Hypothesis Testing

Outline

- Introduction to Hypothesis Testing
  - t-test and p-value
  - Chi-squared text
- Comparing means
- Investigating relationships
  - Correlation
  - R-square
- Choosing the right test
- Non-parametric tests
  - Mann-Whitney test
  - ANOVA

Hypothesis testing

- An **objective** method of making decisions or **inferences** from sample data (evidence)
- Sample data used to choose between two choices i.e. **hypotheses** or statements about a population
- We typically do this by comparing what we have observed to what we expected if one of the statements (**Null Hypothesis**) was true

Hypothesis-1: There is a difference between vaccinated and not-vaccinated people, so I accept to get vaccinated

Hypothesis-2: There is no difference between vaccinated and not-vaccinated people, so I refuse to get vaccinated

Hypothesis testing Framework

- Always two hypotheses

H_A: Research (Alternative) Hypothesis

- What we aim to gather evidence of
- Typically that there **is** a difference/effect/relationship etc.

H_0: Null Hypothesis

- What we assume is true to begin with
- Typically that there is **no** difference/effect/relationship etc.

The Court

- Members of a jury have to decide whether a person is guilty or innocent based on evidence

  Null: The person is innocent

  Alternative: The person is not innocent (i.e. guilty)

- The null can only be rejected if there is enough evidence to doubt it
- i.e. the jury can only convict if there is beyond reasonable doubt for the null of innocence
- They do not know whether the person is really guilty or innocent so they may make a mistake

Types of Errors

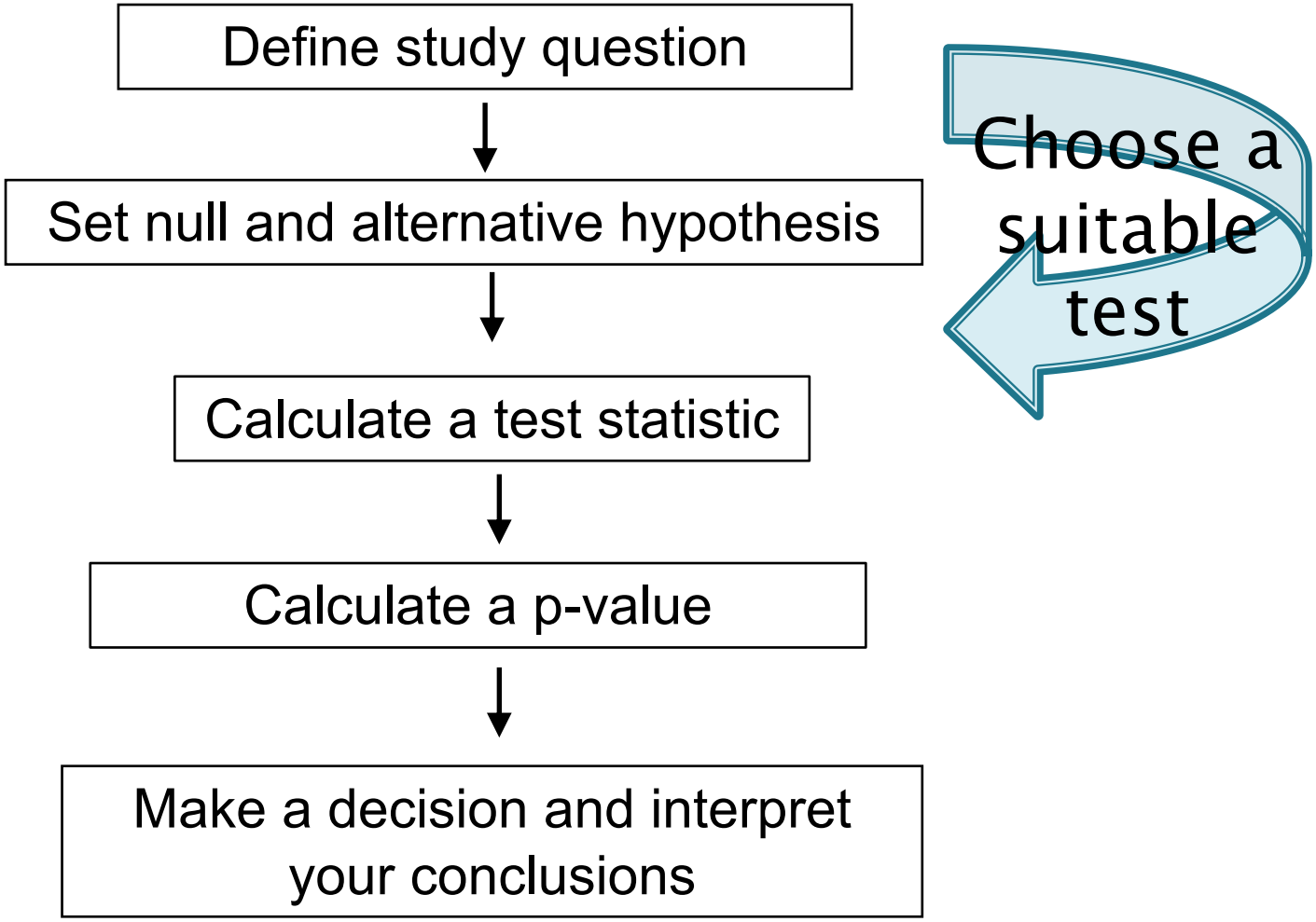
Controlled via sample size (=1−Power of test)

Typically restrict to a 5% Risk = level of significance

	Study reports **NO** difference (Do not reject H_0)	Study reports **IS** a difference (Reject H_0)
H_0 is true Difference Does **NOT** exist in population	✓	X Type I Error
H_A is true Difference **DOES** exist in population	X Type II Error	✓

Prob of this = Power of test

Steps to undertaking a Hypothesis test



Example: Titanic

- The ship Titanic sank in 1912 with the loss of most of its passengers
- 809 of the 1,309 passengers and crew died
  = 61.8%
- **Research question:** Did class of travel (i.e. being 1st, 2nd, or 3rd class) affect survival?

Chi squared Test?

- **Null:** There is **NO** association between class and survival
- **Alternative:** There **IS** an association between class and survival

3 x 2 contingency table

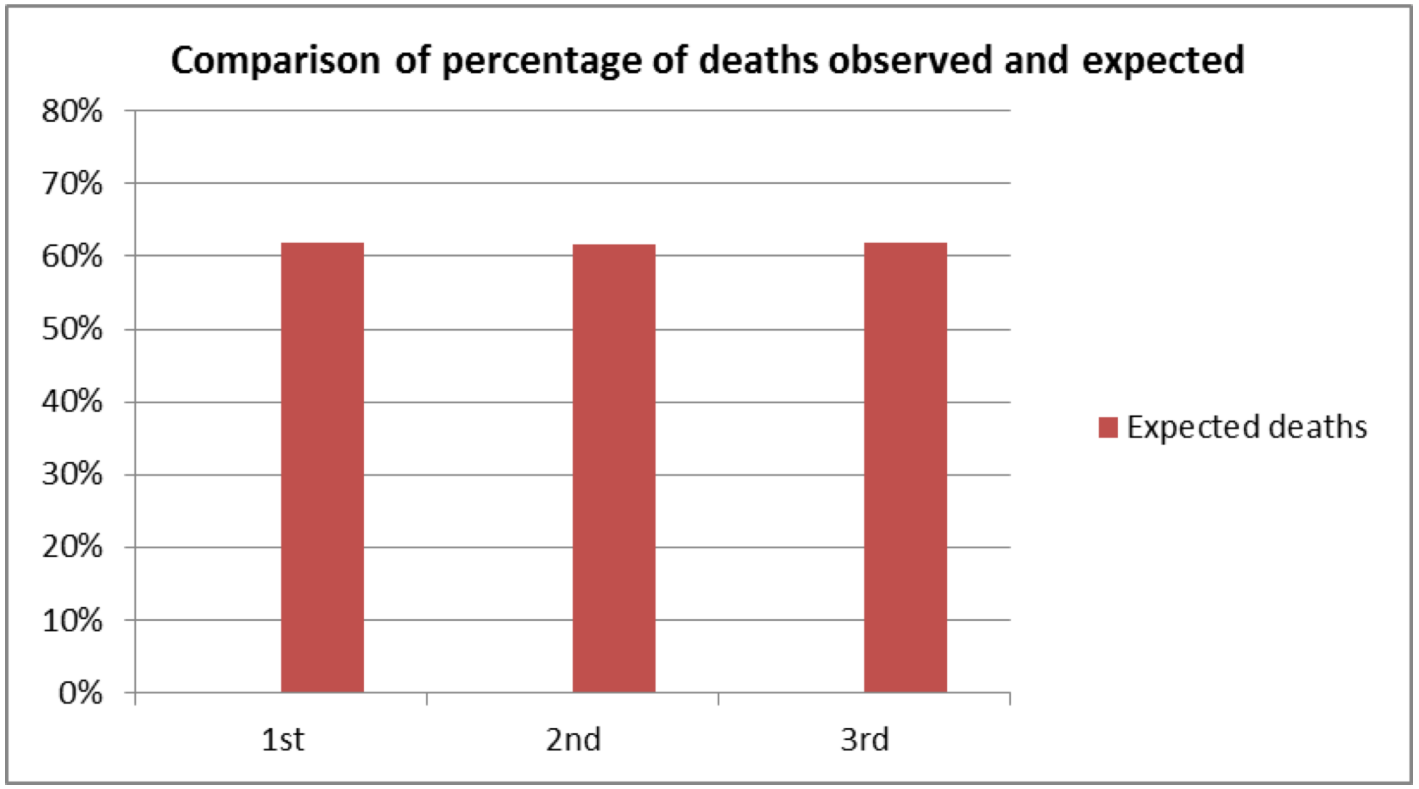
Class * Survived? Crosstabulation

Count

		Survived?		
		Died	Survived	Total
Class	1st	123	200	323
	2nd	158	119	277
	3rd	528	181	709
Total		809	500	1309

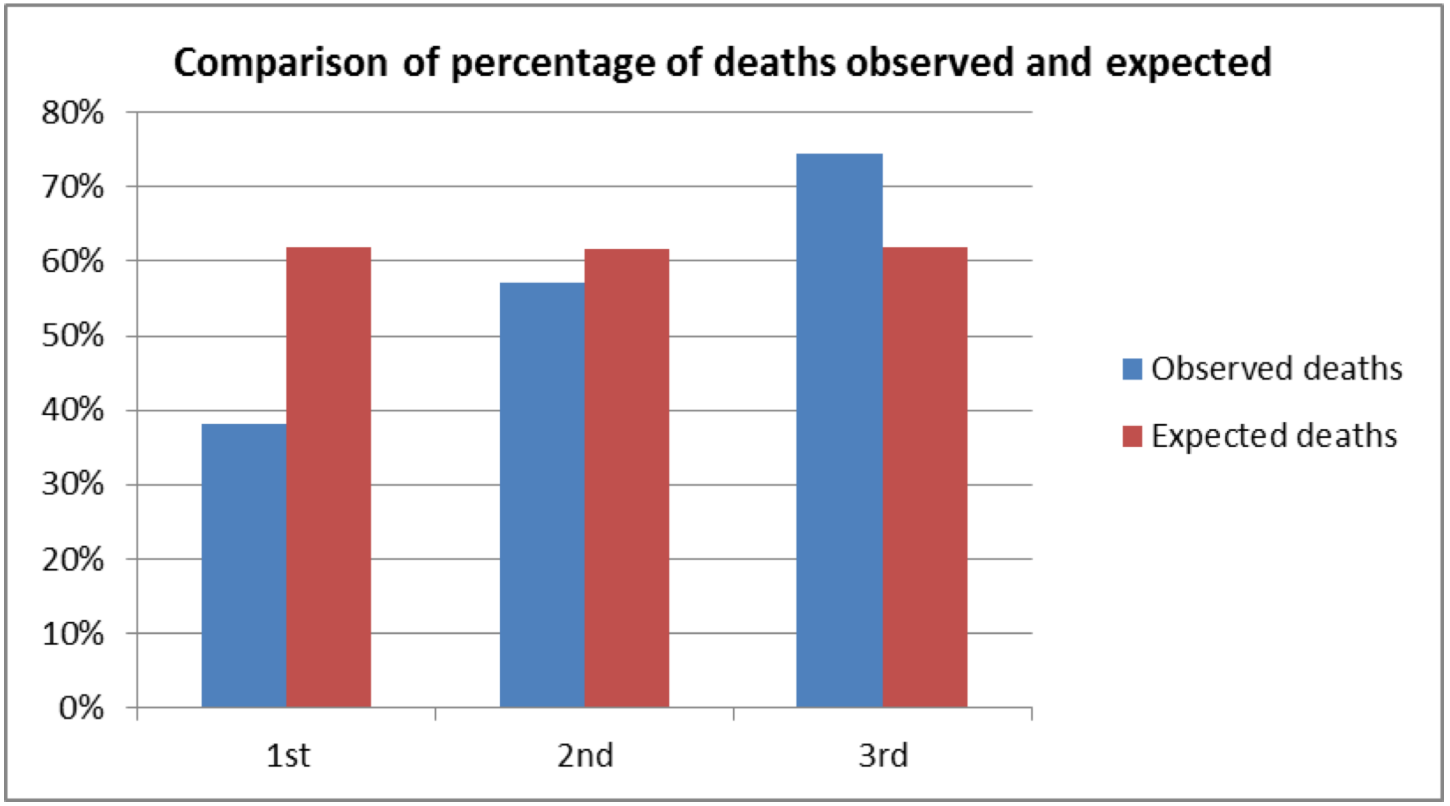
What would be expected if the null is true?

- Same proportion of people would have died in each class!
- Overall, 809 people died out of 1309 = 61.8%

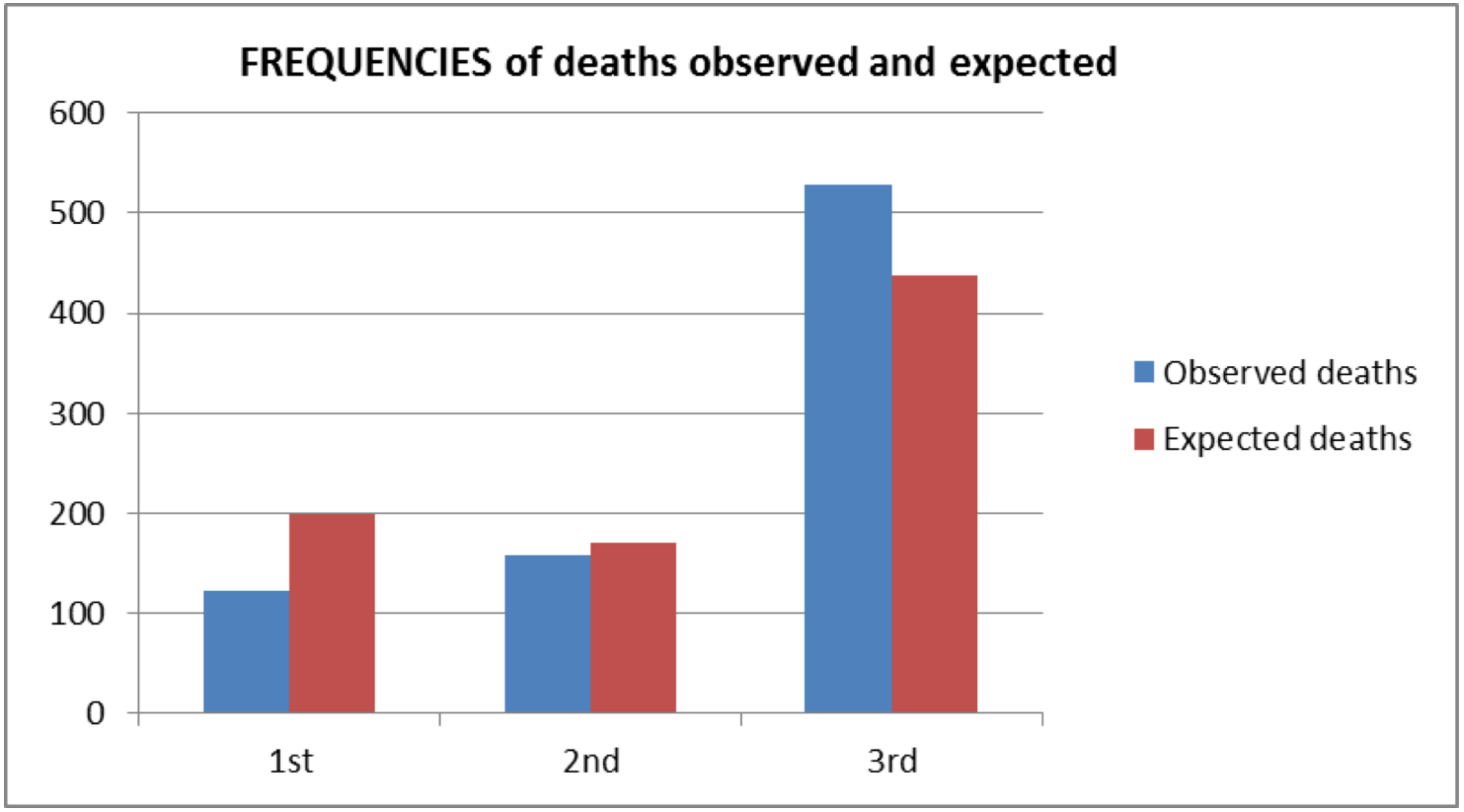


What would be expected if the null is true?

- Same proportion of people would have died in each class!
- Overall, 809 people died out of 1309 = 61.8%



Chi-Squared Test Actually Compares Observed and Expected Frequencies



Expected number dying in each class = 0.618 * no. in class

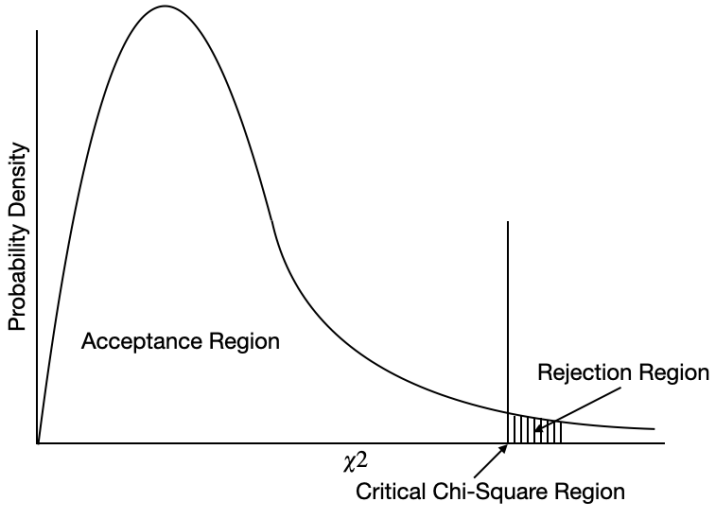
Chi-squared test statistic

- The chi-squared test is used when we want to see if two categorical variables are related
- The test statistic for the Chi-squared test uses the sum of the squared differences between each pair of observed (O) and expected values (E)

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Chi-squared to p-value

Chi-squared test assume chi-squared distribution.
Depending on the degrees of freedom, Chi-squared value can be converted into a p-value



Chi-square Distribution Table

p-values

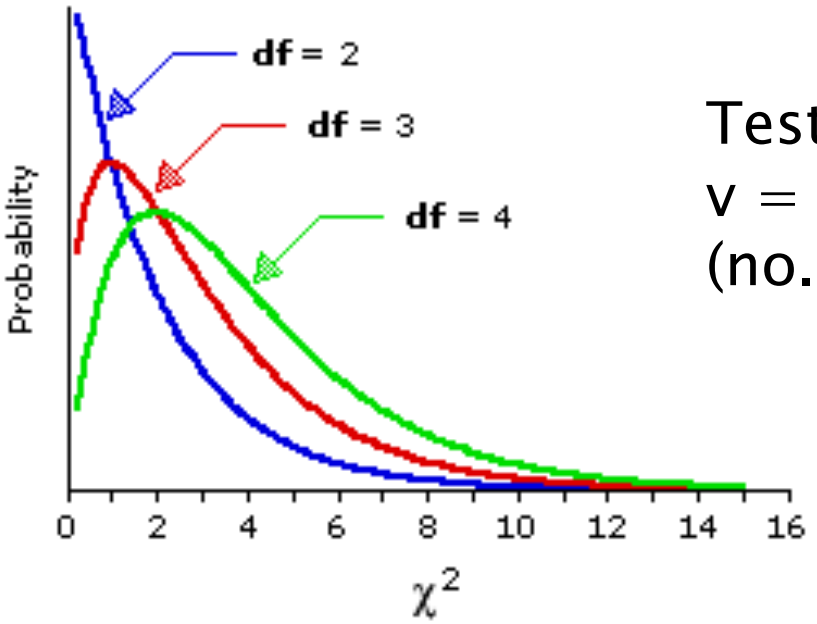
d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58

Degrees of freedom

Chi-Squared values

Chi squared distribution

- P-value: Probability of getting a test statistic at least as extreme as the one calculated **if the null is true**
- The p-value is calculated using the Chi-squared distribution for this test
- Chi-squared is a skewed distribution which varies depending on the degrees of freedom



Testing relationships between 2:
v = degrees of freedom
(no. of rows – 1) x (no. of columns – 1)

Note: One sample test:
v = df = outcomes – 1

Hypothesis Testing: Decision Rule

- We can use statistical software or python modules to undertake a hypothesis test

- One part of the output is the p-value (P)
- If $P < 0.05$ **reject** H_0 $=>$ **Evidence** of H_A being true (i.e. **IS** association)
- If $P > 0.05$ **do not** reject H_0 (i.e. **NO** association)

Open Hypothesis_Testing.ipynb and discuss

Low EXPECTED Cell Counts with the Chi-squared test

We have no cells with expected counts below 5

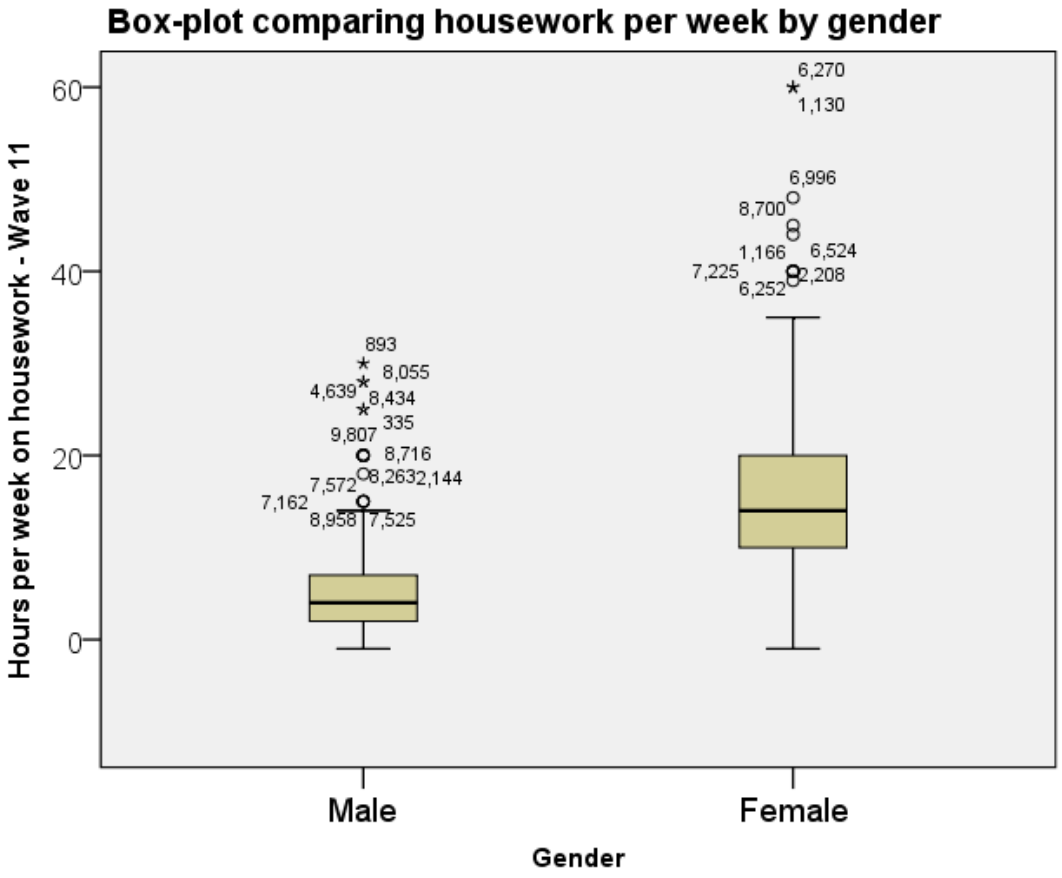
	Died	Survived	Total
1st Class	200	123	323
2nd Class	171	106	277
3rd Class	438	271	709
Total	809	500	1,309

- Check no. of cells with EXPECTED counts less than 5
- If this number corresponds to 20% or higher, then the test statistic does not approximate a chi-squared distribution very well
- In this case
  - if have a 2x2 table, then use **Fishers' Exact test**
  - For larger tables (3x2 etc.), combine categories to make cell counts larger (providing it's meaningful)

Comparing means

Summarising means

- Calculate summary statistics by group
- Look for outliers/ errors
- Use a box-plot or confidence interval plot



T-tests
Paired or Independent (Unpaired) Data?

T-tests are used to compare two population means

We are often interested in comparing two sets of data.

Prior to analysis you must determine whether this data is paired or independent

- **Paired data:** same individuals studied at two different times or under two conditions
  - PAIRED T-TEST
- **Independent:** data collected from two separate groups
  - INDEPENDENT SAMPLES T-TEST

Comparison of hours worked in 1988 to today

Paired or unpaired?

If the same people have reported their hours for 1988 and 2014 have PAIRED measurements of the same variable (hours)

Paired Null hypothesis: The mean of the paired differences = 0

$$H_0 : \mu_d = 0$$

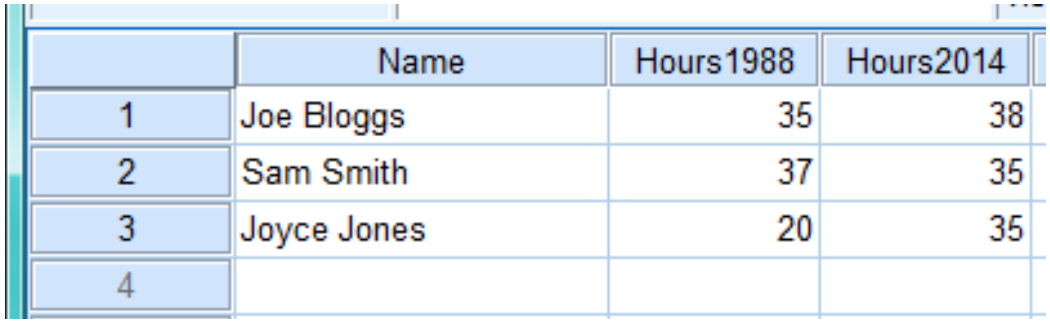
If different people are used in 1988 and 2014 have independent measurements

Independent Null hypothesis: The mean hours worked in 1988 is equal to the mean for 2014

$$H_0 : \mu_{1988} = \mu_{2014}$$

DATA

Paired Data



	Name	Hours1988	Hours2014
1	Joe Bloggs	35	38
2	Sam Smith	37	35
3	Joyce Jones	20	35
4			

Independent Groups

	Name	Hours	Year
1	Joe Bloggs	35	1988
2	Sam Smith	37	1988
3	Joyce Jones	20	1988
4	Li Yu	38	2014
5	Sally McGregor	35	2014
6	Balvinder Singh	35	2014
7			

What is the t-distribution?

- The t-distribution is similar to the standard normal distribution but has an additional parameter called degrees of freedom (df or v)

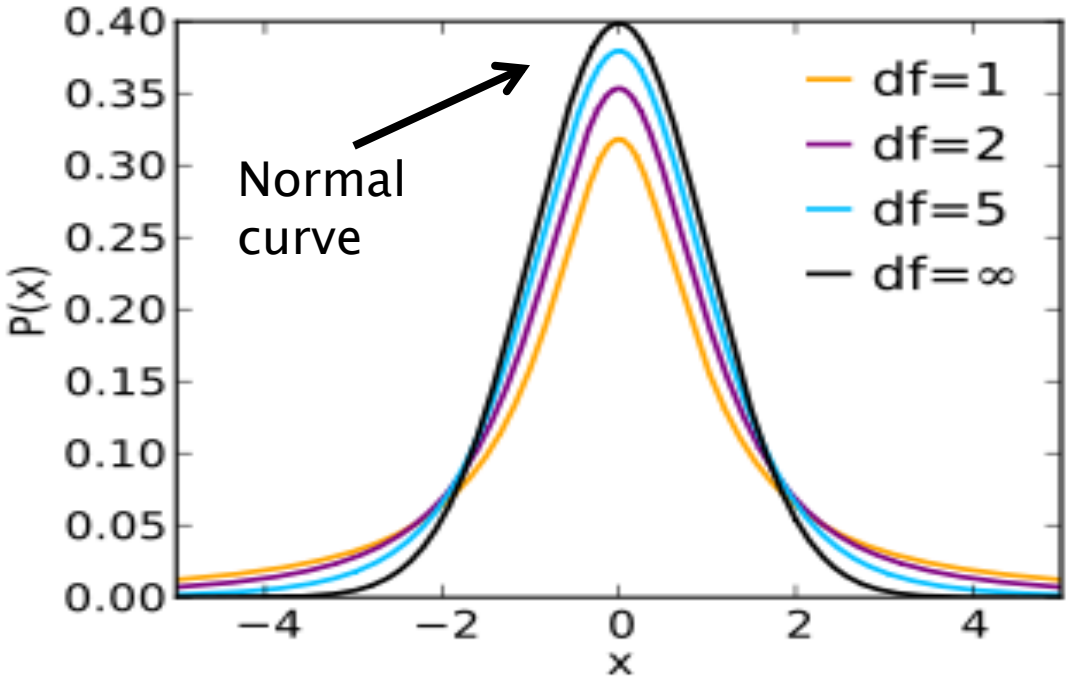
For a paired t-test, v = number of pairs – 1

For an independent t-test, $v = n_{group1} + n_{group2} - 2$

- Used for small samples and when the population standard deviation is not known
- Small sample sizes have heavier tails

Relationship to normal

- As the sample size gets big, the t-distribution matches the normal distribution

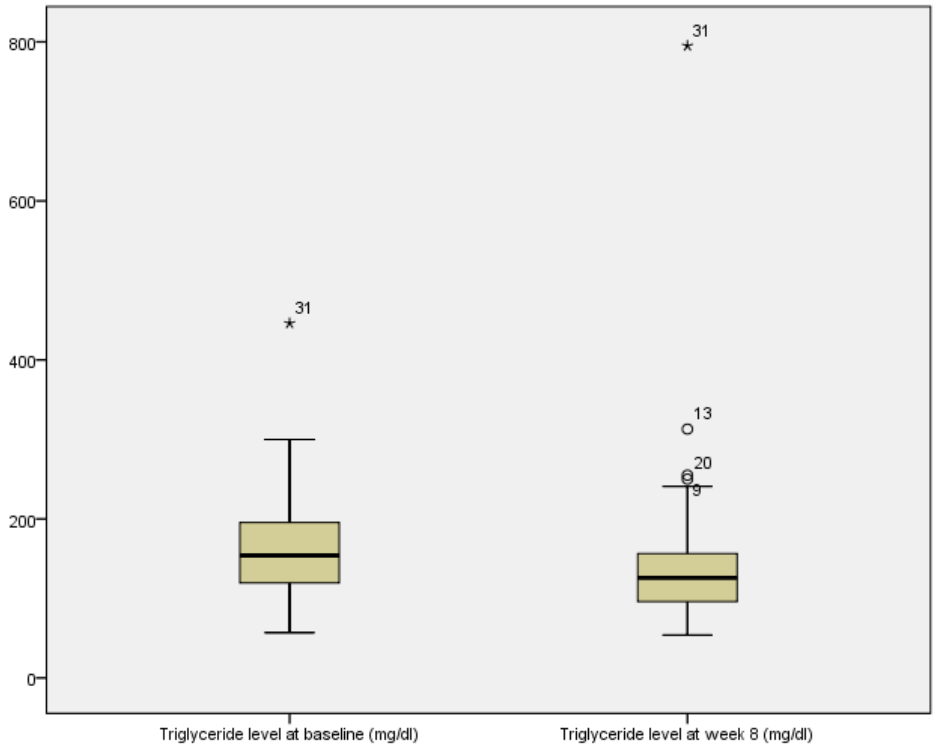


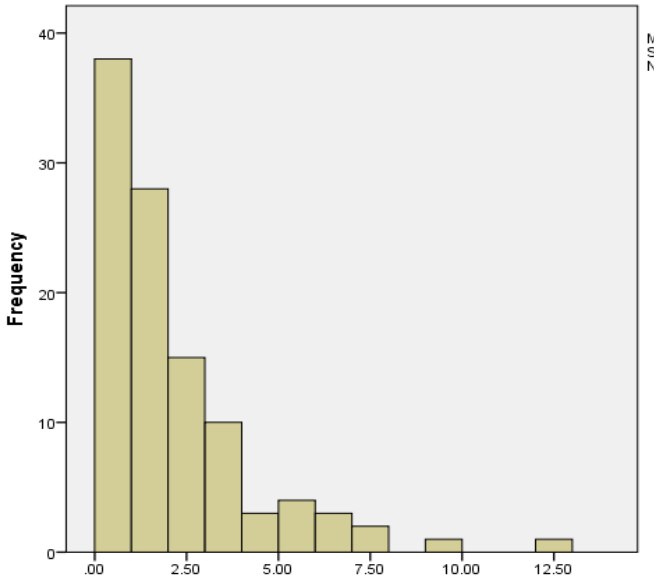
Exercise

- For Examples 1 and 2 (on the following four slides) discuss the answers to the following:
  - State a suitable null hypothesis
  - State whether it's a Paired or Independent Samples t-test
  - Decide whether to reject the null hypothesis
  - State a conclusion in words

Example 1: Triglycerides

- In a weight loss study, Triglyceride levels were measured at baseline and again after 8 weeks of taking a new weight loss treatment.





Triglyceride level difference

Example 1: t-Test Results

	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Triglyceride level at week 8 (mg/dl) - Triglyceride level at baseline (mg/dl)	-11.371	80.360	13.583	-38.976	16.233	-.837	34	.408

Null Hypothesis is:

P-value =
Decision (circle correct answer): Reject Null/ Do not reject Null

Conclusion:

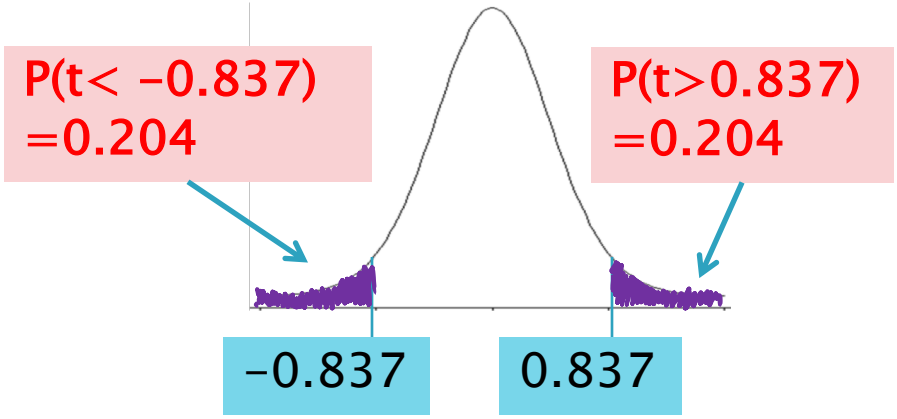
Example 1: Solution

	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Triglyceride level at week 8 (mg/dl) - Triglyceride level at baseline (mg/dl)	-11.371	80.360	13.583	-38.976	16.233	-.837	34	.408

$$H_0 : \mu_d = 0$$

As p > 0.05, do NOT reject the null

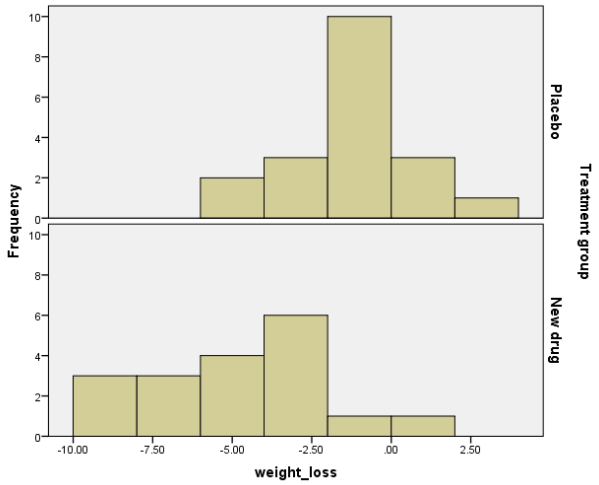
NO evidence of a difference in the mean triglyceride before and after treatment

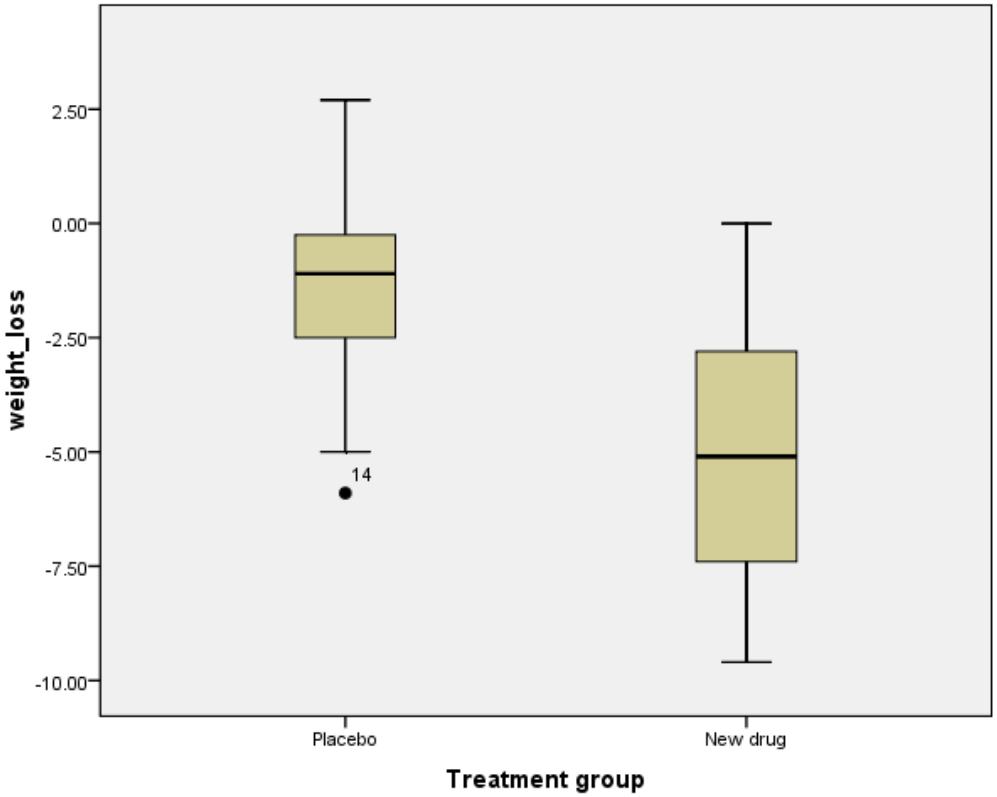


Example 2: Weight Loss

- Weight loss was measured after taking **either** a new weight loss treatment **or** placebo for 8 weeks

Treatment group	N	Mean	Std. Deviation
Placebo	19	-1.36	2.148
New drug	18	-5.01	2.722





Ignore the shaded part of the output for now!

Example 2: t-Test Results

	Levene's Test for Equality of Variances		T-test results					95% CI of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	2.328	.136	4.539	35	.000	3.648	.804	2.016	5.280
Equal variances not assumed			4.510	32.342	.000	3.648	.809	2.001	5.295

Null Hypothesis is:

P-value =

Decision (circle correct answer): Reject Null/ Do not reject Null

Conclusion:

Ignore the shaded part of the output for now!

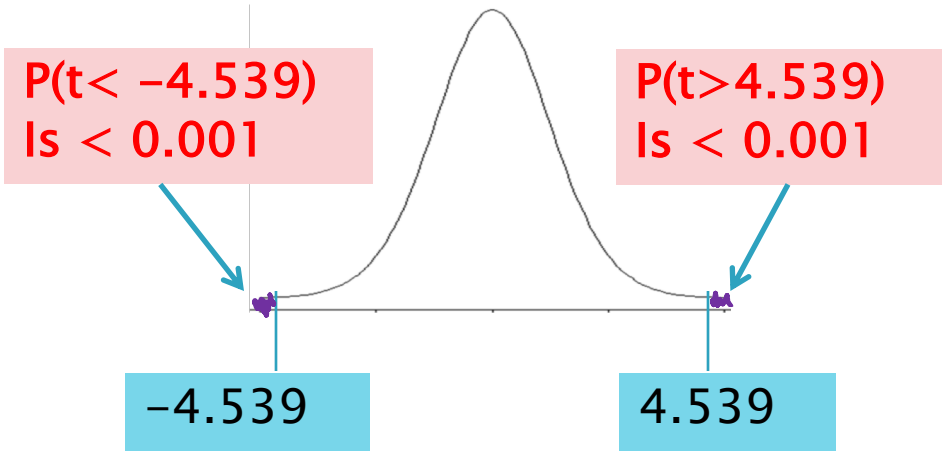
Example 2: Solution

	Levene's Test for Equality of Variances		T-test results					95% CI of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	2.328	.136	4.539	35	.000	3.648	.804	2.016	5.280
Equal variances not assumed			4.510	32.342	.000	3.648	.809	2.001	5.295

H_0: $\mu_{new} = \mu_{placebo}$

As $p < 0.05$, DO reject the null

IS evidence of a difference in weight loss between treatment and placebo



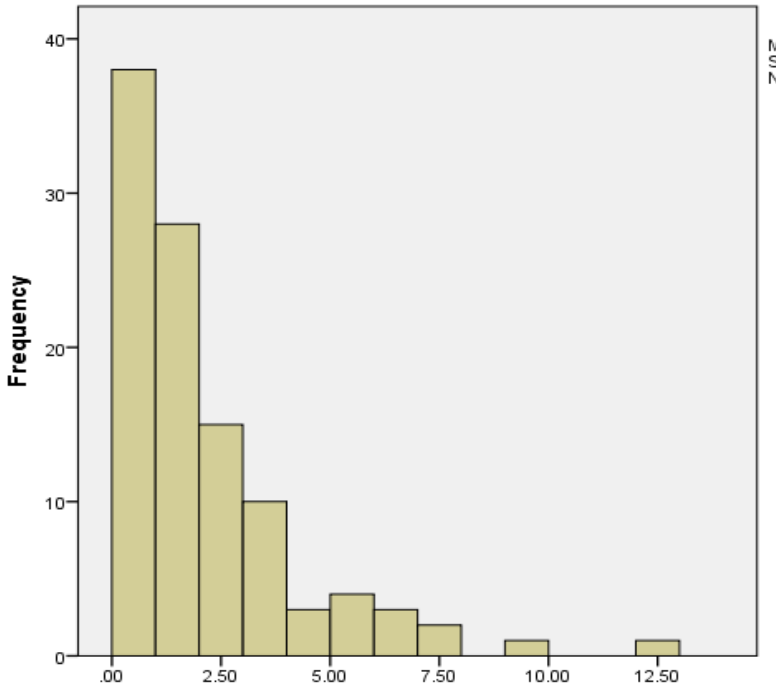
Assumptions in t-Tests

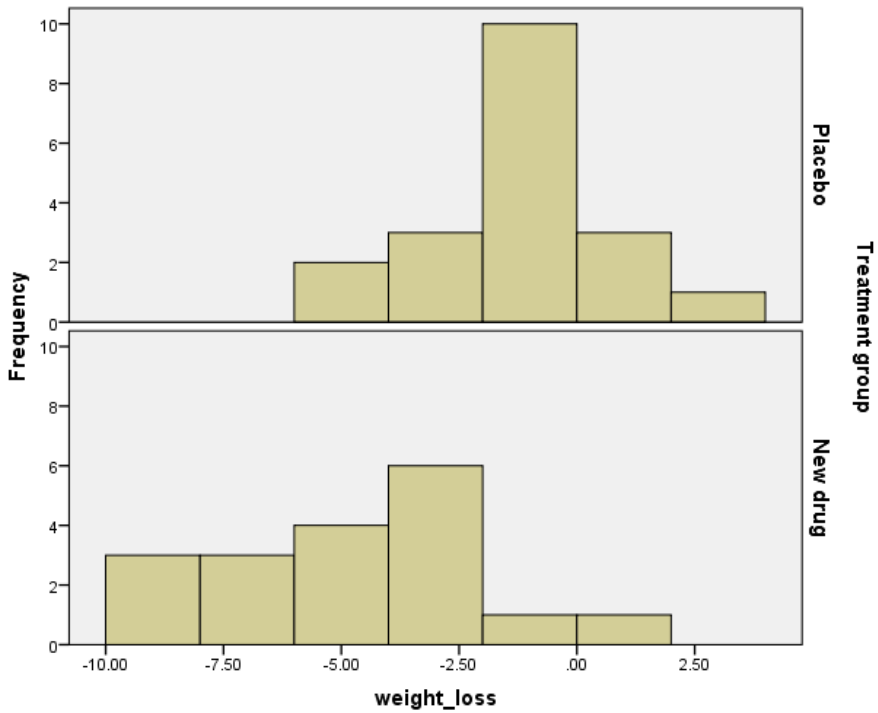
- **Normality**: Plot histograms
  - One plot of the paired differences for any paired data
  - Two (One for each group) for independent samples
  - Don't have to be perfect, just roughly symmetric

- **Equal Population variances**: Compare sample standard deviations
  - As a rough estimate, one should be no more than twice the other
  - Do an F-test to formally test for differences

- However the *t*-test is very robust to violations of the assumptions of Normality and equal variances, particularly for moderate (i.e. >30) and larger sample sizes

Histograms from Examples 1 and 2





Do these histograms look approximately normally distributed?

Levene's Test for Equal Variances from Examples 2

	Levene's Test for Equality of Variances		T-test results					95% CI of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	2.328	.136	4.539	35	.000	3.648	.804	2.016	5.280
Equal variances not assumed			4.510	32.342	.000	3.648	.809	2.001	5.295

Null hypothesis is that pop variances are equal

i.e. H_0: $\sigma^2_{new} = \sigma^2_{placebo}$

Since p = 0.136 and so is >0.05 we do not reject the null

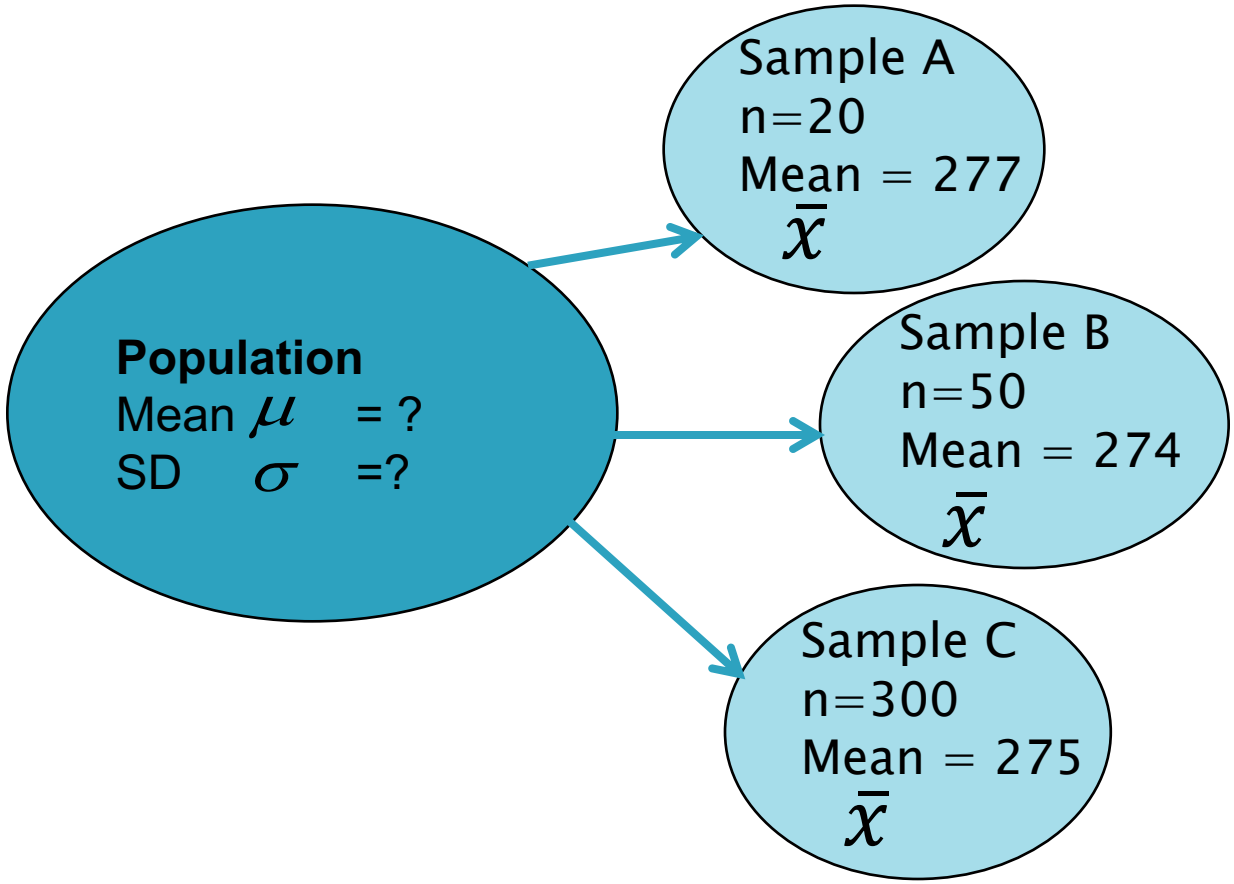
i.e. we can assume equal variances ☺

What if the assumptions are not met?

- There are alternative tests which do not have these assumptions

Test	Check	Equivalent non-parametric test
Independent t-test	Histograms of data by group	Mann-Whitney
Paired t-test	Histogram of paired differences	Wilcoxon signed rank

Sampling Variation



Every sample taken from a population, will contain different numbers so the mean varies.

Which estimate is most reliable?

How certain or uncertain are we?

Confidence Intervals

- A range of values within which we are confident (in terms of probability) that the true value of a pop parameter lies
- A 95% CI is interpreted as 95% of the time the CI would contain the true value of the pop parameter
- i.e. 5% of the time the CI would fail to contain the true value of the pop parameter

	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Triglyceride level at week 8 (mg/dl) - Triglyceride level at baseline (mg/dl)	-11.371	80.360	13.583	-38.976	16.233	-.837	34	.408

Exercise

- Discuss what the interpretation is for the confidence interval from Example 2 (Weight loss was measured after taking **either** a new weight loss treatment **or** placebo for 8 weeks) highlighted below:

	Levene's Test for Equality of Variances		T-test results					95% CI of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	2.328	.136	4.539	35	.000	3.648	.804	2.016	5.280

The true mean weight loss would be between about 2 to 5 kg with the new treatment.

Investigating relationships

Two categorical variables

Are boys more likely to prefer maths and science than girls?

Variables:

- Favourite subject (Nominal)
- Gender (Binary/ Nominal)

Summarise using %'s/ stacked or multiple bar charts

Test: Chi-squared

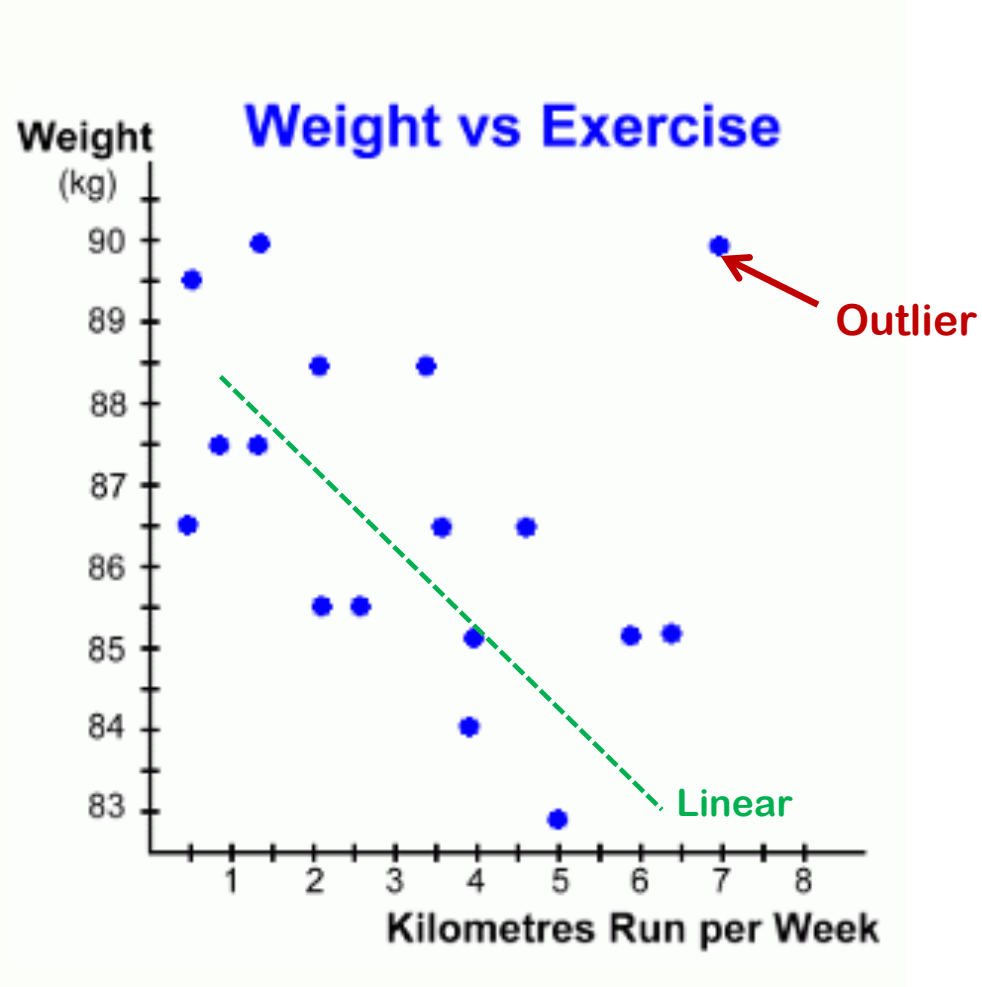
Tests for a relationship between two categorical variables

Scatterplot

Relationship between two scale variables:

- Explores the way the two co-vary: (correlate)
  - Positive / negative
  - Linear / non-linear
  - Strong / weak
- Presence of outliers
- Statistic used:

r = correlation coefficient

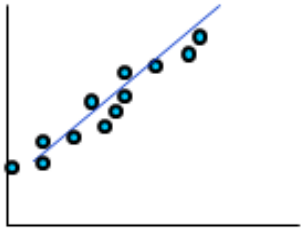


Correlation Coefficient r

- **Measures strength of a relationship between two continuous variables**

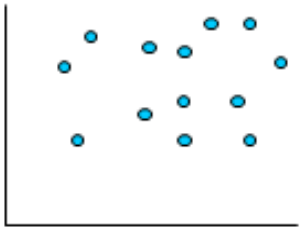
$-1 \leq r \leq 1$

Strong positive linear relationship



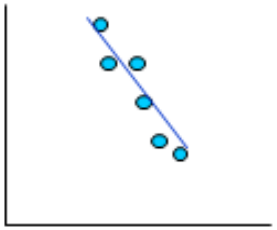
$r = 0.9$

No linear relationship



$r = 0.01$

Strong negative linear relationship



$r = -0.9$

Correlation Interpretation

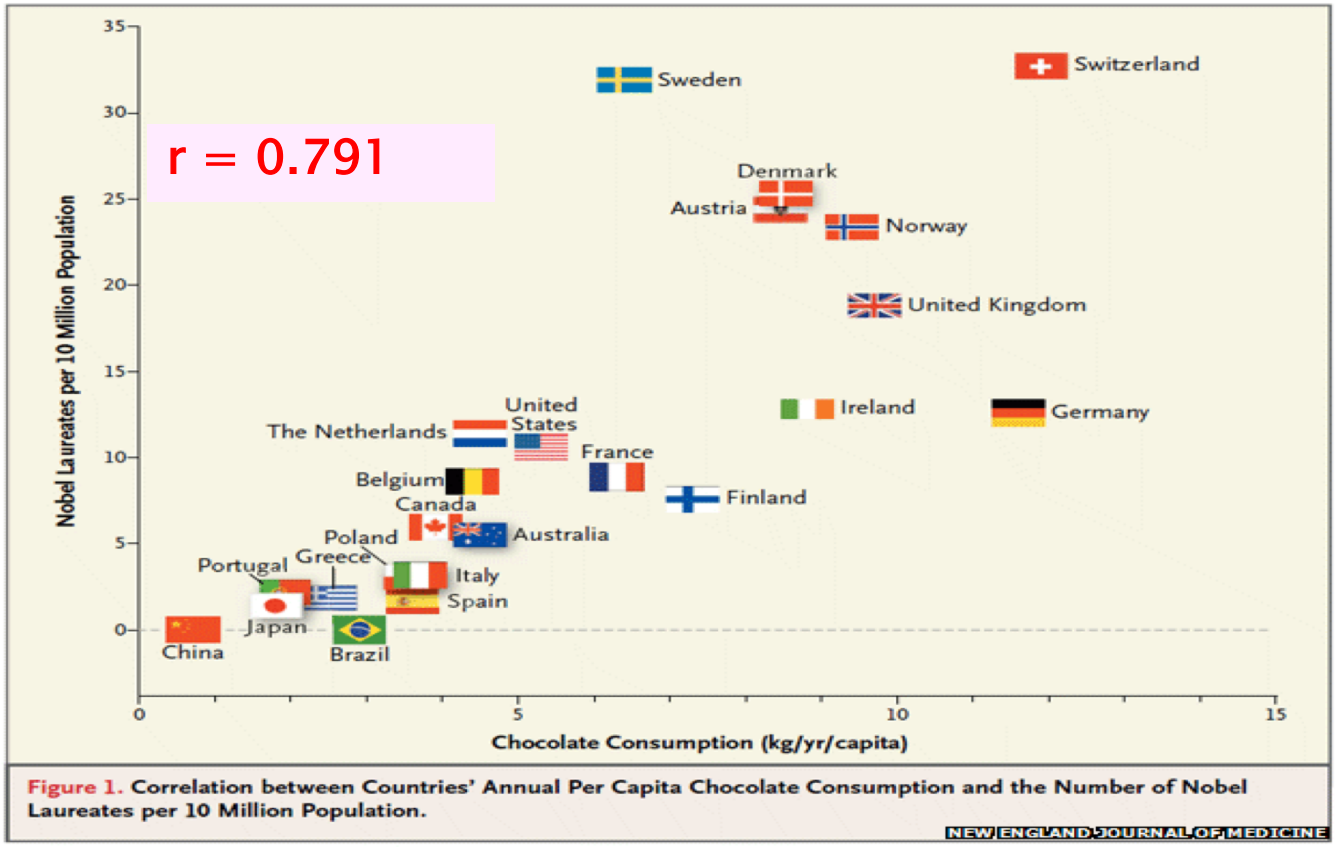
An interpretation of the size of the coefficient has been described by Cohen (1992) as:

Correlation coefficient value			Relationship
−0.3 to +0.3			Weak
−0.5 to −0.3	or	0.3 to 0.5	Moderate
−0.9 to −0.5	or	0.5 to 0.9	Strong
−1.0 to −0.9	or	0.9 to 1.0	Very strong

Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159

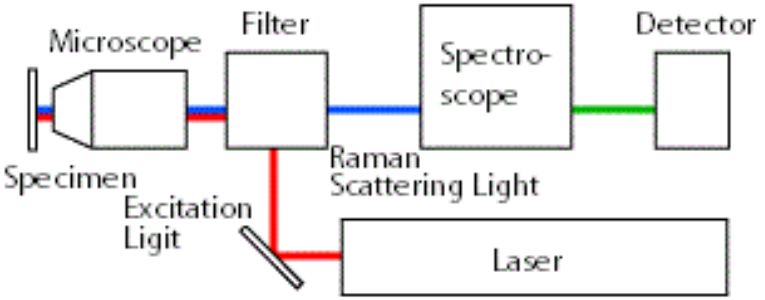
Does chocolate make you clever or crazy?

- A paper in the New England Journal of Medicine claimed a relationship between chocolate and Nobel Prize winners



http://www.nejm.org/doi/full/10.1056/NEJMon1211064

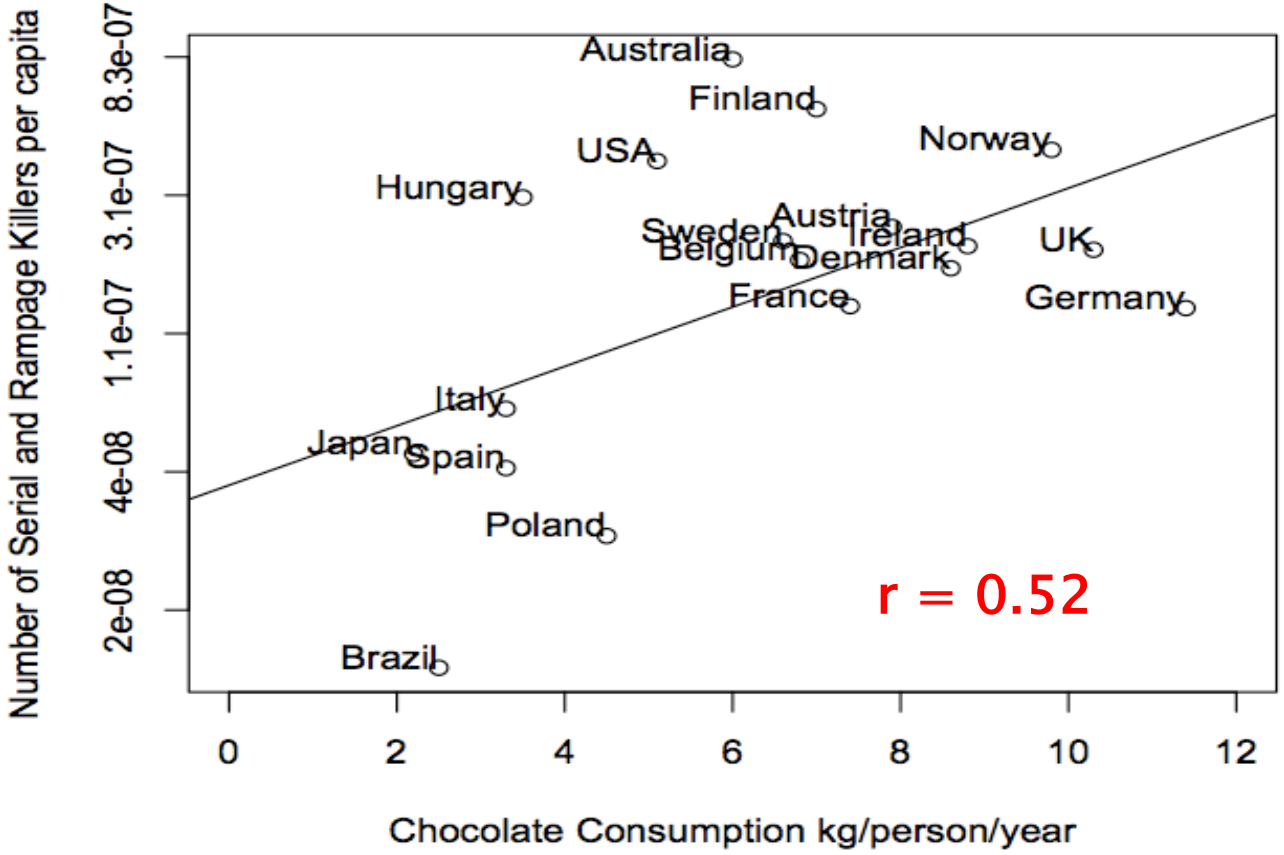
Sir Chandrasekhara Venkata Raman



▲ Components of Raman Microscope

Chocolate and serial killers

- What else is related to chocolate consumption?



http://www.replicatedtypo.com/chocolate-consumption-traffic-accidents-and-serial-killers/5718.html

Hypothesis tests for r

- Tests the null hypothesis that the population correlation r = 0 NOT that there is a strong relationship!
- Sample size plays a big part in determining what level of r is significant.
  - The larger the sample size, the lower the coefficient needs to be to be significant
  - e.g. sample size of 150 will classify a correlation of 0.16 as significant!
- Better to use Cohen's interpretation

Exercise

- Interpret the following correlation coefficients using Cohen's and explain what it means

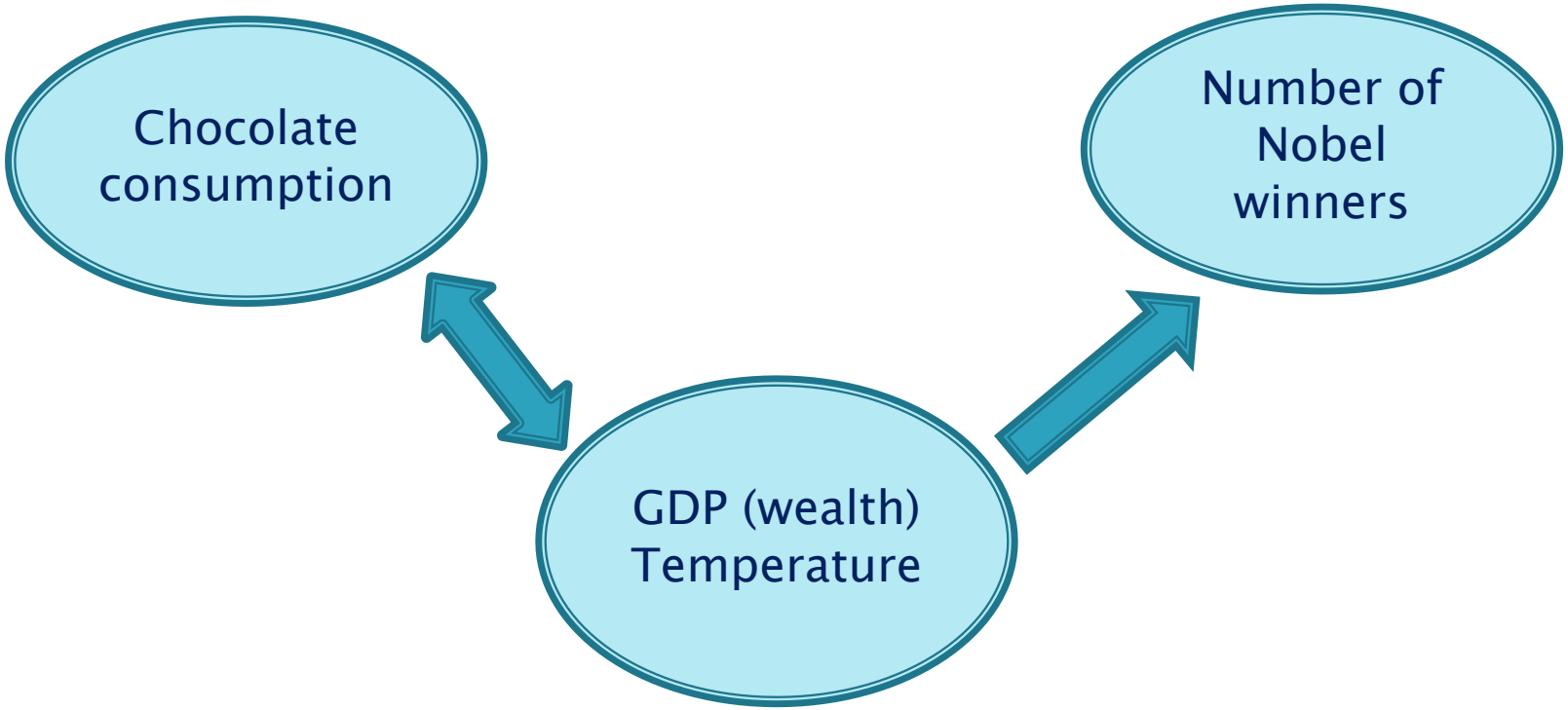
Relationship	Correlation
Average IQ and chocolate consumption	0.27
Road fatalities and Nobel winners	0.55
Gross Domestic Product and Nobel winners	0.7
Mean temperature and Nobel winners	-0.6

Exercise – solution

Relationship	Correlation	Interpretation
Average IQ and chocolate consumption	0.27	Weak positive relationship. More chocolate per capita = higher average IQ
Road fatalities and Nobel winners	0.55	Strong positive. More accidents = more prizes!
Gross Domestic Product and Nobel winners	0.7	Strong positive. Wealthy countries = more prizes
Mean temperature and Nobel winners	-0.6	Strong negative. Colder countries = more prizes.

Confounding

Is there something else affecting both chocolate consumption and Nobel prize winners?



When looking into correlations, it's possible that a relationship between 2 variables may actually be caused by something else related to both. These are called confounding variables.

Birth Weight Dataset

- Factors affecting birth weight of babies

id	headcirumference	length	Birthweight	Gestation	smoker	motherage
1313	12	17	5.8	33	0	24
431	12	19	4.2	33	1	20
808	13	19	6.4	34	0	26
300	12	18	4.5	35	1	
516	13	18	5.8	35	1	20
321	13	19	6.8	37	0	28
1363	12	19	5.2	37	1	20
575	12	19	6.1	37	1	19
822	13	19	7.5	38	0	20
1081	14	21	8.0	38	0	18
1636	14	20	8.6	38	0	29

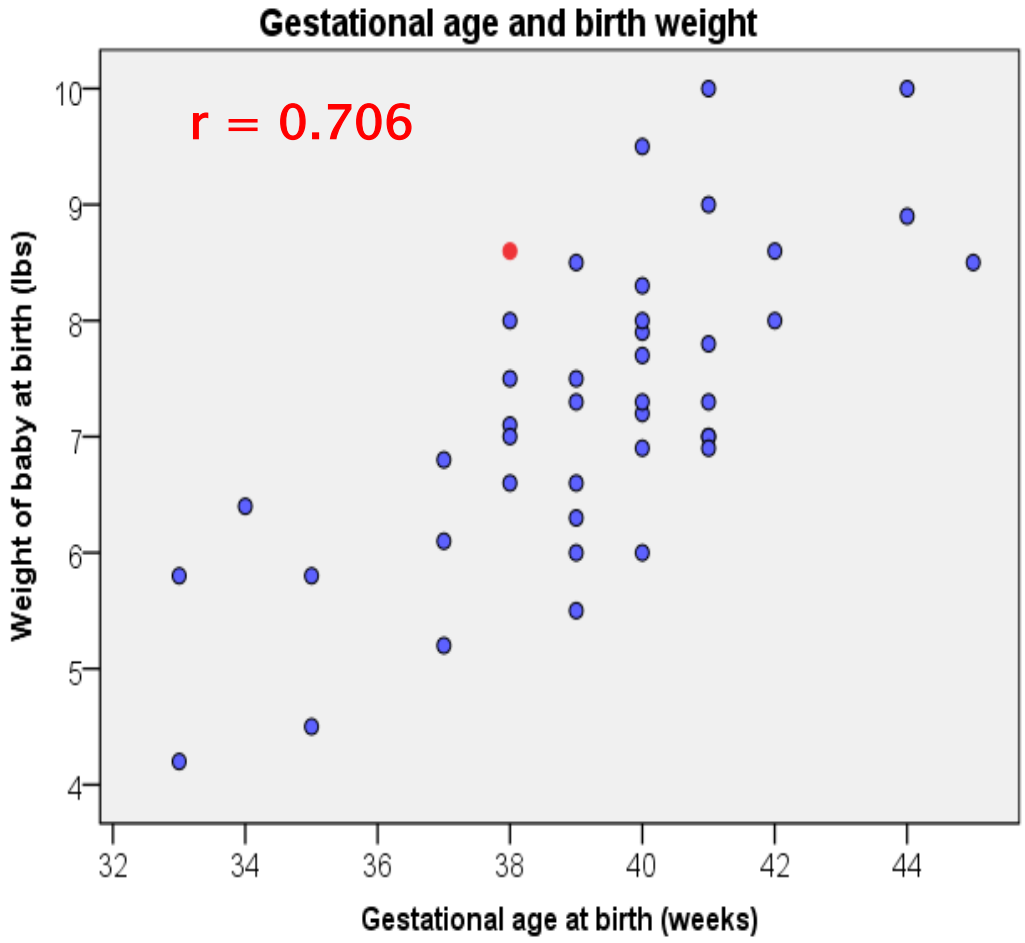
Mother smokes = 1

Standard gestation = 40 weeks

Exercise: Gestational age and birth weight

a) Describe the relationship between the gestational age of a baby and their weight at birth.

b) Draw a line of best fit through the data (with roughly half the points above and half below)



Exercise – Solution

Describe the relationship between the gestational age of a baby and their weight at birth.

There is a strong positive relationship which is linear

Regression: Association between two variables

- Regression is useful when we want to

a) look for significant relationships between two variables

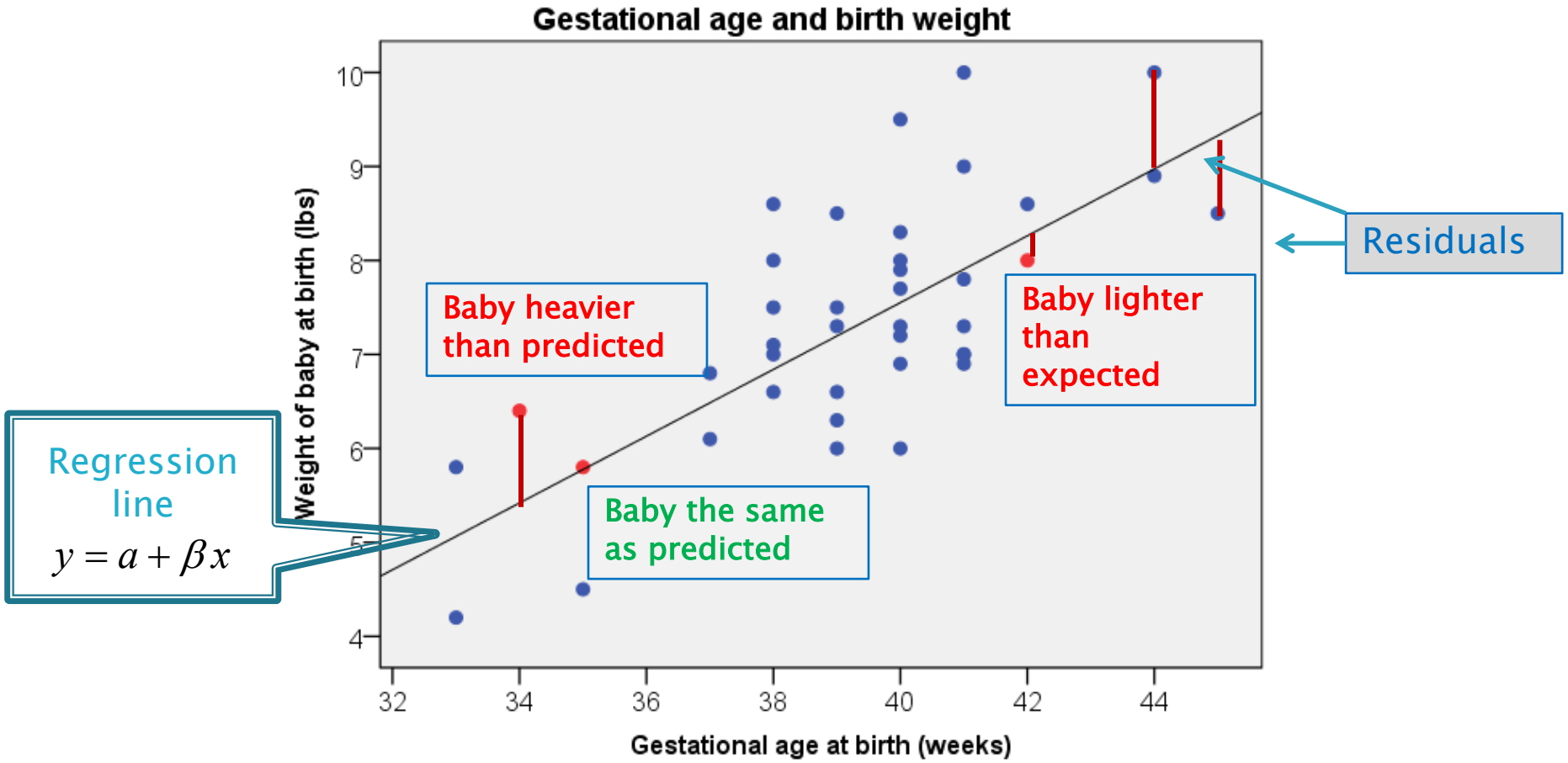
b) predict a value of one variable for a given value of the other

It involves estimating the line of best fit through the data which minimises the sum of the squared residuals

What are the residuals?

Residuals

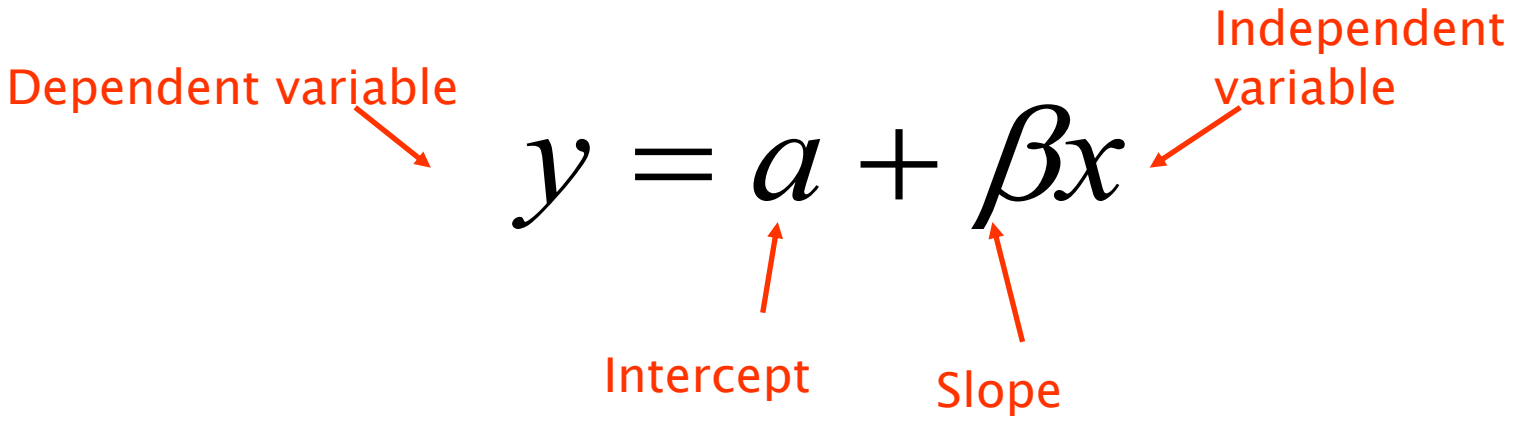
- Residuals are the differences between the observed and predicted weights



X Predictor / explanatory variable (independent variable)

Regression

Simple linear regression looks at the relationship between two Scale variables by producing an equation for a straight line of the form



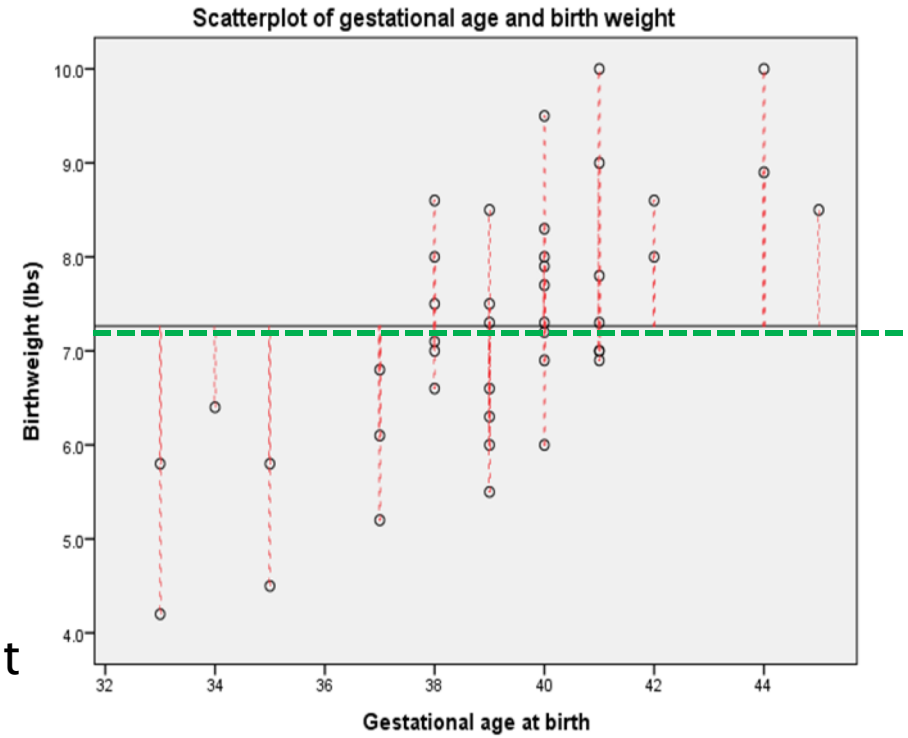
Which uses the independent variable to predict the dependent variable

Hypothesis testing

- We are often interested in how likely we are to obtain our estimated value of β if there is actually no relationship between x and y in the population

One way to do this is to do a test of significance for the slope

$$H_0 : \beta = 0$$



This means that every baby would be classified as the same weight irrelevant of what gestational age they are.

Sample Regression Table

- Key regression table:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-6.660	2.212		-3.011	.004
	Gestational age at birth	.355	.056	.706	6.310	.000

a. Dependent Variable: Birthweight (lbs)

$Y = -6.66 + 0.36x$

P – value < 0.001

- As $p < 0.05$, gestational age is a significant predictor of birth weight. Weight increases by 0.36 lbs for each week of gestation.

How reliable are predictions? – R^2

How much of the variation in birth weight is explained by the model including Gestational age?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.706[a]	.499	.486	.9530

a. Predictors: (Constant), Gestational age at birth

b. Dependent Variable: Birth weight (lbs)

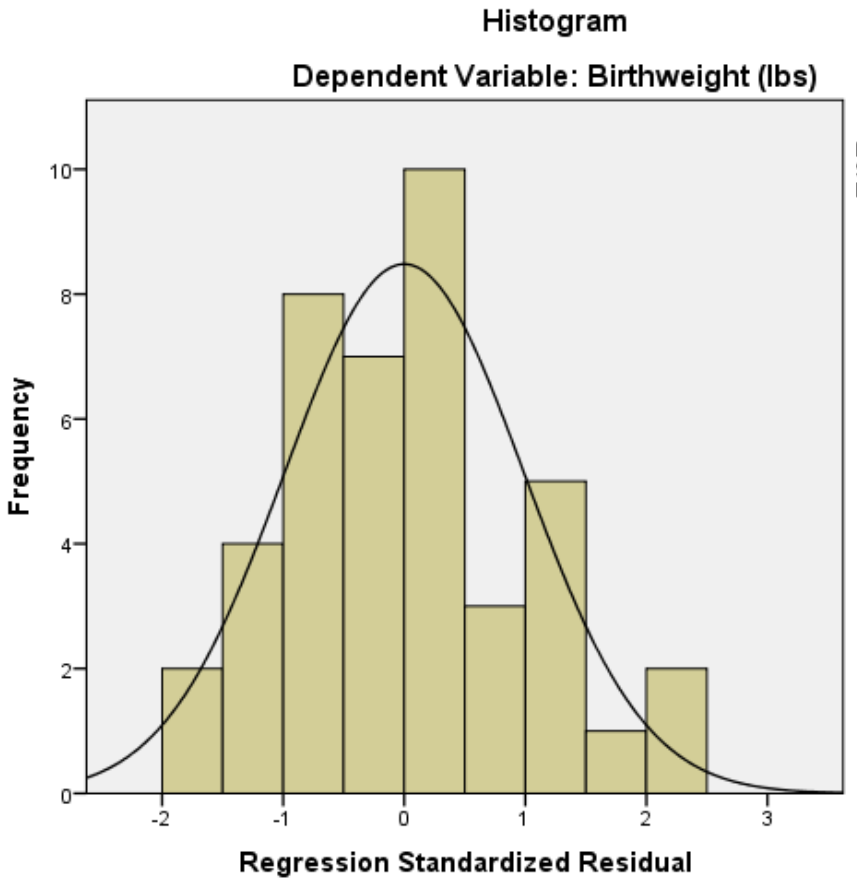
Proportion of the variation in birth weight explained by the model $R^2 = 0.499 = 50\%$
Predictions using the model are fairly reliable.

Which variables may help improve the fit of the model?
Compare models using Adjusted R^2

Assumptions for regression

Assumption	Plot to check
The relationship between the independent and dependent variables is linear.	Original scatter plot of the independent and dependent variables
Homoscedasticity: The variance of the residuals about predicted responses should be the same for all predicted responses.	Scatterplot of standardised predicted values and residuals
The residuals are independently normally distributed	Plot the residuals in a histogram

Checking normality



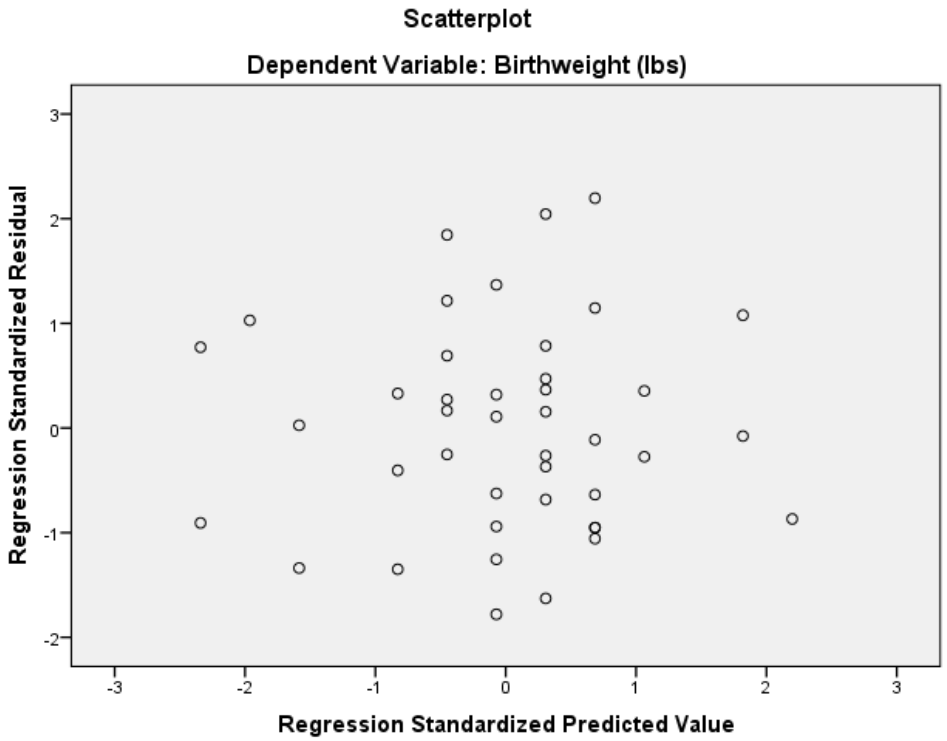
Outliers are outside ± 3

Histogram of the residuals looks approximately normally distributed

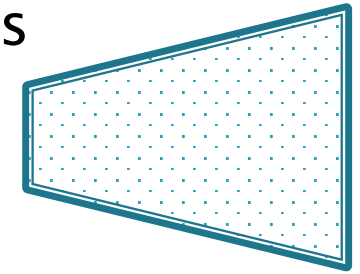
When writing up, just say 'normality checks were carried out on the residuals and the assumption of normality was met'

Predicted values against residuals

Are there any patterns as the predicted values increases?



There is a problem with Homoscedasticity if the scatter is not random. A "funnelling" shape such as this suggests problems.



What if assumptions are not met?

- If the residuals are heavily skewed or the residuals show different variances as predicted values increase, the data needs to be transformed
- Try taking the natural log (ln) of the dependent variable. Then repeat the analysis and check the assumptions