Correlation

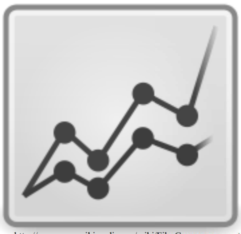


Image source: http://commons.wikimedia.org/wiki/File:Gnome-power-statistics.svg, GPL

Lecture 4

Survey Research & Design in Psychology

James Neill, 2017

Readings

Howitt & Cramer (2014)

- Ch 7: Relationships between two or more variables: Diagrams and tables
- Ch 8: Correlation coefficients: Pearson correlation and Spearman's rho
- Ch 11: Statistical significance for the correlation coefficient: A practical introduction to statistical inference
- Ch 15: Chi-square: Differences between samples of frequency data
- Note: Howitt and Cramer doesn't cover point bi-serial correlation

Overview

1. Covariation
2. Purpose of correlation
3. Linear correlation
4. Types of correlation
5. Interpreting correlation
6. Assumptions / limitations

Covariation

4

e.g., pollen and bees

e.g., study and grades

e.g., nutrients and growth

The world is made of co-variations

5

e.g., depictions of violence in the environment.

We can measure our observations and analyse the extent to which they co-occur.

e.g., psychological states such as stress and depression.

Co-variations are the basis of more complex models.

7

Purpose of correlation

8

Purpose of correlation

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

What is the

- **relationship** or
- **association** or
- **shared variance** or
- **co-relation**

between **two variables**?

9

Purpose of correlation

Other ways of expressing the underlying correlational question include:

To what extent do variables

- **covary**?
- **depend** on one another?
- **explain** one another?

10

Linear correlation

11

Linear correlation

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

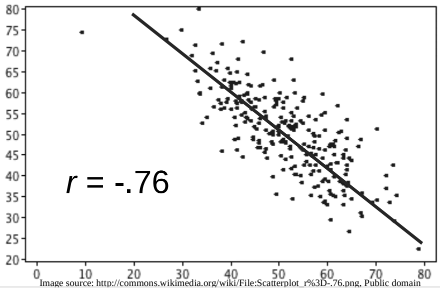


Image source: http://commons.wikimedia.org/wiki/File:Scatterplot_r%3D-.76.png, Public domain

12

Linear correlation

The linear relation between two variables is indicated by a correlation:

- **Direction:** Sign (+ / -) indicates direction of relationship (+ve or -ve slope)
- **Strength:** Size indicates strength (values closer to -1 or +1 indicate greater strength)
- **Statistical significance:** *p* indicates likelihood that the observed relationship could have occurred by chance

13

Types of relationships

- No relationship ($r \sim 0$) (X and Y are independent)
- Linear relationship (X and Y are dependent)
  - As X ↑s, so does Y ($r > 0$)
  - As X ↑s, Y ↓s ($r < 0$)
- Non-linear relationship

14

Types of correlation

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

15

Types of correlation

- Nominal by nominal:
  Phi (Φ) / Cramer's *V*, Chi-square
- 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*

16

Types of correlation and LOM

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

17

Nominal by nominal

18

Nominal by nominal correlational approaches

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

Contingency tables

- Bivariate frequency tables
- Marginal totals (blue)
- Observed cell frequencies (red)

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

Contingency table: Example

Snoring Do you snore? * Smokingr Smoking status Crosstabulation

Count

		Smokingr Smoking status		
		0 Non-smoker	1 Smoker	Total
Snoring Do you snore?	0 yes	50	16	66
	1 no	111	11	122
Total		161	27	188

BLUE = Marginal totals
RED = Cell frequencies

Contingency table: Example

χ^2 = sum of ((observed – expected)^2 / expected)

Snoring Do you snore? * Smokingr Smoking status Crosstabulation

			Smokingr Smoking status		
			0 Non-smoker	1 Smoker	Total
Snoring Do you snore?	0 yes	Count	50	16	66
		Expected Count	56.5	9.5	66.0
	1 no	Count	111	11	122
		Expected Count	104.5	17.5	122.0
Total		Count	161	27	188
		Expected Count	161.0	27.0	188.0

•Expected counts are the cell frequencies that should occur if the variables are not correlated.
•Chi-square is based on the squared differences between the actual and expected cell counts.

22

Cell percentages

Row and/or column cell percentages can also be useful
e.g., ~60% of smokers snore, whereas only ~30%[d] of non-smokers snore.

Snoring Do you snore? * Smokingr Smoking status Crosstabulation

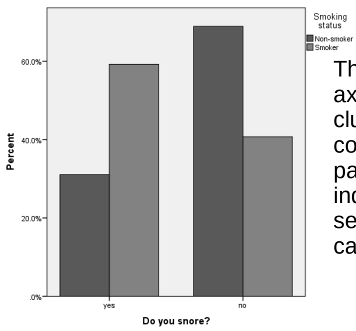
% within Smokingr Smoking status

		Smokingr Smoking status		
		0 Non-smoker	1 Smoker	Total
Snoring Do you snore?	0 yes	31.1%	59.3%	35.1%
	1 no	68.9%	40.7%	64.9%
Total		100.0%	100.0%	100.0%

23

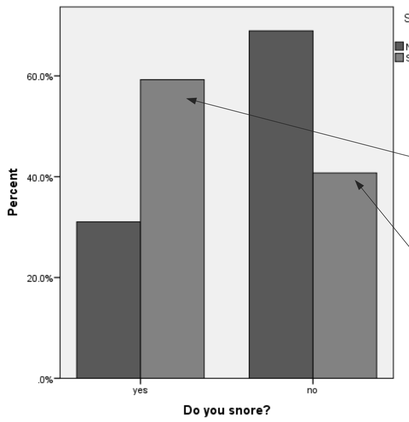
Clustered bar graph

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.

24



~60% of snorers are smokers, whereas only ~30% 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

Smoking (2) x Snoring (2)

Chi-Square Tests

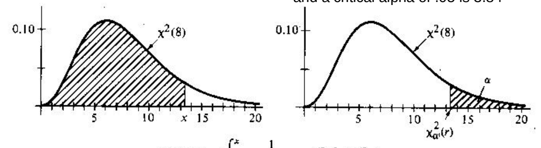
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.073[a]	1	.004		
Continuity Correction[b]	6.883	1	.009		
Likelihood Ratio	7.694	1	.006		
Fisher's Exact Test				.008	.005
Linear-by-Linear Association	8.030	1	.005		
N of Valid Cases	188				

Write-up: χ^2 (1, 188) = 8.07, *p* = .004

Chi-square distribution: Example

The Chi-Square Distribution

The critical value for chi-square with 1 *df* and a critical alpha of .05 is 3.84



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

	$P(X \le x)$							
	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 2 x 2, 2 x 3, 3 x 2 analyses
  e.g., Gender (2) & Pass/Fail (2)

Cramer's *V*

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

Phi (ϕ) & Cramer's *V*: Example

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-.207	.004
	Cramer's V	.207	.004
N of Valid Cases		188	

χ^2 (1, 188) = 8.07, *p* = .004, ϕ = .21

Note that the sign is ignored here (because nominal coding is arbitrary, the researcher should explain the direction of the relationship)

Ordinal by ordinal

Ordinal by ordinal correlational approaches

- Spearman's rho (r_s)
- Kendall tau (τ)
- Alternatively, use nominal by nominal techniques (i.e., recode the variables or treat them as having a lower level of measurement)

Graphing ordinal by ordinal data

- Ordinal by ordinal data is difficult to visualise because its non-parametric, with many points.
- Consider using:
  - Non-parametric approaches (e.g., clustered bar chart)
  - Parametric approaches (e.g., scatterplot with line of best fit)

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

34

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

35

Ordinal correlation example

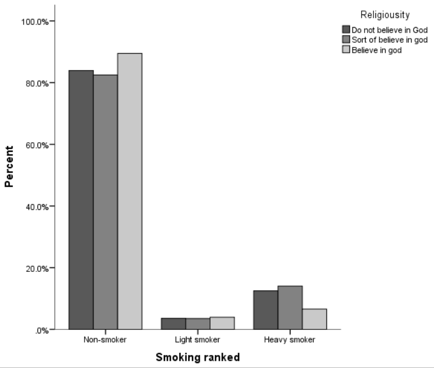
Godranked Religiousity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Do not believe in God	56	29.5	29.5	29.5
	1 Sort of believe in god	57	30.0	30.0	59.5
	2 Believe in god	77	40.5	40.5	100.0
	Total	190	100.0	100.0	

Smokingranked Smoking ranked

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Non-smoker	162	85.3	85.7	85.7
	1 Light smoker	7	3.7	3.7	89.4
	2 Heavy smoker	20	10.5	10.6	100.0
	Total	189	99.5	100.0	
Missing	System	1	.5		
Total		190	100.0		

Ordinal correlation example



Ordinal correlation example

Correlations

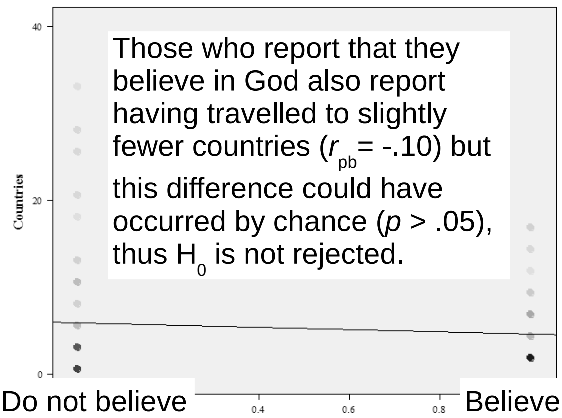
			Godranked Religiousity	Smokingranked Smoking ranked
Kendall's tau_b	Godranked Religiousity	Correlation Coefficient	1.000	-.071
		Sig. (2-tailed)	.	.298
		N	190	189
	Smokingranked Smoking ranked	Correlation Coefficient	-.071	1.000
		Sig. (2-tailed)	.298	.
		N	189	189

$\tau_b = -.07$, $p = .298$

Dichotomous by scale (interval/ratio)

Point-biserial correlation (r_{pb})

- One dichotomous & one interval/ratio variable
  - e.g., belief in god (yes/no) and number of countries visited
- Calculate as for Pearson's product-moment *r*
- Adjust interpretation to consider the direction of the dichotomous scales



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.

Do not believe — Believe

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

Correlations

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

Scale (interval/ratio) by Scale (interval/ratio)

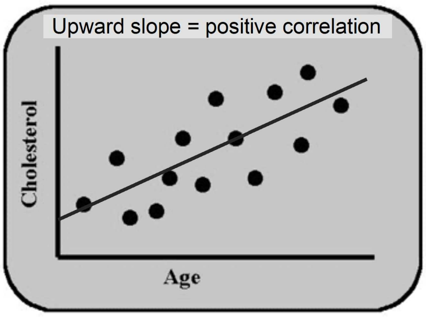
43

Scatterplot

- Plot each pair of observations (X, Y)
  - x = predictor variable (independent; IV)
  - y = criterion variable (dependent; DV)
- By convention:
  - IV on the x (horizontal) axis
  - DV on the y (vertical) axis
- Direction of relationship:
  - +ve = trend from bottom left to top right
  - -ve = trend from top left to bottom right

44

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



45

Line of best fit

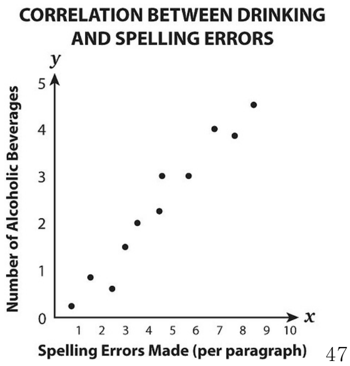
- The correlation between 2 variables is a measure of the degree to which pairs of numbers (points) cluster together around a best-fitting straight line
- Line of best fit: y = a + bx
- Check for:
  - outliers
  - linearity

46

What's wrong with this scatterplot?

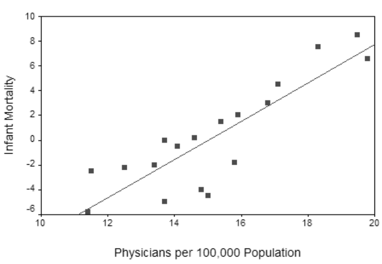
Y-axis should be DV (outcome)

X-axis should be IV (predictor)



47

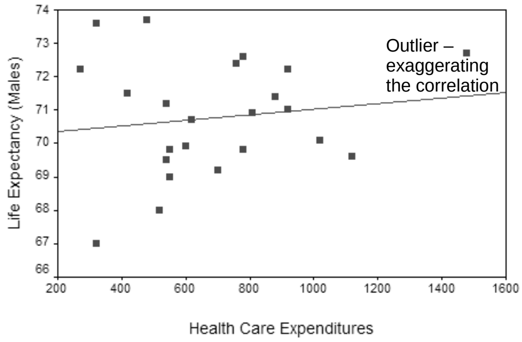
Scatterplot example: Strong positive (.81)



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

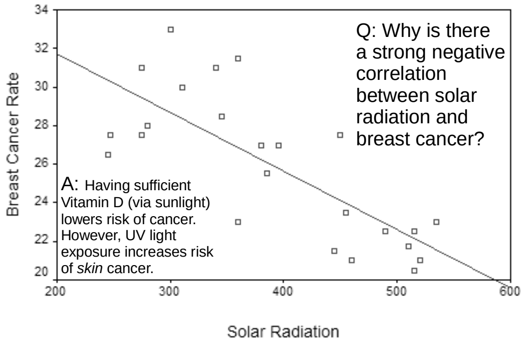
A: Because more doctors tend to be deployed to areas with infant mortality (socio-economic status aside).

Scatterplot example: Weak positive (.14)



49

Scatterplot example: Moderately strong negative (-.76)



50

Pearson product-moment correlation (*r*)

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

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

51

Covariance

- Variance shared by 2 variables

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

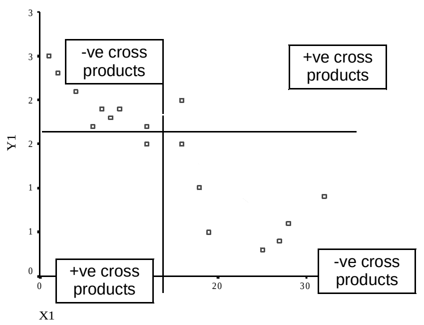
Cross products

N - 1 for the sample; *N* for the population

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

52

Covariance: Cross-products



53

Covariance → Correlation

- Size depends on the measurement scale → 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

54

Covariance, *SD*, and correlation: Example quiz question

The covariance between *X* and *Y* is 1.2. The *SD* of *X* is 2 and the *SD* of *Y* is 3. The correlation is:

a. 0.2
b. 0.3
c. 0.4
d. 1.2

Answer:
1.2 / 2 x 3 = 0.2

55

Hypothesis testing

Almost all correlations are not 0.

So, hypothesis testing seeks to answer:

- What is the **likelihood** that an observed relationship between two variables is “true” or “real”?
- What is the **likelihood** that an observed relationship is simply due to chance?

56

Significance of correlation

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

57

How to test the null hypothesis

- Select a critical value (alpha (α)); commonly .05
- Use a 1- or 2-tailed test; 1-tailed if hypothesis is directional
- Calculate correlation and its *p* value. Compare to the critical alpha value.
- If *p* < critical alpha, correlation is statistically significant, i.e., there is less than critical alpha chance that the observed relationship is due to random sampling variability.

58

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).

59

Errors in hypothesis testing

- **Type I error**:
  decision to reject $\mathbf{H_0}$ when H_0 is true
- **Type II error**:
  decision to not reject $\mathbf{H_0}$ when H_0 is false
- A significance test outcome depends on the statistical power which is a function of:
  - Effect size (*r*)
  - Sample size (*N*)
  - Critical alpha level (α_{crit})

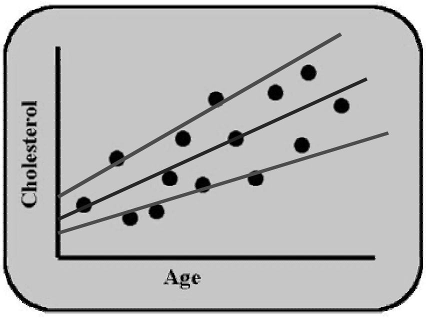
60

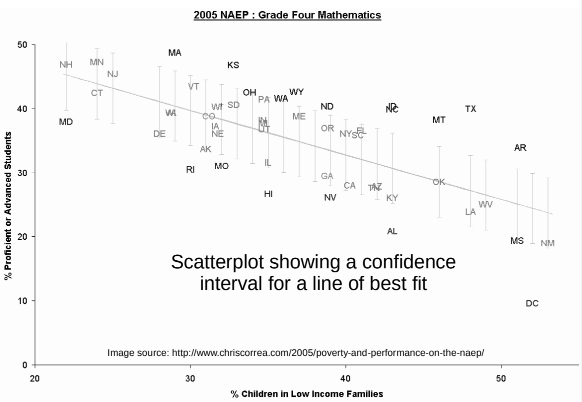
Significance of correlation

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 higher the *N*, the smaller the correlation required for a statistically significant result – why?

Scatterplot showing a confidence interval for a line of best fit





Practice quiz question:
Significance of correlation

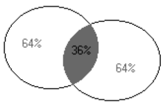
If the correlation between Age and Performance is statistically significant, it means that:

a. there is an important relationship between the variables
b. the true correlation between the variables in the population is equal to 0
c. the true correlation between the variables in the population is not equal to 0
d. getting older causes you to do poorly on tests

Interpreting correlation

Coefficient of Determination (r^2)

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



Interpreting correlation
(Cohen, 1988)

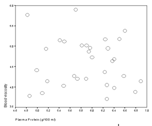
- A correlation is an **effect size**
- Rule of thumb:

Strength	***r***	***r*²**
Weak:	.1 - .3	1 - 9%
Moderate:	.3 - .5	10 - 25%
Strong:	>.5	> 25%

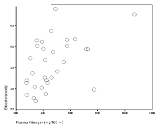
67

Size of correlation (Cohen, 1988)

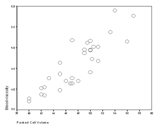
WEAK (.1 - .3)



MODERATE (.3 - .5)



STRONG (> .5)

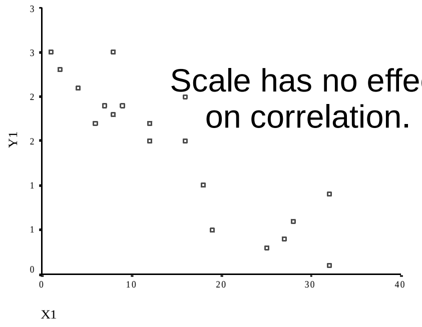


Interpreting correlation
(Evans, 1996)

Strength	***r***	***r*²**
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%)

69

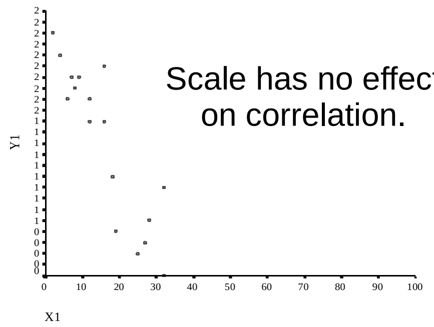
Correlation of this scatterplot = -.9



Scale has no effect on correlation.

70

Correlation of this scatterplot = -.9

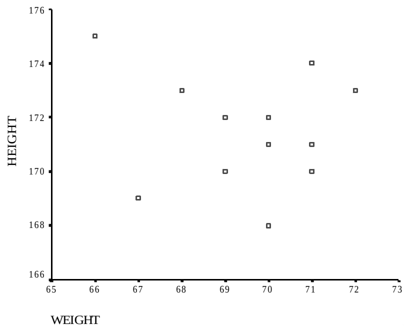


Scale has no effect on correlation.

71

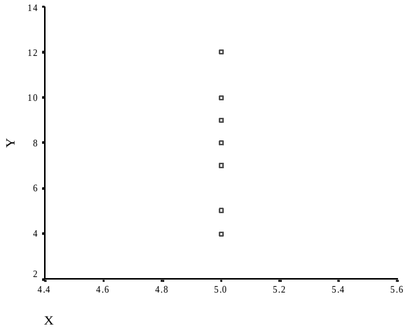
What do you estimate the correlation of this scatterplot of height and weight to be?

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



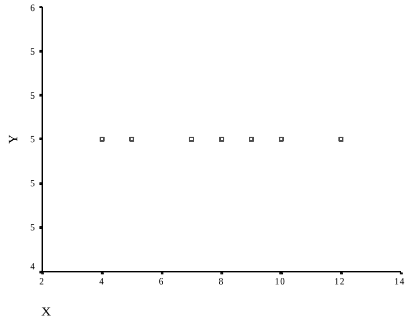
What do you estimate the correlation of this scatterplot to be?

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



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

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



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.”

Assumptions and limitations

(Pearson product-moment linear correlation)

Assumptions and limitations

1. Levels of measurement
2. Normality
3. Linearity
   1. Effects of outliers
   2. Non-linearity
4. Homoscedasticity
5. No range restriction
6. Homogenous samples
7. Correlation is not causation
8. Dealing with multiple correlations

Normality

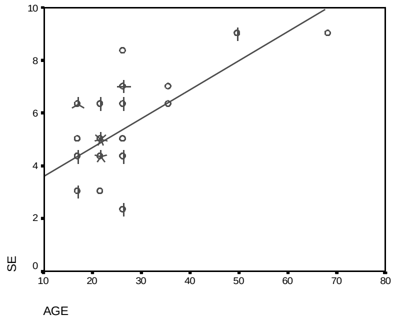
- The X and Y data should be sampled from populations with normal distributions
- Do not overly rely on any single indicator of normality; use histograms, skewness and kurtosis (e.g., within -1 and +1)
- Inferential tests of normality (e.g., Shapiro-Wilks) are overly sensitive when sample is large

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 and a non-parametric approach

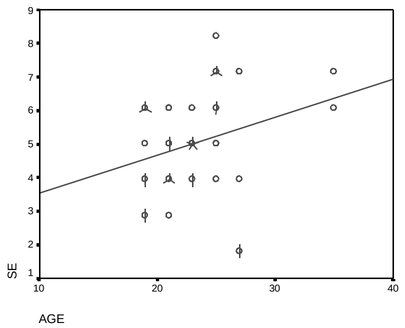
79

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



80

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



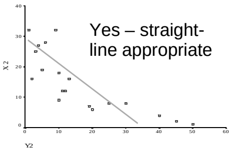
81

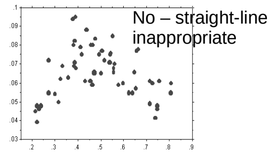
Non-linear relationships

Check scatterplot

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

If so,use *r*.





82

Non-linear relationships

If non-linear, consider:

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

83

Scedasticity

- **<u>Homo</u>**scedasticity refers to even spread of observations about a line of best fit
- **<u>Hetero</u>**scedasticity refers to uneven spread of observations about a line of best fit
- Assess visually and with Levene's test

84

Scedasticity

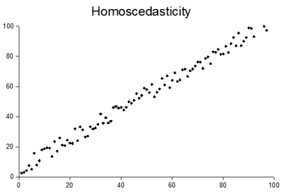


Image source:
https://commons.wikimedia.org/wiki/File:Homoscedasticity.png

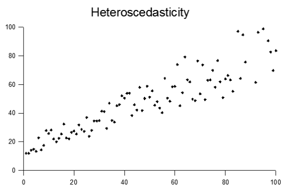
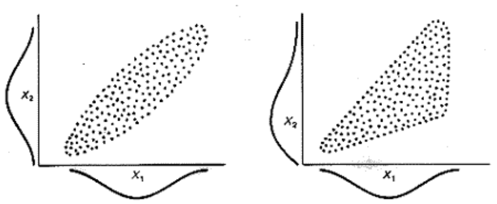


Image source:
https://commons.wikimedia.org/wiki/File:Heteroscedasticity.png

Scedasticity



Homoscedasticity with both variables normally distributed

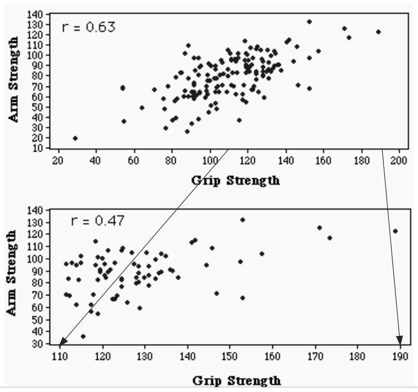
Heteroscedasticity with skewness on one variable

Image source: Unknown

Range restriction

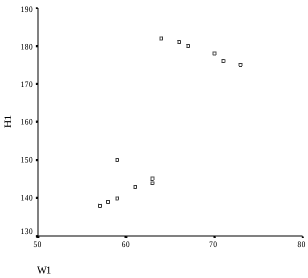
- Range restriction is when the sample contains a restricted (or truncated) range of scores
  - e.g., level of hormone X and age < 18 might have linear relationship
- If range is restricted, be cautious about generalising beyond the range for which data is available
  - e.g., level of hormone X may not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

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



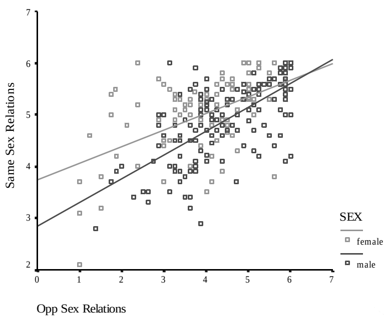
Scatterplot of Same-sex & Opposite-sex Relations by Gender

♂ $r = .67$

$r^2 = .45$

♀ $r = .52$

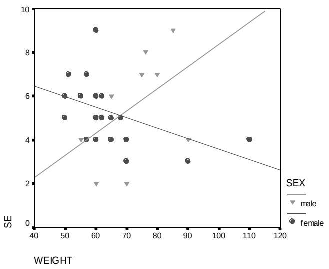
$r^2 = .27$



Scatterplot of Weight and Self-esteem by Gender

♂ *r* = .50

♀ *r* = -.48



Correlation is not causation e.g.,:

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

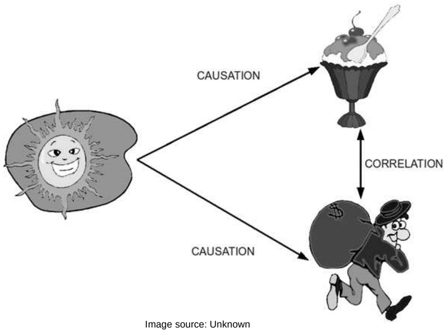
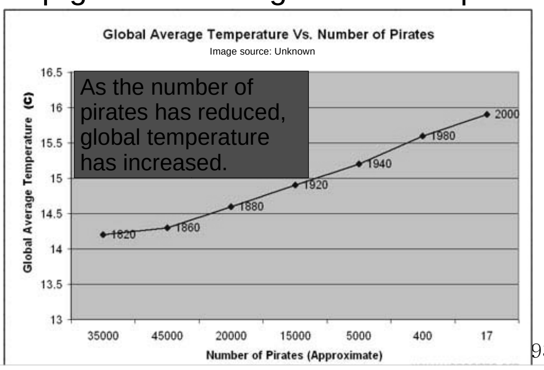


Image source: Unknown

Correlation is not causation e.g.,: Stop global warming: Become a pirate



Dealing with several correlations

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

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

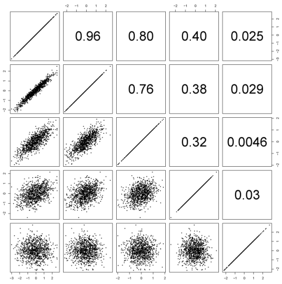


Image source: Unknown

Correlation matrix: Example of an APA Style Correlation Table

Table 1.

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

	Time Manage-ment	Social Compet-ence	Achieve-ment 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					

Summary

Summary: Correlation

1. The world is made of covariations.
2. Covariations are the building blocks of more complex multivariate relationships.
3. Correlation is a standardised measure of the covariance (extent to which two phenomenon co-relate).
4. Correlation does not prove causation - may be opposite causality, bi-directional, or due to other variables.

97

Summary: Types of correlation

- Nominal by nominal: Phi (Φ) / Cramer's *V,* Chi-square
- 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*

98

Summary: Correlation steps

1. Choose measure of correlation and graphs based on levels of measurement.
2. Check graphs (e.g., scatterplot):
   - Linear or non-linear?
   - Outliers?
   - Homoscedasticity?
   - Range restriction?
   - Sub-samples to consider?

99

Summary: Correlation steps

3. Consider
   - Effect size (e.g., Φ, Cramer's *V*, *r*, r^2)
   - Direction
   - Inferential test (*p*)
4. Interpret/Discuss
   - Relate back to hypothesis
   - Size, direction, significance
   - Limitations e.g.,
     - Heterogeneity (sub-samples)
     - Range restriction
     - Causality?

Summary: Interpreting correlation

- Coefficient of determination
  - Correlation squared
  - Indicates % of shared variance

Strength	***r***	r^2
Weak:	.1 - .3	1 – 10%
Moderate:	.3 - .5	10 - 25%
Strong:	> .5	> 25%

Summary: Asssumptions & limitations

1. Levels of measurement
2. Normality
3. Linearity
4. Homoscedasticity
5. No range restriction
6. Homogenous samples
7. Correlation is not causation
8. Dealing with multliple correlations

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.

Howitt, D. & Cramer, D. (2011). *Introduction to statistics in psychology* (5th ed.). Harlow, UK: Pearson.

Open Office Impress

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