Correlation

Probability and Statistical Inference

Bojan Božić

TU Dublin

Winter 2020

Before Inferential Statistics . . .

- We need to establish evidence to support going ahead with building a predictive model.
- If we are asserting a relationship:
    - We need to investigate if there is any evidence of a relationship using the appropriate test and make a decision based on the results (strength, direction etc.).
- If we are asserting a differential effect for different groups:
    - We need to investigate if there is any difference using the appropriate test and make a decision based on the result.

The General Linear Statistical Model

- Concepts of interest (measured by their variables) are hypothesised to be related to each other in some way.
- Goal: summarise/describe accurately what's happening in the data.
- Easiest way to think about it is for bivariate data:
  - Looking to model the pattern in ordered pairs where each member is the value of one of the variables of interest.
  - Our statistical model takes the form of an equation:
    - $Outcome_i = model + error_i$ (i = i^{th} case)
    - Outcome in the data we observed = model we built + an error
  - In the linear model.
    - Trying to fit a line to it to model the pattern.
    - Using the equation of a line $y = b_0 + b_1 x + e$.
    - Where y is our dependent (or outcome) variable and x is the independent (or predictor).
    - b_0 is the intercept (value of y when x is 0)
    - e is the error term - the degree to which the line is in error in describing each data point.

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.

Scatterplots

- When to Use?
  - Bivariate numerical data (two variables).
  - Plot the relationship between two variables:
    - One independent, one dependent.
  - Collection of ordered pairs.
- How to costruct?
  - 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.

Example

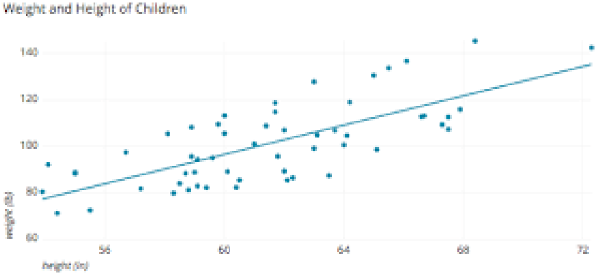
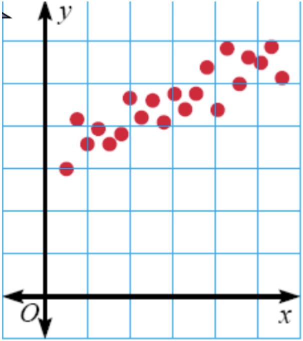


Figure: Example of a scatterplot.

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.



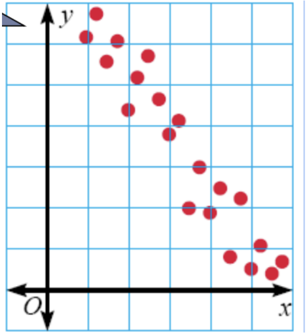
Positive Correlation

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.



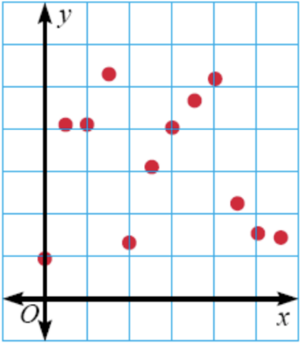
Negative Correlation

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.



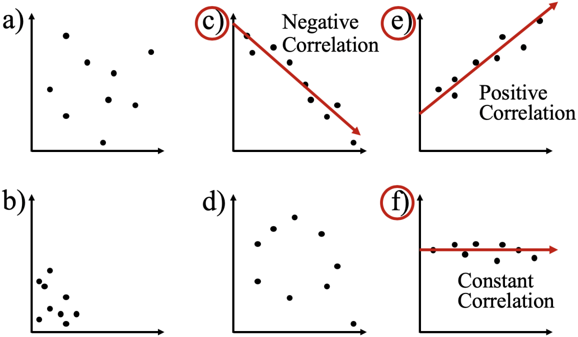
No Correlation

If you look at colour of boots a premier league footballer wears and their scoring average, you will find that there is no correlation between the two.

Scatterplot

Examples

Which scatterplots below show a linear trend?



Example

Suppose we found the age and weight for each person in a sample of 10 adults. Is there any relationship between the age and weight of these adults?

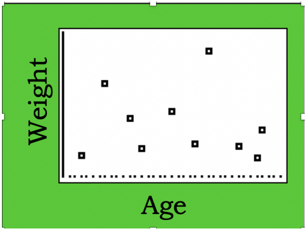
Create a scatterplot of the data below.

Age	24	30	41	28	50	46	49	35	20	39
Wt	256	124	320	185	158	129	103	196	110	130

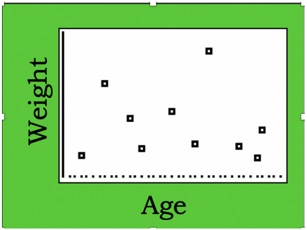
Example

Suppose we found the age and weight for each person in a sample of 10 adults. Is there any relationship between the age and weight of these adults?

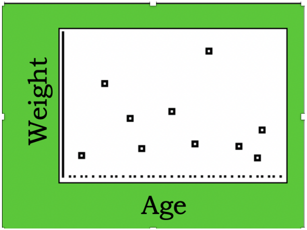
Create a scatterplot of the data below.



Example

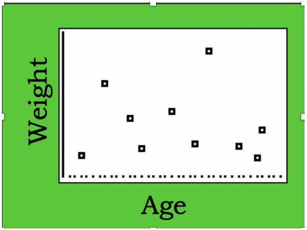


Example



- Do you think there is a relationship? If so, what kind? If not, why not?

Example



- Do you think there is a relationship? If so, what kind? If not, why not?
- There does not appear to be a relationship between age and weight in adults.

Example

Suppose we found the age and weight for each person in a sample of 10 adults. Is there any relationship between the age and weight of these adults?

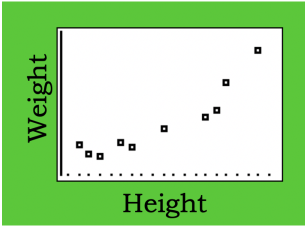
Create a scatterplot of the data below.

Ht	74	65	77	72	68	60	62	73	61	64
Wt	256	124	320	185	158	129	103	196	110	130

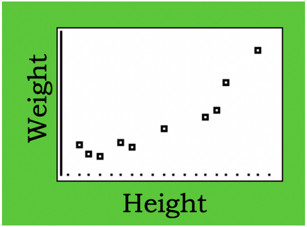
Example

Suppose we found the age and weight for each person in a sample of 10 adults. Is there any relationship between the age and weight of these adults?

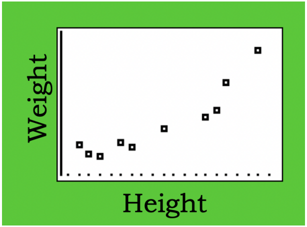
Create a scatterplot of the data below.



Example

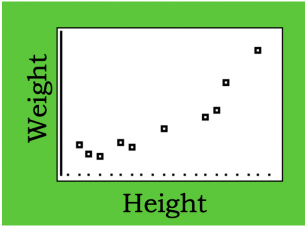


Example



- Do you think there is a relationship? If so, what kind? If not, why not?

Example



- Do you think there is a relationship? If so, what kind? If not, why not?
- Is it positive or negative? Weak or strong?

Correlation

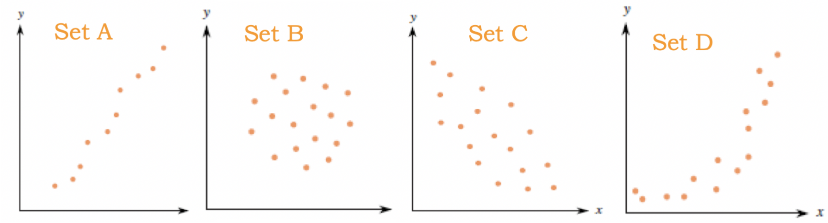
The relationship between bivariate numerical variables:

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

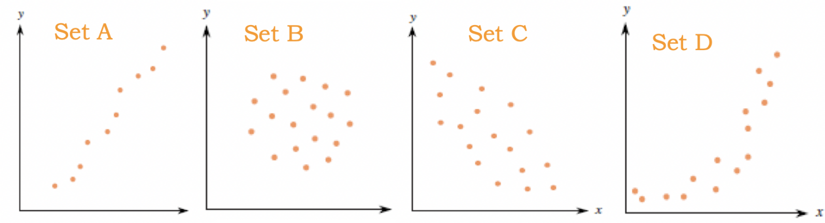
Questions

- What does it mean if the relationship is positive or negative?
- What feature(s) of the graph would indicate a weak or strong relationship?

Identify the strength and direction of the following data sets.

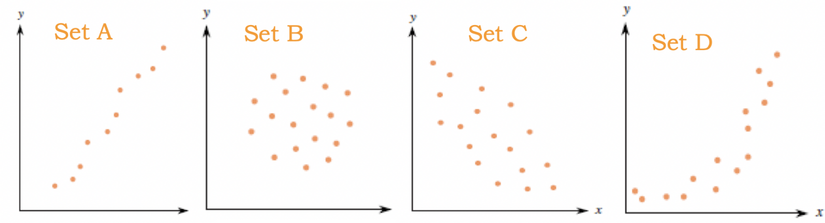


Identify the strength and direction of the following data sets.



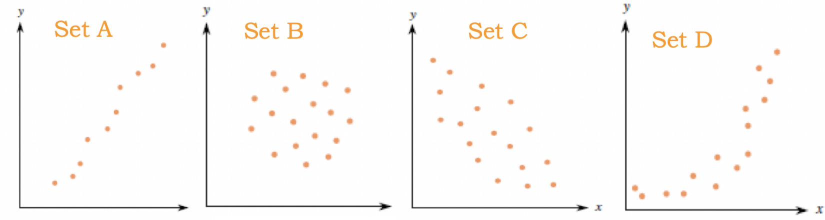
- Set A shows a strong, positive linear relationship.

Identify the strength and direction of the following data sets.



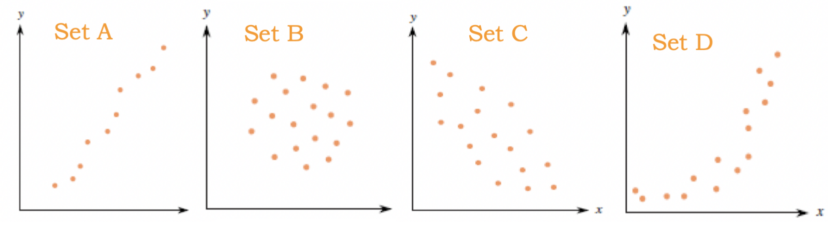
- Set A shows a strong, positive linear relationship.
- Set B shows little or no relationship.

Identify the strength and direction of the following data sets.



- Set A shows a strong, positive linear relationship.
- Set B shows little or no relationship.
- Set C shows a weaker (moderate), negative linear relationship.

Identify the strength and direction of the following data sets.



- Set A shows a strong, positive linear relationship.
- Set B shows little or no relationship.
- Set C shows a weaker (moderate), negative linear relationship.
- Set D shows a strong, positive curved relationship.

Simple Example

- Suppose we want to look at children's maths scores at age 16 and their achievement of a standard maths test at age 7.
- We are interested to see if the score a child achieves at age 7 is related to the score they achieve at age 16.
- And if so what is the direction and strength of the relationship.

Simple Example

$$y = a + bx + e \qquad (1)$$

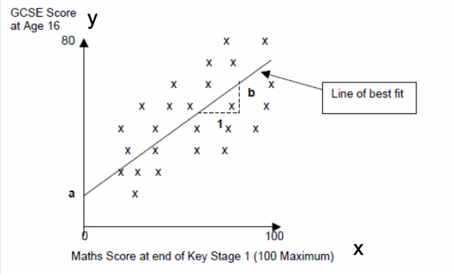


Figure: Relationship between Maths score age 7 with GCSE result at age 16 (for 25 students)

- y is the outcome variable (in this case GCSE Score) - the dependent variable.

Simple Example

$$y = a + bx + e \qquad (1)$$

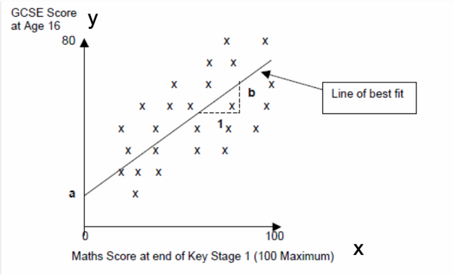


Figure: Relationship between Maths score age 7 with GCSE result at age 16 (for 25 students)

- y is the outcome variable (in this case GCSE Score) - the dependent variable.
- x is the independent variable (in this case Maths Score aged 7).

Simple Example

$$y = a + bx + e \qquad (1)$$

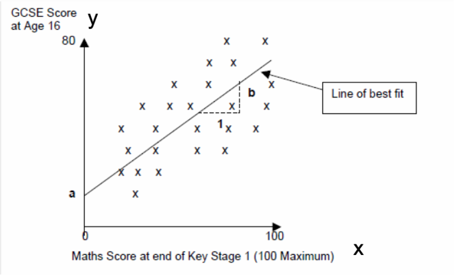


Figure: Relationship between Maths score age 7 with GCSE result at age 16 (for 25 students)

- y is the outcome variable (in this case GCSE Score) - the dependent variable.
- x is the independent variable (in this case Maths Score aged 7).
- **a** is the intercept or the point where the line crosses the y-axis (y value when x=0).

Simple Example

$$y = a + bx + e \tag{2}$$

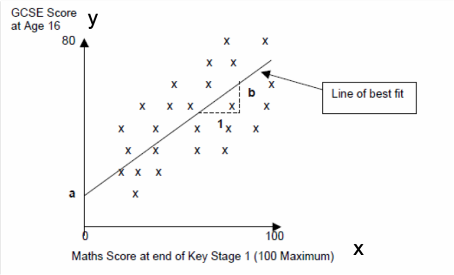


Figure: Relationship between Maths score age 7 with GCSE result at age 16 (for 25 students)

- **b** is the gradient of the line. Represents the amount that the outcome variable changes for one unit change in the independent variable (e.g. for every one percentage point increase in a child's Maths Test score, the line suggests that the child's GCSE Score increases by **b** points).

Simple Example

$$y = a + bx + e \tag{3}$$

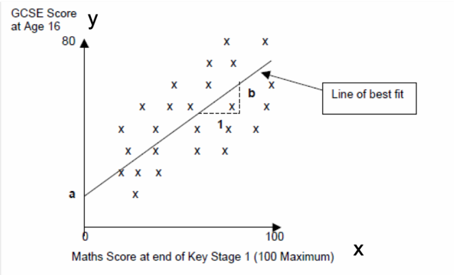


Figure: Relationship between Maths score age 7 with GCSE result at age 16 (for 25 students)

- If the line did completely model the data then all of the points would rest exactly on the line. **e** is basically the vertical distance between each point and the line itself. Model is a best fit. **e** obviously varies. Without any further information about confounding variables we cannot explain this variation - so we include it as an error.

Simple Example

$$y = a + bx \qquad (4)$$

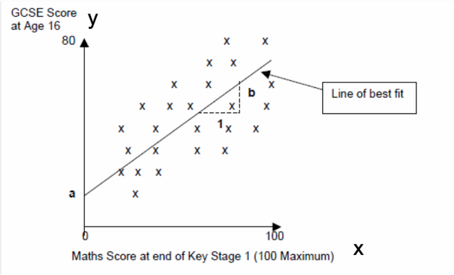


Figure: Relationship between Maths score age 7 with GCSE result at age 16 (for 25 students)

- 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.

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?

Parametric vs 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.

What is 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.
  - Pearson Correlation (Parametric).
  - Spearman Rank Order Correlation (Non-Parametric).

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**.
- And 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.
- We can look at a bi-variate correlation or a partial correlation.
    - Bi-variate - two variables.
    - Partial - two variables while controlling for another.

Linear Correlation

- The extent to which two variables have a straight line 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.

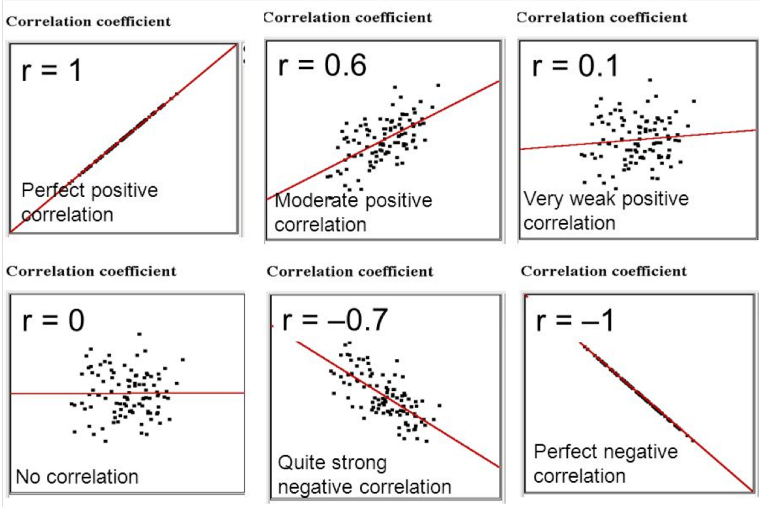
Modeling Relationships

- First, look at some scatterplots of the variables that have been measured.

$$Outcome_i = (model) + error_i \tag{5}$$

$$Outcome_i = (bX_i) + error_i \tag{6}$$

Strength of relationships



Assumptions Parametric Correlation

- Pearson's Correlation
- Level of measurement
  - Interval or ratio
  - 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.
- Spearman's Rho can be used for ordinal or ranked data

Assumptions Parametric Correlation

- 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.

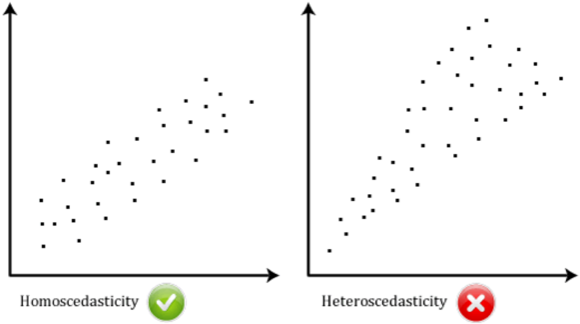
Assumptions for Parametric Correlation

- 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.

Homoscedasticity

- 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

Homoscedasticity



Getting Started With Analysis

- Inspect your data.
- Generate your descriptive statistics.
- Generate your visuals (graphs).
- Make decisions about normality.
    - Choose the correct tests.

Example Pearson Correlation

For the datset survey.dat:

- Dataset created from a survey designed to explore the factors that impact on respondents' psychological adjustment and wellbeing.
- To load the data in R the command is (as usual change the location to reflect where you have the file saved:

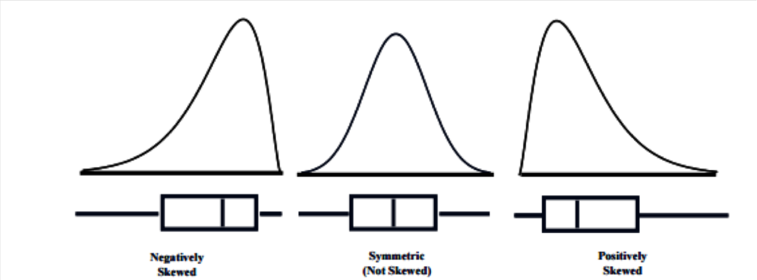
```
survey <- read.table("survey.dat")
```

- PSIWeek4-Lecture.rmd contains all the commands.

Example Pearson Correlation

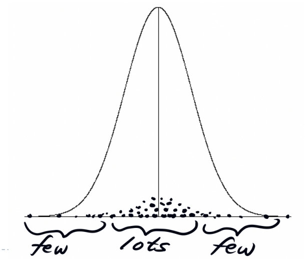
- Question:
  - Is there a relationship between respondents' feelings of control and their level of perceived stress?
- Variables:
  - Feeing of control.
    - **tpcoiss**
    - Total PCOISS derived from the PCOISS questionnaire (tpcoiss).
    - **Dependent variable**
  - Total Perceived Stress
    - **tpstress**
    - Derived from the perceived stress questionnaire.
    - **Independent variable**

Plots and Histograms



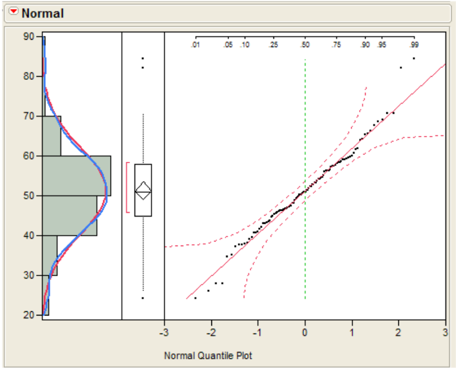
Normal Quantile Plot

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 spacing similar to what is shown 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.

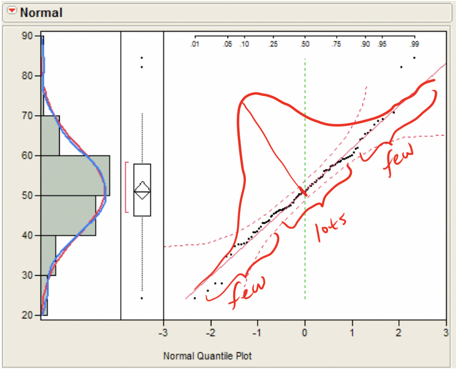
Normal Quantile Plot



THE IDEAL PLOT:

Here is an example where the data is perfectly normal. The plot on right is a normal quantile plot with the data on the vertical axis 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. The points should lie within the dashed lines.

Normal Quantile Plot



THE IDEAL PLOT:

Here is an example where the data is perfectly normal. The plot on right is a normal quantile plot with the data on the vertical axis 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. The points should lie within the dashed lines.

How do I inspect a scale variable?

1. Generate summary statistics.
2. Make sure you include skewness and kurtosis.
3. Generate a histogram with a normal curve showing.
4. Generate a Q-Q plot.
5. Review your statistics and plots to see how far away from normal your data is.

How do I inspect a scale variable in R?

- Use packages pastecs, ggplot2 and semTools:
  - Install them using the install.packages e.g. `install.packages("pastecs")`
  - Thereafter tell R you want to use them e.g. `library(pastecs)`
- Statistical summary:
  - Use `stat.desc` function form.
  - Will give you a set of summary statistics for a variable. `stat.desc(varname, basic=F)`
- Skewness and kurtosis with standard error Use package semTools:
  - `library(semTools)`
  - `skew(varname)`
  - `kurtosis(varname)`
  - (we need the standard error to facilitate creating standardised score of skewness and kurtosis)

How do I inspect a scale variable in R?

- Create a histogram:
  - `library(ggplot2)`
  - `g <- ggplot(survey, aes(x=varname))`
  - `gg <- gg + labs(x="The Label of the X Axis")`
  - `gg <- gg + geom_histogram(binwidth=2, colour="black", aes(y=..density.., fill=..count..))`
  - `gg <- gg + scale_fill_gradient("Count", low="#DCDCDC", high="#7C7C7C")`
- qqnorm will create a qqplot:
  - `qqnorm(varname)`
  - `qqline(varname, col=2)`

Tests of Normality

- As with all estimates we are unlikely to ever see the values of zero in either skewness or kurtosis statistics for the standardised scores of the variable of interest.
- The real question is whether the given estimates vary significantly from zero.
- We need to look at the standard error of skewness and kurtosis.
- What we are looking for is whether the value of 'zero' is within the 95% confidence interval.

CI

How to calculate?

- Need to know the limits within which 95% of the means fall.
- Go back to the normal distribution - 95% of scores fall between $+/$-1.96.
- Once we know the mean and standard deviation we can calculate any score and therefore the CI.

Skewness and Kurtosis

- Standardised scores for skewness and kurtosis between -2 and +2 are considered acceptable in order to prove normal univariate distribution.
- Within R use library semTools.
  - `skew(survey$tpcoiss)`
  - `kurtosis(survey$tpcoiss)`
  - Be careful when loading libraries, some have functions with the same names and later loads will override others.

Tests of Normality

- Standardised scores (value/std.error) for skewness between -2 and +2 are considered acceptable in order to prove normal univariate distribution.
- Tpcois:
  - Standardised Skew = -.401/.118=-3.40
  - Standardised Kurtosis=.257/.236=1.08
- Skewness is not acceptable so we need to look into this further.
- Need to look at the outliers, how many of them there are or whether we can transform it to become more normal.

So our data has failed the standardised skew

- Does this mean we can't use parametric tests?
- No.
- We can do some additional checks.
- First create a histogram to see how much skew there is.

So our data fails the standardised skew

- 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.
- `sort (scale(survey$tpcoiss))`: will sort a list in ascending order.

Deciding Normality

- Check your Q-Q Plot.
- Check skewness and kurtosis standardised scores.
- Check impact of outliers.
  - At 0.05 level if 95% of your data is within +/- 1.96 when converted to standardised scores – it is likely your data is safe to treat as normal.
  - If the sample size is small (80 or fewer cases), a case is an outlier if its standard score is +/-2.5 or beyond.
  - If the sample size is larger than 80 cases, a case is an outlier if its standard score is +/-3.29 or beyond.

So our data fails the standardised skew

For tpcoiss:

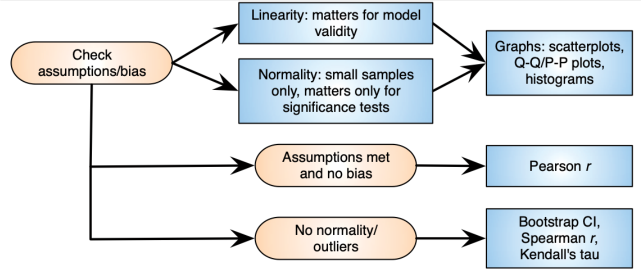
- 23 values fall outside +/- 1.96 (including missing data).
- 23/439=5.2% of our data.
- 10/439=2% of our data (if we exclude missing data).
- Since the data is larger than 80 cases we can use +/- 3.29 as our measure.
- 11/439=2.5% (including missing data).
- 2/439=0.04% (excluding missing data).
- So it is ok to treat as normal.

Example Pearson Correlation

- Look at the distribution of both variables.
- Create a scatterplot.
    - Look at outliers.
    - Look at distribution of the data points.
- Run the correlation.
- Interpret the output.
    - Check the information you have been given about the sample.
    - Determine the direction of the relationship.
    - Determine the strength of the relationship.
    - Calculate the coefficient of determination.
    - Assess the significance level.

Conducting Correlation Analysis

FIGURE 7.5 The general process for conducting correlation analysis



Check assumptions/bias

Linearity: matters for model validity

Normality: small samples only, matters only for significance tests

Graphs: scatterplots, Q-Q/P-P plots, histograms

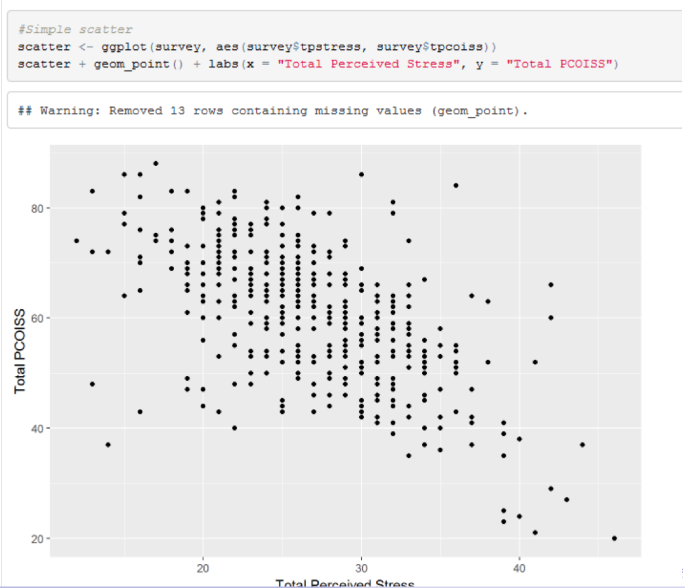
Assumptions met and no bias

Pearson *r*

No normality/ outliers

Bootstrap CI, Spearman *r*, Kendall's tau

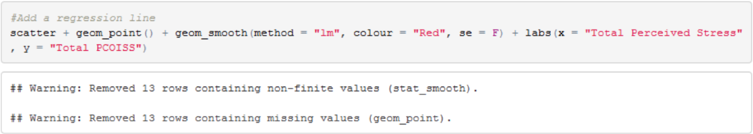
Total PCOISS and Total Perceived Stress



```
#Simple scatter
scatter <- ggplot(survey, aes(survey$tpstress, survey$tpcoiss))
scatter + geom_point() + labs(x = "Total Perceived Stress", y = "Total PCOISS")
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```

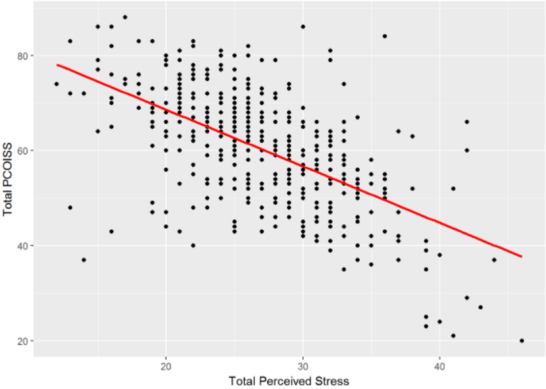
Total PCOISS and Total Perceived Stress



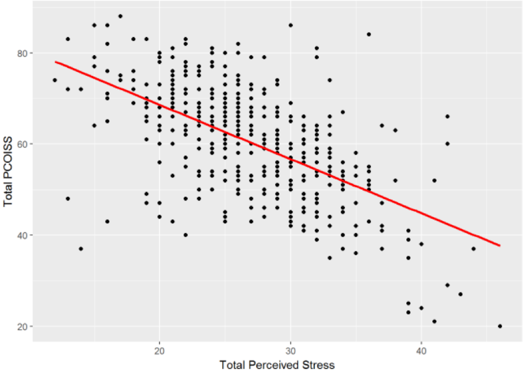
```
#Add a regression line
scatter + geom_point() + geom_smooth(method = "lm", colour = "Red", se = F) + labs(x = "Total Perceived Stress"
, y = "Total PCOISS")
```

```
## Warning: Removed 13 rows containing non-finite values (stat_smooth).

## Warning: Removed 13 rows containing missing values (geom_point).
```



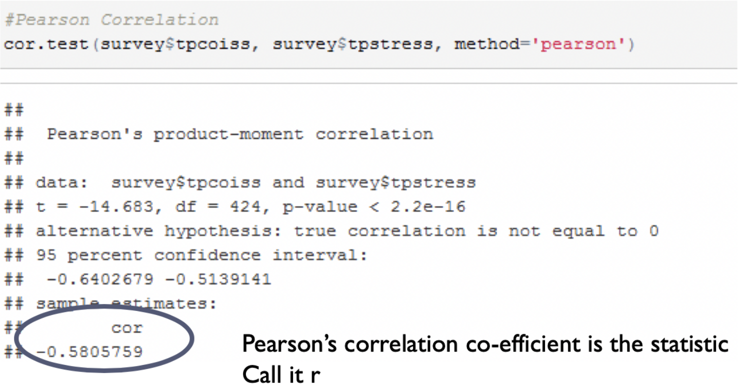
Total PCOISS and Total Perceived Stress



There appears to be negative correlation.

As stress increases, perceived control decreases.

Doing an Correlation in R Total PCOISS, Total Perceived Stress



```
#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
```

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

- **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.

Things to know about the Correlation Co-efficient

- It varies between -1 and +1.
    - 0 = no relationship.
- It is an effect size (ignore sign for magnitude of effect).
    - +/-.1 = small/weak.
    - +/-.3 = medium/moderate.
    - +/-.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.

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.
- You can report this 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.
- For our example r^2 for -.580*-.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.

Reporting a Pearson Correlation in words

"The relationship between Total PCOISS (derived from the PCOISS questionnaire) and Total Perceived Stress (derived from the perceived stress questionnaire) was investigated using a Pearson correlation. A strong negative correlation was found ($r = -.580, n = 424, p < .001$)."

- **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.

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

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

What is a test of hypotheses?

A test of hypotheses is a method that uses sample data to decide between two competing claims (hypotheses) about the population characteristic. Is the value of the sample statistic . . .

- a random occurrence due to natural variation? OR
- a value that would be considered surprising?

Hypothesis Testing

- Goal : Make statement(s) regarding unknown population parameter values based on sample data.
- : Elements of a hypothesis test:
  - Null hypothesis (H_0)
    - Statement regarding the value(s) of unknown parameter(s).
    - Typically will imply no association between independent and dependent variables in our theory (will always contain an equality).
  - Alternative hypothesis (H_a)
    - Statement contradictory to the null hypothesis (will always contain an inequality).
    - Collect data and seek evidence against H_0 as a way of bolstering H_a (deduction).
  - Test statistic
    - Quantity based on sample data and null hypothesis which allows you to determine between null and alternative hypotheses.

Hypothesis Testing

- Hypothesis may concern an effect (e.g. correlation) in the population or a difference between groups in a population.
- The general goal of a hypothesis test is to rule out chance (sampling error) as a plausible explanation for the results from a research study.
- All hypothesis testing starts with the null hypothesis : that there is no effect or difference in the population.

Hypothesis Statements

Null Hypothesis

The **null hypothesis**, denoted by H_0, is a claim about a **population** characteristic that is initially assumed to be true.

Alternative Hypothesis

The alternative hypothesis, denoted by H_a, is the competing claim.

You are usually trying to determine if this claim is believable.

Statements

- The hypothesis statements are ALWAYS about the population – NEVER about a sample!
- To determine what the alternative hypothesis should be, you need to keep the research objectives in mind.

The Form of Hypotheses

Null Hypothesis:

- H_0: population characteristic $=$ hypothesised value.

The Form of Hypotheses

Null Hypothesis:

- H_0: population characteristic $=$ hypothesised value.
- The Null Hypothesis always includes the equal case!

The Form of Hypotheses

Null Hypothesis:

- H_0: population characteristic $=$ hypothesised value.
- The Null Hypothesis always includes the equal case!
- This hypothesised value is a specific number determined by the context of the problem.

Alternative Hypothesis:

- H_a: population characteristic $>$ hypothesised value.

The Form of Hypotheses

Null Hypothesis:

- H_0: population characteristic $=$ hypothesised value.
- The Null Hypothesis always includes the equal case!
- This hypothesised value is a specific number determined by the context of the problem.

Alternative Hypothesis:

- H_a: population characteristic $>$ hypothesised value.
- Notice that the alternative hypothesis uses the same population characteristic and the same hypothesised value as the null hypothesis.

The Form of Hypotheses

Null Hypothesis:

- H_0: population characteristic $=$ hypothesised value.
- The Null Hypothesis always includes the equal case!
- This hypothesised value is a specific number determined by the context of the problem.

Alternative Hypothesis:

- H_a: population characteristic $>$ hypothesised value.
- Notice that the alternative hypothesis uses the same population characteristic and the same hypothesised value as the null hypothesis.
- The sign is determined by the context of the problem.

The Form of Hypotheses

Null Hypothesis:

- H_0: population characteristic $=$ hypothesised value.
- The Null Hypothesis always includes the equal case!
- This hypothesised value is a specific number determined by the context of the problem.

Alternative Hypothesis:

- H_a: population characteristic $>$ hypothesised value.
- Notice that the alternative hypothesis uses the same population characteristic and the same hypothesised value as the null hypothesis.
- The sign is determined by the context of the problem.
- H_a: population characteristic $<$ hypothesised value.
- H_a: population characteristic $\neq$ hypothesised value.

The Form of Hypotheses

Null Hypothesis:

- H_0: population characteristic $=$ hypothesised value.
- The Null Hypothesis always includes the equal case!
- This hypothesised value is a specific number determined by the context of the problem.

Alternative Hypothesis:

- H_a: population characteristic $>$ hypothesised value.
- Notice that the alternative hypothesis uses the same population characteristic and the same hypothesised value as the null hypothesis.
- The sign is determined by the context of the problem.
- H_a: population characteristic $<$ hypothesised value.
- H_a: population characteristic $\neq$ hypothesised value.
- These are considered one-tailed tests, becuase you're interested in only one direction.
- This is considered a two-tailed test, because you're interested

Our Example

- H_0: There is no relationship between Total PCOISS and Total Perceived Stress.
- H_a: There is a relationship between Total PCOISS and Total Perceived Stress.
  - Two-tailed hypothesis.

Hypothesis Test

When you perform a hypothesis test you make a decision:

reject H0 or fail to reject H0

When you make one of these decisions, there is a possibility that you could be wrong! That you made an error!

Each could possibly be a wrong decision; therefore, there are **two types of errors**.

Type I Error

- The error of **rejecting** H_0 when H_0 is **true**.
- The probability of a Type I error is denoted by α. α is called the **significance level** of the test.

Type II Error

- The error of **failing** to reject H_0 when H_0 is **false**.
- The probability of a Type II error is denoted by β.

Here is another way to look at the types of errors

	H_0 is true	H_0 is false
Reject H_0	Type I Error	Correct
Fail to reject H_0	Correct	Type II Error

Here is another way to look at the types of errors

	H_0 is true	H_0 is false
Reject H_0	Type I Error	Correct
Fail to reject H_0	Correct	Type II Error

- Suppose H_0 is true and we fail to reject it, what type of decision wss made?

Here is another way to look at the types of errors

	H_0 is true	H_0 is false
Reject H_0	Type I Error	Correct
Fail to reject H_0	Correct	Type II Error

- Suppose H_0 is true and we fail to reject it, what type of decision wss made?
- Suppose H_0 is true and we reject it, what type of decision wss made?

Here is another way to look at the types of errors

	H_0 is true	H_0 is false
Reject H_0	Type I Error	Correct
Fail to reject H_0	Correct	Type II Error

- Suppose H_0 is true and we fail to reject it, what type of decision wss made?
- Suppose H_0 is true and we reject it, what type of decision wss made?
- Suppose H_0 is false and we reject it, what type of decision wss made?

Here is another way to look at the types of errors

	H_0 is true	H_0 is false
Reject H_0	Type I Error	Correct
Fail to reject H_0	Correct	Type II Error

- Suppose H_0 is true and we fail to reject it, what type of decision wss made?
- Suppose H_0 is true and we reject it, what type of decision wss made?
- Suppose H_0 is false and we reject it, what type of decision wss made?
- Suppose H_0 is false and we fail to reject it, what type of decision wss made?

P-value, α - statistical significance

- A probability measure of evidence about H_0.
- Put simply it is the probability, given the null hypothesis is true, that the results could have been obtained purely on the basis of chance alone.
- The probability (under presumption that H_0 true) the test statistic equals observed value or value even more extreme predicted by H_a.
- The **P-value** allows us to answer the question:
  - Do our sample results allow us to reject H_0 in favour of H_a?
  - If that probability (p-value) is small, it suggests the observed result cannot be easily explained by chance.

Statistical Significance

- Working with random samples can never have 100% certainty that findings we derive from the sample will reflect real differences in the population as a whole.
- Convention is that (for your field of study) there is an accepted level of probability such that it is considered so small that the finding from your sample is unlikely to have occurred by chance or sampling error.
  - Normally, that line is drawn at $p = 0.05$ or $p = 0.01$.
  - In other words, when a statistical test tells us that the finding has less than a 5% or 1% chance of occurring due to sampling error then we tend to conclude that we can be sufficiently confident that the finding is therefore likely to reflect a 'real' characteristic of the population as a whole.
  - When this occurs, you can say that your finding is **statistically significant**.

Statistical Significance

A range of statistical tests can be used:

- Each will tell you how likely it is that a finding you get from your sample would occur simply by chance if no such difference actually existed in the population as a whole.
- ie. the probability that your finding is simply a fluke occurrence deriving from the random selection of your sample.

One-tailed and Two-tailed Tests

- P values are calculated for you are based upon a non-directional alternative hypothesis.
  - **Two-tailed tests.**
- In other words ...
  - While you state that a difference may exist between men and women, you don't state what that difference is i.e. is it likely to be more men than women or more women than men?

One-tailed and Two-tailed Tests

- If you can be more specific prior to analysing your data, usually on the basis of a theory you may have, and thus state a directional hypothesis then you can cut the probability calculated for you in half.
  - This is called a **one-tailed test**.
- Thus if your statistical test gives you a value of $p = 0.07$ then, if you have stated a directional hypothesis and your findings confirm this, then you can quote the probability as $p = 0.035$.
  - (which then may become statistically significant depending on the level you are working with).

Hypothesis Testing

- You start with the assumption (the null hypothesis H_0) that there are no differences or relationships in the population as a whole.
- You then state an alternative hypothesis (H_A) that there is a difference or a relationship.
- You select a sample and find a difference/relationship in it.
- You can then use a variety of tests for statistical significance to work out the probability of the difference/relationship you have found in your sample simply occurring by chance.

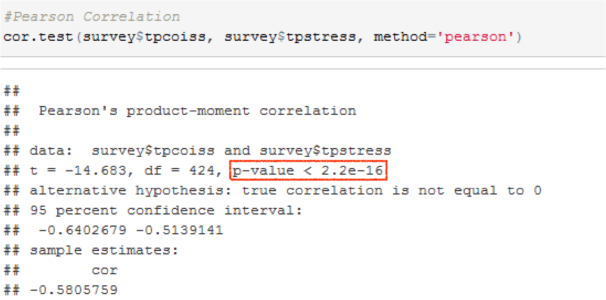
Hypothesis Testing

- Using the standard level accepted by your domain (e.g. $p \leq 0.05$ or $p \leq 0.01$).
- If the probability less than this value then you reject the null hypothesis and thus accept the alternative hypothesis and you can state that your findings are 'statistically significant'.
- If the probability is greater this value then you conclude that there is no evidence to reject the null hypothesis and your findings are not 'statistically significant'.
  - N.B This is different from concluding that you have evidence to accept the null hypothesis. In these cases, your findings are said to be 'not significant'.
- Caveat:
  - If we get a p-value of 0.051 should we accept the null hypothesis?
  - Should we reject the null hypothesis if we get a p-value of 0.049?
  - Need to allow for some flexibility in interpretation.

Accepting and Rejecting Hypotheses

- A non-statistically significant test result **does not** mean that the null hypothesis is true.
- A significant result **does not** mean that the null hypothesis is false.

Our Example - Doing a Correlation



```r
#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
```

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.

P-value

- If no decision needed, report and interpret P-value.
- If decision needed, select a cutoff point (such as 0.05 or 0.01) and reject H_0 if P-value $\leq$ that value.

Reporting the results

If you wish to report in text:

"The relationship between Total PCOISS (derived from the PCOISS questionnaire) and Total Perceived Stress (derived from the perceived stress questionnaire) was investigated using a Pearson correlation. A strong negative correlation was found ($r = -.58, n = 424, p < .001$)."

- 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.

Things to know about the Correlation Co-efficient

- It varies between -1 and $+1$
    - $0 =$ no relationship
- It is an effect size (ignore sign for magnitude of effect)
    - $+/$-.1 $=$ small/weak
    - $+/$-.3 $=$ medium/moderate
    - $+/$-.5 $=$ large/strong
    - Cohen's effect size heuristic is standard.

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 $r = -.58$ so $r^2 = 0.3364 = 33.64\%$
  - What does this mean?
  - Our concepts have 33.64% of their variation in common.

Things to know about the Correlation Co-efficient

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

Correlation and Causality

The third-variable problem:

- Causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results.

In our example:

- Perceived stress does relate significantly to perception of control.
- There may be other factors that are influencing perception of control.
- And some of these variables may not have been measured by the researcher.

Correlation and Causality

Direction of causality:

- Correlation coefficients say nothing about which variable causes the other to change.
- The correlation coefficient doesn't indicate in which direction causality operates.
- So, although it is intuitively appealing to conclude that perceived stress causes perceived control to change, there is no statistical reason why perceived control cannot cause perceived stress.
- This is where your knowledge of the constructs and existing theories in your area come in to play
  - You interpret your statistical findings through the lens of your subject area.
  - Why including relevant background and related research is important in your reporting.

Statistical vs Practical Significance

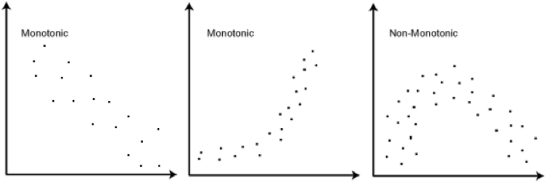
- A small correlation co-efficient can reach statistical significance.
- This doesn't mean anything practically.
- Need to consider both the co-efficient and the amount of shared variance (squaring the co-efficient).
  - E.g. a coefficient of 0.2 explains 4% of the shared variance.
- Need also to consider 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

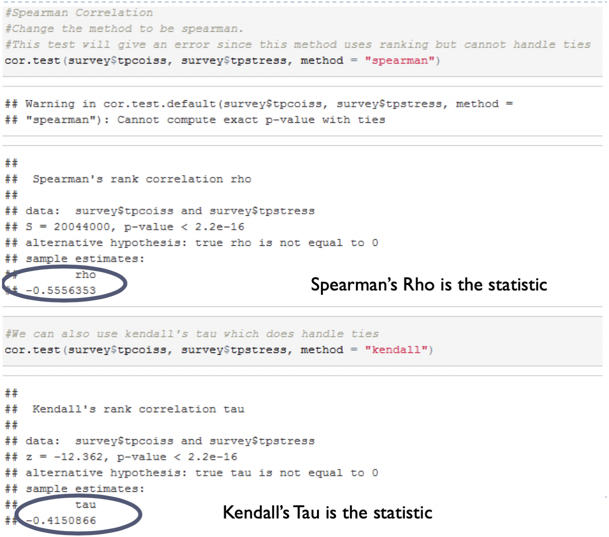
- In small samples (e.g. $n = 30$) you may have moderate correlation that does not reach statistical significance.
- In larger samples ($n = 100+$) small correlations may reach statistical significance.
- You need to report statistical significance but also the strength of the relationship and the amount of shared variance.

Spearman (Non-parametric)

- Doesn’t require normality.
- Requires independent observations.
- Use when assumptions of Pearson are violated or when data is not scale.
- Spearman - Requires a monontonic 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

```
#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

10min Break

Lab Setting

- Random groups in breakout rooms.
- Time: 40mins
- Random speaker shares screen and presents results in 2min discussion afterwards.
- Use slack for questions.

Lab Assignment

- Install and load discussed packages (*pastecs*, *ggplot2*, *semTools*).
- Read in the file *Regression.sav*.
- Inspect data (description: `https://rdrr.io/cran/mlmRev/man/Exam.html`) and discuss possible correlations.
- Create summary statistics for variables of interest.
- Find and evaluate skew and kurtosis.
- Plot a histogram with average as line.
- Create QQPlot with normality line.
- Run a correlation test and report the result as discussed.