Correlation

Lecture 4

Survey Research & Design in Psychology

James Neill, 2012

Overview

1. Purpose of correlation
2. Covariation
3. Linear correlation
4. Types of correlation
5. Interpreting correlation
6. Assumptions / limitations
7. Dealing with several correlations

Readings

Howell (2010)

- Ch6 Categorical Data & Chi-Square
- Ch9 Correlation & Regression
- Ch10 Alternative Correlational Techniques
  - 10.1 Point-Biserial Correlation and Phi: Pearson Correlation by Another Name
  - 10.3 Correlation Coefficients for Ranked Data

Purpose of correlation

4

Purpose of correlation

The underlying purpose of correlation is to help address the question:

What is the
- **relationship** or
- degree of **association** or
- amount of **shared variance**

between **two variables**?

5

Purpose of correlation

Other ways of expressing the underlying correlational question include:

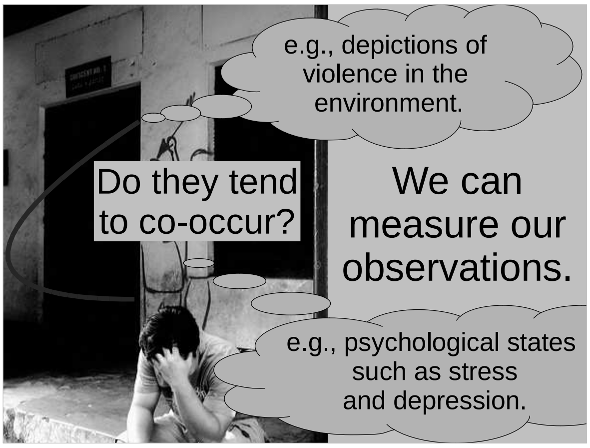
To what extent
- do two variables **covary**?
- are two variables **dependent** or **independent** of one another?
- can one variable be **predicted** from another?

6

Covariation

The world is made of covariation.

We observe covariations in the psycho-social world.



Covariations are the basis of more complex models.

Linear correlation

Linear correlation

The extent to which two variables have a simple **linear** (straight-line) relationship.

Linear correlations provide the building blocks for multivariate correlational analyses, such as:

- Factor analysis
- Reliability
- Multiple linear regression

Linear correlation

Linear relations between variables are indicated by correlations:

- **Direction:** Correlation sign (+ / -) indicates direction of linear relationship
- **Strength:** Correlation size indicates strength (ranges from -1 to +1)
- **Statistical significance:** *p* indicates likelihood that observed relationship could have occurred by chance

What is the linear correlation? Types of answers

- No relationship (independence)
- Linear relationship:
  - As one variable ↑s, so does the other (+ve)
  - As one variable ↑s, the other ↓s (-ve)
- Non-linear relationship
- Pay caution due to:
  - Heteroscedasticity
  - Restricted range
  - Heterogeneous samples

Types of correlation

To decide which type of correlation to use, consider the **levels of measurement** for each variable

Types of correlation

- Nominal by nominal:
  Phi (Φ) / Cramer's *V*, Chi-squared
- Ordinal by ordinal:
  Spearman's rank / Kendall's Tau *b*
- Dichotomous by interval/ratio:
  Point bi-serial r_{pb}
- Interval/ratio by interval/ratio:
  Product-moment or Pearson's *r*

Types of correlation and LOM

	Nominal	**Ordinal**	**Int/Ratio**
Nominal	Clustered bar-chart, Chi-square, Phi (φ) or Cramer's *V*	⇐ Recode	Scatterplot, bar chart or error-bar chart Point bi-serial correlation (r_{pb})
Ordinal		Scatterplot or clustered bar chart Spearman's Rho or Kendall's Tau	⇐⇑ Recode
Int/Ratio			Scatterplot Product-moment correlation (*r*)

Nominal by nominal

Nominal by nominal correlational approaches

- Contingency (or cross-tab) tables
  - Observed
  - Expected
  - Row and/or column %s
  - Marginal totals
- Clustered bar chart
- Chi-square
- Phi/Cramer's V

Contingency tables

- Bivariate frequency tables
- Cell frequencies (red)
- Marginal totals (blue)

	Diseased	Disease Free	
Exposed	a	b	n_1
Not Exposed	c	d	n_0
	m_1	m_0	n

Contingency table: Example

b2 Do you snore? * b3r Smoker Crosstabulation

Count

		b3r Smoker		
		0 No	1 Yes	Total
b2 Do you snore?	0 yes	50	16	66
	1 no	111	9	120
Total		161	25	186

RED = Contingency cells
BLUE = Marginal totals

Contingency table: Example

b2 Do you snore? * b3r Smoker Crosstabulation

			b3r Smoker		
			0 No	1 Yes	Total
b2 Do you snore?	0 yes	Count	50	16	66
		Expected Count	57.1	8.9	66.0
	1 no	Count	111	9	120
		Expected Count	103.9	16.1	120.0
Total		Count	161	25	186
		Expected Count	161.0	25.0	186.0

Chi-square is based on the differences between the actual and expected cell counts.

b2 Do you snore? * b3r Smoker Crosstabulation

% within b2 Do you snore?

		b3r Smoker		
		0 No	1 Yes	Total
b2 Do you snore?	0 yes	75.8%	24.2%	100.0%
	1 no	92.5%	7.5%	100.0%
Total		86.6%	13.4%	100.0%

Row and/or column cell percentages may also aid interpretation e.g., ~2/3rds of smokers snore, whereas only ~1/3rd of non-smokers snore.

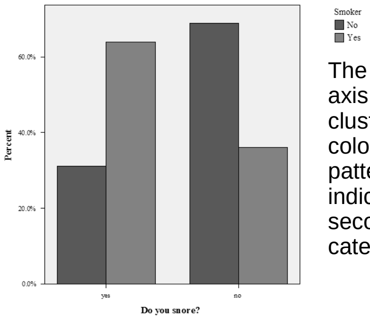
b2 Do you snore? * b3r Smoker Crosstabulation

% within b3r Smoker

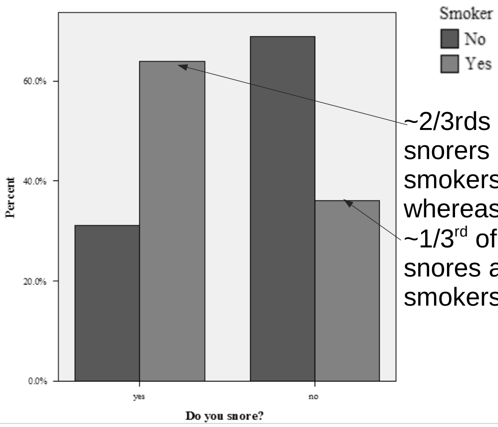
		b3r Smoker		
		0 No	1 Yes	Total
b2 Do you snore?	0 yes	31.1%	64.0%	35.5%
	1 no	68.9%	36.0%	64.5%
Total		100.0%	100.0%	100.0%

Clustered bar graph

Bivariate bar graph of frequencies or percentages.



The category axis bars are clustered (by colour or fill pattern) to indicate the the second variable's categories.



~2/3rds of snorers are smokers, whereas only ~1/3rd of non-snores are smokers.

Pearson chi-square test

The value of the test-statistic is

$$X^2 = \sum \frac{(O - E)^2}{E},$$

where

X^2 = the test statistic that approaches a χ^2 distribution.

O = frequencies observed;

E = frequencies expected (asserted by the null hypothesis).

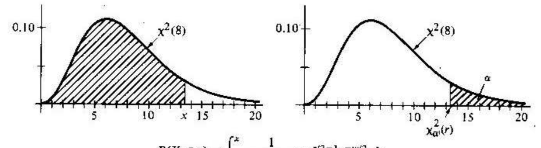
Pearson chi-square test: Example

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10.259[b]	1	.001
Continuity Correction[a]	8.870	1	.003
Likelihood Ratio	9.780	1	.002
Fisher's Exact Test			
Linear-by-Linear Association	10.204	1	.001
N of Valid Cases	186		

Write-up: $\chi 2$ (1, 186) = 10.26, *p* = .001

Chi-square distribution: Example

The Chi-Square Distribution



$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2}\, dw$$

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09

Phi (ϕ) & Cramer's *V*

(non-parametric measures of correlation)

Phi (ϕ)

- Use for 2x2, 2x3, 3x2 analyses
  e.g., Gender (2) & Pass/Fail (2)

Cramer's *V*

- Use for 3x3 or greater analyses
  e.g., Favourite Season (4) x Favourite Sense (5)

Phi (ϕ) & Cramer's *V*: Example

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-.235	.001
	Cramer's V	.235	.001
N of Valid Cases		186	

χ^2 (1, 186) = 10.26, p = .001, φ = .24

Ordinal by ordinal

Ordinal by ordinal correlational approaches

- Spearman's rho (r_s)
- Kendall tau (τ)
- Alternatively, use nominal by nominal techniques (i.e., treat as lower level of measurement)

Graphing ordinal by ordinal data

- Ordinal by ordinal data is difficult to visualise because its non-parametric, yet there may be many points.
- Consider using:
  - Non-parametric approaches (e.g., clustered bar chart)
  - Parametric approaches (e.g., scatterplot with binning)

Spearman's rho (r_s) or Spearman's rank order correlation

- For ranked (ordinal) data
  - e.g. Olympic Placing correlated with World Ranking
- Uses product-moment correlation formula
- Interpretation is adjusted to consider the underlying ranked scales

Kendall's Tau (τ)

- Tau a
  - Does not take joint ranks into account
- Tau b
  - Takes joint ranks into account
  - For square tables
- Tau c
  - Takes joint ranks into account
  - For rectangular tables

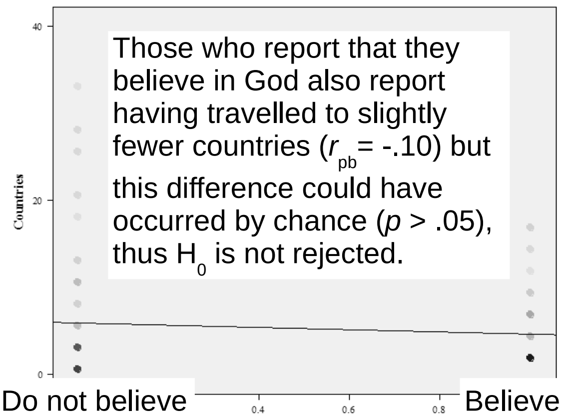
Dichotomous by interval/ratio

37

Point-biserial correlation (r_{pb})

- One dichotomous & one continuous variable
  - e.g., belief in god (yes/no) and amount of international travel
- Calculate as for Pearson's product-moment *r*,
- Adjust interpretation to consider the underlying scales

38



Those who report that they believe in God also report having travelled to slightly fewer countries (r_{pb} = -.10) but this difference could have occurred by chance (p > .05), thus H_0 is not rejected.

Point-biserial correlation (r_{pb}): Example

Correlations

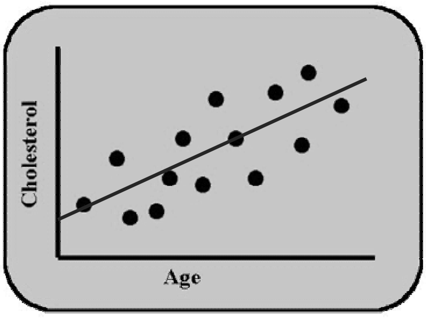
		b4r God	b8 Countries
b4r God	Pearson Correlation	1	-.095
0 = No	Sig. (2-tailed)		.288
1 = Yes	N	127	127
b8 Countries	Pearson Correlation	-.095	1
	Sig. (2-tailed)	.288	
	N	127	190

Interval/ratio by Interval/ratio

Scatterplot

- Plot each pair of observations (X, Y)
  - x = predictor variable (independent)
  - y = criterion variable (dependent)
- By convention:
  - the IV should be plotted on the x (horizontal) axis
  - the DV on the y (vertical) axis.

Scatterplot showing relationship between age & cholesterol with line of best fit



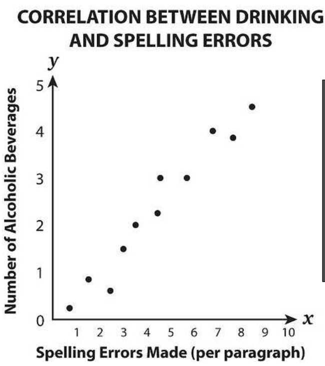
43

Line of best fit

- The correlation between 2 variables is a measure of the degree to which pairs of numbers (points) cluster together around a best-fitting straight line
- Line of best fit: y = a + bx
- Check for:
  - –outliers
  - –linearity

44

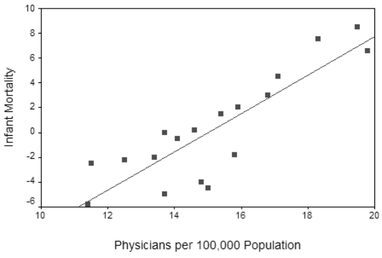
What's wrong with this scatterplot?



IV should treated as X and DV as Y, although this is not always distinct.

45

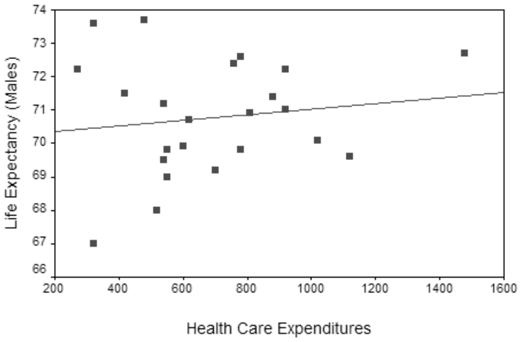
Scatterplot example: Strong positive (.81)



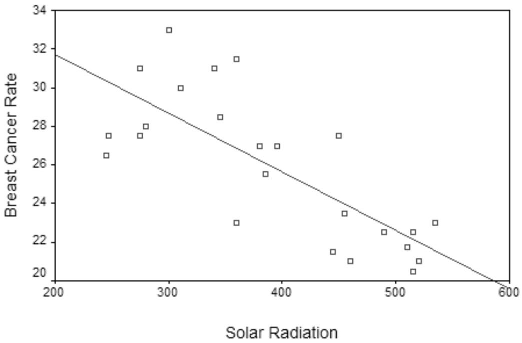
Q: Why is infant mortality positively linearly associated with the number of physicians (with the effects of GDP removed)?

A: Because more doctors tend to be deployed to areas with infant mortality (socio-economic status aside).

Scatterplot example: Weak positive (.14)



Scatterplot example: Moderately strong negative (-.76)



Pearson product-moment correlation (*r*)

- The product-moment correlation is the **standardised covariance**.

$$r_{X,Y} = \frac{\text{cov}(X,Y)}{S_X S_Y}$$

49

Covariance

- Variance shared by 2 variables

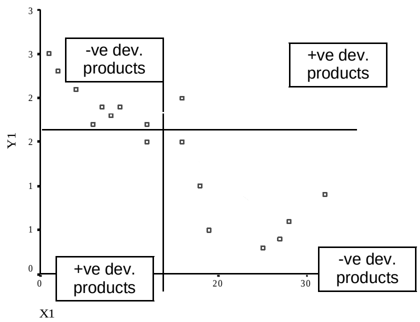
$$Cov_{XY} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{N - 1}$$

Cross products

- Covariance reflects the direction of the relationship:
  - +ve cov indicates + relationship
  - -ve cov indicates - relationship.

50

Covariance: Cross-products



51

Covariance

- Dependent on the scale of measurement → Can't compare covariance across different scales of measurement (e.g., age by weight in kilos <u>versus</u> age by weight in grams).
- Therefore, **standardise** covariance (divide by the cross-product of the Sds) → correlation
- Correlation is an effect size – i.e., standardised measure of strength of linear relationship

52

Covariance, *SD*, and correlation: Quiz

For a given set of data the covariance between *X* and *Y* is 1.20. The *SD* of *X* is 2 and the *SD* of *Y* is 3. The resulting correlation is:

a. .20
b. .30
c. .40
d. 1.20

Answer:
1.20 / 2 x 3 = .20

53

Hypothesis testing

Almost all correlations are not 0, therefore the question is:

"What is the **likelihood** that a relationship between variables is a 'true' relationship - or could it simply be a result of random sampling variability or 'chance'?"

54

Significance of correlation

- **Null hypothesis (H_0)**: ρ = 0: assumes that there is no ‘true’ relationship (in the population)
- **Alternative hypothesis (H_1)**: ρ <> 0: assumes that the relationship is real (in the population)
- Initially assume $\mathbf{H_0}$ is true, and evaluate whether the data support $\mathbf{H_1}$.
- **ρ (rho)** = population product-moment correlation coefficient

How to test the null hypothesis

- Select a critical value (alpha (α)); commonly .05
- Can use a 1 or 2-tailed test
- Calculate correlation and its *p* value. Compare this to the critical value.
- If *p* < critical value, the correlation is statistically significant, i.e., that there is less than a x% chance that the relationship being tested is due to random sampling variability.

Correlation – SPSS output

Correlations

		Cigarette Consumption per Adult per Day	CHD Mortality per 10,000
Cigarette Consumption per Adult per Day	Pearson Correlation		
	Sig. (2-tailed)		
	N		
CHD Mortality per 10,000	Pearson Correlation	.713*	
	Sig. (2-tailed)	.000	
	N	21	

**. Correlation is significant at the 0.01 level (2-tailed).

Imprecision in hypothesis testing

- **Type I error**: rejects $\mathbf{H_0}$ when it is true
- **Type II error**: Accepts $\mathbf{H_0}$ when it is false
- Significance test result will depend on the power of study, which is a function of:
  - Effect size (r)
  - Sample size (N)
  - Critical alpha level (α_{crit})

58

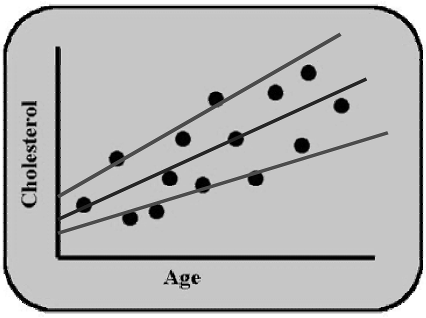
Significance of correlation

df (N-2)	critical p = .05
5	.67
10	.50
15	.41
20	.36
25	.32
30	.30
50	.23
200	.11
500	.07
1000	.05

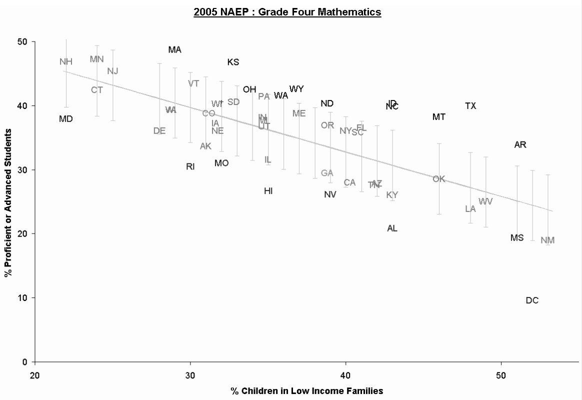
The size of correlation required to be significant decreases as N increases – why?

59

Scatterplot showing a confidence interval for a line of best fit



60



Practice quiz question:
Significance of correlation

If the correlation between Age and test Performance is statistically significant, it means that:

a. there is an important relationship between Age and test Performance

b. the true correlation between Age and Performance in the population is equal to 0

c. the true correlation between Age and Performance in the population is not equal to 0

d. getting older causes you to do poorly on tests

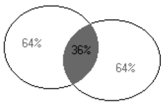
62

Interpreting correlation

63

Coefficient of Determination (r^2)

- CoD = The proportion of variance or change in one variable that can be accounted for by another variable.
- e.g., r = .60, r^2 = .36



64

Interpreting correlation (Cohen, 1988)

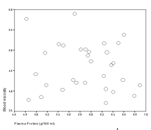
A correlation is an **effect size**, so guidelines re strength can be suggested.

Strength	r	r^2
weak:	.1 to .3	(1 to 10%)
moderate:	.3 to .5	(10 to 25%)
strong:	>.5	(> 25%)

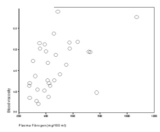
65

Size of correlation (Cohen, 1988)

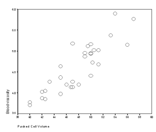
WEAK (.1 - .3)



MODERATE (.3-.5)



STRONG (>.5)

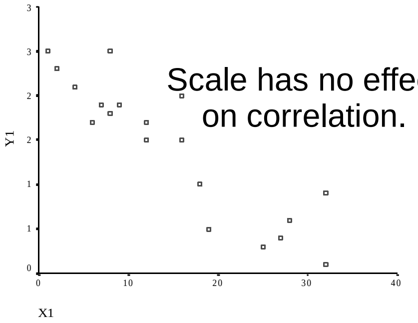


Interpreting correlation
(Evans, 1996)

Strength	***r***	***r²***
very weak	0 - .19	(0 to 4%)
weak	.20 - .39	(4 to 16%)
moderate	.40 - .59	(16 to 36%)
strong	.60 - .79	(36% to 64%)
very strong	.80 - 1.00	(64% to 100%)

67

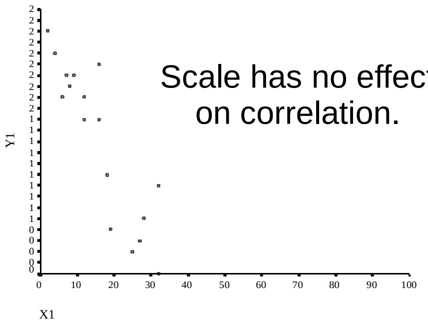
Correlation of this scatterplot = -.9



Scale has no effect on correlation.

68

Correlation of this scatterplot = -.9

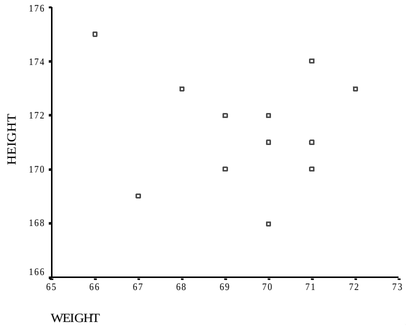


Scale has no effect on correlation.

69

What do you estimate the correlation of this scatterplot of height and weight to be?

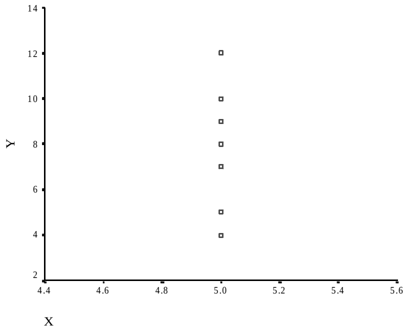
a. -.5
b. -1
c. 0
d. .5
e. 1



)

What do you estimate the correlation of this scatterplot to be?

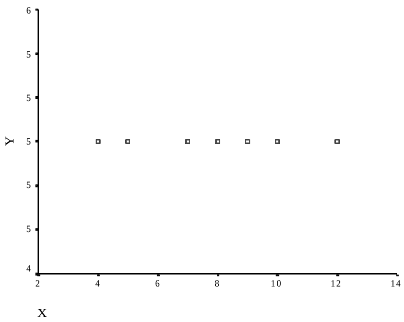
a. -.5
b. -1
c. 0
d. .5
e. 1



71

What do you estimate the correlation of this scatterplot to be?

a. -.5
b. -1
c. 0
d. .5
e. 1



2

Write-up: Example

"Number of children and marital satisfaction were inversely related (*r* (48) = -.35, *p* < .05), such that contentment in marriage tended to be lower for couples with more children. Number of children explained approximately 10% of the variance in marital satisfaction, a small-moderate effect (see Figure 1)."

73

Assumptions and limitations (Pearson product-moment linear correlation)

74

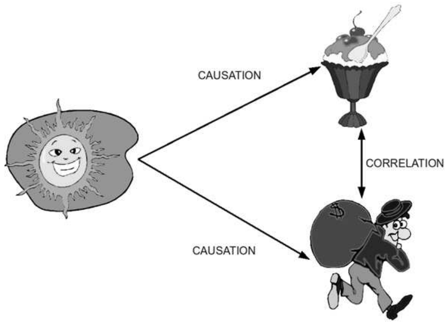
Assumptions and limitations

1. Levels of measurement ≥ interval
2. Correlation is not causation
3. Linearity
   1. Effects of outliers
   2. Non-linearity
4. Normality
5. Homoscedasticity
6. Range restriction
7. Heterogenous samples

75

Correlation is not causation e.g.,:

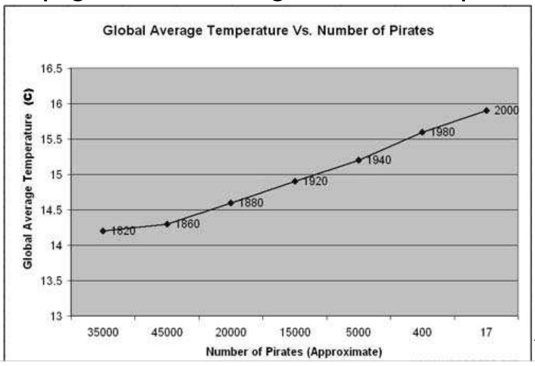
correlation between ice cream consumption and crime, but actual cause is temperature



76

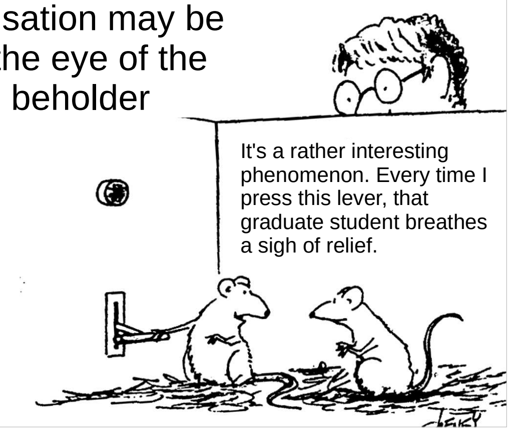
Correlation is not causation e.g.,:

Stop global warming: Become a pirate



77

Causation may be in the eye of the beholder

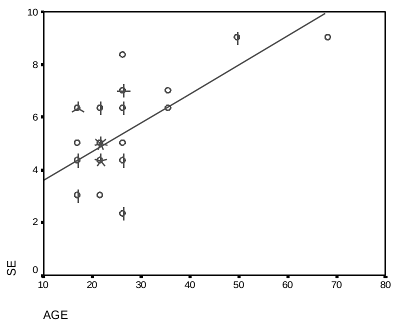


Effect of outliers

- Outliers can disproportionately increase or decrease *r*.
- Options
  - compute *r* with & without outliers
  - get more data for outlying values
  - recode outliers as having more conservative scores
  - transformation
  - recode variable into lower level of measurement

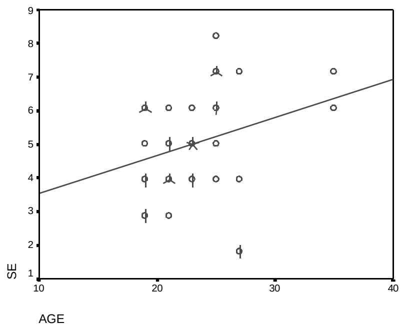
79

Age & self-esteem (*r* = .63)



80

Age & self-esteem (outliers removed) *r* = .23



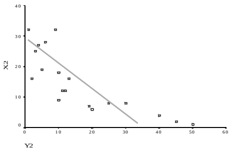
81

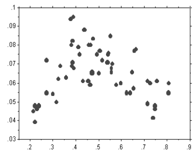
Non-linear relationships

Check scatterplot

Can a linear relationship ‘capture’ the lion’s share of the variance?

If so,use *r*.





Non-linear relationships

If non-linear, consider

- Does a linear relation help?
- Transforming variables to ‘create’ linear relationship
- Use a non-linear mathematical function to describe the relationship between the variables

Normality

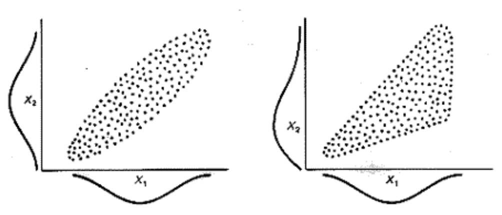
- The X and Y data should be sampled from populations with normal distributions
- Do not overly rely on a single indicator of normality; use histograms, skewness and kurtosis, and inferential tests (e.g., Shapiro-Wilks)
- Note that inferential tests of normality are overly sensitive when sample is large

Homoscedasticity

- Homoscedasticity refers to even spread about a line of best fit
- Heteroscedasticity refers to uneven spread about a line of best fit
- Assess visually and with Levene's test

85

Homoscedasticity



Homoscedasticity with both variables normally distributed **Heteroscedasticity with skewness on one variable**

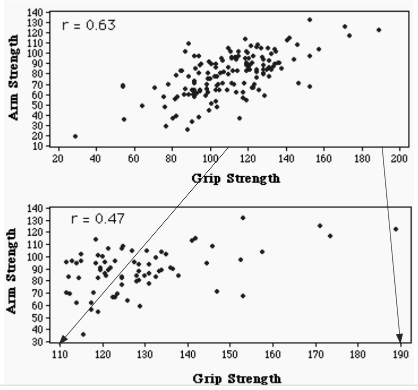
86

Range restriction

- Range restriction is when the sample contains restricted (or truncated) range of scores
  - e.g., cognitive capacity and age < 18 might have linear relationship
- If range restriction, be cautious in generalising beyond the range for which data is available
  - E.g., cognitive capacity does not continue to increase linearly with age after age 18

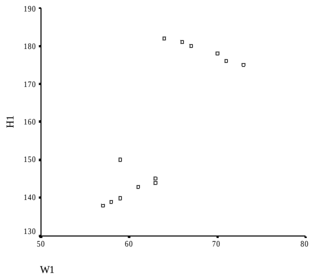
87

Range restriction



Heterogenous samples

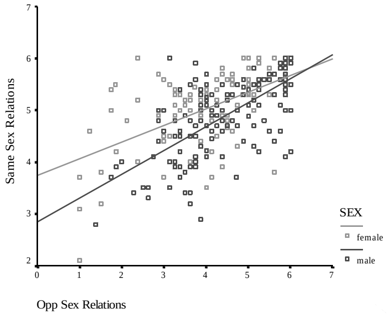
- Sub-samples (e.g., males & females) may artificially increase or decrease overall r.
- Solution - calculate r separately for sub-samples & overall, look for differences



Scatterplot of Same-sex & Opposite-sex Relations by Gender

♂ *r* = .67

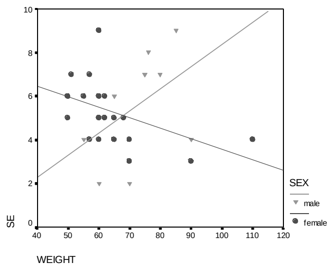
♀ *r* = .52



Scatterplot of Weight and Self-esteem by Gender

♂ $r = .50$

♀ $r = -.48$

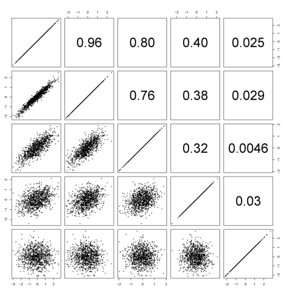


Dealing with several correlations

Dealing with several correlations

Scatterplot matrices organise scatterplots and correlations amongst several variables at once.

However, they are not detailed over for more than about five variables at a time.



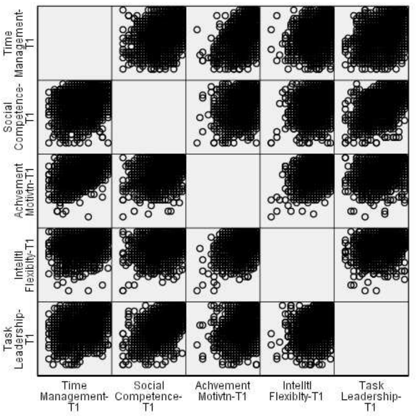
Correlation matrix: Example of an APA Style Correlation Table

Table 1.

Correlations Between Five Life Effectiveness Factors for Adolescents and Adults (N *= 3640)*

	Time Management	Social Competence	Achievement Motivation	Intellectual Flexibility	Task Leadership
Time Management		.36	.53	.31	.42
Social Competence			.37	.32	.57
Achievement Motivation				.42	.41
Intellectual Flexibility					.37
Task Leadership					

Scatterplot matrix



Summary

Key points

1. Covariations are the building blocks of more complex analyses, e.g., reliability analysis, factor analysis, multiple regression
2. Correlation does not prove causation – may be in opposite direction, co-causal, or due to other variables.

Key points

3. Choose measure of correlation and graphs based on levels of measurement.
4. Check graphs (e.g., scatterplot):
   - Outliers?
   - Linear?
   - Range?
   - Homoscedasticity?
   - Sub-samples to consider?

Key points

5. Consider effect size (e.g., Φ, Cramer's *V*, r, r^2) and direction of relationship
6. Conduct inferential test (if needed).

Key points

7. Interpret/Discuss
   - Relate back to research hypothesis
   - Describe & interpret correlation **(direction, size, significance)**
   - **Acknowledge limitations e.g.,**
     - Heterogeneity (sub-samples)
     - Range restriction
     - Causality?

100

References

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Howell, D. C. (2007). *Fundamental statistics for the behavioral sciences*. Belmont, CA: Wadsworth.

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Wadsworth.

101

Open Office Impress

- This presentation was made using Open Office Impress.
- Free and open source software.
- http://www.openoffice.org/product/impress.html

102