s are:

\[
\begin{align*}
\sigma_X &= \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{1}} = \sqrt{\frac{1}{1}} = 1.0 \quad (3-2) \cdot 8.165 = 1.2245 \\
\sigma_Y &= \sqrt{\frac{(0-2)^2 + (1-2)^2 + (5-2)^2}{1}} = \sqrt{\frac{14}{1}} = 3.742 \quad (2-2) \cdot 8.165 = 1.2245
\end{align*}
\]

The \( z \)-transformation subtracts the mean and divides by the standard deviation for each of the \( X \) and \( Y \) values, giving:

\[
\begin{align*}
Z_X &= \frac{Z - \mu_X}{\sigma_X} \\
Z_Y &= \frac{Y - \mu_Y}{\sigma_Y}
\end{align*}
\]

\[
\begin{align*}
Z_X &= \frac{1 - 2}{1} = -1 \\
Z_Y &= \frac{0 - 2}{3.742} = -0.532
\end{align*}
\]
Then the correlation coefficient is

\[
\begin{align*}
\text{Do we really believe this?} \\
\text{How much of the variation in } Y \text{ is explained by knowing } X \text{? This is}\end{align*}
\[
\begin{align*}
\begin{align*}
\frac{3}{1} = \mu
\end{align*}
\end{align*}
\]
Spreads in data vs. spreads under normality

\* Heavier, approx. higher than normal tails or spread

\* Under normality, spread of middle half of data is 1.35

\* So, consider \( \frac{2.3}{\text{Middle 3/4}} \), consider \( \frac{1.35}{\text{Middle 3/4}} \), etc.

\* Spread of middle 3/4 of data is 2.35

\* Under normality, spread of middle half of data is 1.35

\* Heavier, approx. higher than normal tails or spread