Outline for Ch. 12 - Correlation

1. Overview of when to use correlation
2. **What are correlations, conceptually?**
   - null and alternative hypotheses
   - **calculation of Pearson's *r*, the *correlation coefficient***
   - interpretation of *r*
3. The coefficient of determination, R^2
4. Range restriction
5. Outliers
6. Correlation matrices

Variance vs. Covariance

- The variance tells us: how much scores of a ***single*** *variable* deviate from that variable's mean
- The covariance tells us: how much scores on *two* variables (X & Y) differ from *their respective means*.
  - For each participant, if their scores on X and Y deviate from X's and Y's means by the *same* amount – and if this is the case across all your Ps – the two variables are likely to be *related*.

Watch Commercials, Buy Candy?

Imagine that 5 people were asked how many advertisements for a certain candy they saw one week, and then researchers measured how many packets of that candy they purchased the next week . . .

TABLE 6.1

Subject	1	2	3	4	5	Mean	S
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

$$\text{Variance } (s^2) = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{N - 1}$$

$$Cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

The covariance is, broadly speaking, a measure of the **size of the relationship** between X and Y, how much of Y's variation is associated with X's variation

Calculating the correlation coefficient, *r*

Use the covariance - *cov(x,y)* - to calculate the correlation coefficient, *r*

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

standard deviation of variable X

standard deviation of variable Y

TABLE 6.1

Subject	1	2	3	4	5	Mean	S
Adverts Watched	5	4	4	6	8	**5.4**	**1.67**
Packets Bought	8	9	10	13	15	**11.0**	**2.92**

Trust me that the covariance = 4.25 here…

$$r = \frac{4.25}{1.67 \times 2.92}$$

$$r = .87$$

Practice – indicate true or false, & correct the false statements

1. Correlational analyses can be used to analyze data when researchers have measured two quantitative variables.
2. The covariance indicates how different the mean scores of two different variables are from each other.
3. Computing the correlation coefficient (*r*) involves dividing the covariance of variables X and Y, by the product of X's and Y's standard deviations.
4. $\rho \neq 0$ is a symbolic statement of the alternative hypothesis for a correlational analysis.
5. Higher standard deviations for X and Y mean a larger amount of error, or unsystematic variation, or "noise" in the data.

Practice – indicate true or false, & correct the false statements

1. Correlational analyses can be used to analyze data when researchers have measured two quantitative variables.
2. The covariance indicates how different the mean scores of two different variables are from each other.
3. Computing the correlation coefficient (*r*) involves dividing the covariance of variables X and Y, by the product of X's and Y's standard deviations.
4. $\rho \neq 0$ is a symbolic statement of the alternative hypothesis for a correlational analysis.
5. Higher standard deviations for X and Y mean a larger amount of error, or unsystematic variation, or "noise" in the data.

1. True
2. False - how much scores on *two* variables differ from *their respective means*
3. True
4. True
5. True

Outline for Ch. 12 - Correlation

1. Overview of when to use correlation
2. **What are correlations, conceptually?**
   - null and alternative hypotheses
   - calculation of Pearson's *r*, the *correlation coefficient*
   - **interpretation of *r*, including how NOT to interpret r**
3. The coefficient of determination, R^2
4. Range restriction
5. Outliers
6. Correlation matrices

Going back to candy ads and candy bought example, what does $r = .87$ mean?

- First, what are the possible values of *r*?

  $-1 \leq r \leq +1$

- *r* expresses the **direction** & strength (magnitude) of relationship btwn two variables
- Positive correlations ($r > 0$)
  - As x ↑, y ↑
  - As x ↓, y ↓

  } hours spent reading newspaper & knowledge of world events (assuming no fake news)
- Negative correlations ($r < 0$)
  - As x ↑, y ↓
  - As x ↓, y ↑

  } age of car and its resale value
  or
  rainfall amount and attendance at football games

How do we interpret the correlation coefficient, *r* ?

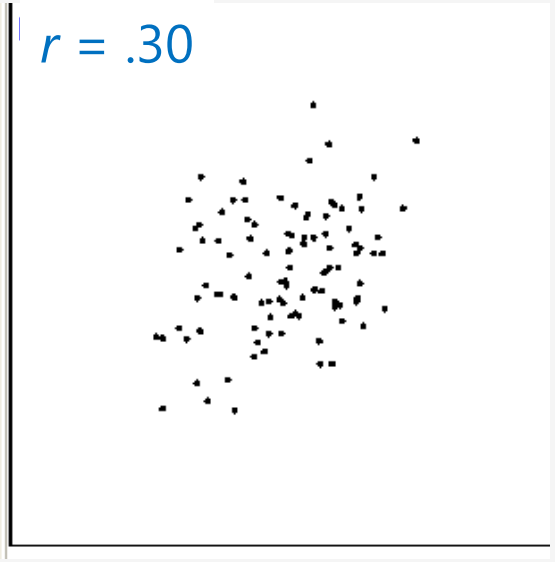
- *r* expresses the direction & **strength (magnitude)** of relationship btwn two variables
  - Examine the **absolute value** of *r* :
    - the larger it is (the closer to 1), the stronger the relationship.

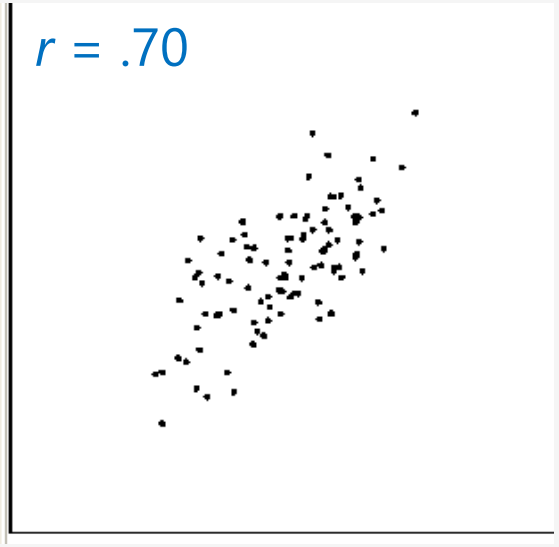
    (the smaller it is (the closer to 0), the weaker the relationship)

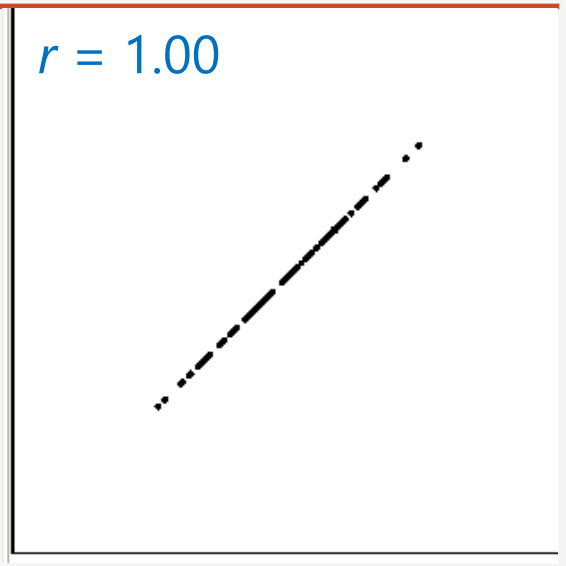
For which sample is the correlation *stronger*?
Sample 1: *r* = **-.85** ⇐
Sample 2: *r* = **.68**

Scatterplot depictions of correlations: direction & strength

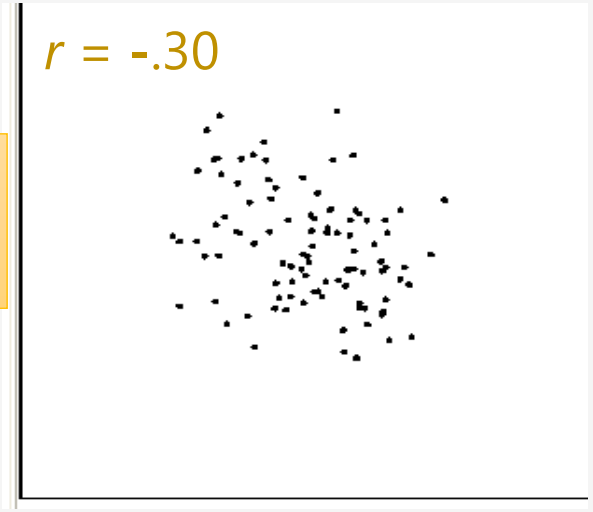
POSITIVE CORRELATIONS

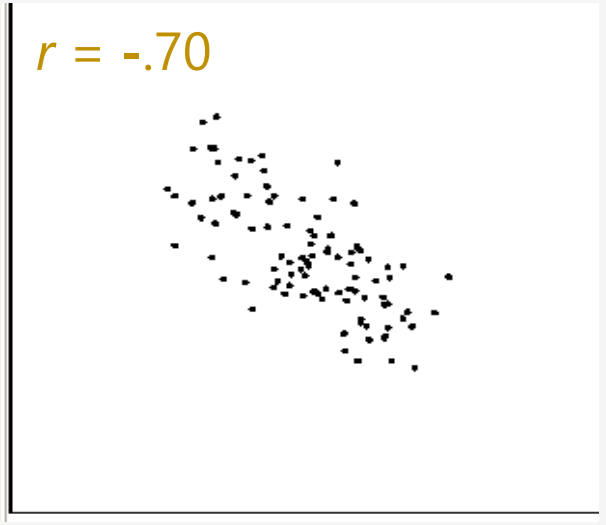


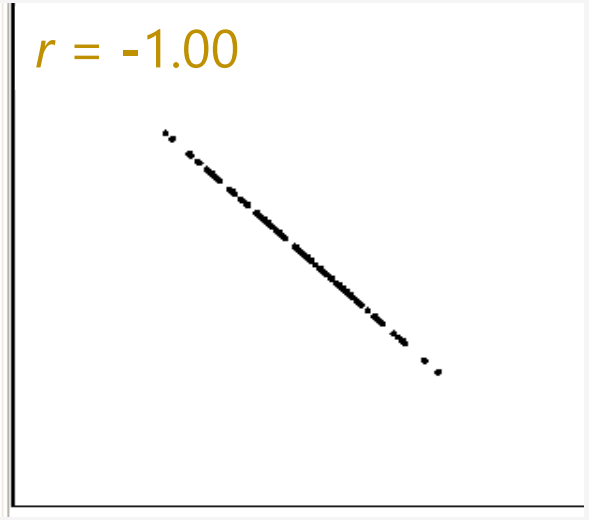




NEGATIVE CORRELATIONS



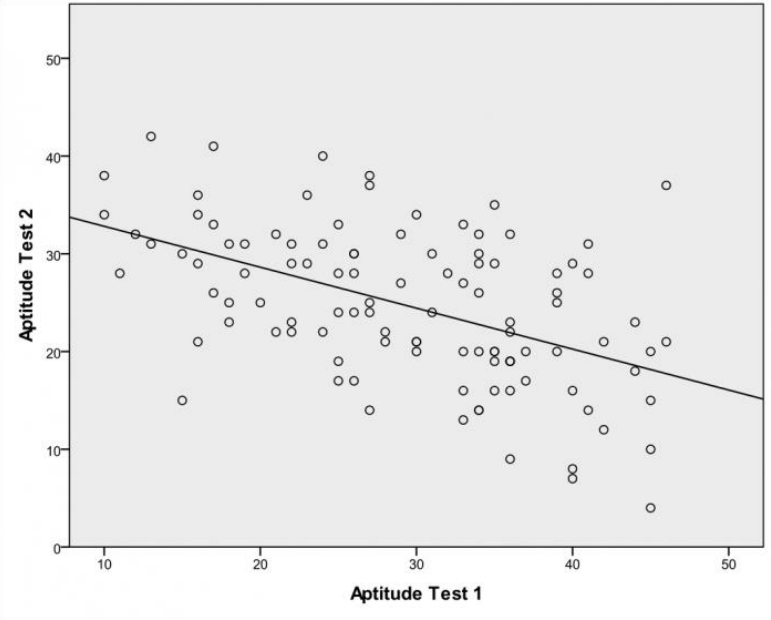




Practice – indicate true or false, & correct the false statements

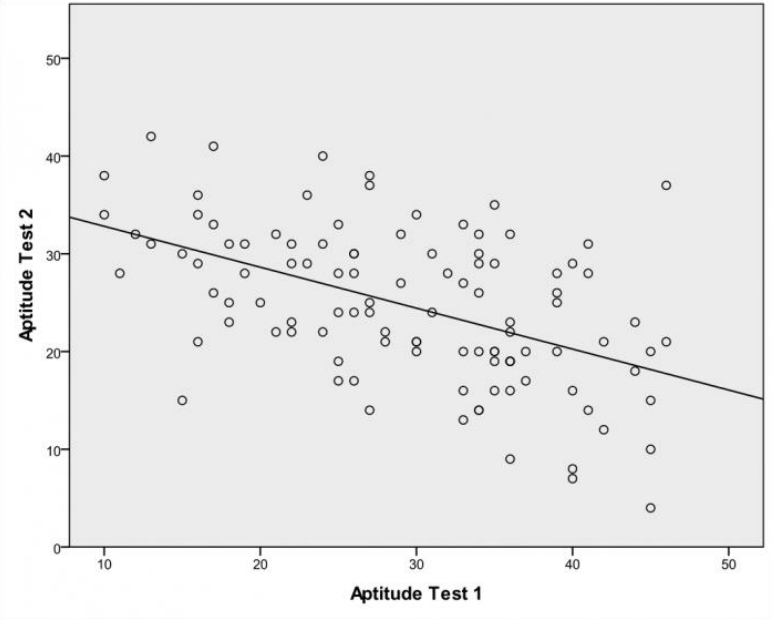
1. An *r* of **-.58** indicates a *stronger* relationship than an *r* of **.20**
2. If *r* = **-.84**, it suggests that the variables are not related.
3. As variable x goes down, variable y goes up; it is possible that this relationship produced a correlation coefficient, *r*, of .76.

4. The graph to the right displays a negative correlation.



Practice – indicate true or false, & correct the false statements

1. An *r* of **-.58** indicates a *stronger* relationship than an *r* of **.20**
2. If *r* = **-.84**, it suggests that the variables are not related.
3. As variable x goes down, variable y goes up; it is possible that this relationship produced a correlation coefficient, *r*, of .76.
4. The graph to the right displays a negative correlation.



1. True
2. False – they *are* related - negatively related. Only correlations close to zero are *not related*.
3. False – the *r* value must be a negative number (e.g., *r* = -.76)
4. True

How do we interpret the correlation coefficient, *r* ?

- *r* expresses the direction & **strength** of relationship btwn two variables
  - Examine the absolute value of *r* : the larger it is (the closer to 1), the stronger the relationship.
  - ***r*** can also be interpreted as an *effect size*

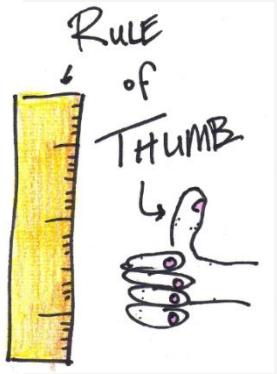
… of the relationship between two variables

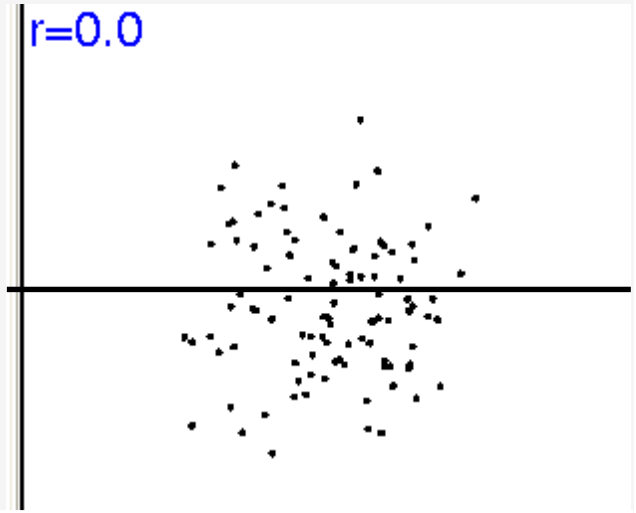
- Remember, *effect sizes* in statistics are…
  - standardized measures of the magnitude of an effect

    (*standardized* simply implies that effect sizes are comparable across studies, and across different types of variables)

How do we interpret the correlation coefficient, *r* ?

- *r* expresses the direction & **strength** of relationship btwn two variables
  - Examine the absolute value of *r* : the larger it is (the closer to 1), the stronger the relationship.
  - ***r*** is an example of an *effect size*
    - An *r* of 0 indicates *no systematic linear relationship (no effect)*
    - *Rough* guidelines for interpretation of *r* (its absolute value)
      - ***r*** ~ .10 = small effect (weak relationship)
      - ***r*** ~ .30 = medium effect (moderate relationship)
      - ***r*** ~ .50 or larger = large effect (strong relationship)





An *r* = .87 means the variables have a **strong positive** relationship.
Later on we will discuss how we know whether to reject or retain the null hypothesis based on the p-value associated with r.

Outline for Ch. 12 - Correlation

1. Overview of when to use correlation
2. **What are correlations, conceptually?**
   - null and alternative hypotheses
   - calculation of Pearson's *r*, the *correlation coefficient*
   - **interpretation of *r*, including how NOT to interpret *r***
3. The coefficient of determination, R^2
4. Range restriction
5. Outliers
6. Correlation matrices

correlational *designs* vs. correlational *analyses*

- A *design* refers to the procedures/method used to run a study (i.e., used to collect data).

- An *analysis* refers to how you analyze the data, once it's been collected.

Review of correlational *designs*

Q: How do researchers run a study with a correlational *design*?

A: Measure both (or all) of the variables (no manipulations)

Q: And *why* might researchers run a study with a correlational design, rather than a study with an *experimental* design?

A: When variables are:

- *naturally occurring* (e.g., age, whether or not person has schizophrenia)
- *unable to be manipulated for practical or ethical reasons* (e.g., level of marijuana use)

correlational *designs* vs. correlational *analyses*

- *Think Critically:* If you run a study with a correlational *design*, you will not always analyze the data with a correlational *analysis*. **Why not?**

- Studies with correlational designs may involve qualitative, measured variables. E.g.,
  - measure year in school (freshman, sophomore, jr, sr) and measure GPA to see if there are differences across groups → requires **analysis of variance (ANOVA)**

VS.

  - measure ACT score for entering freshmen and measure GPA at end of freshmen year to see if there is a relationship between ACT score and GPA → requires **correlational analysis**

How NOT to interpret *r* :

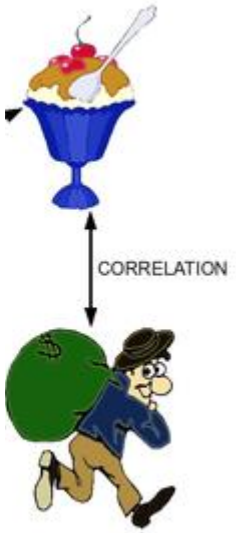
- . . . in terms of a causal relationship. Why not?

Correlational studies may be characterized by the …

1. **third variable problem**: there may be other measured or unmeasured variables affecting the relationship between X and Y.

The value of *r* says nothing about whether third variables (confounds) exist.

EX: There is a positive correlation between ice-cream sales & armed robberies (as sales go up, # of armed robberies go up).



How NOT to interpret *r* :

- . . . in terms of a causal relationship. Why not?

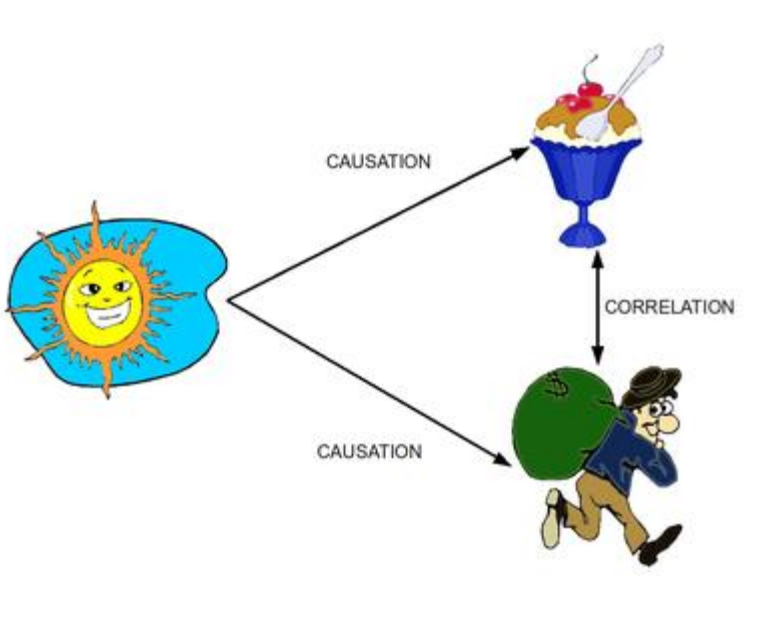
Correlational studies may be characterized by the …

1. **third variable problem**: there may be other measured or unmeasured variables affecting the relationship between X and Y.

The value of *r* says nothing about whether third variables (confounds) exist.

EX: There is a positive correlation between ice-cream sales & armed robberies (as sales go up, # of armed robberies go up.

This relationship may be attributable to a confound (aka "third variable") – **hotter temperatures** – which causes both ice-cream sales & armed robberies to rise.



How NOT to interpret *r* :

- . . . in terms of a causal relationship. Why not?

Correlational studies may be characterized by the . . .

2. **directionality problem**: even if there were to be a causal relationship between X and Y, the value of *r* says nothing about which variable causes the other to change
   - EX: Suppose teens' IQ is positively correlated with # of years spent in school.
     - Does higher IQ cause teens to spend more years in school, or vice-versa?

How NOT to interpret *r* :

- Take-home point: Since studies with correlational designs (which require correlational analyses) cannot establish causality, make sure not to use cause-effect language in your hypotheses or interpretations.

- ***Say:* "More commercials watched is associated with more candy bought"**
- ***Say:* "Number of commercials watched is positively related to amount of candy bought"**
- *Rather than:* "Watching more commercials *causes* people to buy more candy" *or* "Watching commercials has a positive *effect on* candy purchased"

Outline for Ch. 12 - Correlation

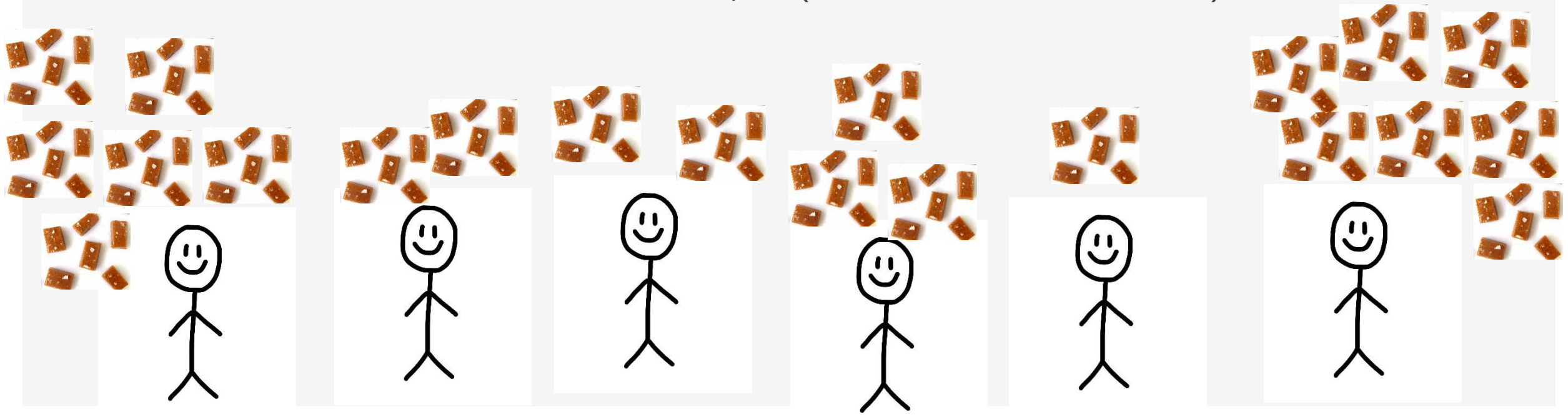
1. Overview of when to use correlation
2. What are correlations, conceptually?
   - null and alternative hypotheses
   - calculation of Pearson's *r*, the *correlation coefficient*
   - interpretation of *r,* including how NOT to interpret *r*
3. **The coefficient of determination, R^2 (r^2)**
4. Range restriction
5. Outliers
6. Correlation matrices

Amount of variation shared

- Back to example . . . The correlation between # of candy ads watched and # of packages of candy bought is $r = .87$
  - What % of the total variation in the # of packages people bought ***is shared by (is explained by)*** the # of ads they watched?

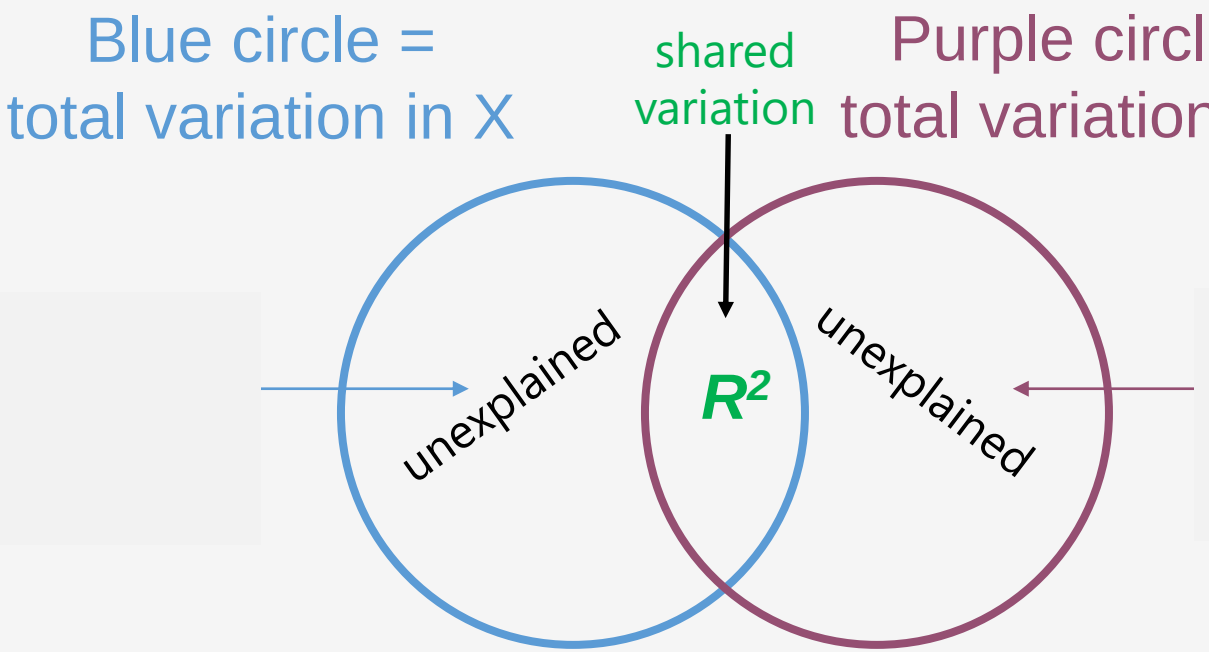
    → coefficient of determination, R^2 (book uses lowercase r^2)

Should sound similar to <u>systematic variation</u>



The coefficient of determination, R^2, tells us . . .

- how much of the total variation in one variable can be explained by knowing its relationship to the other variable (vs. left unexplained)
  - R^2 tells us the % of the variation in one variable (e.g., Y) that is shared by the other variable (e.g., X)
- calculate R^2 by squaring the correlation coefficient, *r*

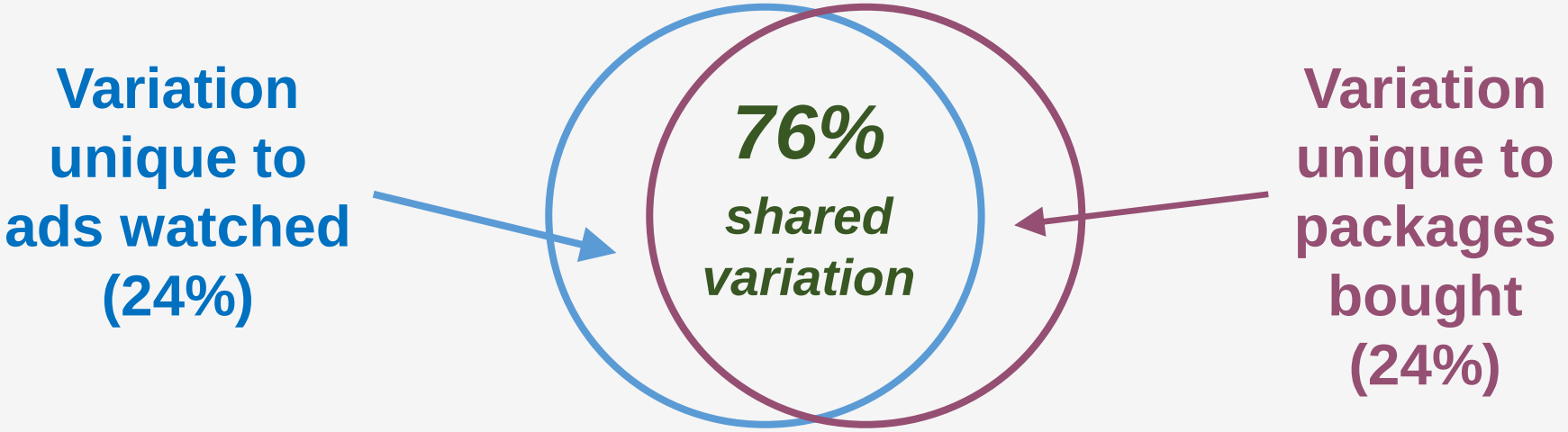


coefficient of determination, R^2 – example 1

- Back to example . . . The correlation between # of candy ads watched and # of packages of candy bought is r =.87.
- $R^2 = (.87)^2 = .87 * .87 =$ **.76**
  - 76% of the variation in the # of packages people bought is shared by the # of ads they watched

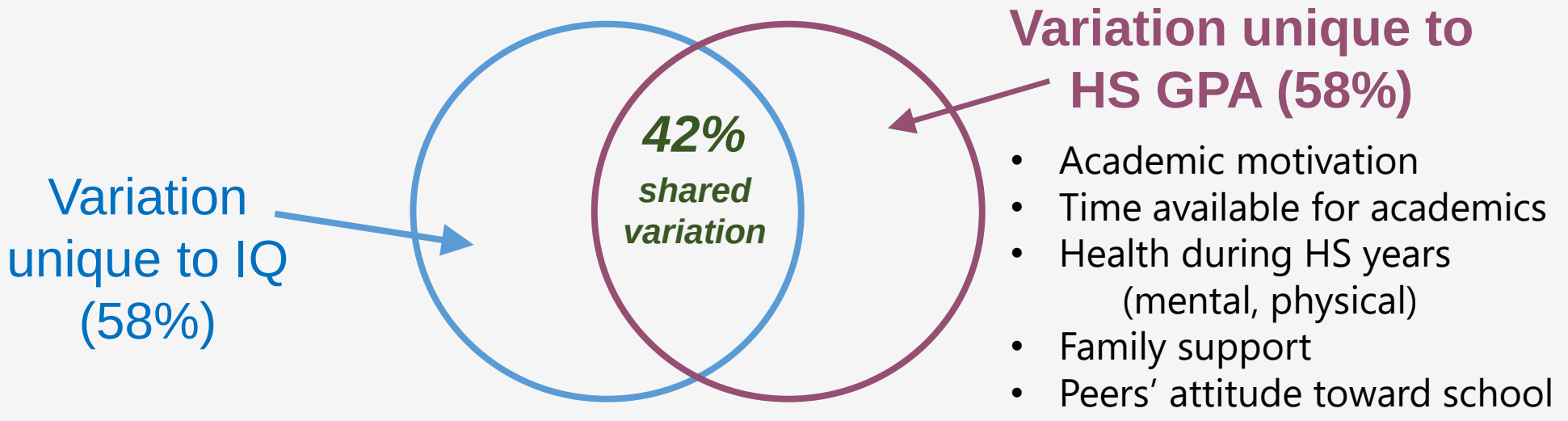
variable Y

variable X



coefficient of determination, R^2 – example 2

- Suppose students took an IQ test prior to entering high school (HS). We also have a measure of their graduating HS GPA.
- The variables are positively correlated, but not perfectly. $r = .65$, $R^2 = .42$
  - 42% of the variance in students' HS GPA is shared by their IQ scores.
  - What might help explain the 58% of variance in HS GPA that is not explained by IQ score?



Practice

Q1. How much variance in Y has been explained by X, if $r = .40$; and what is this calculation called?

A. 2%; correlation coefficient

B. 16%; coefficient of determination

C. 40%; correlation coefficient

D. 80%; coefficient of determination

Q2 Suppose a study showed that interest in Greek life is related to marijuana use, such that **higher interest in Greek life is related to less interest in marijuana**. Which of the following *may* be true, in reality, given the info above?

A. Using marijuana makes students less likely to participate in Greek life.

B. A third variable – for example, personality of student – influences both; certain personalities are both more likely to have interest in Greek life and less likely to be interested in marijuana use.

C. Joining a fraternity/sorority makes students less likely to use marijuana.

D. Any of the above may be true given the information in the question.

Q1. B
Q2. D

Outline for Ch. 12 - Correlation

1. Overview of when to use correlation
2. What are correlations, conceptually?
   - null and alternative hypotheses
   - calculation of Pearson's *r*, the *correlation coefficient*
   - interpretation of *r,* including how NOT to interpret *r*
3. The coefficient of determination, R^2 (r^2)
4. **Range restriction**
5. **Outliers**
6. **Correlation matrices**

Range restriction

rho, the correlation in the population

- *r* may severely misrepresent ρ when your sample has a restricted range for one or both measured variables

- Ex. Suppose HS GPA and college GPA are correlated in the population at ρ = .40, and you collect data from 8 college students…

ID#	X (high school GPA)	Y (college GPA)
1	4.0	2.7
2	3.9	2.9
3	4.0	1.6
4	3.8	3.0
5	4.0	2.2
6	4.0	3.5
7	4.0	3.8
8	3.8	1.9