Correlation: Assumptions, issues and intercorrelation

Dr. Margriet Groen

Overview

- Assumptions and issues with correlation analysis
  - Variables, independence, bivariate normality
  - Outliers, range restrictions
- Overcoming such problems
  - Spearman's rho
- More complex correlation analysis
  - Intercorrelation

Correlation Assumptions

- A statistical test works by comparing your result to what we would expect to see if only chance were operating, and to describe that expectation we have to build a model of how the data behave.
- Assumptions are the conditions that model relies on.
- If the data meet them, the p-value is an accurate answer to “how surprising is this result?”, and if they don’t, the p-value may still be calculated but can be misleading.

Pearson's *r*: Assumptions

About how the data were collected

1. Both variables are continuous (interval or ratio)
2. Observations are independent of one another

About the shape of the data

3. The pairs of scores come from a **bivariate normal distribution**

In practice, we check what this implies:

- the relationship is a straight line (**linearity**)
- the spread of Y is the same at every value of X (**homoscedasticity**)
- the Y scores are roughly normal at every value of X (**normality**)

Types of variables

Nominal: Categories with no inherent order (e.g., eye colour, diagnosis type)
Ordinal: Categories with a meaningful order, but the gaps between them are not necessarily equal (e.g., level of education, mild/moderate/severe symptoms)
Interval: Ordered values with equal gaps between them, but no true zero (e.g., calendar year)
Ratio: Ordered, equal intervals, and a true zero (e.g., reaction times in ms, number of errors, height)

Assumption 1: Types of variables

Pearson's *r* describes the relationship between two variables on an interval or ratio scale

✘ Nominal ✘ Ordinal ✔ Interval ✔ Ratio

Example 1: You would **not** try to correlate a continuous variable (*smelliness*) with a categorical variable (*type of cheese*).

Example 2: You would **not** try to correlate a continuous variable (*reading ability*) with an ordinal variable (*highest level of education*)

Assumption 2: Independence of observations

Each participant's data should be independent of every other participant's data. Repeated measure from the same person, or data form related people (such as siblings), violate this assumption.

✔ Independent: 100 university students each complete a stress questionnaire and a sleep diary once.

✘ Not independent: The same 20 students complete both measures on 5 separate days, giving 100 data points.

Pearson's *r*: Assumptions

About how the data were collected

1. Both variables are continuous (interval or ratio) 2. Observations are independent of one another

About the shape of the data

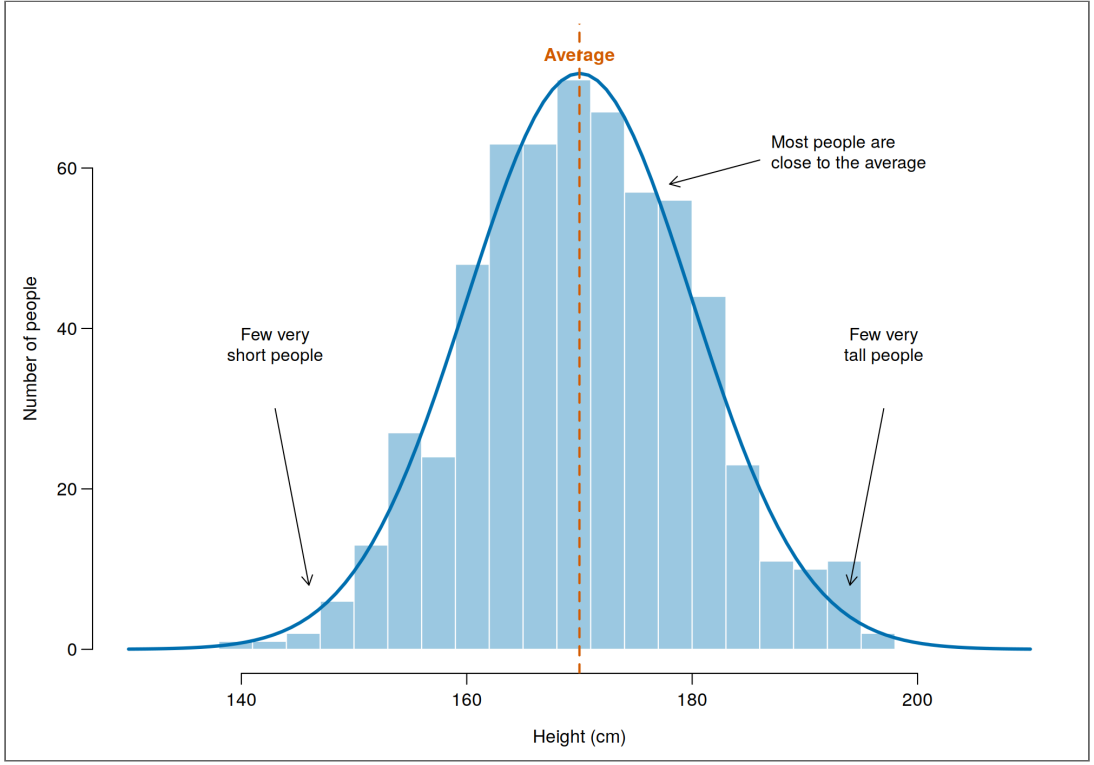
3. The pairs of scores come from a **bivariate normal distribution**

In practice, we check what this implies:

- the relationship is a straight line (**linearity**)
- the spread of Y is the same at every value of X (**homoscedasticity**)
- the Y scores are roughly normal at every value of X (**normality**)

What is a normal distribution?

- A **bell-shaped** pattern that many measurements follow
- Most values cluster around a **typical middle value**
- The further from the middle, the **rarer** the values become
- **Symmetrical**: being 10 cm taller than average is about as unusual as being 10 cm shorter



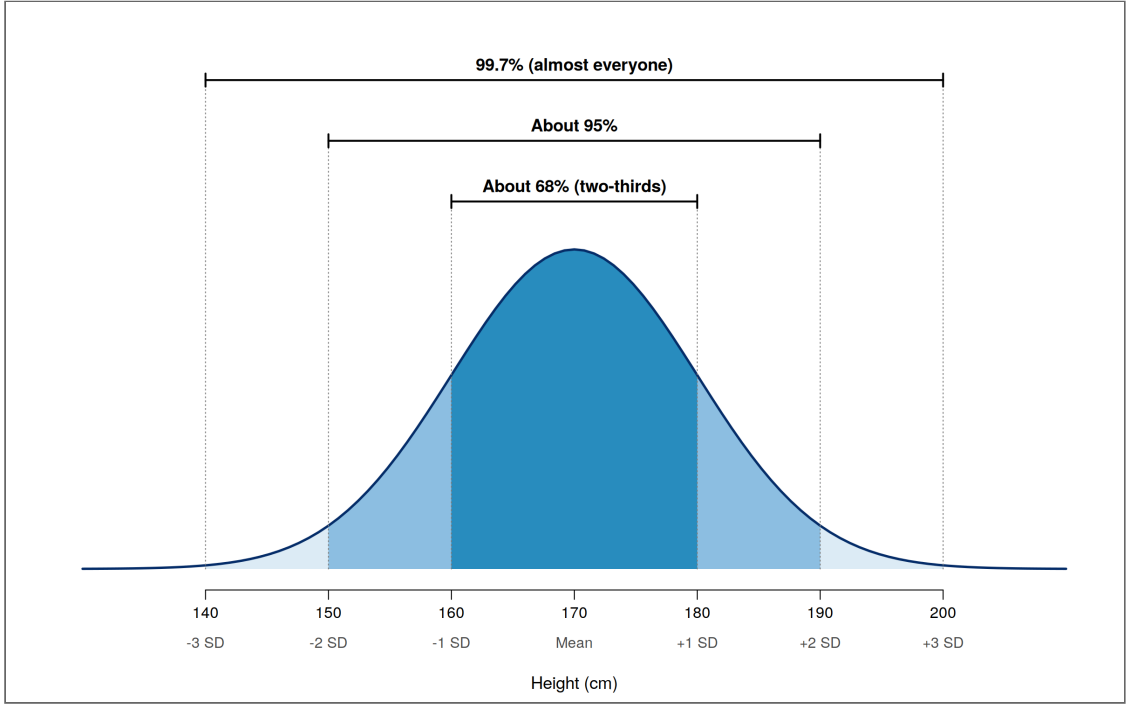
Example: Most people are close to average, and very tall or very short people are rare.

Two numbers describe the bell

- **Mean**: where the middle of the bell sits
- **Standard deviation (SD)**: how spread out the bell is, roughly the typical distance from the mean

Rule of thumb

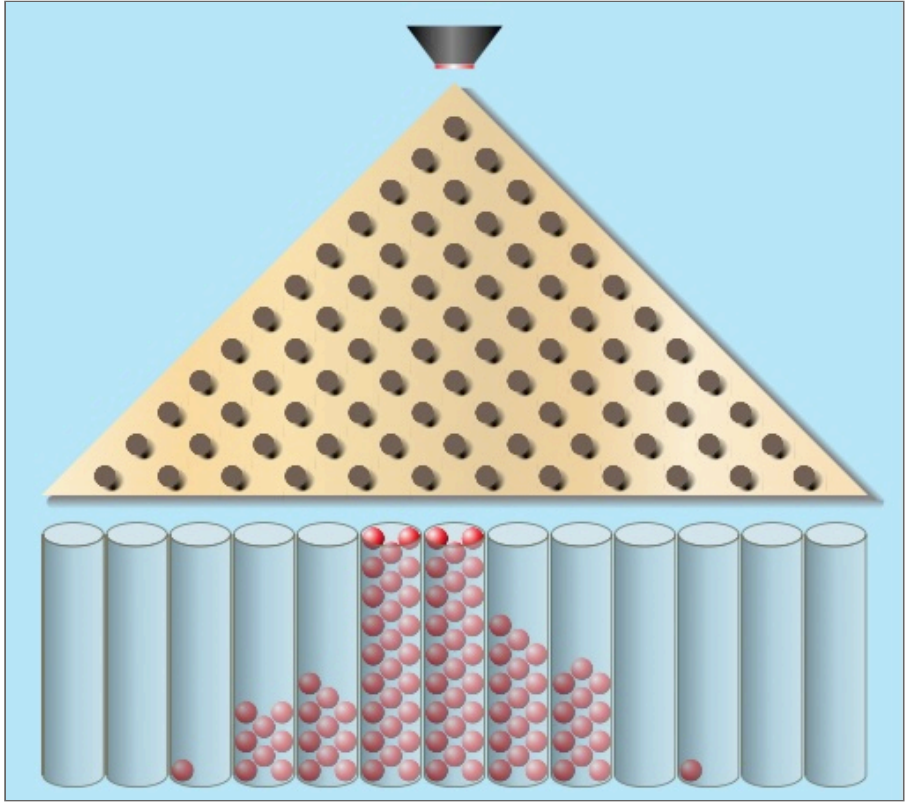
- About **two-thirds** of values fall within 1 SD of the mean
- About **95%** fall within 2 SDs
- Almost **all** fall within 3 SDs



Example: if height has a mean of 170 cm and an SD of 10 cm, about 95% of adults are between 150 cm and 190 cm.

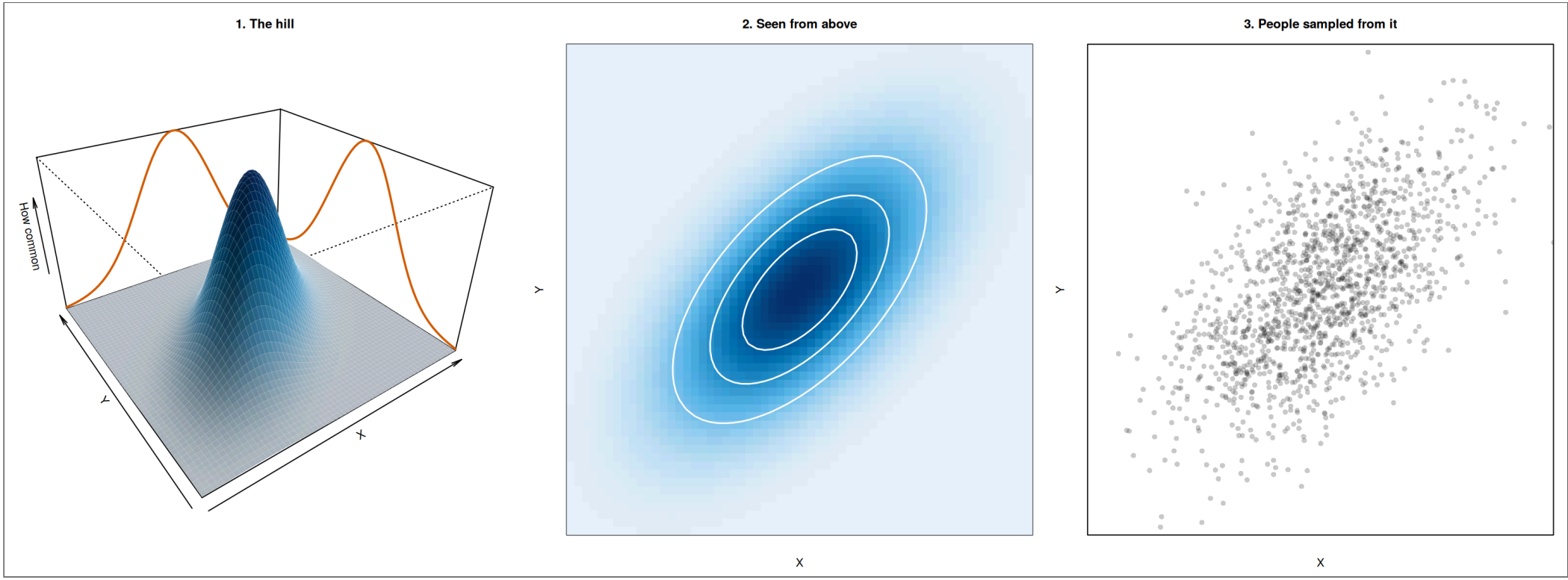
Why does the bell shape appear so often?

- The **central limit theorem** is a result from **probability theory**: When many small, independent influences are added up or averaged, the result is approximately **normally distributed**, whatever the shape of the individual influences
- In the 1870s Francis Galton invented what is referred to as the ‘Galton board’ to illustrate it
  - Balls fall through rows of pegs, bounce left or right at random, and pile up in bins at the bottom
  - The pile forms a bell shape: many small chance nudges add up to a normal distribution

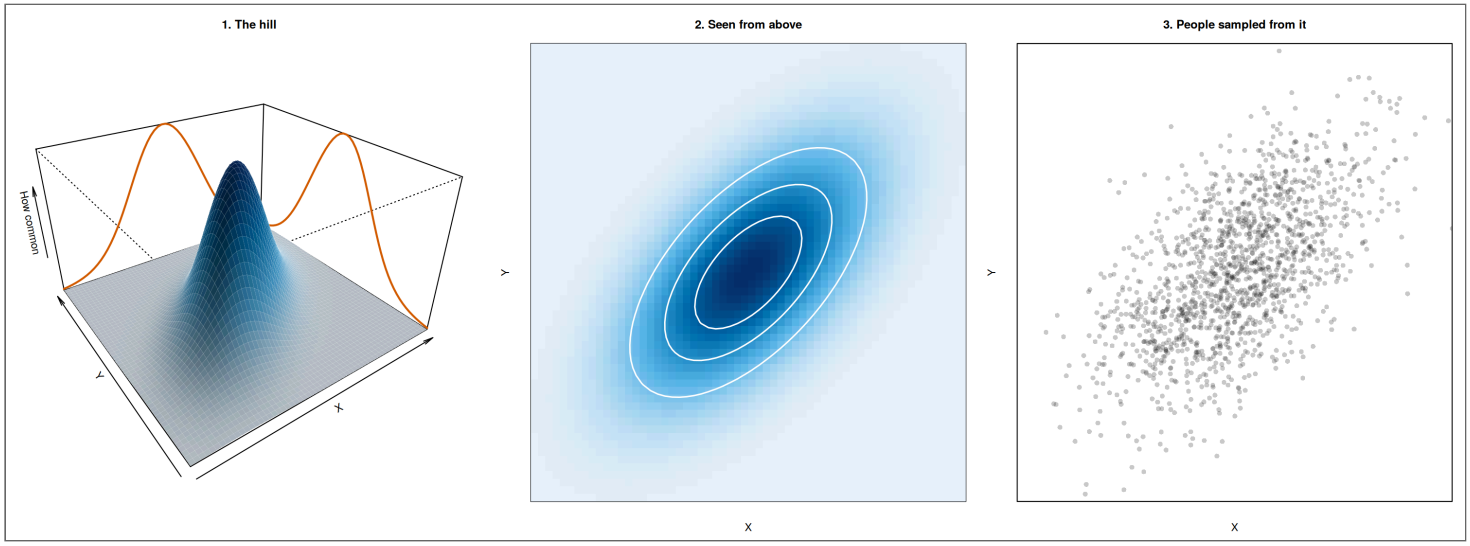


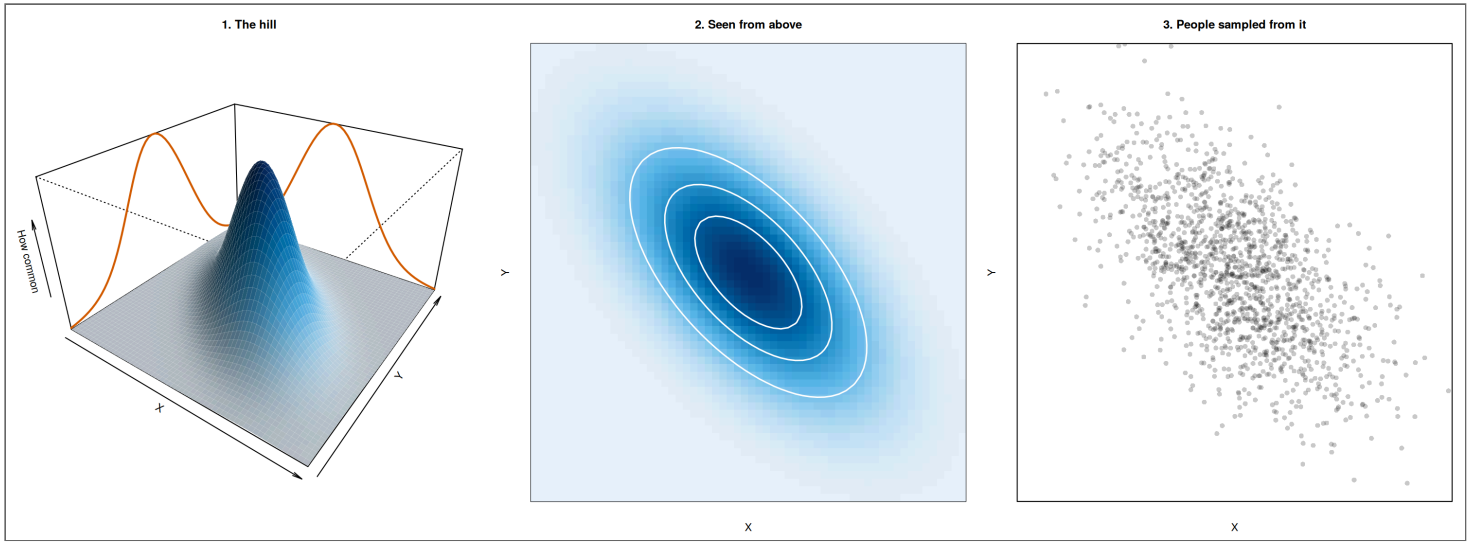
What is a bivariate normal distribution?

The **normal distribution for two variables together**: for any value of X, the Y scores are **normal**, centred on a **straight line**, with the **same spread**



The bivariate normal distribution





Pearson's *r*: Assumptions

About how the data were collected

1. Both variables are continuous (interval or ratio) 2. Observations are independent of one another

About the shape of the data

3. The pairs of scores come from a **bivariate normal distribution**

In practice, we check what this implies:

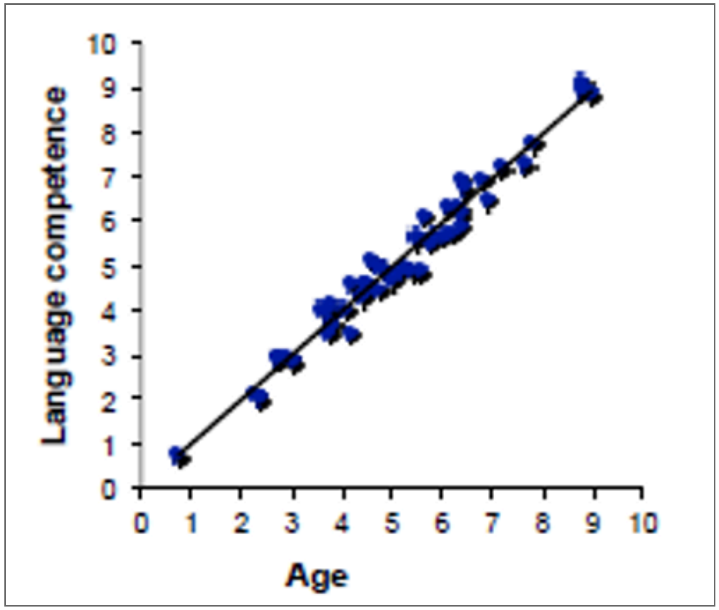
- the relationship is a straight line (**linearity**)
- the spread of Y is the same at every value of X (**homoscedasticity**)
- the Y scores are roughly normal at every value of X (**normality**)

Linearity (1)

Pearson's *r* assumes that the relationship between X and Y is linear (a straight line).

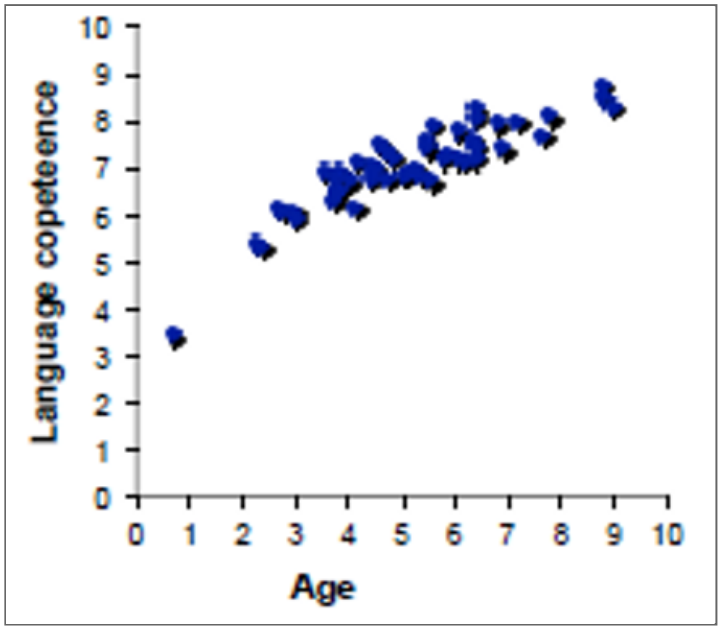
Note: This doesn't mean that all the points need to fall on the straight line, rather, the general 'trend' should be described by a straight line.

Example: Relationship between age and language competence is linear



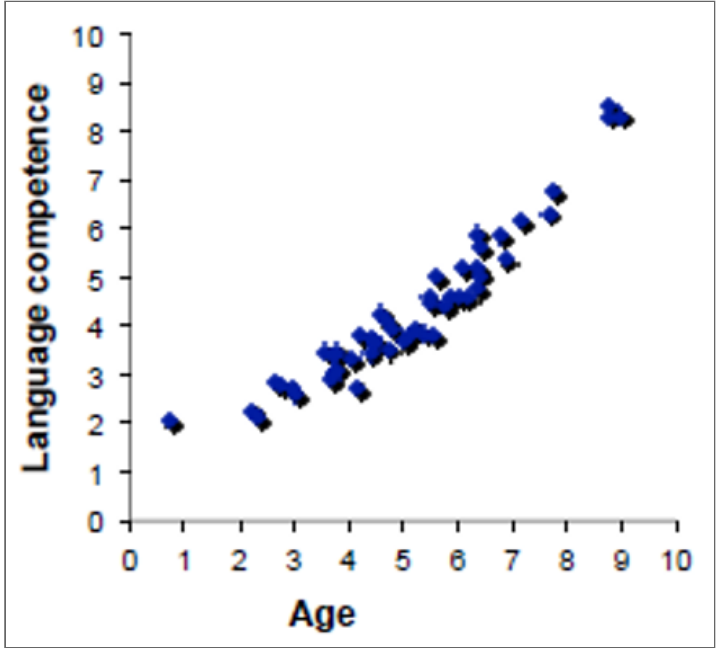
Linearity (2)

Pearson's *r* cannot provide a full picture of **curvilinear** relationships.



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

then flattens out

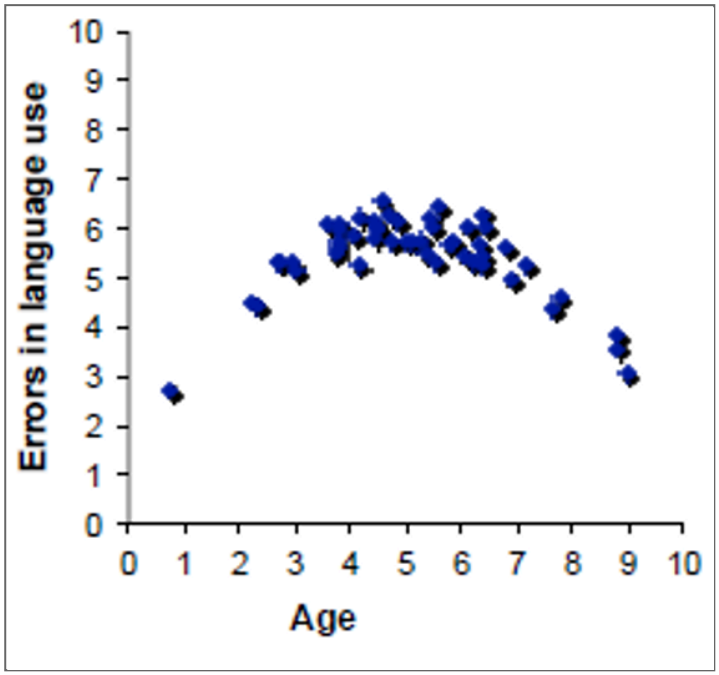


Whereas in others (e.g., compound sentences) there is a later growth spurt

Linearity (3)

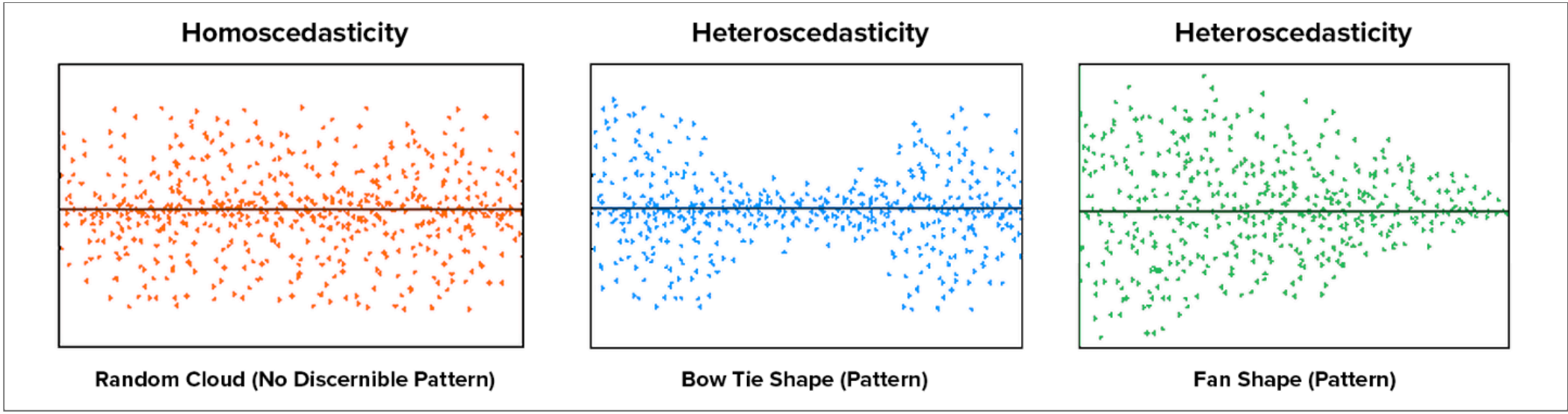
… and in others (language errors) there are more odd developmental patterns

Example: Pearson's *r* = -.18 and not significant. However, looking at the scatterplot, we can see a clear relationship between the variables



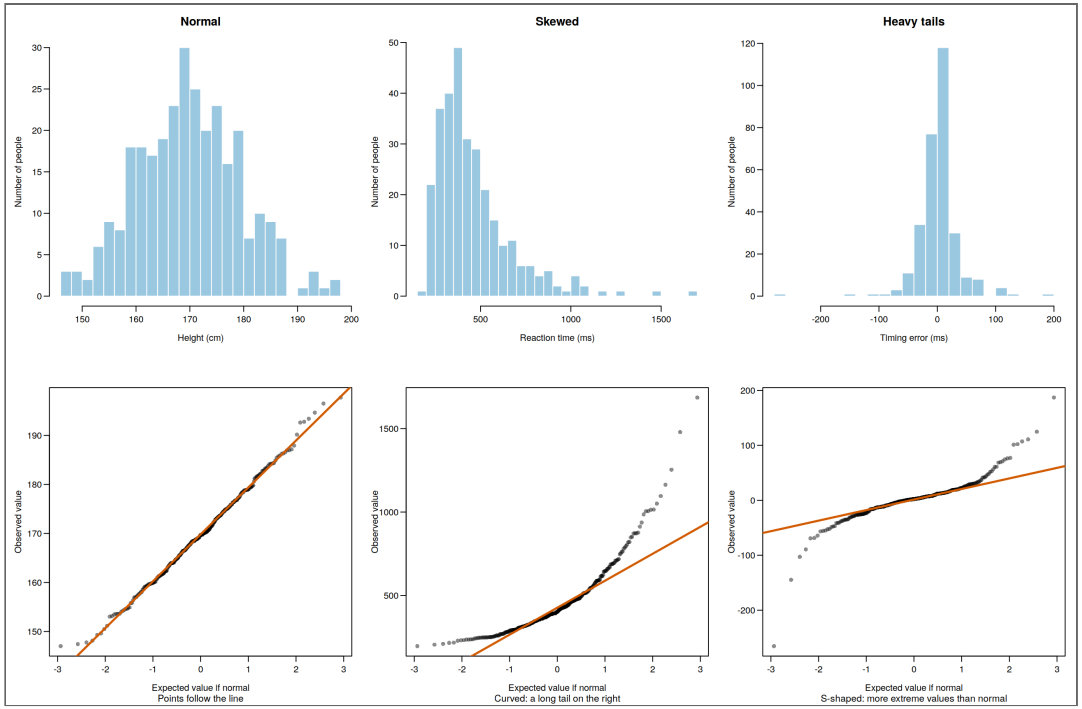
Homoscedasticity

Does the spread have homoscedasticity? Homoscedasticity means the spread of Y is similar at every value of X.

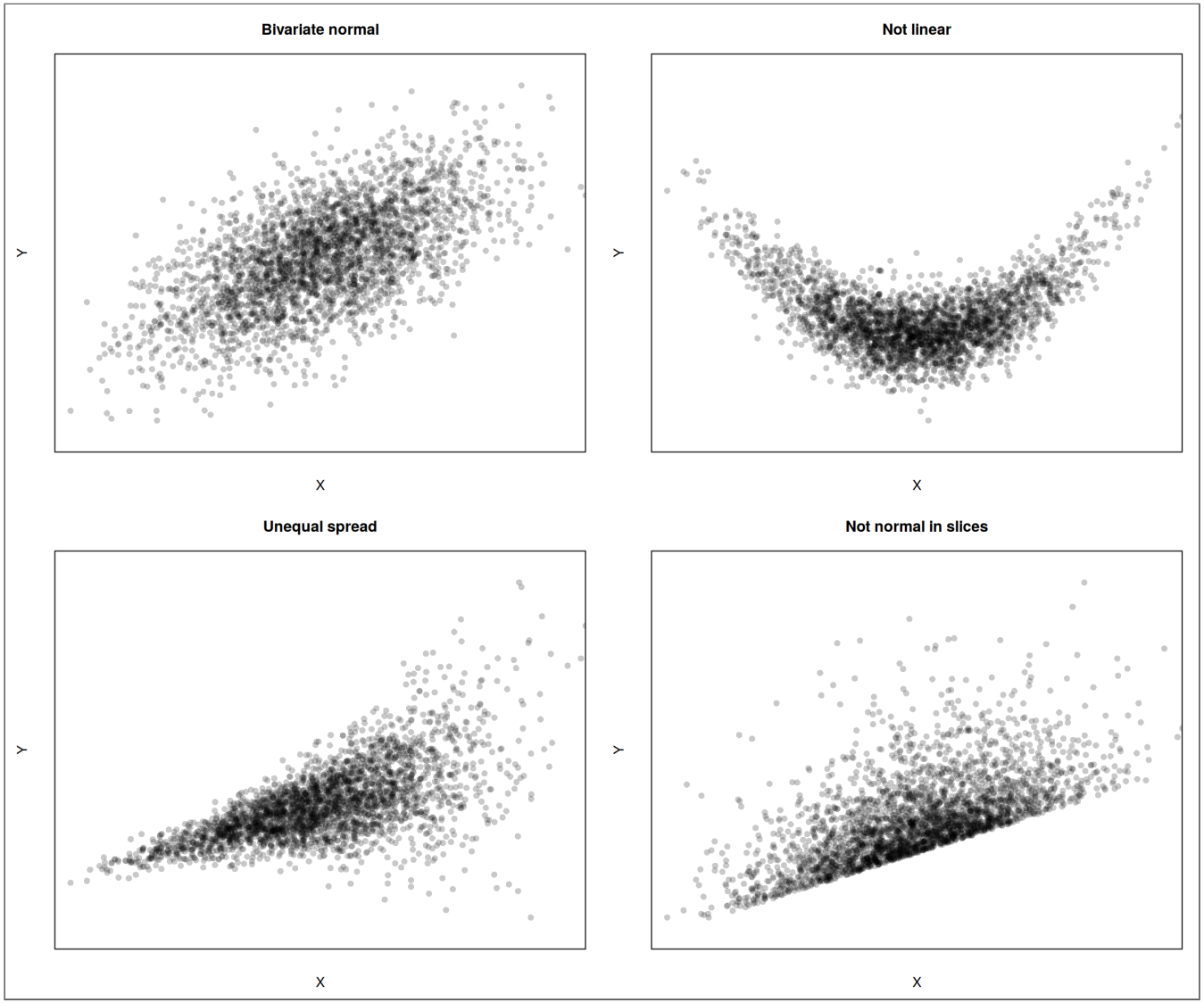


Normality

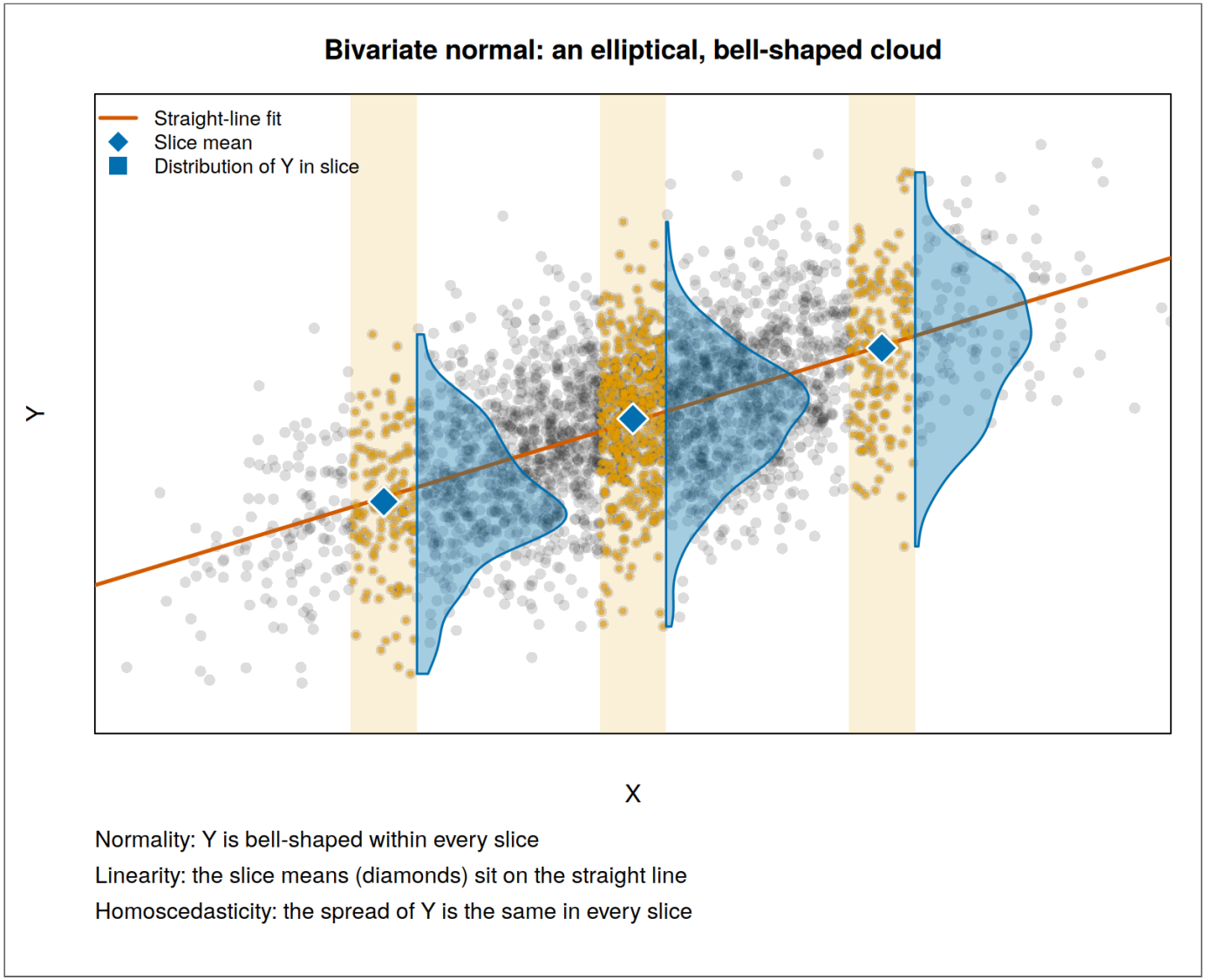
- A good approximation of whether the normality assumption is met is to check whether both variables are normally distributed using **histograms**.
- A **Q-Q plot** compares your data with what a perfectly normal distribution looks like. Eacch point is one observation. If the points follow the diagaonal line, the data are close to normal. Systematic curves away from the line show skew or heavy tails.



So what happens?



Bivariate normal



Normality: Y is bell-shaped within every slice

Linearity: the slice means (diamonds) sit on the straight line

Homoscedasticity: the spread of Y is the same in every slice

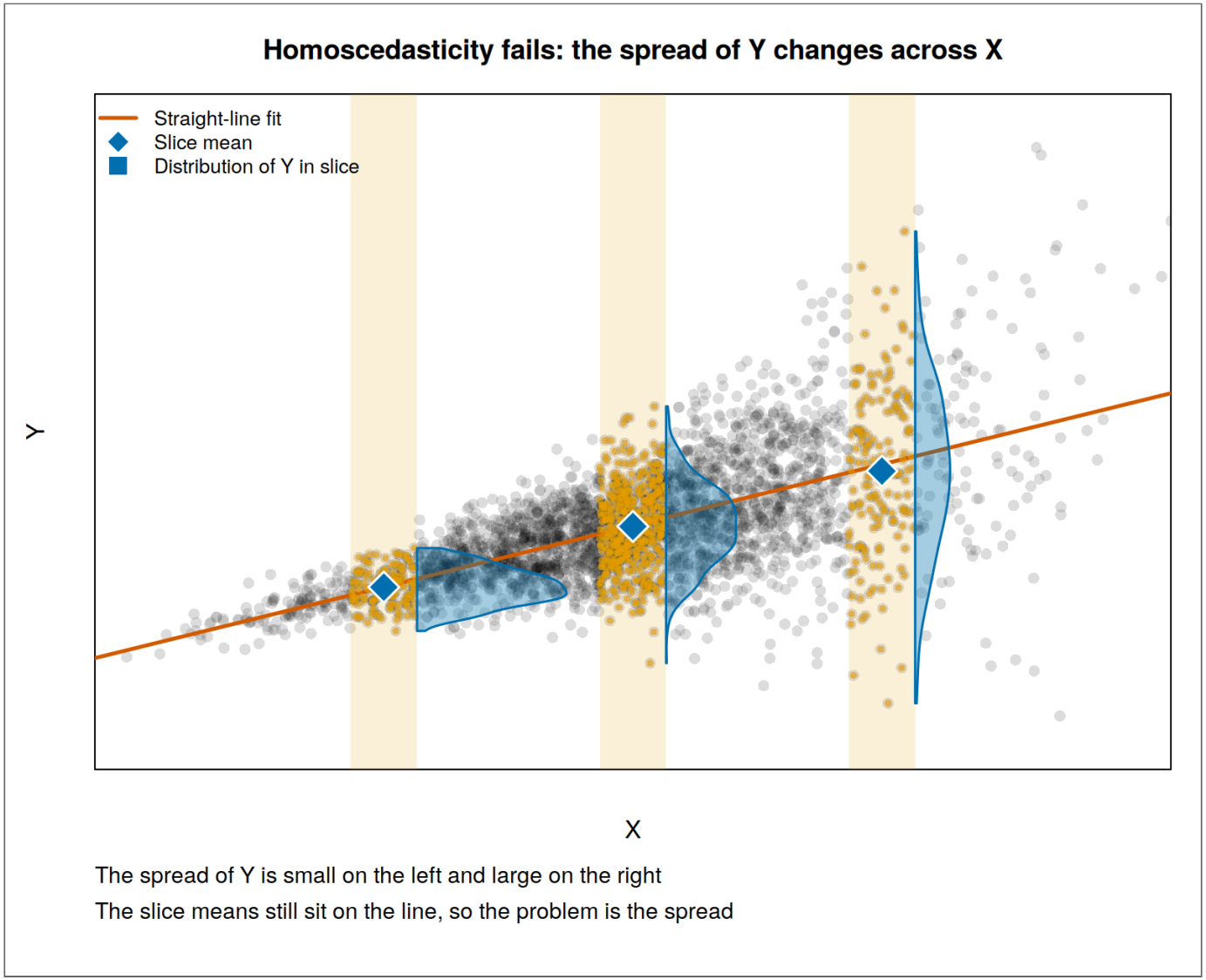
Not linear



The slice means (diamonds) leave the straight line

Spread and shape within slices look fine, but r misdescribes this relationship

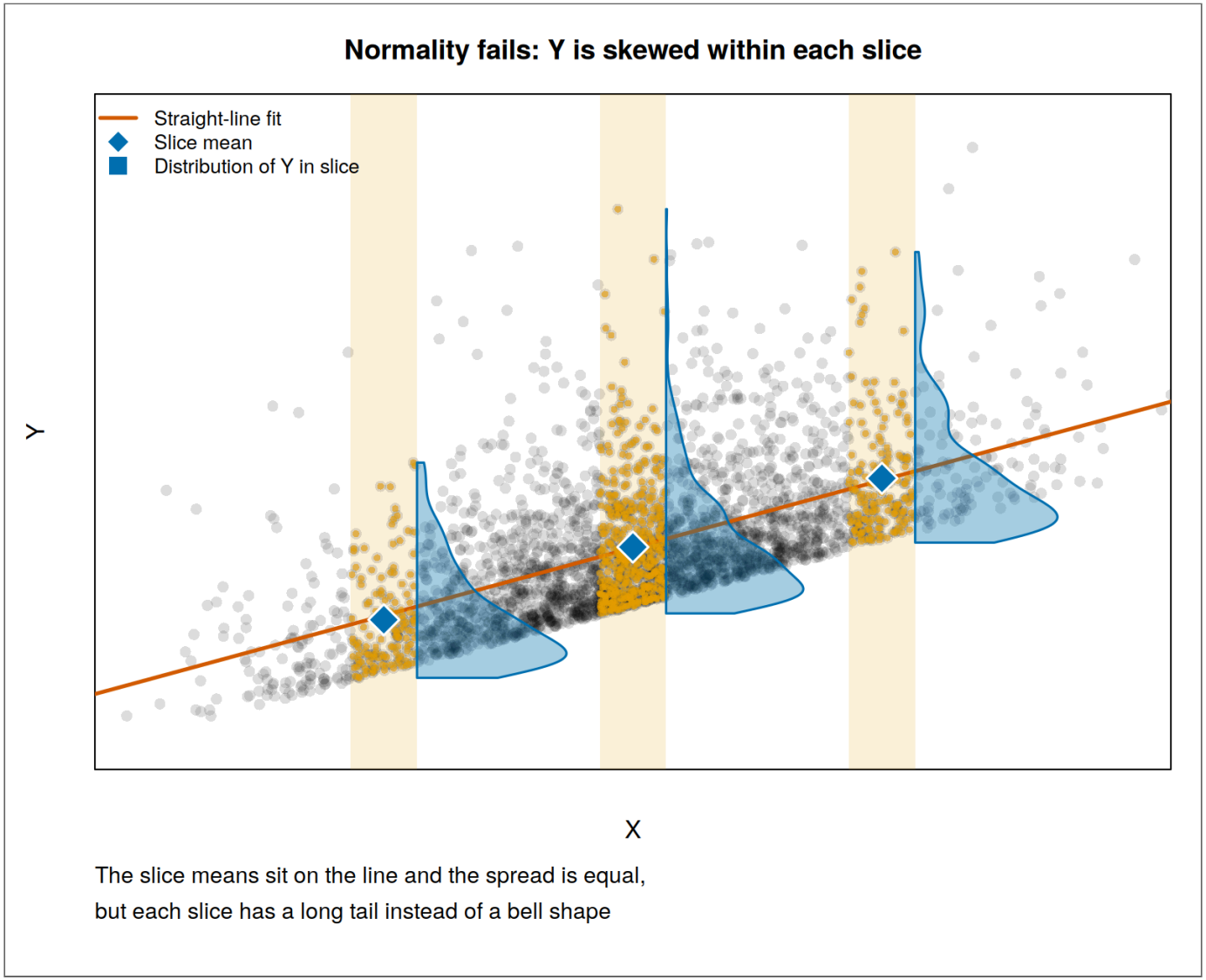
Unequal spread



The spread of Y is small on the left and large on the right

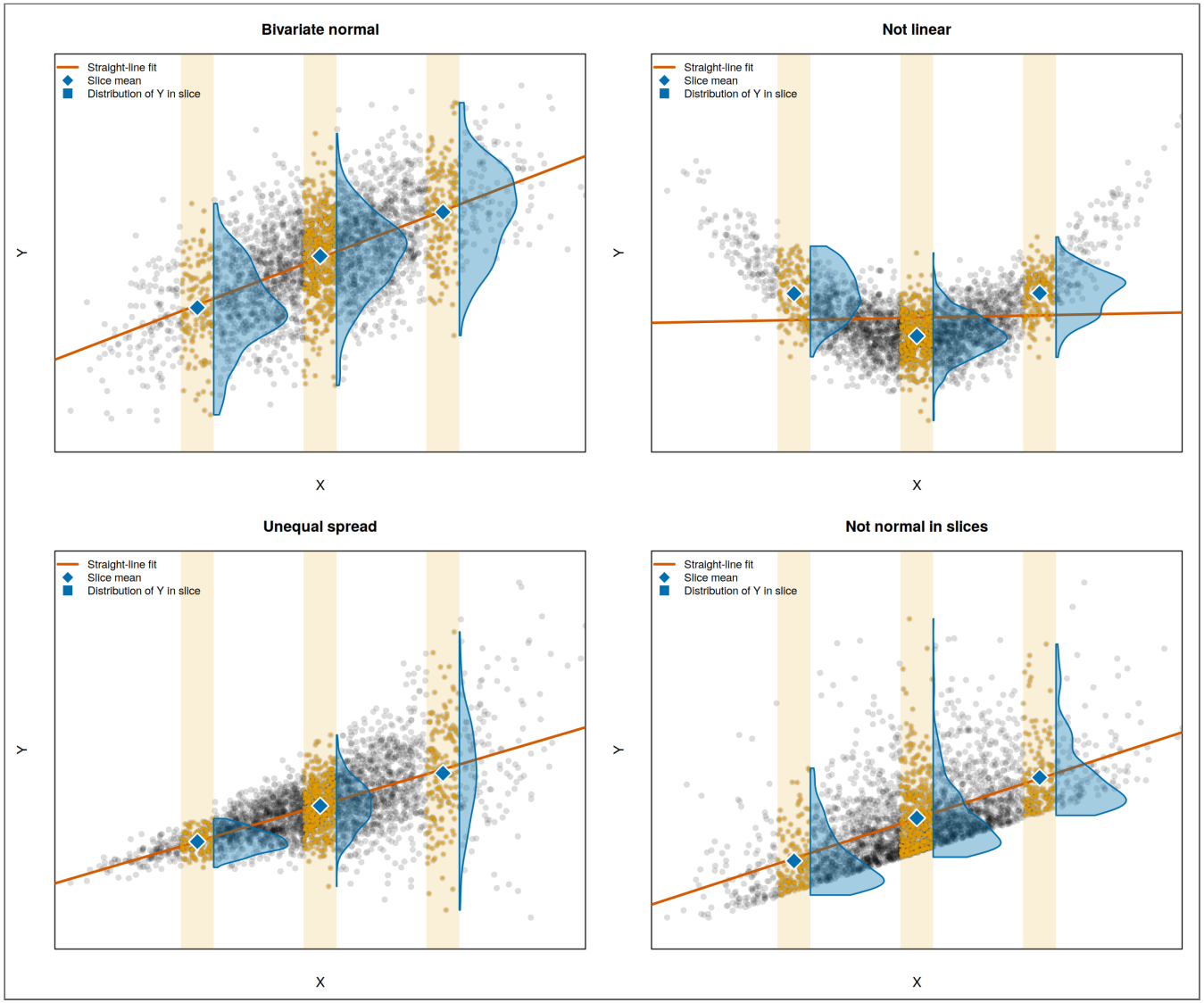
The slice means still sit on the line, so the problem is the spread

Not normal



The slice means sit on the line and the spread is equal,
but each slice has a long tail instead of a bell shape

You can tell a lot from a scatterplot!



Why is any of this a problem?

What is the p-value actually asking?

> If there were truly no relationship between X and Y in the population, how often would a sample of a particular size produce an *r* as large as yours just by chance?

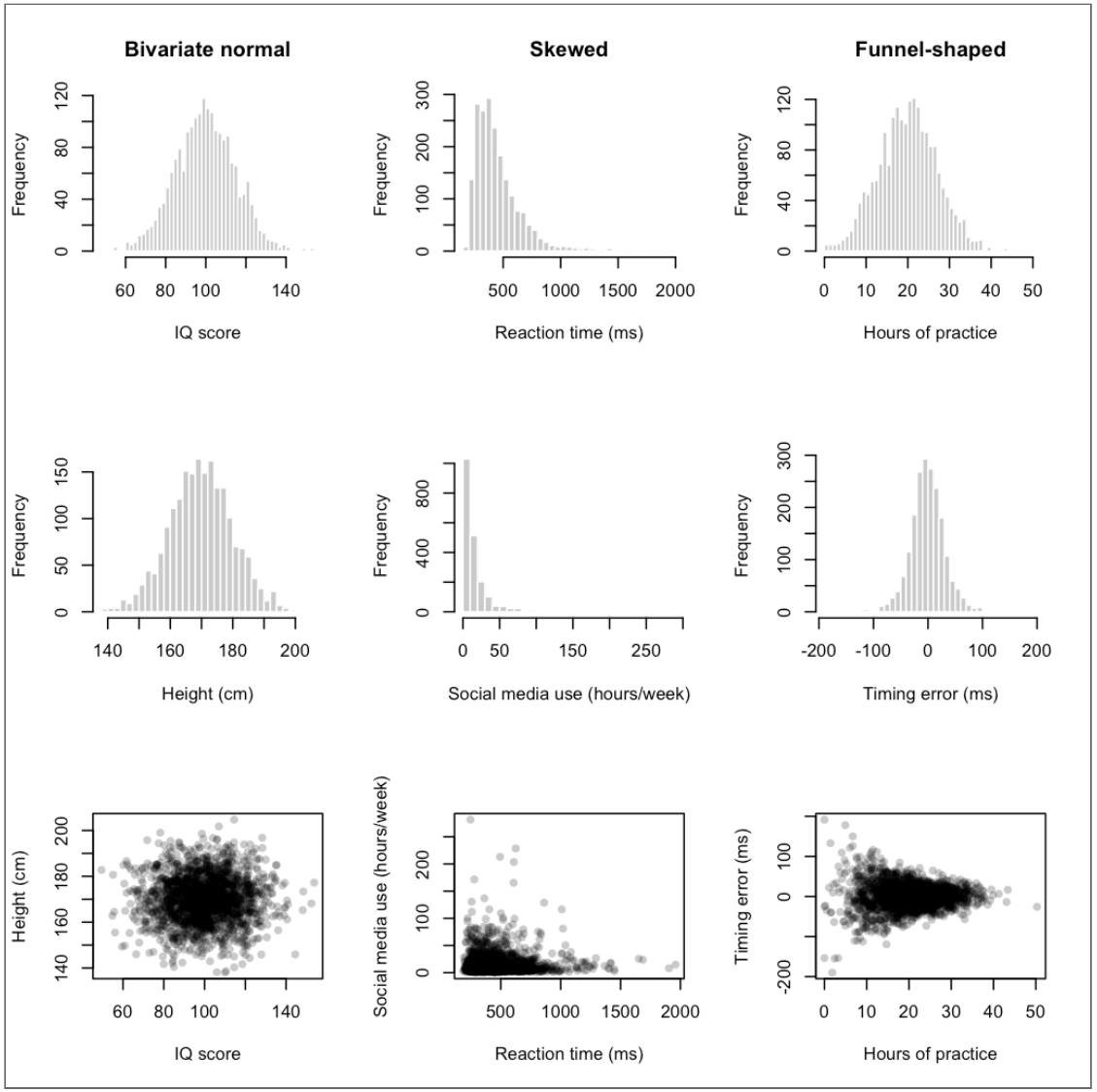
We need a model of the population

> We can't calculate chance behaviour without some description of how the data are generated. Bivariate normality is that description: the population of (X, Y) pairs forms a smooth, elliptical, bell-shaped cloud.

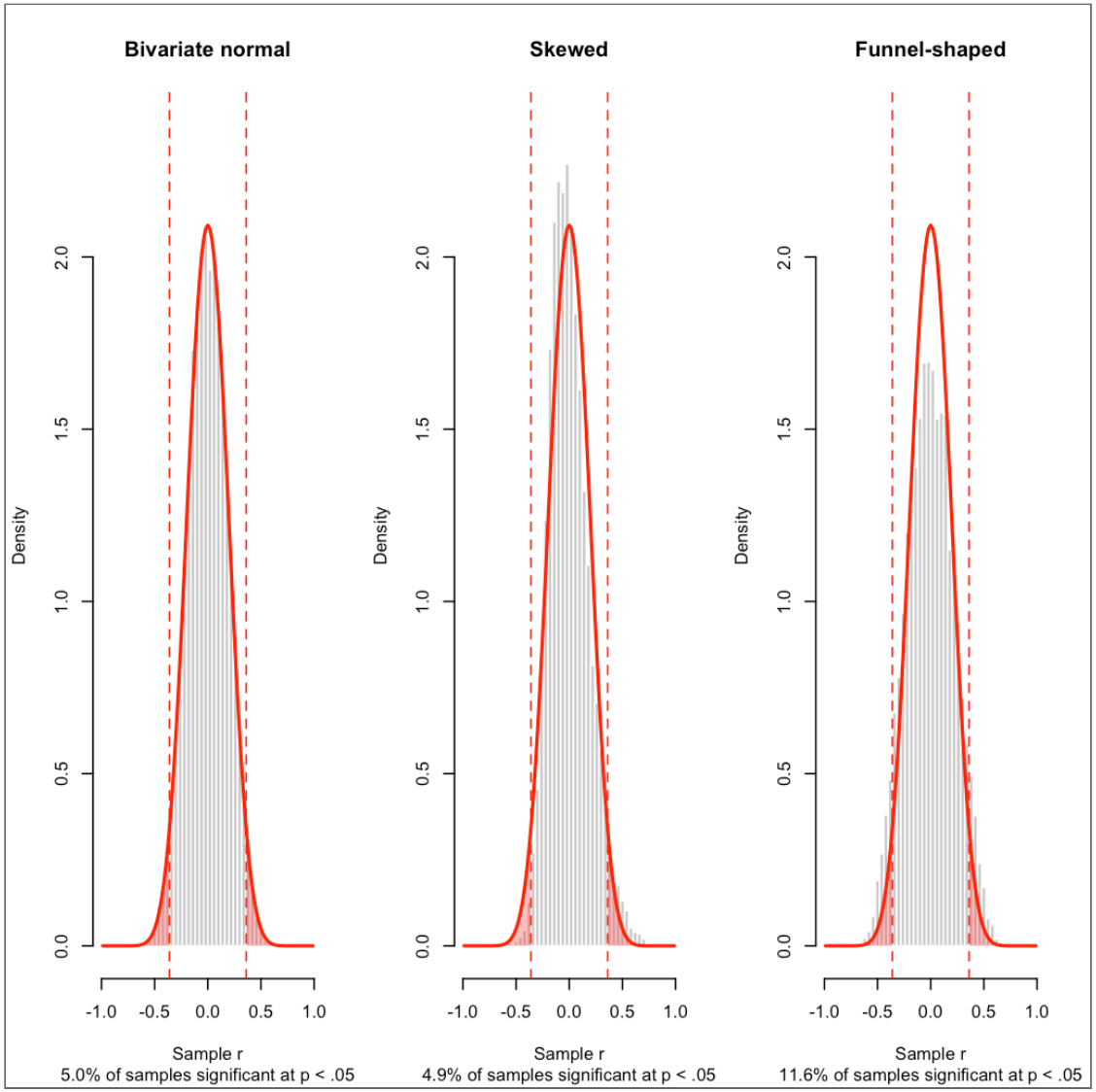
What goes wrong when the assumption fails

> If the population isn't bivariate normal, the sampling distribution of *r* can differ from the one the t-test assumes. The p-value is then calculated against the wrong 'chance' picture, so it can be too optimistic or too conservative.

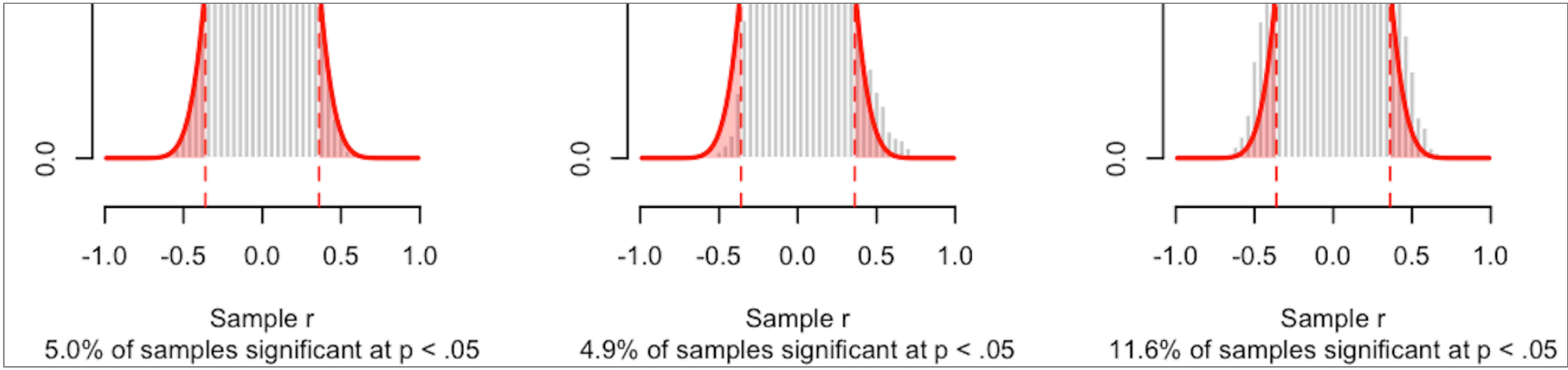
Let's look at some scenarios (1)



Let's look at some scenarios (2)



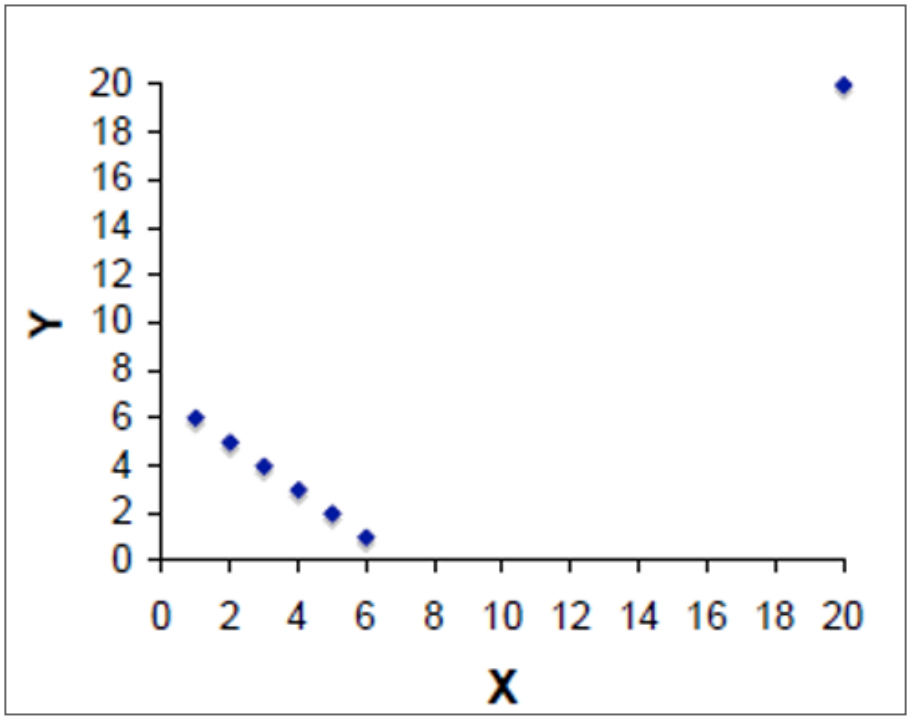
Let's look at some scenarios (3)



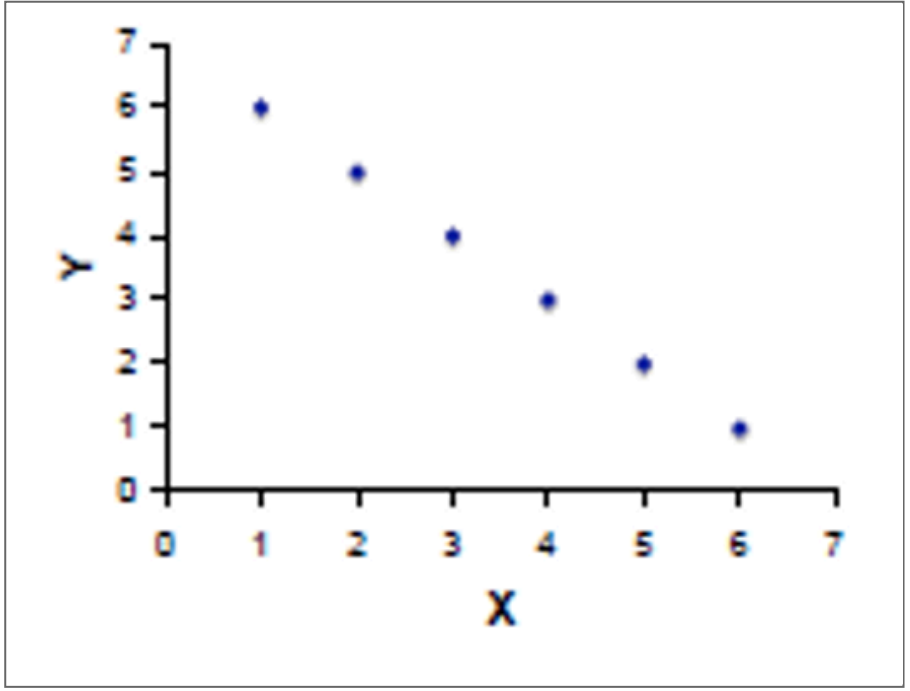
Correlation issues: Outliers (1)

Correlation may also be distorted by outliers with extreme scores (usually more than 3 SDs from the mean).

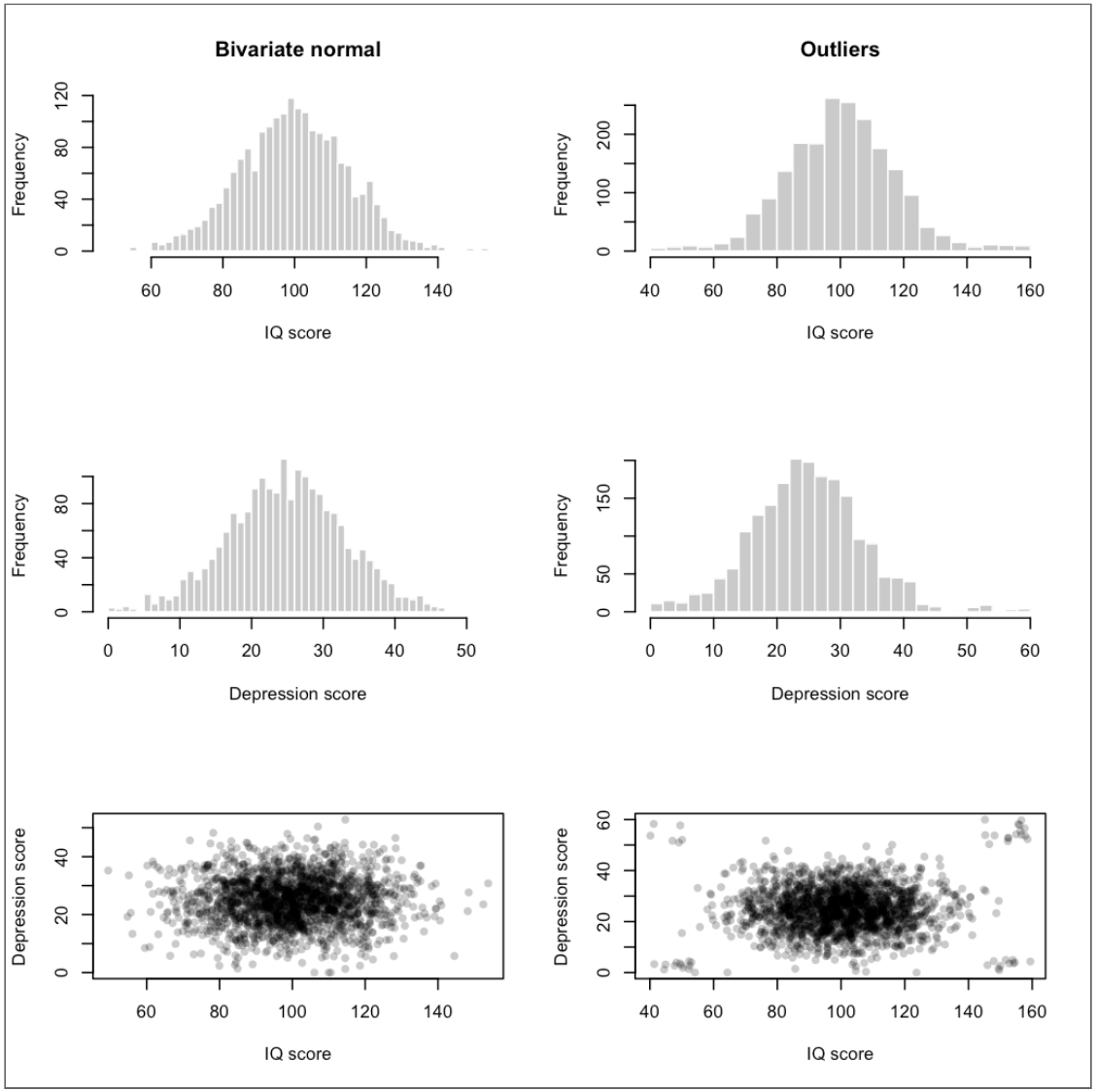
r = .86



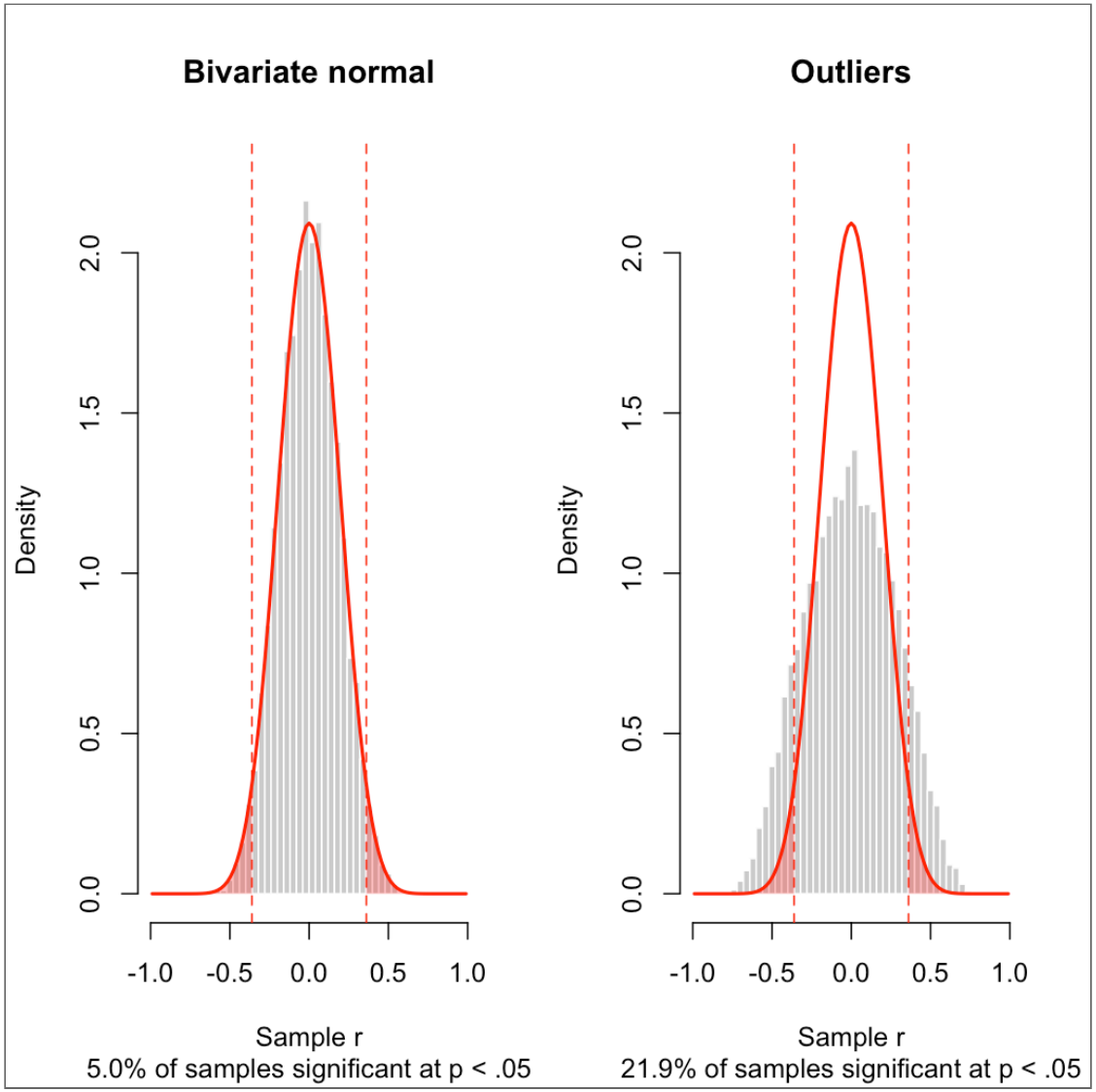
r = -1


7
6
5
4
3
2
1
0
Y
0
1
2
3
4
5
6
7
X

Correlation issues: Outliers (2)



Correlation issues: Outliers (3)



An alternative: Spearman's rho

- A correlation coefficient that works on **ranks** instead of the raw scores
- Spearman's rho (ρ) is simply **Pearson's *r* calculated on the ranks**
- It measures how **consistently** Y rises (or falls) as X rises: a **monotonic** relationship
- It runs from −1 to +1, and the sign and strength are read in the same way as *r*
- It can be used with **ordinal** data as well as interval or ratio data

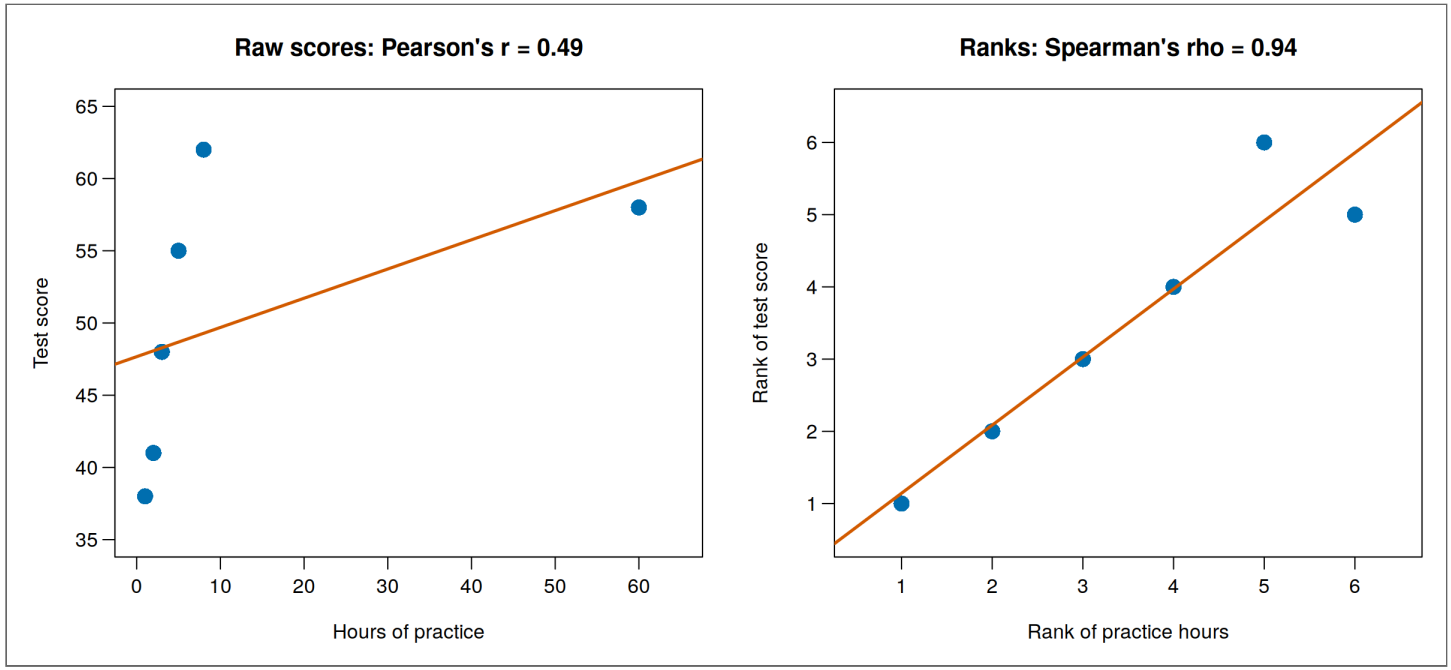
How is it calculated?

1. **Rank** the X scores from lowest (1) to highest
2. **Rank** the Y scores separately in the same way
3. Calculate **Pearson's *r*** on the two sets of ranks

Participant	Hours of practice	Test score	Rank (practice)	Rank (score)
A	2	41	2	2
B	5	55	4	4
C	3	48	3	3
D	8	62	5	6
E	1	38	1	1
F	60	58	6	5

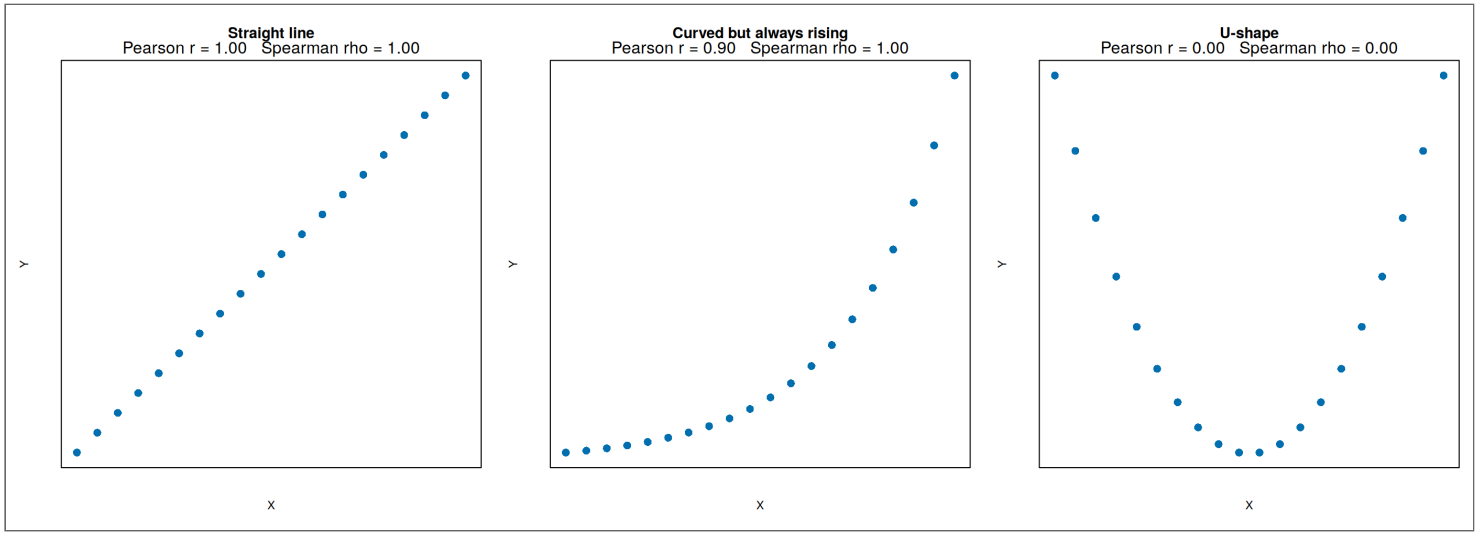
Tied scores share the average of the ranks they would have taken. Two scores tied for 3rd and 4th place both get 3.5.

Why ranks help: one extreme score



- Participant F practised far more than anyone else, which pulls the straight line towards them
- In the ranks, F is just “highest”, whether they practised for 10 hours or 60

Monotonic is not the same as linear

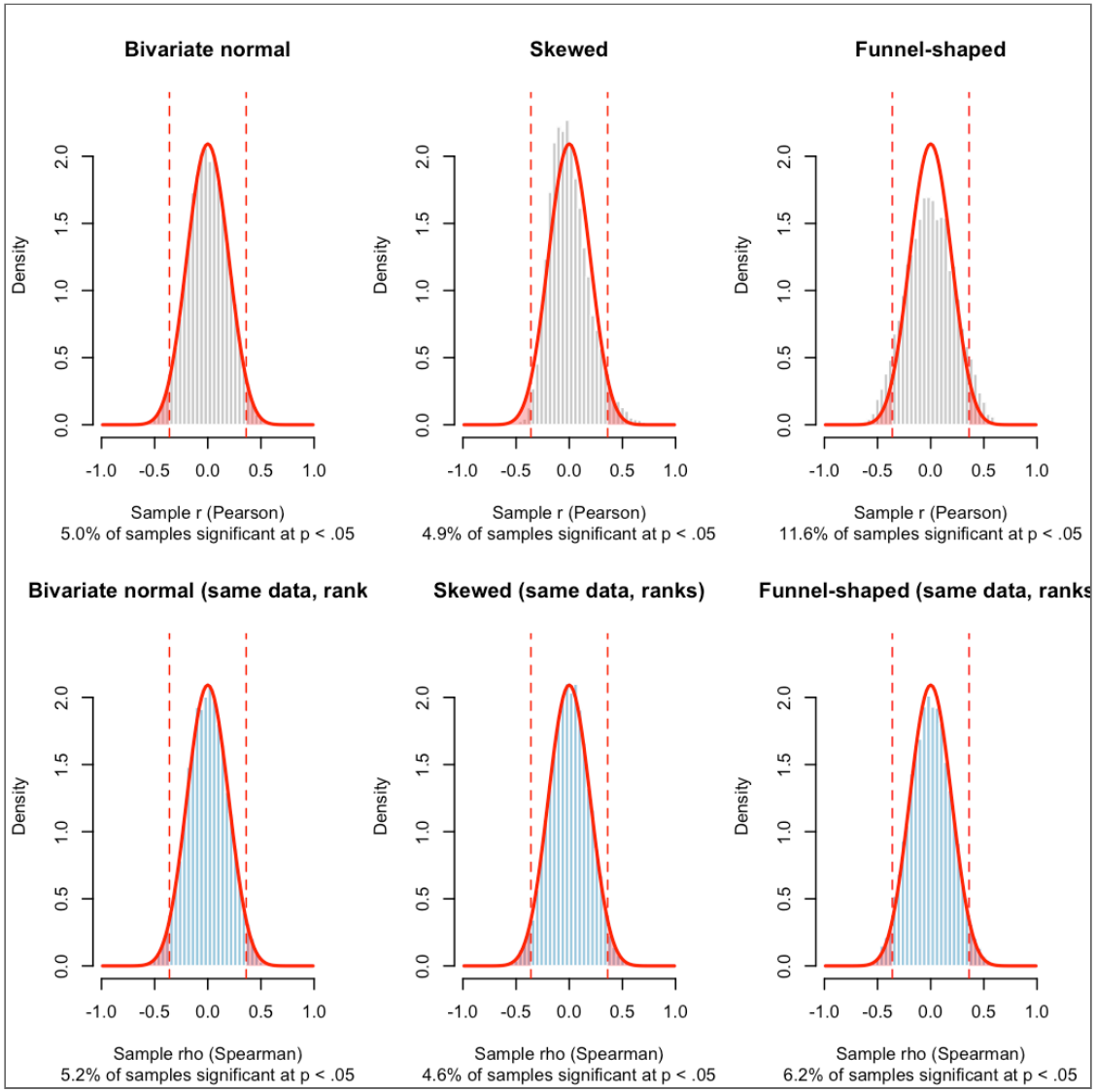


- **Straight line:** both coefficients agree
- **Curved but always rising:** rho is perfect (1.00), but *r* is lower, because the points are not on a straight line
- **U-shape:** both are zero, because the relationship does not consistently rise or fall

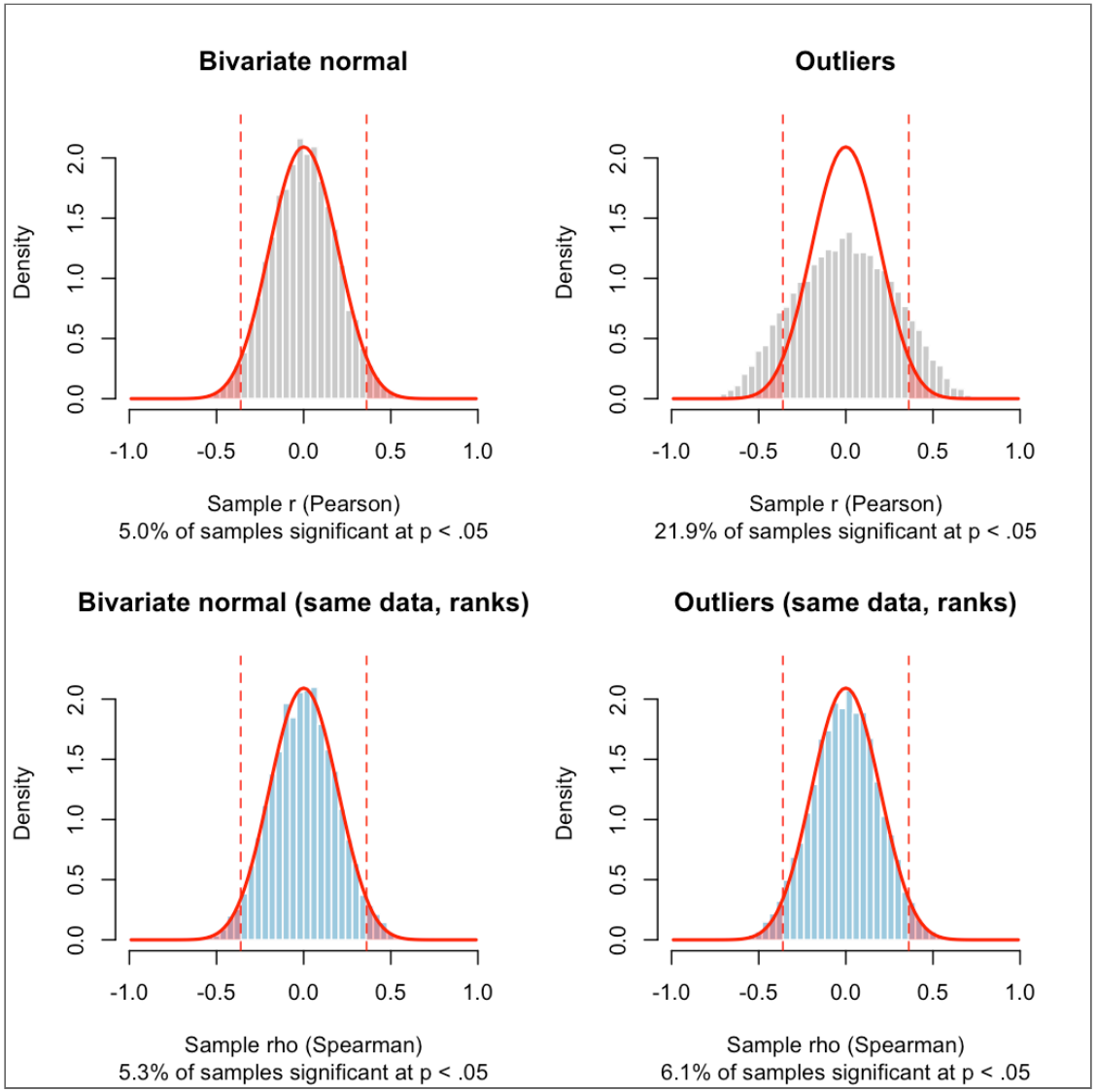
Pearson's *r* vs Spearman's rho

	Pearson's *r*	**Spearman's rho**
Calculated from	The raw scores	The ranks of the scores
Measures	A straight-line (linear) relationship	A consistently rising or falling (monotonic) relationship
Data	Interval or ratio	Ordinal, interval or ratio
Outliers	Can have a large influence	Much less influence
For the *p*-value	Independence, plus bivariate normality (linearity, equal spread, normality)	Independence and a monotonic relationship; no normality needed

Skew and funnel scenarios



Outlier scenario



When should you use Spearman's rho?

- Your data are **ordinal** (e.g. a single rating item, or rankings)
- The scatterplot shows **outliers** that you can't justify removing
- The relationship is **curved but consistently rising or falling**

But it isn't a cure-all

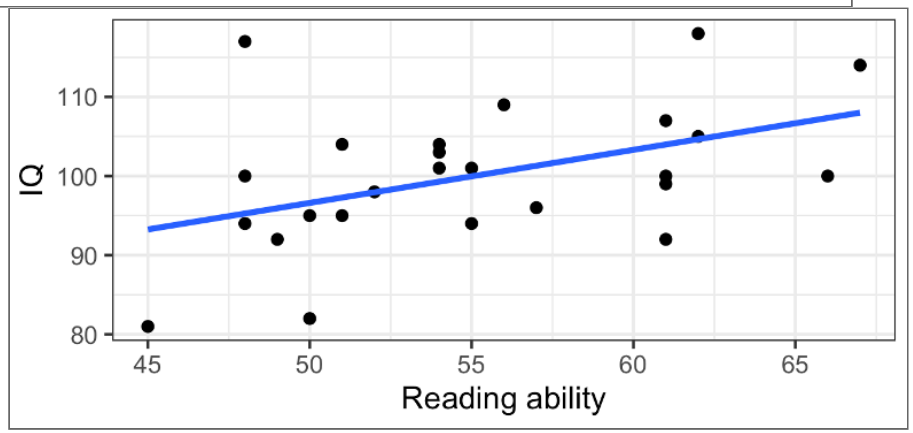
- It still needs **independent** observations
- It can't describe a **U-shaped** relationship
- It only **partly** helps when the spread of Y changes across X

Choose from the scatterplot and your research question, not by trying both and keeping the smaller p-value. Or report both.

Correlation issues: Missing data

A correlation can only be computed for participants who contribute a data point on both variables.

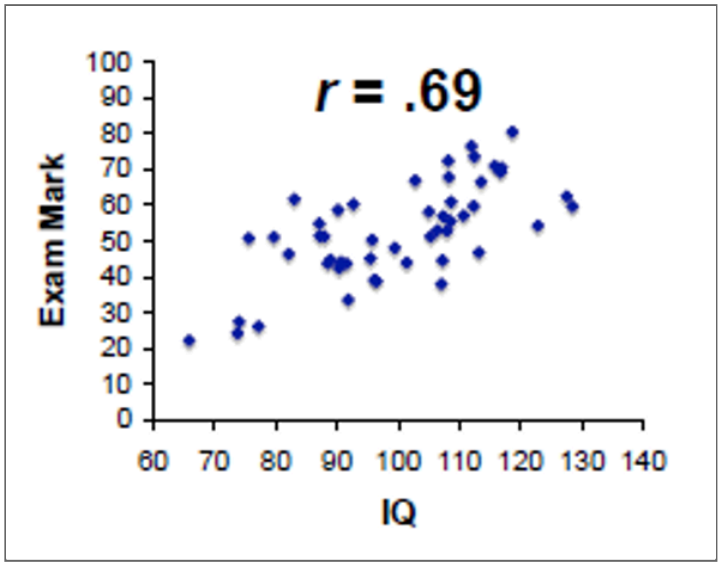
Case	Abil	IQ	Home	TV
1	61	107	144	487
2	56	109	123	608
3	45	81	108	640
4	66	100	155	493
5	49	92	103	636
6	62	105	161	407
7	61	92	138	463
8	55	101	119	717
9	62	118	155	643
10	61	99	121	674
11	51	104	93	675
12	48	100	127	595
13	50	95	97	673
14	50	82	140	523
15	67	114	151	665
16	51	95	112	663
17	55	94	102	684
18	54	103	142	505
19	57	96	127	541
20	54	104	102	678
21	52	98	124	564
22	48	117	87	787
23	61	100	141	582
24	54	101	117	647
25	48	94	111	448


IQ
110
100
90
80
45
50
55
60
65
Reading ability

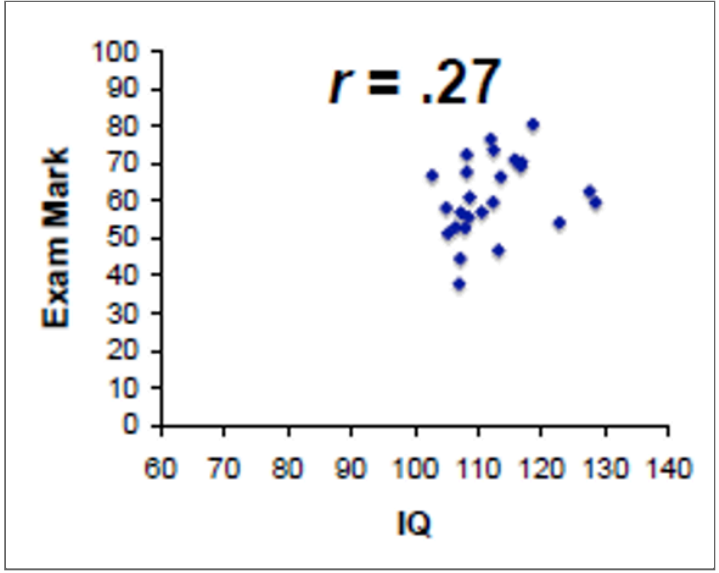
Correlation issues: Range restrictions (1)

The sample used may not represent the true variation 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 …

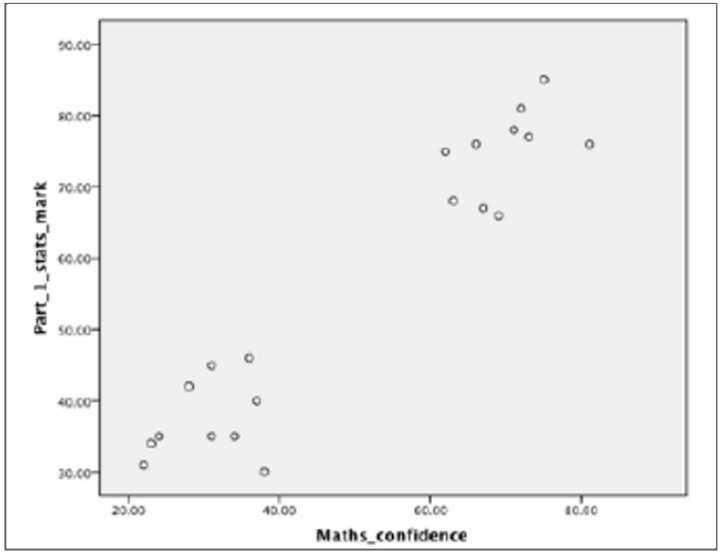


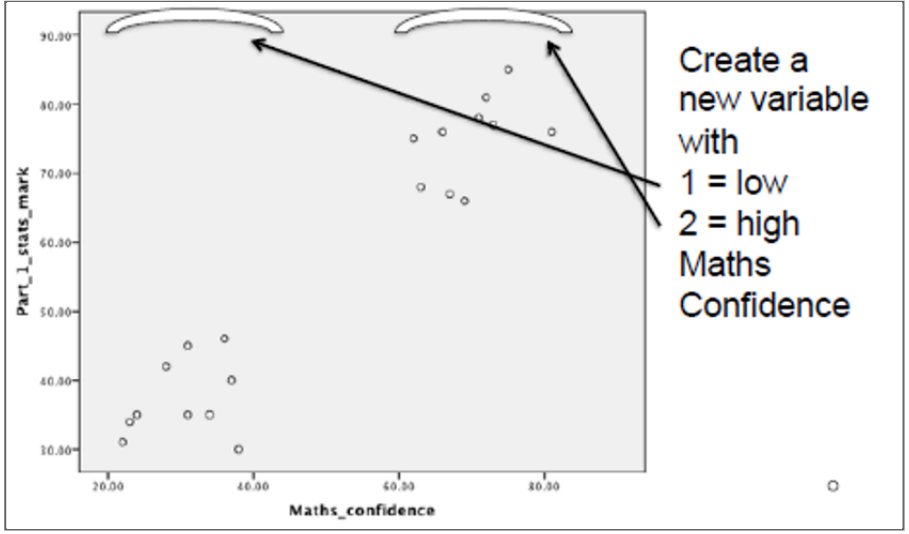
…but the correlation is likely to be much smaller if only university students are considered.


r = .27
Exam Mark
100
90
80
70
60
50
40
30
20
10
0
60
70
80
90
100
110
120
130
140
IQ

Correlation issues: Range restrictions (2)

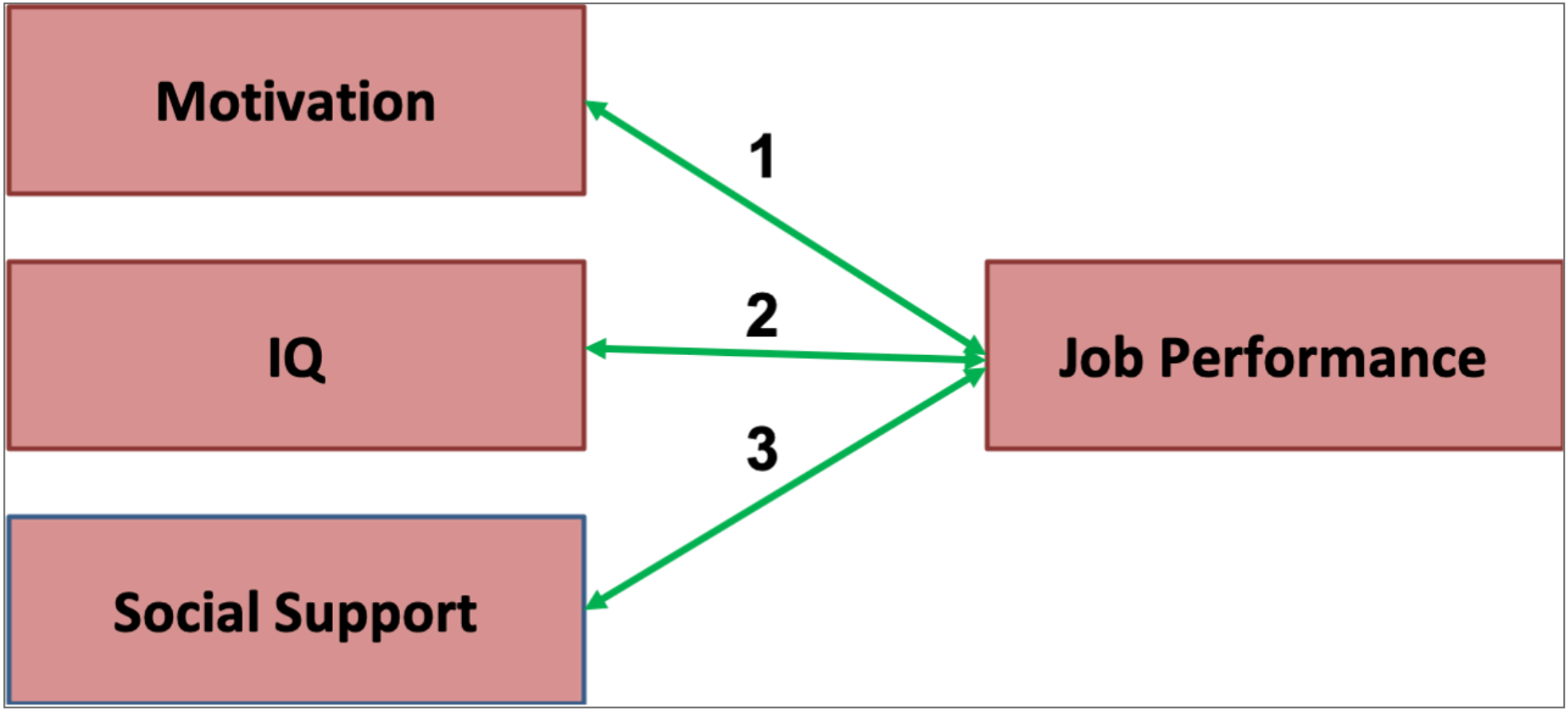
If the range on one variable is unusually large (and there are two very distinct clusters), it is sometimes more beneficial to create a new variable



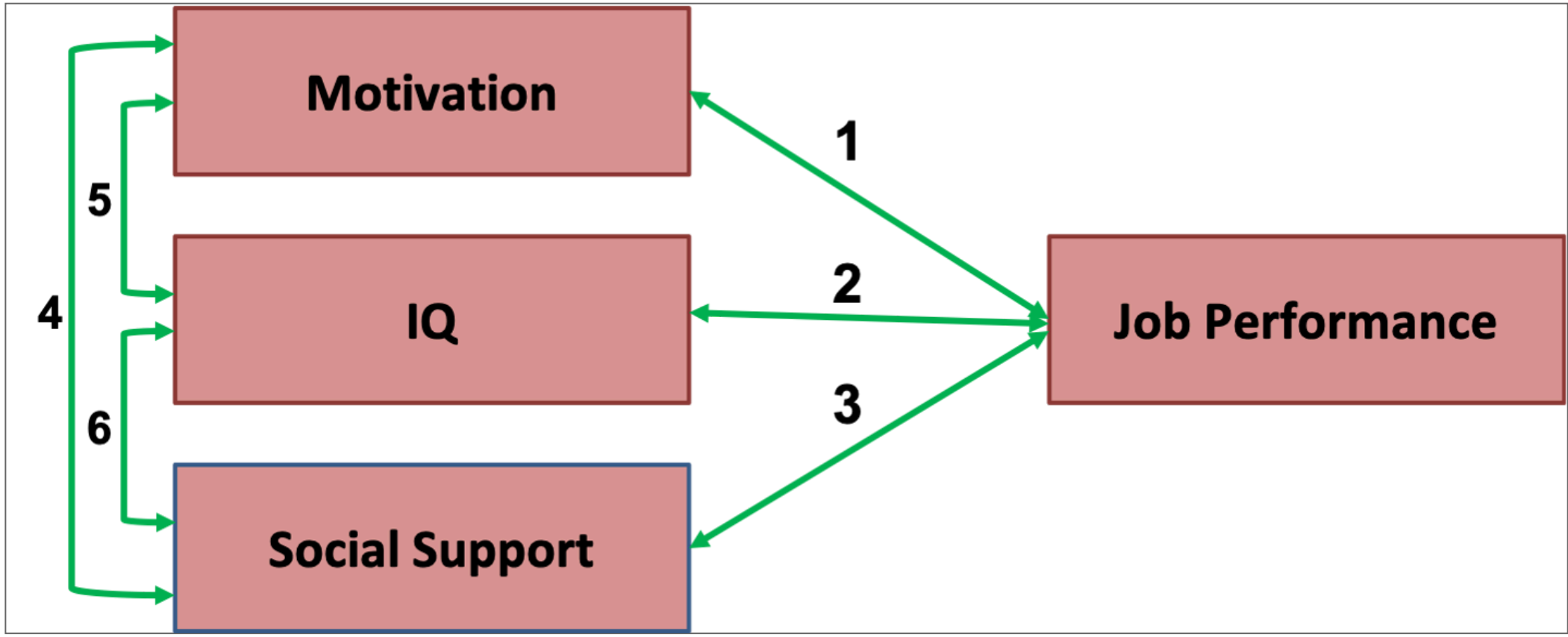

Part_1_stats_mark
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
40.00
60.00
80.00
Maths_confidence
Create a new variable with
1 = low
2 = high
Maths Confidence

Intercorrelation (1)

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



Intercorrelation (2)



Constructing a 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)

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

Many correlations, more false positives

- A correlation matrix tests **every pair** of variables: with k variables, that is $k(k - 1) / 2$ tests
- Each test has a **5% chance** of a false positive when there is no real relationship
- Across many tests, those chances **add up**

Variables	Correlations tested	Expected false positives	Chance of at least one
3	3	0.15	14%
5	10	0.5	40%
10	45	2.25	90%
20	190	9.5	over 99.9%

Intercorrelation: What can we do?

- **Decide in advance** which correlations test your hypotheses, and test only those
- **Adjust the threshold** for the number of tests, for example Bonferroni: .05 ÷ number of tests (45 tests gives about .001)

  Example: Each varaible is correlated with 3 other variables: .05 / 3 = .016. Therefore, for a correlation coefficient to be 'significant' $p<.016$.

- **Report all the tests you ran**, not just the significant ones
- Treat unexpected significant correlations in a large matrix with caution.

Summary

Assumptions

About how the data were collected Variables need to be continuous and observations need to be independent

About the shape of the data Bivariate normal distribution (linearity, homoscedasticity, normality)

Issues Beware of missing data and check the scatterplot for influence of outliers and possible range restrictions.

Spearman's *rho* as an alternative to Pearson's *r*

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