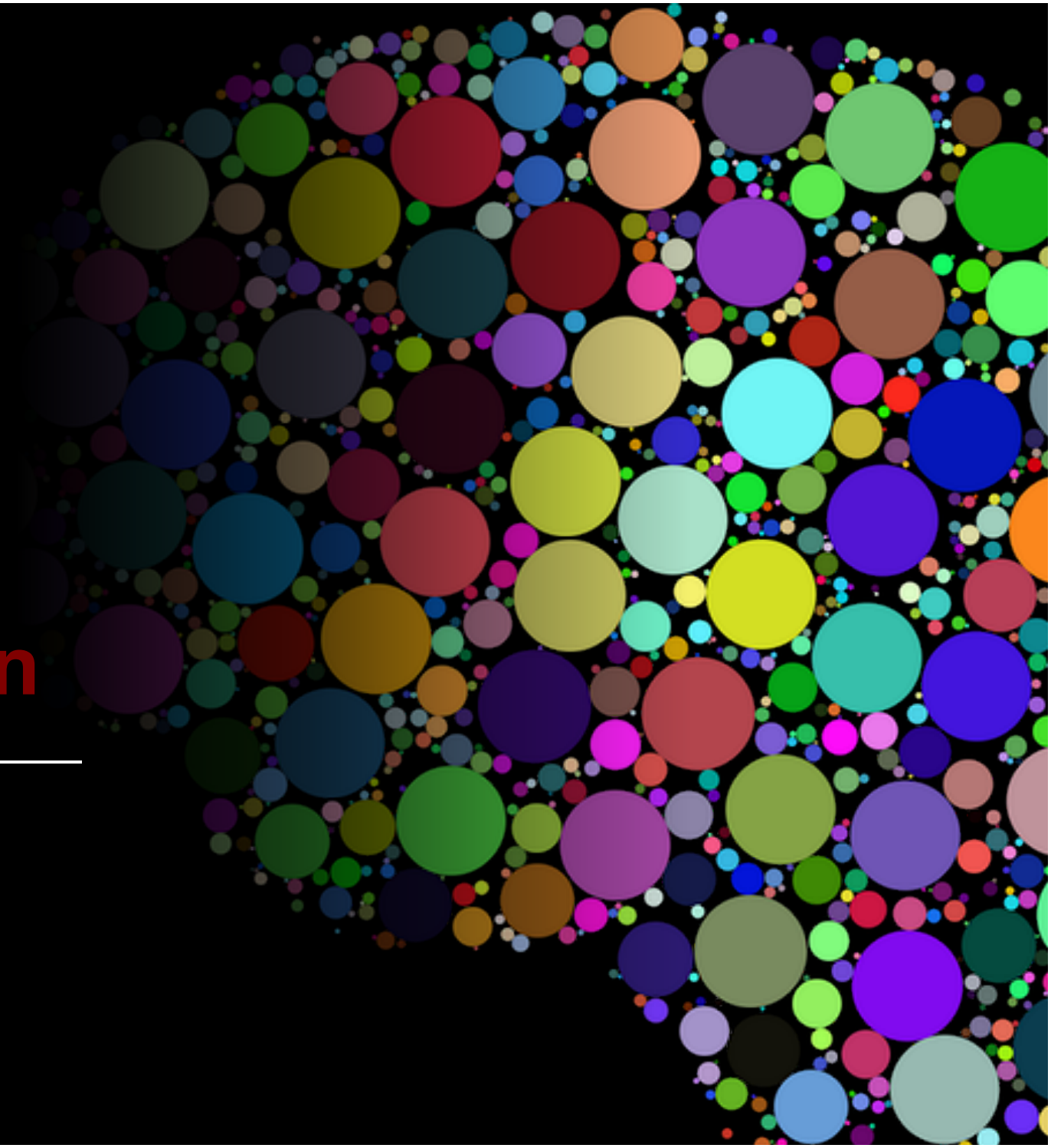
Correlation

Chapter 15



A research design reminder

> Experimental designs

- You directly manipulated the independent variable.

> Quasi-experimental designs

- You examined naturally occurring groups.

> Correlational designs

- You just observed the independent variable.

Defining Correlation

> Co-variation or co-relation between two variables

> These variables change together

> Usually scale (interval or ratio) variables

> Correlation does NOT mean causation!

Correlation versus Correlation

> Correlational designs = research that doesn't manipulate things because you can't ethically or don't want to.
  - You can use all kinds of statistics on these depending on what you types of variables you have.

> Correlation the statistic can be used in any design type with two continuous variables.

Correlation Coefficient

> A statistic that quantifies a relation between two variables

> Tells you two pieces of information:

- Direction
- Magnitude

Correlation Coefficient

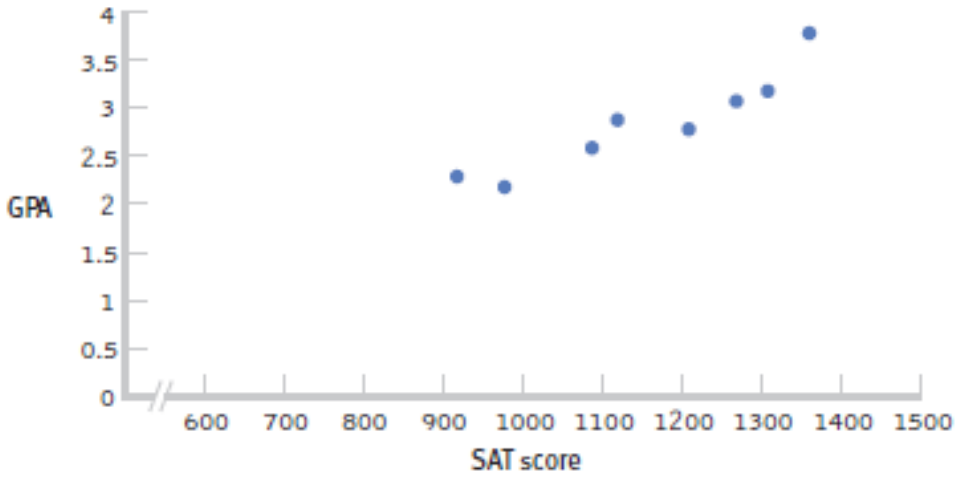
> Direction

- Can be either positive or negative
- Positive: as one variable goes up, the other variable goes up
- Negative: as one variable goes up, the other variable goes down.

Positive Correlation

Association between variables such that high scores on one variable tend to have high scores on the other variable

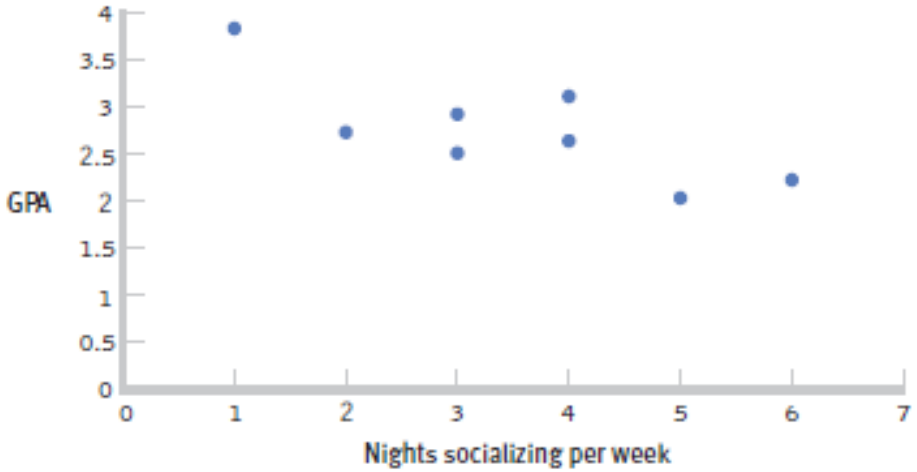
A direct relation between the variables



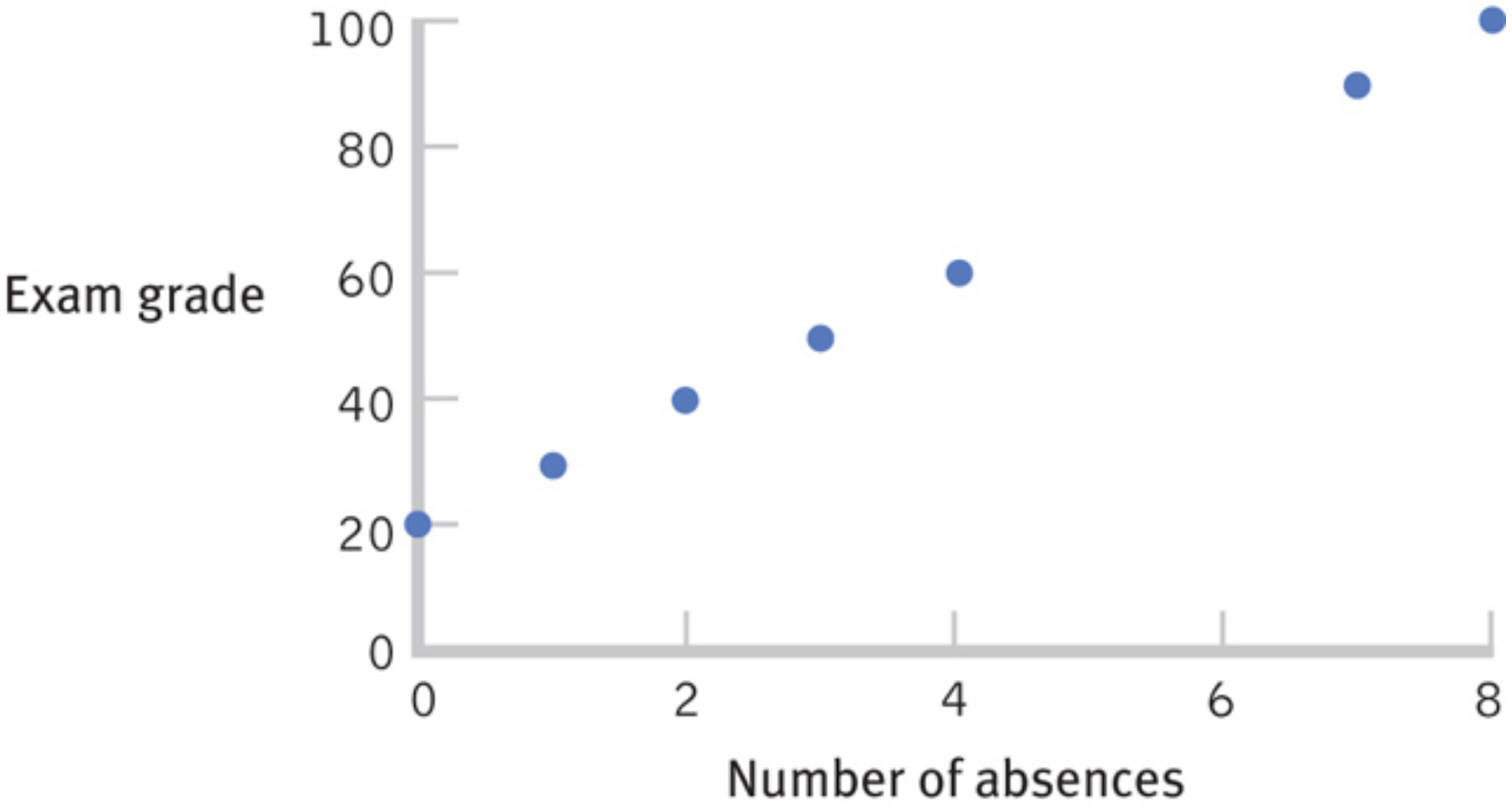
Negative Correlation

Association between variables such that high scores on one variable tend to have low scores on the other variable

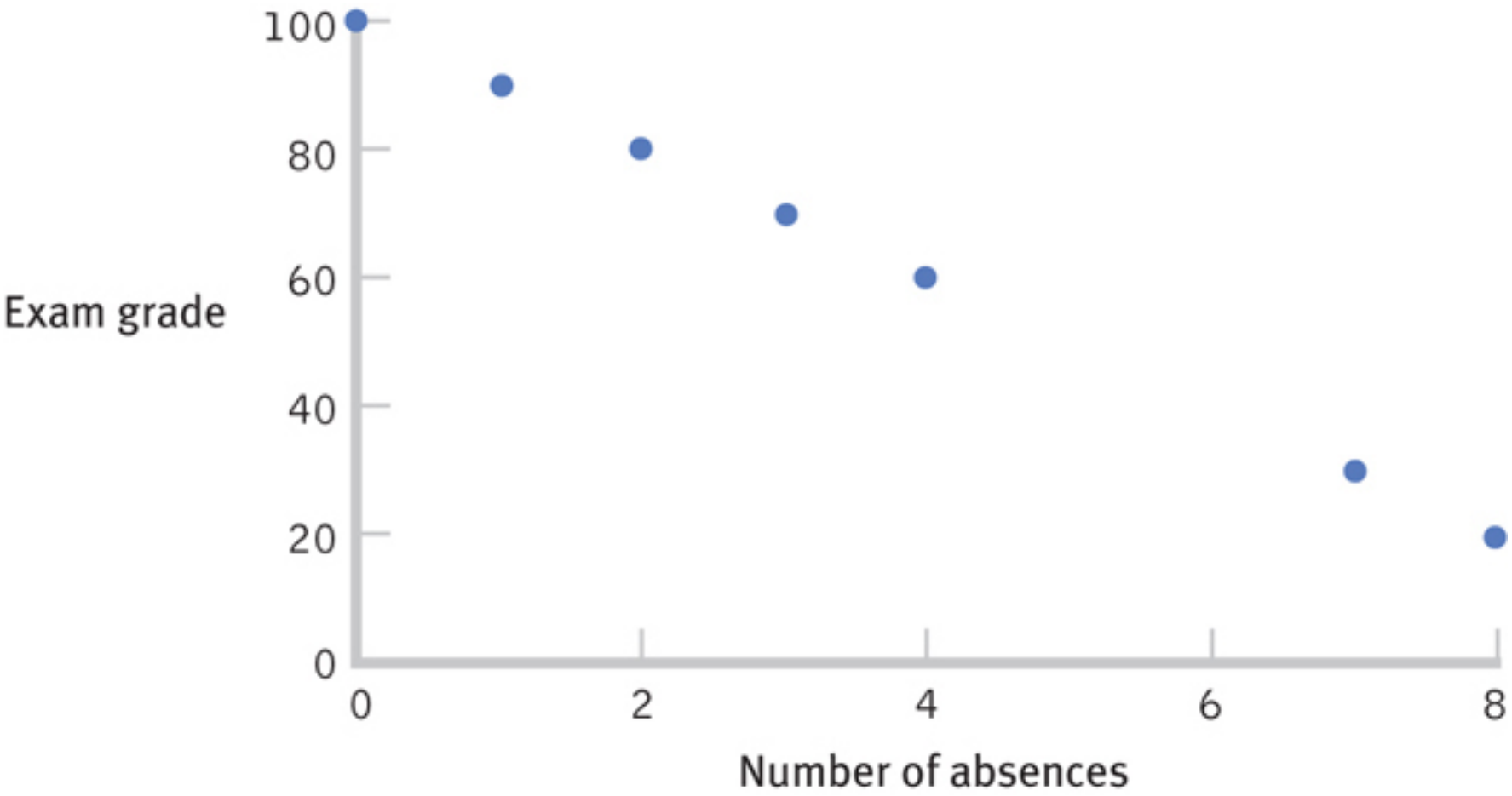
An inverse relation between the variables



A Perfect Positive Correlation



A Perfect Negative Correlation



Correlation Coefficient

> Magnitude

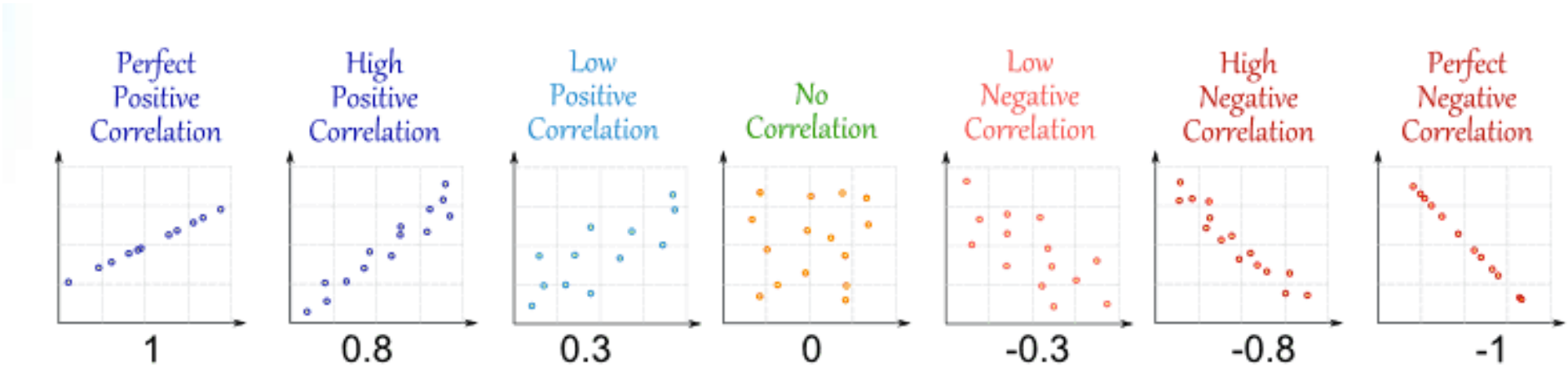
- Falls between -1.00 and 1.00
- The value of the number (not the sign) indicates the strength of the relation

TABLE 15-1. How Strong Is an Association?

Cohen (1988) published guidelines to help researchers determine the strength of a correlation from the correlation coefficient. In social science research, however, it is extremely unusual to have a correlation as high as 0.50, and many have disputed the utility of Cohen's conventions for many social science contexts.

Size of the Correlation	Correlation Coefficient
Small	0.10
Medium	0.30
Large	0.50

A visual guide



Let's play guess the correlation!

Check Your Learning

> Which is stronger?

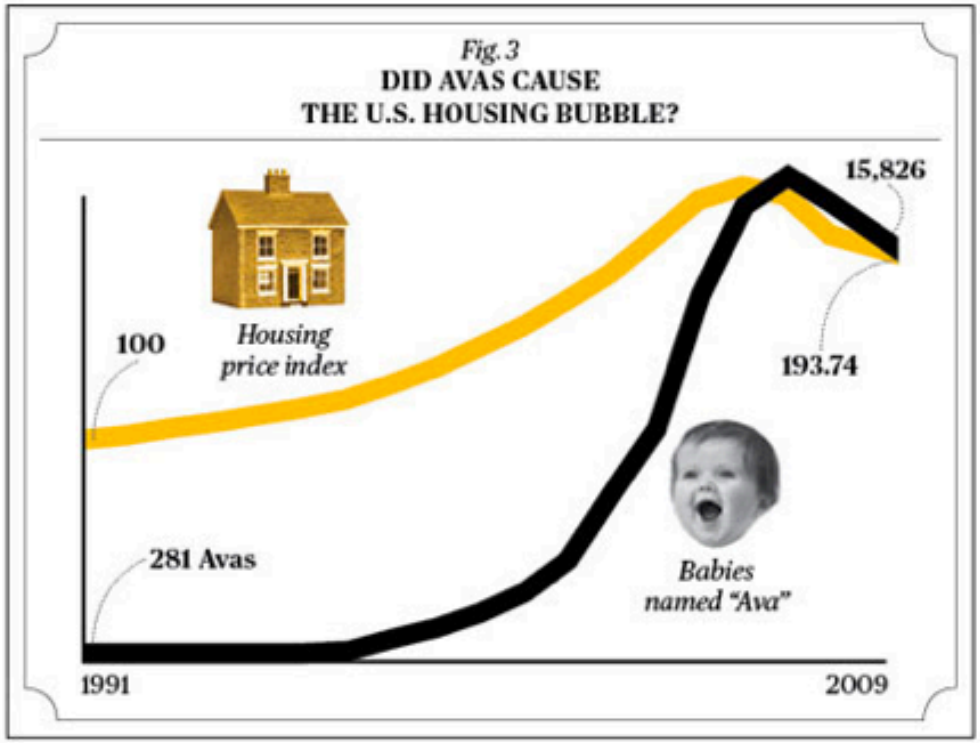
- A correlation of 0.25 or -0.74?

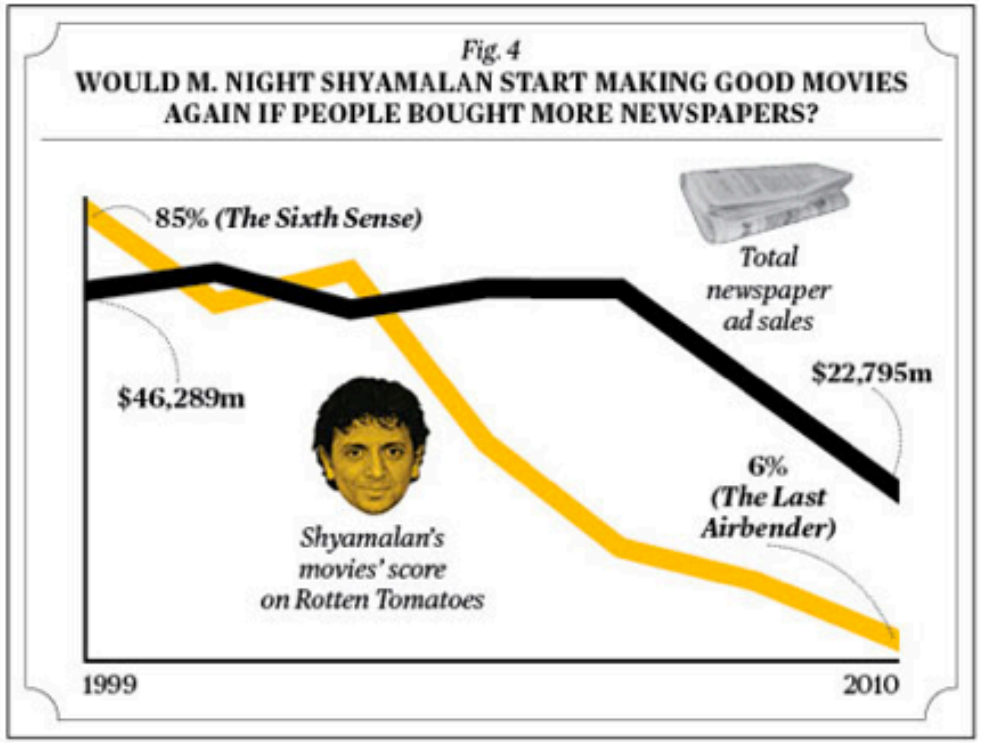
Misleading Correlations

> Something to think about

- There is a 0.91 correlation between ice cream consumption and drowning deaths.
  - > Does eating ice cream cause drowning?
  - > Does grief cause us to eat more ice cream?

Another silly example





Correlation Overall

> Tells us that two variables are related

- How they are related (positive, negative)
- Strength of relationship (closer to l 1 l is stronger).

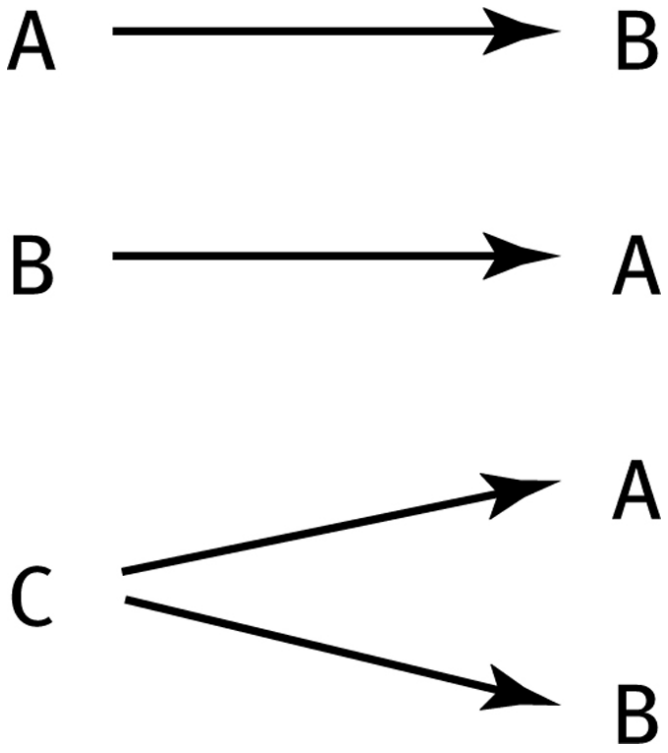
> Lots of things can be correlated – must think about what that means.

The Limitations of Correlation

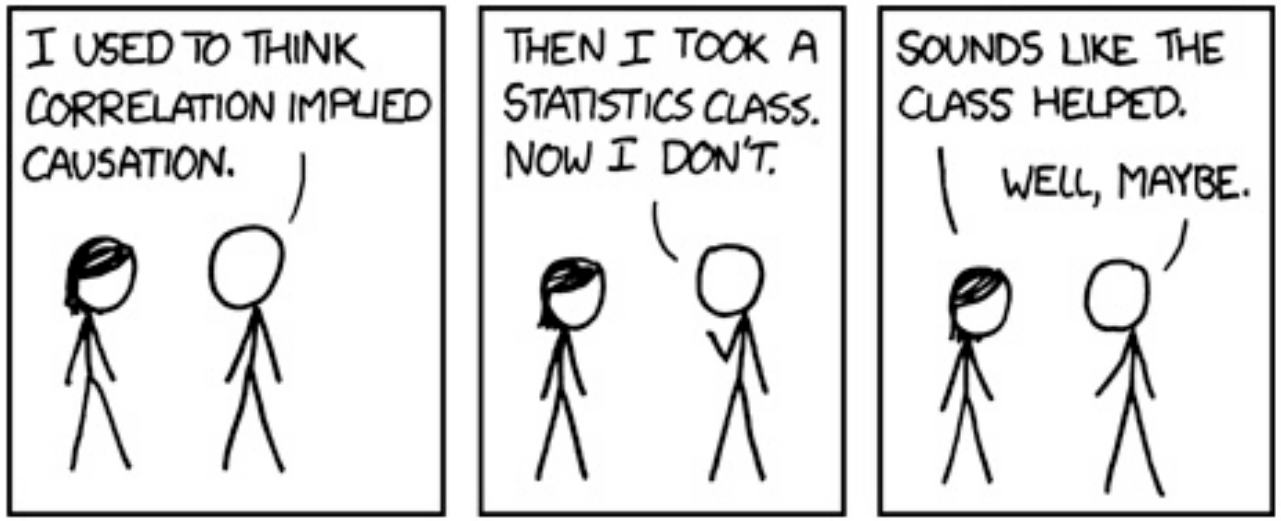
> Correlation is not causation.

- Invisible third variables

Three Possible Causal Explanations for a Correlation



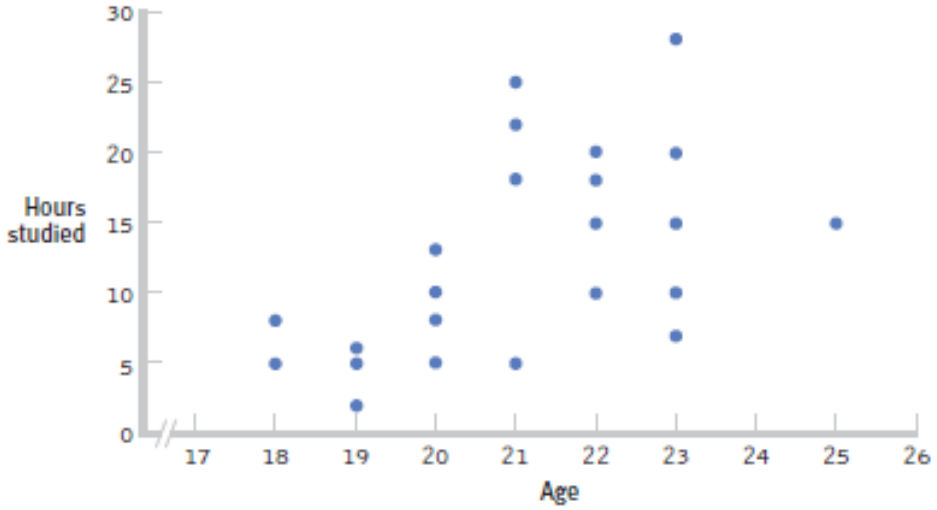
Causation Direction



The Limitations of Correlation, cont.

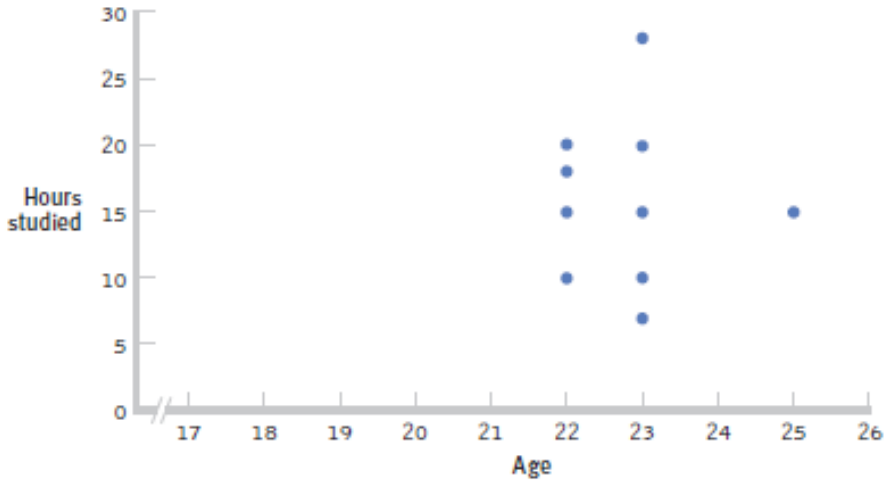
> Restricted Range.

A sample of boys and girls who performed in the top 2% to 3% on standardized tests - a much smaller range than the full population from which the researchers could have drawn their sample.



> Restricted Range, cont.

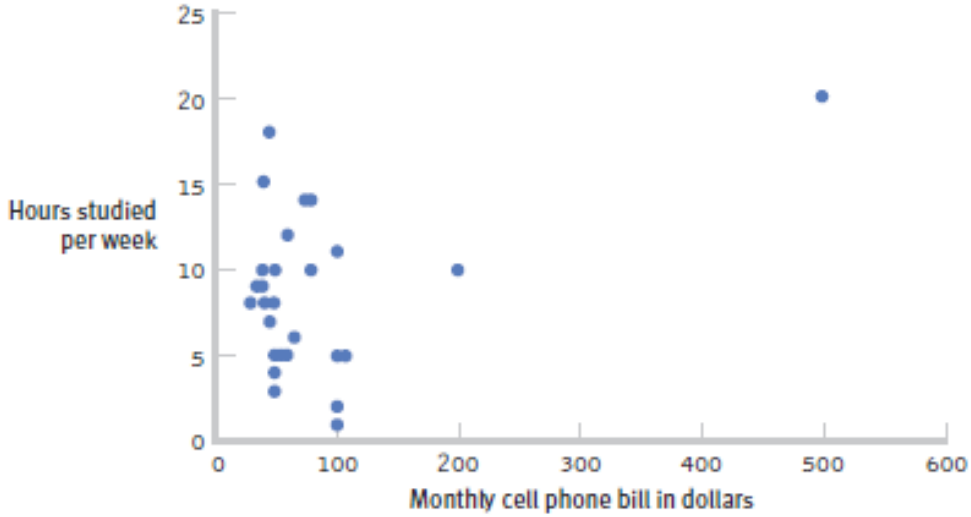
If we only look at the older students between the ages of 22 and 25, the strength of this correlation is now far smaller, just 0.05.



The Limitations of Correlation, cont.

> The effect of an outlier.

One individual who both studies and uses her cell phone more than any other individual in the sample changed the correlation from 0.14, a negative correlation, to 0.39, a much stronger and positive correlation



The Pearson Correlation Coefficient

- A statistic that quantifies a linear relation between two scale variables.
- Symbolized by the italic letter *r* when it is a statistic based on sample data.
- Symbolized by the italic letter *p* “rho” when it is a population parameter.

> Pearson correlation coefficient

- *r*
- Linear relationship

$$r = \frac{\sum[(X - M_X)(Y - M_Y)]}{\sqrt{(SS_X)(SS_Y)}}$$

Correlation Hypothesis Testing

- Step 1. Identify the population, distribution, and assumptions
- Step 2. State the null and research hypotheses.
- Step 3. Determine the characteristics of the comparison distribution.
- Step 4. Determine the critical values.
- Step 5. Calculate the test statistic
- Step 6. Make a decision.

Assumptions - Step 1

> Random selection

> X and Y are at least scale variables

> X and Y are both normal

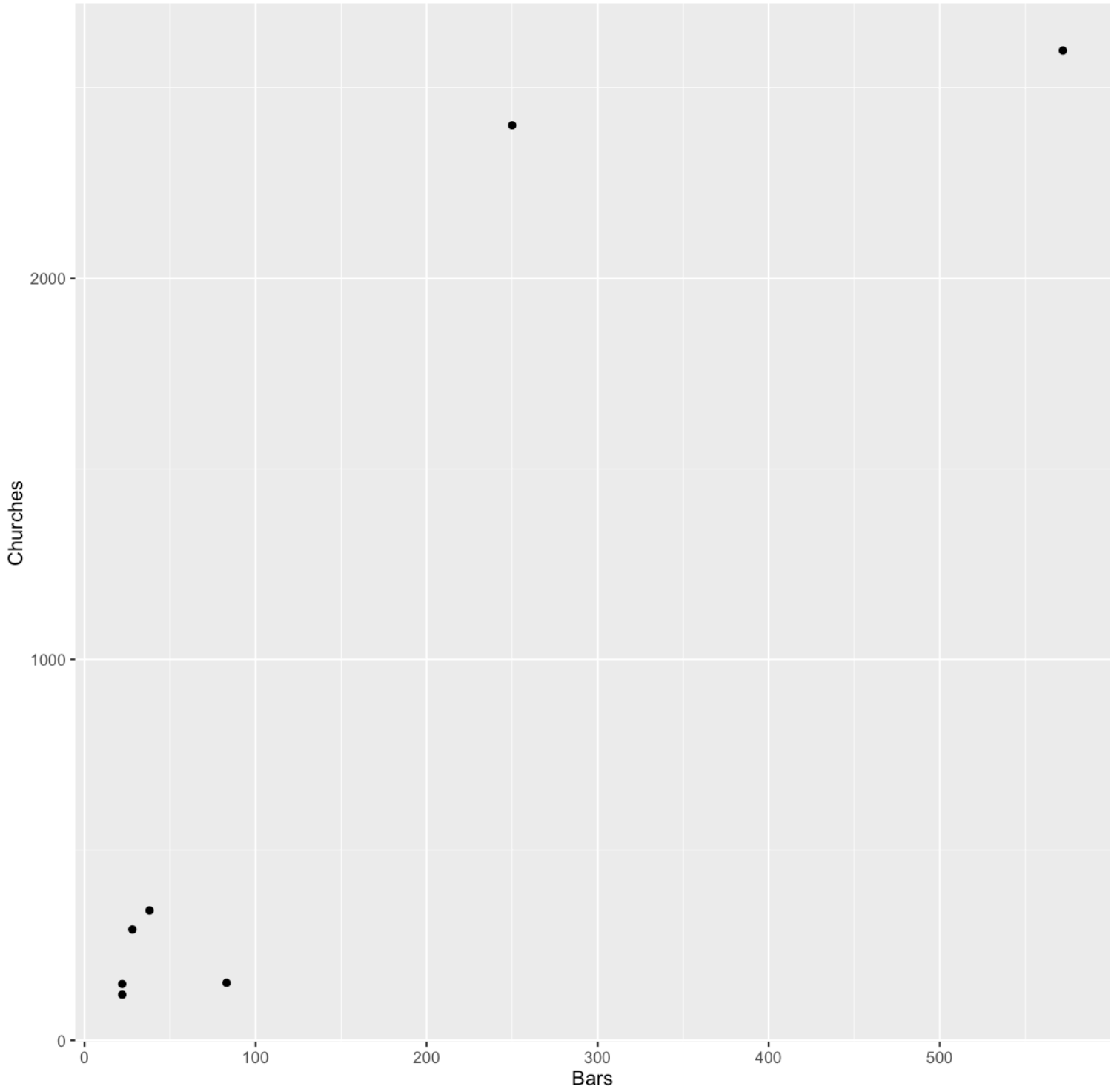
> Homoscedasticity (for real this time!)

- Each variable must vary approximately the same at each point of the other variable.
- See scatterplot for UFOs, megaphones, snakes eating dinner.

Assumptions – Step 1

> How to assess homoscedasticity?

- You can make a plot!
- Ggplot to the rescue!


Churches
2000
1000
0
0
100
200
300
400
500
Bars

Step 2

Population: no correlation between (var 1) and (var 2)

Sample: correlation between (var 1) and (var 2)

Hypothesis Steps

Step 2 (with our example):

Null: *r* for bars and churches = 0

Research: *r* for bars and churches /= 0

(can also use the rho symbol).

Step 3

```
        Pearson's product-moment correlation

data:  chapter15$Bars and chapter15$Churches
t = 4.7366, df = 5, p-value = 0.005166
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4740891 0.9859411
sample estimates:
      cor
0.9042978
```

Step 4

> (note: different from your book)

> Find the t-critical value using qt

- qt(.05/2, 5, lower.tail = F)
- + and – 2.57

Step 5

Calculate the *t*-found given your *r* value.

$$t = \frac{r}{\sqrt{(1 - r^2)/(N-2)}}$$

Or just use output: t found = 4.74

Step 6

> Reject the null!

- There is a significant correlation.

> If you fail to reject the null:

- There was not a significant relationship between the variables in this dataset.

Effect Size?

> *r* is often considered an effect size.

- Most people square it though to r^2, which is the same as ANOVA effect size.
- You can also switch *r* to Cohen's *d*, but people are moving away from doing this because *d* normally is reserved for nominal variables.

Confidence Intervals

> Involves Fisher's r to z transform and is generally pretty yucky.
  - Which means people do not do them unless they are required to.

> BUT cor.test gives them to you! ☺

Correlation and Psychometrics

- Psychometrics is used in the development of tests and measures.
- Psychometricians use correlation to examine two important aspects of the development of measures—reliability and validity.

Reliability

> A reliable measure is one that is consistent.

> Example types of reliability:

- Test–retest reliability.
- Split-half reliability

> Cronbach's alpa (aka coefficient alpha)

- Want to be bigger than .80.

Validity

- A valid measure is one that measures what it was designed or intended to measure.
- Correlation is used to calculate validity, often by correlating a new measure with existing measures known to assess the variable of interest.

> Correlation can also be used to establish the validity of a personality test.

> Establishing validity is usually much more difficult than establishing reliability.

> Buzzfeed!

Partial Correlation

- A technique that quantifies the degree of association between two variables after statistically removing the association of a third variable with both of those two variables.
- Allows us to quantify the relation between two variables, controlling for the correlation of each of these variables with a third related variable.

> We can assess the correlation between number of absences and exam grade, over and above the correlation of percentage of completed homework assignments with these variables.