PSYC122 week 11: Correlation

EVERYBODY WHO WENT TO THE MOON HAS EATEN CHICKEN!
GOOD GRIEF. CHICKEN MAKES YOU GO TO THE MOON!

Dr. Margriet A. Groen

Topics for week 11-15: Testing for associations

Week 11. Correlation

Week 12. Correlation (continued)

Week 13: Simple linear regression

Week 14. Chi-square

Week 15. Recap and class test

Outline

- What is correlation? Correlation coefficient and scatterplots
- Types of correlation Interpreting correlation coefficient and scatterplots
- Covariance and correlation How correlation is derived from covariance
- Hypothesis testing Critical values for significance
- Coefficient of determination R^2
- Why correlation does not infer causation
- How to conduct correlation R
- How to report correlation APA

What is correlation?

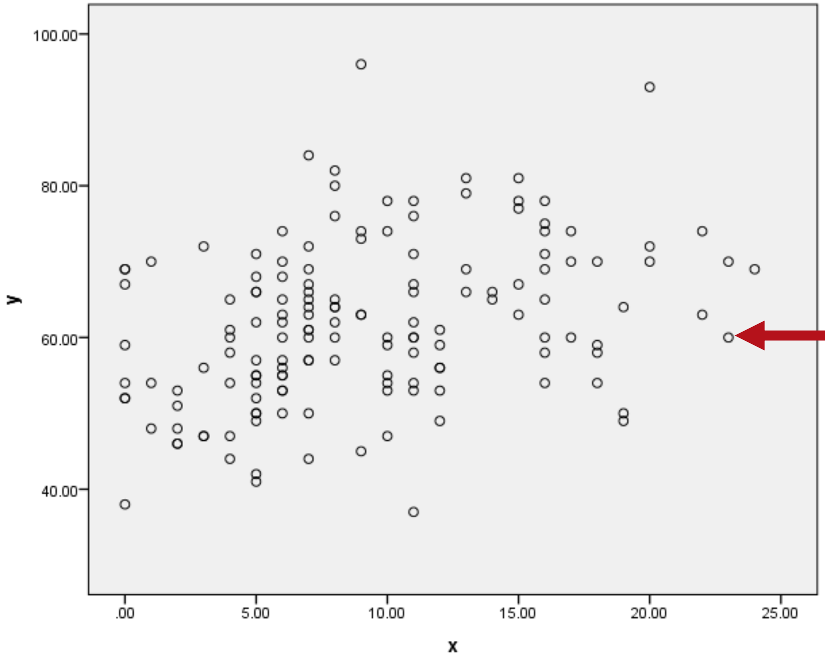
Correlation measures the *relationship* between two *continuous (or numerical)* variables of interest

- ***Does height relate to weight?***
- ***Is seminar absence related to WBA scores?***
- ***Does the number of cats you own relate to how violent you are?***

In other words, if something happens to *X*, what happens to *Y?*

The scatterplot

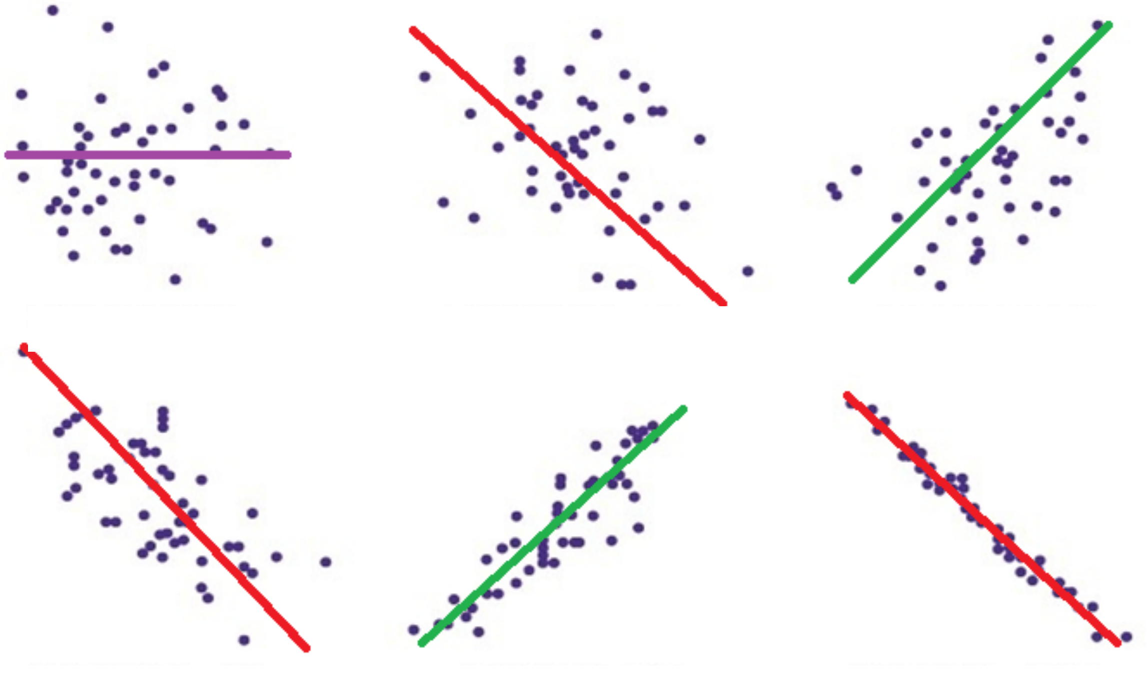
We assess the relationship between two variables *visually* via a **scatterplot.**



The 1st step of correlation analysis is constructing and interpreting a **scatterplot** displaying scores for the two variables

The two values for each individual are used to plot a single point on the graph

Scatterplots: strength and direction



1. **The strength of the relationship**: closeness of points to the line of best fit
2. **The direction of the relationship:** positive, negative, or null

Use of scatterplots

Graphs are not just an end product or a ‘pretty’ addition to your paper. They allow us to:

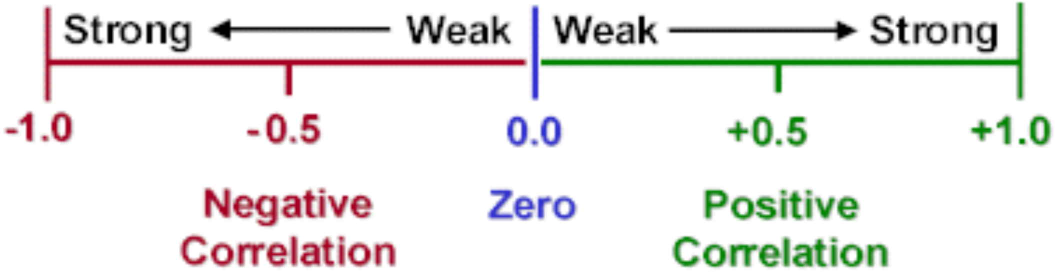
- Familiarise ourselves with the data
- Identify the distribution of data and any initial relationships
- Identify any outliers (more on this next week!)

The Pearson Product-Moment Correlation Coefficient (*r*)

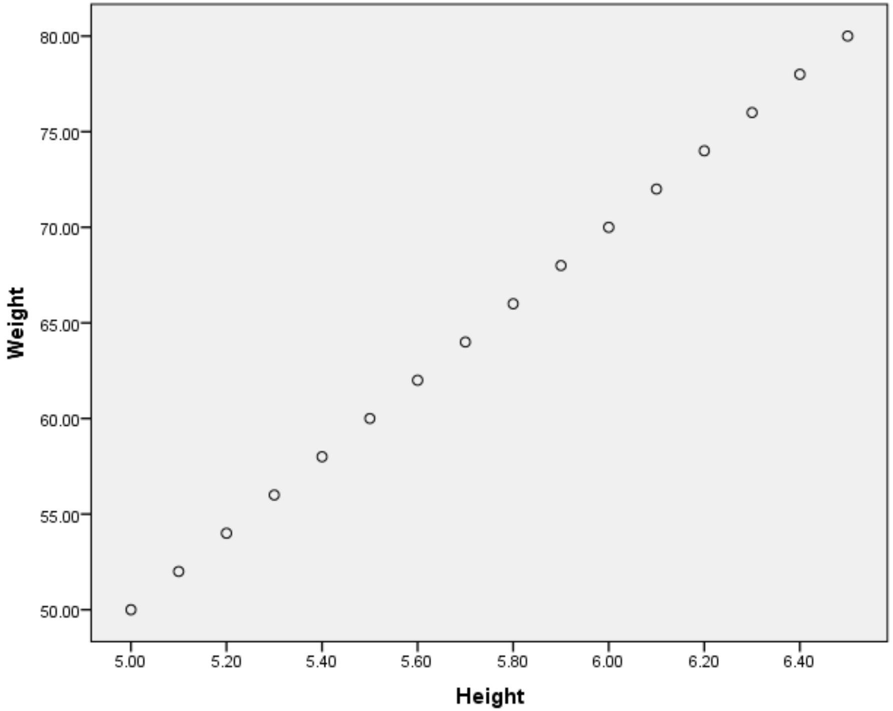
‘Pearson’s *r*’ is a statistic that quantifies the *linear* correlation between two variables ranging from -1 to 1.

The same two aspects that are visible in a scatterplot are also reflected in the correlation coefficient:

1. **The strength of the relationship**: the value (ignoring -/+)
2. **The direction of the relationship:** positive, negative, or null



Types of correlation: Positive correlation



When X increases, Y also increases

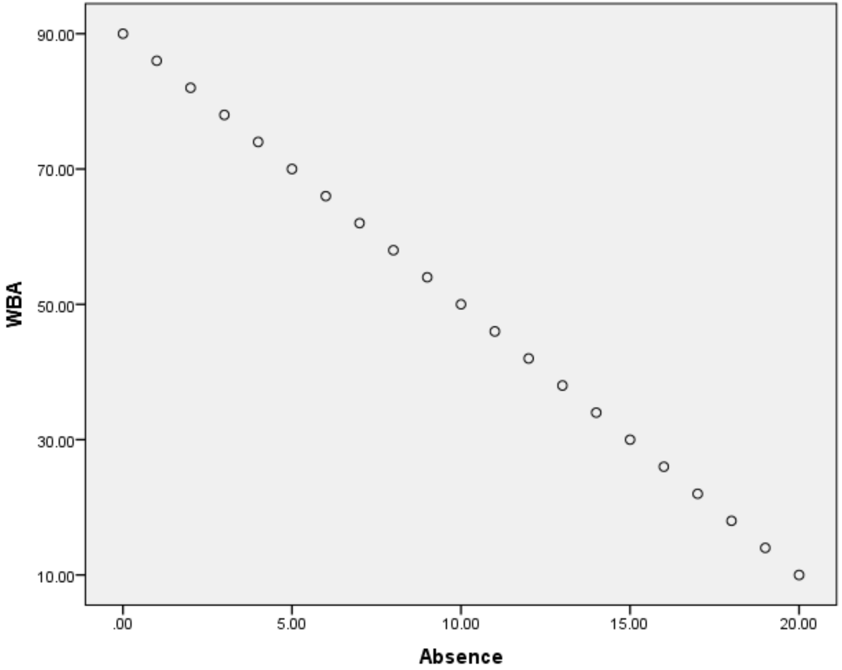
A perfect positive correlation, *r* = 1

(a positive value)

As height increases, weight increases.

Height is positively correlated with weight.

Types of correlation: Negative correlation



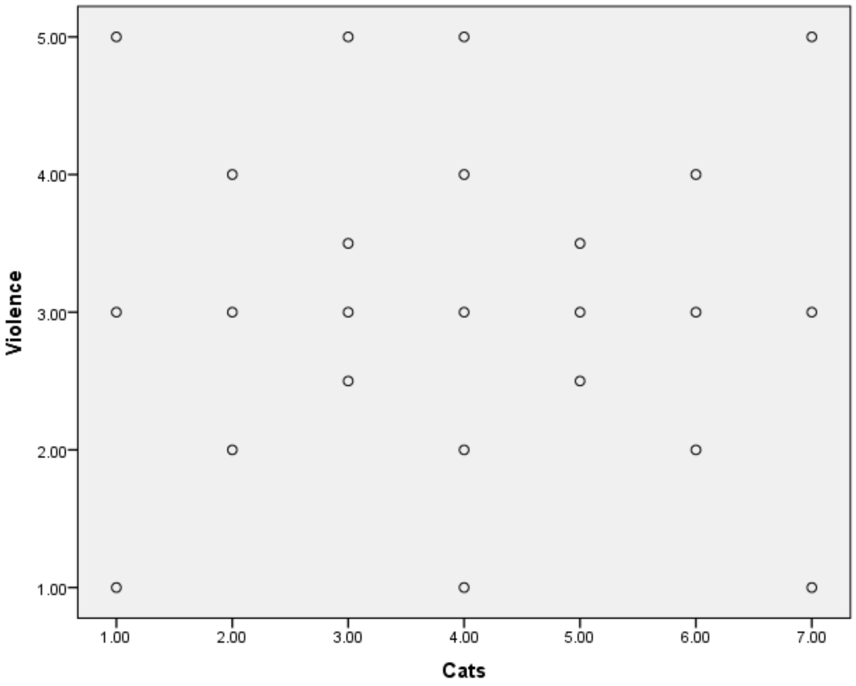
When X increases, Y decreases

A perfect negative correlation, *r* = -1

(a negative value)

As seminar absence increases, WBA scores decrease. Seminar absence is negatively correlated with WBA score.

Types of correlation: Null correlation

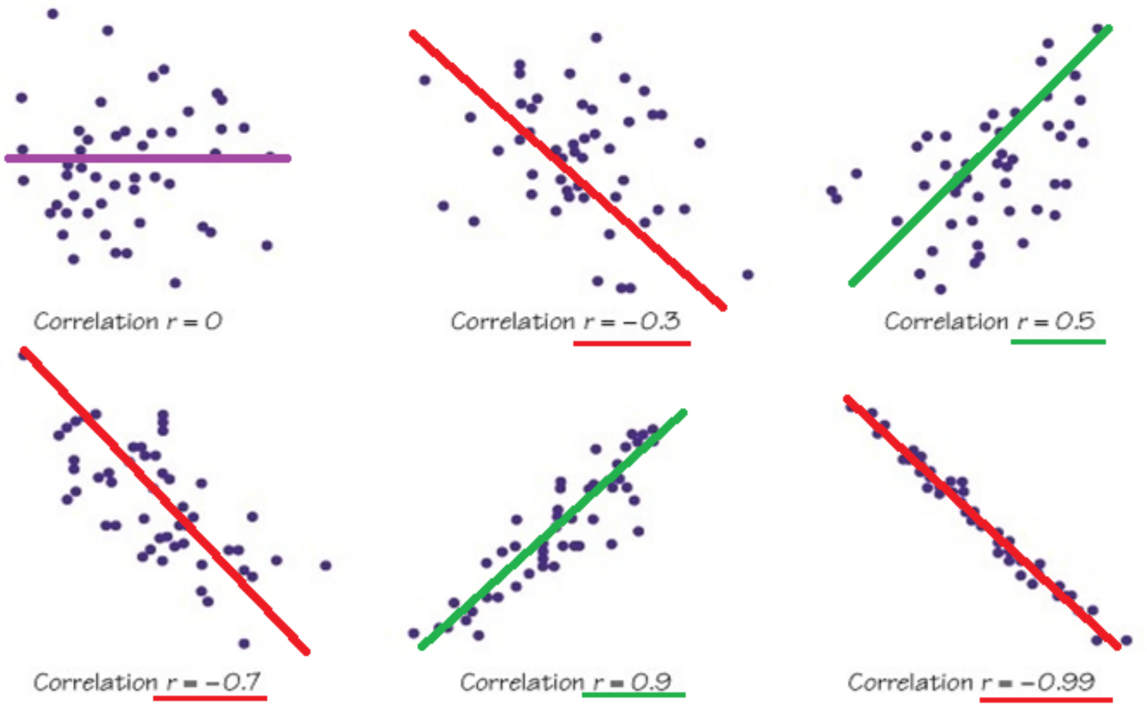


When X increases, Y shows no consistent change

A null correlation, *r* = 0

As the number of cats owned increases, level of violence does not change. There is no relationship between number of cats owned and violence score.

The Pearson Product-Moment Correlation Coefficient (*r*) (*cont.*)



Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$

Correction Coefficient

Small/weak: *r* > .1

Medium/moderate: *r* > .3

Large/strong: *r* > .5

****Pearson's r is an effect size in itself!***

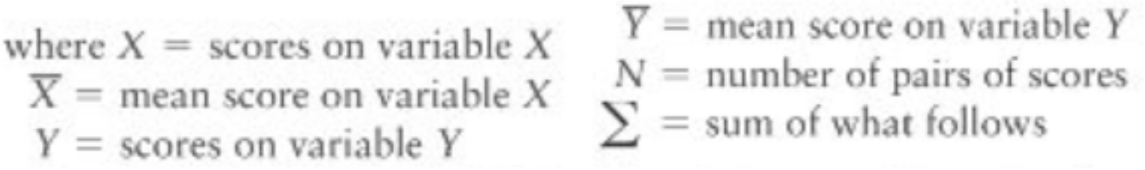
Understanding correlation by covariance

- **Recap:** Think back to the formula to measure variance (on one variable, e.g. *X*)

$$\boldsymbol{variance} = \frac{\sum(\boldsymbol{X} - \bar{\boldsymbol{X}})^{2}}{\boldsymbol{N}}$$

- *Covariance is the extent to which two variables vary together. So* Instead of multiplying the scores by itself $(X - \bar{X})^2$ we multiple it with the other variable (e.g. *Y*)

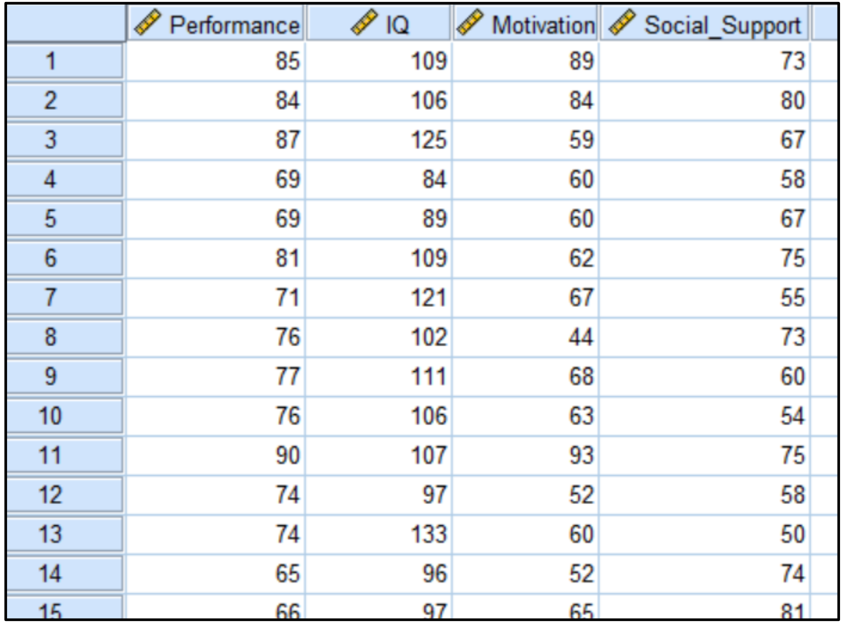
$$\boldsymbol{covariance} = \frac{\sum(\boldsymbol{X} - \bar{\boldsymbol{X}})(\boldsymbol{Y} - \bar{\boldsymbol{Y}})}{\boldsymbol{N} - \boldsymbol{1}}$$



where X = scores on variable X
$\bar{X}$ = mean score on variable X
Y = scores on variable Y
$\bar{Y}$ = mean score on variable Y
N = number of pairs of scores
Σ = sum of what follows

Understanding correlation by covariance (*cont.*)

$$covariance = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N-1}$$



	Performance	IQ	Motivation	Social_Support
1	85	109	89	73
2	84	106	84	80
3	87	125	59	67
4	69	84	60	58
5	69	89	60	67
6	81	109	62	75
7	71	121	67	55
8	76	102	44	73
9	77	111	68	60
10	76	106	63	54
11	90	107	93	75
12	74	97	52	58
13	74	133	60	50
14	65	96	52	74
15	66	97	65	81

Mean	**78**	**67**
SD	**8**	**13.7**

<u>Looking at job performance and motivation</u>

(85 – 78) X (89 – 67) = 154
(84 – 78) X (84 – 67) = 102
Etc… (for each participant)
Then add all of the totals and divide by N-1

From covariance to correlation

- Unlike standard variance, *covariance* may have a positive or a negative value, suggesting the direction of the variance (similar to Pearson's *r*).

Covariance for job performance and motivation = 69.24

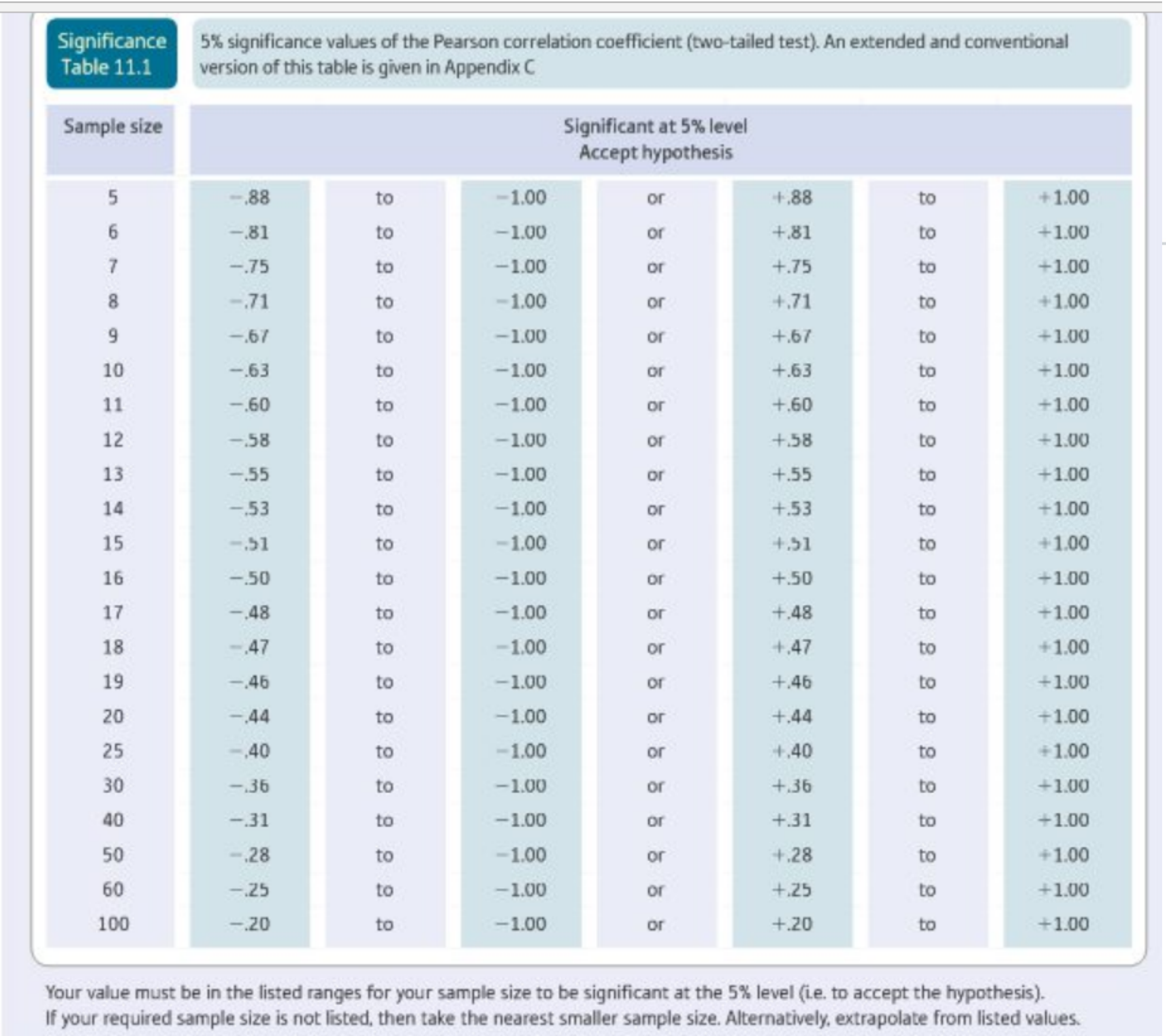
- **So why do Pearson's correlation?** The size of the covariance is affected by the size of variances of the two separate variables which can make comparisons difficult. The correlation formula improved this by replacing N with the SD's

$$r = \frac{covariance}{SD(for\,X)\times SD(for\,Y)} \qquad r = \frac{69.24}{8\times 13.7} \qquad r = .63$$

Hypothesis testing

Null hypothesis: There is no correlation

Significance: For a correlation to be significant it needs to be bigger than the ***critical value***



Significance Table 11.1

5% significance values of the Pearson correlation coefficient (two-tailed test). An extended and conventional version of this table is given in Appendix C

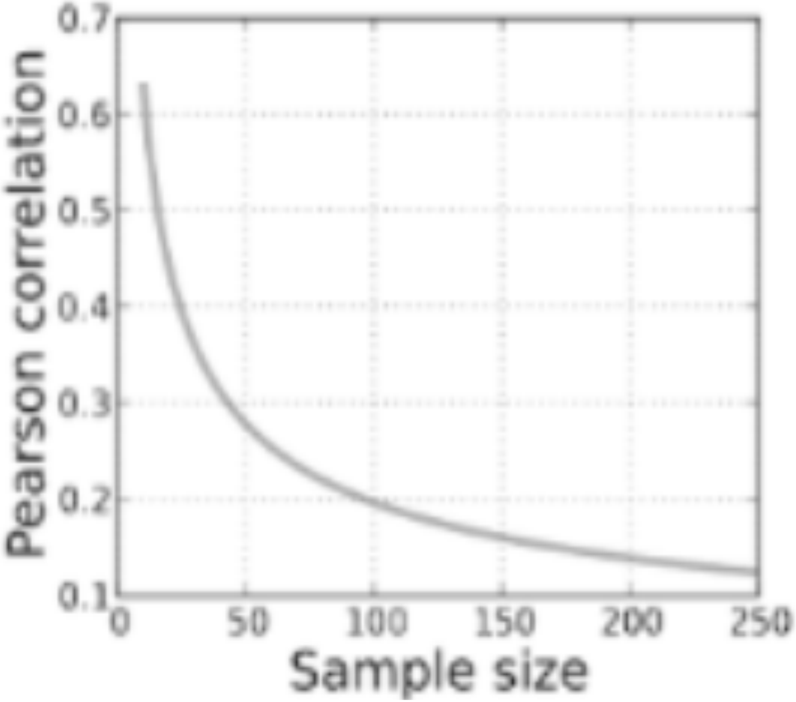
Sample size	Significant at 5% level Accept hypothesis						
5	−.88	to	−1.00	or	+.88	to	+1.00
6	−.81	to	−1.00	or	+.81	to	+1.00
7	−.75	to	−1.00	or	+.75	to	+1.00
8	−.71	to	−1.00	or	+.71	to	+1.00
9	−.67	to	−1.00	or	+.67	to	+1.00
10	−.63	to	−1.00	or	+.63	to	+1.00
11	−.60	to	−1.00	or	+.60	to	+1.00
12	−.58	to	−1.00	or	+.58	to	+1.00
13	−.55	to	−1.00	or	+.55	to	+1.00
14	−.53	to	−1.00	or	+.53	to	+1.00
15	−.51	to	−1.00	or	+.51	to	+1.00
16	−.50	to	−1.00	or	+.50	to	+1.00
17	−.48	to	−1.00	or	+.48	to	+1.00
18	−.47	to	−1.00	or	+.47	to	+1.00
19	−.46	to	−1.00	or	+.46	to	+1.00
20	−.44	to	−1.00	or	+.44	to	+1.00
25	−.40	to	−1.00	or	+.40	to	+1.00
30	−.36	to	−1.00	or	+.36	to	+1.00
40	−.31	to	−1.00	or	+.31	to	+1.00
50	−.28	to	−1.00	or	+.28	to	+1.00
60	−.25	to	−1.00	or	+.25	to	+1.00
100	−.20	to	−1.00	or	+.20	to	+1.00

Your value must be in the listed ranges for your sample size to be significant at the 5% level (i.e. to accept the hypothesis). If your required sample size is not listed, then take the nearest smaller sample size. Alternatively, extrapolate from listed values.

Pearson’s *r* correlation coefficient critical values for *p* = .05

From Howitt & Cramer, 2017

Hypothesis testing (*cont.*)



For 10 participants, *r* = .63 to 1.00 (ignore +/-) for significance at *p* < .05

20 participants, *r* = .44 to 1.00

50 participants, *r* = .28 to 1.00

100 participants, *r* = .20 to 1.00

Coefficient of determination

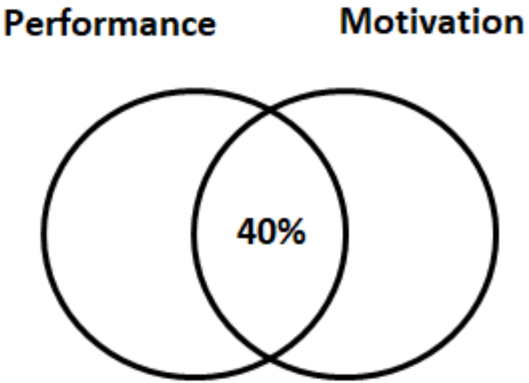
The ***coefficient of determination*** tells us the proportion of variance in one variable that can be accounted for by the other variable.

- This is established by squaring *r* and as *r* is below 1, the squared value will always be less. (e.g. if ***r*** = .3, ***R²*** = .09)

Job performance and motivation

r = .63, *R*² = .63 X .63 = .40

40% of the variance in job performance is accounted for by the variance in motivation



Why does correlation not infer causation?

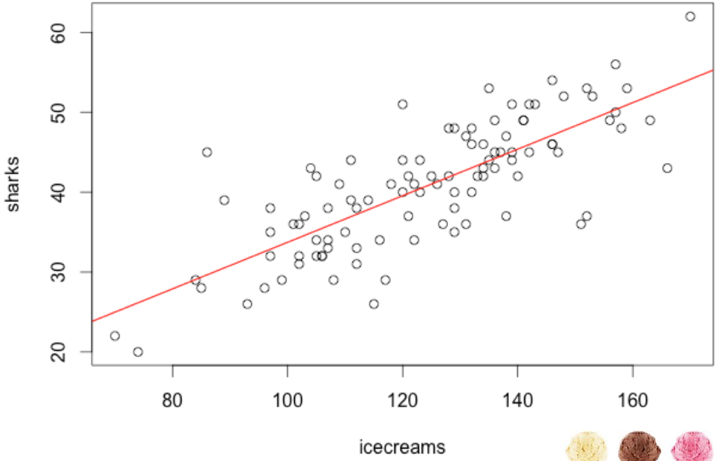
Correlation can only infer relationships between variables and not causation

Why?

The only way to infer causation is to experimentally control/manipulate an IV, and then measure a DV. Then you can say that this manipulation caused any variation observed in the DV

Correlation tells us how two variables relate to one another but does not tell us whether one causes the other (even though it might!)

Why does correlation not infer causation? *(cont.)*



Shark attacks are related to ice cream sales.

This does not suggest that ice-cream filled bellies are a delicacy for sharks.

However, there could another factor influencing this relationship…

Hot weather → more people swim in the sea → more shark attacks

Hot weather → more people want ice cream

Summary

- Correlation measures the relationship between two numerical or continuous variables.
- A scatterplot is useful to construct **before** the correlation analysis to interpret the relationship and assumptions (more of this next week).
- Pearson’s correlation coefficient gives us information on the strength and direction of the relationship.
- The significance of the correlation is partly dependent on the sample size.
- The coefficient of determination tells us the proportion of variance that can be accounted for by the other variable.
- Do not confuse correlation with causation and think!

  What else could be influencing the correlation?