Correlation

Greg C Elvers

What Is Correlation?

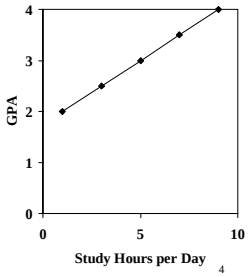
- *Correlation* is a descriptive statistic that tells you if two variables are related to each other
  - E.g. Is your GPA related to how much you study?
- When two variables are correlated, knowing the value of one variable allows you to predict the value of the other variable

Perfect Correlation

- When two variables are perfectly correlated, knowing the value of one variable allows you to exactly predict the value of the other variable

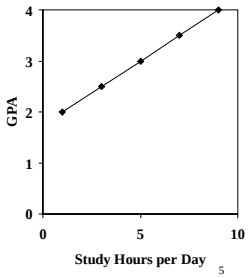
Perfect Correlation

- For example, if the only thing that determined your GPA was the amount of time that you studied, then the two would be perfectly correlated
- If you know the value of one variable, you can exactly determine the value of the other variable



Perfect Correlation

- That is, all the variability in one variable is explained by the variability in the other variable



5

Perfect Correlations

- Few, if any, psychological variables are perfectly correlated with each other
- Many non-psychological variables do have a perfect correlation
  - E.g. Time since the beginning of class and the time remaining in the class are perfectly correlated
- What are other examples of perfectly correlated variables?

6

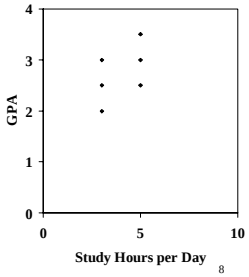
Less Than Perfect Correlations

- Even if two variables are correlated, most of the time you cannot perfectly predict the value of one variable given the other
  - E.g., other variables besides amount of time spent studying influence your GPA
  - Some of the variability is people's GPA is due to the amount of time spent studying, but not all the variability is due to it

7

Less Than Perfect Correlations

Study Time	IQ	GPA
3	80	2.0
3	100	2.5
3	120	3.0
5	80	2.5
5	100	3.0
5	120	3.5



8

Less Than Perfect Correlations

- With a less than perfect correlation, we can no longer perfectly predict the value of one variable given the other variable
- We cannot explain all the variability in one variable with the variability in the other variable

The Correlation Coefficient

- Correlation coefficients tell us how perfectly two (or more) variables are related to each other
- They can also be used to determine how much variability in one variable is explainable by variation in the other variable.

Pearson's Product Moment Correlation Coefficient

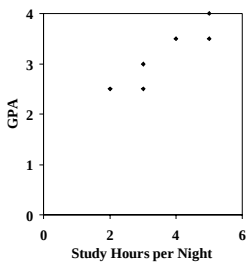
- *Pearson's product moment correlation coefficient*, or *Pearson's r*, for short is a very common measure of how strongly two variables are related to each other
- Pearson's r must lie in the range of -1 to +1 inclusive

Interpretation of Pearson's r

- To interpret Pearson's r, you must consider two parts of it:
  - The sign of r
  - The *magnitude*, or absolute value of r

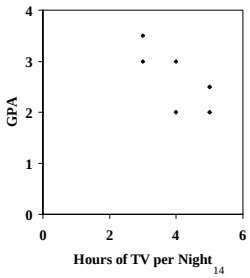
The Sign of r

- When r is greater than 0 (I.e. its sign is positive) the variables are said to have a *direct relation*
- In a direct relation, as the value of one variable increases, the value of the other variable also tends to increase



The Sign of r

- When r is less than 0 (i.e., its sign is negative) the variables are said to have an *indirect relation*
- In an indirect relation, as the value of one variable increases, the value of the other variable tends to decrease



Is the Sign of r + or -?

- As the number of cigarettes smoked per day increases, GPA tends to decrease
- As the number of cats in a farm yard increases, the number of mice tends to decrease
- As the weight of a cat increases, the length of its whiskers tends to increase

Is the Sign of r + or -?

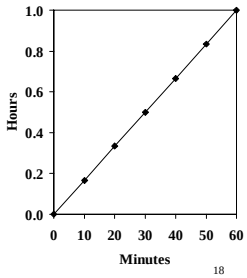
- Create two examples of correlations and determine if the sign of r is positive or negative

The Magnitude of r

- The *magnitude* refers to the size of the correlation coefficient ignoring the sign of r
- The magnitude is equivalent to taking the absolute value of r
- The larger the magnitude of r is, the more perfectly the two variables are related to each other
- The smaller the magnitude of r is, the less perfectly the two variables are related to each other

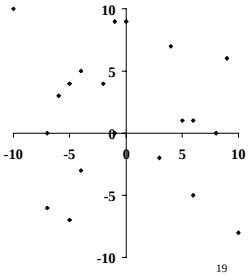
r = 1

- When r equals 1.0, there is a perfect correlation between the variables
- Knowing the value of one variable exactly predicts the value of the other variable



r = 0

- When r equals 0, either the assumptions of correlation have been violated or there is no relation between the two variables
- The points in a scatter plot with r = 0 will tend to form a circular cluster



0 < | r | < 1

- The larger the magnitude or r is, the more the scatter plot's points will tend to cluster tightly about a line

0 < | r | < 1

Magnitude of r

- Cohen (1988) recommends the following values of r for "small", "medium", and "large" effects

Correlation	Negative	Positive
Small	-.29 to -.10	.10 to .29
Medium	-.49 to -.30	.30 to .49
Large	-1.00 to -.50	.50 to 1.00

Magnitude of r

- List a couple of pairs of variables and guess whether the magnitude of r is closer to 0 or closer to 1

Pearson's r

- Pearson's r makes several assumptions about the data
- When these assumptions are violated, r must be interpreted with extreme caution
- Assumptions:
  - Linear relation
  - Non-truncated range
  - Sufficiently large sample size

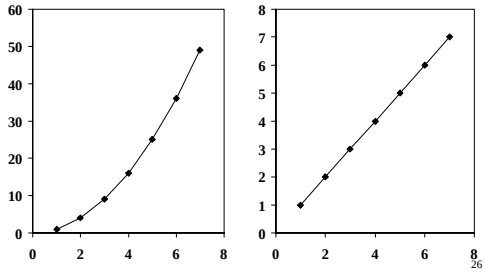
Linear Relation

- Pearson's r, in its simplest form, only works for variables that are linearly related
  - That is, the equation that allows us to predict the value of one variable from the value of the other is a line:
    Y = slope * X + intercept
  - Always look at the scatter plot to determine if the two variables are approximately linearly related

Linear Relation

- If the variables are not linearly related, Pearson's r will indicate a smaller relation than actually exists
- Often, non-linear relations can be transformed into linear ones by taking the appropriate mathematical transformation

Square Root of Y Transformation

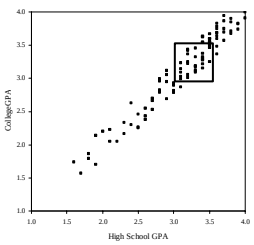


Non-Truncated Range

- A *truncated range* occurs when the range of one of the variables is very small
- When the range is truncated, Pearson's r will indicate a smaller relation between the variables than what actually exists
- Once a range truncation occurs, there is little that you can do; be careful not to design studies that will lead to a truncated range

Truncated Range

- A linear relation clearly exists in this data
- Consider only the data in the square (thereby truncating the range)
- Is the linear relation as clear as it was?
- No



Sample Size

- If the size of the sample is too small, relations can appear due to chance
  - These relations disappear when a larger sample is considered
- Too large of a sample can make near 0 correlations statistically significant, even though they have very little explanatory power

Sample Size

- The magnitude of r does not depend on sample size
- The likelihood of finding a statistically significant r does depend on sample size
- The sample should be large enough to generalize to the population of interest