Review of Standard Deviation

Population (N observations) Sample (sample size n)

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \qquad s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

μ = mean $\bar{x}$ = mean

1

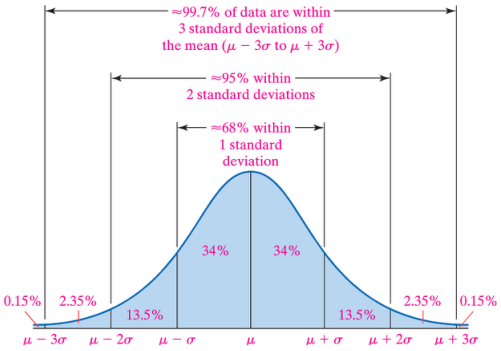
(1) Why $n-1$ in the sample formula, not n?

Because the population formula when applied to samples tends to give lower values for the std. dev. than the actual value.

2 See the data on the right.

TEST 1	Scores	Sample
	16.5	16.5
	19.0	20.0
	14.0	19.0
	21.0	20.0
	20.0	19.0
	12.0	16.0
	16.0	20.0
	19.0	16.0
	18.0	18.0
	19.0	16.0
	20.0	
	21.0	
	19.0	
	16.0	
	20.0	
	21.0	
	20.0	
	20.0	
	18.0	
	21.0	
	16.0	
	17.0	
	20.0	
	16.5	
	15.0	
	18.0	
	17.0	
	21.0	
	16.0	
	18.0	
Mean	18.2	18.1
Pop. Std Dev.	2.3	1.7
Samp. Std Dev.		1.8

Where are most of the data?



Empirical Rule: Page 138, fig 13.

3

The standard deviation (sd) of data is a measure of how dispersed the values are.

The larger the sd, relative to the mean, the more spread out the data are from it.

The smaller the sd, relative to the mean, the more sharply peaked the data are around it.

4

Examples

(2) What is the mean (μ) and the standard deviation (σ) of the data set below?

$\{1, 1, 1, 1, 1, 1\}$ $\mu =$ 1. $\sigma =$ 0.

(3) What is the mean (μ) and the standard deviation (σ) of the data set below?

$\{1, 1, 1, -1, -1, -1\}$ $\mu =$ 0. $\sigma =$ 1.

5

(4) What is the mean (μ) and the standard deviation (σ) of the data set below?

$$\{4, 4, 4, 2, 2, 2\}$$

$(=\{1+3, 1+3, 1+3, -1+3, -1+3, -1+3\})$

$\mu =$ 3. $\sigma =$ 1.

6

ADDITIONAL NOTES

(5) What is the mean (μ) and the standard deviation (σ) of the data set below?

$$\{20, 20, 20, 10, 10, 10\}$$

$(=\{5\times4, 5\times4, 5\times4, 5\times2, 5\times2, 5\times2\})$

$\mu =$ $5 \times 3 = 15.$ $\sigma =$ $5 \times 1 = 5.$

7

(6) What is the mean (μ) and the standard deviation (σ) of the data set below?

$$\{-20, -20, -20, -10, -10, -10\}$$

$(=\{-5\times4, -5\times4, -5\times4, -5\times2, -5\times2, -5\times2\})$

$\mu =$ $-5 \times 3 = -15.$ $\sigma =$ $|-5| \times 1 = 5.$

8

OK, Sherlocks, you've uncovered these properties:

- $\mu(c) = c.$
  $\sigma(c) = 0.$
- $\mu(X + c) = \mu(X) + c.$
  $\sigma(X + c) = \sigma(X).$
- $\mu(c \cdot X) = c \cdot \mu(X).$
  $\sigma(c \cdot X) = |c| \cdot \sigma(X).$

Where c is a constant and X is the data set.

9

Relationships

So far we've discussed problems with a single variable: univariate data.

We'll now discuss problems with two variables: bivariate data.

We'll be interested in possible relationships between the two variables.

10

A scatter diagram is a graph that shows the relationship between two quantitative variables on the same individual. Each individual is represented by a point in the diagram.

11

We view one variable as the response (dependent) variable and the other as explanatory (independent).

The explanatory variable is plotted on the horizontal axis, and the response variable on the vertical axis.

12

ADDITIONAL NOTES

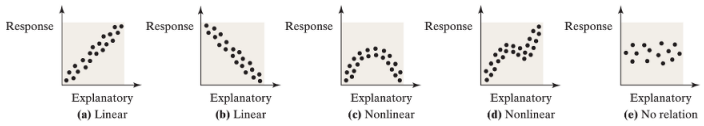
It is not always clear which should be considered the response variable and which explanatory.

For example, does high school GPA predict a student's SAT score or can SAT score predict GPA?

The researcher must determine which variable plays the role of explanatory variable based on the questions he or she wants answered.

13

Examples of scatter diagrams



We're interested in how variables that are linearly related might be associated.

14

Linearly related variables are **positively** associated when above-average values of one are associated with above-average values of the other, and below-average values of one are associated with below-average values of the other.

That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.

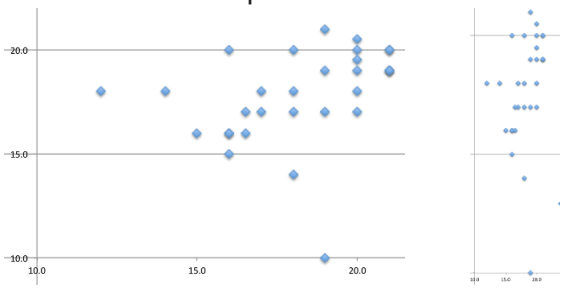
15

Linearly related variables are **negatively** associated when above-average values of one are associated with below-average values of the other.

That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

16

Trying to spot an association visually is not reliable. Here's a scatter plot. Is there a correlation?



17

The **linear** correlation coefficient is a measure of the strength and direction of the linear relation between two quantitative variables. The letter r represents the sample correlation coefficient (and ρ for population).

18

ADDITIONAL NOTES

Sample Linear Correlation Coefficient

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

$\bar{x}$ = sample mean of explanatory variable

s_x = sample std. dev. of explanatory variable

$\bar{y}$ = sample mean of response variable

s_y = sample std. dev. of response variable

n = number of individuals

19

Properties of the Linear Correlation Coefficient, r

- $-1 \leqslant r \leqslant 1$.
- If $r = +1$, then a perfect positive linear relation exists between the variables.
- If $r = -1$, then a perfect negative linear relation exists between the variables.

20

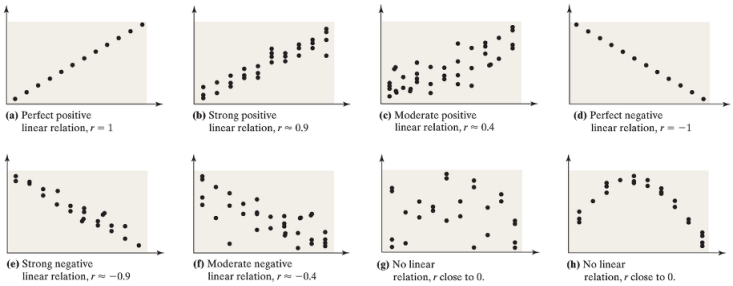
- The closer r is to $+1$, the stronger the positive association between the variables.
- The closer r is to -1, the stronger the negative association between the variables.
- If r is close to 0, then little evidence exists of a *linear* relation between the variables.

Note: r close to 0 doesn't imply no relation, just no linear relation.

21

- The linear correlation coefficient is unitless. The unit of measure for x and y plays no role in the interpretation of r.
- The correlation coefficient is not resistant. An observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

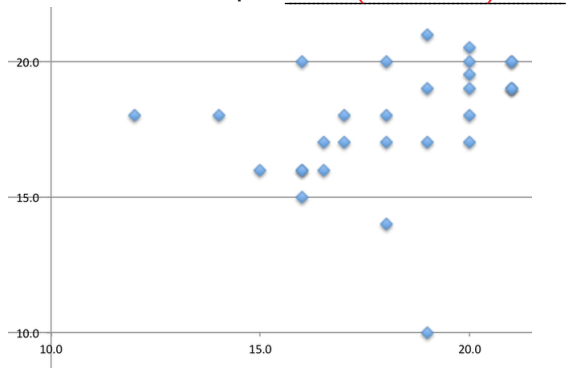
22



How close to zero does r have to be for you to know there's no linear relation?

23

Back to that example $(r = 0.33)$



24

ADDITIONAL NOTES

Is $r = 0.33$ close enough to zero to say there's no linear relation?

Depends on the sample size.
If the value of $|r|$ is lower than the critical value for *that sample size* then there's low evidence for a linear relation.

25

Critical values for r					
n	r-crit	n	r-crit	n	r-crit
4	0.950	13	0.553	22	0.423
5	0.878	14	0.532	23	0.413
6	0.811	15	0.514	24	0.404
7	0.754	16	0.497	25	0.396
8	0.707	17	0.482	26	0.388
9	0.666	18	0.468	27	0.381
10	0.632	19	0.456	28	0.374
11	0.602	20	0.444	29	0.367
12	0.576	21	0.433	30	0.361

Appendix A, table 2

26

(7) OK, you statistical sluggers, is there a (positive) linear relation in our example?
The critical value for $n = 30$ is $r_{\text{crit}} = 0.361$. Our calculated value is $r = 0.33 < r_{\text{crit}}$. Therefore, low evidence for a linear relation.

27

Correlation vs Causation

The existence of an association does not prove causation.

The behavior of two variables may seem related, but neither may *cause* the behavior of the other.

For example, there's a positive association between air-conditioning bills and high crime.

28

(8) Does that show causation? Do people go crazy with high bills and go out and commit crimes?
No. There's an underlying variable: temperature. High temperatures cause greater AC use and are also more likely to lead to people getting irritable, etc., and therefore more likely to commit crime.
A lurking variable is an initially hidden variable that's related to both explanatory and response variables.

29

(9) Dandelions and daisies tend to be found together on sports fields or in public parks in numbers that increase or decrease together (a positive association as they vary together in the same way). Causation or correlation?
Likely to be correlation. No evidence one causes the other, but both flourish or not under the same lurking conditions – for example, nature of soil, mowing frequency.

30

ADDITIONAL NOTES

(10) Thyme tends to be found where the soil has low acidity. Is the relationship between the number of thyme plants and soil acidity a positive or negative association? Negative.

(11) Is the association between the number of thyme plants and soil acidity likely to be correlation or causation? Causation.

31

(12) Studies show a positive association between listening to loud music and active acne. Does listening to loud music cause acne? Unlikely.

(13) Does acne cause listening to loud music? Unlikely.

(14) If the positive association between listening to loud music and active acne is correlation, not causation, what might be a lurking variable? Age.

32

Studies show a positive association between hand size and reading ability.

(15) Do bigger hands cause better reading ability? Unlikely.

(16) Does better reading give you bigger hands? Unlikely.

33

(17) If the positive association between hand size and reading ability is correlation, not causation, what might be a lurking variable? Holding books easily.

34

(18) Studies show a negative association between cell phone use and sperm count. As cell phone use goes up, does sperm count seem (according to these studies) go up or down? Down.

(19) Hard to know, for sure, but does lower sperm count cause higher cell phone use? Unlikely.

(20) Hard to know, for sure, but does higher cell phone use cause lower sperm count? Unlikely.

35

(21) If the negative association between cell phone use and sperm count is correlation, not causation, what might be a lurking variable?

Stress?

It's a complex issue and no *biological reason* for the association has been found.

36

ADDITIONAL NOTES