Data and variables

DATA: the answers to questions or measurements from the experiment

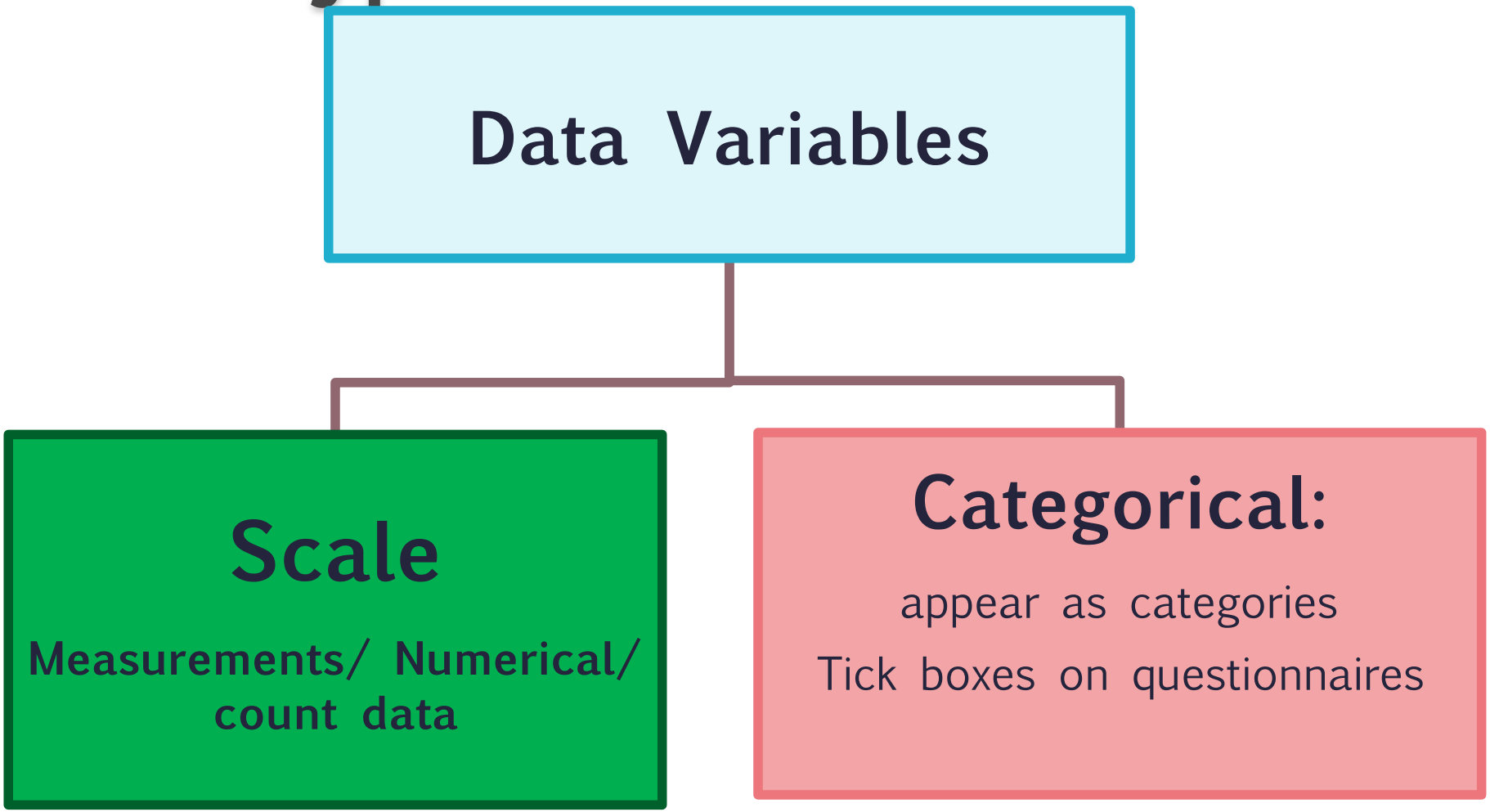
VARIABLE = measurement which varies between subjects e.g. height or gender

One variable per column

One row per subject

	A	B	C	D
1	Subject ID	Gender	Year of study	Height
2	1	Male	1	170
3	2	Female	2	160
4	3	Female	3	165
5	4	Male	PG	175
6	5	Female	3	168

Data types


Data Variables
Scale
Measurements/ Numerical/ count data
Categorical:
appear as categories
Tick boxes on questionnaires

Data types

Variables

- **Scale**
  - **Continuous**
    Measurements
    takes any value
  - **Discrete:**
    Counts/ integers
- **Categorical**
  - **Ordinal:**
    obvious order
  - **Nominal:**
    no meaningful order

What data types relate to following questions?

- Q1: What is your favourite subject?

Maths	English	Science	Art	French

- Q2: Gender:

Male	Female

- Q3: I consider myself to be good at mathematics:

Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree

- Q4: Score in a recent mock GCSE maths exam:

Score between 0% and 100%

What data types relate to following questions?

- Q1: What is your favourite subject?

Nominal

Maths	English	Science	Art	French

- Q2: Gender:

Male	Female

Binary/ Nominal

- Q3: I consider myself to be good at mathematics:

Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree

Ordinal

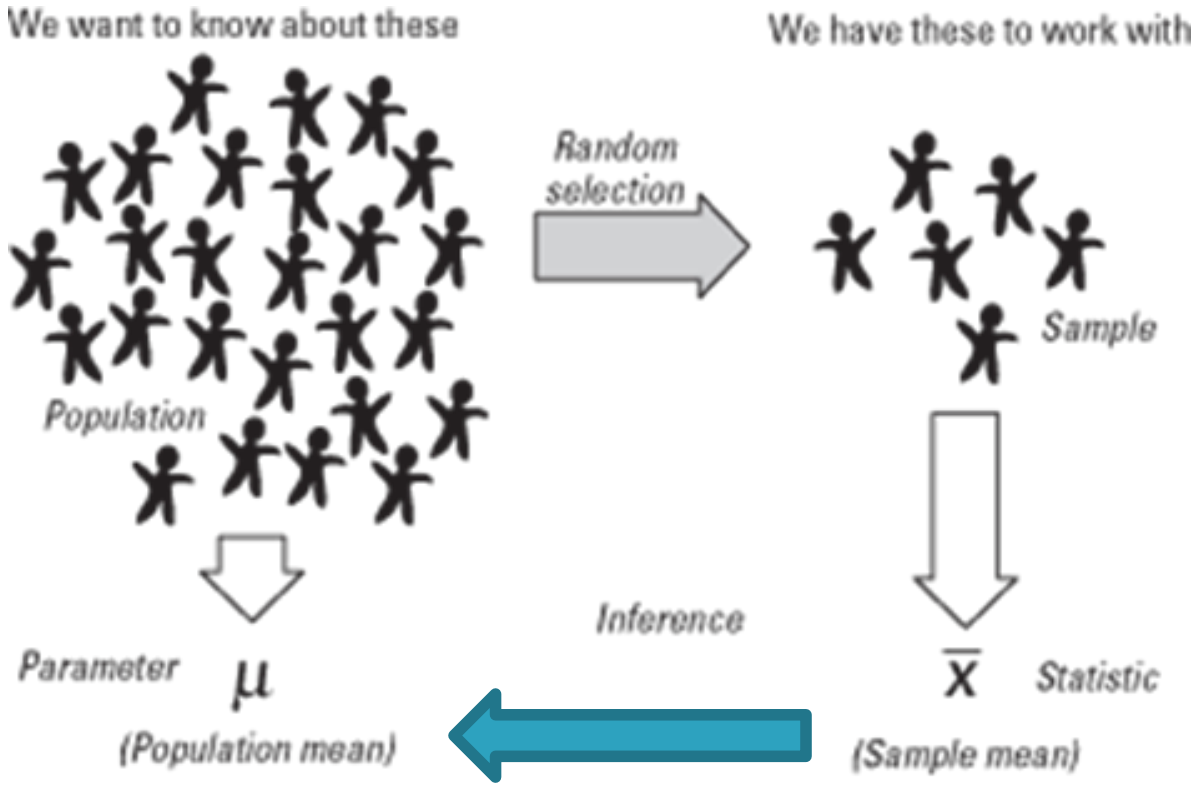
- Q4: Score in a recent mock GCSE maths exam:

Score between 0% and 100%

Scale

Populations and samples

- Taking a sample from a population



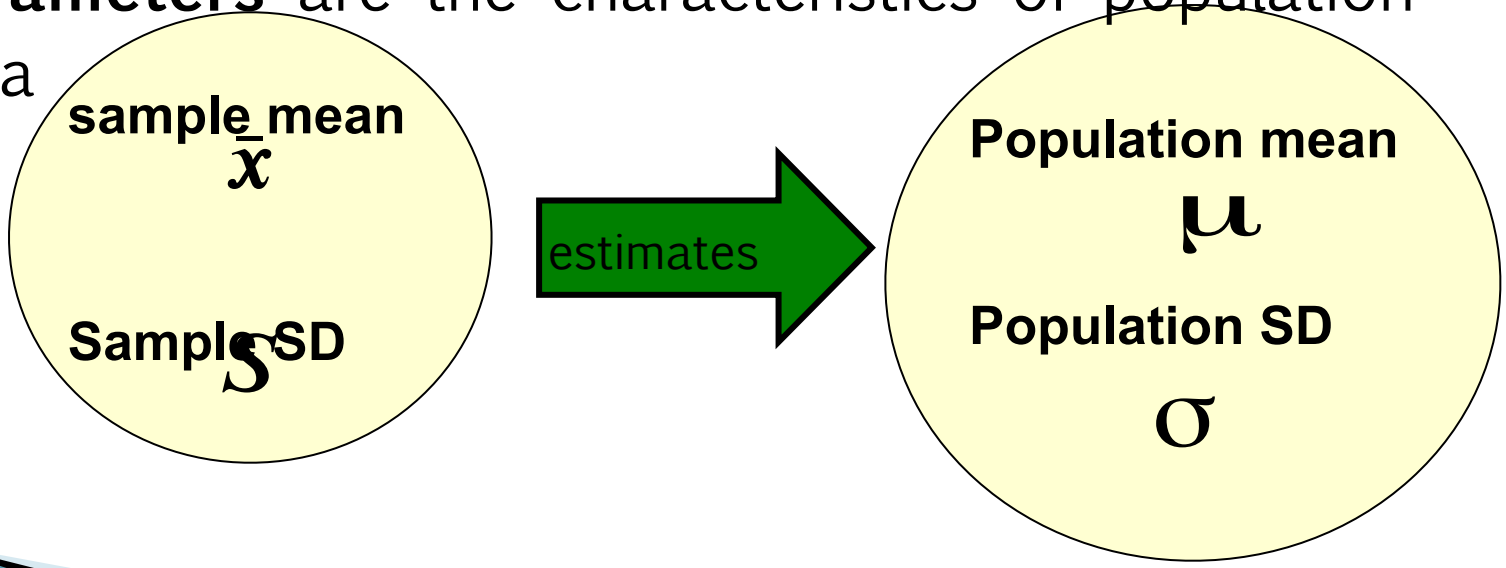
Sample data 'represents' the whole population

Point estimation

Sample data is used to estimate parameters of a population

Statistics are calculated using sample data.

Parameters are the characteristics of population data



How can exam score data be summarized?

Exam marks for 60 students (marked out of 65)

48	37	1	33	26	22	15	22	40	30	12	36
21	20	29	13	44	52	28	39	16	48	56	27
47	12	35	24	10	36	18	34	9	25	31	42
31	27	64	25	58	17	26	38	28	43	33	5
25	55	7	32	39	23	49	43	11	37	22	54

Summary statistics

- Mean = $\frac{\sum_{i=1}^{n} x}{n} = \bar{x}$

Standard deviation (s) is a measure of how much the individuals differ from the mean

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

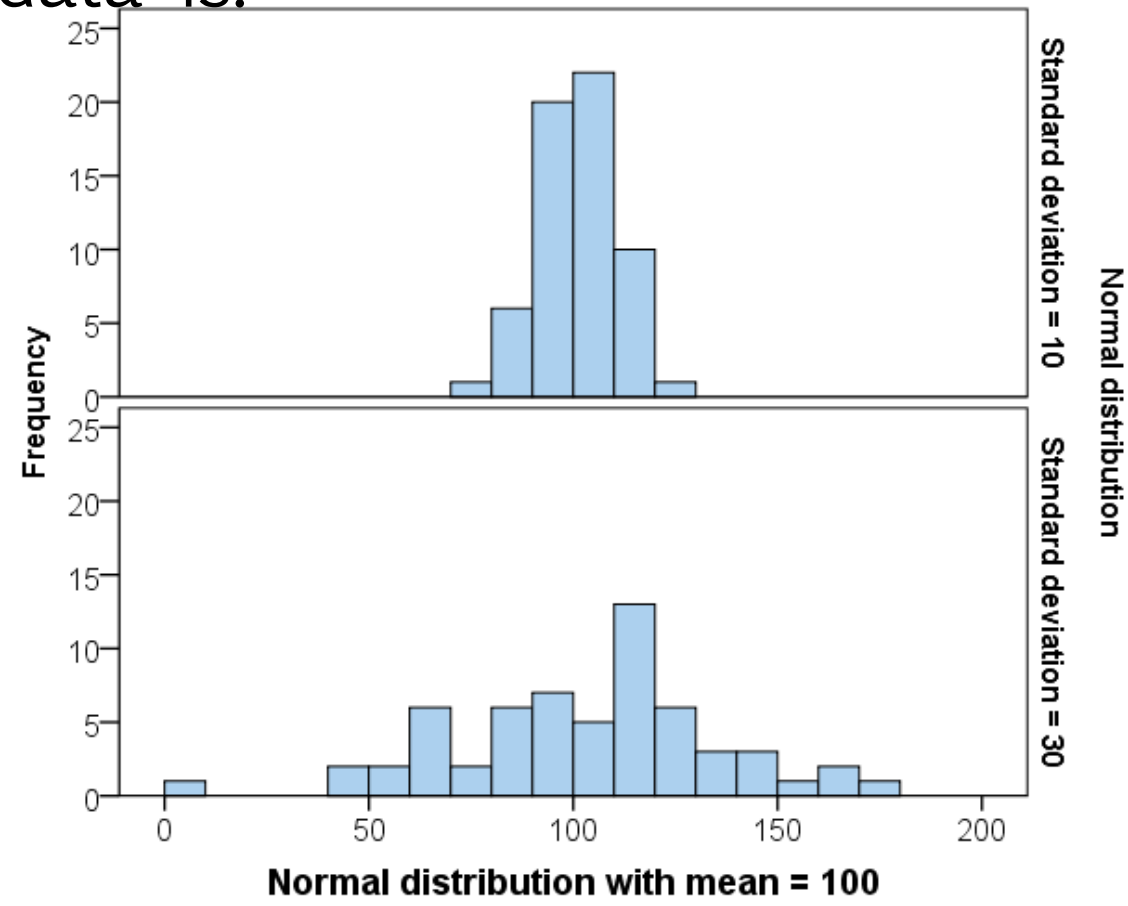
Large SD = very spread out data

Small SD = there is little variation from the mean

For exam scores, mean = 30.5, SD = 14.46

Interpretation of standard deviation

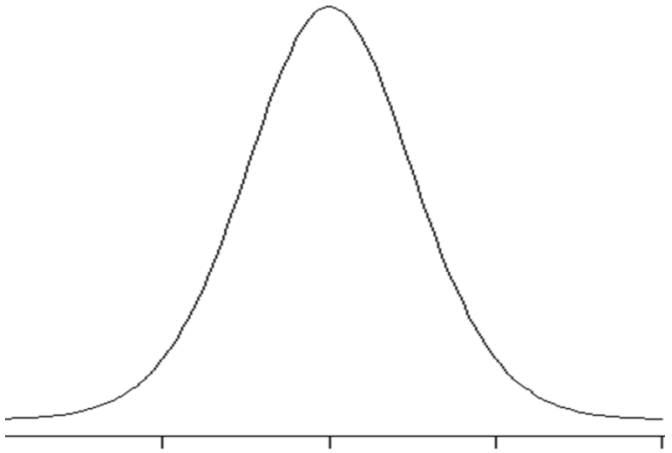
- The larger the standard deviation, the more spread out the data is.



Scale data

If have scale data it we often assume it follows a Normal distribution

To analyse it we often assume it follows a Normal distribution



Normal

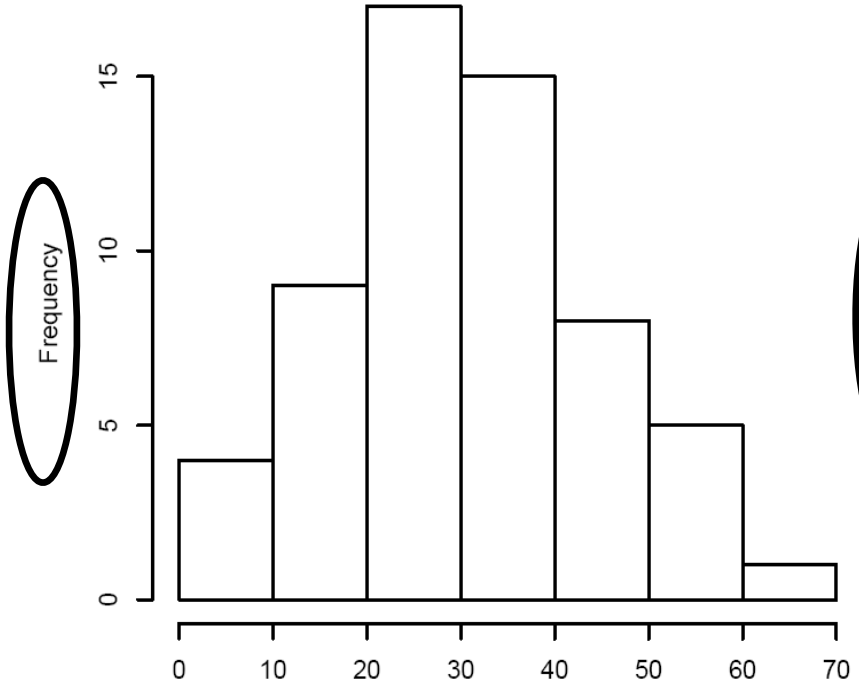
Discussion

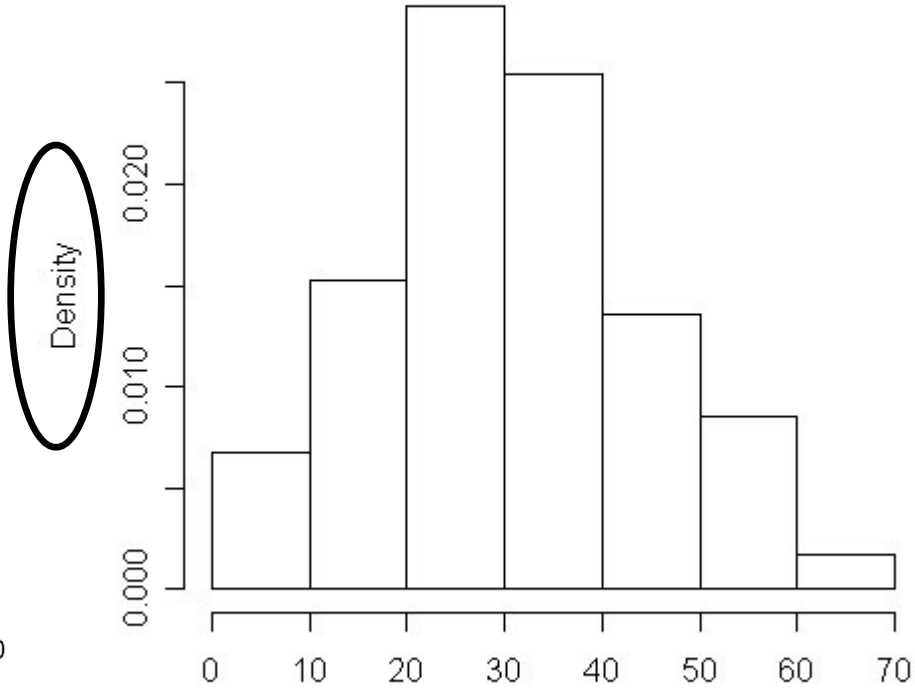
- How could you explain to a someone what we mean by data being assumed to follow a Normal Distribution?

Group Frequency Table

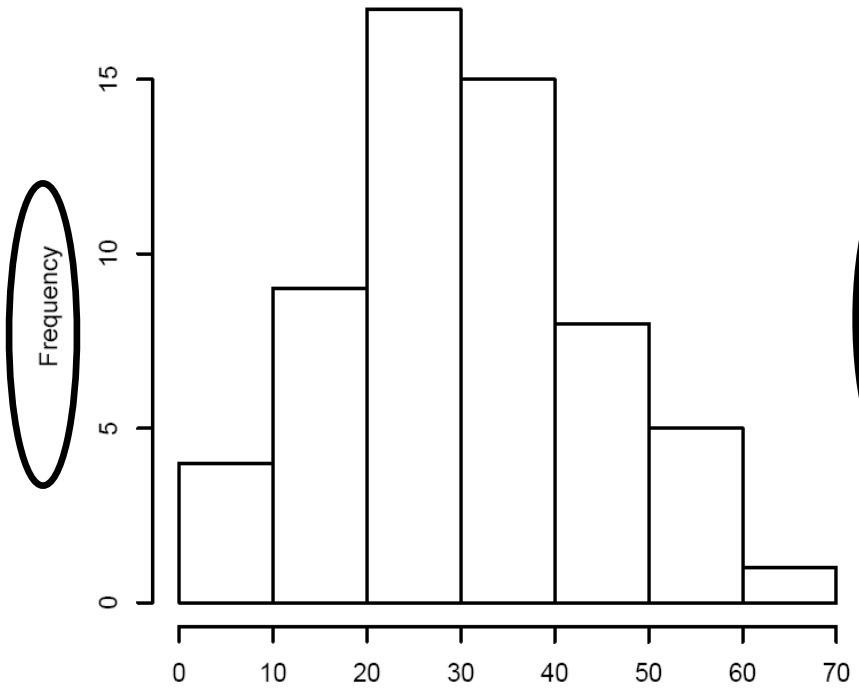
	Frequency	Percent
0 but less than 10	4	6.7
10 but less than 20	9	15.0
20 but less than 30	17	28.3
30 but less than 40	15	25.0
40 but less than 50	9	15.0
50 but less than 60	5	8.3
60 or over	1	1.7
Total	**60**	**100.0**

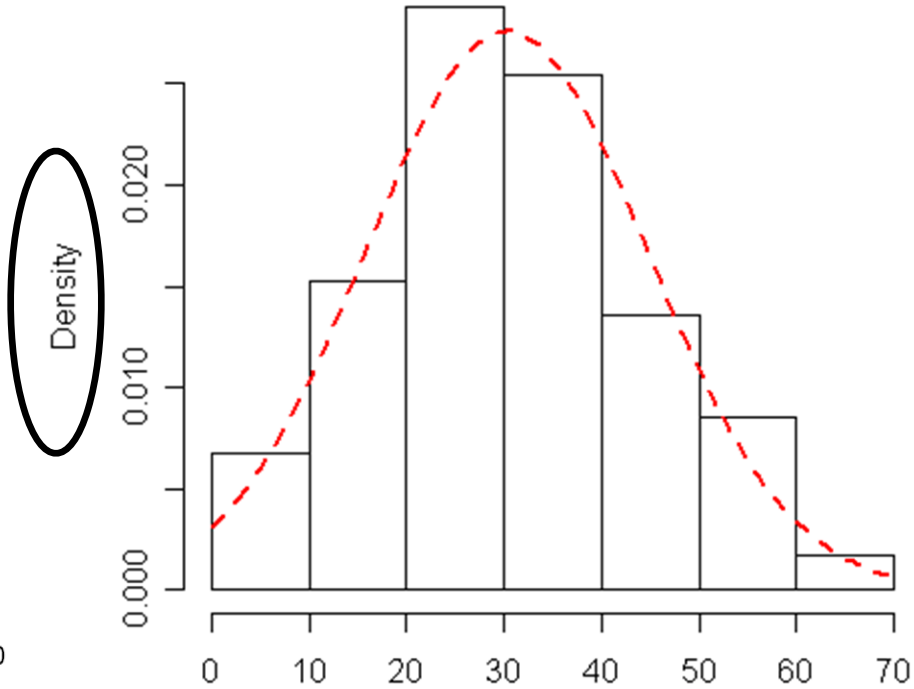
Histogram and Probability Distribution for Exam Marks Data



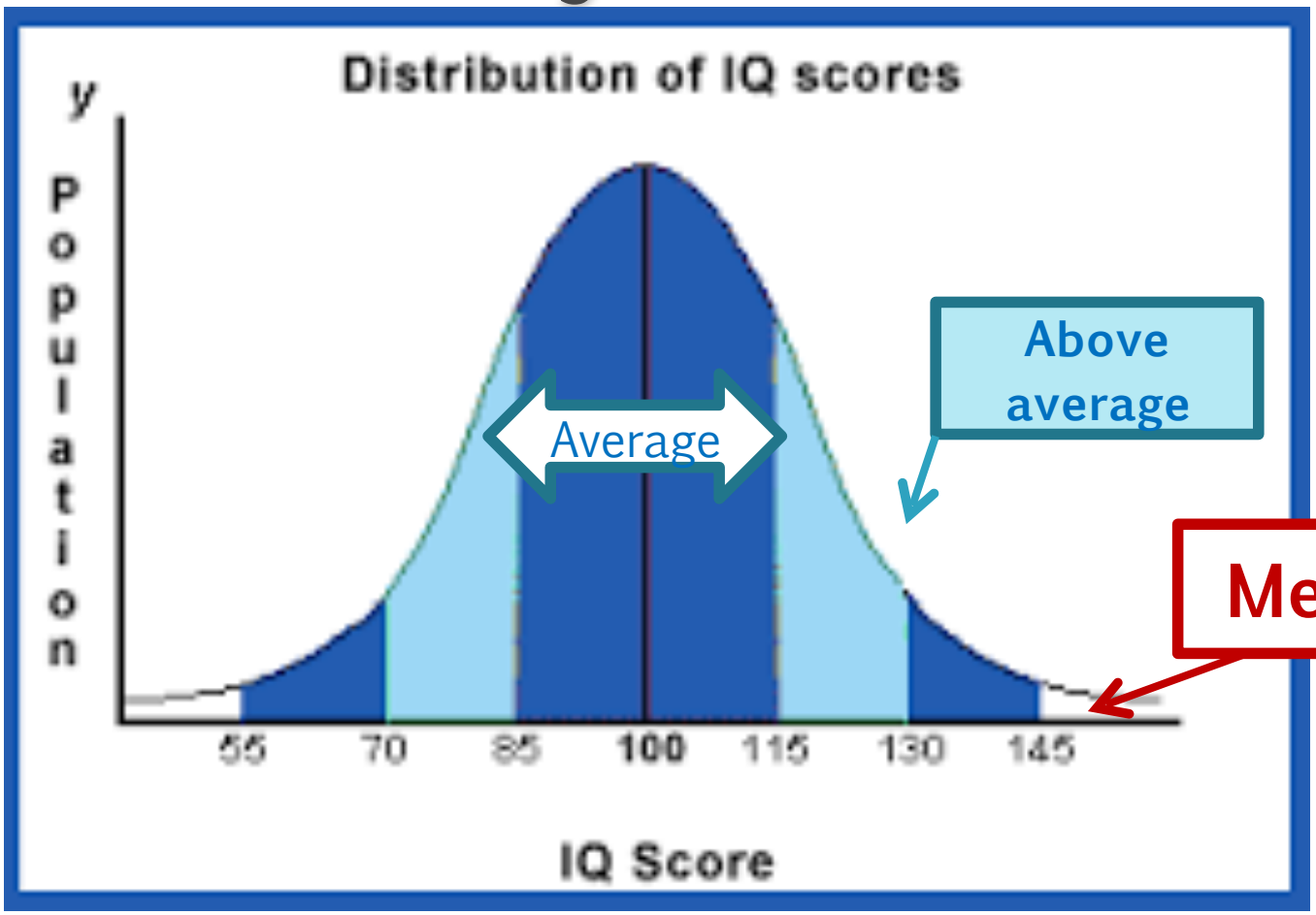


Histogram and Probability Distribution for Exam Marks Data



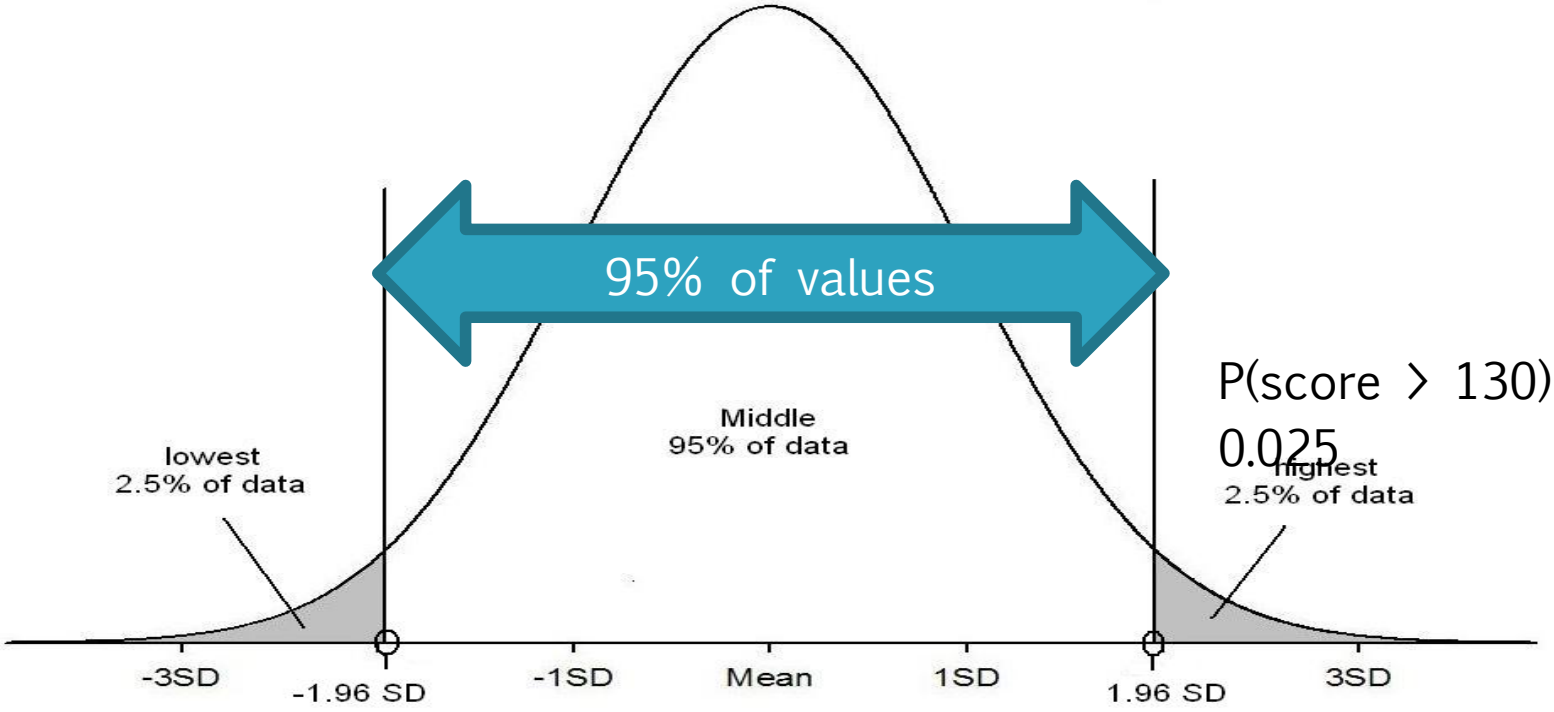


IQ is normally distributed



Mean = 100, SD = 15.3

95% 1.96 x SD's from the mean



P(score > 130) = 0.025

70 100 130

$$mean - (1.96 \times SD)$$

$$100 - (1.96 \times 15.3) = 70$$

$$mean + (1.96 \times SD)$$

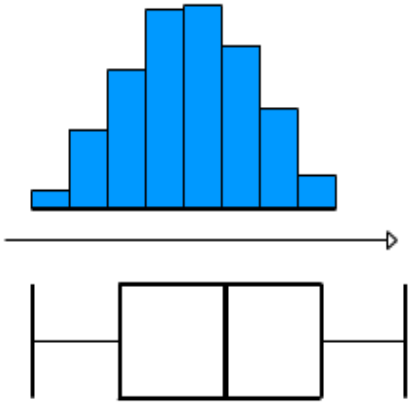
$$100 + (1.96 \times 15.3) = 130$$

95% of people have an IQ between 70 and 130

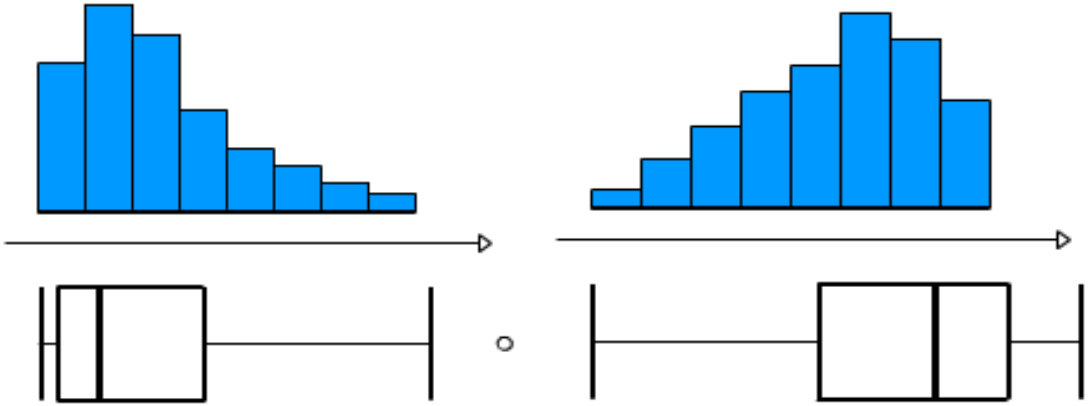
Assessing Normality

Charts can be used to **informally** assess whether data is:

Normally distributed



Or….Skewed



The mean and median are very different for skewed data.

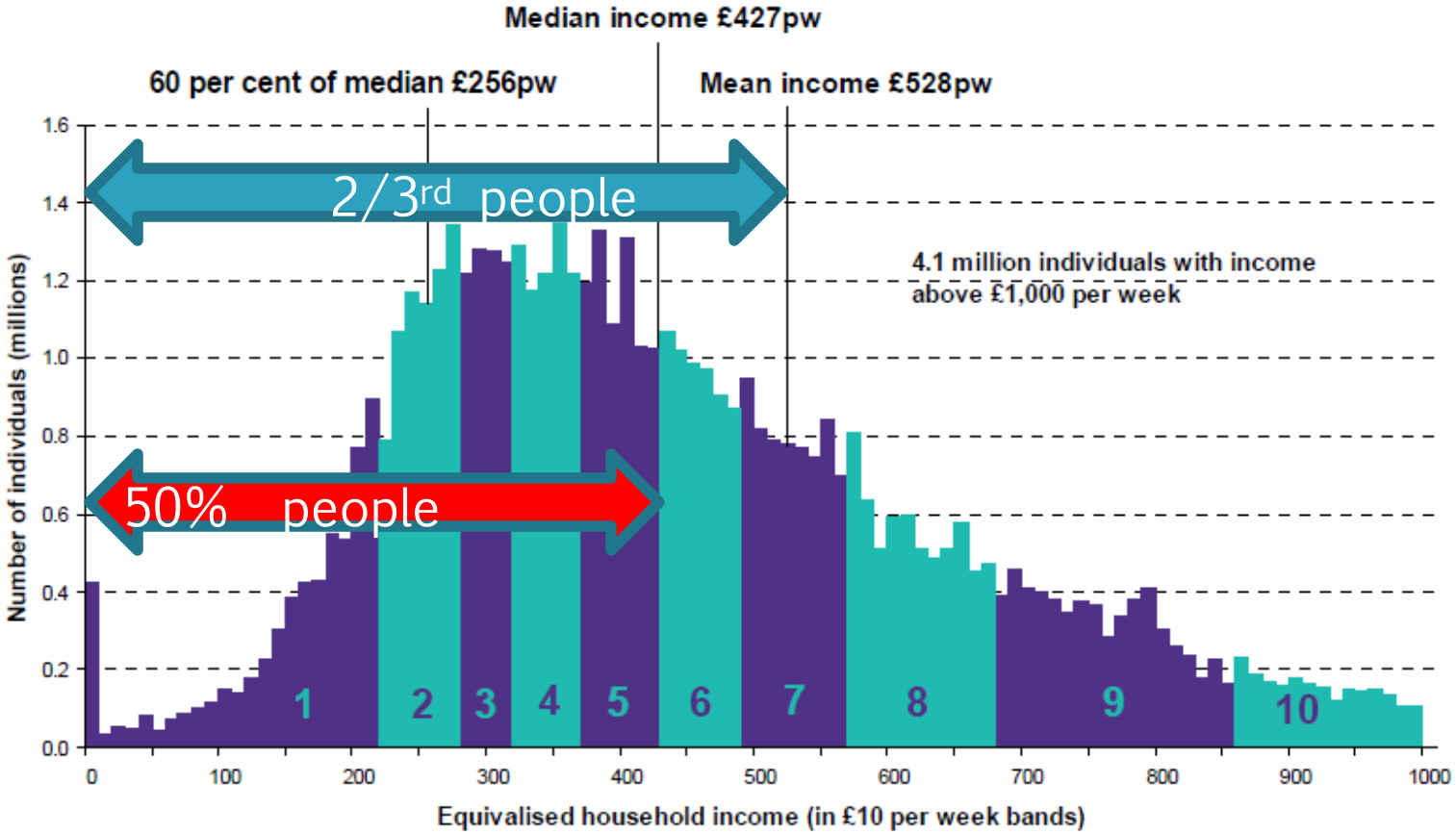
Discussion

Is the following statement:

"2 out of 3 people earn less than the average income"

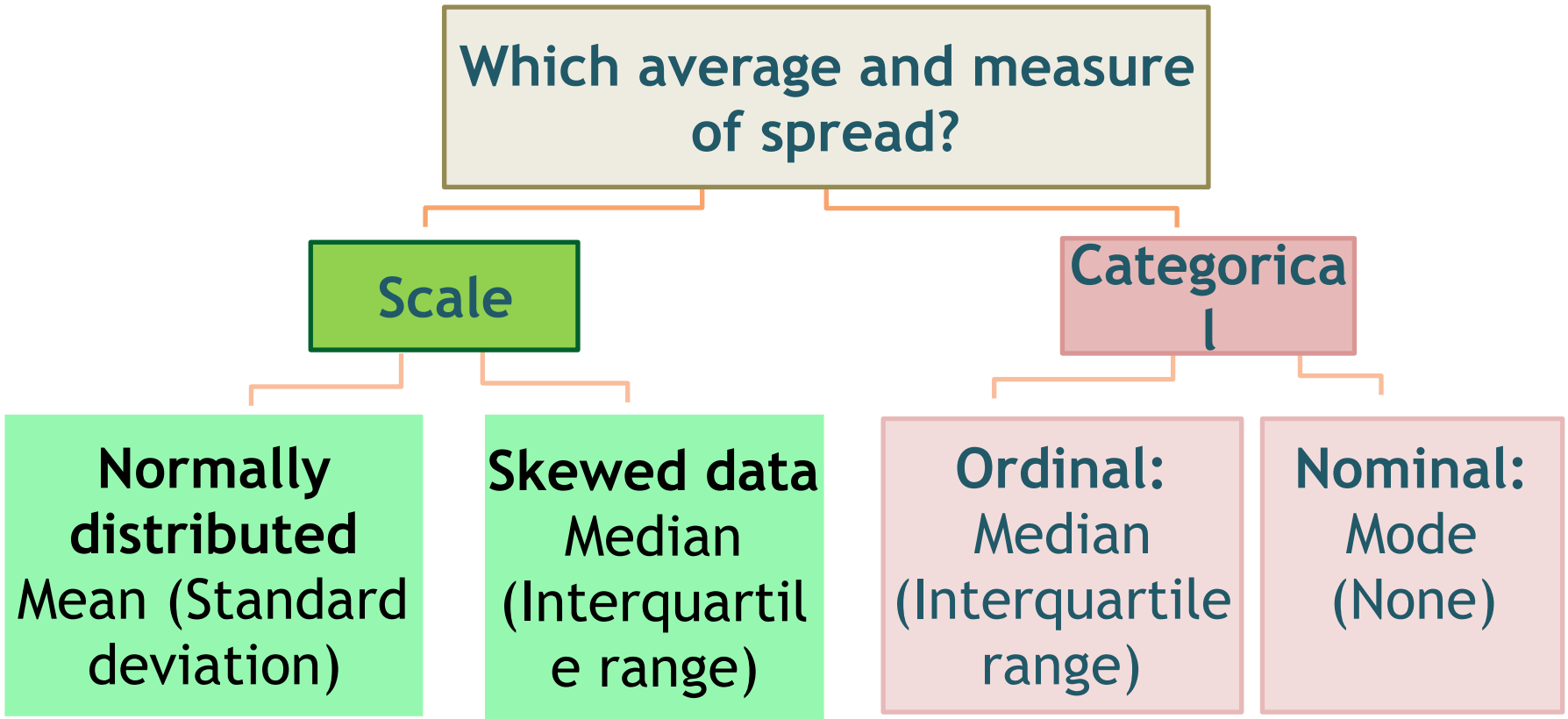
A. True
B. False
C. Don't know

Sometimes the median makes more sense!



Source: Households Below Average Income: An analysis of the income distribution1994/95 – 2011/12, Department for Work and Pensions

Choosing summary statistics



Which graph? Exercise

1st variable	Only 1 variable	Scale	Categorical
Scale	Histogram	Scatter plot	Box-plot/ Confidence interval plot
Categorical	Pie/ Bar	Box-plot/ Confidence interval plot	Stacked/ multiple bar chart

Which graph would you use when investigating:

1) Whether daily temperature and ice cream sales were related?

2) Comparison of mean reaction time for a group having alcohol and a group drinking water

Exercise: Ticket cost comparison

Summary statistics for cost of Titanic ticket by survival

Cost of ticket	Survived?	
	Died	Survived
Mean	23.4	49.4
Median	10.5	26
Standard Deviation	34.2	68.7
Interquartile range	18.2	46.6
Minimum	0	0
Maximum	263	512.33

a) **Is there a big difference in average ticket price by group?**

b) **Which group has data which is more spread out?**

c) **Is the data skewed?**

d) **Is the mean or median a better summary measure?**

Which graph? Solution

1st variable	Only 1 variable	Scale	Categorical
Scale	Histogram	Scatter plot	Box-plot/ Confidence interval plot
Categorical	Pie/ Bar	Box-plot/ Confidence interval plot	Stacked/ multiple bar chart

Which graph would you use when investigating:

1) Whether daily temperature and ice cream sales were related? Scatter
2) Comparison of mean reaction time for a group having alcohol and a group drinking water Boxplot or confidence interval plot

Exercise: Ticket cost comparison Solution

a) Is there a big difference in average ticket price by group?

The mean and median are much bigger in those who survived.

b) Which group has data which is more spread out?

The standard deviation and interquartile range are much bigger for those who survived so that data is more spread out

c) Is the data skewed?

Yes. The medians are much smaller than the means and the plots show the data is positively skewed.

d) Is the mean or median a better summary measure?

The median as the data is skewed

Hypothesis Testing

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 testing Framework What the text books might say!

- 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.

Discussion

- How could you help a someone understand what hypothesis testing is and why they need to use it?

Could try explaining things in the context of "The Court Case"?

- 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

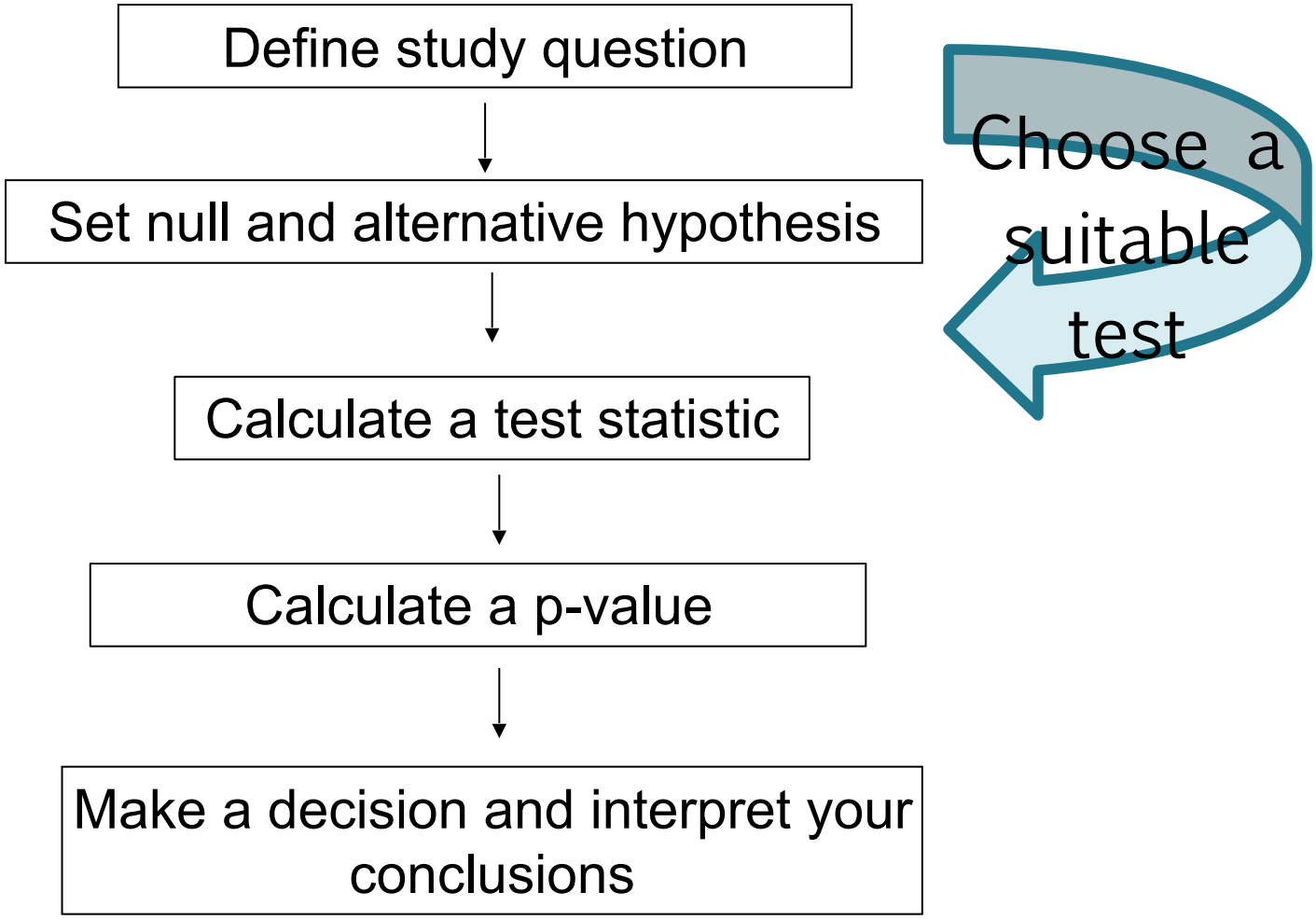
Controlled via sample size (=1-Power of test)

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) 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

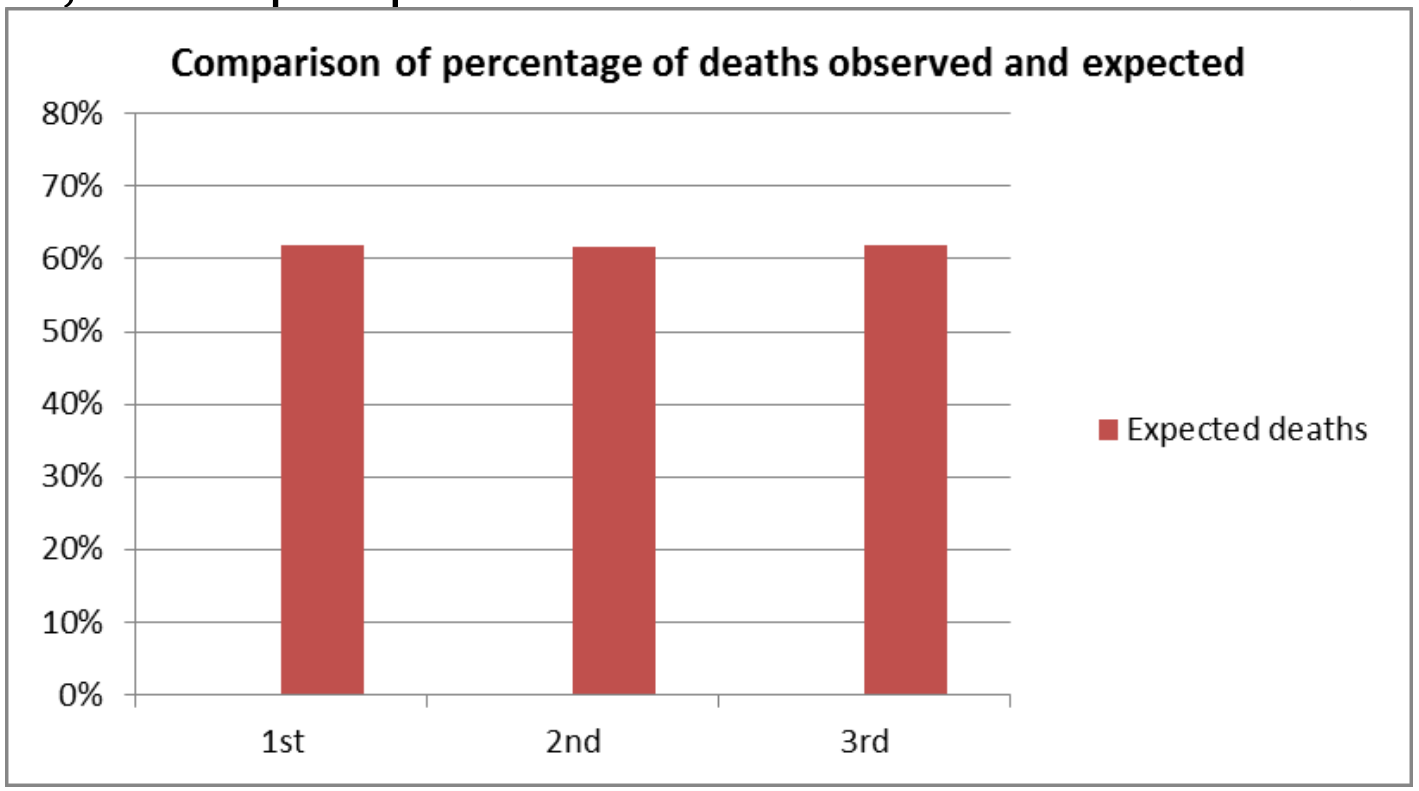
Class * Survived? Crosstabulation

Count

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

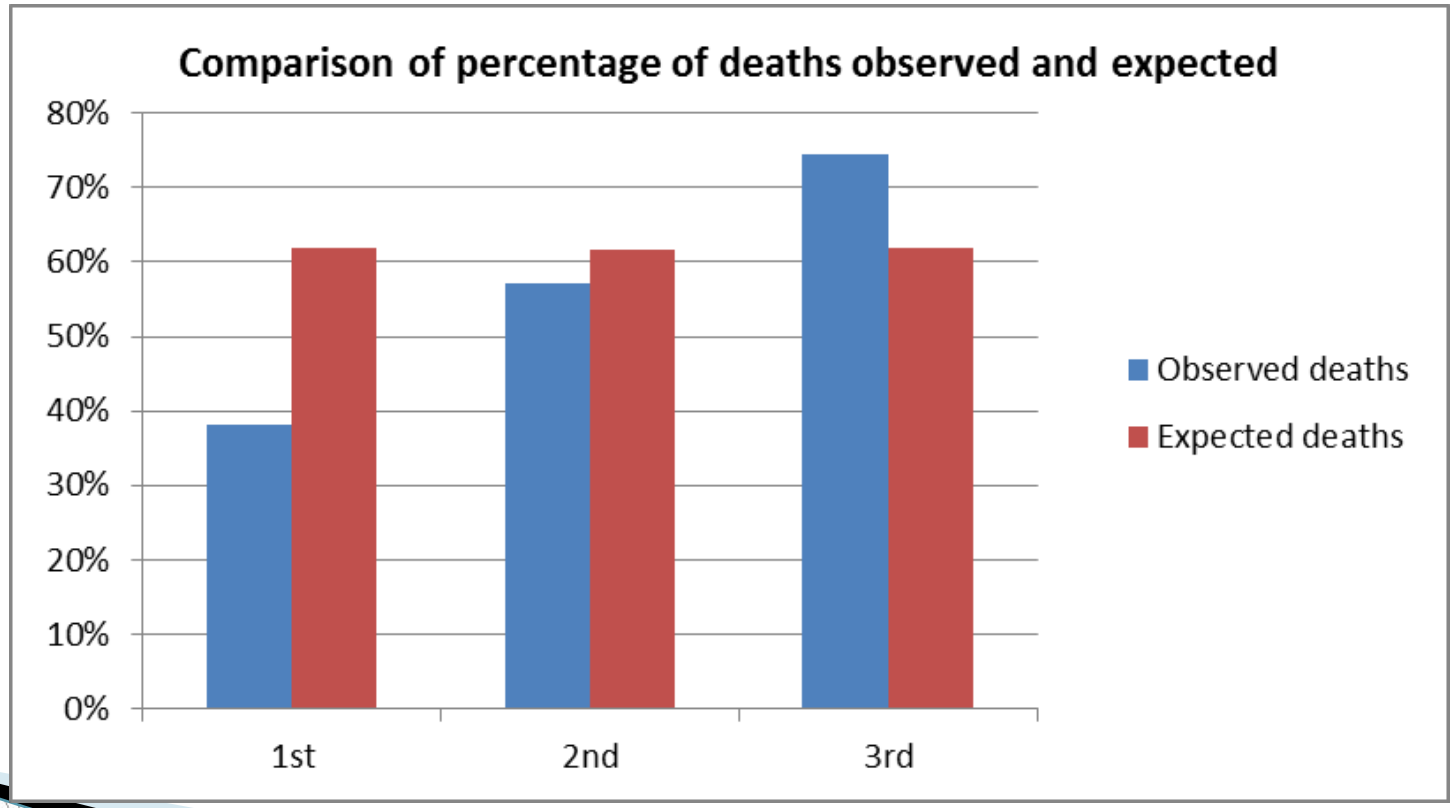
What would be expected if the null is true?

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

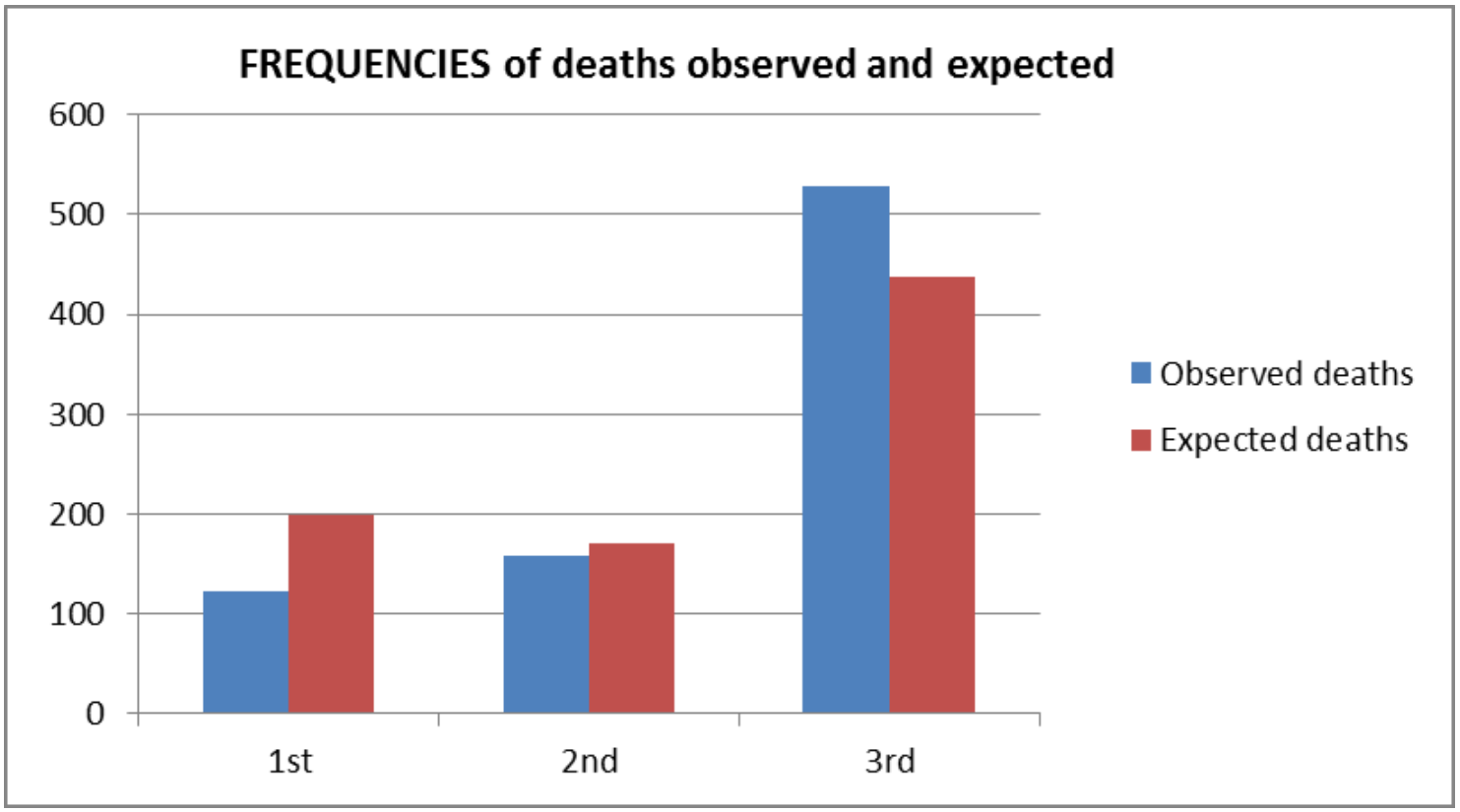


What would be expected if the null is true?

- 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)

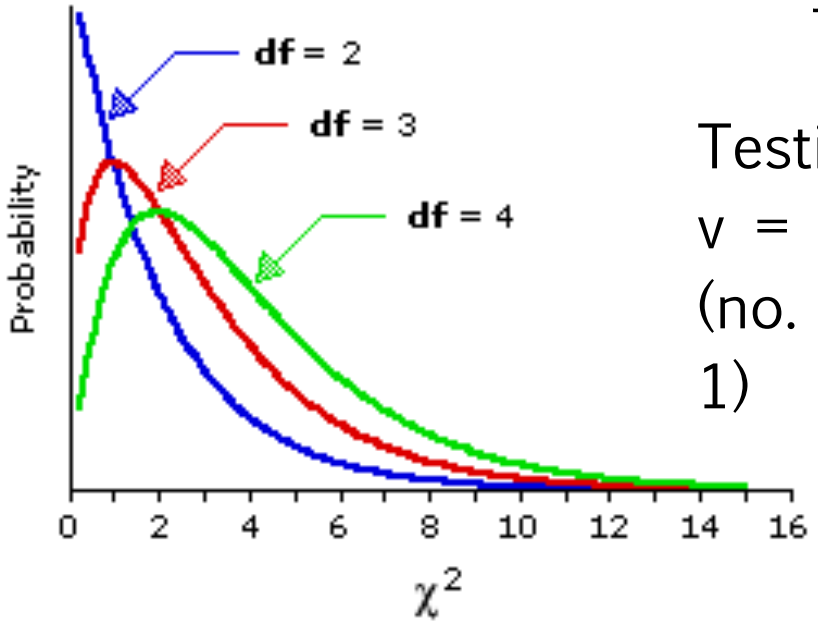
$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Hypothesis Testing: Decision Rule

- We can use statistical software to undertake a hypothesis test e.g. SPSS
- 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)

Chi squared distribution

- 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)

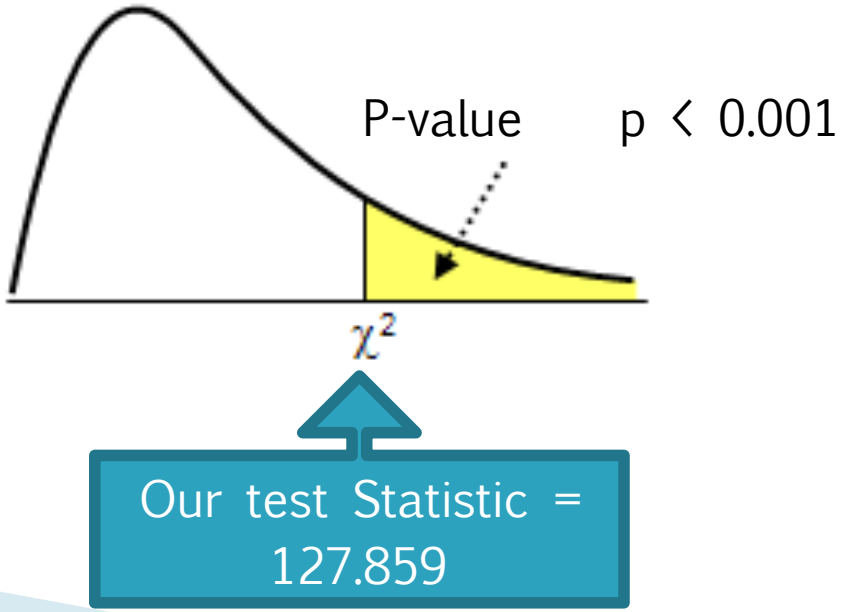
Note: One sample test:
v = df = outcomes – 1

What's a p-value?
The technical answer!

Probability of getting a test statistic at least as extreme as the one calculated **if the null is true**

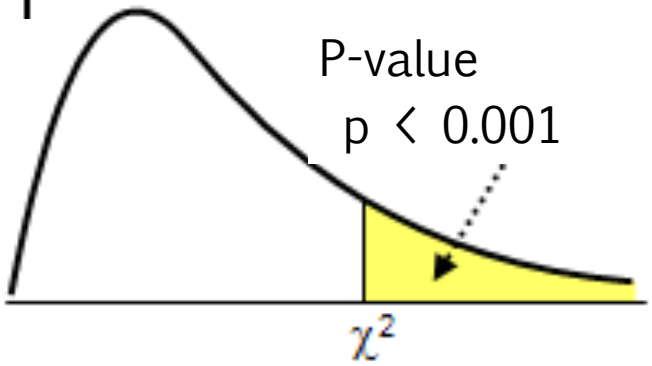
In Titanic example, the probability of getting a test statistic of 127.859 or above (**if the null is true**) is < 0.001

Distribution of test statistics



Interpretation

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	127.859[a]	2	.000



Test Statistic = 127.859

Since $p < 0.05$ we reject the null

There is evidence (χ^2_2=127.86, $p < 0.001$) to suggest that there is an association between class and survival

But… what is the nature of this association/relationship?

Titanic exercise

Were 'wealthy' people more likely to survive on board the Titanic?

Option 1:

- Choose the right percentages from the next slide to investigate
- Fill in the stacked bar chart with the chosen %'s
- Write a summary to go with the chart

Contingency tables exercise

Which percentages are better for investigating whether class had an effect on survival?

Column

Class * Survived? Crosstabulation

			Survived?		
			Died	Survived	Total
Class	1st	Count	123	200	323
		% within Survived?	15.2%	40.0%	24.7%
	2nd	Count	158	119	277
		% within Survived?	19.5%	23.8%	21.2%
	3rd	Count	528	181	709
		% within Survived?	65.3%	36.2%	54.2%
Total		Count	809	500	1309
		% within Survived?	100.0%	100.0%	100.0%

Row

Class * Survived? Crosstabulation

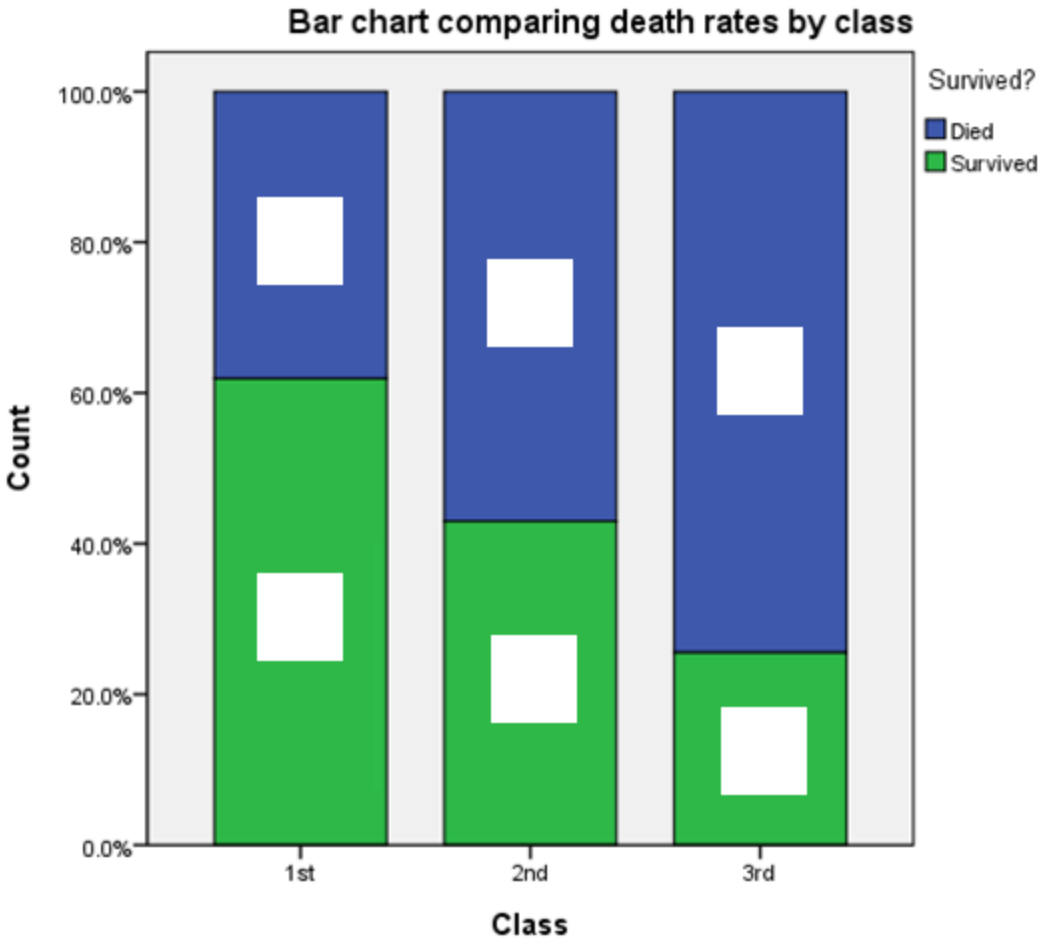
			Survived?		
			Died	Survived	Total
Class	1st	Count	123	200	323
		% within Class	38.1%	61.9%	100.0%
	2nd	Count	158	119	277
		% within Class	57.0%	43.0%	100.0%
	3rd	Count	528	181	709
		% within Class	74.5%	25.5%	100.0%
Total		Count	809	500	1309
		% within Class	61.8%	38.2%	100.0%

65.3% of those who died were in 3rd class

74.5% of those in 3rd class died

Did class affect survival? Question

Fill in the %'s on the stacked bar chart and interpret



Did class affect survival? Solution

%'s within each class are preferable due to different class frequencies

pclass * survived Crosstabulation

			survived		Total
			Died	Survived	
pclass	1st	Count	123	200	323
		% within pclass	38.1%	61.9%	100.0%
	2nd	Count	158	119	277
		% within pclass	57.0%	43.0%	100.0%
	3rd	Count	528	181	709
		% within pclass	74.5%	25.5%	100.0%
Total		Count	809	500	1309
		% within pclass	61.8%	38.2%	100.0%

Did class affect survival? Solution

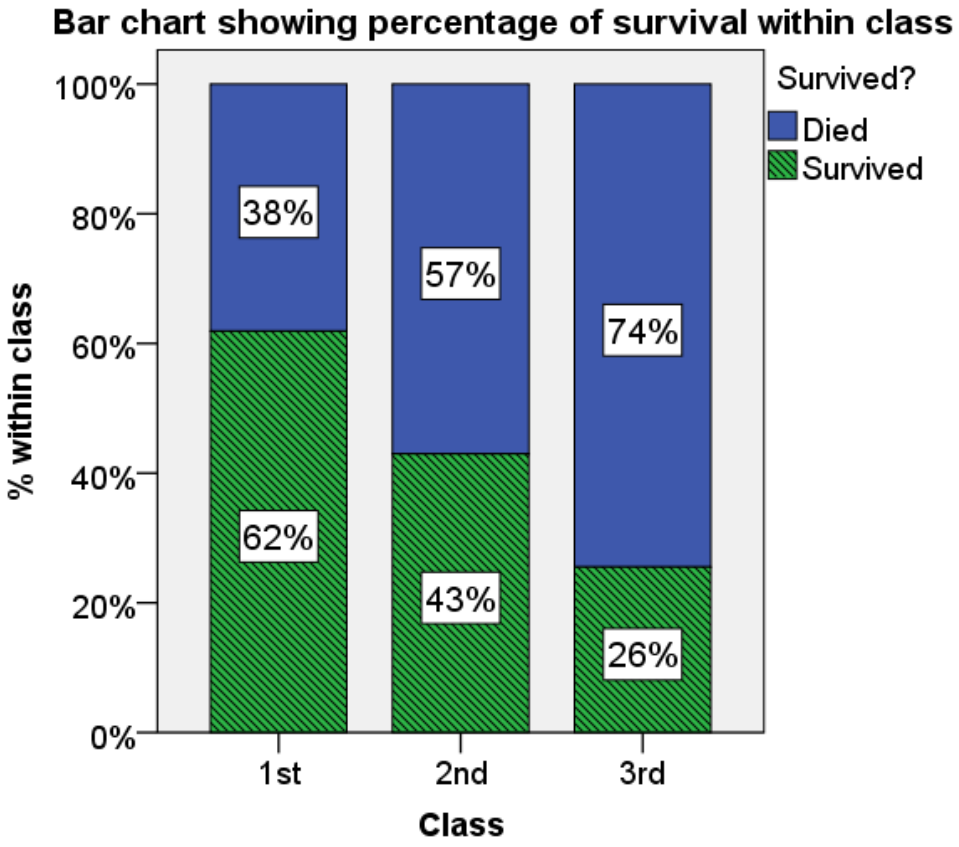


Figure 1: Bar chart showing % of passengers surviving within each class

Data collected on 1309 passengers aboard the Titanic was used to investigate whether class had an effect on chances of survival. There was evidence (χ^2_2=127.86, p < 0.001) to suggest that there is an association between class and survival.

Figure 1 shows that class and chances of survival were related. As class decreases, the percentage of those surviving also decreases from 62% in 1st Class to 26% in 3rd Class.

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

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	127.859[a]	2	.000
Likelihood Ratio	127.765	2	.000
Linear-by-Linear Association	127.709	1	.000
N of Valid Cases	1309		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 105.81.

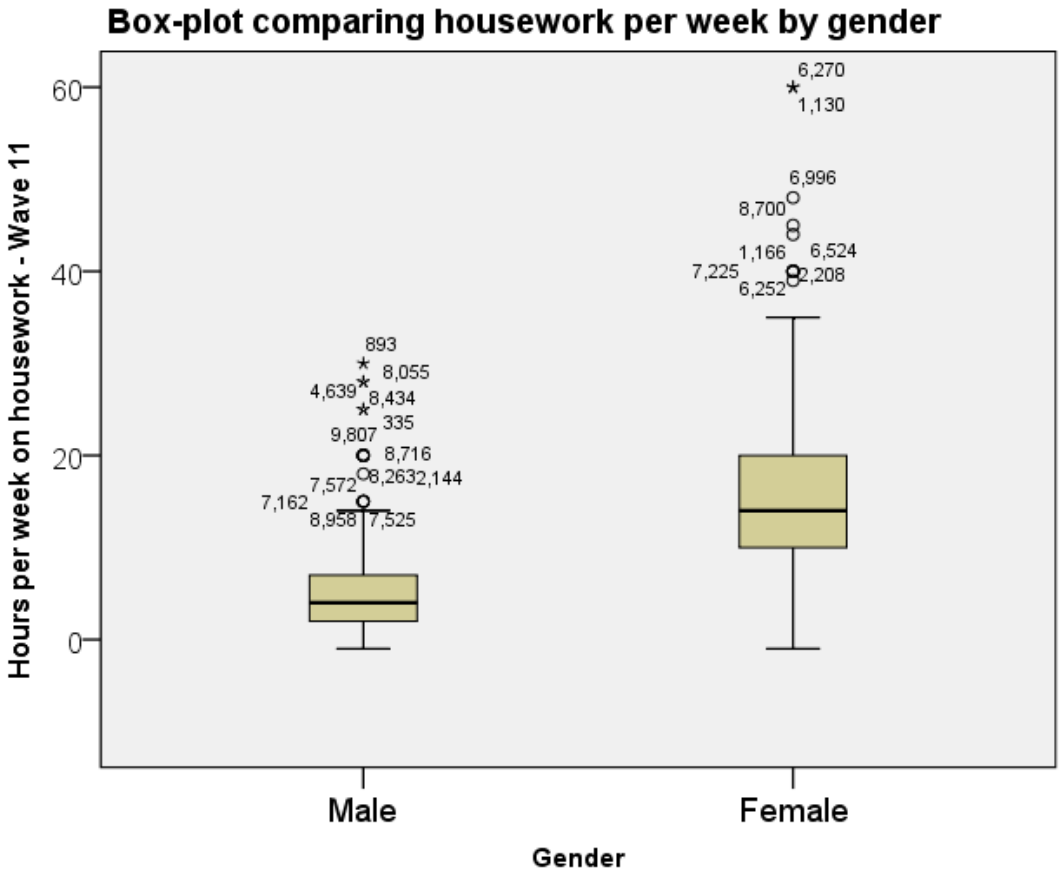
Low Cell Counts with the Chi-squared test

- Check no. of cells with EXPECTED counts less than 5
- SPSS reports the % of cells with an expected count <5
- If more than 20% then the test statistic does not approximate a chi-squared distribution very well
- If any expected cell counts are <1 then cannot use the chi-squared distribution
- In either case if have a 2x2 table use **Fishers' Exact test** (SPSS reports this for 2x2 tables)
- In 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

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

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

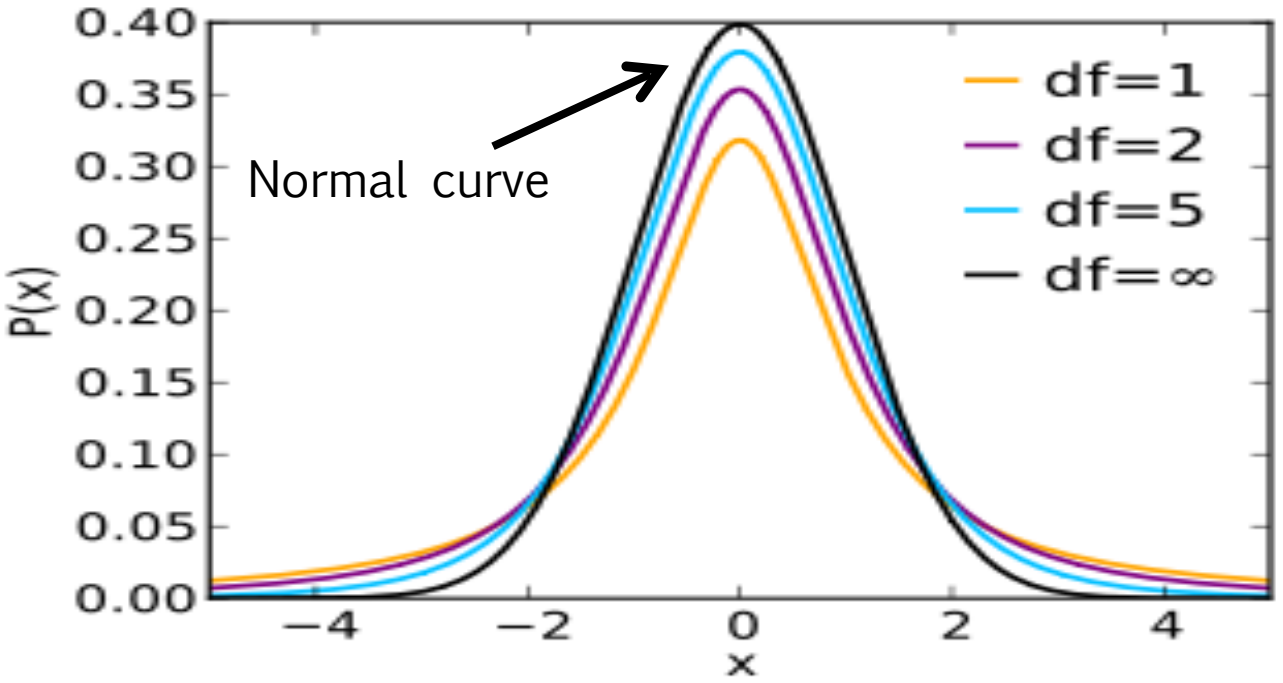
For an independent t-test,

$$v = n_{group1} + n_{group2} - 2$$

- 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



CAST e-books in statistics: t-distribution

http://cast.massey.ac.nz/collection_public.html

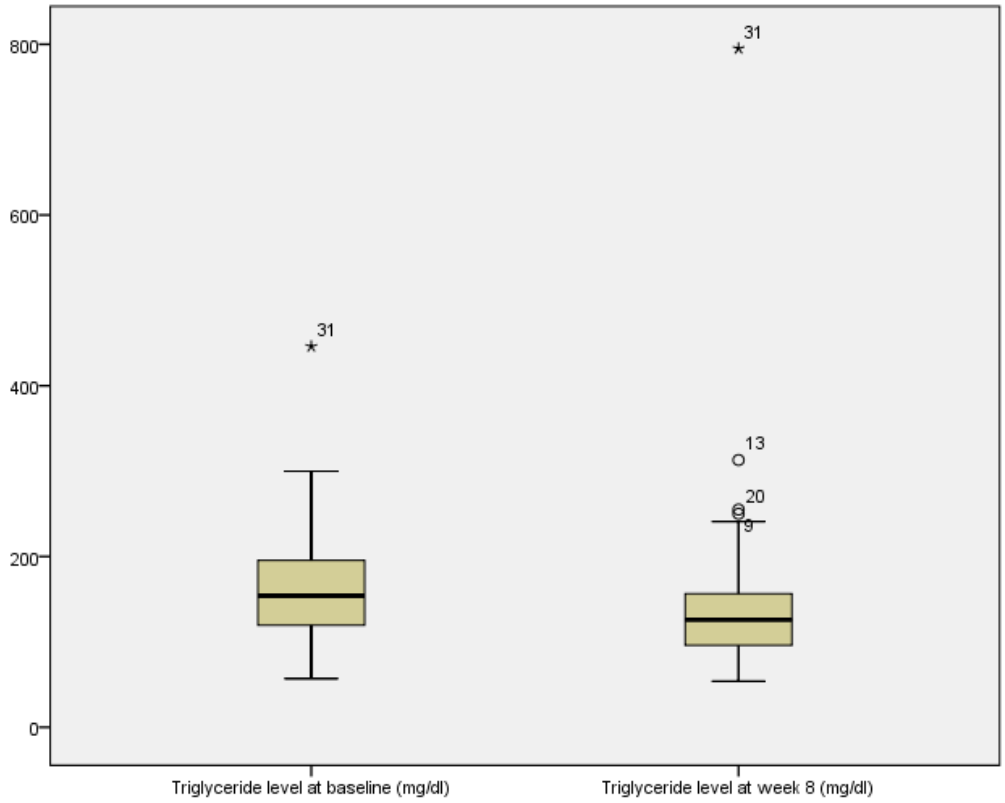
- Introductory e-book (general) and select
  10. Testing Hypotheses
      3. Tests about means
         4. The t-distribution
      5. The t-test for a mean

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.



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:

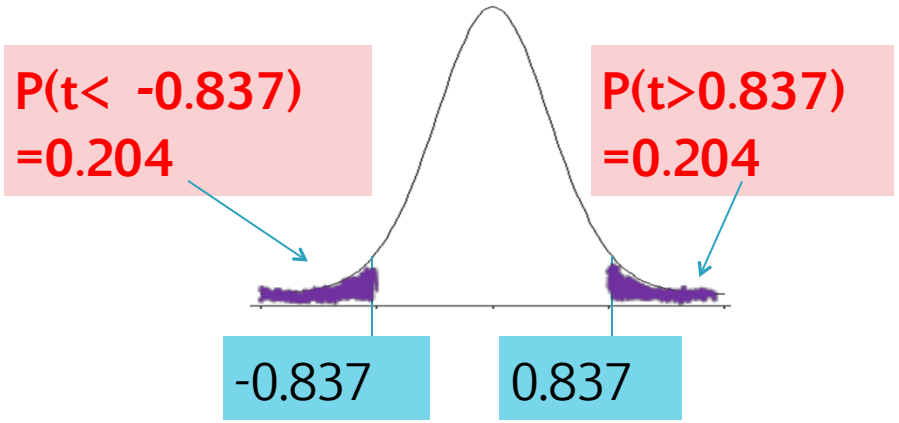
Example 1: Solution

	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

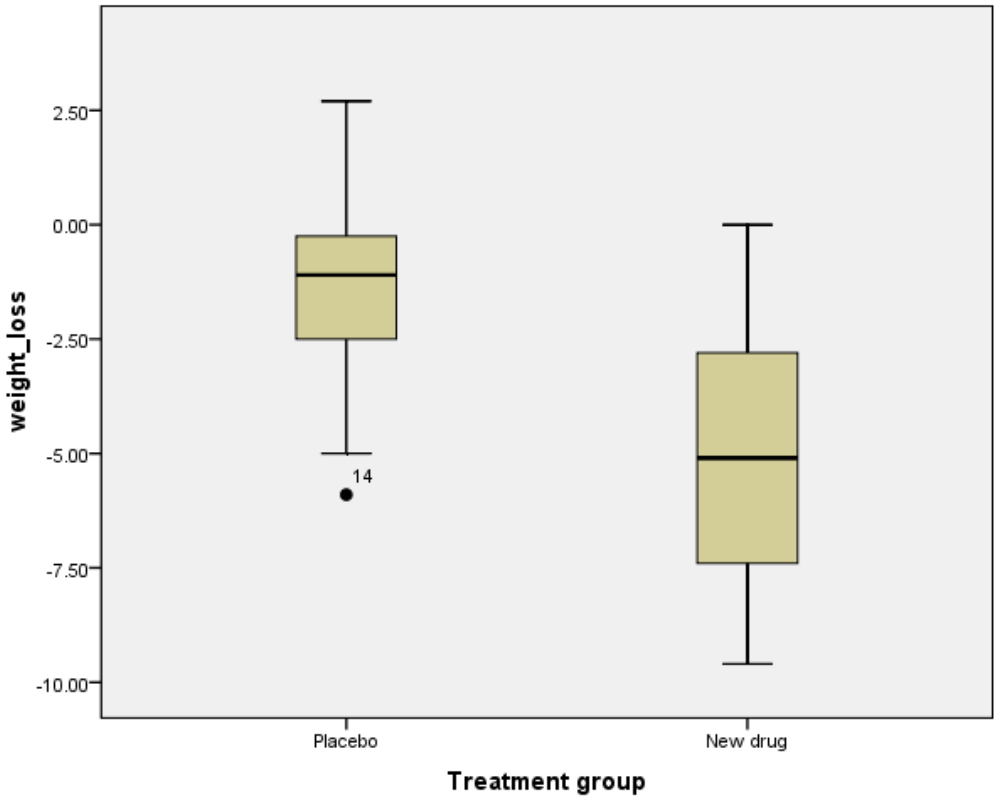
NO evidence of a difference in the mean triglyceride before and after treatment



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:

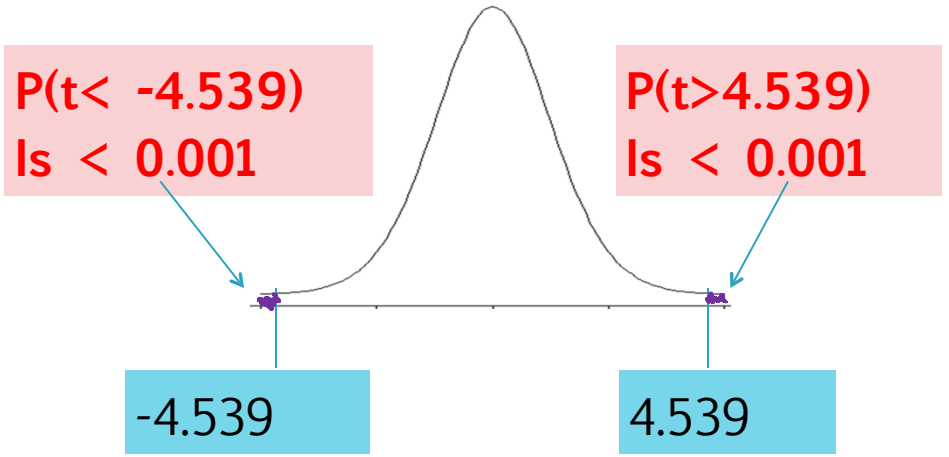
Ignore the shaded part of the output for now!

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

- Every test has assumptions
- Tutors quick guide shows assumptions for each test and what to do if those assumptions are not met

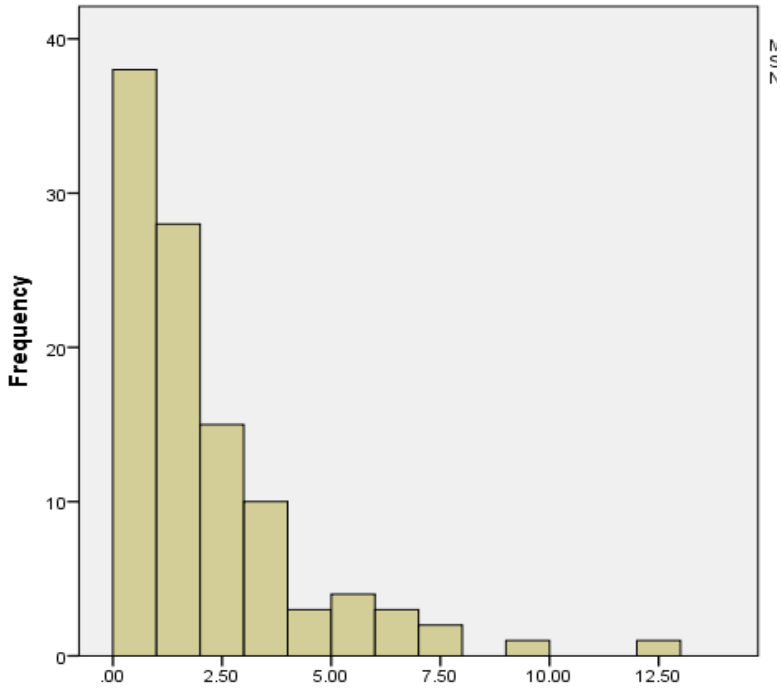
Assumptions in t-Tests

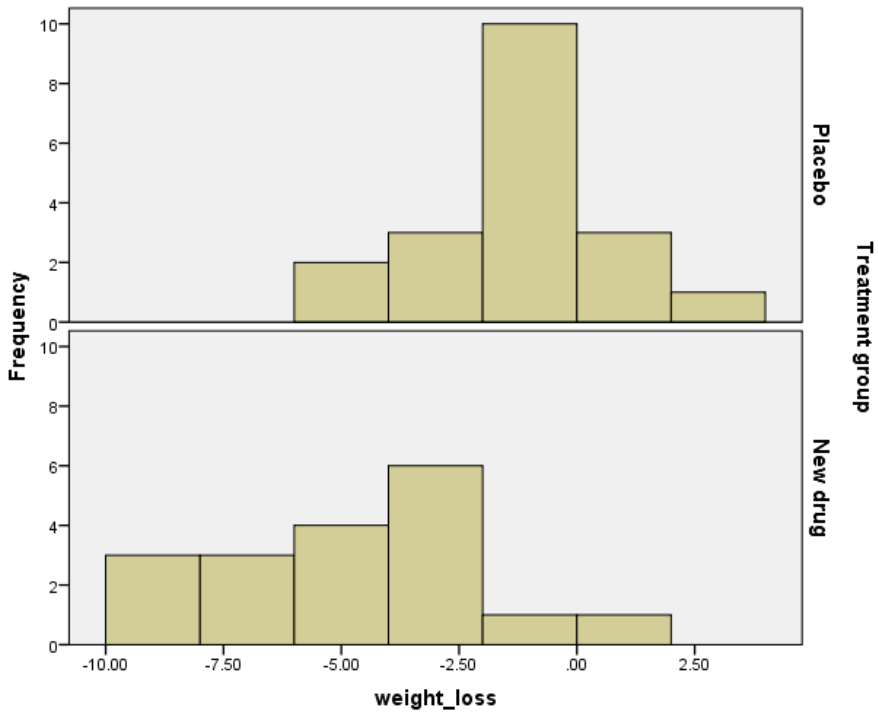
- **Normality:** Plot histograms
  - One plot of the paired differences for any paired data
  - Two (One for each group) for independent samples
  - Don't have to be perfect, just roughly symmetric

- **Equal Population variances:** Compare sample standard deviations
  - As a rough estimate, one should be no more than twice the other
  - Do an F-test (Levene's in SPSS) 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

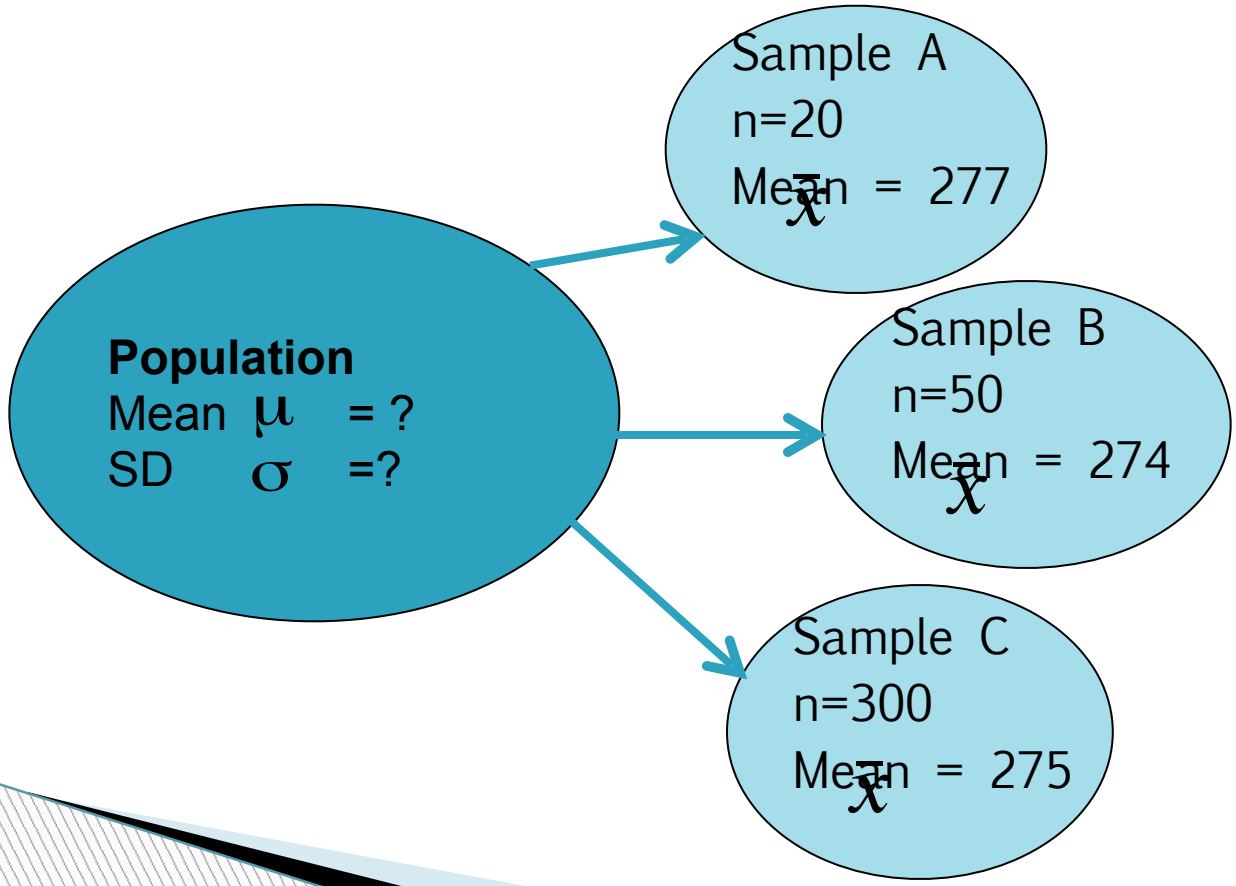
i.e. we can assume equal variances ☺️

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

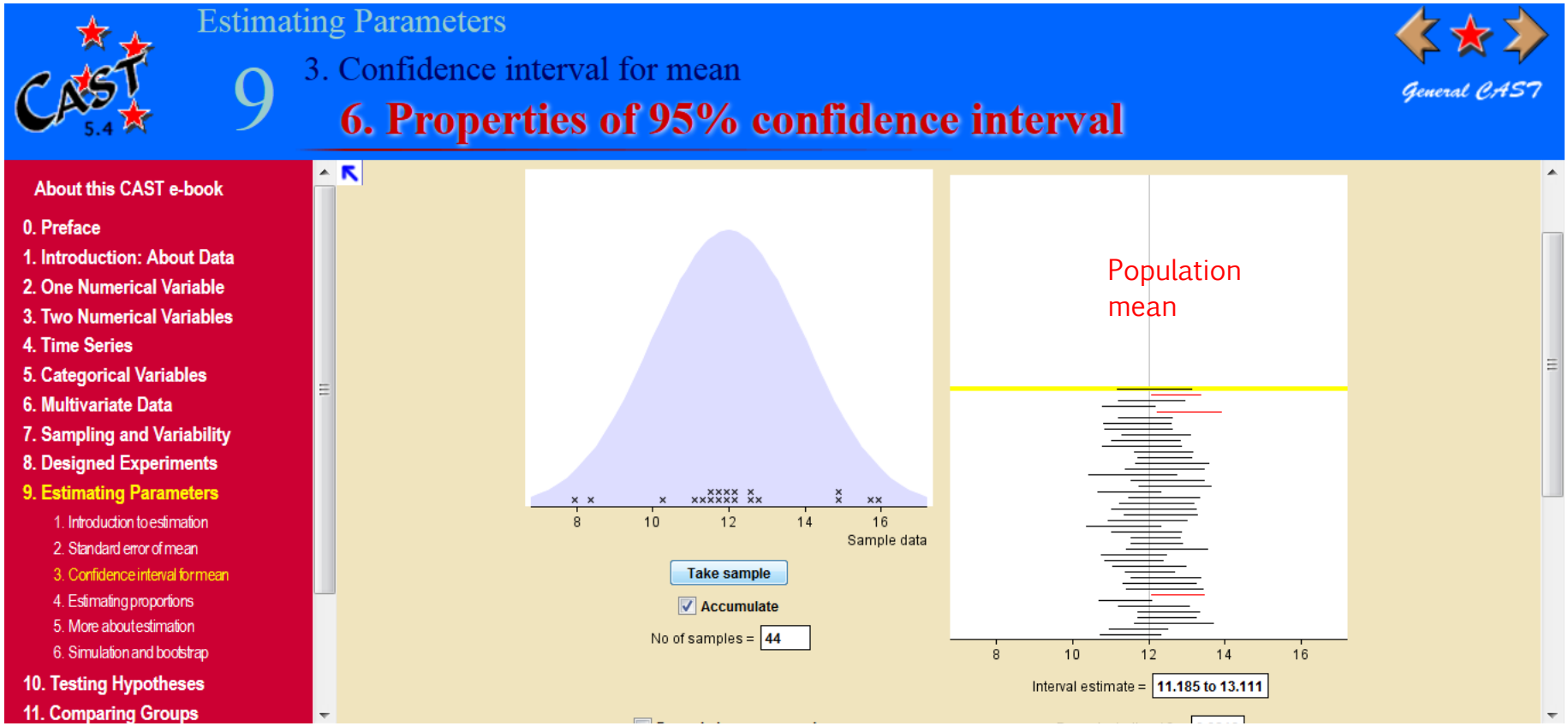
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Triglyceride level at week 8 (mg/dl) - Triglyceride level at baseline (mg/dl)	-11.371	80.360	13.583	-38.976	16.233	-.837	34	.408

CAST e-books in statistics: Confidence Intervals

http://cast.massey.ac.nz/collection_public.html

- Choose introductory e-book (general) and select
    9. Estimating Parameters
        3. Conf. Interval for mean
            6. Properties of 95% C.I.

Confidence Interval simulation from CAST



Seven do not include 141.1mmHg - we would expect that the 95% CI will not include the true population mean 5% of the time

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

Exercise: Solution

- Discuss what the interpretation is for the confidence interval from Example 2 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.

This is always positive hence the hypothesis test rejected the null that the difference is zero

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

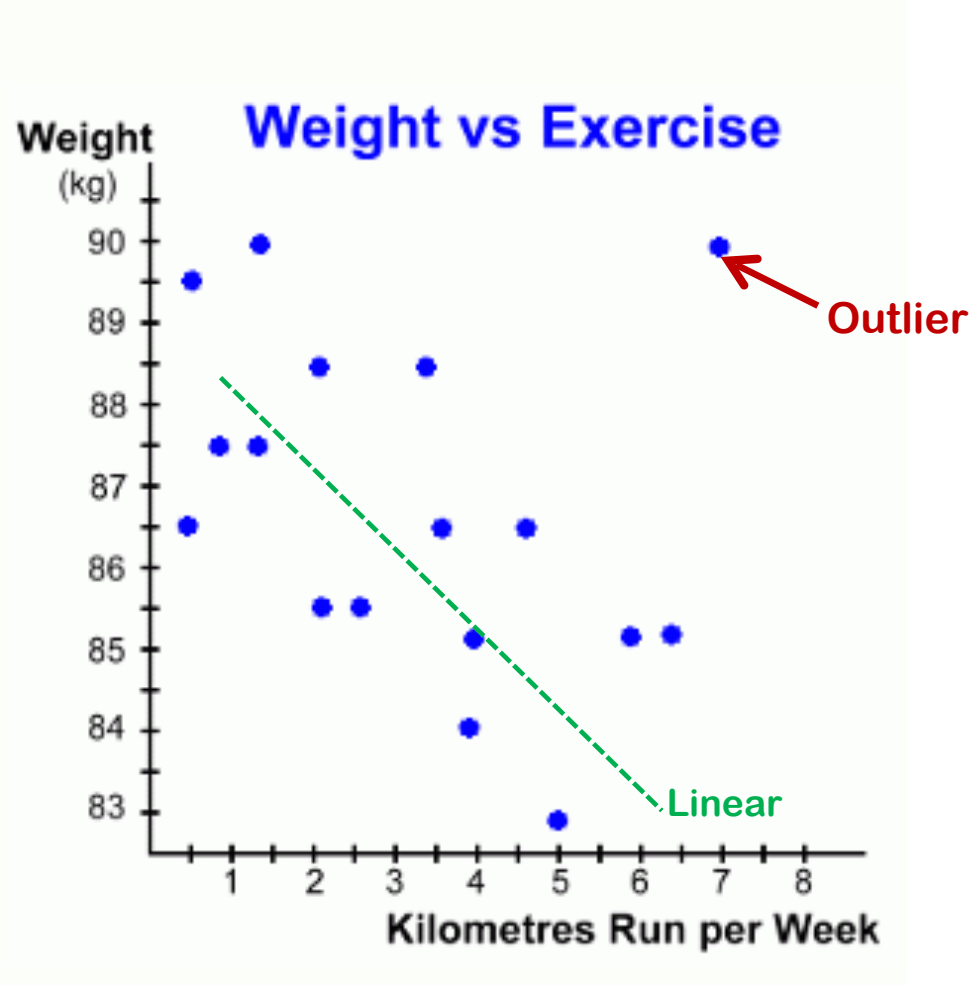
Tests for a relationship between two categorical variables

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

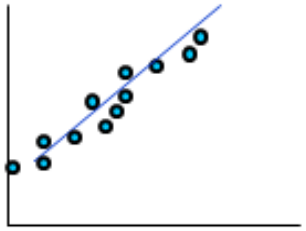


Correlation Coefficient r

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

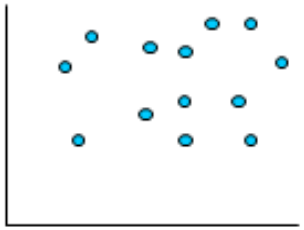
-1 r 1

Strong positive linear relationship



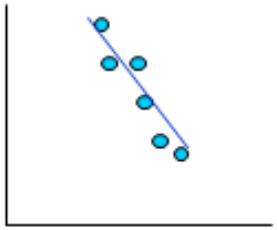
r = 0.9

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

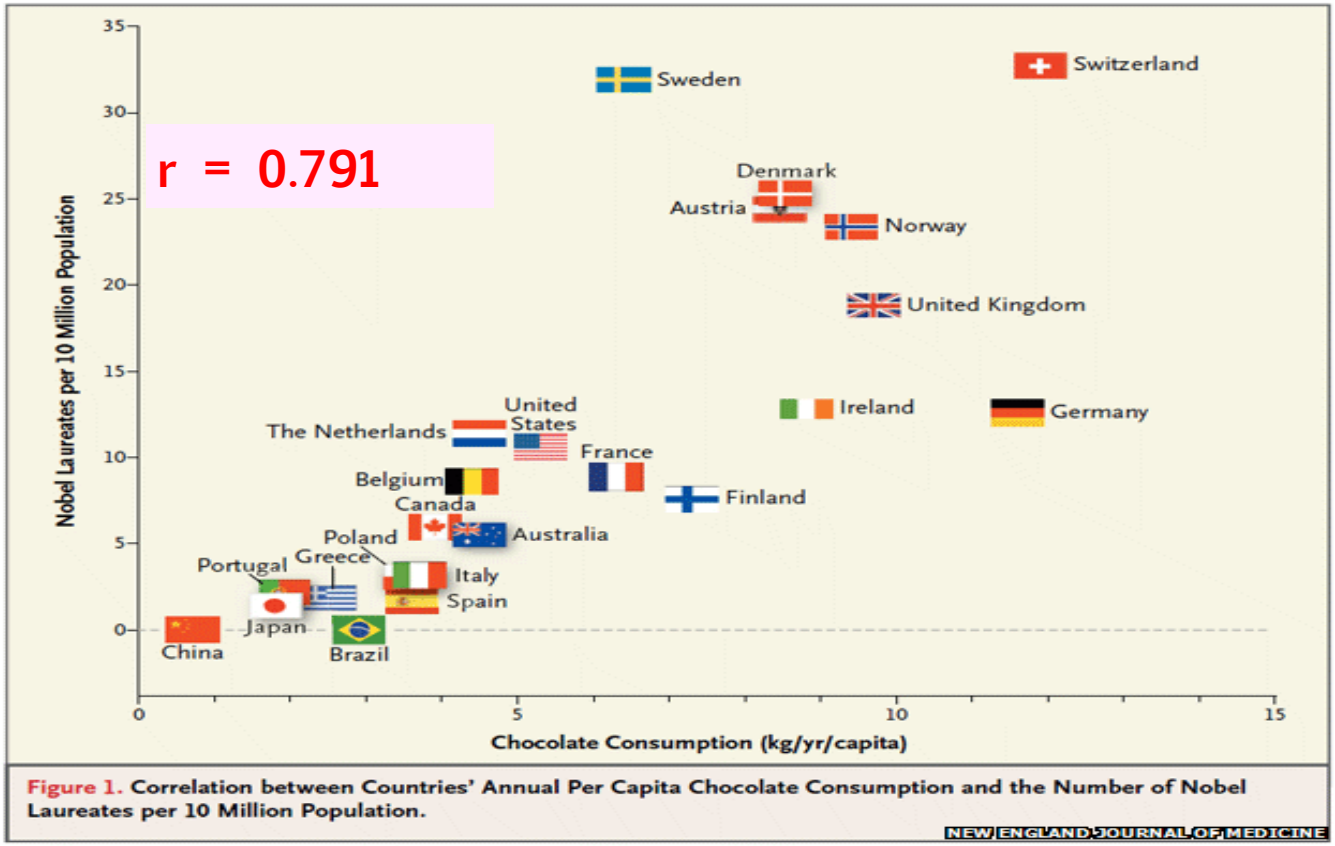


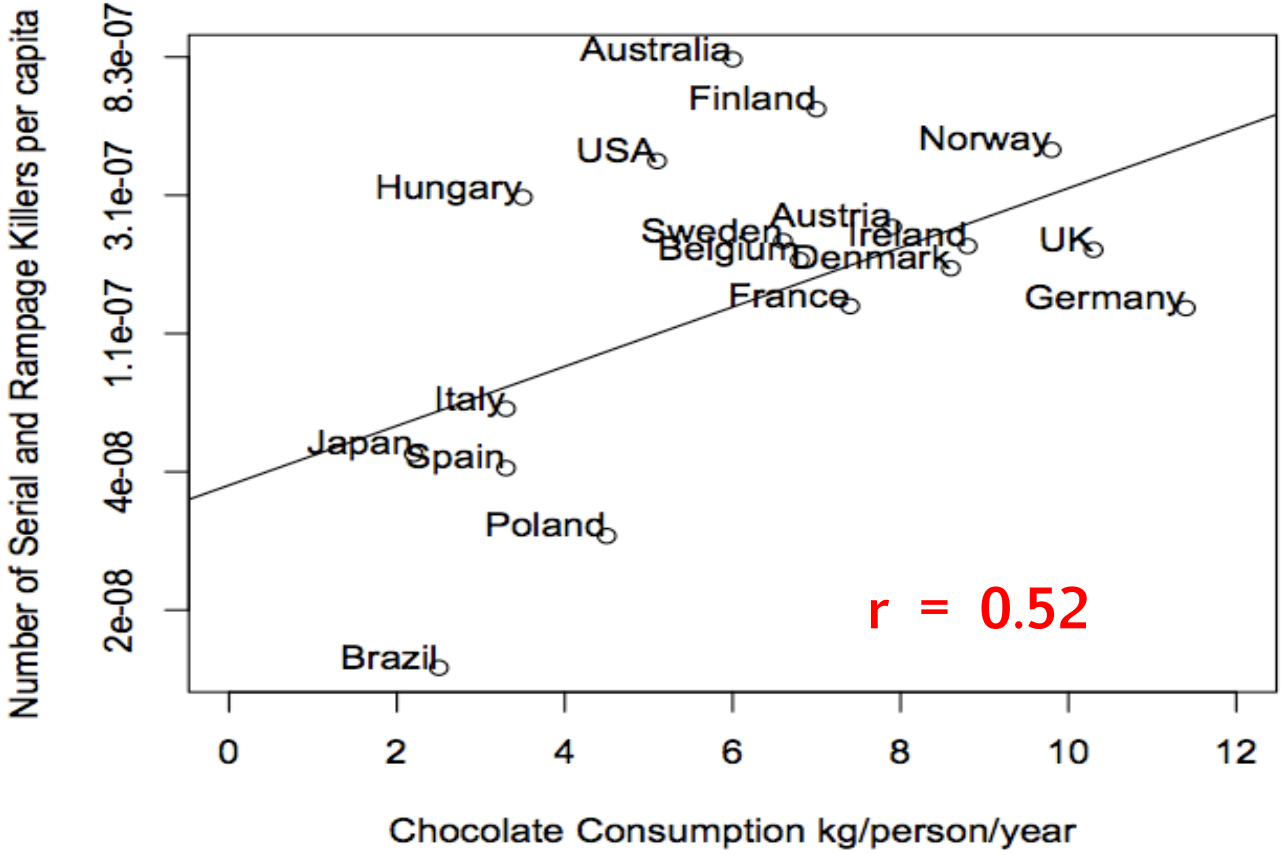
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

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

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!

It is highly influenced by the number of observations 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

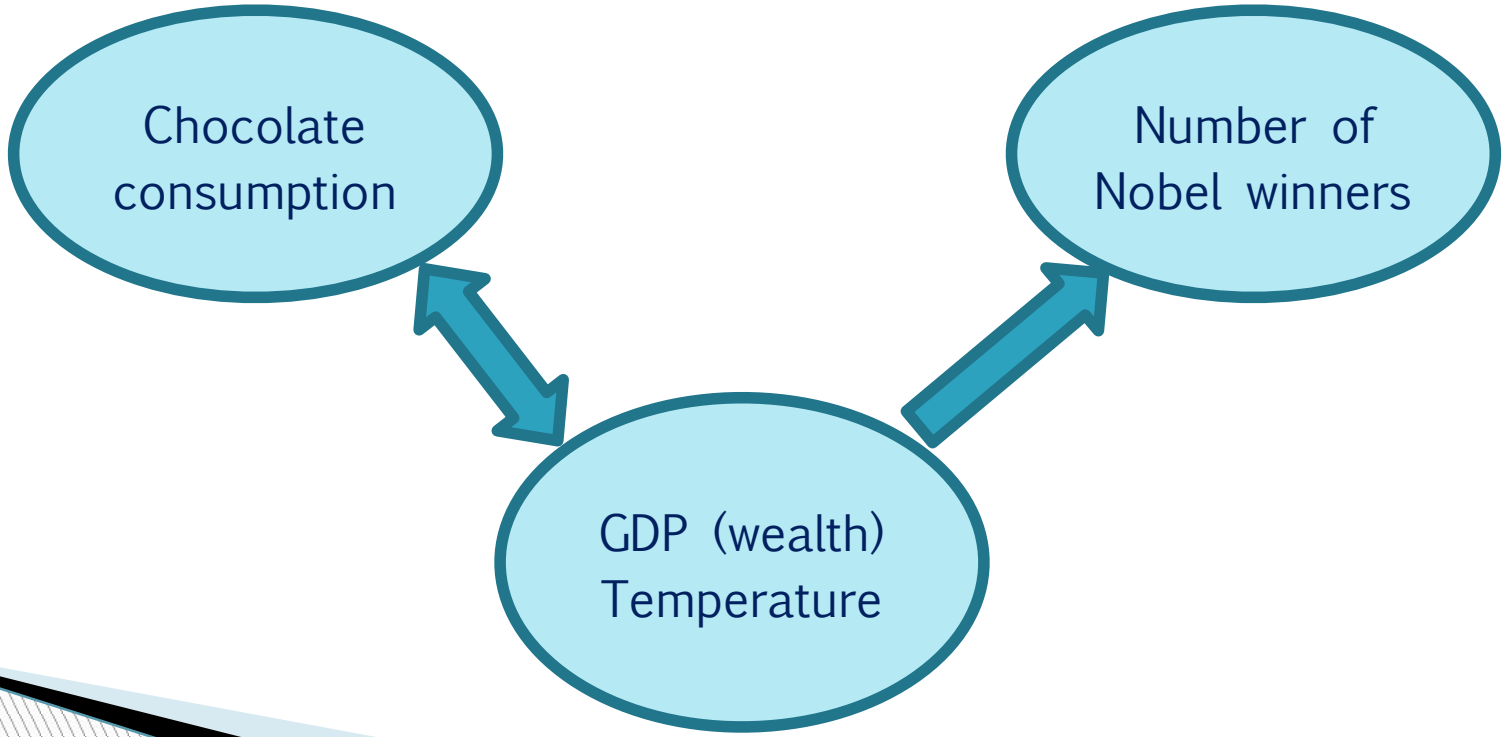
Relationship	Correlation
Average IQ and chocolate consumption	0.27
Road fatalities and Nobel winners	0.55
Gross Domestic Product and Nobel winners	0.7
Mean temperature and Nobel winners	-0.6

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?



Dataset for today

- 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

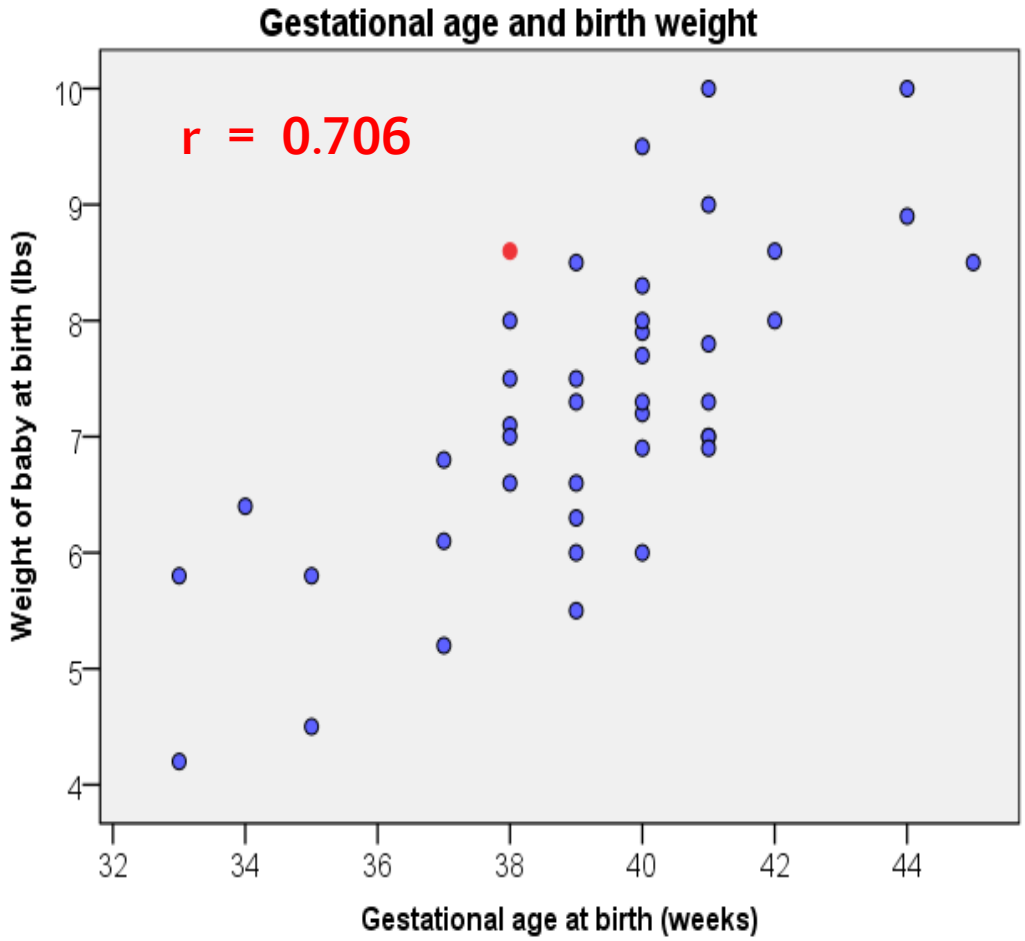
Mother smokes = 1

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.

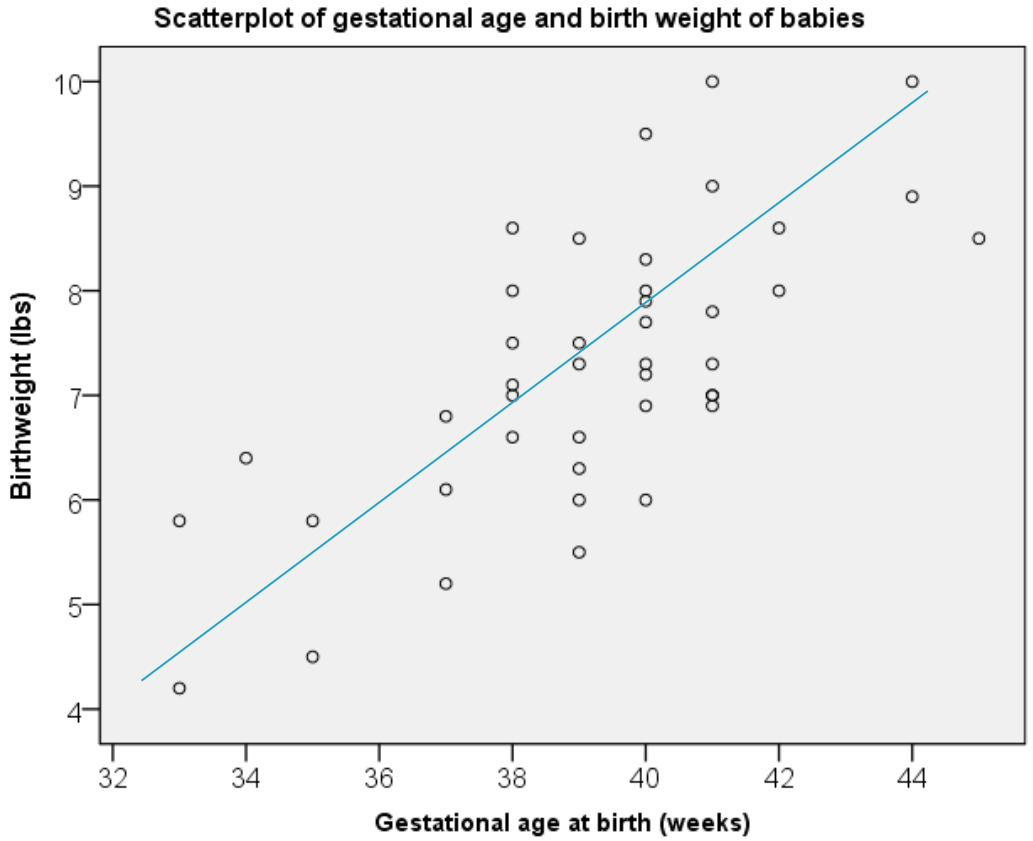
b) Draw a line of best fit through the data (with roughly half the points above and half below)



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

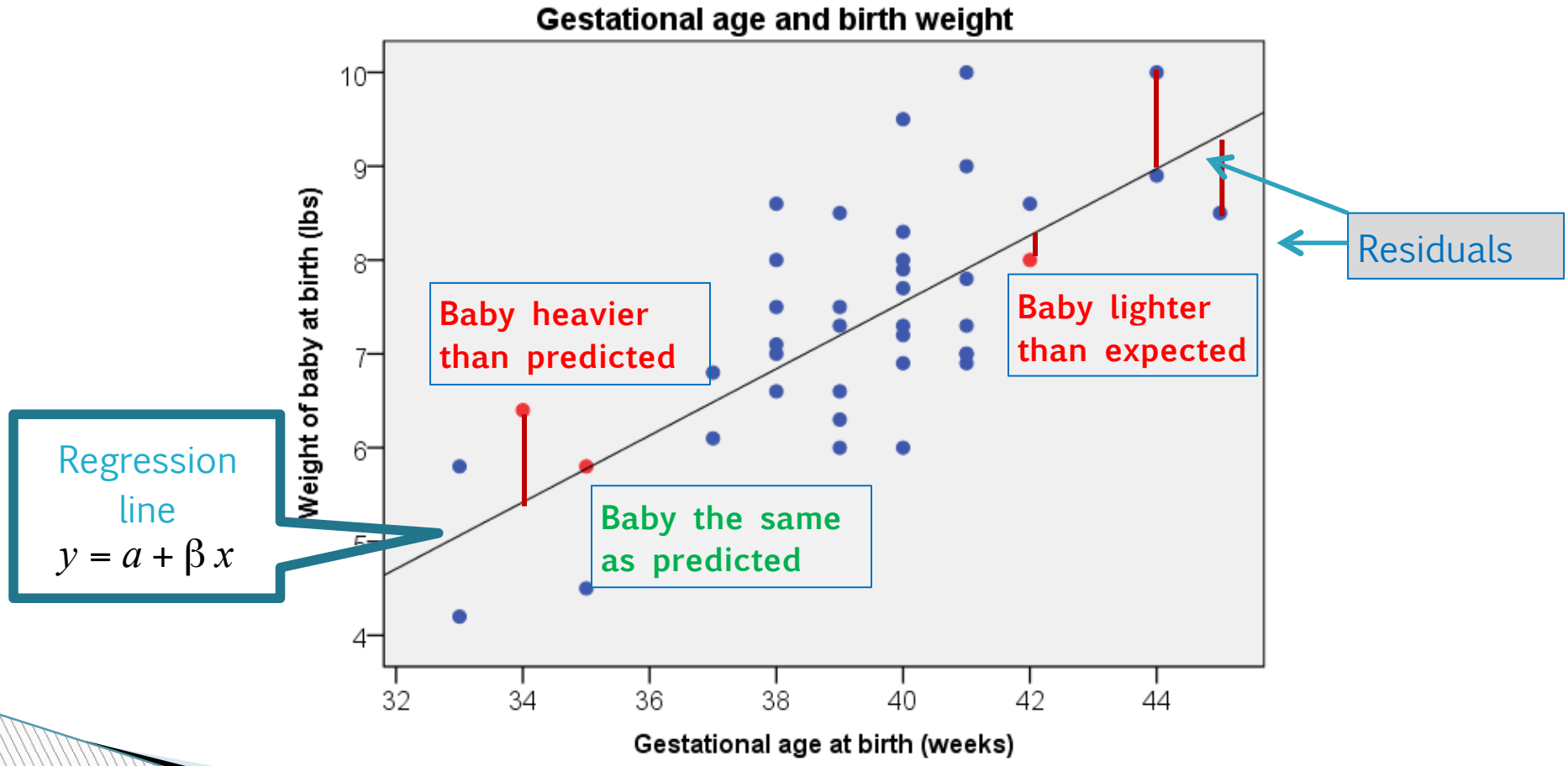
b) *predict* a value of one variable for a given value of the other

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

Dependent variable

Independent variable

$$y = a + \beta x$$

Intercept

Slope

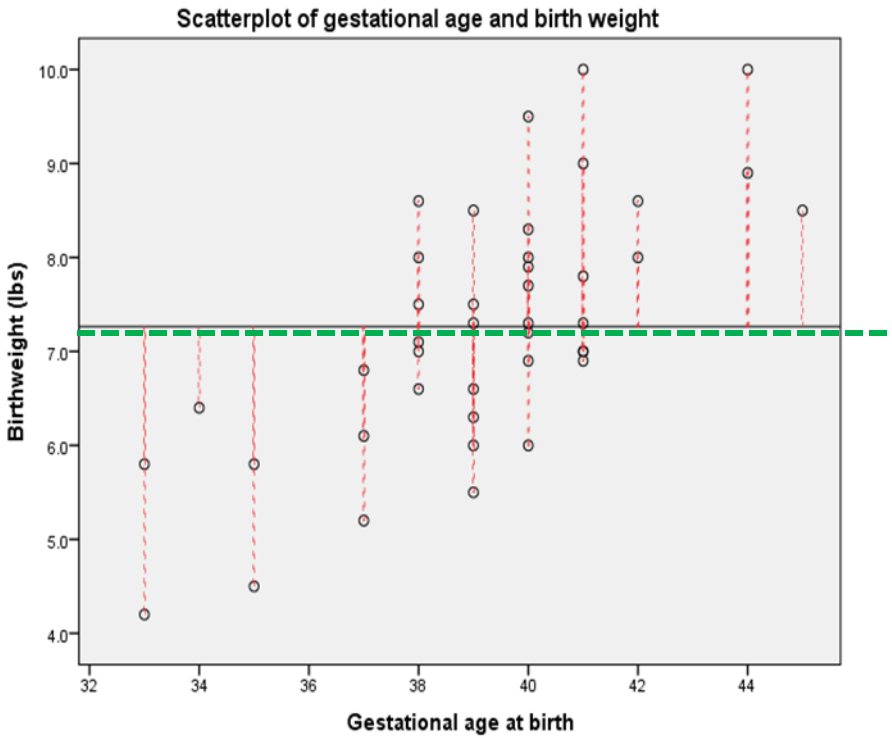
Which uses the independent variable to predict the dependent variable

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

$$H_0 : \beta = 0$$



Output from SPSS

- 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)

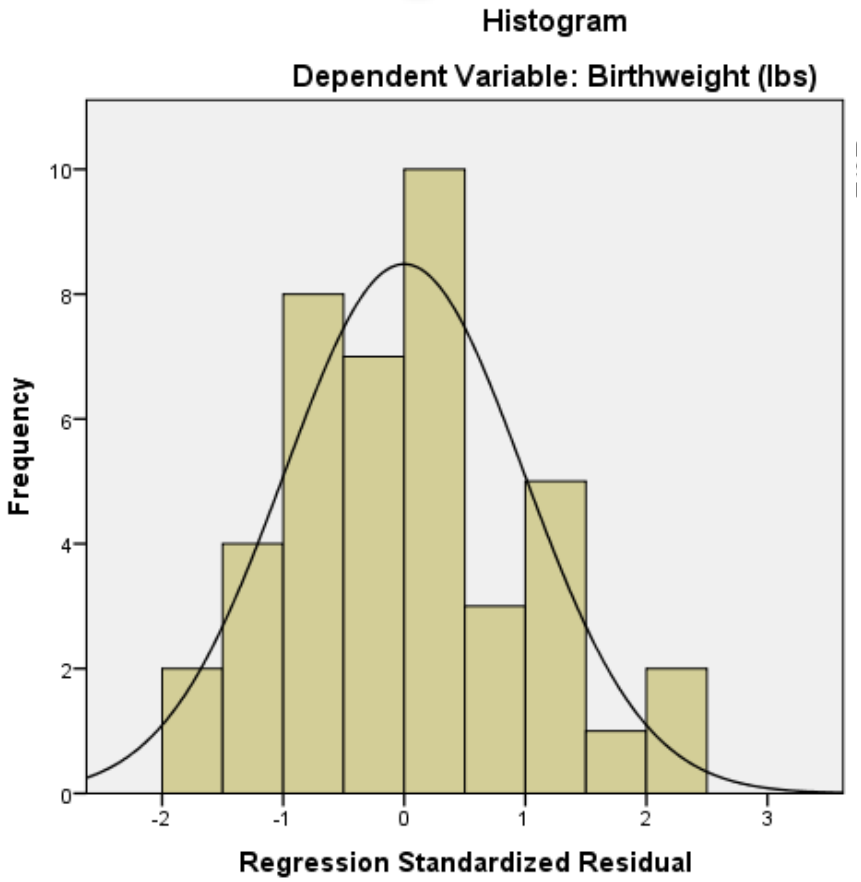
Proportion of the variation in birth weight explained by the model R^2 = 0.499 = 50%
Predictions using the model are fairly reliable.

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



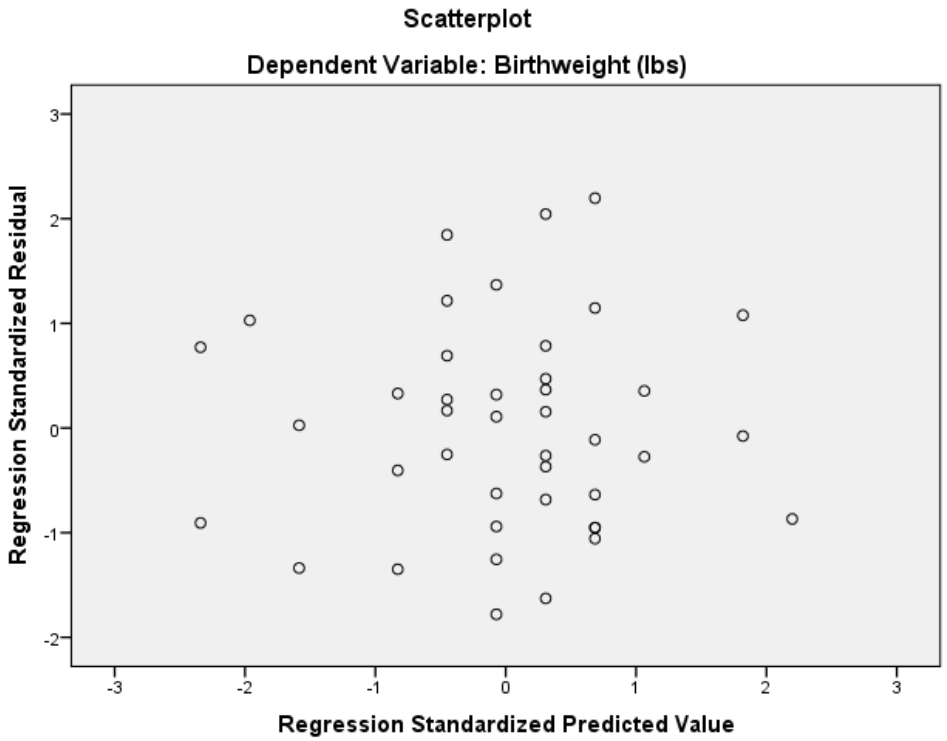
Histogram of the residuals looks approximately normally distributed

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

Outliers are outside ± 3

Predicted values against residuals

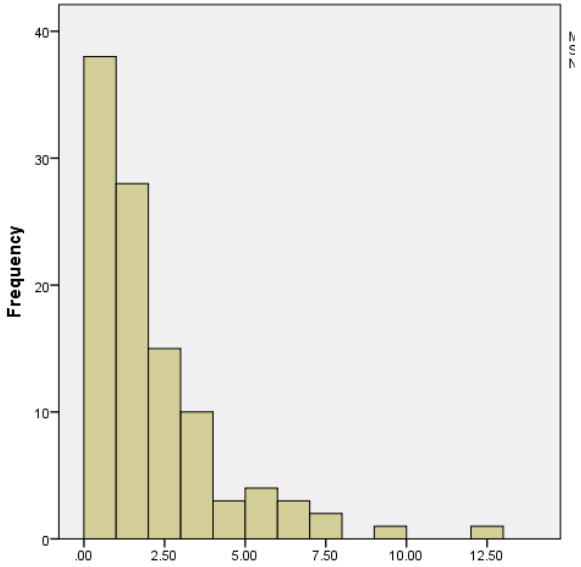
Are there any patterns as the predicted values increases?

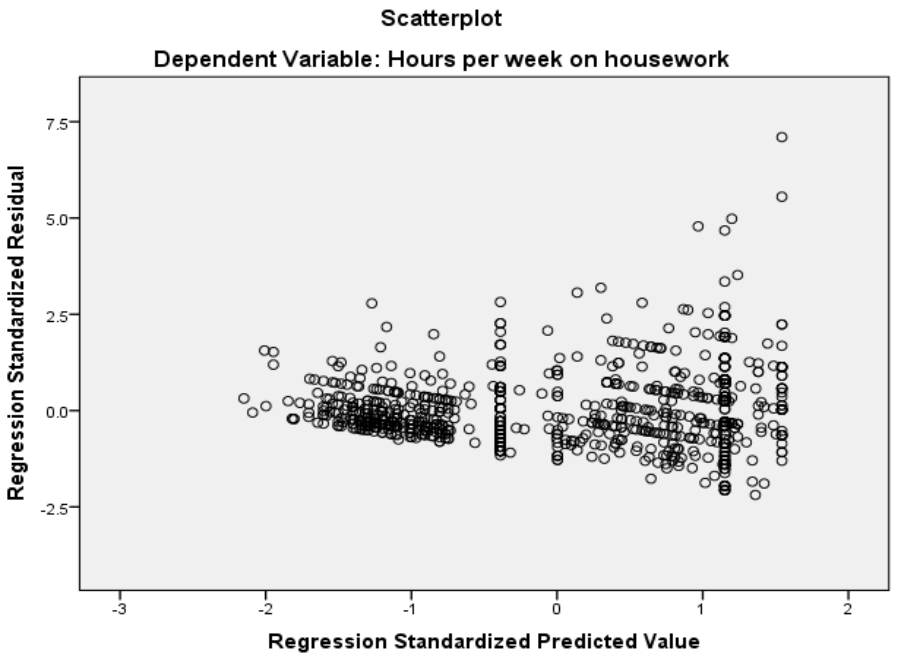


There is a problem with Homoscedasticity if the scatter is not random. A "funnelling" shape such as this suggests problems.

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 a



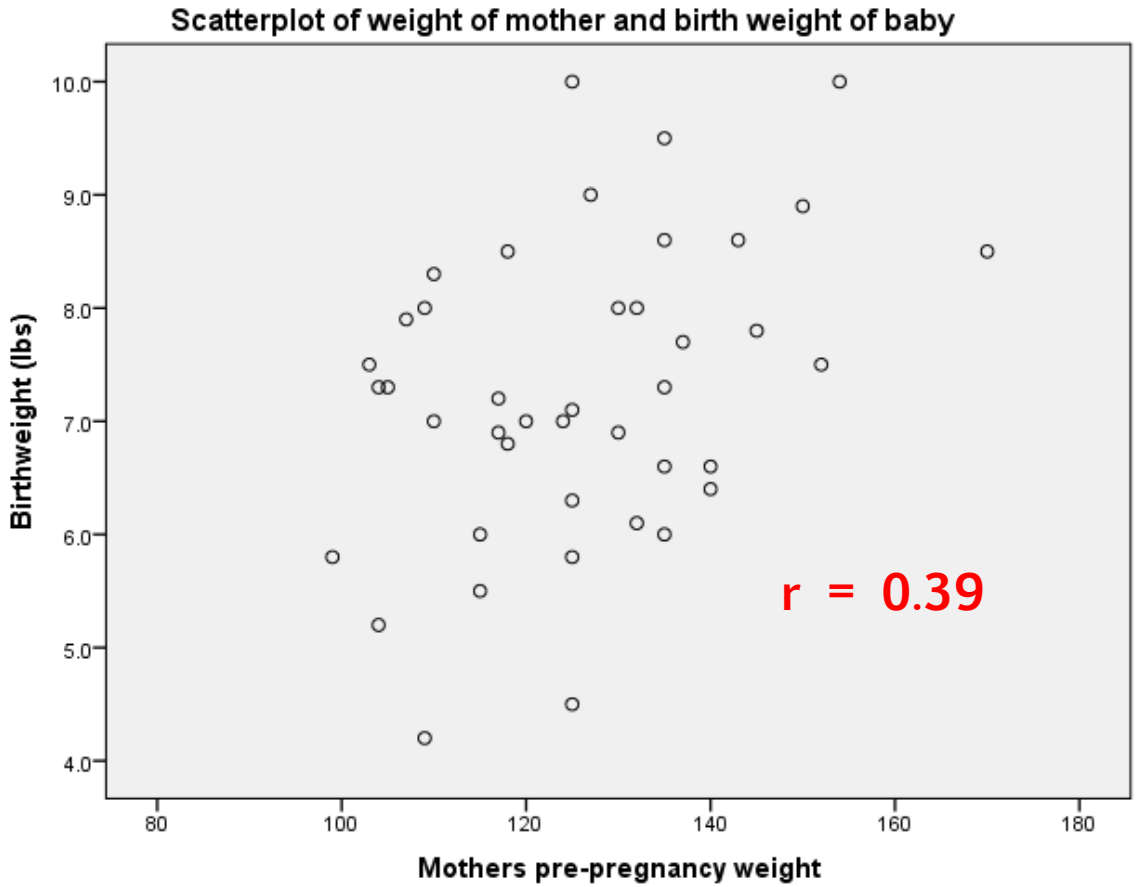


Exercise

- Investigate whether mothers pre-pregnancy weight and birth weight are associated using a scatterplot, correlation and simple regression.

Exercise - scatterplot

- Describe the relationship using the scatterplot and correlation coefficient



Regression question

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.159	1.547		2.042	.048
	Mothers pre-pregnancy weight	.033	.012	.390	2.675	.011

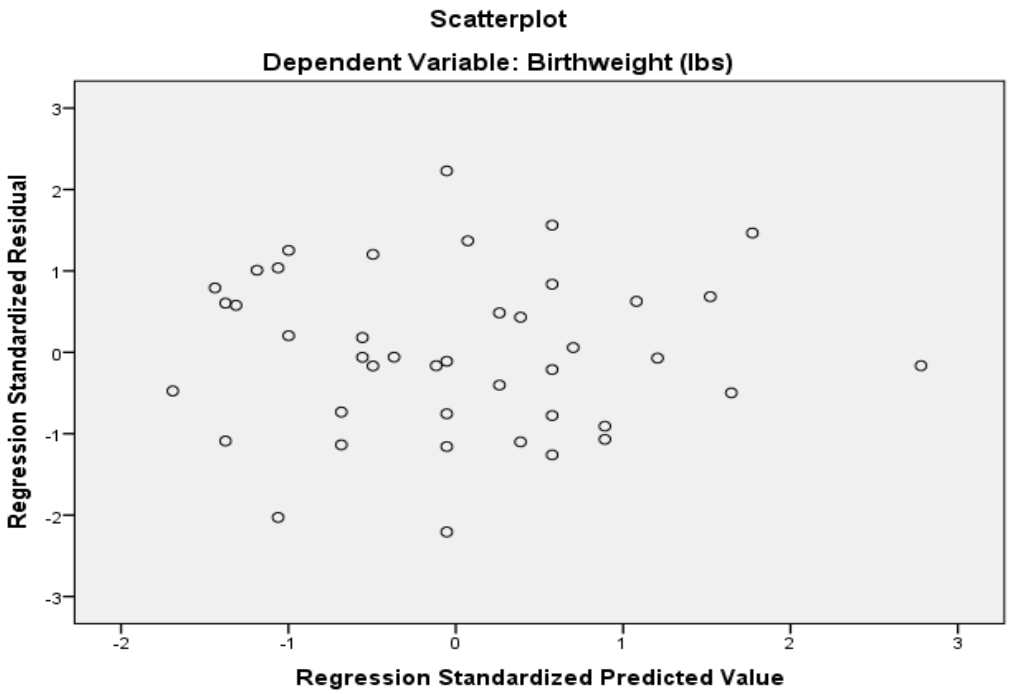
a. Dependent Variable: Birthweight (lbs)

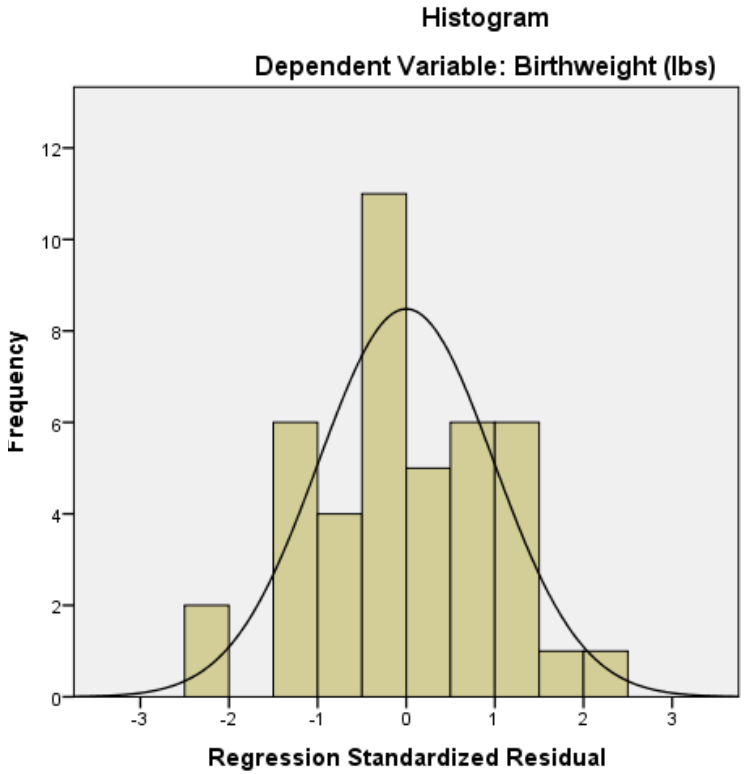
- Pre-pregnancy weight p-value:
- Regression equation:
- Interpretation:

$R^2 = 0.152$

Does the model result in reliable predictions?

Check the assumptions





Correlation

- Pearson's correlation = 0.39
- Describe the relationship using the scatterplot and correlation coefficient
- There is a moderate positive **linear** relationship between mothers' pre-pregnancy weight and birth weight (r = 0.39). Generally, birth weight increases as mothers weight increases

Regression

Pre-pregnancy weight p-value: $p = 0.011$

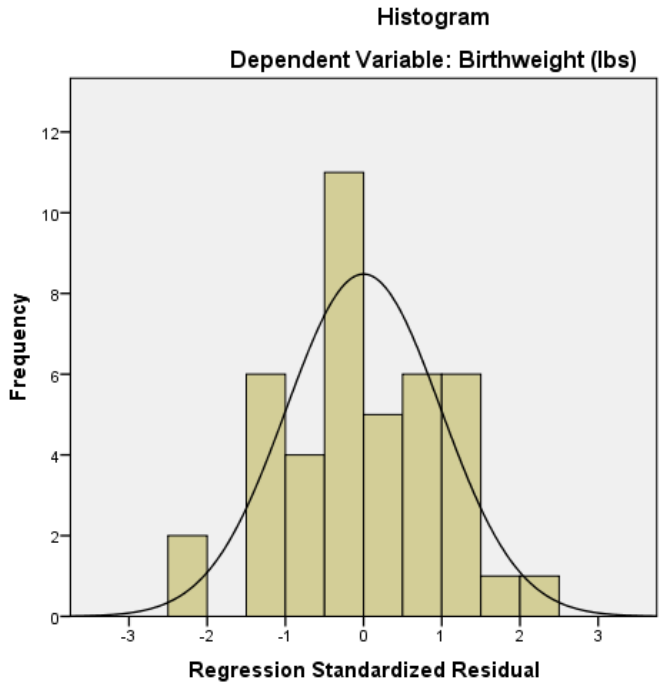
- Regression equation: $y = 3.16 + 0.03x$

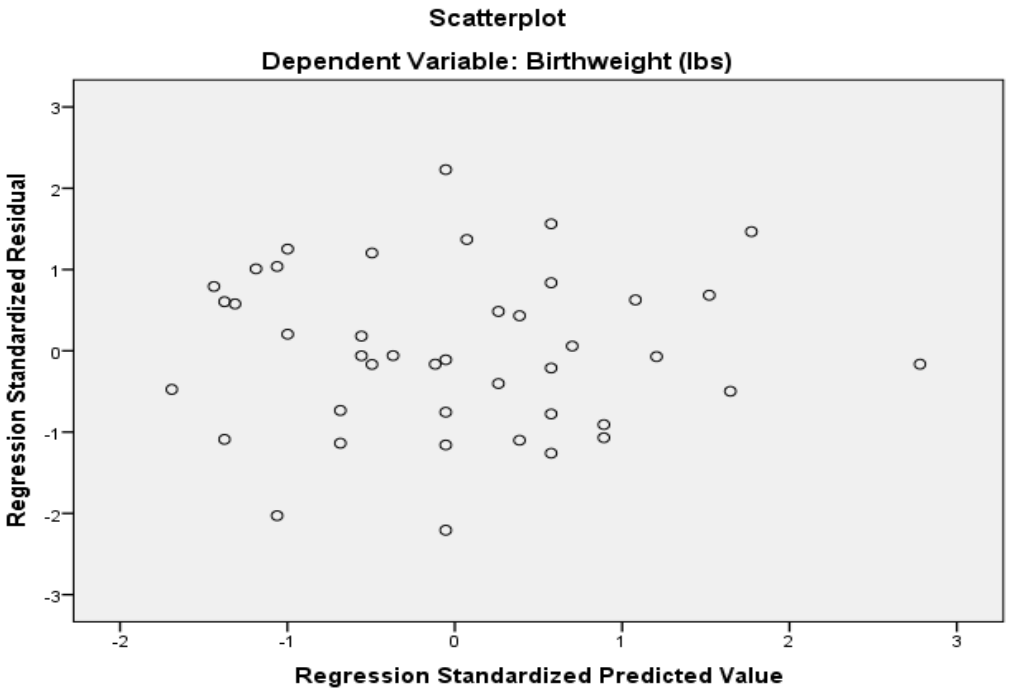
- Interpretation:
- There is a significant relationship between a mothers' pre-pregnancy weight and the weight of her baby ($p = 0.011$). Pre-pregnancy weight has a positive affect on a baby's weight with an increase of 0.03 lbs for each extra pound a mother weighs.

- Does the model result in reliable predictions?
- Not really. Only 15.2% of the variation in birth weight is accounted for using this model.

Checking assumptions

- Linear relationship
- Histogram roughly peaks in the middle
- No patterns in residuals





Multiple regression

- Multiple regression has several binary or Scale independent variables

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- Categorical variables need to be recoded as binary dummy variables
- Effect of other variables is removed (controlled for) when assessing relationships

Multiple regression

What affects the number of Nobel prize winners?

Dependent: Number of Nobel prize winners

Possible independents: Chocolate consumption, GDP and mean temperature

- Chocolate consumption is significantly related to Nobel prize winners in simple linear regression
- Once the effect of a country's GDP and temperature were taken into account, there was no relationship

Multiple regression

- In addition to the standard linear regression checks, relationships BETWEEN independent variables should be assessed
- Multicollinearity is a problem where continuous independent variables are too correlated ($r > 0.8$)
- Relationships can be assessed using scatterplots and correlation for scale variables
- SPSS can also report collinearity statistics on request. The VIF should be close to 1 but under 5 is fine whereas 10 + needs checking

Exercise

- Which variables are most strongly related?

Correlations

		Birthweight (lbs)	Gestational age at birth	Maternal height	Mothers pre-pregnancy weight
Birthweight (lbs)	Pearson Correlation	1	.706**	.368*	.390*
	Sig. (2-tailed)		.000	.017	.011
	N	42	42	42	42
Gestational age at birth	Pearson Correlation	.706**	1	.231	.251
	Sig. (2-tailed)	.000		.141	.110
	N	42	42	42	42
Maternal height	Pearson Correlation	.368*	.231	1	.671**
	Sig. (2-tailed)	.017	.141		.000
	N	42	42	42	42
Mothers pre-pregnancy weight	Pearson Correlation	.390*	.251	.671**	1
	Sig. (2-tailed)	.011	.110	.000	
	N	42	42	42	42

Exercise - Solution

- Which variables are most strongly related?
- Gestation and birth weight (0.709)

- Mothers height and weight (0.671)

Mothers height and weight are strongly related. They don't exceed the problem correlation of 0.8 but try the model with and without height in case it's a problem.

- When both were included in regression, neither were significant but alone they were

Logistic regression

- Logistic regression has a binary dependent variable
- The model can be used to estimate probabilities

- Example: insurance quotes are based on the likelihood of you having an accident
- Dependent = Have an accident/ do not have accident
- Independents: Age (preferably Scale), gender, occupation, marital status, annual mileage

- Ordinal regression is for ordinal dependent variables

Choosing the right test

Choosing the right test

- One of the most common queries in stats support is ‘Which analysis should I use’
- There are several steps to help the student decide
- When a student is explaining their project, these are the questions you need answers for

Choosing the right test

1) A clearly defined research question
2) What is the dependent variable and what type of variable is it?
3) How many independent variables are there and what data types are they?
4) Are you interested in comparing means or investigating relationships?
5) Do you have repeated measurements of the same variable for each subject?

Research question

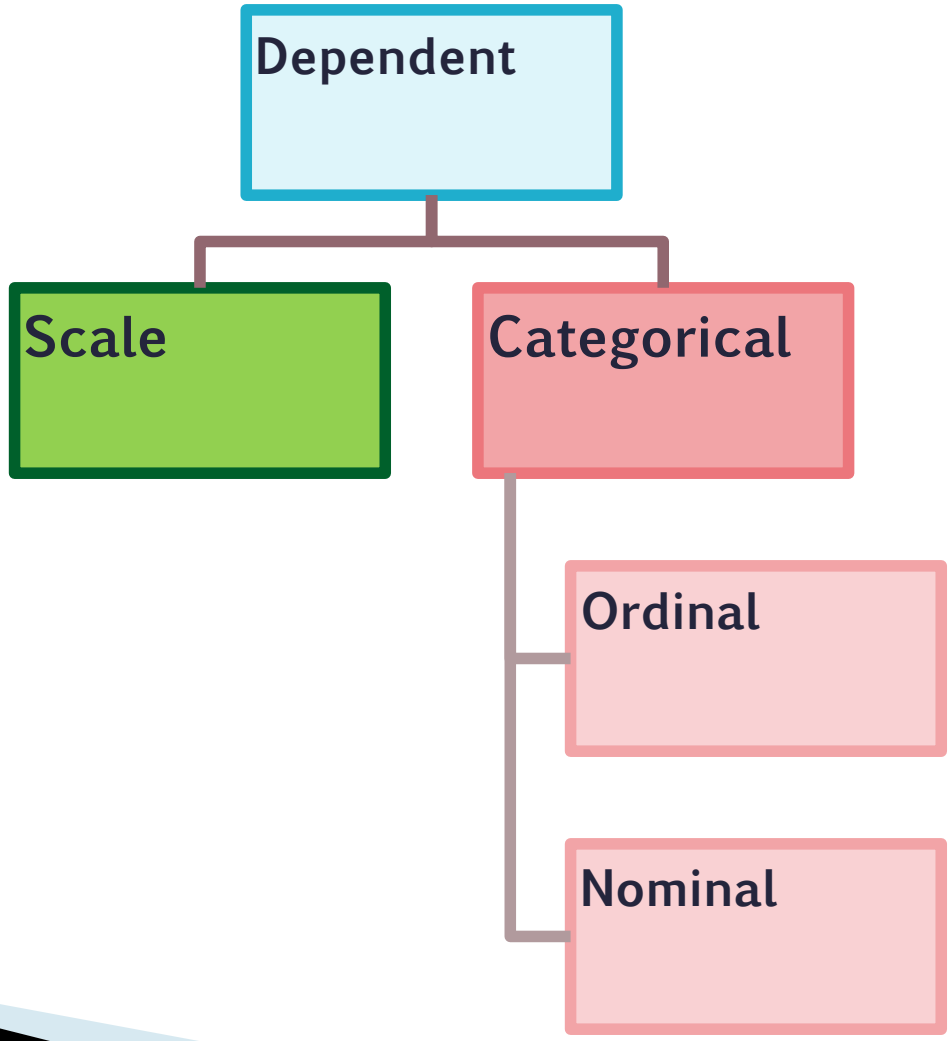
- Clear questions with measurable quantities
- Which variables will help answer these questions
- Think about what test is needed before carrying out a study so that the right type of variables are collected

Dependent variables

Does **attendance** have an association with **exam score**?

Do **women** do more **housework** than **men**?

What variable type is the dependent?



Are boys **better** at math?

- How can 'better' be measured and what type of variable is it?

Exam score (Scale)

- Do boys think they are better at math??

I consider myself to be good at math (ordinal)

How many variables are involved?

- Two – interested in the relationship
- One dependent and one independent
- One dependent and several independent variables: some may be controls
- Relationships between more than two: multivariate techniques (not covered here)

Data types

Research question	Dependent/ outcome variable	Independent/ explanatory variable
Does attendance have an association with exam score?	Exam score (scale)	Attendance (Scale)
Do women do more housework than men?	Hours of housework per week (Scale)	Gender (binary)

Exercise:

How would you investigate the following topics? State the dependent and independent variables and their variable types.

Research question	Dependent/ outcome variable	Independent/ explanatory variable
Were Americans more likely to survive on board the Titanic?		
Does weekly hours of work influence the amount of time spent on housework?		
Which of 3 diets is best for losing weight?		

Exercise: Solution

How would you investigate the following topics? State the dependent and independent variables and their variable types.

Research question	Dependent/ outcome variable	Independent/ explanatory variable
Were Americans more likely to survive on board the Titanic?	Survival (Binary)	Nationality (Nominal)
Does weekly hours of work influence the amount of time spent on housework?	Hours of housework (Scale)	Hours of work (Scale)
Which of 3 diets is best for losing weight?	Weight lost on diet (Scale)	Diet (Nominal)

Comparing means

- Dependent = Scale
- Independent = Categorical

- How many means are you comparing?

- Do you have independent groups or repeated measurements on each person?

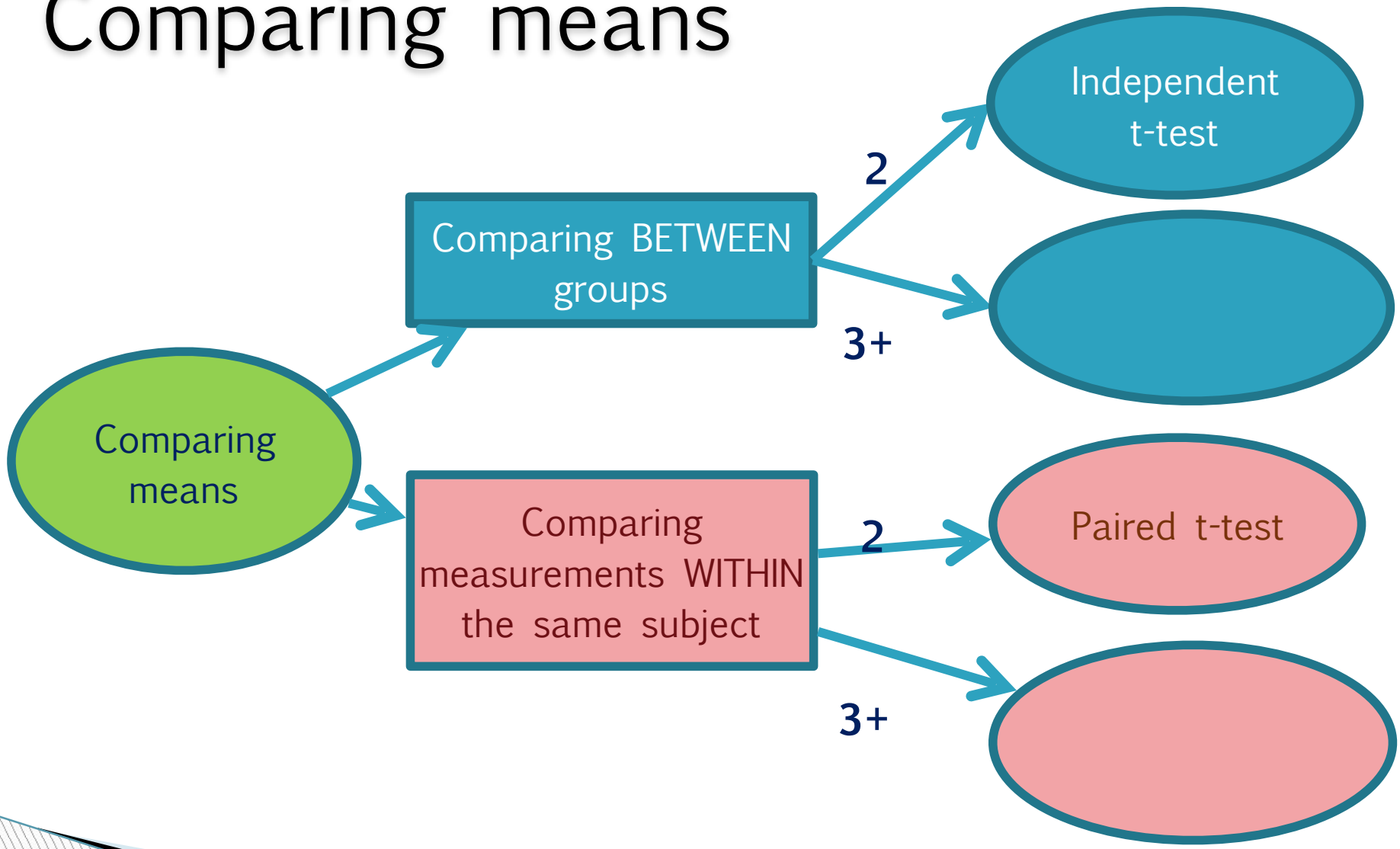
Comparing measurements on the same people

Also known as within group comparisons or repeated measures.

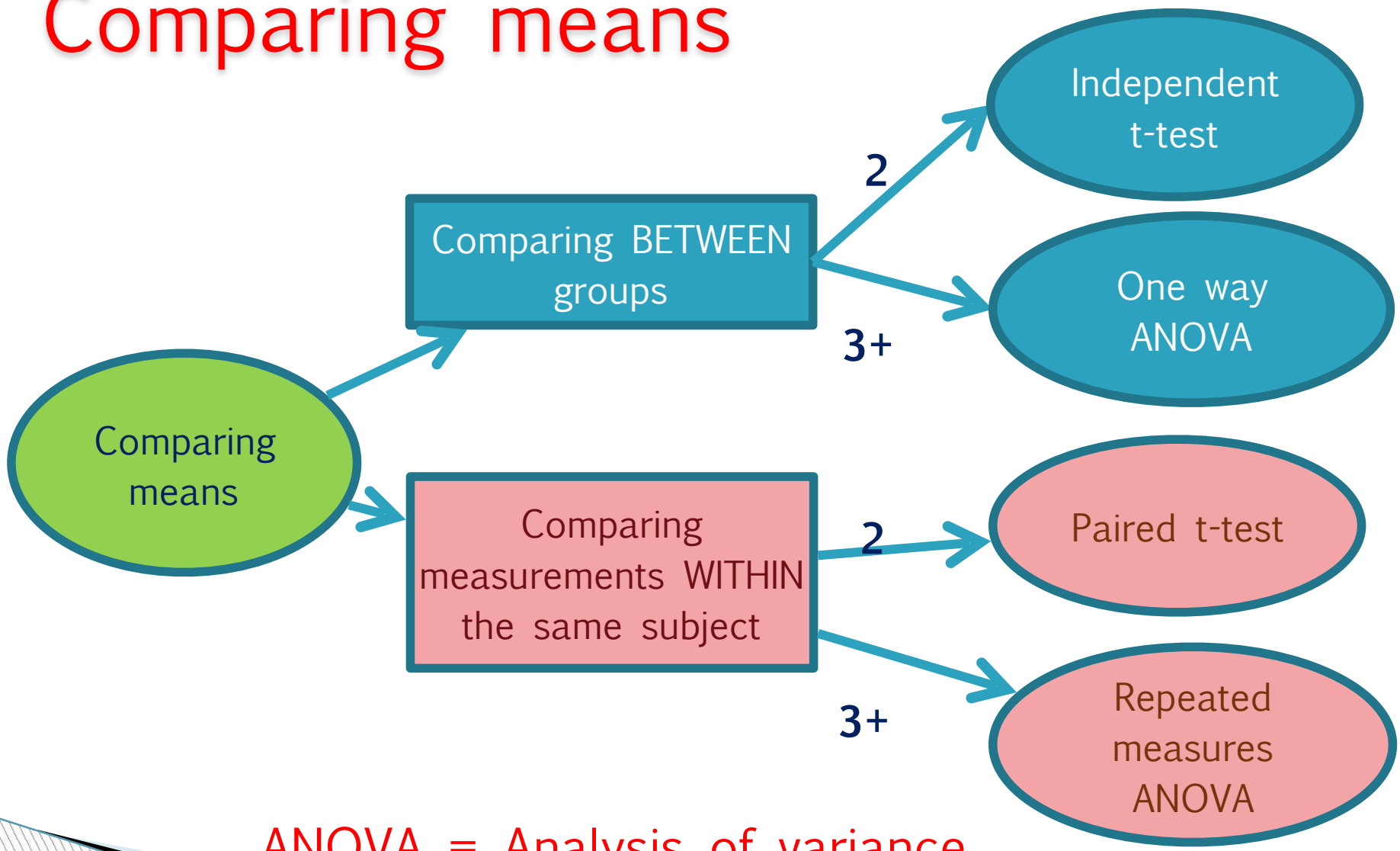
Can be used to look at differences in mean score:
(1) over 2 or more time points e.g. 1988 vs 2014
(2) under 2 or more conditions e.g. taste scores

Participants are asked to taste 2 types of cola and give each scores out of 100.
Dependent = taste score
Independent = type of cola

Comparing means



Comparing means



ANOVA = Analysis of variance

Exercise – Comparing means

Research question	Dependent variable	Independent variable	Test
Do women do more housework than men?	Housework (hrs per week) (Scale)	Gender (Nominal)	
Does Margarine X reduce cholesterol? Everyone has cholesterol measured on 3 occasions	Cholesterol (Scale)	Occasion (Nominal)	
Which of 3 diets is best for losing weight?	Weight lost on diet (Scale)	Diet (Nominal)	

Exercise: Solution

Research question	Dependent variable	Independent variable	Test
Do women do more housework than men?	Housework (hrs per week) (Scale)	Gender (Nominal)	Independent t-test
Does Margarine X reduce cholesterol? Everyone has cholesterol measured on 3 occasions	Cholesterol (Scale)	Occasion (Nominal)	Repeated measures ANOVA
Which of 3 diets is best for losing weight?	Weight lost on diet (Scale)	Diet (Nominal)	One-way ANOVA

Tests investigating relationships

Investigating relationships between	Dependent variable	Independent variable	Test
2 categorical variables	Categorical	Categorical	Chi-squared test
2 Scale variables	Scale	Scale	Pearson's correlation
Predicting the value of an dependent variable from the value of a independent variable	Scale	Scale/binary	Simple Linear Regression
	Binary	Scale/ binary	Logistic regression

Note: Multiple linear regression is when there are several independent variables

Exercise: Relationships

Research question	Dependent variable	Independent variables	Test
Does attendance affect exam score?	Exam score (Scale)	Attendance (Scale)	
Do women do more housework than men?	Housework (hrs per week) (scale)	Gender (Binary) Hours worked (Scale)	
Were Americans more likely to survive on board the Titanic?	Survival (Binary)	Nationality (Nominal)	
	Survival (Binary)	Nationality , Gender, class	

Note: There may be 2 appropriate tests for some questions

Exercise: Solution

Research question	Dependent variable	Independent variables	Test
Does attendance affect exam score?	Exam score (Scale)	Attendance (Scale)	Correlation/ regression
Do women do more housework than men?	Housework (hrs per week) (scale)	Gender (Binary) Hours worked (Scale)	Regression
Were Americans more likely to survive on board the Titanic?	Survival (Binary)	Nationality (Nominal)	Chi-squared
	Survival (Binary)	Nationality , Gender, class	Logistic regression

Note: There may be 2 appropriate tests for some questions