CHARGE!
YOU'LL NEVER REACH YOUR OBJECTIVES IF YOU DON'T GET OFF YOUR ASSUMPTIONS
©2007

Lancaster University

Correlation

Assumptions, issues and intercorrelation

Dr. Margriet A. Groen

Outline: Aims and Objectives

Assumptions and issues with correlation analysis
variables, linearity, normality, outliers, range restrictions

Overcoming such problems Spearman’s Rho

More complex correlation analysis Inter-correlation

Correlation Assumptions

- In order to conduct a correlation analysis, the data must meet pre-specified assumptions.

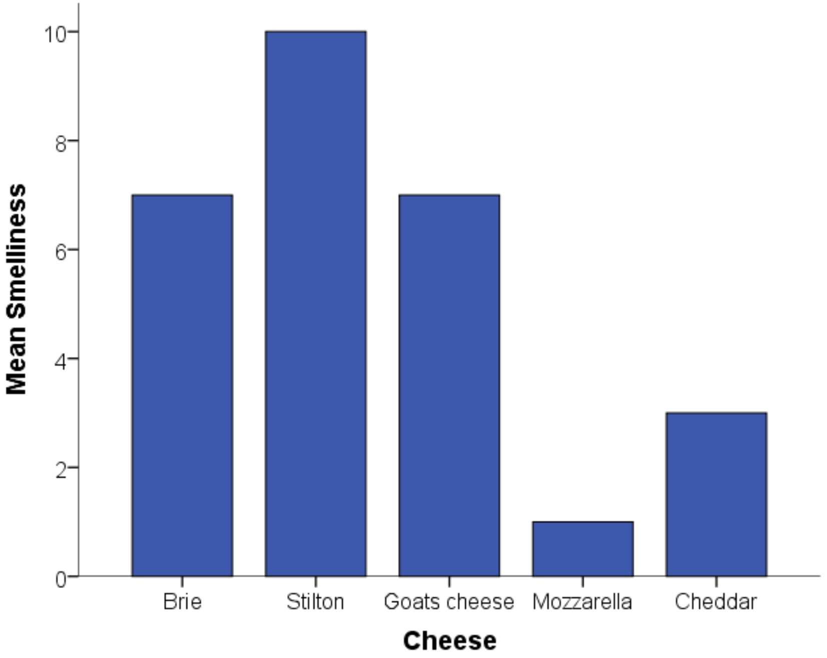
- If any of these assumptions are violated, then Pearson's *r* may provide misleading information regarding the relationship between the two variables of interest.

Correlation Assumptions
Type of variables

1. Correlation describes the relationship between <u>equal interval numeric</u> variables

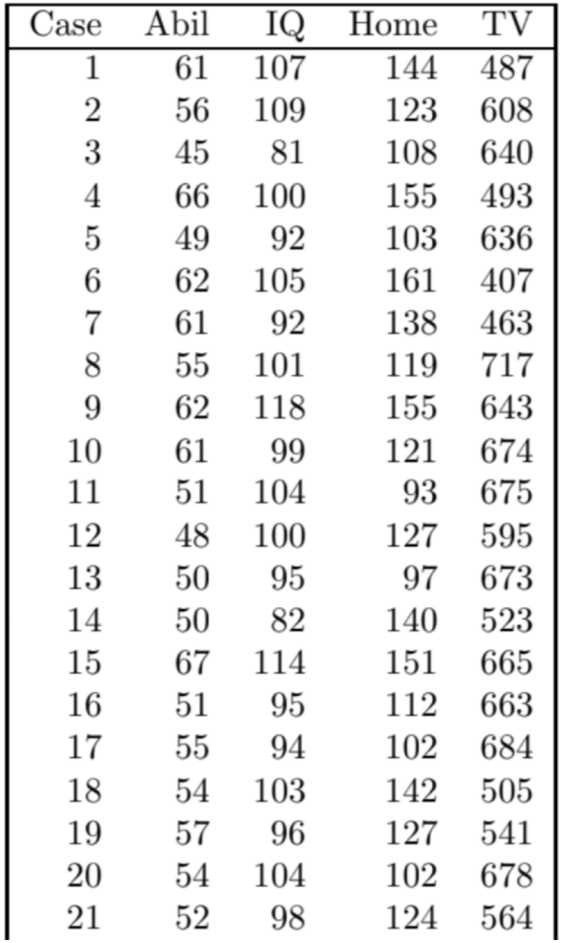
- Variations in *X* and *Y* should relate to the variation in the **magnitude** of the variables, and not variations between different categories

- **Example:** you would not try to correlate a continuous variable *(mean smelliness)* with a categorical variable *(type of cheese).*

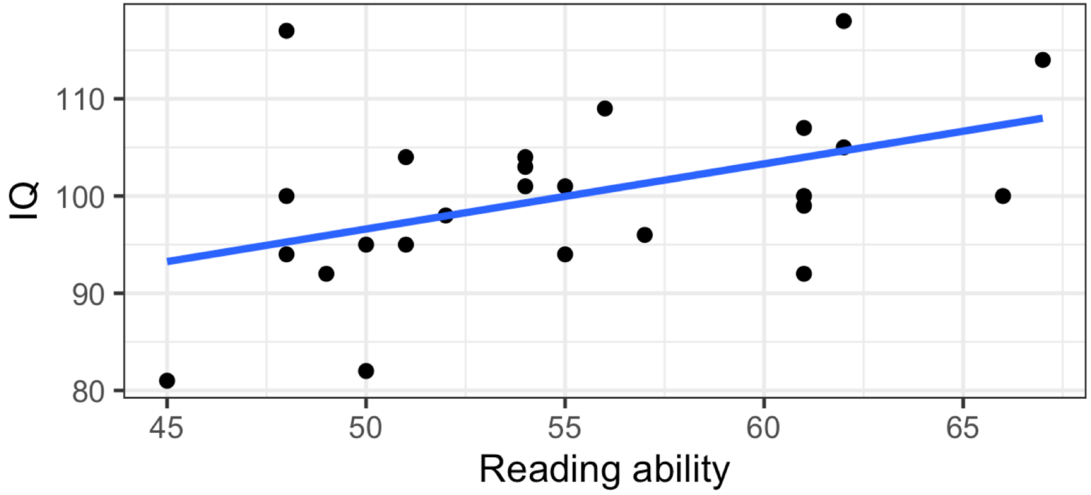


Correlation Assumptions
Missing data

2. Is there a data point for each participant on both variables?



Case	Abil	IQ	Home	TV
1	61	107	144	487
2	56	109	123	608
3	45	81	108	640
4	66	100	155	493
5	49	92	103	636
6	62	105	161	407
7	61	92	138	463
8	55	101	119	717
9	62	118	155	643
10	61	99	121	674
11	51	104	93	675
12	48	100	127	595
13	50	95	97	673
14	50	82	140	523
15	67	114	151	665
16	51	95	112	663
17	55	94	102	684
18	54	103	142	505
19	57	96	127	541
20	54	104	102	678
21	52	98	124	564

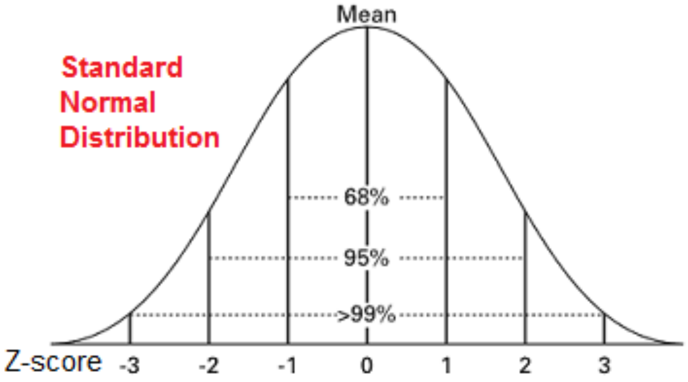


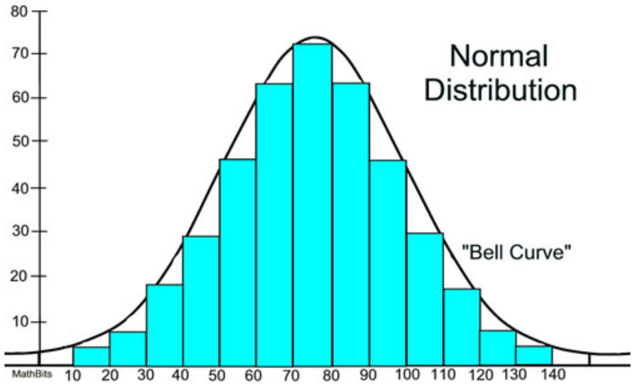
Correlation Assumptions
Normality (1)

3. Data should be normally distributed

If either of the distributions (*X* or *Y*) are not normal, the correlation may be distorted.

You can do a visual check of this by looking at a histogram.

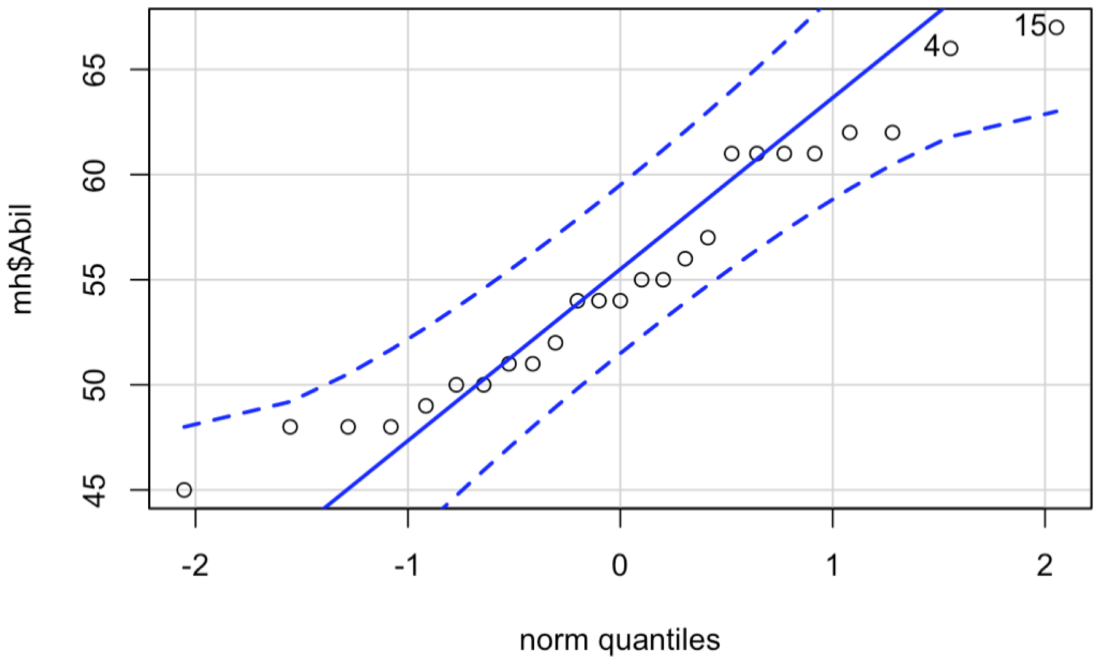




Correlation Assumptions
Normality (2)

3. Data should be normally distributed

The 'normal' quantile-quantile plot (or qq-plot) also indicates whether a variable is normally distributed.

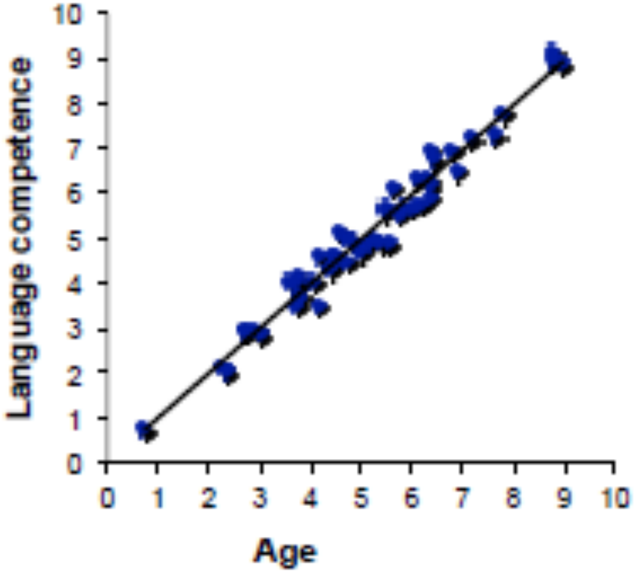


Correlation Assumptions
Linearity (1)

4. Correlation analysis assumes that the relationship between X and Y is linear (a straight line).

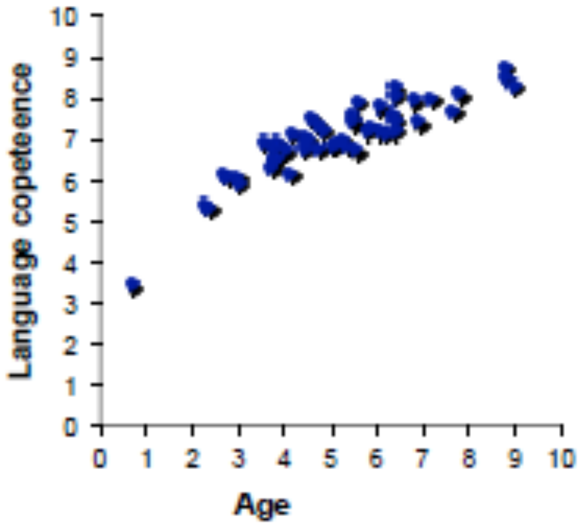
Note: This doesn't mean that all of the points need to fall on the straight line, rather, the general 'trend' should be described by a straight line *e.g. positive / negative relationship (a null relationship is a horizontal straight line)*

Example: the relationship between age and language skill is linear

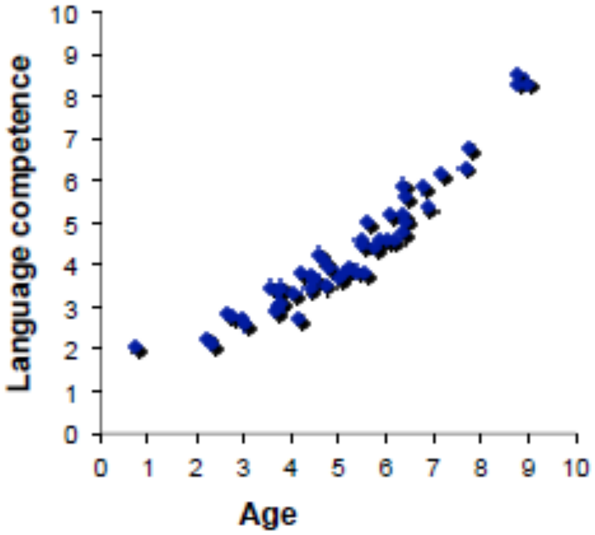


Correlation Assumptions
Linearity (2)

- However, correlation analysis can not provide a full picture of ***curvilinear*** relations.
- **Example:** the relationship varies in different aspects of language



Competence in some aspects (e.g., vocabulary) grows rapidly and then flattens out



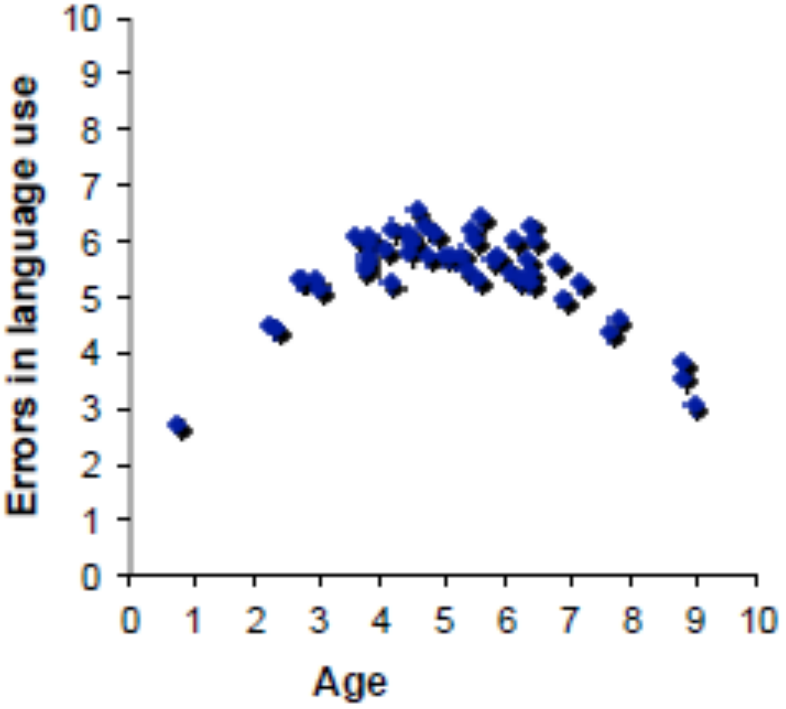
Whereas in others (e.g., compound sentences) there is a later growth spurt

Correlation Assumptions
Linearity (3)

…and in others (language errors) there are more odd developmental patterns

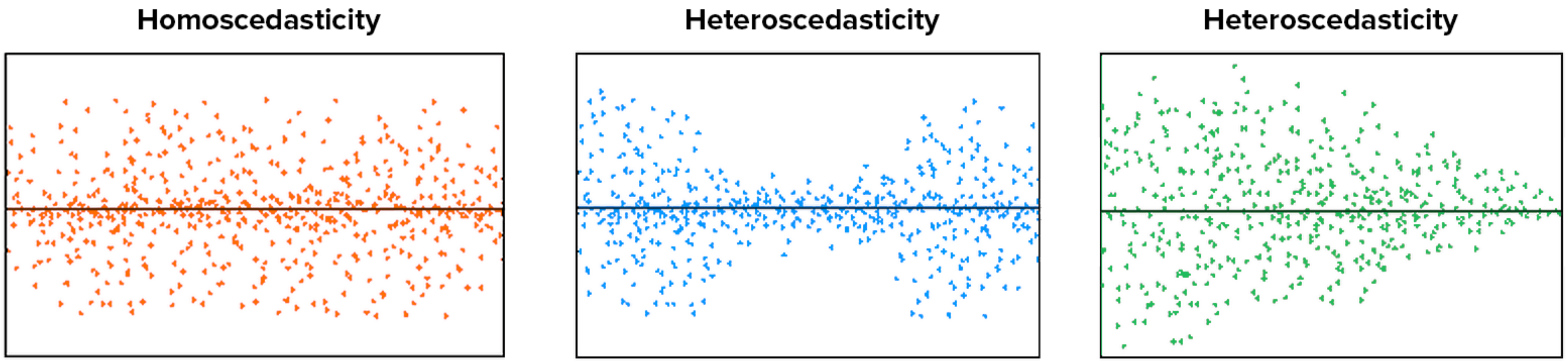
In this correlation r = -.18, which is **not significant**. However, looking at the scatterplot, we can see a clear relationship between the variables.

Curvilinear relationships will be discussed in more depth next year



Correlation Assumptions
Homoscedasticity

5. Does the spread have homoscedasticity?



Homoscedasticity

Heteroscedasticity

Heteroscedasticity

Random Cloud (No Discernible Pattern)

Bow Tie Shape (Pattern)

Fan Shape (Pattern)

Correlation Issues
Dealing with distribution problems

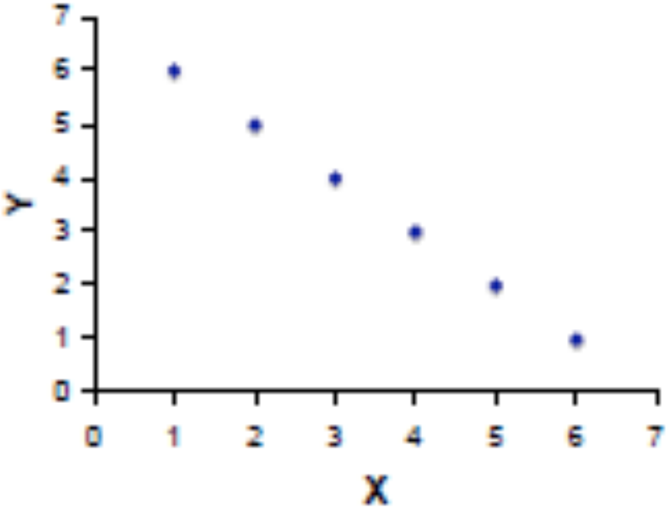
- When dealing with data which is not normally distributed, a non-parametric test correlation coefficient can be used ***(Spearman's Rho)***.
- ***Spearman's Rho*** is based on the ranking of scores (lowest→highest)

X	Y
15	10
17	12
20	11
22	15
25	15

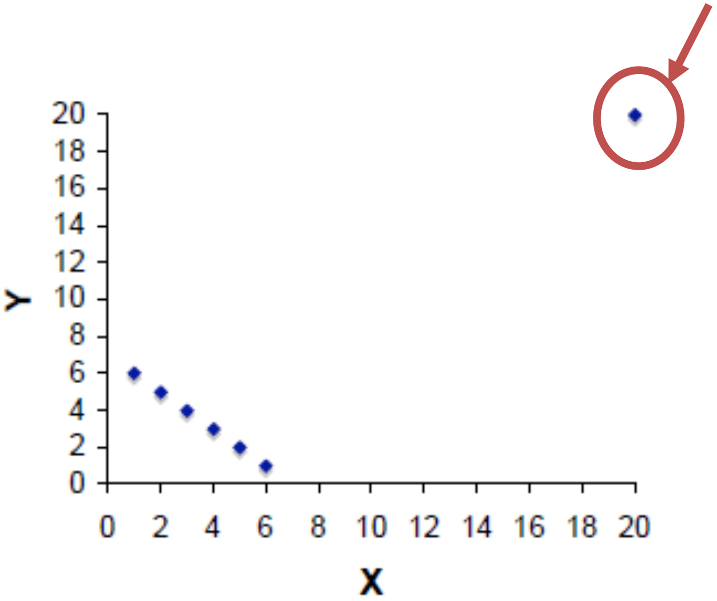
Correlation Issues
Outliers

As with non-parametric data, correlation may also be distorted by outliers with extreme scores (usually more than 3 SD's away from the mean)

***r* = -1**



***r* = .86**

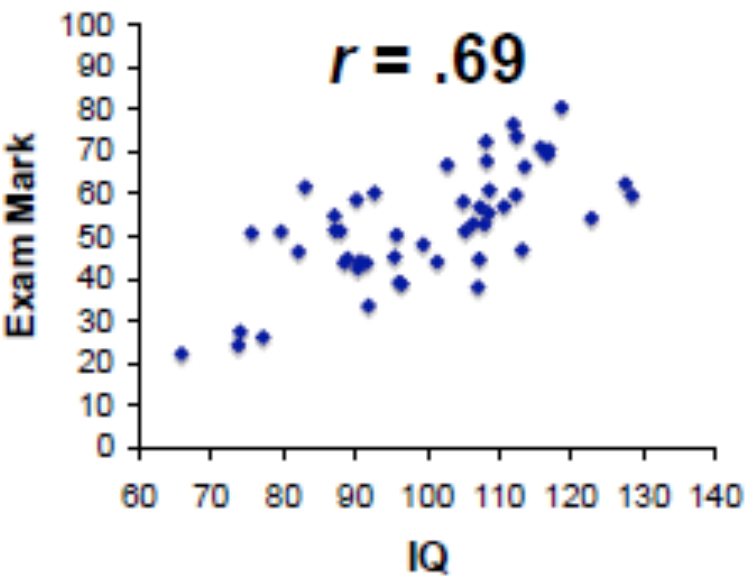


Outlier

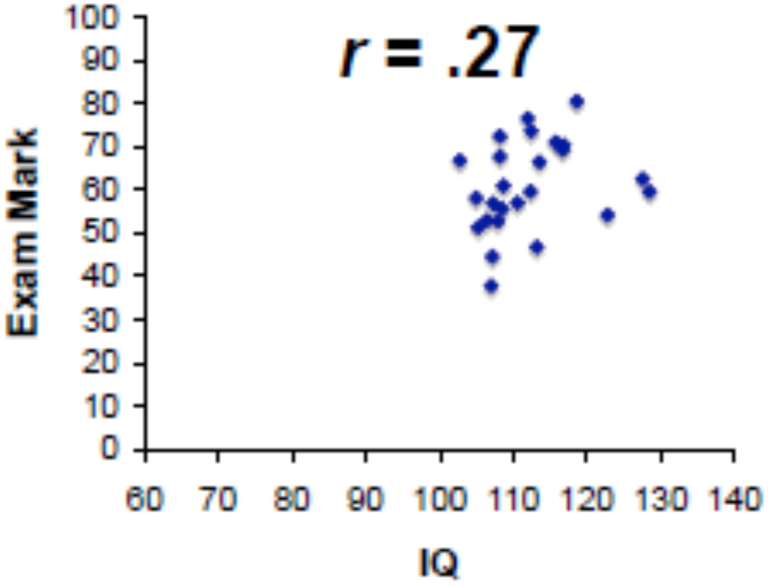
Correlation Issues
Range restrictions (1)

The sample used may not represent the true variation present in the two variables present in the population

Exam mark is likely to be correlated with IQ if it was measured across the whole population

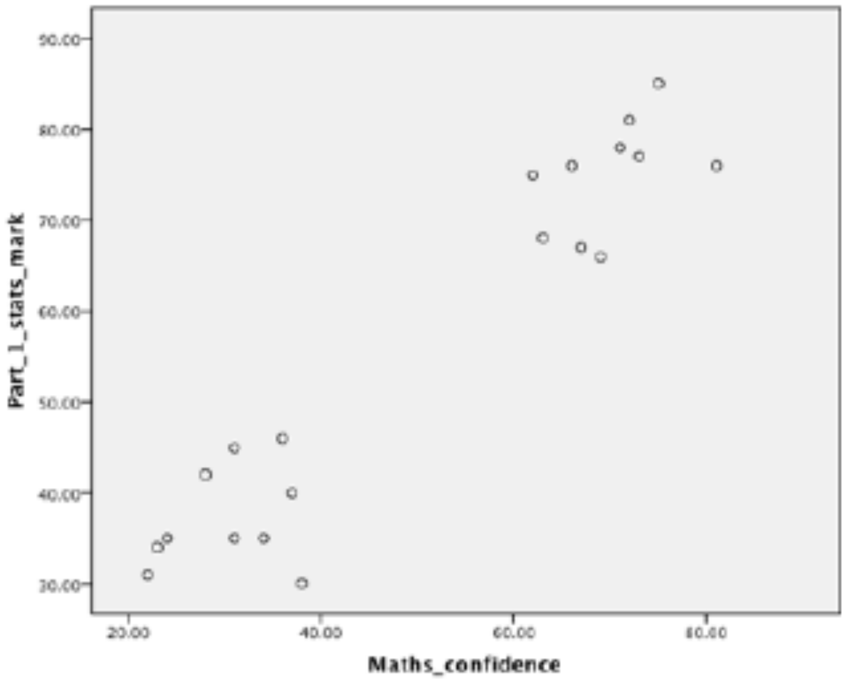


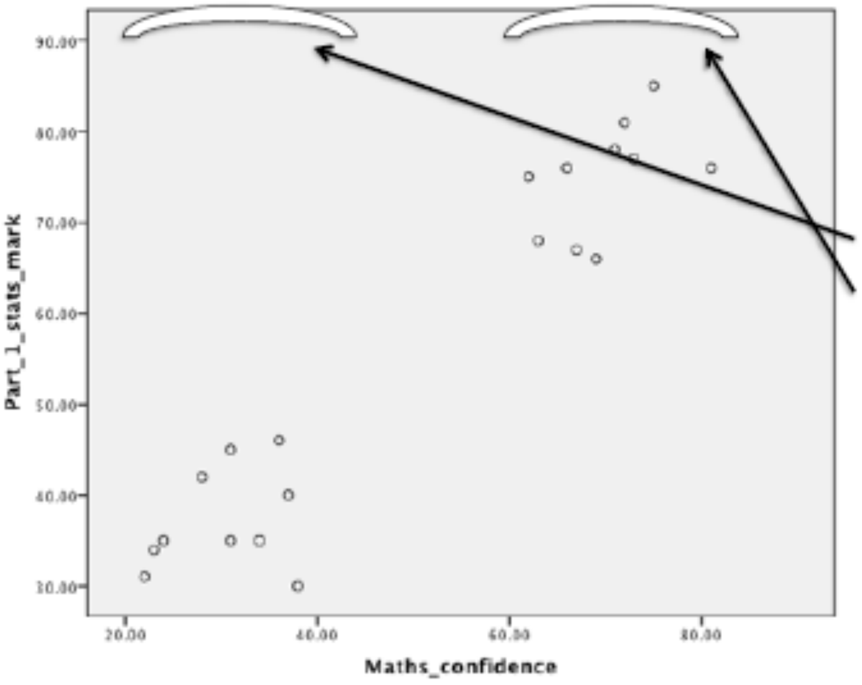
…but the correlation is likely to be much smaller if only Lancaster students are considered



Correlation Issues
Range restrictions (2)

If one range on one variable is unusually large (and there are two very distinct clusters), it is sometimes more beneficial to create a new variable.

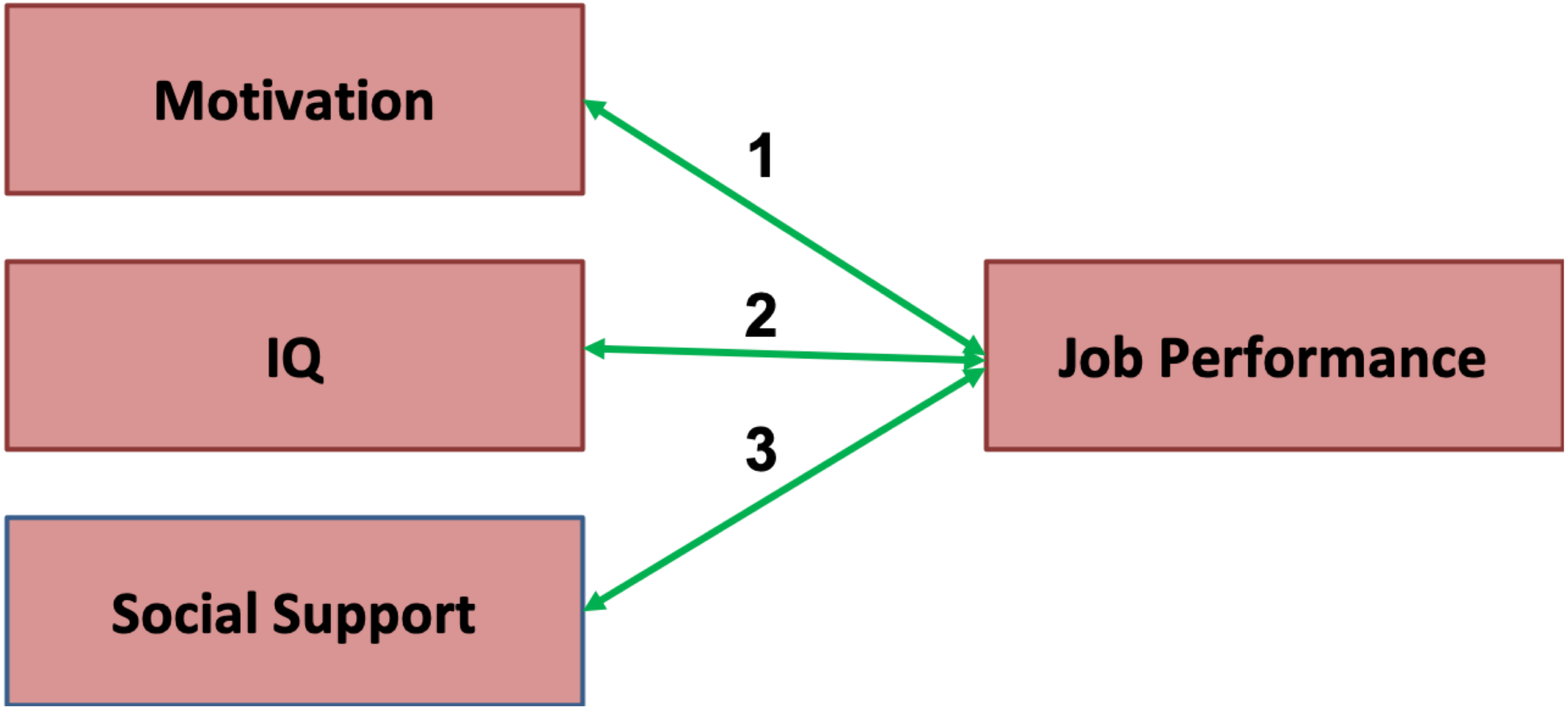




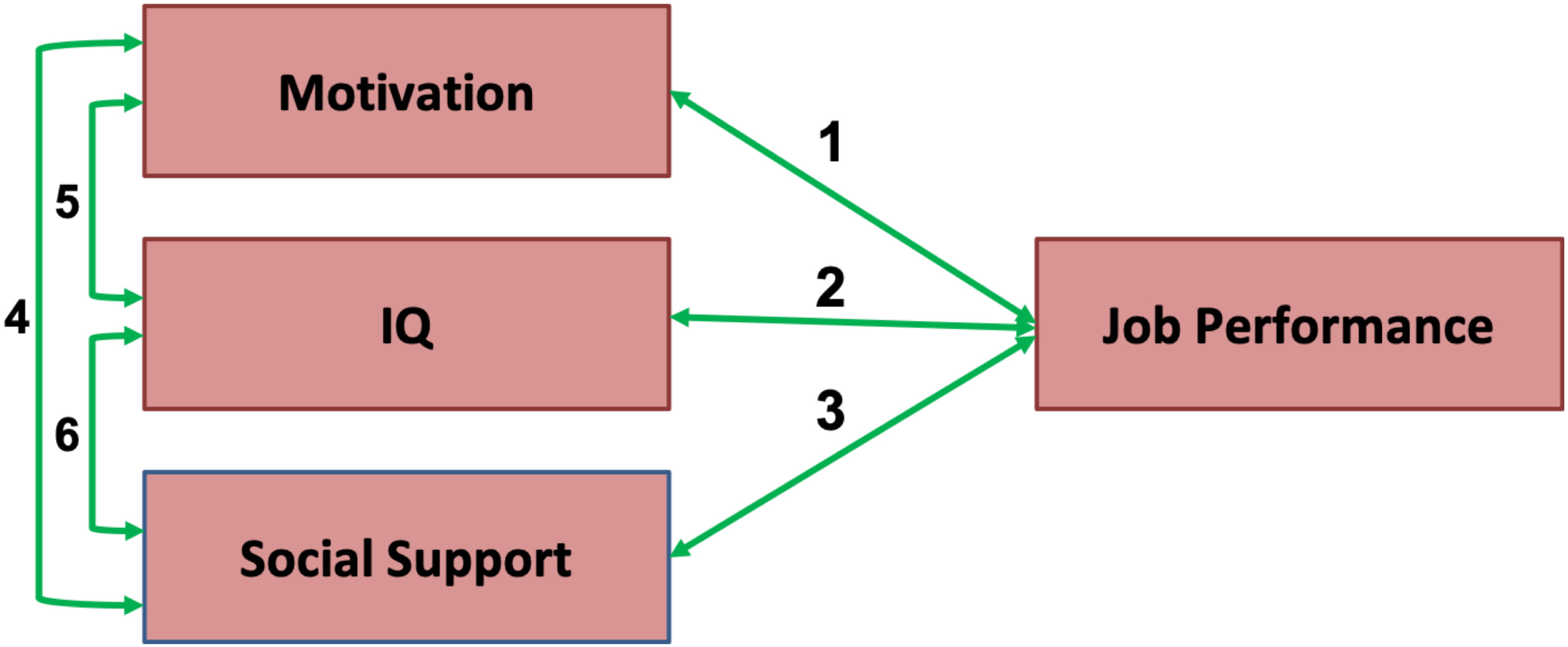
Create a new variable with
1 = low
2 = high
Maths Confidence

Intercorrelation (1)

What if you are interested in the association between more than just two variables?



Intercorrelation (2)



Intercorrelation
Constructing an APA correlation matrix

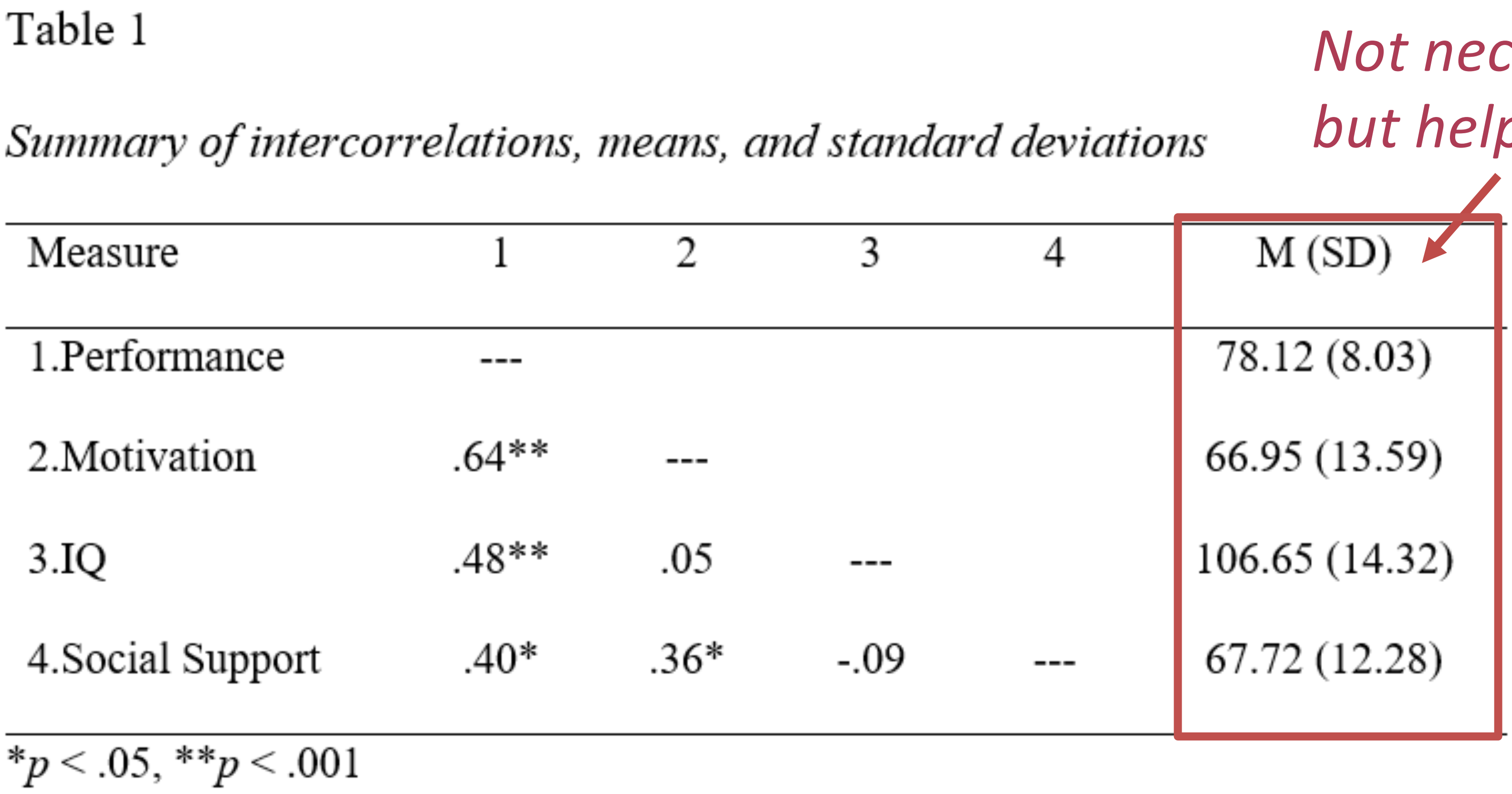


Table 1

Summary of intercorrelations, means, and standard deviations

Measure	1	2	3	4	M (SD)
1.Performance	---				78.12 (8.03)
2.Motivation	.64**	---			66.95 (13.59)
3.IQ	.48**	.05	---		106.65 (14.32)
4.Social Support	.40*	.36*	-.09	---	67.72 (12.28)

*$p < .05$, **$p < .001$

Not necessary but helpful!

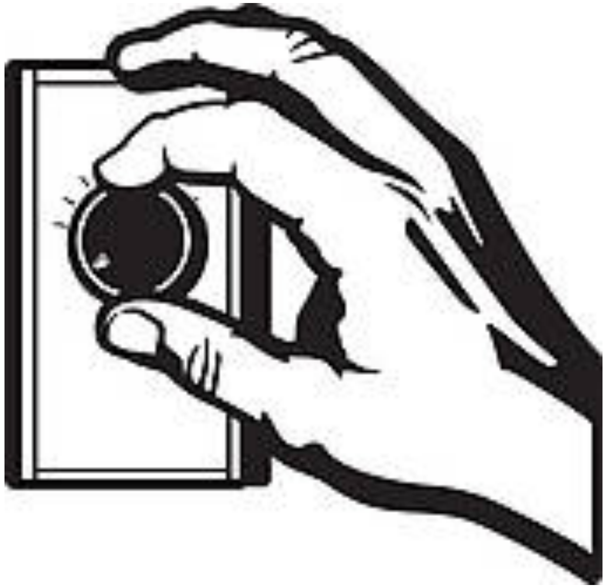
Intercorrelation
Type one error issues

- Before we tackle that: brief recap of what a *p*-value is …

Intercorrelation
Type one error issues (4)

Conducting many correlations at once increases the chances of Type 1 error. We therefore need to apply a ***Bonferroni adjustment*** which changes the significance level of *p*.

Bonferroni adjustment – The significance level is adjusted by dividing the normal significance value by number of tests performed.



*Each variable is correlated with **3** other variables* – ***0.05 / 3 = 0.016***

Therefore, for a correlation coefficient to be 'significant' p < .016

Summary

Assumptions

1. Variables need to be at interval level
2. Each participant needs to have a data point for each variable
3. Both variables need to normally distributed
4. The relationship needs to be linear
5. Spread needs to be homoscedastic

Issues Check the scatterplot for influence of outliers and possible range restrictions

Non-parametric correlation Spearman's Rho

More complex correlation analysis A correlation matrix can show you multiple correlations between variables at once, but beware of the multiple comparison problem (use the Bonferroni adjustment).