Lecture 5

- **Compute numerical measures of association (i.e. the covariance and correlation)**

- **Being careful about the presence of outliers as well as spurious correlations.**

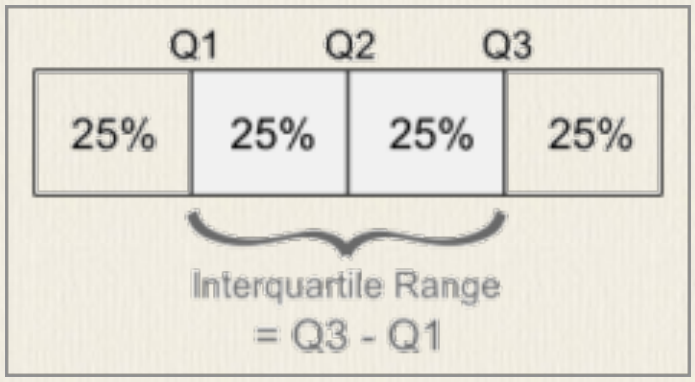
Measures of relative standing

A **quartile** is a statistical term describing a division of observations into four defined intervals based upon the values of the data and how they compare to the entire set of observations.

Each quartile contains 25% of the total observations. Generally, the data is ordered from smallest to largest with those observations falling below 25% of all the data analyzed allocated within the 1st quartile, observations falling between 25.1% and 50% and allocated in the 2nd quartile, then the observations falling between 51% and 75% allocated in the 3rd quartile, and finally the remaining observations allocated in the 4th quartile.

IQR

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by *Q1*, *Q2*, and *Q3*, respectively.



Example:

Suppose that the city where the high temperature was 77 failed to report, so that we are left with the following 11 numbers arranged according to size:

	72	74	75	78	79	82	85	86	90	93	94

Find *Q1*, *Q2* (the median), and *Q3*.

Solution:

For n = 11, the median position is 6 and, referring to the preceding data, which are already arranged according to size, we find that the median is *Q2* = 82. For the five values below 82 the median position is 3, and *Q1*, the third value, equals 75. Counting from the other end, *Q3*, the third value, equals 90. One can see that there are two values below 75, two values between 75 and 82, two values between 82 and 90, and two values above 90. Again, this satisfies the requirement for the three quartiles, *Q1*, *Q2*, and *Q3*.

Statistics assumes that your data points (the numbers in your list) are clustered around some central value. The "box" in the box-and-whisker plot contains, and thereby highlights, the middle half of these data points.

Measure of relationship

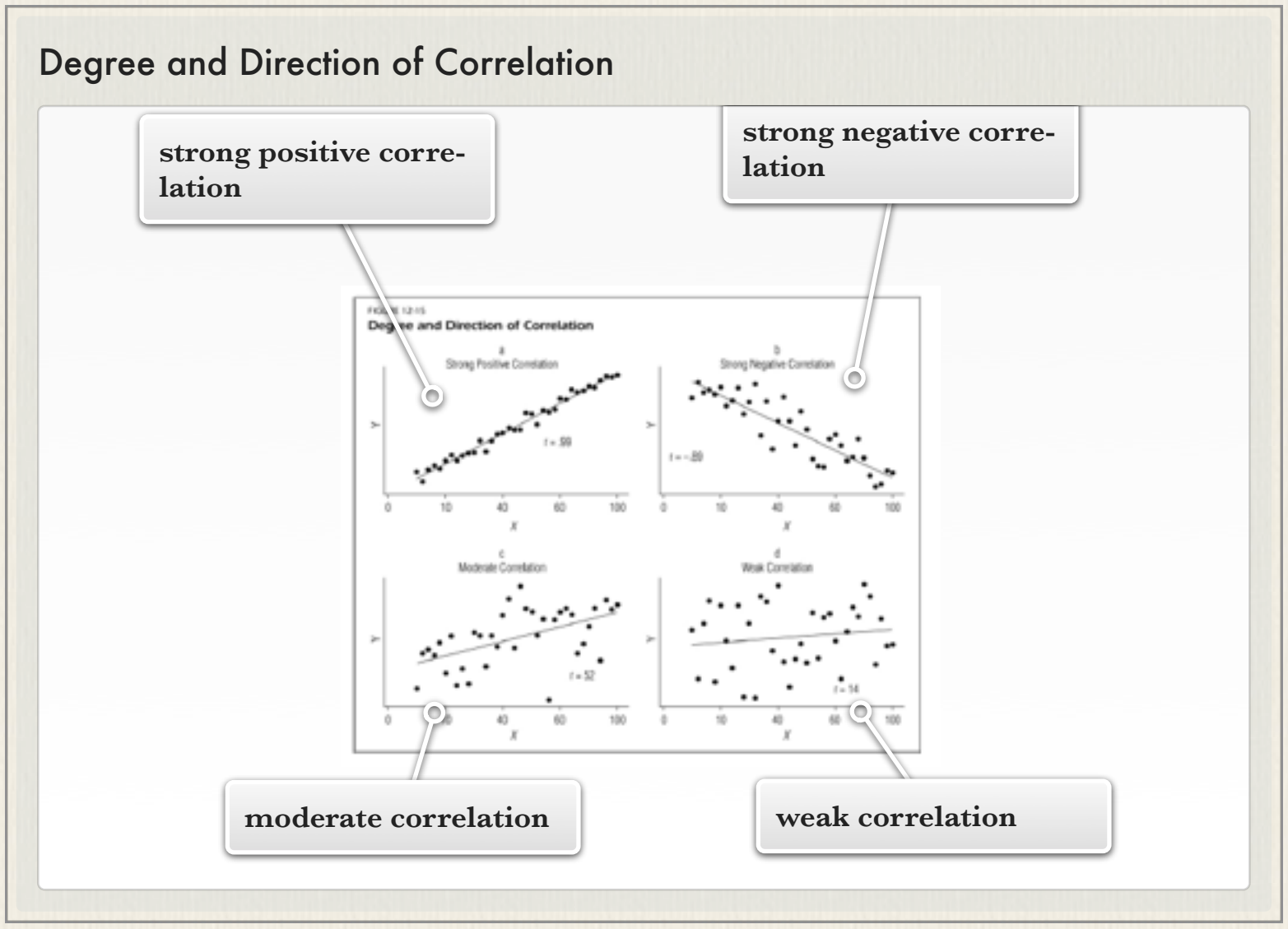
Covariance: In probability theory and statistics, covariance is a measure of how much two random variables change together.

$$cov(x, y) = s_x y = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

Correlation Coefficient

The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables.

$$r = \frac{s_{xy}}{s_x s_y}$$



The correlation coefficient tells us two key things about the association:

Direction - A positive correlation tells us that as one variable gets bigger the other tends to get bigger. A negative correlation means that if one variable gets bigger the other tends to get smaller (e.g. as a student's level of economic deprivation decreases their academic performance increases).

Strength - The weakest linear relationship is indicated by a correlation coefficient equal to 0 (actually this represents no correlation!). The strongest linear correlation is indicated by a correlation of -1 or 1. The strength of the relationship is indicated by the magnitude of the value regardless of the sign (+ or -), so a correlation of -0.6 is equally as strong as a correlation of +0.6. Only the direction of the relationship differs.

Difference

Correlation is scaled to be between -1 and +1 depending on whether there is positive or negative correlation, and is dimensionless. The covariance however, ranges from zero, in the case of two independent variables, to *Var(X)*, in the case where the two sets of data are equal.

Example:

How to interpret the difference between the two measurements?

Solution:

Random experiment