1. Overview

Up to now we have only looked at what are called "univariate" statistics. This means we are only studying a single variable in a given population or a given sample. For example, we might talk about mean height, mean weight, the standard deviation of an LSAT score. But now, we turn to relationships between two variables.

2. Basic Definitions & Graphical Summary

A. Scatterplot or Scatter Diagram -- a graphical representation of a two-variable analysis.

- A scatterplot or scatter diagram is a two dimensional plot of data. The horizontal dimension is called x, and the vertical dimension is called y.
- Each point on a scatterplot or scatter diagram shows two values, an x value and a y value. Each point represents a single case. A single case could be a single person or object, but a single case could be a matched pair (e.g. father-son, twins, husband-wife)

B. Positive and negative relationships

There is a POSITIVE relationship if above-average values of x are associated with above-average values of y.

Conversely, there is a NEGATIVE relationship if above-average values of x are associated with below average values of y.

C. Warning! Scatter diagrams only show association, but association does not mean causation (firefighters, fire damage)

In many disciplines, X and Y are usually called the INDEPENDENT and DEPENDENT variables respectively. They are given these names because the independent variable is thought to influence the dependent variable. There is nothing to stop us from reversing the relationship. Designation of independent and dependent rely strongly on how the question is being asked.

3. Numerical Summary: The correlation coefficient r

The CORRELATION COEFFICIENT, denoted r, measures how close the data are to a straight line or in other words it measures the strength of association. It is a numerical summary of the scatter diagram graphic.

The correlation coefficient can take values from -1 to +1. Values near zero mean that the data is not close to a straight line. Values near the ones (both positive and negative) mean that the data is very close to a straight line.

Formula

Your text gives you a very long formula for calculating the correlation coefficient (pp 132-134) and I am not certain how useful it is. Instead, read the technical note on p. 134, the formula is reproduced here:

$$r = \frac{(\text{average of the products } xy) - ((\text{average } x) * (\text{average } y))}{(\text{Standard Deviation } x) * (\text{Standard Deviation } y)}$$

4. The SD Line

On the original front cover of your book is a scatter plot of father's height (x) vs. son's height (y). The dashed line is a SD line. The SD line, on a scatterplot, is a line that goes through the point of averages

(the point with values = mean of X and mean of Y). It's slope (Chapter 10) equal to the SD of Y divided by the SD of X. This will become more useful when we get to Chapter 10.

5. Properties of the correlation coefficient

- -1 <= r <= 1
- If r is close to 1 or -1, the data are close to a line
- If r is close to 0, the data are not close to a line
- Pictures! (see pages 127 and 129 of your text)
- The correlation r measures how close the data are to a line
- r does NOT tell what percentage of the data fall on the line
- The correlation between x and y is the same as the correlation between y and x.
- Invariant under addition. If some constant "a" is added to every one of the X or the Y values, the correlation is unchanged.
- Invariant under multiplication: if all of the x or the y values are multiplied by some positive constant "b", the correlation is unchanged. The correlation can change very dramatically if only ONE of the data points is changed.

	show	receipts	capacity
1.	Angels in America	326121	7456
2.	Blood Brothers	154064	7936
3.	Cats	346723	11856
4.	Crazy for You	463377	11720
5.	Falsettos	86864	6440
6.	Fool Moon	163802	10696
7.	The Goodbye Girl	429158	12736
8.	Guys and Dolls	457087	10256
9.	Jelly's Last Jam	253951	9864
10.	Kiss of the Spider Woman	406498	9048
11.	Les Miserables	481973	11304
12.	Miss Saigon	625804	14088
13.	Phantom of the Opera	674609	12872
14.	Shakespeare for my Father	78898	4520
15.	The Sisters Rosensweig	340862	8768
16.	Someone Who'll Watch over Me	73903	6248
17.	Tommy	590334	12784
18.	The Will Rogers Follies	265561	11360

The means and SDs are below:

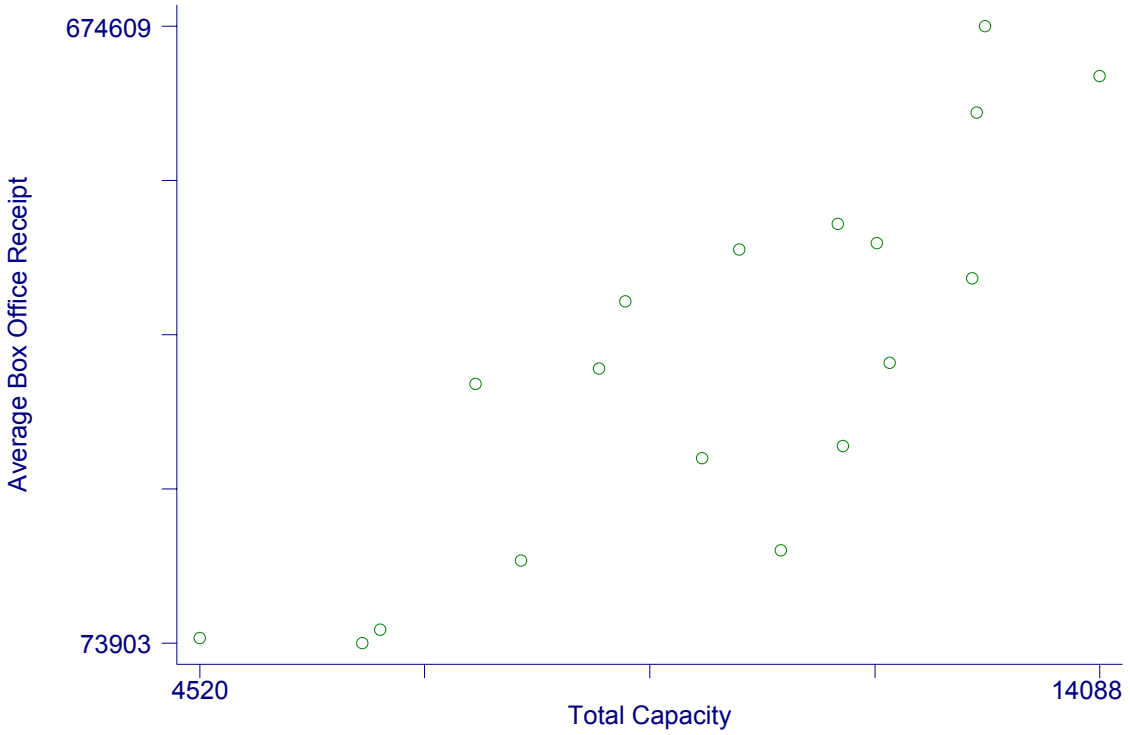
Variable	Obs	Mean	Std. Dev.	Min	Max
receipts	18	345532.7	187723.8	73903	674609
capacity	18	9997.333	2664.093	4520	14088

A. Compute the sample correlation coefficient r for the data.

The correlation coefficient is:

$$\frac{3842363105.777778 - (345532.7 \times 9997.333)}{187723.8 \times 2664.093} = .82$$

The correlation r of theater capacity and receipts is .82.



3. Where the correlation can fail you or deceive you

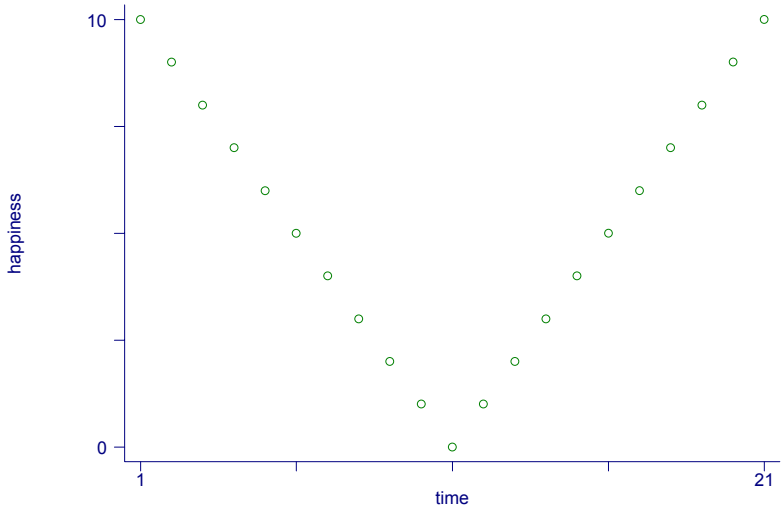
Different SDs

Appearances can be deceptive, the overall appearance of a scatter diagram depends on the Standard Deviations of the individual X and Y variables. Smaller standard deviations make the scatter diagram look "tighter" or more closely placed together. This happens because r is defined by how far individual points deviate from their means divided by their SDs. Clustering is relative. So beware, your eye can be fooled. Be sure to:

- examine the ranges of the X and Y variables when comparing two sets of data
- look at the standard deviation for the X variable and the Y variable. Are most points within a standard deviation or not?
- calculate the correlation coefficient (if it hasn't been done yet) or review them (if they have been calculated for you)

Outliers --The correlation is useful when your data points are football shaped. But sometimes r will mislead. A single outlier can counterbalance a strong linear relationship. .

Non-Linear relationships -- Two variables can be related, but their relationship is not well described by a straight line. Correlations are good for straight-line relationships, they are terrible for curved or other non-linear relationships. The correlation coefficient for a curved relationship will be near zero even though there is clearly a relationship.



Solutions? Always take a look at the scatter diagram if possible.

Ecological Correlations

Correlations for aggregated information (e.g. information by state) are almost always stronger than correlations for individuals. Often people calculate correlations based on rates or averages that are based on many individuals combined. This tends to make a relationship look stronger than it is. Beware when people do this.

Association is NOT causation

Remember: a correlation tells you how strongly two variables are related in a linear fashion. It doesn't mean one variable causes another to happen. The problem here is one of CONFOUNDING, that is, a third variable may be present that has not be taken into account.