Math 146 2.4 - Scatter Plots, Linear Regression and Correlation

This section introduces us to *paired sample data*. This will be expanded greatly in section 10.1 and 10.2.

DEFINITIONS

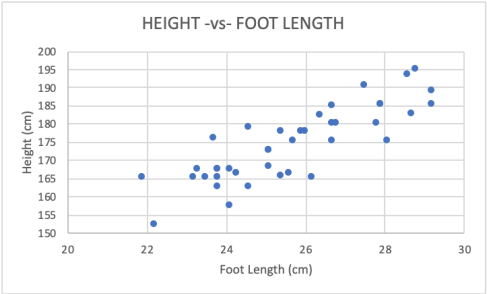
A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

Warning: Correlation does not imply causation as we have seen a number of times already.

Scatter Plots

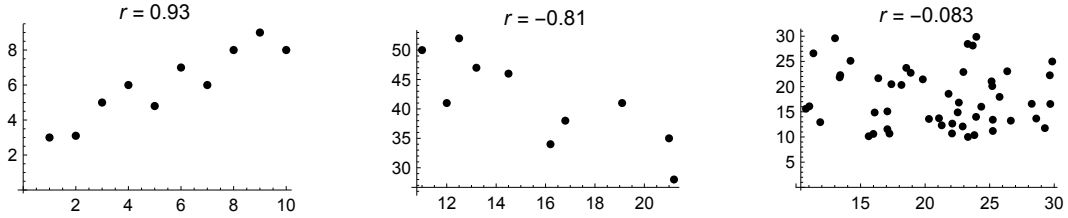
A scatterplot is a graphical representation of paired data. We treat each pair as an (x, y) pair. The following are 40 subjects foot length in cm, and their height.



There is obviously a slight upward trend for taller heights having larger foot lengths, and it appears to be linear. This is called linear correlation.

Correlation

Correlation is measured using the **coefficient of correlation**, ***r***, and can have values between –1 and 1 inclusive.



The closer r is to 1, the better the data fits a positive sloped line; the closer r is to –1, the closer to a negative sloped line; and the closer r is to 0 means there is no correlation. But, how do we know when there is *significant correlation* or not? We'll explore this in more detail in chapter 1, but for now, we can say it really determines on the size of the data set. A small data set will need a correlation coefficient very close to 1 or –1, while a very large data set may have significant correlation with r as small as $r = 0.4$. The following table allows us to determine if the is significant linear correlation or not. This is just a portion of Table A-6 in the back of our text.

n	4	5	6	7	8	9	10	15	20	50	100
critical r	0.950	0.878	0.811	0.754	0.707	0.666	0.632	0.514	0.444	0.275	0.196

To use this table, suppose we have 8 data points, that is $n = 8$. If $r > 0.707$ or $r < -0.707$, we can conclude there is sufficient evidence that there exists linear correlation. If $-0.707 < r < 0.707$ we conclude there is not sufficient evidence to conclude linear correlation.

Example 1 Determine the significance of correlation in the scatter plots above.

Regression

Once we conclude there is linear correlation, we can find the line that best fits the data. This is called the **line of best fit**, **regression line**, or **least squares fit**.

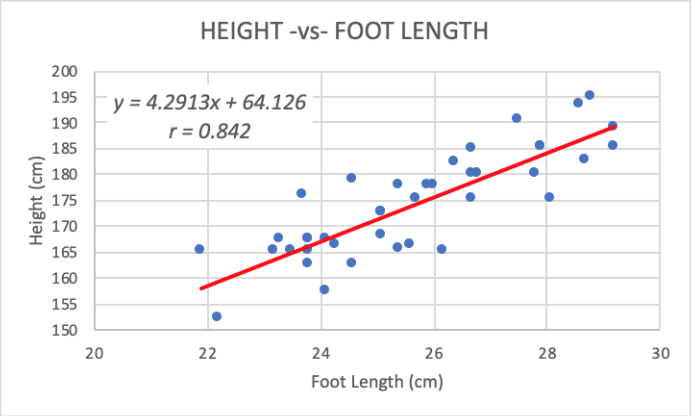
Definition

The *linear regression equation* is given by:

$$\hat{y} = b_0 + b_1 x$$

Where b_1 is the slope of the line, and b_0 is the y intercept.

Note that this is really a relabeled slope intercept form for a line $y = m x + b$. For, now, we'll let technology calculate both the correlation coefficient and the linear regression line (see the corresponding Technology Insights on how to use the various technology.) For our height -vs- foot length data above we get the following correlation coefficient and regression line:



Example 2 If Chuck has a foot length of 26 cm, predict his height.