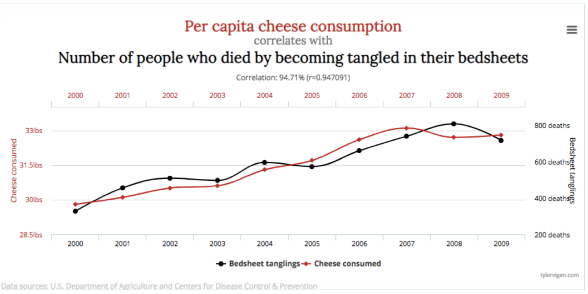
Pick Up the Warm Up



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

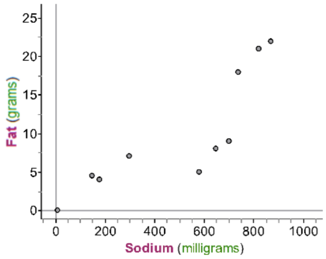
What four characteristics should you consider when interpreting a scatterplot?

DUFS

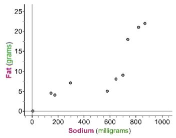
- *Direction: positive, negative, none*
- *Form: linear, non-linear*
  - *Don't let one or two points sway you!*
- *Strength: how closely the points follow the form*
- *Unusual features*
  - *Outliers: outside the overall pattern (no specific rule)*
  - *Clusters*

ALWAYS IN CONTEXT!!

2. The scatterplot shows the amount of sodium (in milligrams) and amount of fat (in grams) in salads from McDonalds (with no dressing). Describe the relationship between sodium and fat.

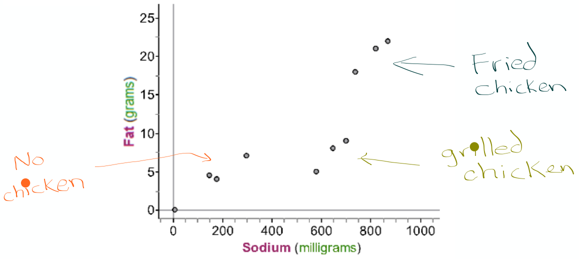


Description



There is a moderately strong, positive, linear association between Sodium and fat in salads at McDonalds. There are three distinct clusters.

2. The scatterplot shows the amount of sodium (in milligrams) and amount of fat (in grams) in salads from McDonalds (with no dressing). Describe the relationship between sodium and fat.



Which of the following examples shows causation?

none necessesarily

Correlation does not mean Causation

This is so important because so often, attention seeking media will infer a causal relationship to make a better story or...

or a company might try to sell their products by inferring a causal relationship

TARGETS

- Understand the basic properties of correlation, including how the correlation is influenced by outliers.
- Distinguish correlation from causation.

Just watch for now

For a linear association between two quantitative variables, the **correlation *r*** measures the direction and strength of the association.

CAUTION:

It is only appropriate to use the correlation to describe strength and direction for a ***linear*** relationship.

Measuring Linear Association: Correlation (pages 160–162)

- Guess the Correlation Activity

Teams

Round 1 A → B → C → etc

Round 2 E → D → C → B → A

Pick Up

AP Stats – Class Notes 3.1 Day 2 – Measuring Correlation

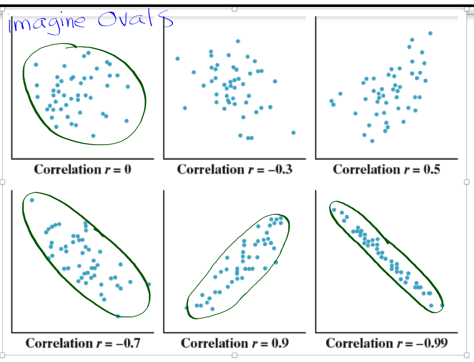
1. Correlation addresses Direction and Strength of a Linear Association. It does not address Form !

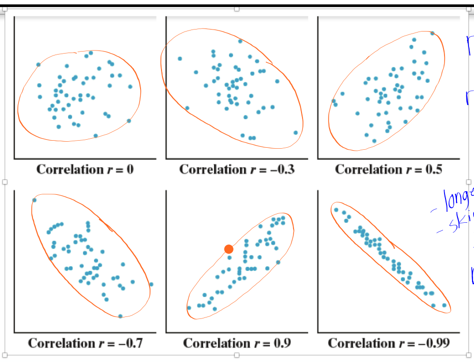
It is used only for 2 quantitative variables --- otherwise use "association".

Some Important Properties of the Correlation *r*

- The correlation *r* is always a number between –1 and 1 ($-1 \le r \le 1$).
- The correlation *r* indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association.
- The extreme values $r = -1$ and $r = 1$ occur *only* in the case of a perfect linear relationship, when the points lie exactly along a straight line.
- If the linear relationship is strong, the correlation *r* will be close to 1 or –1.
- If the linear relationship is weak, the correlation *r* will be close to 0.

Can imagine Ovals

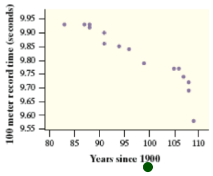




round
↓
$r \approx 0$

- longer
- skinny
↓
$r \approx \pm 1$

3. Here is a scatterplot that shows the relationship between the years since 1900 and the 100-meter sprint record time (in seconds) for the years 1983 to 2010. For these data, $r = -0.927$. Interpret the value of r.



The correlation of r = -.927 confirms that the linear association between years since 1900 and 100-meter record time is strong and negative

↑ strength ↑ direction

Cautions about Correlation

(pages 163–166)

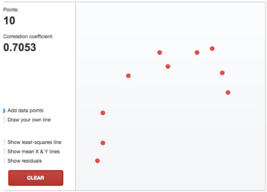
- Correlation doesn't imply causation

p. 163-164 ✓

Activity

~ Laptop

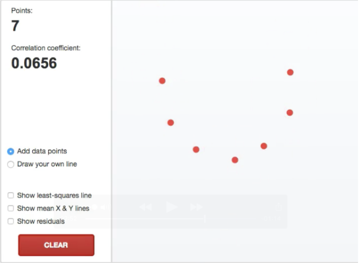
Develop important understanding about aspects of correlation.



Points: 10

Correlation coefficient: 0.7053

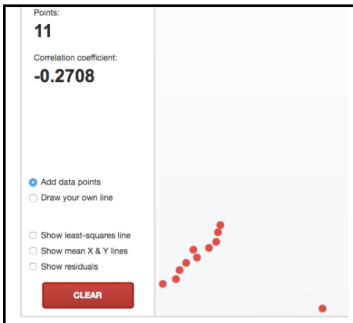
Correlation does not tell us anything about the form of a relationship.



Points: 7

Correlation coefficient: 0.0656

there may be a very strong **Non-linear** relationship but there is virtually no linear relationship.



Points: 11

Correlation coefficient: -0.2708

A single point can have a very large effect on correlation.

Page 163 Activity

4. Summary

Correlation doesn't ____________________ →

Correlation doesn't measure ____________

Correlation should only be used to describe a linear association.

Correlation ____________ a resistant measure of strength.

Correlation is just a _________________ to a scatter plot. --- Don't start with correlation

(start with a picture!)

- Correlation doesn't imply causation
- Correlation doesn't measure form
- Correlation should only be used to describe a linea association
- Correlation isn't a resistant measure of strength
- Correlation is just a supplement to a scatterplot—don't start with correlation

5. Which association is more linear: one with $r = 0.50$ or one with $r = 0.90$?

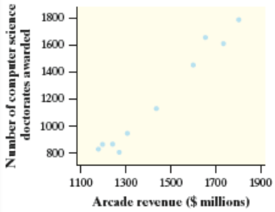
You CAN'T determine the form of a relationship using only r.

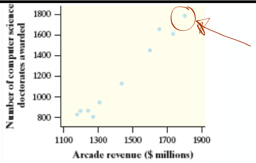
(r is a supplement for a graph, not a replacement)

Cautions About Correlation

6. Does playing video games help you get a degree?

Computer science majors are often gamers, frequenting the local arcades. Are the arcade games helping them toward earning a doctorate in computer science? Below is a scatterplot created using U.S. data for the years 2000 to 2009. The explanatory variable is the arcade revenue (in millions of dollars) and the response variable is the number of doctoral degrees awarded in computer science.





(a) Will playing lots of video games at the arcade make you more likely to get a computer science doctorate? Explain your answer.

NO; even though there is a strong correlation between arcade revenue and # of computer doctorites, causation should not be inferred. It may be that both of these variables are changing due to another variable such as overall progress of technology.

(b) What effect does the year 2008 ($1803 million in arcade revenue and 1787 computer science doctorates awarded) have on the correlation? Explain your answer.

Because the point for 2008 is in the positive linear pattern of the rest of the points, it makes the correlation closer to 1.

- To calculate or not to calculate?

HOW TO CALCULATE THE CORRELATION r

Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between x and y is

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_x}\right)\left(\frac{y_1-\bar{y}}{s_y}\right) + \left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right) + \cdots + \left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

or, more compactly,

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$= \frac{1}{n-1}\sum z_x z_y$$

Calculating Correlation, Additional Facts

(pages 166–169)

Notes from 3.1 Day 2

October 01, 2018

1. Correlation requires that both variables be quantitative.

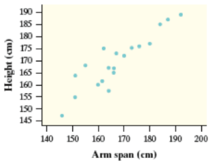
1. Correlation requires that both variables be quantitative.
2. Correlation makes no distinction between explanatory and response variables.

1. Correlation requires that both variables be quantitative.
2. Correlation makes no distinction between explanatory and response variables.
3. *r* does not change when we change the units of measurement of *x*, *y*, or both.

1. Correlation requires that both variables be quantitative.
2. Correlation makes no distinction between explanatory and response variables.
3. *r* does not change when we change the units of measurement of *x*, *y*, or both.
4. The correlation *r* has no unit of measurement. It's just a number.

9. Arm span versus height

The following scatterplot shows the arm span (in centimeters) and height (in centimeters) for a random sample of 19 twelfth graders. The correlation is r = 0.902.



(a) Explain why it is incorrect to say that the correlation is 0.902 centimeter.

Because correlation is calculated using standardized values, it does not have units.

$\frac{\text{Value-mean}}{\text{SD}}$

(b) What would happen to the correlation if height was measured in inches instead of centimeters? Explain your answer.

The correlation would be the same. Because it is calculated with standardized values, changes of units don't affect correlation

(c) What would happen to the correlation if height was used as the explanatory variable and arm span was used as the response variable?

The correlation would be the same because correlation doesn't make a distinction between explanatory and response variables.

Why r ?

↑ Karl Pearson Pearson's Correlation Coeff.

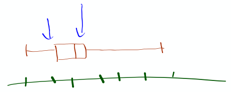
Why not c ?

- Actually originally developed by the cousin of Charles Darwin, Francis Galton.
- He studied linear regression
↑

Ch. 2 Test

Tidbits

Outliers are based on 5-number summary (not the mean)



Lower $Q_1 - 1.5(IQR)$

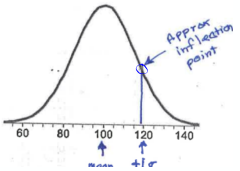
Upper $Q_3 + 1.5(IQR)$

Free Response

You don't have to interpret unless asked but all answers should be in context.

9. Estimate the standard deviation of the Normal density curve to the right.

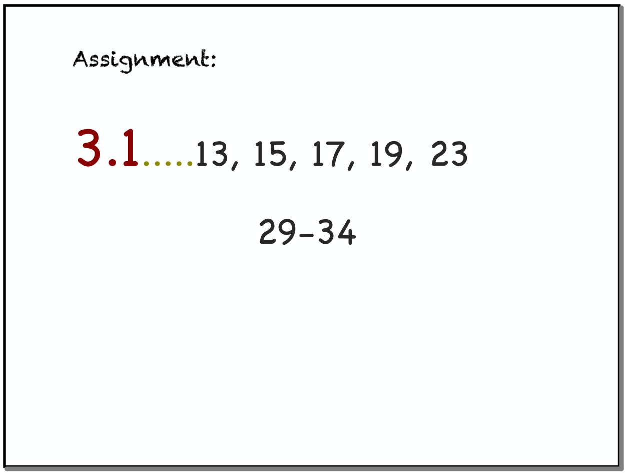
a) 5
b) 10
c) 15 ←
[✓ d) 20]
e) 25



Rubric for #17 (Pink Sheet)

M/C $11/15 = 73\%$ → 84% ↘ $\boxed{79.5\%}$

F/R avg $\boxed{2.0}$ → 75% ↗



Assignment:

3.1.....13, 15, 17, 19, 23

29-34