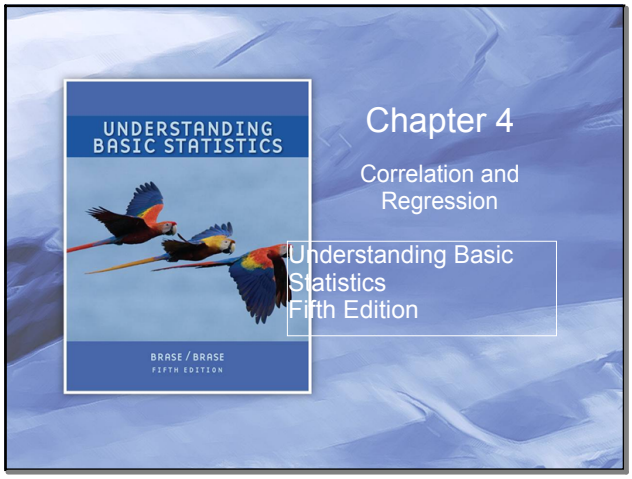
Login your clickers & need calculators.

Have out the 4.1 vocabulary to collect.

ABSENT STUDENTS:
Need: pg. 93-98 & pg. 109 - 111

You can make-up the test:
1) LUNCH & homeroom
2) AFTER-School

Apr 20-7:34 AM



Oct 8-11:45 AM

Scatter Diagrams

- A graph in which pairs of points, (x, y), are plotted with x on the horizontal axis and y on the vertical axis.
- The explanatory variable is x.
- The response variable is y.
- One goal of plotting paired data is to determine if there is a linear relationship between x and y.

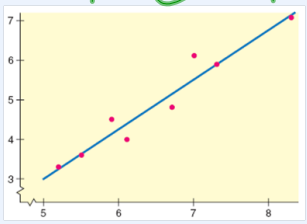
Oct 8-11:45 AM

Paired Data (x, y)

Important Questions

How strong is the linear correlation between x and y?

What line best represents the data?

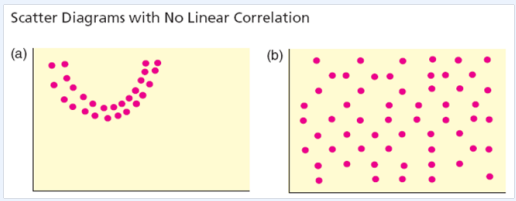


Oct 8-11:45 AM

How Strong Is the Linear Correlation?

Not all relationships are linearly-correlated.

Scatter Diagrams with No Linear Correlation



Statisticians need a quantitative measure of the strength of the linear association.

Oct 8-11:45 AM

The Sample Correlation Coefficient *r*

Statisticians use the sample correlation coefficient *r* to measure the strength of the linear correlation between paired data.

- *r* has no units.
- $-1 \le r \le 1$
- $r > 0$ indicates a positive relationship between *x* and *y*, $r < 0$ indicates a negative relationship.
- $r = 0$ indicates no *linear* relationship.
- Switching the explanatory variable and response variable does not change *r*.
- Changing the units of the variables does not change *r*.

Oct 8-11:45 AM

A Computational Formula for *r*

Obtain a random sample of *n* data pairs (*x*, *y*). The data pairs should have a *bivariate normal distribution*. This means that for a fixed value of *x*, the *y* values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed *y*, the *x* values should have their own (approximately) normal distribution.

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy.

2. With *n* = sample size, Σx, Σy, Σx^2, Σy^2, and Σxy, you are ready to compute the sample correlation coefficient *r* using the computation formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Be careful! The notation Σx^2 means first square *x* and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the *x* values, then square the result.

Oct 8-11:45 AM

Illustration

Caribou (*x*, in hundreds) and wolf (*y*) populations

x	*y*	x^2	y^2	*xy*
70	3	4900	9	210
115	45	13,225	2025	5175
105	21	11,025	441	2205
82	7	6724	49	574
93	16	8649	256	1488
125	62	15,625	3844	7750
88	12	7744	144	1056
$\Sigma x = 678$	$\Sigma y = 166$	$\Sigma x^2 = 67{,}892$	$\Sigma y^2 = 6768$	$\Sigma xy = 18{,}458$

Oct 8-11:45 AM

Illustration

Caribou (x, in hundreds) and wolf (y) populations

x	y	x^2	y^2	xy
70	3	4900	9	210
115	45	13,225	2025	5175
105	21	11,025	441	2205
82	7	6724	49	574
93	16	8649	256	1488
125	62	15,625	3844	7750
88	12	7744	144	1056
$\Sigma x = 678$	$\Sigma y = 166$	$\Sigma x^2 = 67{,}892$	$\Sigma y^2 = 6768$	$\Sigma xy = 18{,}458$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \quad (2)$$

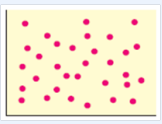
$$= \frac{7(18{,}458) - (678)(166)}{\sqrt{7(67{,}892) - (678)^2}\sqrt{7(6768) - (166)^2}} \approx \frac{16{,}658}{(124.74)(140.78)} \approx 0.949$$

Oct 8-11:45 AM

Interpreting the Value of r

$r = 0$

There is no linear relation for the points of the scatter diagram.

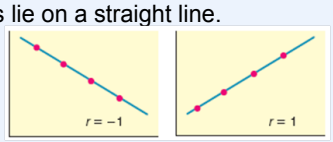


Oct 8-11:45 AM

Interpreting the Value of r

$r = 1$ or $r = -1$

There is a perfect linear relation between x and y; all points lie on a straight line.

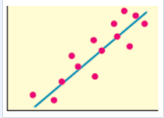


Oct 8-11:45 AM

Interpreting the Value of r

$0 < r < 1$

The x and y values has a *positive correlation*. As x increases, y tends to increase.

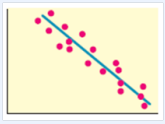


Oct 8-11:45 AM

Interpreting the Value of *r*

$-1 < r < 0$

The *x* and *y* values have a negative correlation. As *x* increases, *y* tends to decrease.



Oct 8-11:45 AM

1 Which of the following shows a strong negative correlation?

Answer

A

a).

b).

c).

d).

Oct 8-11:45 AM

Critical Thinking

r = sample correlation coefficient computed from a random sample of (x, y) data pairs.

ρ = population correlation coefficient computed from all population data pairs (x, y).

- Expect *r* to vary from sample to sample.
- So, consider the *significance* of *r* as well as its value when assessing the strength of a linear correlation. (Section 11.4)

Oct 8-11:45 AM

Critical Thinking

- $|r| \approx 1$ only implies a linear relationship between *x* and *y*.
- It **does not** imply a *cause and effect* relationship between *x* and *y*.
- The values of *x* and *y* may both depend linearly on some third *lurking variable*.

Oct 8-11:45 AM

Critical Thinking

Over the past few years, there has been a strong positive relationship between the annual consumption of coffee and the number of computers sold per year. Which conclusion is the best one to draw from this strong correlation?

Answer?

a). Coffee consumption stimulates computer sales.
b). Computer users are sophisticated and thus are inclined to drinking coffee.
c). The correlation is purely accidental.
d). The responses of both variables probably reflect the increasing wealth of the citizenry.

Oct 8-11:45 AM

Have out your 4.1 vocabulary

Make a scatter plot for Example 1 and answer the questions.

Make a scatter plot for Example 2 and answer the questions.

Oct 9-10:39 AM

4.1 pg. 37 true/false

Grade: «12»
Subject: Statistics
Date: «10-9-12»

Oct 9-10:38 AM

1 Answer?

True

False

Oct 9-10:35 AM

2 Answer?

True

False

Oct 9-10:36 AM

3 Answer?

True

False

Oct 9-10:36 AM

4 Answer?

True

False

Oct 9-10:36 AM

Linear Regression

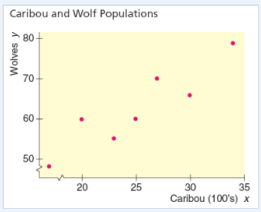
- Linear Regression - a mathematical technique for creating a linear model for paired data.
- Based on the “least-squares” criterion of best fit.

Oct 8-11:45 AM

Caribou and wolf populations in Denali National Park

Questions

- Do the data points have a linear relationship?
- How do we find an equation for the best fitting line?
- Can we predict the value of the response variable for a new value of the predictor variable?
- What fractional part of the variability in *y* is associated with the variability in *x*?

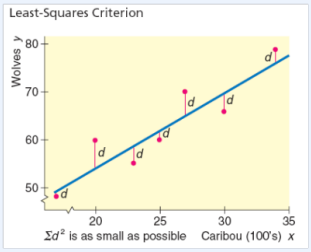


Oct 8-11:45 AM

Least-Squares Criterion

Least-squares criterion

The sum of the squares of the vertical distances from the data points (x, y) to the line is made as small as possible.



Oct 8-11:45 AM

How to find the equation for the least-squares line $\hat{y} = a + bx$

Obtain a random sample of n data pairs (x, y), where x is the *explanatory variable* and y is the *response variable*.

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy. Then compute the sample means $\bar{x}$ and $\bar{y}$.
2. With n = sample size, Σx, Σy, Σx^2, Σy^2, Σxy, $\bar{x}$, and $\bar{y}$, you are ready to compute the slope b and intercept a using the computation formulas

$$\text{Slope:} \quad b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \tag{3}$$

$$\text{Intercept:} \quad a = \bar{y} - b\bar{x} \tag{4}$$

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values, then square the result.

Oct 8-11:45 AM

3. The equation of the least-squares line computed from your sample data is

$$\hat{y} = a + bx \tag{5}$$

Note: Inferences for the population slope (Section 11.4) require the data pairs to have a *bivariate normal distribution*. That is, for a fixed value of x, the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed value of y, the x values should have their own (approximately) normal distribution. Chapter 6 discusses normal distributions.

Oct 8-11:45 AM

Properties of the Regression Equation

- The point $(\bar{x}, \bar{y})$ is always on the least-squares line.
- The slope tells us the amount that *y* changes when *x* increases by one unit.

Oct 8-11:45 AM

Illustration

Caribou (*x*, in hundreds) and wolf (*y*) populations

x	30	34	27	25	17	23	20
y	66	79	70	60	48	55	60

x	*y*	x^2	y^2	*xy*
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\Sigma x = 176$	$\Sigma y = 438$	$\Sigma x^2 = 4628$	$\Sigma y^2 = 28{,}026$	$\Sigma xy = 11{,}337$

Oct 8-11:45 AM

Illustration

x	*y*	x^2	y^2	*xy*
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\Sigma x = 176$	$\Sigma y = 438$	$\Sigma x^2 = 4628$	$\Sigma y^2 = 28{,}026$	$\Sigma xy = 11{,}337$

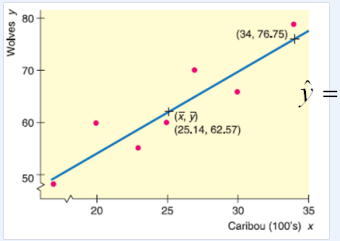
$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(11{,}337) - (176)(438)}{7(4628) - (176)^2} = \frac{2271}{1420} \approx 1.60$$

$$a = \bar{y} - b\bar{x} \approx 62.57 - 1.60(25.14) \approx 22.35$$

Oct 8-11:45 AM

Illustration

Least-squares linear relationship between caribou and wolf populations:



$$\hat{y} = 22.35 + 1.60x$$

Oct 8-11:45 AM

Critical Thinking: Making Predictions

- We can simply plug in *x* values into the regression equation to calculate *y* values.

Predicting $\hat{y}$ values for x values that are **between** observed x values in the data set is called **interpolation.**

Predicting $\hat{y}$ values of x values that are **beyond** observed x values in the data set is called **extrapolation.**

- Extrapolation may produce unrealistic forecasts.

Oct 8-11:45 AM

Coefficient of Determination

- Another way to gauge the fit of the regression equation is to calculate the coefficient of determination, r^2.

1). Compute *r*. Simply square this value to get r^2.
2). r^2 is the fractional amount of total variation in *y* that can be explained using the linear model.
3). $1 - r^2$ is the fractional amount of total variation in *y* that is due to random chance (or possibly due to lurking variables).

Oct 8-11:45 AM

Coefficient of Determination

The linear correlation coefficient for a set of paired data is $r = 0.86$.

What fractional amount of the total variation in *y* is due to random chance and/or to lurking variables?

a). 0.86 b). 0.14 c). 0.74 d). 0.26

Oct 8-11:45 AM

Coefficient of Determination

The linear correlation coefficient for a set of paired data is $r = 0.86$.

What fractional amount of the total variation in *y* is due to random chance and/or to lurking variables?

a). 0.86 b). 0.14 c). 0.74 d). 0.26

Oct 8-11:45 AM

Oct 9-4:03 PM