PSY 201: Statistics in Psychology

Lecture 10

Correlation

How changes in one variable correspond to change in another variable.

Greg Francis

Purdue University

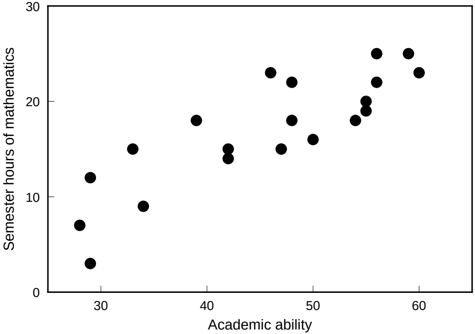
Fall 2023

CORRELATION

- two variables may be related
  - SAT scores, GPA
  - hours in therapy, self-esteem
  - grade on homeworks, grade on exams
  - number of risk factors, probability of getting AIDS
  - height, points in basketball
  - ...
- how do we show the relationship?
- scattergrams

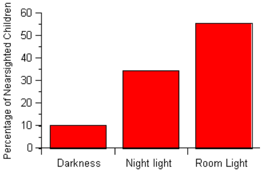
SCATTERGRAMS

- plot value of one variable against the value of the other variable



RELATIONSHIPS

- Identifying these types of relationships is one of the key issues in statistical analysis
- Consider a 1999 study that reported a relationship between the use of nightlights in a child's room and the tendency of the child to need glasses



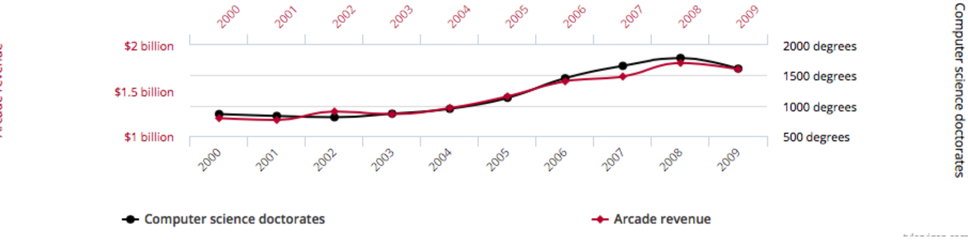
- My daughter slept with a nightlight. Was I a bad father?

COMPLICATIONS

- Clearly there is a relationship between using a nightlight and needing glasses
- However, it's not clear what the nature of the relationship involves
- It *could* be that the extra light somehow influences the child's eyes and causes the need for glasses
- Or it could be that needing glasses will somehow co-occur with the use of a nightlight (e.g., children who need glasses will want a night light, or their parents will want a nightlight)
- Finding a relationship is necessary for establishing causation, but it is not enough

SPURIOUS CORRELATION

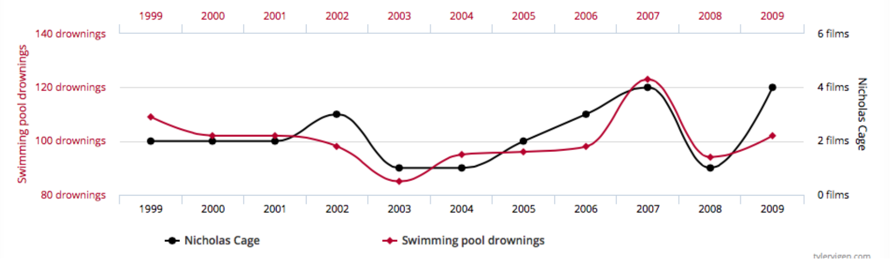
- Since so many variables get measured, it is easy to identify spurious correlations
- Sometimes there is an explanation for the relationship:



- (increased use of technology)

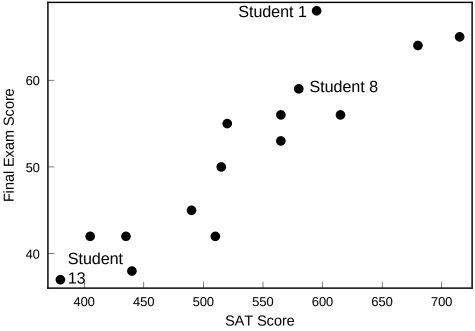
SPURIOUS CORRELATION

- Since so many variables get measured, it is easy to identify spurious correlations
- Sometimes there is no explanation for the relationship:



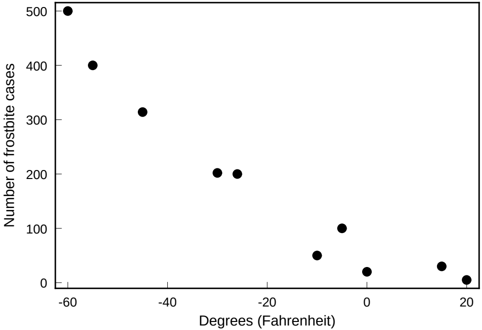
POSITIVE CORRELATION

- First, we need to understand how to quantify the existence of a relationship.
- Increases in the value of one variable tend to occur with increases in the value of the other variable
- SAT scores and exam scores



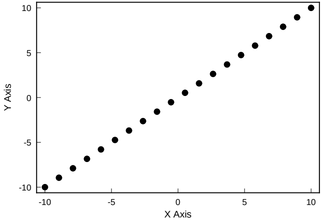
NEGATIVE CORRELATION

- Increases in the value of one variable tend to occur with **decreases** in the value of the other variable
- temperature and number of people with frostbite

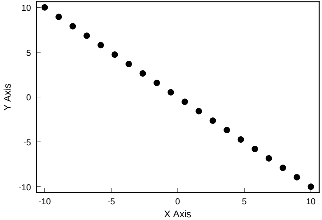


PERFECT CORRELATIONS

- perfect positive correlation

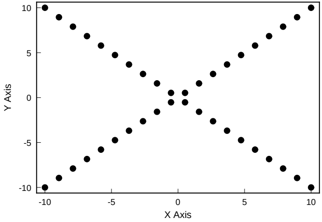


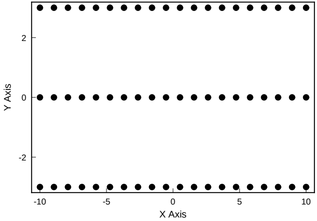
- perfect negative correlation



NO CORRELATION

- no correlation
- balance of larger and smaller values





CORRELATION COEFFICIENT

- quantitative measure of correlation
- bounded between

$$-1.0 \ \& \ +1.0$$

- correlation coefficient of -1.0 indicates perfect negative correlation
- correlation coefficient of $+1.0$ indicates perfect positive correlation
- correlation coefficient of 0.0 indicates **no** correlation
- values in between give ordinal measures of relationship

PEARSON r

- Pearson product-moment correlation coefficient
- one correlation coefficient for **quantitative** data (the most important one)

$$r = \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}}$$

- several formulas
  - z-scores
  - Deviation scores
  - Raw scores
  - Covariance
- all give the same result!

z SCORES

- Two steps
  - Convert raw scores into z scores
  - Find the mean of cross-products

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

z SCORES

- what does this calculation do?
- suppose you have two distributions that have a positive correlation
- then a large value of X will be above $\overline{X}$ and have a positive z_x score
- and a corresponding Y will be above $\overline{Y}$ and have a positive z_y score
- Thus the cross-product

$$z_x z_y$$

- will be positive

PEARSON r

- also a small value of X will be below $\overline{X}$ and have a negative z_x score
- and the corresponding Y will be below $\overline{Y}$ and have a negative z_y score
- Thus

$$z_x z_y$$

- will again be positive
- to find the average, sum all the products (positive numbers) we divide by $n-1$

$$r_{xy} = \frac{\Sigma z_x z_y}{n-1}$$

- still a positive number!

PEARSON r

- exactly the opposite is true for negatively correlated distributions
- then a large value of X will be above $\overline{X}$ and have a positive z_x score
- and a corresponding Y will be **below** $\overline{Y}$ and have a **negative** z_y score
- Thus

$$z_x z_y$$

- will be negative

PEARSON r

- while a small value of X will be below $\overline{X}$ and have a negative z_x score
- and the corresponding Y will be **above** $\overline{Y}$ and have a **positive** z_y score
- Thus

$$z_x z_y$$

- will again be negative
- to find the average, sum all the products (negative numbers) we divide by $n-1$

$$r_{xy} = \frac{\Sigma z_x z_y}{n-1}$$

- still a negative number!

DEVIATION FORMULA

- it is awkward to convert to z scores
- we can get the same number with deviation scores

$$x = X - \overline{X}$$

$$y = Y - \overline{Y}$$

- deviation score formula

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

RAW SCORE FORMULA

- it is awkward to calculate deviation scores
- raw score formula

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left[n\Sigma X^2 - (\Sigma X)^2\right]\left[n\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

COVARIANCE FORMULA

$$\text{covariance} = s_{xy} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{n - 1}$$

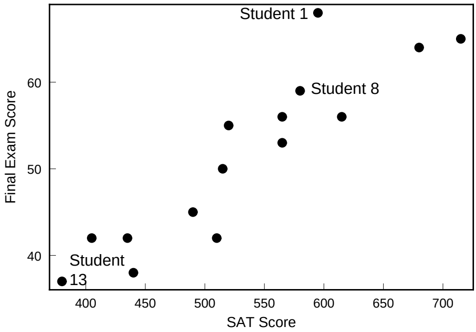
- average cross-product of deviation scores (similar to variance)
- Pearson r turns out to be:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- where s_x and s_y are the standard deviations of their respective distributions

EXAMPLE

X	Y
595	68
520	55
715	65
405	42
680	64
490	45
565	56
580	59
615	56
435	42
440	38
515	50
380	37
510	42
565	53



EXAMPLE

- standard score formula

$$r_{xy} = \frac{\Sigma z_x z_y}{n-1} = \frac{12.67}{14} = 0.905$$

X	Y	z_x	z_y	$z_x z_y$
595	68	0.63	1.64	1.03
520	55	-0.15	0.35	-0.05
715	65	1.88	1.34	2.52
405	42	-1.34	-0.94	1.26
680	64	1.51	1.24	1.87
490	45	-0.46	-0.64	0.29
565	56	0.32	0.45	0.14
580	59	0.48	0.74	0.36
615	56	0.84	0.45	0.38
435	42	-1.03	-0.94	0.97
440	38	-0.97	-1.33	1.29
515	50	-0.20	-0.15	0.03
380	37	-1.60	-1.43	2.29
510	42	-0.25	-0.94	0.24
565	53	0.32	0.15	0.05
$\sum X = 8010$	$\sum Y = 772$	$\sum z_x = 0.0$	$\sum z_y = 0.0$	$\sum z_x z_y = 12.67$

EXAMPLE

- deviation score formula

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{12332.00}{\sqrt{(130460.0)(1429.72)}} = 0.903$$

X	Y	x	y	xy
595	68	61.0	16.53	1008.33
520	55	-14.0	3.53	-49.42
715	65	181.0	13.53	2448.93
405	42	-129.0	-9.47	1221.63
680	64	146.0	12.53	1829.38
490	45	-44.0	-6.47	284.68
565	56	31.0	4.53	140.43
580	59	46.0	7.53	346.38
615	56	81.0	4.53	366.93
435	42	-99.0	-9.47	937.53
440	38	-94.0	-13.47	1266.18
515	50	-19.0	-1.47	27.93
380	37	-154.0	-14.47	2228.38
510	42	-24.0	-9.47	227.28
565	53	31.0	1.53	47.43
$\sum X = 8010$	$\sum Y = 772$	$\sum x = 0.0$	$\sum y = 0.0$	$\sum xy = 12332.00$

- $\Sigma x^2 = 130460.0$ and $\Sigma y^2 = 1429.72$

EXAMPLE

- raw score formula

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left[n\Sigma X^2 - (\Sigma X)^2\right]\left[n\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

$$\frac{(15)(424580) - (8010)(772)}{\sqrt{[(15)(4407800) - (8010)^2][(15)(41162) - (772)^2]}} = 0.903$$

X	Y	XY
595	68	40460
520	55	28600
715	65	46475
405	42	17010
680	64	43520
490	45	22050
565	56	31640
580	59	34220
615	56	34440
435	42	18270
440	38	16720
515	50	25750
380	37	14060
510	42	21420
565	53	29945
$\sum X = 8010$	$\sum Y = 772$	$\sum XY = 424580$

- $\Sigma X^2 = 4407800$ and $\Sigma Y^2 = 41162$

EXAMPLE

- covariance formula

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{880.86}{(96.53)(10.11)} = 0.903$$

- where,

$$s_{xy} = \frac{\Sigma xy}{n-1} = \frac{12332}{14} = 880.86$$

$$s_x = \sqrt{\frac{\Sigma x^2}{n-1}} = \sqrt{\frac{130460}{14}} = 96.53$$

$$s_y = \sqrt{\frac{\Sigma y^2}{n-1}} = \sqrt{\frac{1429.72}{14}} = 10.11$$

CORRELATION

- r measures correlation between two variables
- **not** just any two variables
  - The two variables must be **paired observations**.
  - Variables must be quantitative (interval or ratio scale).

CONCLUSIONS

- correlation
- scattergrams
- Pearson r
- formulas

NEXT TIME

- factors affecting r
- interpreting r

Is there a link between IQ and problem solving ability?