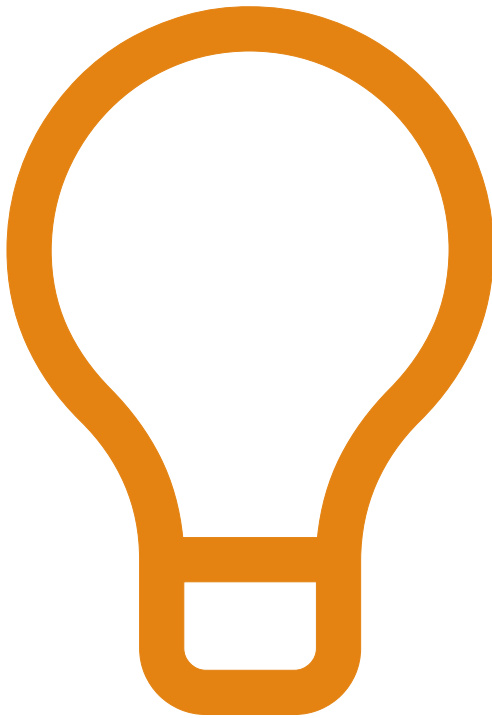


MATH9102

CORRELATION

Sources used in creation of this lecture:

Statistics and Data Analysis, Peck, Olsen and Devore; Discovering Statistics Using R, Field, Miles and Field; Understanding Basic Statistics, Brase and Brase; SPSS Survival Manual, Julie Pallant



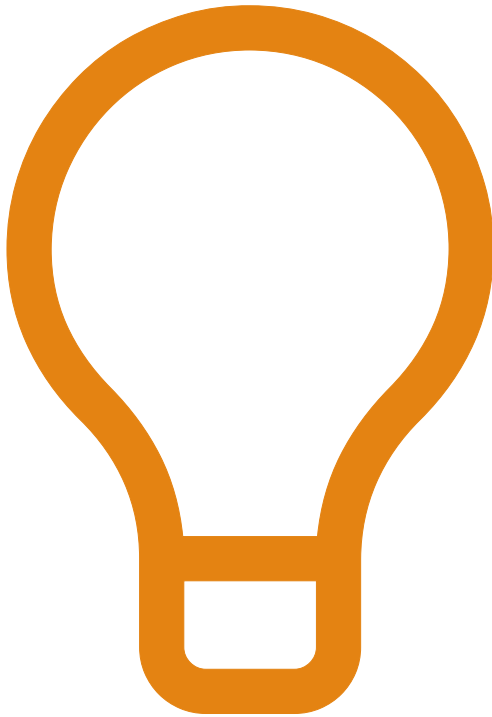
Correlation - what are we interested in?

We are interested in two concepts in the real world.

We hypothesize that they vary in common (change at a constant rate)

We are interested in calculating a statistical measure that expresses the extent to which two variables are **linearly related**.

Linear: A change in one variable is reflected by a consistent change in the other.



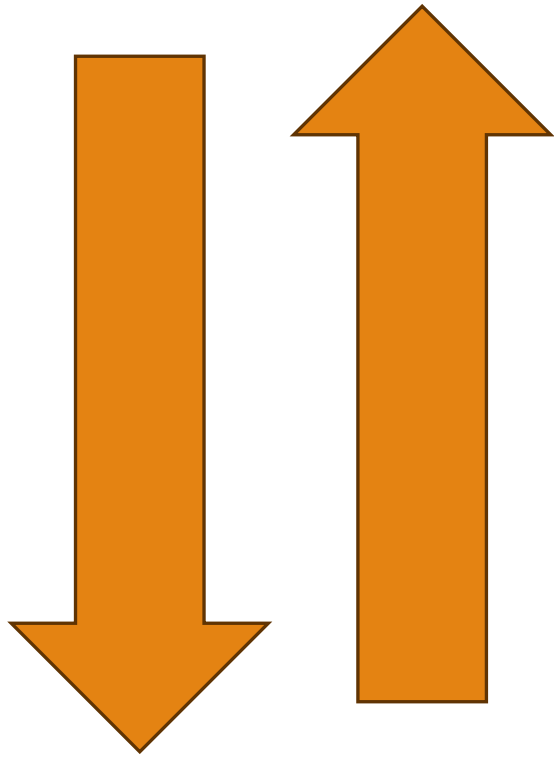
Correlation - what are we interested in?

We are interested in

- the **strength** of the relationship

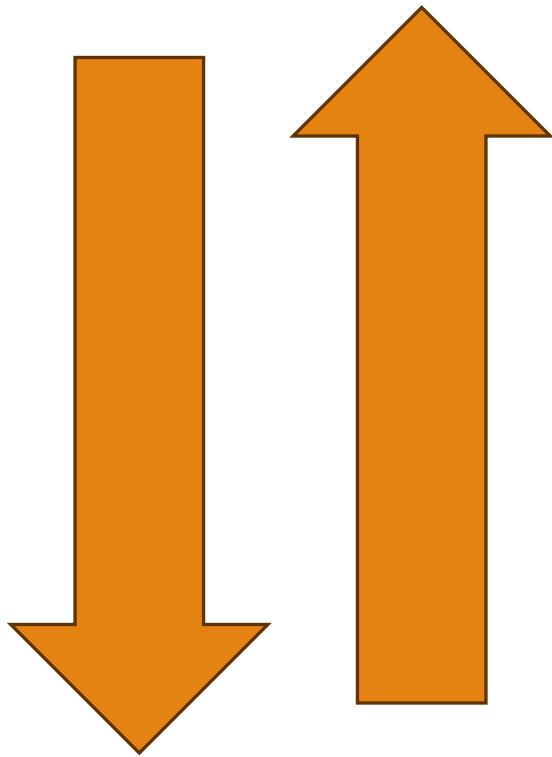
AND

- the **direction**



Covariance

- Covariance is a measure of how two variables change together.
- It tells us whether two variables tend to increase or decrease together, or if one increases while the other decreases.

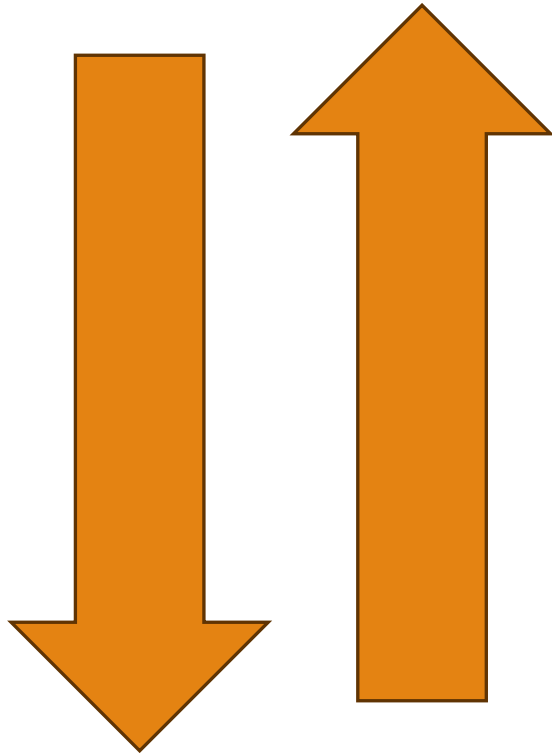


Covariance

Variance tells us by how much scores deviate from the mean for a single variable.

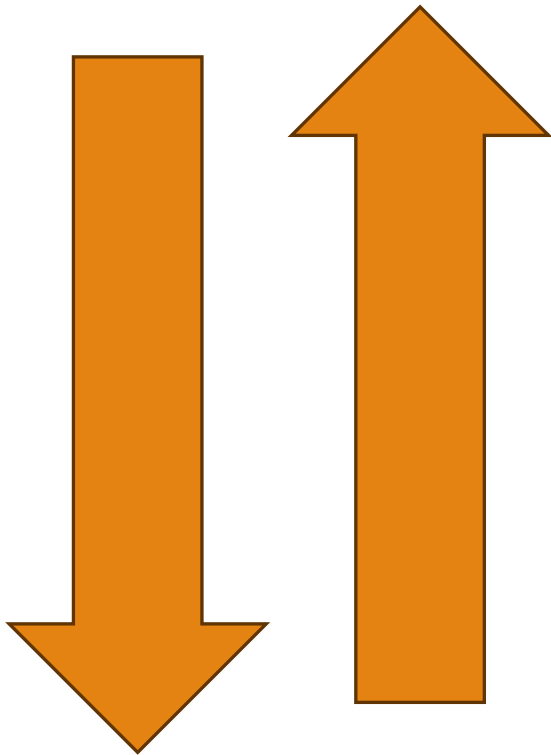
Covariance = Scaled version of variance

- Calculate the error between the mean and each observations score for the first variable (x).
- Calculate the error between the mean and their score for the second variable (y).
- Multiply these error values.
- Add these values and you get the cross-product deviations.
- The covariance is the average cross-product deviations.



Covariance

- **Positive covariance:**
 - If two variables have a positive covariance, it means that as one variable increases, the other tends to increase as well (they move in the same direction).
- **Negative covariance:**
 - If two variables have a negative covariance, it means that as one variable increases, the other tends to decrease (they move in opposite directions).

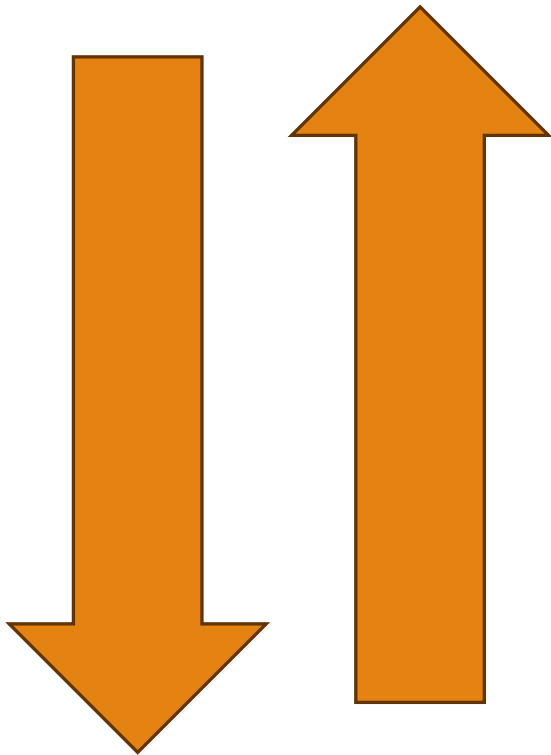


Covariance

However, covariance doesn't provide a standardized measure

It only gives a sense of the direction of the relationship (positive or negative)

And the magnitude of their joint variability, but the magnitude can be hard to interpret since it's not scaled.



Correlation

Correlation is a **standardized** version of covariance, meaning it's scaled to always fall between -1 and 1.

Correlation coefficient (r) ranges from:

- **+1: Perfect positive** relationship (as one variable increases, the other increases in a perfectly linear fashion).
- **-1: Perfect negative** relationship (as one variable increases, the other decreases in a perfectly linear fashion).
- **0: No relationship** (the variables are not related).

Correlation:

What are we interested in?

Modelling this in the linear model

- We are using a line as the model

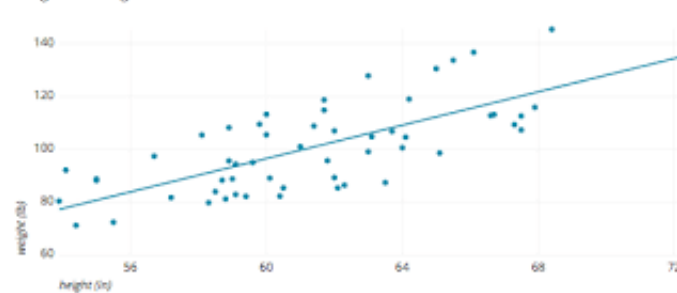
Direction of the relationship

- Positive or negative
- Slope of the line

Strength of the relationship

- How close are the data points to the line
 - Very close = strong
 - Very dispersed = weak

Weight and Height of Children



Correlation: Visualize using a Scatterplot

When to Use:

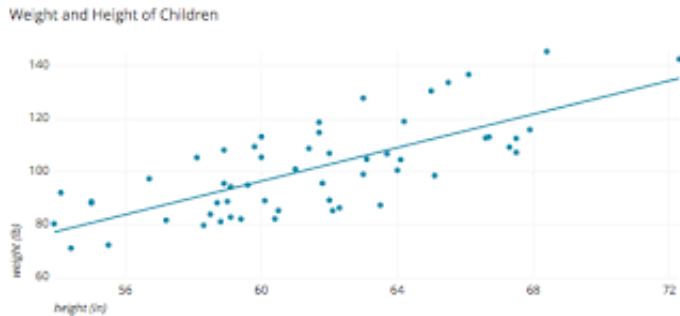
- Bivariate numerical data (two variables)
- Plot the relationship between two variables
 - One independent, one dependent
- Collection of ordered pairs

How to construct:

- Draw a horizontal scale and mark it with appropriate values of the independent variable
- Draw a vertical scale and mark it appropriate values of the dependent variable
- Plot each point corresponding to the observations

To describe

- Comment the relationship between the variables



Correlation:

What are we interested in?

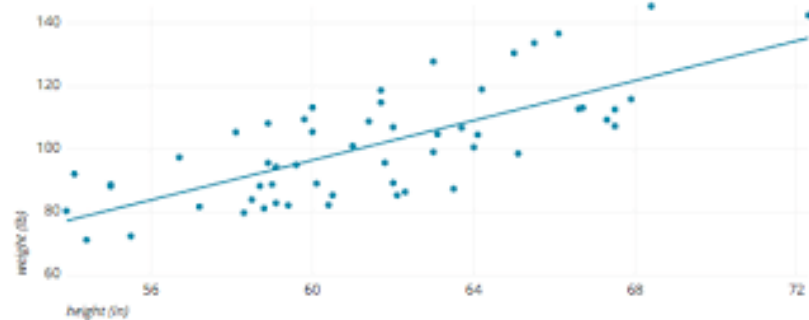
Direction

- Positive or negative
- Slope of the line

Strength

- How close are the data points to the line
 - Very close = strong
 - Very dispersed = weak

Weight and Height of Children

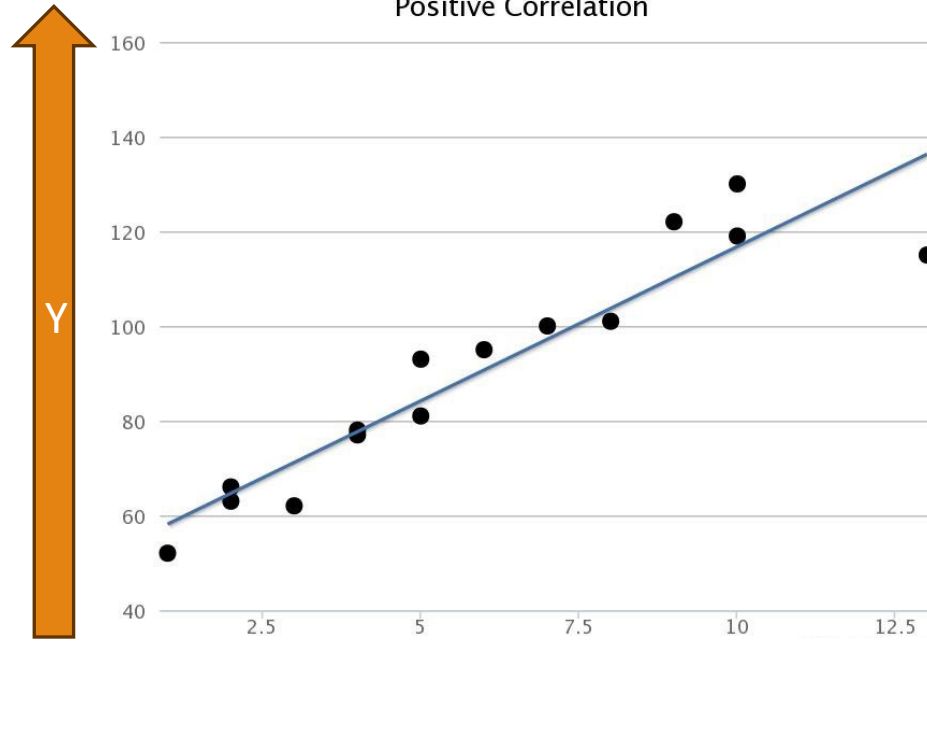


Positive Correlation

If the x-coordinates and the y-coordinates both increase, then it is **POSITIVE CORRELATION**.

This means that both are going up, and they are related.

Positive Correlation



Positive Correlation

If you look at the age of a child and the child's height, you will find that as the child gets older, the child gets taller.

Because both are going up, it is positive correlation.

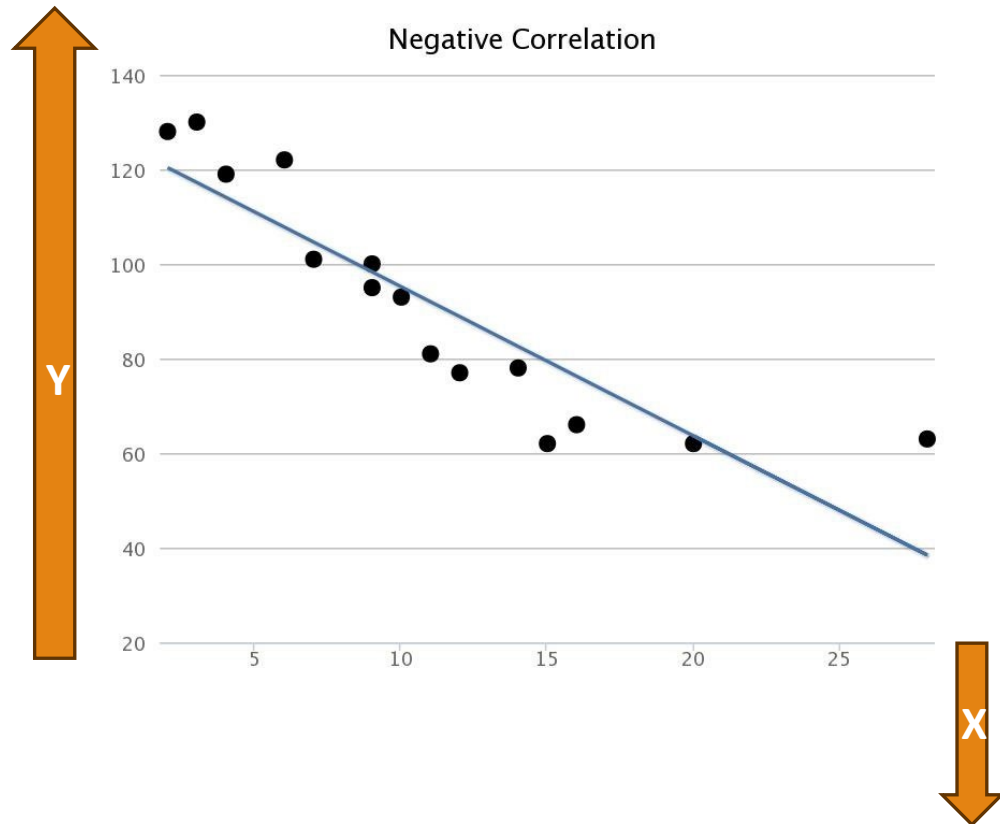
Age	1	2	3	4	5	6	7	8
Height	25	31	34	36	40	41	47	55

Negative Correlation

If the x-coordinates and the y-coordinates have one increasing and one decreasing, then it is **NEGATIVE CORRELATION**.

This means that 1 is going up and 1 is going down, making a downhill graph.

This means the two are related as opposites.

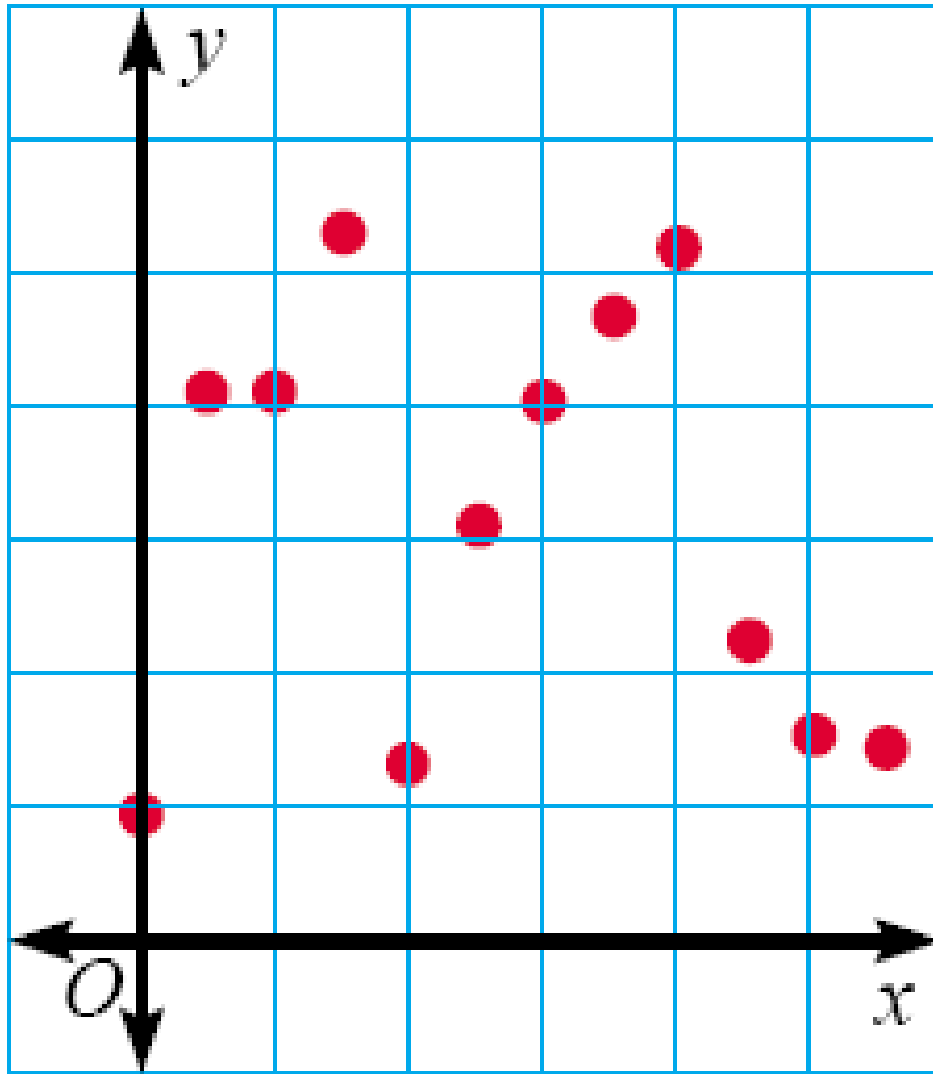


Negative Correlation

If you look at the age of a car and its value, you will find as the car gets older, the car is worth less.

This is negative correlation.

Age of car	1	2	3	4	5
Value	€30,000	€27,000	€23,500	€18,700	€15,350



No Correlation

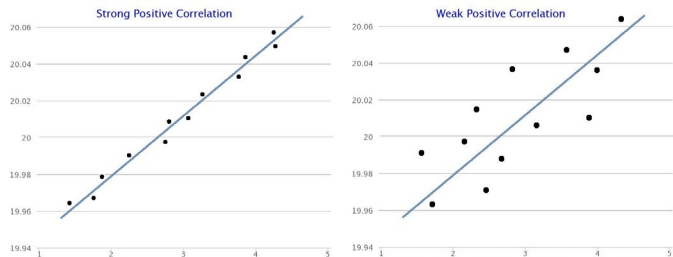
If there seems to be no pattern, and the points looked scattered, then it is no correlation.

This means the two are not related.

Strength of the relationship

Consider how close the Data Points are to the line:

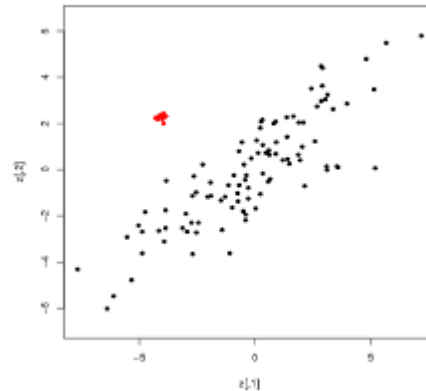
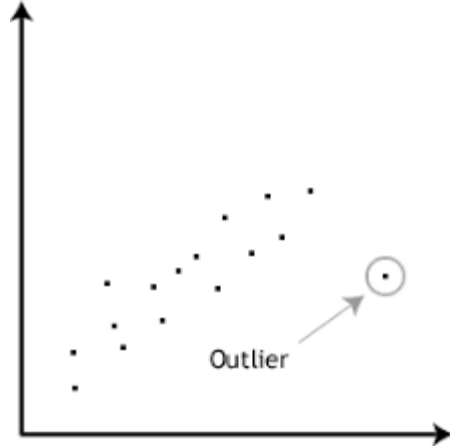
- If the points are widely spread from the trend line, it indicates more variability in the data and thus a weaker relationship.
- If the points are tightly packed along the line, it indicates less variability and a stronger relationship.
- Strong Relationship:
 - If the points are closely clustered around a straight line (for a linear relationship) or a curve (for a nonlinear relationship), it indicates a strong correlation. The points follow a clear pattern.
- Weak Relationship:
 - If the points are more scattered and don't form a clear pattern, the relationship between the variables is weak or even nonexistent.



Outliers

Presence of outliers (points far away from the general trend) can weaken the strength of the relationship, especially if the rest of the points follow a pattern.


You need to investigate these outliers in your initial data inspection to ensure they could not unduly influence your analysis.



Correlation

The relationship between bivariate numerical variables

- May be positive or negative
- May be weak or strong

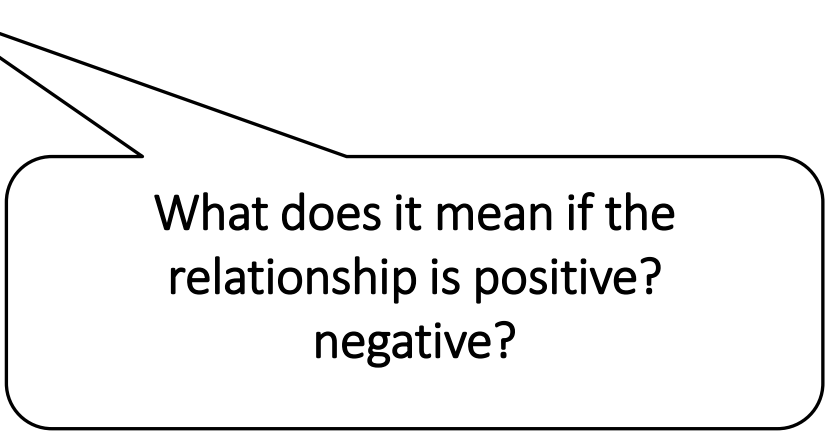


What does it mean if the
relationship is positive?
negative?

Correlation

The relationship between bivariate numerical variables

- May be positive or negative
- May be weak or strong



What does it mean if the
relationship is positive?
negative?

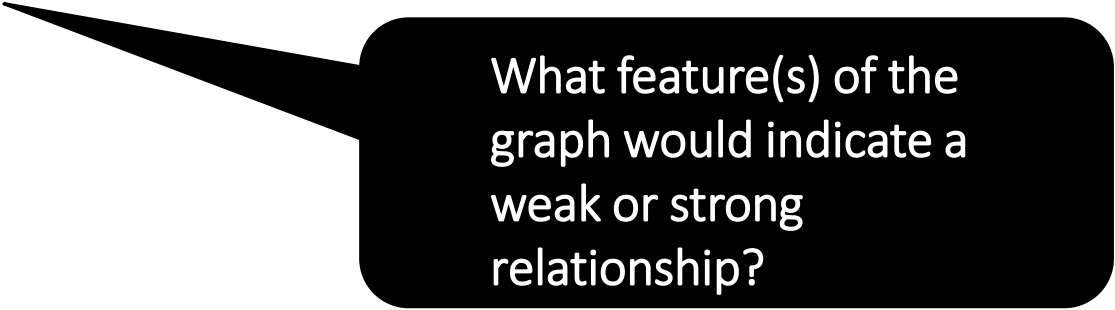
A positive relationship is one where as x increases, y increases.

A negative relationship is one where as x increases, y decreases.

Correlation

The relationship between bivariate numerical variables

- May be positive or negative
- May be weak or strong

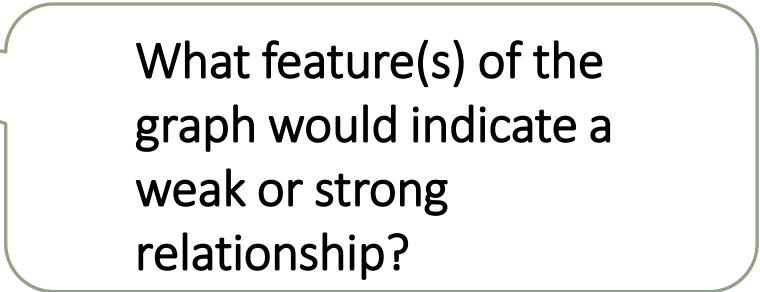


What feature(s) of the graph would indicate a weak or strong relationship?

Correlation

The relationship between bivariate numerical variables

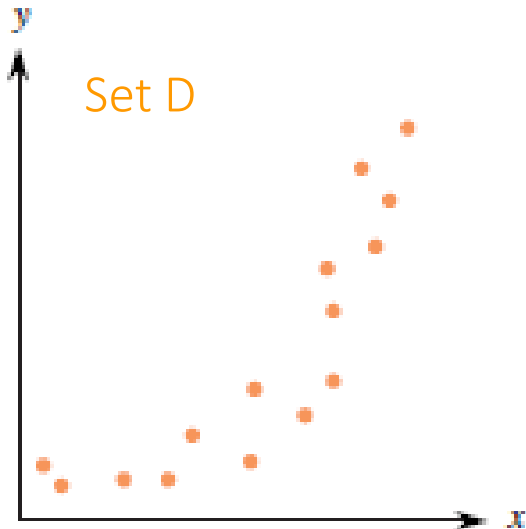
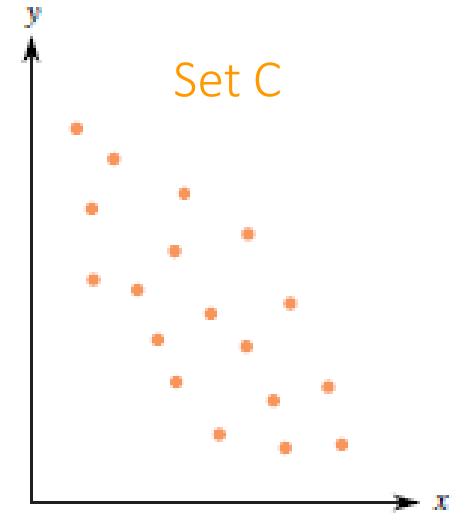
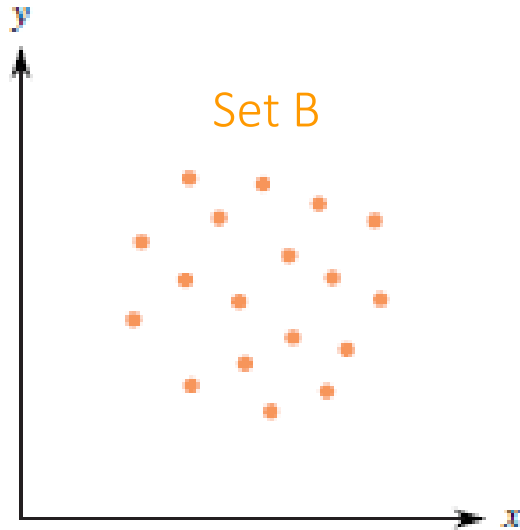
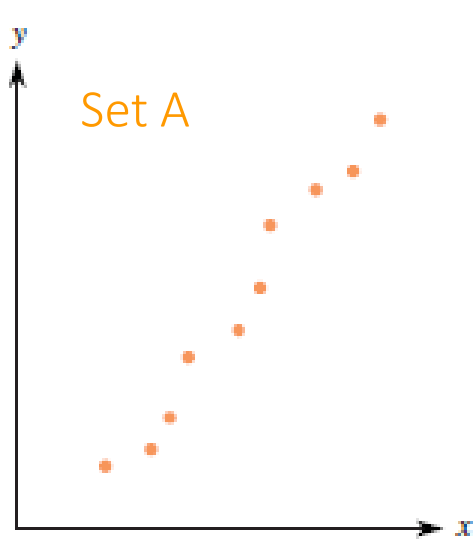
- May be positive or negative
- May be weak or strong



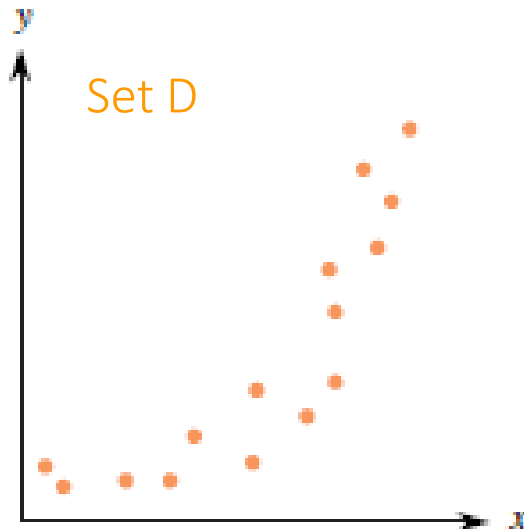
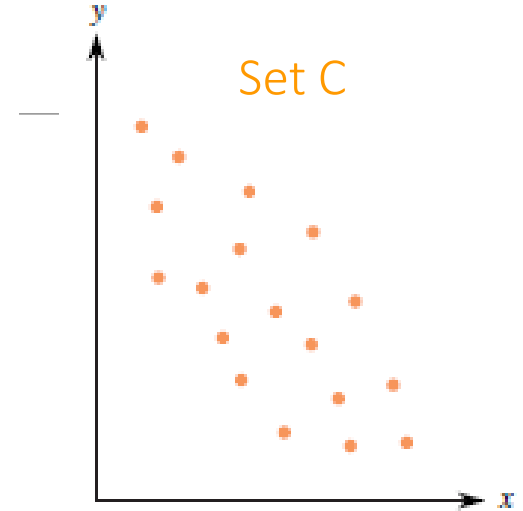
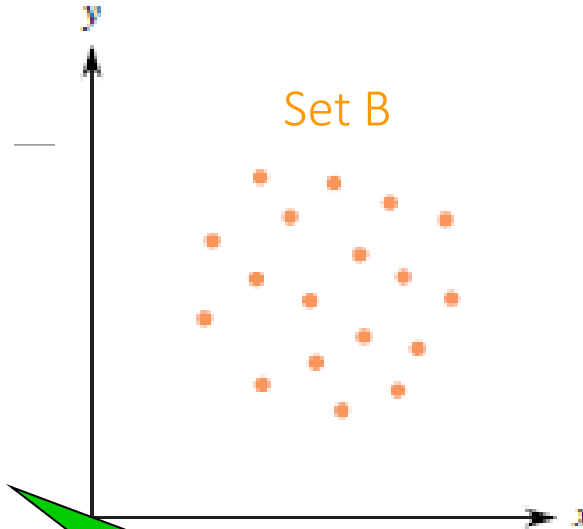
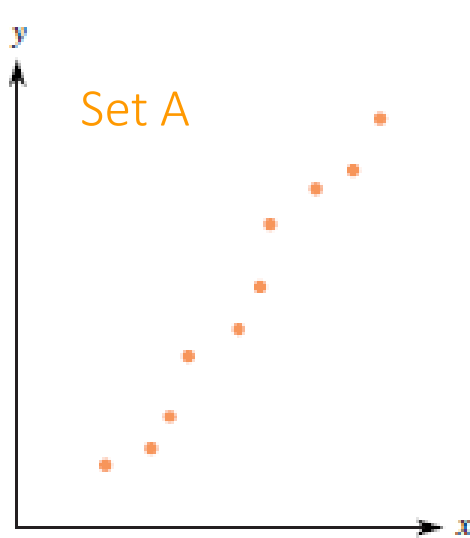
What feature(s) of the graph would indicate a weak or strong relationship?

Closeness of Data Points to a Line
Spread of the Data
Outliers

Identify the strength and direction of the following data sets:

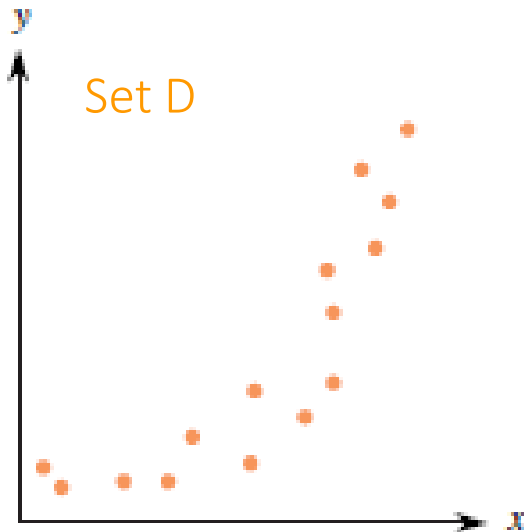
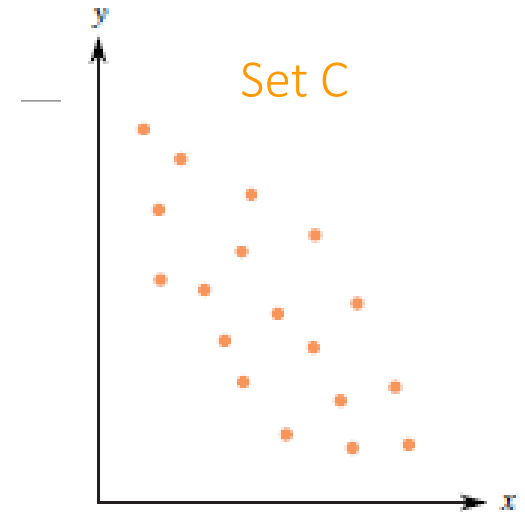
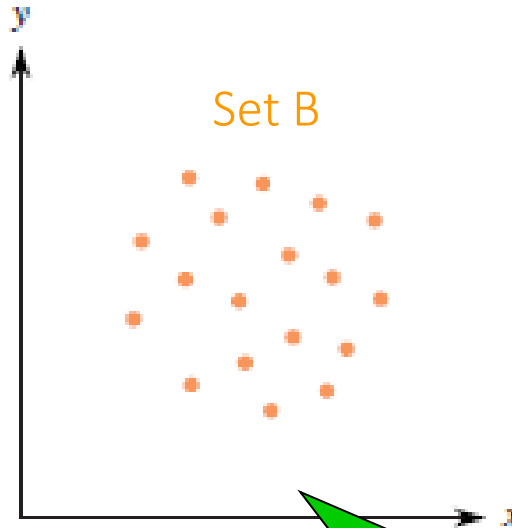
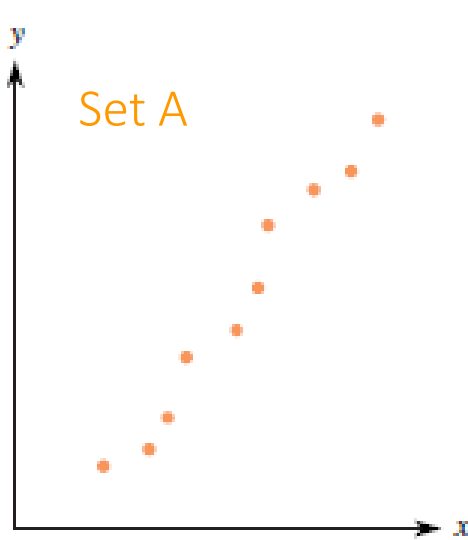


Identify the strength and direction of the following data sets.



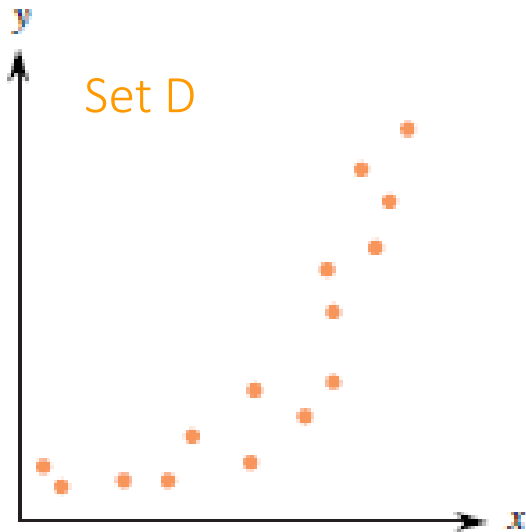
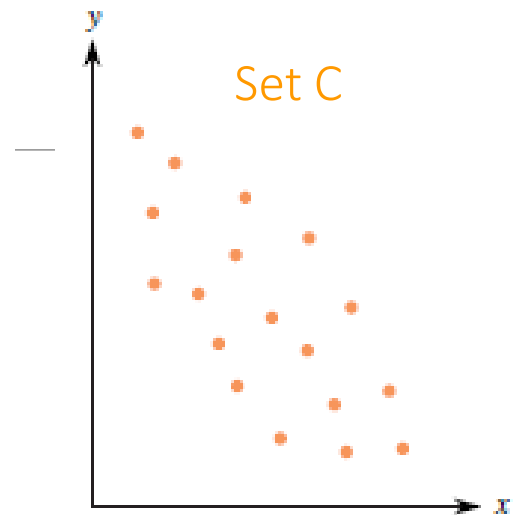
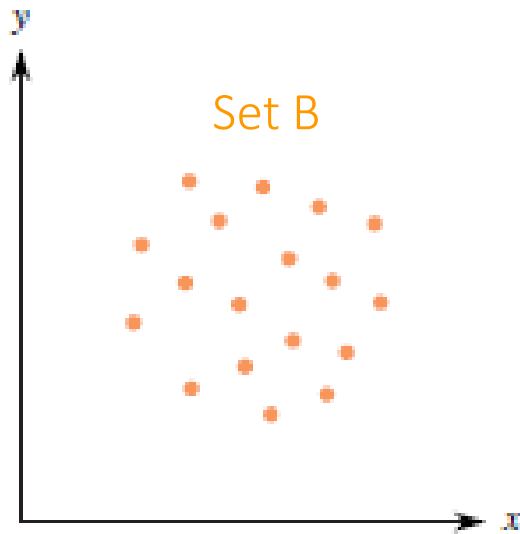
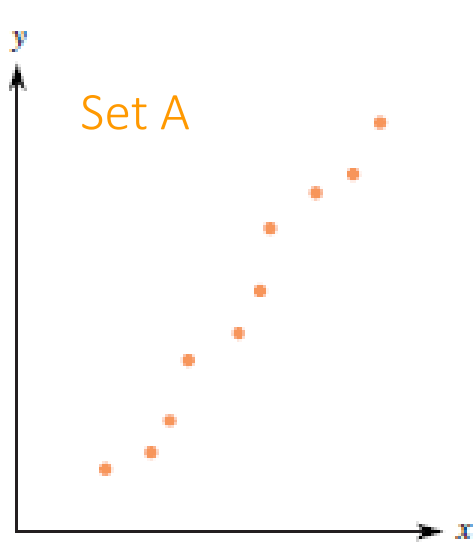
Set A shows a strong, positive linear relationship.

Identify the strength and direction of the following data sets.



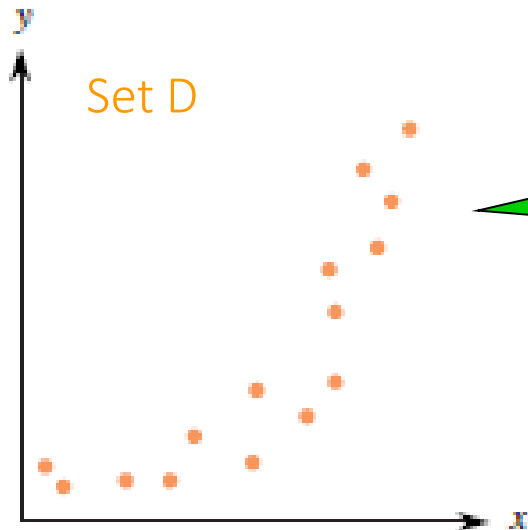
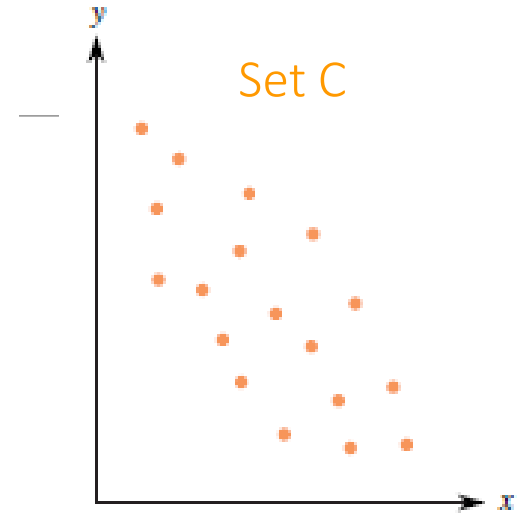
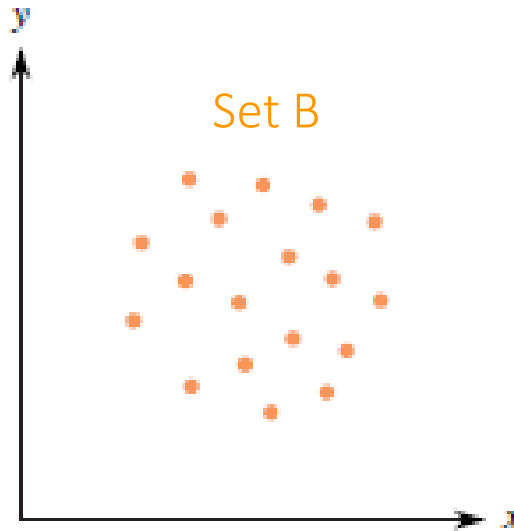
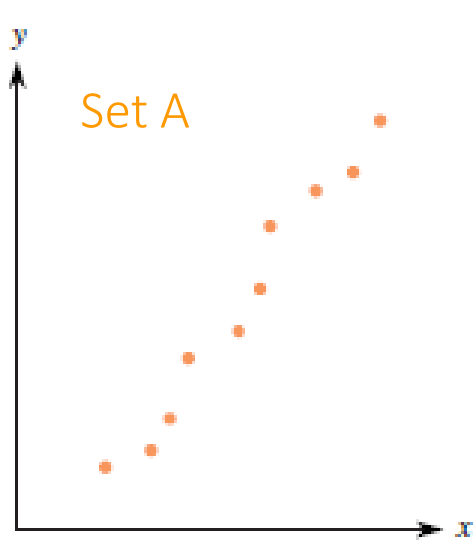
Set B shows little or no relationship.

Identify the strength and direction of the following data sets.



Set C show a weaker (moderate), negative linear relationship.

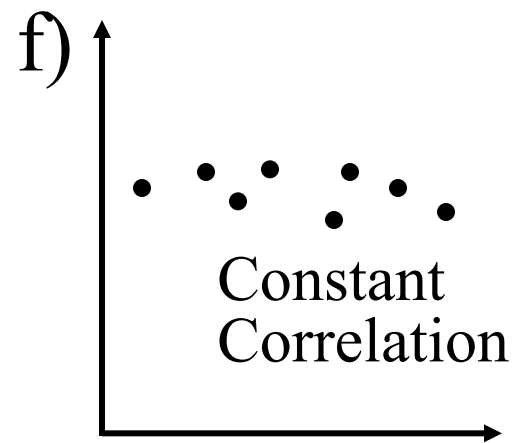
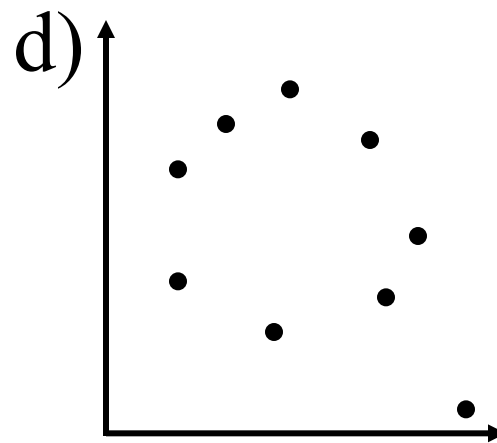
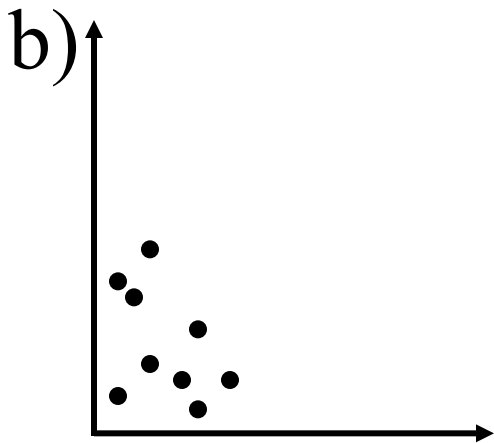
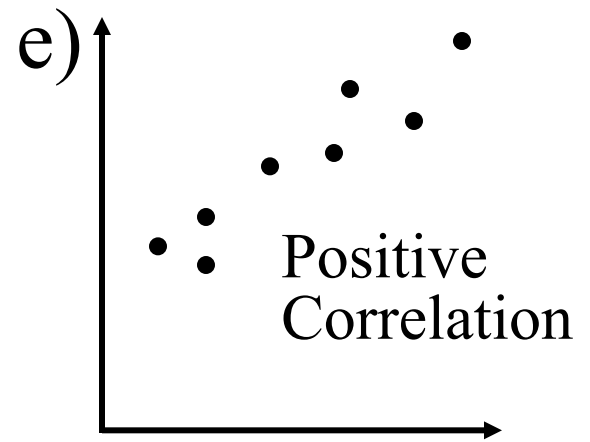
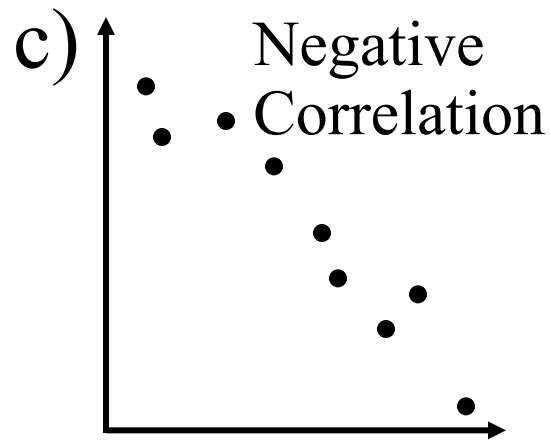
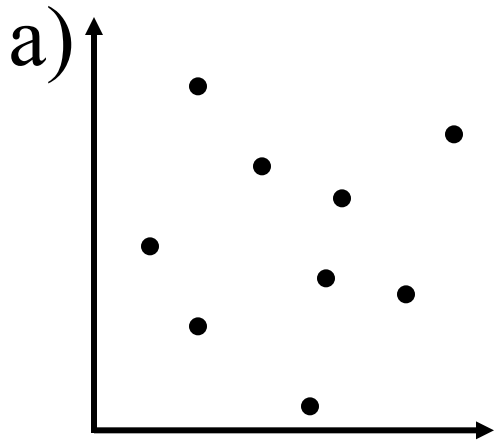
Identify the strength and direction of the following data sets.



Set D shows a strong,
positive **curved**
relationship.

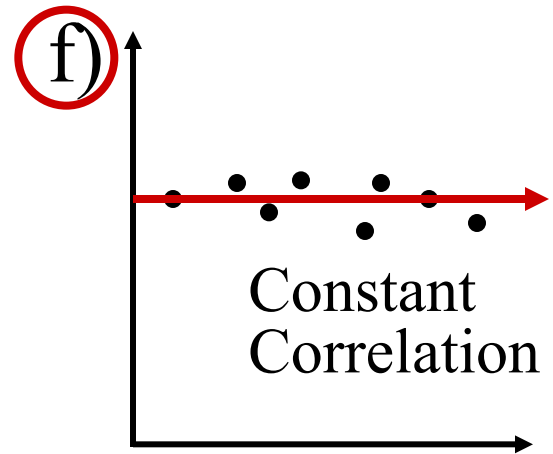
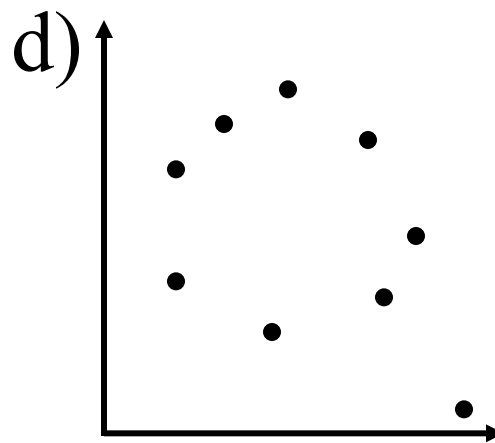
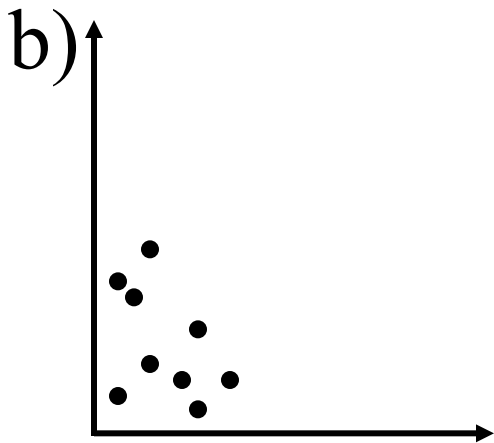
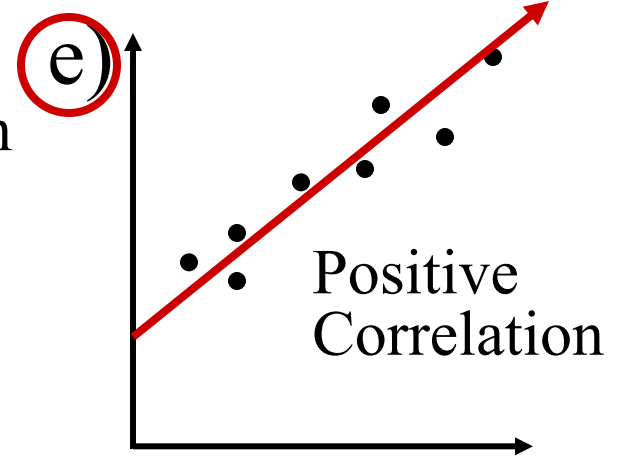
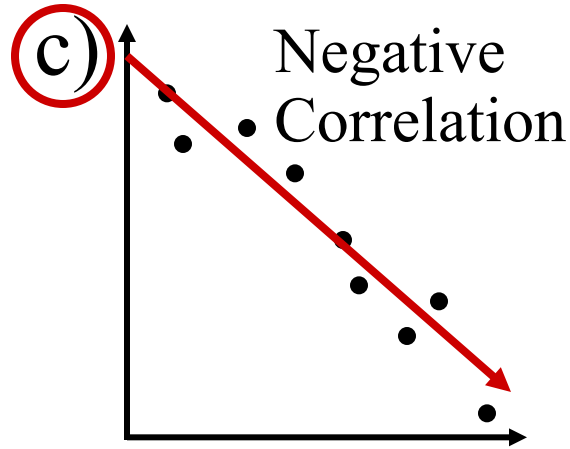
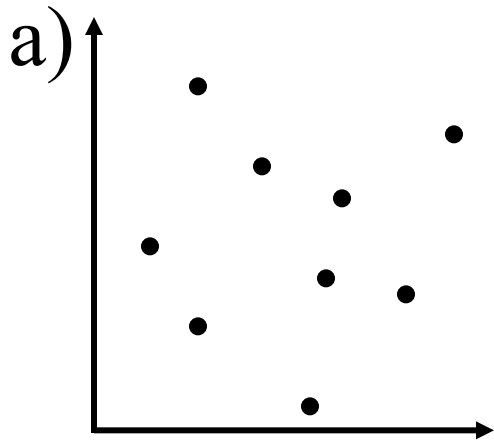
Scatterplots

Which scatterplots below show a linear trend?



Scatterplots

Which scatterplots below show a linear trend?



Statistical Model

Equation for any model:

$$\text{Outcome}_i = (\text{model}) + \text{error}_i$$

For a Linear model:

$$\text{Outcome}_i = \Sigma(\beta \times \text{predictor variable}_i) + \text{error}_i$$

$$\text{Dependent}_i = \Sigma(\beta \times \text{independent variable}_i) + \text{error}_i$$

β = weight that describes the relationship between the dependent and independent variable

Represents the amount that the outcome variable changes for one unit change in the independent variable

For a linear relationship:

$$y(\text{dependent/outcome variable}) = \Sigma(b_0 + b(\text{slope of line}) \times x(\text{predictor/independent variable}))$$

Simple Example

Suppose we want to look at children's maths scores at age 16 and their achievement of a standard maths test at aged 7 (key stage 1).

We are interested to see if the score a child achieves at age 7 is related to the score that they achieve at age 16.

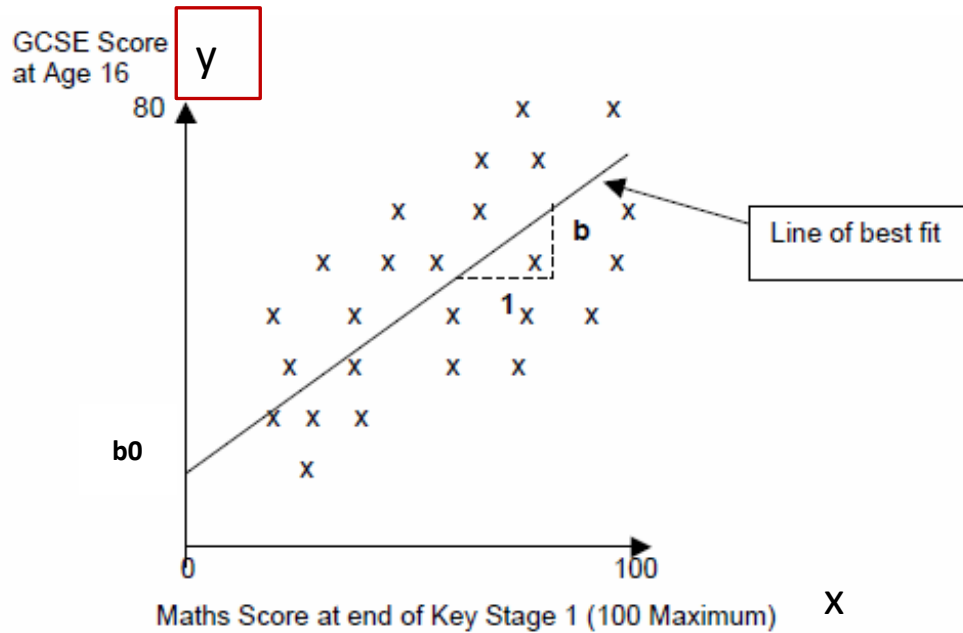
- And if they are related, what is the direction and strength of the relationship

$$y = b_0 + bx + e$$

'y' is the **outcome** variable (in this case maths score age 16) – the **dependent** variable

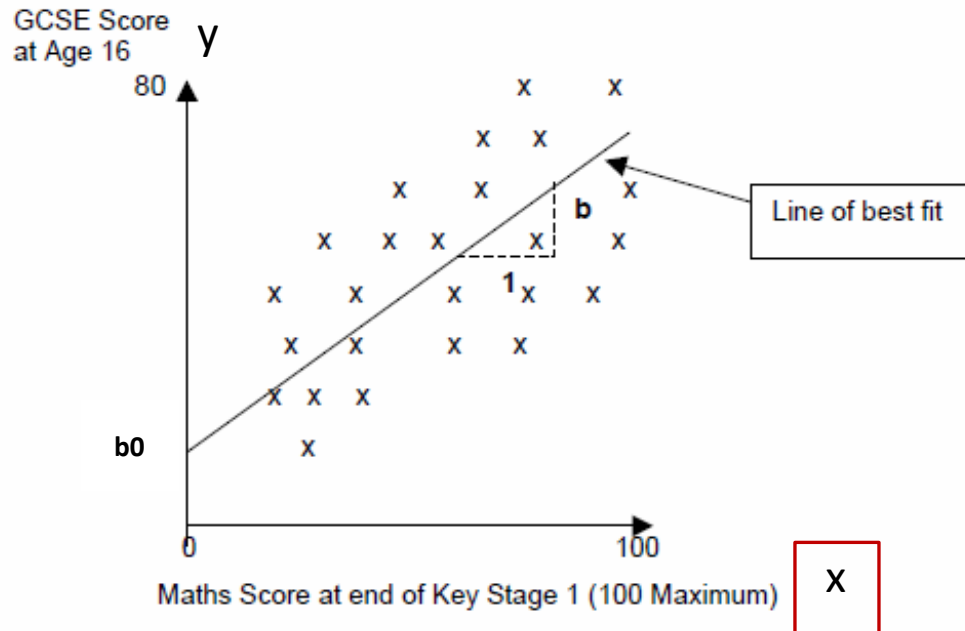
Simple Example

Relationship between
Maths score age 7
with score at age 16
(for 25 students)



Simple Example

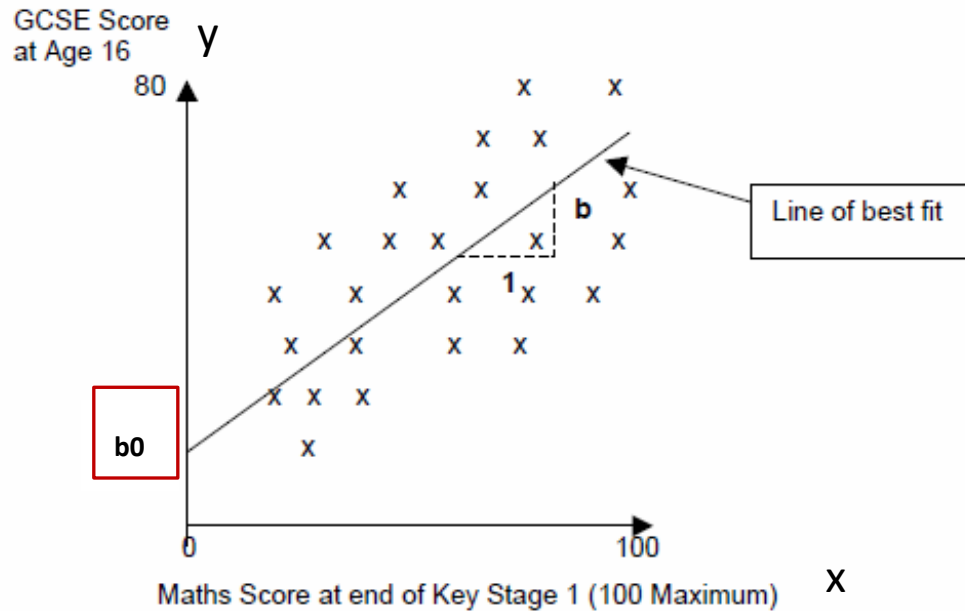
$$y = b_0 + bx + e$$



'x' is the **predictor** variable (in this case score age 7) – the **independent** variable

$$y = b_0 + bx + e$$

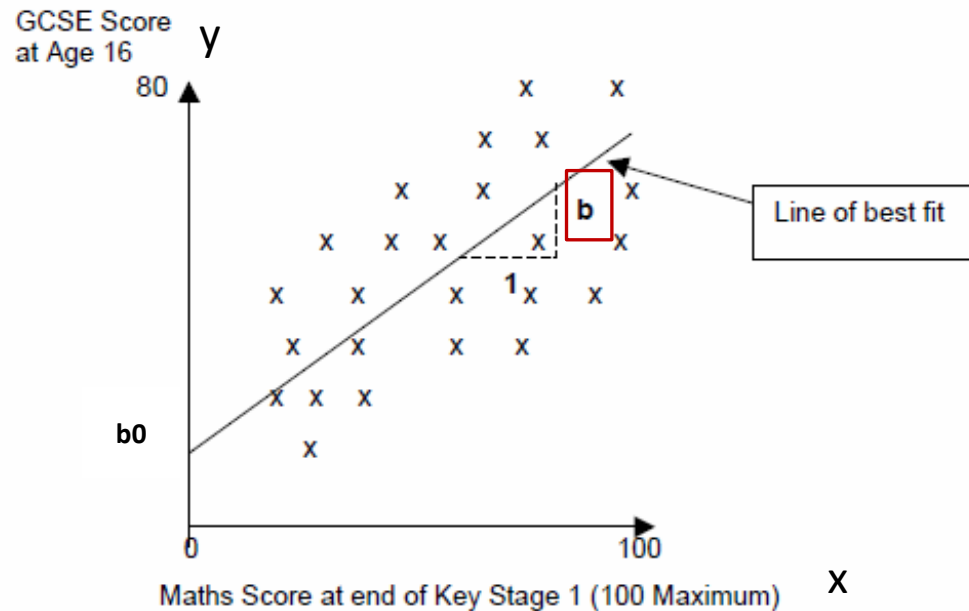
Simple Example



'b0' is the **intercept** or the point where the line crosses the y-axis (y value when $x=0$ i.e. value of maths score at age 16 if score at age 7 is 0)

$$y = b_0 + bx + e$$

Simple Example

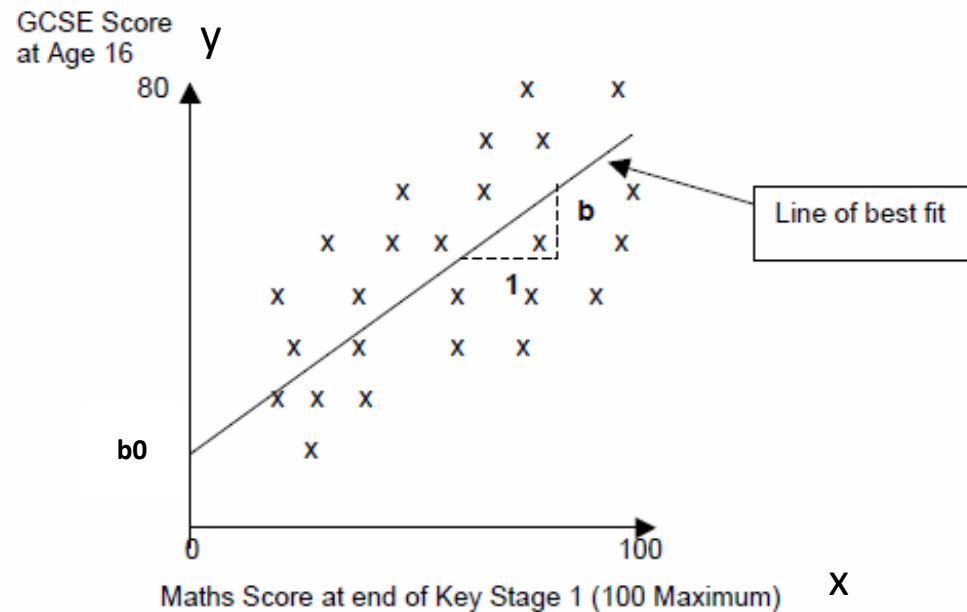


'b' is the **gradient** of the line

- Represents the amount that the outcome variable changes for one unit change in the independent variable
 - i.e. for every one percentage point increase in a child's Maths Test score, the line suggests that the child's maths score age 16 increases by 'b' points

$$y = b_0 + bx + e$$

Simple Example



If the line did completely model the data, then all of the points would rest exactly on the line.

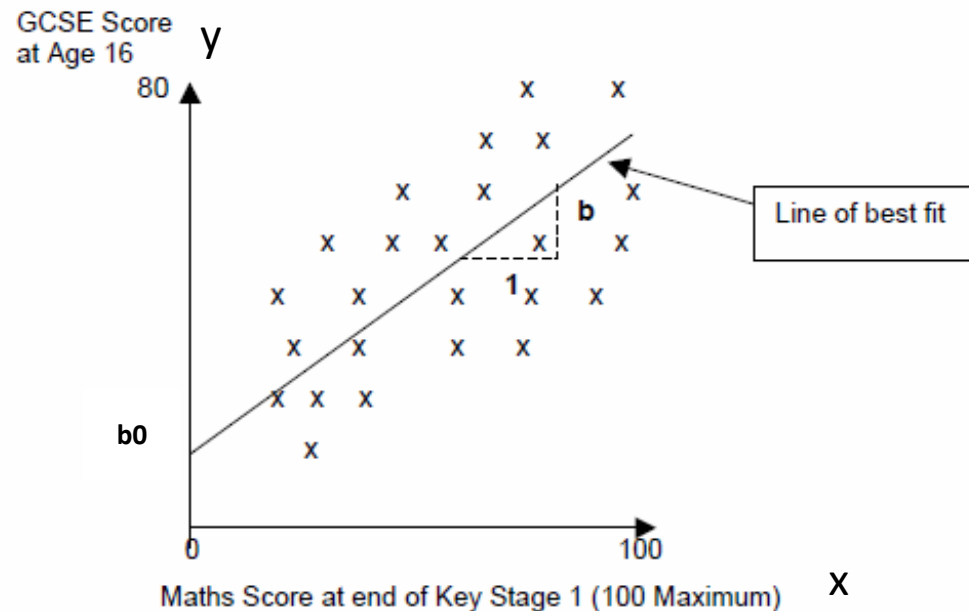
But they don't - e is the vertical distance between each point and the line itself.

- The model is a best fit: ' e ' obviously varies for each data point.
 - Without any further information about confounding variables, we cannot explain this variation – so we include it as an error.

Simple Example

However, if we are dealing with a normal distribution these error terms will cancel each other out and we do not need to include it in the equation.

$$y = b_0 + bx$$



We are really looking at Co-variation

World is full of co-variation

- Nutrition and growth
- Pollen and bees
- Violence on TV and violence in Society ?

What is a Correlation?

It is a way of measuring the extent to which two variables are related.

It measures the pattern of values across variables.

It is used to describe the strength and direction of the linear relationship between two variables.

- Parametric Test (normal distribution): Pearson Correlation (statistic is r)
- Non-parametric Test (non-normal distribution): Spearman Rank Order Correlation (statistic is ρ) or Kendall's Tau (statistic is τ)

$$r = \frac{1}{n-1} \left(\frac{\sum_x \sum_y (x - \bar{x})(y - \bar{y})}{s_x s_y} \right)$$

Measuring Relationships

We are investigating whether as one variable increases, the other increases, decreases or stays the same.

We assess the relationship via the **correlation coefficient**.

- The **sign** of will give us the direction (positive or negative)
- The **magnitude** will give us the strength

.

We can look at a bi-variate correlation or a partial correlation

- Bi-variate – two variables
- Partial – two variables while controlling for another

Measuring Relationships

By calculating the **Covariance**.

- We look at how much each score deviates from the mean.
- If both variables deviate from the mean by the same amount, they are likely to be related.

Linear Correlation

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

We are interested in

- Direction (+/-)
- Strength (Weak/Moderate/Strong)
 - Values closer to +1 or -1 indicate stronger relationship
- Statistical Significance
 - Likelihood the relationship we observe is occurring due to chance

Calculating Pearson Correlation Co-efficient

We have a number (n) of data pairs (x,y)

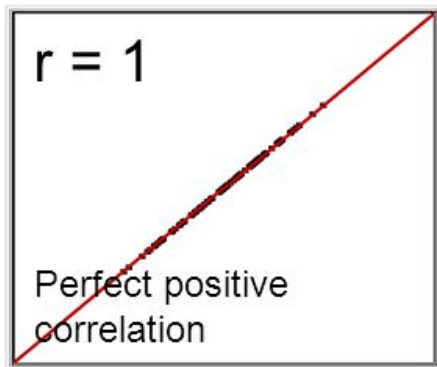
We calculate the mean of x and the mean of y

We calculate the std. deviation of x and y (s_x and s_y)

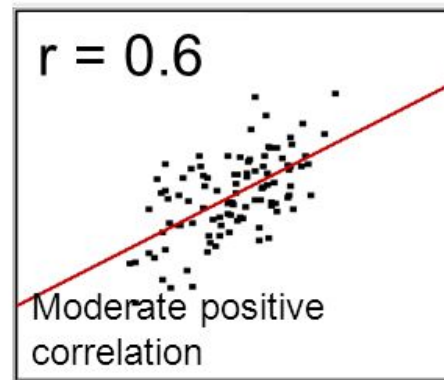
$$r = \frac{1}{n-1} \left(\frac{\sum_x \sum_y (x - \bar{x})(y - \bar{y})}{s_x s_y} \right)$$

Strength of relationships

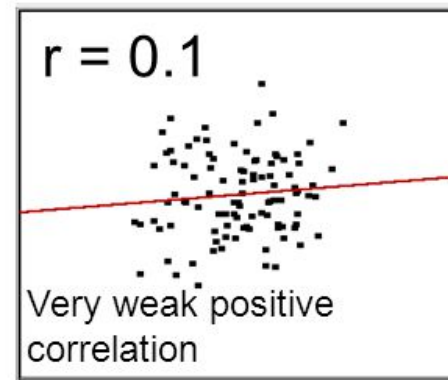
Correlation coefficient



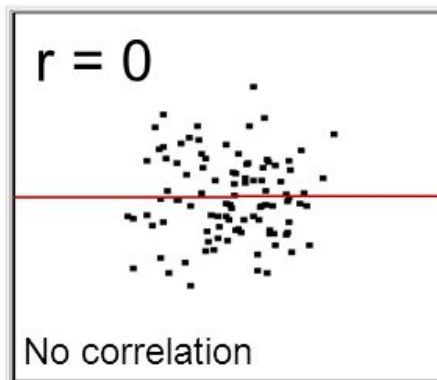
Correlation coefficient



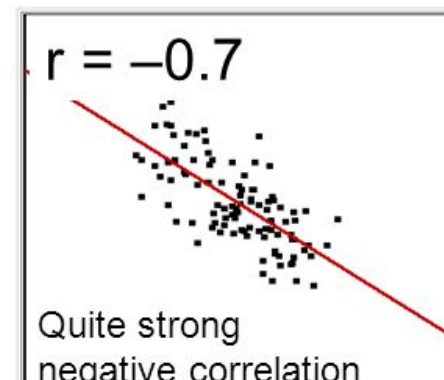
Correlation coefficient



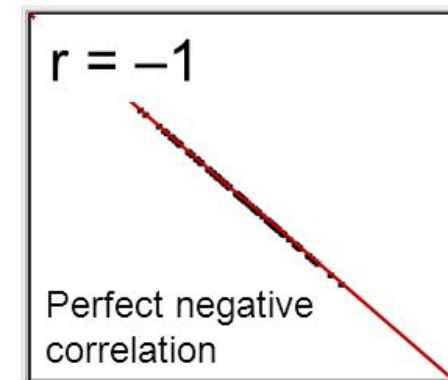
Correlation coefficient



Correlation coefficient



Correlation coefficient



How do we calculate the strength and direction?

Determined by testing the **null hypothesis**

- Usually states that there is no correlation between the two variables
 - i.e., the population correlation coefficient = 0

1. Calculate the Correlation Coefficient

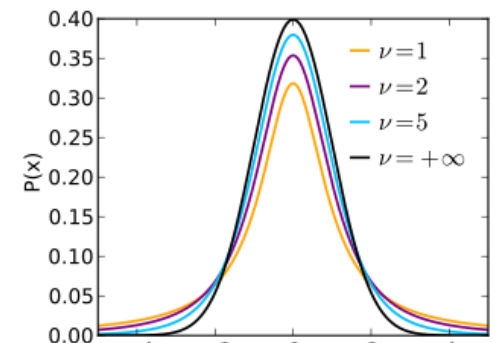
- Computed from the sample data.
- Summarizes the strength and direction of the linear relationship between the two variables.
- Pearson's r , Spearman ρ , Kendall τ

2. A test statistic is calculated to assess how likely it is that the observed sample correlation occurred by chance, under the assumption that the null hypothesis holds

- For Pearson the statistic used to determine this is t , calculated as $t = r\sqrt{(n-2)/(1-r^2)}$
- The degrees of freedom for this test are $n - 2$
- This is because there are two variables involved.

How do we calculate the strength and direction?

3. Compute a p-value for the test statistic
 - P value is determined using the Student's t-distribution with $n - 2$ degrees of freedom.
 - The t statistic is looked up in the associated table to identify the probability.
 - We can compare our statistic to the **critical values** for an idealised distribution
 - We look it up in a table that shows how extreme a statistic needs to be for a given number of degrees of freedom to be considered statistically significant.
 - Indicates the probability of obtaining a correlation at least as extreme as the one observed, assuming the null hypothesis holds.
 - A low p-value (less than the cut-off level for your field of work) suggests that the correlation is statistically significant, meaning it is unlikely that the observed correlation happened by chance.



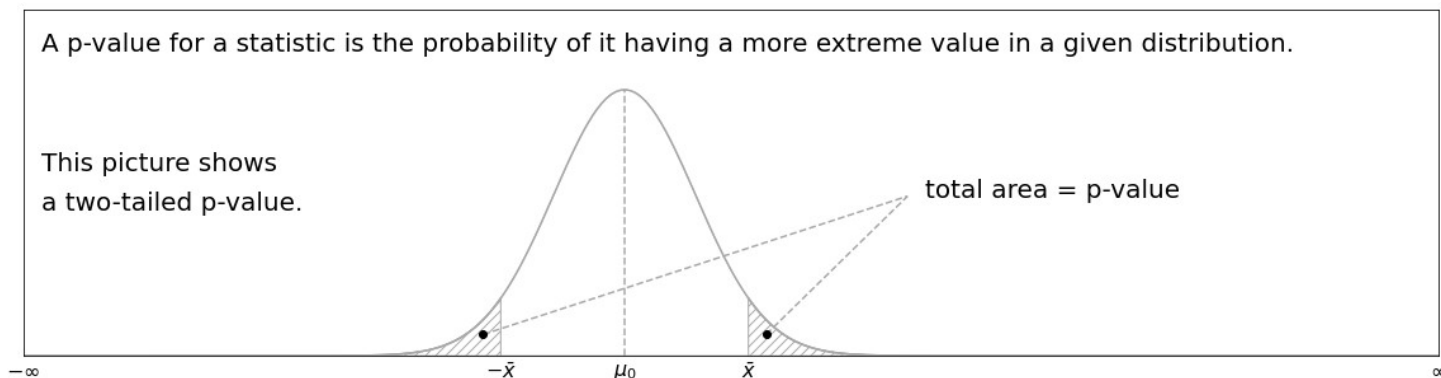
This Photo by Unknown Author is licensed under CC BY-SA

P-value

The p-value is the central concept in statistical hypothesis testing.

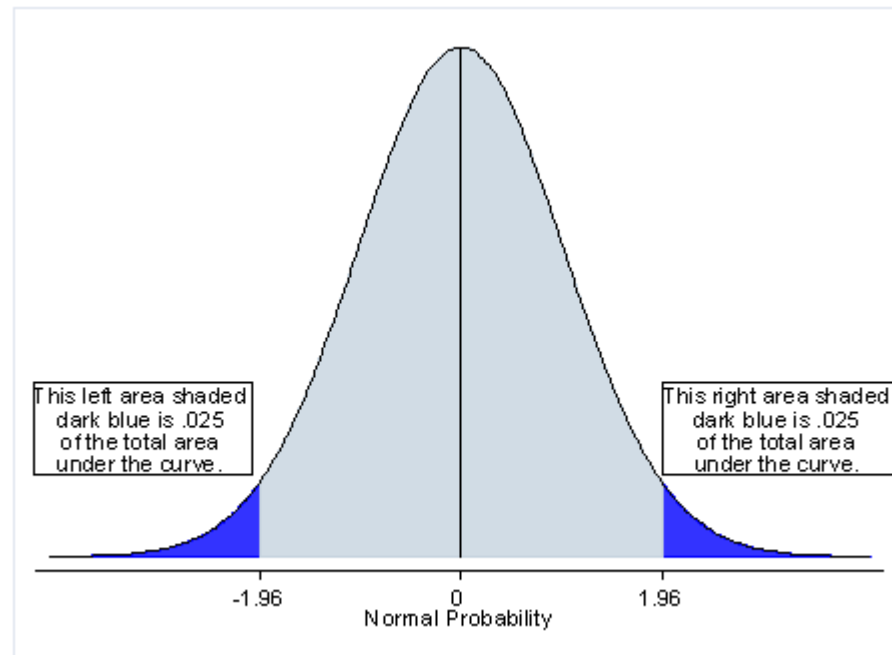
A p-value of a statistic with respect to a distribution is the probability of more extreme values for the statistic being drawn from the distribution

- The extreme values are in the tails of the distribution
- Depending on how the variable is expected to behave or its meaning in the context of the analysis, its p-value may be defined as one-tailed or two-tailed



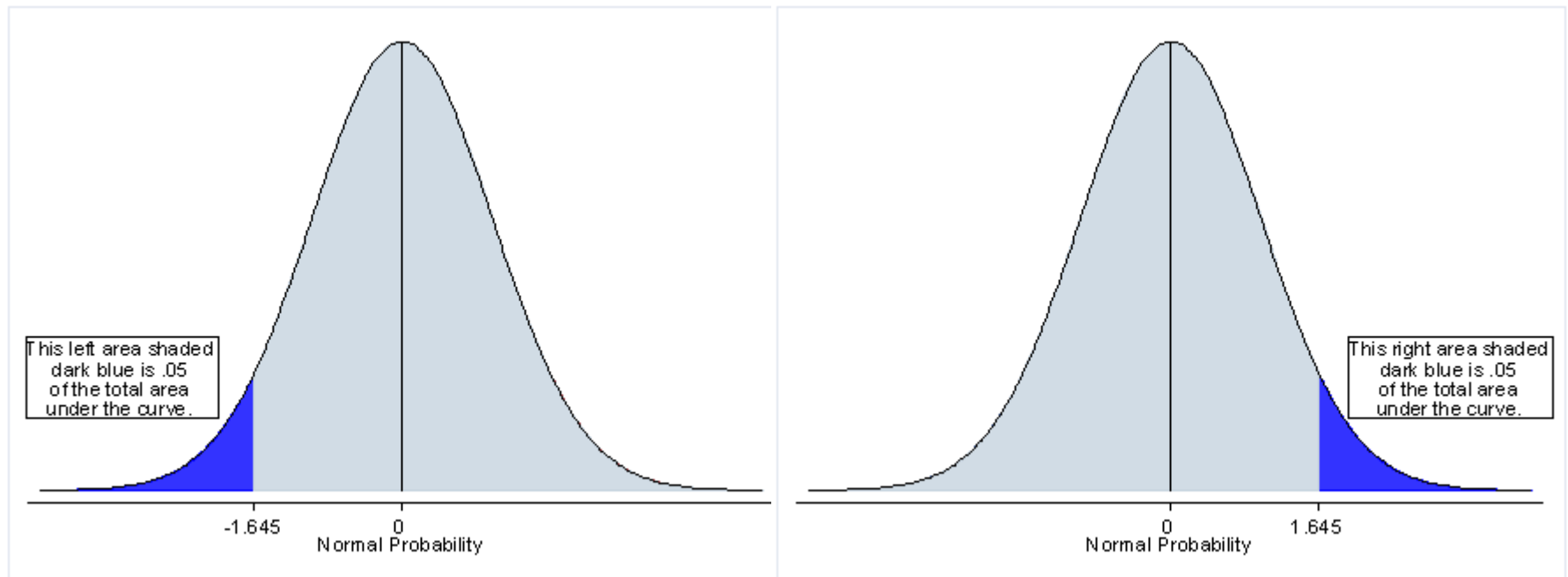
P-value

If we are looking at a two tailed hypothesis (at cut-off level of 0.05) we look at both extremes of the distribution:



P-value

If we are looking at a one tailed hypothesis (at a cut-off of 0.05) we look at only one side of the distribution:



P-value

The **threshold of probability** or **level of significance** (denoted as α) is the level of significance is the threshold at which you decide whether to reject the null hypothesis.

It represents the probability of making the error of incorrectly rejecting a null hypothesis.

Common levels of significance are:

5% (0.05):

- This means that you are willing to accept a 5% chance of incorrectly rejecting the null hypothesis.

1% (0.01):

- This means you are willing to accept only a 1% chance of making that error.

0.1% (0.001):

- This means you are willing to accept only a 0.1% chance of making that error.

P-value and Confidence Interval

A confidence interval is a range of values that is likely to contain the true population parameter (e.g., a mean or proportion) with a certain degree of confidence.

The level of confidence is typically $1-\alpha$:

- If α is 0.05 (5%), the confidence interval would be 95%.
 - A 95% confidence interval means that if you were to repeat the experiment 100 times, in 95 of those experiments, the calculated interval would contain the true population parameter.
- For $\alpha=0.01$, the confidence interval would be 99%.

P-value

Decide in advance your cut-off value (0.05, 0.01)

P-value \geq cut-off

- No evidence to reject null hypothesis
- Report decision and report the actual p-value

P-value $<$ cut-off

- Need to also consider the strength of your statistic
- Have evidence to reject null hypothesis in favour of the alternate
- Report $p <$ your cut-off
 - Note: convention is if p value is $< .000$ then report it as < 0.001 even if working at a level of 0.05.

Parametric v Non-parametric

Parametric

- Make assumptions about the population from which the sample is taken
- Shape of the population (normally distributed)

Non-parametric

- Do not make assumptions about the population and its distribution
- Tolerant set of tests which don't expect your data to anything fancy
 - Not high-powered and don't promise more than they can deliver
 - May fail to detect differences that exist
- Use for nominal or ordinal data
- Use for small samples
- Use for skewed data

Pearson's Correlation - Assumptions

Level of measurement

- Two Interval or ratio (scale) variables
- Exception:
 - You can have one independent variable with two categories (e.g. gender) and one continuous dependent.
 - Caveat: you must have approximately the same numbers of cases for each category of the categorical variable

Pearson Correlation - Assumptions

Related Pairs

- Each case must provide a score on the two independent variables

Independence of observations

- Each measurement must not be influenced by any other
 - E.g. if studying TV habits on children and all children are from same family then behaviours of one child are likely to affect all so observation is unlikely to be independent

Pearson Correlation - Assumptions

Normality

- Scores should be normally distributed
- Inspect histograms for each variable

Linearity

- There must be a linear relationship between the two variables
- Inspect a scatterplot and you should see a straight line not a curve

Homoscedasticity

- Variability of variable 1 should be similar to variable 2
- Check scatterplot
 - Looking at distance between the points to that straight line.
 - The shape of the scatterplot should be tube-like or rectangular in shape.
 - If the shape is cone-like, then homoscedasticity would not be met.
- It is a matter of degree



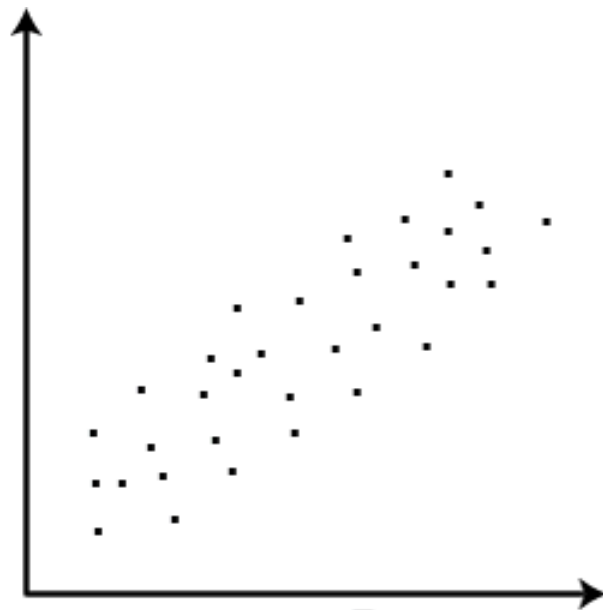
Key
Slide

Homoscedasticity

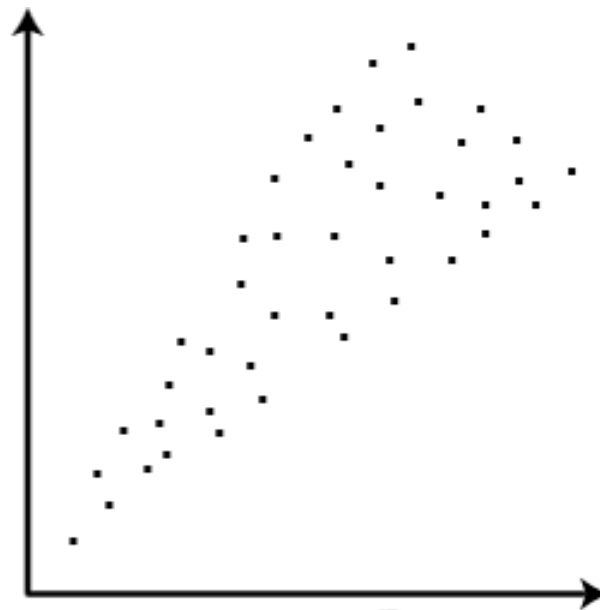
Suppose that we are interested in the relationship between income levels and spending on gadgets.

- We want to investigate if income level could be considered to predict the level of spending on gadgets.
- We find a strong, positive association between income and spending.
 - So far so good.
- But when we look at our pattern graphically, we find the levels of spend are low for low incomes
 - This makes sense people with low incomes don't spend lots of money on luxury items
- And we find the level of spend varies for those with high incomes
 - Again, this makes sense, some people are more moderate in their spending than others
- We therefore have **heteroscedasticity** which means that it doesn't make sense to base any prediction based on this relationship

Homoscedasticity



Homoscedasticity



Heteroscedasticity



Getting Started With Analysis

Inspect your data

Generate your descriptive statistics

Generate your visuals (graphs)

Make decisions about normality

- Choose the correct tests



Conducting Correlation Analysis

Check
assumptions/bias

Example Pearson Correlation

For the dataset survey.dat

- SPSS Survival Manual 6th Edition Julie Pallant
- <http://spss.allenandunwin.com.s3-website-ap-southeast-2.amazonaws.com/data-files.html#.Wb0vvnWP-po>
- Dataset created from a designed to explore the factors that impact on respondents' psychological adjustment and wellbeing.

PSIWeek3.rmd R notebook which contains all the commands used in this lecture.

Hypotheses

H0: There is **no** relationship between a respondent's perception of their ability to regulate and manage their internal state (i.e. emotions, thoughts, and physiological states) and their level of perceived stress.

Ha: There **is a** relationship between a respondent's perception of their ability to regulate and manage their internal state (i.e. emotions, thoughts, and physiological states) and their level of perceived stress.

- Two-tailed hypothesis

Data inspection

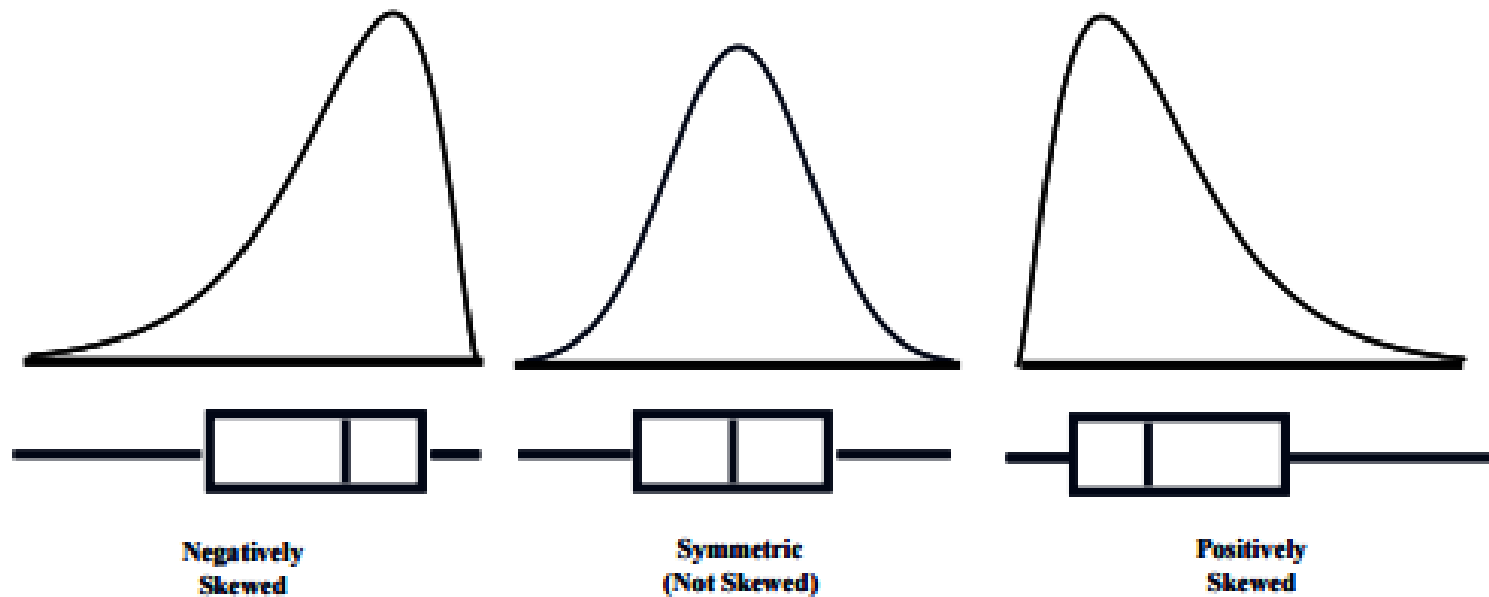
We need to:

- inspect to decide which type of test to use
- report this inspection and the conclusions we have drawn
- provide the appropriate descriptive statistics and visualisations

How do I inspect interval/ratio data?

1. Generate plots
 - Generate a histogram with a normal curve showing
 - Generate a Q-Q plot
2. Generate summary statistics
 - Central tendency and dispersion plus measures of skewness and kurtosis and assessing the percentage of standardised scores falling within acceptable limits.
3. Review your statistics and plots to see how far away from normal your data is
4. Report the correct statistics based on your assessment of whether your data can be considered to follow the normal distribution

Histogram



Normal Quantile Plot (Q-Q Plot)

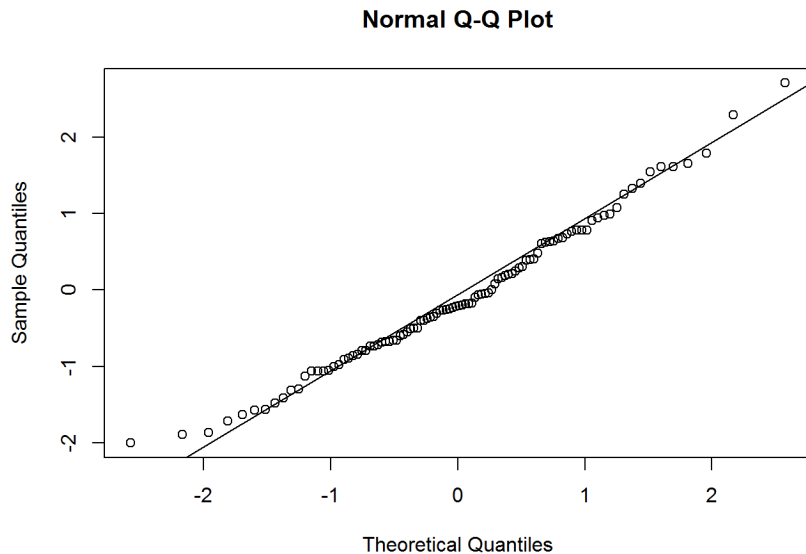
Quantile-Quantile Plot (Q-Q Plot)

Compares the quantiles of your data to the quantiles of a theoretical distribution (often the normal).

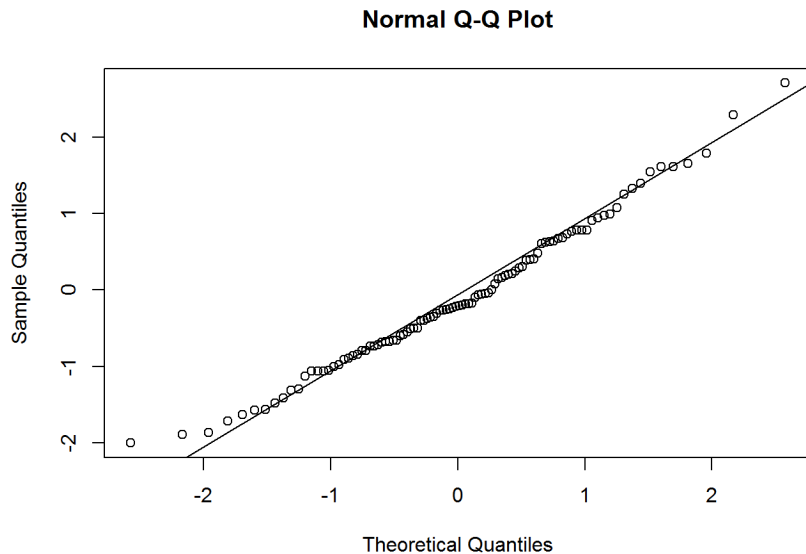
Helps us to assess whether our data can be considered to plausibly come from some theoretical distribution such as a Normal distribution

A scatterplot created by plotting two sets of quantiles against one another.

If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



Normal Quantile Plot (Q-Q Plot)



What are “quantiles”?

- These are often referred to as “percentiles”.

These are points in your data below which a certain proportion of your data fall.

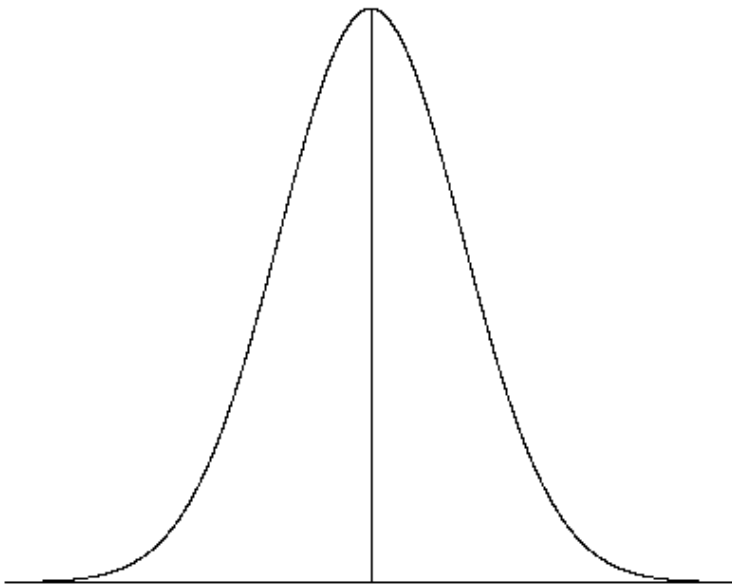
For example:

- For a standard Normal distribution with a mean of 0 (the peak of the bell curve)
- The 0.5 quantile, or 50th percentile, is 0.
 - Half the data lies below 0.
- The 0.95 quantile, or 95th percentile, is about 1.64.
 - 95 percent of the data lies below 1.64.

Quantiles are basically just your data sorted in ascending order, with various data points labelled as being the point below which a certain proportion of the data fall

Normal Quantile Plot (Q-Q Plot)

Basically compares the spacing of our data to what we would expect to see in terms of spacing if our data were approximately normal.



If our data is approximately normally distributed, we should see spacing that is similar that on the normal curve on the right.

Very few observations in both tails and increasingly more observations as we move towards the mean from either side.

Also remember the spacing must be symmetric about the mean.

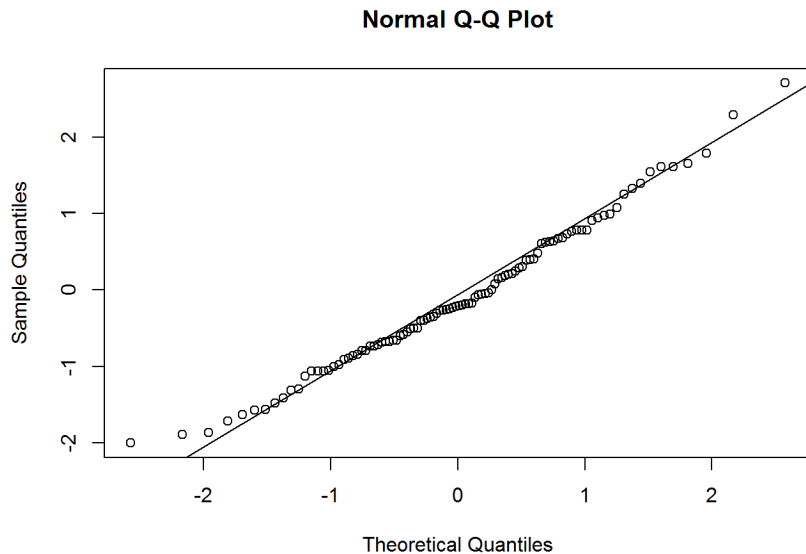
Normal Quantile Plot (Q-Q Plot)

THE IDEAL PLOT:

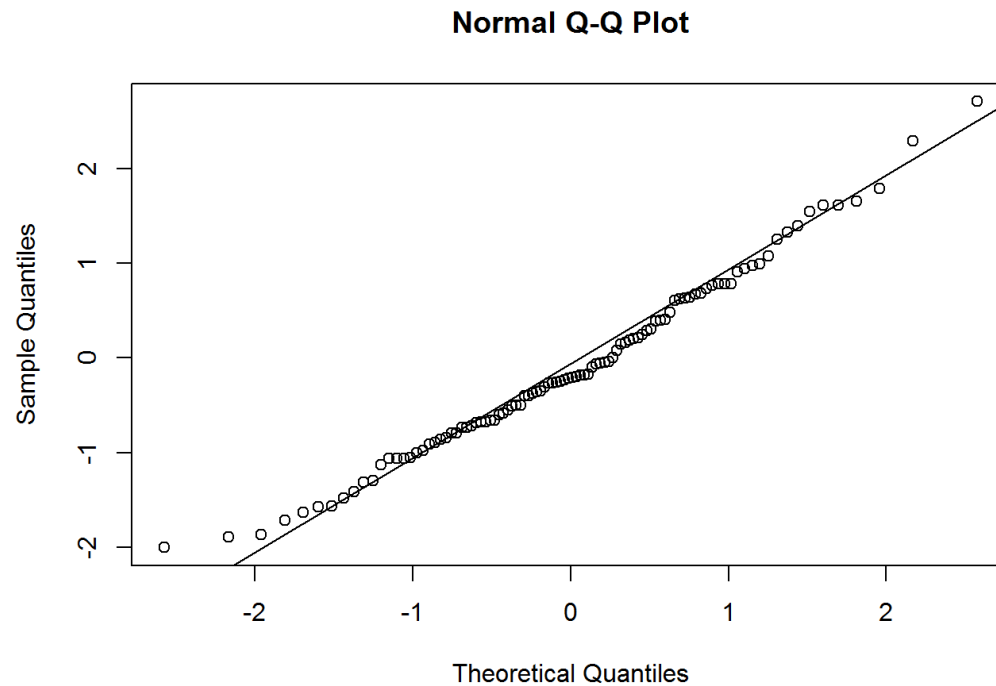
The plot on right is a Q-Q plot with the data on the vertical axis (observed) and the expected z-scores if our data was normal on the horizontal axis.

When our data is approximately normal the spacing of the two will agree resulting in a plot with observations lying on the reference line in the normal quantile plot.

When looking at a Q-Q plot, you should look for points that stray far from the line of expected values, as well as trends in the observed values.



Normal Quantile Plot



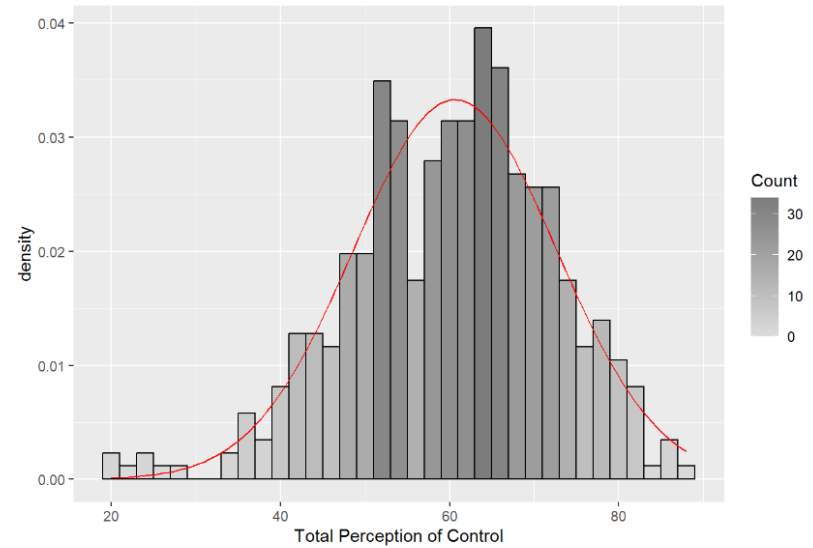
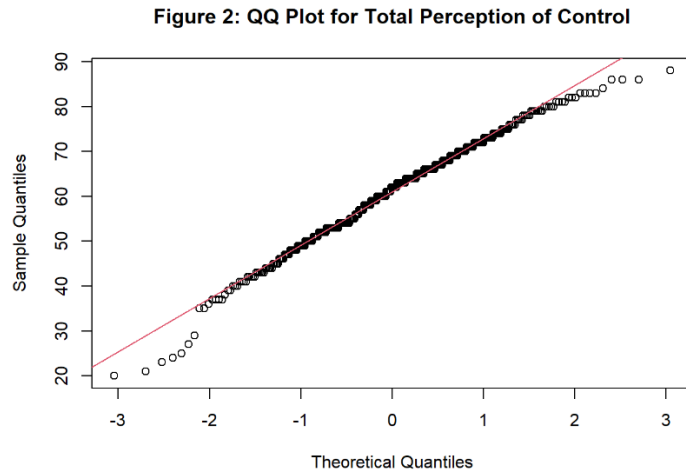


Figure 1: Histogram for Total Perception of Control

Step 1: Generate Plots for Perception of Control

INSPECT BY EYE

Step 2: Generate Summary Statistics for Perception of Control

Generate Descriptive Statistics

Before we do that, we need to see how far away from normal our data is:

- Calculate standardised scores for skew and kurtosis
 - Check if these are within acceptable limits.
- Calculate the percentage of standardised scores for the concept itself are outside the acceptable range
 - 95% within ± 1.96
 - 99.7% within ± 3.29 for larger distributions

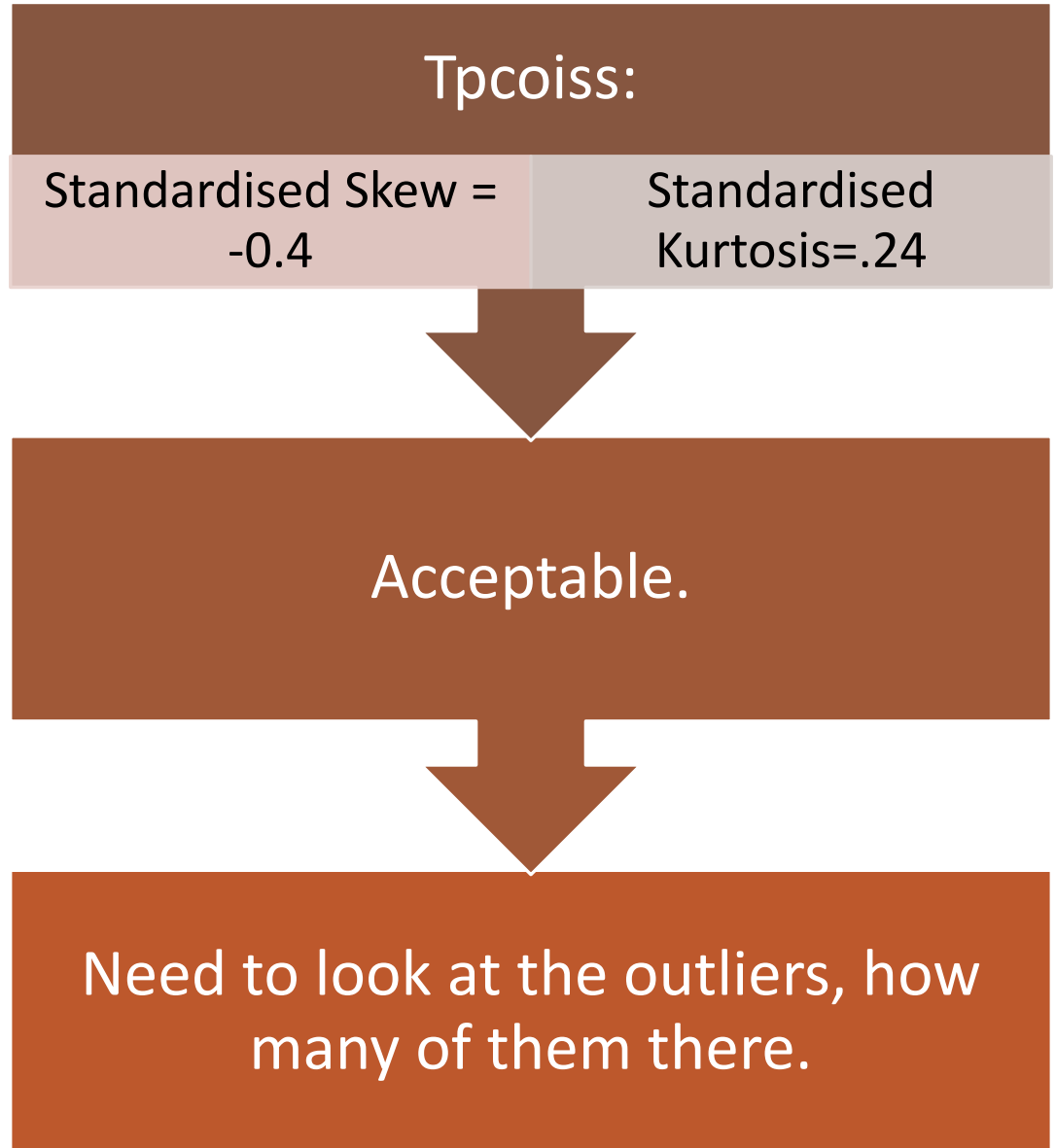
Check Standardized Skew and Kurtosis

Accepted Heuristics ONE

Calculate a value for skew and divide that by the standard error (same for kurtosis)

- If the score is between ± 2 (1.96 rounded) it is acceptable to treat your data as approximating a normal distribution
- Cite:
- Curran, Patrick J., Stephen G. West, and John F. Finch. "The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis." *Psychological methods* 1.1 (1996): 16.

Our example



Check the standardized scores for the variable

Accepted Heuristic TWO

Convert the raw score for TPCOISS to a standardised score

If 95% of our data falls within ± 1.96 then we can treat the data as normal (using the empirical rule):

```
ztpcoiss<- abs(scale(survey$tpcoiss))
```

- will convert of raw scores to absolute value of z scores

```
FSA::perc(as.numeric(ztpcoiss), 1.96, "gt")
```

- Will calculate the percentage that are greater than 1.96
- Cite:
- Field, A., Field Z. and Miles J. (2012). Discovering Statistics Using R.

Check the standardized scores for the variable

Accepted Heuristic THREE

If more than 5% of our data falls outside of ± 1.96

- The size of the dataset has an impact
- If the sample size is larger than 80 cases, a case is an outlier if its standard score is ± 3.29 or beyond
- Therefore if 99% of our data falls between ± 3.29 we can treat our data as approximately normal
- Cite: Field, A., Field Z. and Miles J. (2012). Discovering Statistics Using R.

What does
this mean for
our data?

For TPCOISS

- 18 values fall outside ± 1.96 (ignoring missing data)
- $18/431 = 4.18\%$ of our data
- Since the data is larger than 80 cases we can use ± 3.29 as our measure
 - 0.46% of our data
- It is therefore acceptable to treat as approximately normal
- We will report mean and standard deviation as well as describing our decision making

Reporting

Total Perception of Control on Internal States scores were assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 1 and Figure 2) identified some issues with skewness and kurtosis. While the standardized score for kurtosis (1.09) could be considered acceptable using the criteria proposed by Curran, West and Finch (1996), but the standardized score for skewness (-3.4) was outside the acceptable range. However as over 99% of standardized scores fall within the bounds of ± 3.29 , using the guidance of Field, Miles and Field (2013), the data can be considered to approximate a normal distribution ($m=60.63$, $sd=11.99$, $n=430$).

Repeat for TPStress

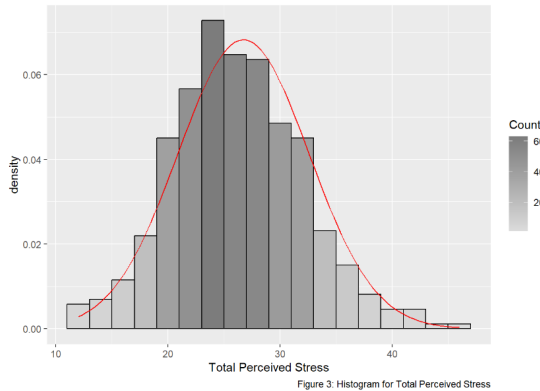


Figure 3: Histogram for Total Perceived Stress

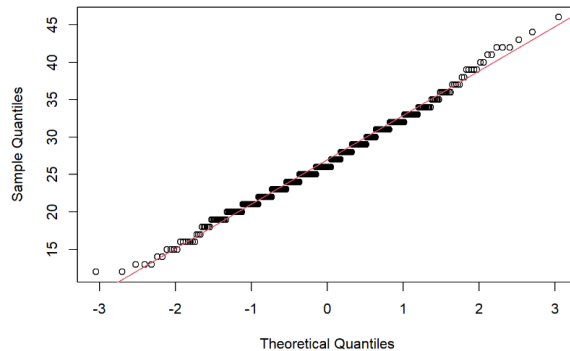
Standardised score for skew : -3.396456

Standardised score for kurtosis: 1.088118

6% outside +/- 1.96

0.23% outside +/- 3.29

Figure 4: QQPlot for Total Perceived Stress



Reporting

Total Perceived Stress were assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 3 and Figure 4) identified some issues with skewness and kurtosis. While the standardized score for kurtosis (0.77) could be considered acceptable using the criteria proposed by Curran, West and Finch (1996), but the standardized score for skewness (2.08) was outside the acceptable range. However as over 99% of standardized scores fall within the bounds of ± 3.29 , using the guidance of Field, Miles and Field (2012), the data can be considered to approximate a normal distribution ($m=26.73$, $sd=5.85$, $n=433$).

Correlation example

Scatterplot

Total PCOISS and Total Perceived Stress

```
```{r, fig.cap="Figure 5: Scatterplot of Correlation between Total  
PCOISS and Total Perceived Stress"}
#Create a simple scatterplot of feeling of control and perceived stress
#aes(x,y)
scatter <- ggplot(survey, aes(tpcoiss, tpstress))
scatter + geom_point() +
 geom_smooth(method = "lm", colour = "Red", se = F) + labs(x = "Total
PCOISS", y = "Total Perceived Stress")
```
```

Total PCOISS and Total Perceived Stress

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 13 rows containing non-finite outside the scale range  
(`stat_smooth()`).
```

```
Warning: Removed 13 rows containing missing values or values outside the scale range  
(`geom_point()`).
```

This is fine.

There is no point including rows where you have a value for only one variable or are missing values for both.

Note: You need to find out why these are missing.

Total PCOISS and Total Perceived Stress

As perceived control increases, stress decreases

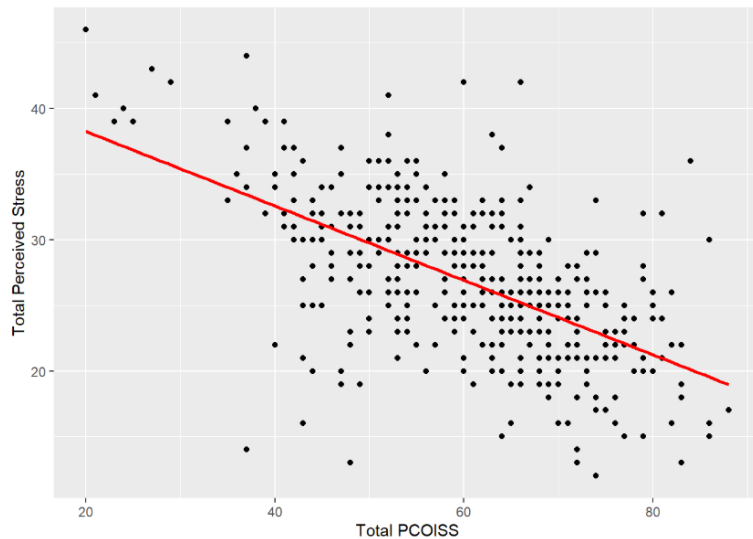


Figure 5: Scatterplot of Correlation between Total PCOISS and Total Perceived Stress

Doing a Correlation in R

Total PCOISS, Total Perceived Stress

```
#Pearson Correlation  
cor.test(survey$tpcoiss, survey$tpstress, method='pearson')
```

Doing a Correlation in R

Total PCOISS, Total Perceived Stress

```
##  
## Pearson's product-moment correlation  
##  
## data: survey$tpcoiss and survey$tpstress  
## t = -14.683, df = 424, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.6402679 -0.5139141  
## sample estimates:  
## cor  
## -0.5805759
```

df – number of independent pieces of information considered in calculating r (report as n)

Pearson's correlation co-efficient is the statistic
Call it r in your reporting

P-value – indicates whether you have a statistically significant result or not

Note 1: 2.2e-16 is 2.2 * e to the power of -16 (very small number)

Note 2: You should round your co-efficient to 2 or three decimal places e.g. -0.581

Doing a Correlation in R

Total PCOISS, Total Perceived Stress

```
##  
## Pearson's product-moment correlation  
##  
## data: survey$tpcoiss and survey$tpstress  
## t = -14.683, df = 424, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.6402679 -0.5139141  
## sample estimates:  
## cor  
## -0.5805759
```

Pearson's correlation co-efficient is the statistic
Call it r in your reporting

Note 1: 2.2e-16 is $2.2 * e$ to the power of -16 (very small number)

Note 2: You should round your co-efficient to 2 or three decimal places e.g. -0.581

Hang on – why is t being reported? Why is it included in the correlation results?

```
#Pearson Correlation  
cor.test(survey$tpcoiss, survey$tpstress, method='pearson')
```

```
##  
## Pearson's product-moment correlation  
##  
## data: survey$tpcoiss and survey$tpstress  
## t = -14.683, df = 424, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.6402679 -0.5139141  
## sample estimates:  
## cor  
## -0.5805759
```

To decide if the correlation is statistically significantly different from zero a t-test is conducted to compare with a group with a zero correlation

Things to know about the Correlation Co-efficient


It varies between -1 and +1

- 0 = no relationship

It is an effect size – how strong is the correlation

- (ignore sign for magnitude of effect)
- ± 0.1 = small/weak
- ± 0.3 = medium/moderate
- ± 0.5 = large/strong
- Cohen's effect size heuristic is standard

Find a book that is respected in your field that discusses this and cite it when stating you used Cohen's convention



Key
Slide

Our Example -Doing a Correlation

```
##  
## Pearson's product-moment correlation  
##  
## data: survey$tpcoiss and survey$tpstress  
## t = -14.683, df = 424, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.6402679 -0.5139141  
## sample estimates:  
## cor  
## -0.5805759
```

This significance value tells us that the probability of this correlation being due to random chance is very low (close to zero in fact).

Hence, we can have confidence that this relationship is genuine and not a chance result (if we have done everything else correctly e.g. representative sample, eliminating bias etc.)

Reporting a Pearson Correlation in words

The relationship between perceived sense of being in control (Total PCOISS derived from the PCOISS questionnaire) and perceived stress (Total Perceived Stress derived from the perceived stress questionnaire) was investigated using a Pearson correlation. A strong, statistically significant, negative correlation was found ($r = -0.58$, $n=424$, $p<.001$). This result suggests that individuals with higher levels perceived sense of control report lower levels perceived stress.

NOTE1:

- Because the significance is $< .000$ in test results, the convention is to report it as $<.001$

NOTE2:

- $N=424$ because it does not include missing values

NOTE3:

- In some cases where you are doing multiple correlation tests, you may report a set of coefficients in a table and discuss the implications in the text.

Reporting p values

If statistically significant (e.g., $p < 0.05$)

- report as: “ $p < .05$ ” or “ $p < .001$ ”
- Without the exact value.
- Common thresholds: 0.05, 0.01, 0.001

If not significant, report the exact value: “ $p = .27$ ”.

The logic:

- When significant, the exact decimal (like $p = .0000123$) isn't very informative — we just need to know it's below the threshold.
- When not significant, the exact value is useful to show how close it was (e.g., $p = .06$ is more informative than just “not significant”).

How to report when no statistically significant correlation was found

If you wish to report in text – this is not based on our dataset

“The relationship between perceived tension(derived from the PTT questionnaire) and perceived stress (derived from the perceived stress questionnaire) was investigated using a Pearson correlation. No statistically significant correlation was found ($r = -.08$, $n=424$, $p=.10$). This results provides no evidence against the null hypothesis of no relationship between perceived tension and perceived stress.

NOTE1:

Because the test is not statistically significant we report the actual p value.



Key
Slide

Things to know about the Correlation Co-efficient

Significance of all co-efficients and covariance depends on the p-value (significance value of the test)

Covariance

Variance tells us by how much scores deviate from the mean for a single variable.

Covariance = Scaled version of variance

- Calculate the error between the mean and each observations score for the first variable (x).
- Calculate the error between the mean and their score for the second variable (y).
- Multiply these error values.
- Add these values and you get the cross product deviations.
- The covariance is the average cross-product deviations

r^2

Shared variance

Coefficient of determination, r^2

- By squaring the value of r you get the proportion of variance in one variable shared by the other.

For our example r^2 for $-.580 \times -.580 = .3364$

This means that time Total PCOISS and Total Perceived stress share 33.64% of their variance

- Always round up to 2 decimal places
- You can report the variance if it is relevant to your domain.
- In some cases, you may report a set of coefficients in a table and discuss the variance in the text.

Covariance

It depends upon the units of measurement.

- E.g. The Covariance of two variables measured in Miles might be 4.25, but if the same scores are converted to Km, the Covariance is 11.

One solution: standardise it!

- Divide by the standard deviations of both variables.
- Create standardised scores
- Analyse -> Descriptive Statistics -> Frequencies (check Save as Standardised Scores)

The standardised version of Covariance is known as the **Correlation coefficient**.

Things to know about the Correlation Co-efficient

Coefficient of determination, r^2

- By squaring the value of r you get the proportion of variance in one variable shared by the other.
- You can report this if it is relevant to your research.
- In some cases you may report a set of coefficients in a table and discuss the variance in the text.

What does this mean?

- In our example $r = -0.58$ so $r^2 = 0.3364 = 33.64\%$
- What does this mean?
 - Our concepts have 33.64% of their variation in common

Statistical v Practical Significance

A small correlation co-efficient can reach statistical significance.

This doesn't mean anything practically.

Need to consider both the co-efficient and the amount of shared variance (squaring the co-efficient) .

- E.g. a coefficient of 0.2 explains 4% of the shared variance

Need also to consider other research into the area and compare your findings with those

- Even though your research explains only a small amount of the variance it may be more than others have found (or less).

Sample size and correlation

In small samples (e.g. $n=30$) you may have moderate correlation that does not reach statistical significance

In larger samples ($n=100+$) small correlations may reach statistical significance

You need to report statistical significance but also the strength of the relationship and the amount of shared variance.

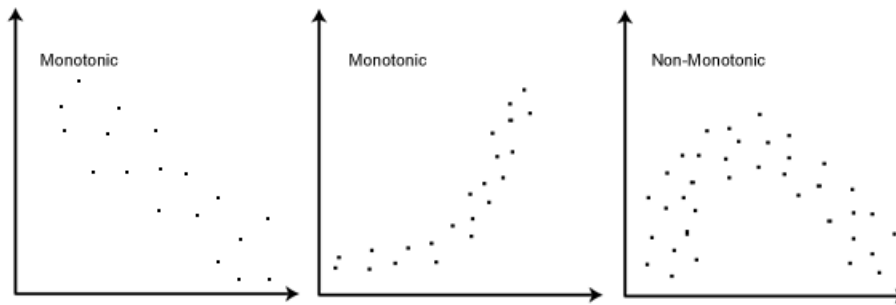
Non-parametric Tests: Spearman or Kendall

Doesn't require normality

Requires independent observations

Use when assumptions of Pearson are violated or when data is not scale

Spearman - Requires a monotonic relationship – as one variable increases, so does the other or as one increases the other decreases



Spearman Correlation in R

```
#Spearman Correlation
#Change the method to be spearman.
#This test will give an error since this method uses ranking but cannot handle ties
cor.test(survey$tpcoiss, survey$tpstress, method = "spearman")
```

```
## Warning in cor.test.default(survey$tpcoiss, survey$tpstress, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
## Spearman's rank correlation rho
##
## data: survey$tpcoiss and survey$tpstress
## S = 20044000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## -0.5556353
```

Spearman's Rho is the statistic

Spearman Correlation in R

```
#We can also use kendall's tau which does handle ties  
cor.test(survey$tpcoiss, survey$tpstress, method = "kendall")
```

```
##  
## Kendall's rank correlation tau  
##  
## data: survey$tpcoiss and survey$tpstress  
## z = -12.362, p-value < 2.2e-16  
## alternative hypothesis: true tau is not equal to 0  
## sample estimates:  
##      tau  
## -0.4150866
```

Kendall's Tau is the statistic

Reporting Spearman and Kendall in words

As you would for Pearson except make sure you cite the correct test and give the correct statistic

- Spearman
 - Spearman's rho
 - or r_s
 - or the Greek letter ρ
- Kendall
 - Kendall's tau
 - Kendall's tau-b
 - τ_b

Spearman V Kendall

Spearman

- More widely reported – check which your discipline prefers
- Based on ranked data
- More sensitive to error and discrepancies in data
- Easier to calculate by hand

Kendall

- More robust
- Based on concordant and discordant pairs (consistent and inconsistent)

Generally, both lead to same inferences