Chapter 1

Exercise 1:

a) **P:** pre-school children, **I:** -, **C:** -, **O:** prevalence of asthma
b) **P:** newly diagnosed cancer patients, **I:** -, **C:** -, **O:** prognosis
c) **P:** epileptic children, **I:** -, **C:** non-epileptic children, **O:** BMI
d) **P:** children who develop epilepsy between 5 and 15 years of age, **I:** -, **C:** -, **O:** prescription of antibiotics in the first year of life
e) **P:** fetuses with head-circumferences above the 99th centile, **I:** -, **C:** -, **O:** likelihood of congenital abnormality
f) **P:** milt to moderate asthmatics, **I:** steroids, **C:** -, **O:** lung function
g) **P:** patients with severe eczema, **I:** patient contact with further therapy, **C:** patient contact without further therapy, **O:** skin condition

Exercise 2:

a) Single sample, observational
b) Single sample, observational
c) Cross-sectional
d) Case-control
e) Cohort
f) RCT, cross-over
g) RCT

Exercise 3:

a) No confounders
b) No confounders
c) Age, sex, socio-economic, co-morbidities, genetic factors, medication
d) Socio-economic factors, genetics and in general whatever could be associated with antibiotic use
e) Ethnicity, dietary habits, family factors, genetic factors, weight and height of fetuses
f) We would hope for no confounding factors because of the randomisation

g) We would hope for no confounding factors because of the randomisation

Exercise 5:

A: Web addiction and depression

a) What research question did the study hope to answer?

Is there a relationship between excessive internet use and depression?

b) What was the target population?

Internet users.

c) How was this population sampled from? Were there any biases in the sampling procedure? Could this affect outcome?

A total of 1,319 respondents were recruited via links placed on UK-based social networking sites. This group is not necessarily representative of the general population of internet users and their responses should be viewed with caution. Recruitment took place via social-networking sites, which older people do not tend to use and therefore has sample a predominantly younger population with an average age of 21.

d) What measurements were made on each subject? Were these measurements adequate to address the research question? Were any spurious and were there others that should've been made that weren't?

Three questionnaires were applied:

- Young's IAT – identified people as mildly, moderately or severely addicted
- Internet Function Questionnaire – measured the different uses people have for the Internet
- BDI – long-standing, widely used, self-evaluation depression scale

As these questionnaires are self reported it is likely that there might be a degree of inaccuracy introduced.

The researchers were not able to examine the wider personal, social, professional and health circumstances of participants, which are likely to be the main influence on mental health. A singly questionnaire cannot be taken as a definite diagnosis of addiction or depression.

e) In the light of the study results what appears to be the answer to the research question?

There was a close correlation between addictive tendencies and depression across the sample, with the higher the depression score, the higher the addiction score. Men showed more addictive tendencies than women and younger people more than older people.

f) Are the authors' conclusions justified?

The authors conclude that the concept of internet addiction is ‘emerging as a construct that must be taken seriously’ and ‘those who regard themselves as dependent on the internet report high levels of depressive symptoms’. They also say that further work is need on assessing this relationship.

Although the study has found an association, this does not prove causation. It’s possible that a person uses the internet more because they are depressed, not the other way around. A link between depression and internet addiction is not out of question but for a causal relationship to be proved further research is needed.

Also, only 18 people were considered to have internet addiction, so examining associations between other factors in this small number of people is likely to involve some inaccuracy.

g) Could the study design have been improved to enable the research question to be answered in a more direct, simple or definite way?

An alternative, more appropriate study would involve a more representative sample of the general population of internet users, more data on their social, medical and personal information, validation of their addiction and depression status via

professionals and comparison between groups that were assigned to increased and decreased use of the internet.

B: Magnesium and depression

a) What research question did the study hope to answer?

Does 6 weeks of oral magnesium chloride supplementation improve symptoms of mild-to-moderate depression in a primary care population?

b) What was the target population?

Adults with mild-to-moderate depression (in primary care?).

c) How was this population sampled from? Were there any biases in the sampling procedure? Could this affect outcome?

There are few details of how patients were sampled. The paper only says:

“Participants were recruited through primary care providers (PCPs) within a single academic medical centre.”

There is only one centre involved and we are assuming that patients from this centre are representative of the population of interest.

If the population of interest is primary care patients then we assume that the sample is representative of patients from all primary care centres. However, the authors seem to want to generalise their findings to all people with mild-to-moderate depression indicating that this is their target population. It is unlikely that patients in this primary care centre are representative of this population. Those in primary care only capture people seeking medical help.

d) What measurements were made on each subject? Were these measurements adequate to address the research question? Were any spurious and were there others that should've been made that weren't?

Primary outcome: PHQ-9 score (baseline and after treatment). The questionnaire appears to be validated: “The PHQ-9 is a validated questionnaire with high sensitivity and specificity for the diagnosis of depression”

Secondary outcomes: GAD-7 score (measure symptoms of anxiety) and side-effects of treatment.

e) In the light of the study results what appears to be the answer to the research question?

The researchers say their results showed that "magnesium supplements may be a fast, safe and easily accessible alternative or adjunct [addition] to starting or increasing the dose of antidepressant medications."

f) Are the authors' conclusions justified?

The study is not blinded and does not have a placebo group. They claim having a placebo group is "not useful when the research seeks to assess the presence and magnitude of the effect of an intervention."

They add: "Whether magnesium works because it induces a physiological change in the subject, or only because of the placebo effect (or a combination of the two), it remains that subjects do report better levels of depression and anxiety when taking magnesium than when not."

The lack of a placebo group in the study means we cannot be sure whether magnesium is a useful treatment for depression. We know that the placebo effect is real, and that it can bias results of clinical trials if not tested for by a placebo group in the study.

This study was relatively small (only 112 people provided data that could be analysed); lasted only 12 weeks and did not include a placebo group. It's entirely possible that the results shown with magnesium pills are due to the placebo effect, and that they would have worn off with a longer study period.

This study seems to have been a wasted opportunity to find out whether magnesium is a useful supplement for people with mild to moderate depression.

g) Could the study design have been improved to enable the research question to be answered in a more direct, simple or definite way?

The clear limitation is blinding. Patients could have reported better outcomes just because they knew they were taking magnesium supplements and knew the purpose of the study. The authors try to argue that a placebo controlled trial is not necessary, but most would disagree with this argument.

Enrolling patients with depression listed on their medical chart resulted in missing people with undiagnosed depression or who do not use Primary Care. A better representative sample could have been derived from the general population.

Although improvement in symptoms occurred within two weeks and was maintained while on treatment, long-term effectiveness is unknown and longer trials are needed.

C: Driving and gambling

a) What research question did the study hope to answer?

Is there a relationship between gambling and risk of collisions?

b) What was the target population?

People with gambling problems.

c) How was this population sampled from? Were there any biases in the sampling procedure? Could this affect outcome?

Adults aged 18 years and older were recruited from six problem gambling treatment centres in Ontario, Canada from 2003 to 2005. Participants were among a treatment population originally recruited for a larger study looking at collision risk related to substance abuse.

The sample is biased towards those that are seeking help and doesn’t capture those that don’t admit to having a gambling problem.

d) What measurements were made on each subject? Were these measurements adequate to address the research question? Were any spurious and were there others that should've been made that weren't?

The study consisted of both quantitative and qualitative measures.

Measurements included:

- Six validated psychosocial scales related to problem gambling
    - aggression
    - risk taking/impulsivity
    - perceived stress
    - sleepiness
    - social support
    - DSM-IV – an assessment of problem gambling
- The Gambling Effects of Driving Scale

The primary outcome was whether they had collision while driving, which they assessed through open-ended questions on driving after gambling (qualitative).

Since the relationship between problem gambling and driving has not been examined in the past, the Gambling Effects on Driving Scale was designed specifically for this study.

There are doubts about the validity of this questionnaire developed for the study itself and insufficient details are given about how it was derived.

e) In the light of the study results what appears to be the answer to the research question?

An association is observed between gambling problems and adverse driving outcomes. Bu the authors conclude that more research is needed to understand the

causal pathways and increased collision risk among problem gamblers and among the general gambling population.

f) Are the authors' conclusions justified?

The cross sectional study design does not allow one to make a causal association between gambling and driving. But since the authors don't make strong conclusions, they appear to be justified. The authors recognise a number of limitations to their research:

- Data is self-reported, so there could be information bias, or social desirability bias.
- The sampling procedure means we can't generalise to the population of gambling addicts.

g) Could the study design have been improved to enable the research question to be answered in a more direct, simple or definite way?

The study would have been better as a cohort study. People could have been sampled from casinos/bookmakers, which would have also captured light gamblers that could have been used a comparison group. Adverse vehicle collision data could be collected from insurance companies if data protection laws would allow.

Exercise 6:

i) Should mention:

- Choice of sample - target group, population sample from - how sampled
- Any exclusion criteria - reasons - generalisability of results
- Outcome measures - what, when, who assesses
- Control group?
- Blinding - whether possible? Who will be blind?
- How allocated to treatment groups? Any potential confounders to consider? How will these be taken into account?
- Mechanism for recording refusers - what information collected on them
- Similarly for dropouts

- Sample size needed - how calculated? How many need to be recruited to achieve these numbers
- Statistical analyses envisaged

ii) Should mention:

- Choice of sample - target group, population sample from - how sampled
- Any exclusion criteria - reasons - generalisability of results
- Outcome measures - what, when, who assesses
- Control group? - does this apply here?
- Blinding - whether possible? Who will be blind?
- mechanism for recording refusers - what information collected on them
- Similarly for dropouts
- Sample size needed - how calculated? How many need to be recruited to achieve these numbers
- Statistical analyses envisaged

Chapter 2

Exercise 1:

i) Height: **numeric**
ii) Blood pressure: **numeric**
iii) Number of male siblings: **numeric (discrete)**
iv) Customer service satisfaction: **ordinal**
v) Number of female siblings: **numeric (discrete)**
vi) Family history of asthma: **binary**
vii) Hair colour: **nominal**
viii) Lung function: **numeric**
ix) Music volume (number of decibels): **numeric**
x) Country of birth: **nominal**
xi) Tolerance rating: **ordinal**

Exercise 2:

Weight (in either kg or lb) is a continuous numeric variable. A person's weight can take any value within a wide range. Weight could be categorised into either two or more categories.

With 2 categories (a dichotomy of weight) the decision may be made to select a weight at which below this is OK and above it is **overweight**.

With more than two categories, there could be a series of weight ranges. For example, 4 categories could be used: **below** 50Kg/ 51-75Kg / 76-100Kg/ **over** 100Kg.

In either of these latter scenarios information is lost. It is preferable to record the actual weights.

Exercise 3:

The factors to be considered when deciding on how to display a set of data is:

- The types of variables
- The number of variables
- And the number of observations

Exercise 4:

		Variables	Type of variables	Values they take	Appropriate	Alternatives
i)	side by side histogram	% of visits	Continuous	0-100%	Yes	Stacked bar chart
		Prescribing choices	Categorical – nominal	7 levels		
		Before and after guideline groups	Categorical - binary	2 levels		
	Trend graph	% of visits	Continuous	0-100%	Yes	
		Time in halves of 5 years	Categorical – Nominal	10 levels		
ii)	Dot diagram	Distal	Continuous	0? – 2.80	Yes	
		Exposure to copolymer	Categorical – binary	2 levels		
iii)	Bar chart	HLH diagn crit	Categorical – nominal	13 levels	Yes	
		Frequency	Continuous	0-100%		
	Table	Ethnicity	Categorical – nominal	6 levels	Yes	Side-by-side and stacked bar chart
		Registry areas	Categorical – binary	2 levels		

		Diagnoses	Categorical – binary	2 levels		
iv)	Stacked 3d bar chart	Leisure activity	Categorical – nominal	9 levels	Yes	
		Frequency of leisure activity	Categorical – ordinal	3 levels		
		Number (and %) of individuals choosing each of the frequency levels	Continuous	0-73 (0-100%)		
	Stacked 3d bar chart	Life satisfaction	Categorical – nominal	3 levels	Yes	
		Time broken to 3 levels	Categorical – nominal	3 levels		
		% of individuals choosing each of the 3 levels of life satisfaction	Continuous	0-100%		
v)	Stacked bar chart	Type of maltreatment	Categorical – nominal	4 levels	Yes	Side by side bar chart
		Certainty of maltreatment	Categorical – nominal	3 levels		
		Rate (%) of individuals reporting in each group	Continuous	0-100		
vi)	Scatter plot	Age	Continuous	0.01-100	Yes	
		CI	Continuous	0-160		

	Scatter plot	Age	Continuous	0.01-100	Yes	
		AYC	Continuous	1200-2800		
		Group of patients	Categorical – nominal	3 levels		
vii)	Stacked bar charts	Gestational age	Continuous	13-31	No	Scatterplot with two continuous variables on x and y axes and different symbols for the binary variable
		Number of placentas	Continuous	0-30		
		Ureaplasma spp	Categorical – binary	2 levels		
viii)	Line diagram	Time	Categorical – binary	2 levels	Yes	
		% active trachoma	Continuous	0-100%		
		Treatment	Categorical – binary	2 levels		
ix)	Dot plot	TAK score	Continuous	0-18	No	IQ should had been kept in a continuous format and presented as scatterplot
		IQ category	Categorical – binary	2 levels		
		Health status	Categorical – binary	2 levels		
x)	Scatter plot	Age	Continuous	0-10	Yes	
		Fat-free mass	Continuous	0-20		
	Scatter plot	Age	Continuous	0-10	Yes	
		Weight	Continuous	0-25		

xi)	Bar chart	Month	Categorical – ordinal	12 levels	Yes	
		No of cases	Continuous	0-40		
		Age	Continuous	1-15	Yes	Age could have been kept in a continuous format and presented in scatterplot against no of cases
		No of cases	Continuous	0-40		
	Picture	Location of bite	Categorical – nominal	8 levels	Yes	
		Frequency of bites	Continuous	0-100%		
xii)	Table	Age group	Categorical – ordinal	6 levels	No	Both variables are continuous and should have been kept in this format and presented in scatterplot
		Haemoglobin level	Categorical – ordinal	4 levels		
	Table	Morphology	Categorical – nominal	4 levels	No	This new variable could have been incorporated into the above scatterplot via different symbol for each morphology group
		Haemoglobin level	Categorical – ordinal	*4 levels*		

Chapter 3

Exercise 1:

	Autistic	**Non-autistic**	**Totals**
Fragile X syndrome	18 (18/144=12.5%)	52 (52/254=20.5)	70 (70/398=17.6%)
No fragile X syndrome	126 (126/144=87.5)	202 (202/254=79.5%)	328 (328/398=82.4%)
Totals	144	254	398

Exercise 2:

Mean = 769/24 = 32.04, Median = 33.5

What is the difference between the mean and the median?

Difference = 33.5 - 32.04 = 1.46

Does this tell us anything about the distribution? What?

The mean age is younger than the median, so, if anything there is some downward skewing of the ages. However the difference is small suggesting that skewing is minimal.

Exercise 3:

Were the times to arrival symmetrically distributed? Are they upwardly or downwardly skew?

The large differences between the means and medians suggest that the data is skew, i.e. they are not normally distributed.

There is upward skewing since the means are greater than the medians.

Exercise 4:

Median = 3

The sum of the injury times (0.25 + 0.25 + 0.5 + ...+ 4) = 146. Calculate the mean recovery time.

Mean = 146/25 = 5.84

Compare this to the median and comment.

The mean is greater than the median suggesting that there may be some upward skewing of the recovery times. Examination of the data shows that the majority recovered within 6 months and there were just a few who took a lot longer to recover (10, 12, 31 and 37 months).

Comment on the statement at the foot of the table.

Since the mean changes dramatically when the two upper times (31 and 37) are removed (i.e. it falls to 3.4 months), this indicates that these are outlying values. However, this is a somewhat arbitrary thing to do - the mean must fall when the two highest values are removed and it is not straightforward to interpret whether the fall is 'large' or not. It would have been better if they had instead given the median value.

Exercise 5:

Range = 40 - 25 = 15 years

25th centile = 26, 75th centile = 36; Interquartile range = 36-26 = 10

The standard deviation of the ages is 4.84, calculate the variance.

Variance = 4.842 = 23.43

How would you expect the distribution to have changed if the standard deviation had been 1.73?

If the standard deviation was 1.73 instead of 4.84 would expect the distribution to be more tightly bunched - taller and thinner.

Exercise 6:

Range = 37 - 0.25 = 36.75 months
25th centile = 2 (i.e. the 7th largest values - which has 6 values less than and 18 values greater than it)

75th centile = 5 (i.e. the 19th largest value - which has 18 values less than and 6 values greater than it)

Interquartile range = 5 - 2 = 3

The variance of these times is 80.39, calculate the standard deviation.

Standard deviation = = 8.966

How would you expect the distribution to have changed if the variance had been 120.47?

If the variance was 120.47 instead of 80.39 would expect the distribution of recovery times to be more spread out.

Exercise 7:

Age of onset is upwardly skew; age of death is downwardly skew. Because of skewness, the mean age will not be representative of the bulk of the patients and the standard deviation will give a distorted measure of spread.

Suitable measures of centre and spread would be the median and interquartile range respectively.

The data could be transformed to normality since age is a continuous variable. Age at onset is upwardly skew, a log, square root or inverse may correct this. Age at death is downwardly skew, so suitable transformations may be squaring, cubing or anti-logging.

Exercise 8:

Median for all 21 boys = 105

Median for Griffiths scale = 98

Median for BAS scale = 114 (half way between 112 and 116)

These are all fairly similar to the mean values indicating approximate symmetry within each of the groupings.

Range for all boys = 128 - 68 = 60

Range for Griffiths scale = 104 - 68 = 36

Range for BAS scale = 128 - 105 = 23

Exercise 9:

Expect approximately 95% to lie in the interval (mean ± 2 sd)
= (32.04 ± 2(4.84)) = (32.04 ± 9.68) = (22.36, 41.72)

None are younger than 22.36 = 0%; None are older than 41.72 = 0%

Exercise 10:

Mean = 795.24/24= 33.14

Median = 6.0 + 4.4 = 5.2

Difference = 33.14-5.2 = 27.94

The mean is much larger than the median, showing strong upward skew.

Range within which expect 95% to lie if normally distributed = (mean ± 2sd)

= (33.14 ± 2(81.224)) = (33.14 ± 162.45)

= (-129.31, 195.59)

No values are less than -129.31; in fact negative values are impossible. 1 value is greater than 195.59 = 4%

The expected range covers impossible values at the lower end, this again suggests positive skewing.

Exercise 11:

There are 21 values in each sample, the median is the middle (or 11th lowest) value.

Median NAA/Ch = 1.55

Median Lactate/Ch = 0.15

95% of the values would be expected to lie in the range (mean ± 1.96sd)

For NAA/Ch this is (1.56 ± 1.96(0.36)) = (0.85, 2.27)

For Lactate/CH the range is (0.28 ± 1.96(0.34)) = (-0.39, 0.95)

For NAA/Ch, none of the values lie below the range (0%), 2 lie above (9.5%)

The lower edge of the Lactate range is negative. Lactate/CH cannot be negative and hence none of the values lie below the range (0%). 1 value (4.8%) lies above the range.

Lactate/Ch is clearly non-normal. There are 9 zero values. The distribution appears to be J-shaped. Mean and standard deviation would not be used to summarise this data.

The range mean ± 2sd for NAA/Ch gives reasonable limits. The fact that 0% lie below and 9% lie above may suggest some upward skewing of the data. However the sample is small and hence each individual forms a relatively large percentage of the total. We would expect given normality to observe 1 individual outside the range in either direction. In this sample none are below and 2 are above, a discrepancy of only 1 either side against the figures expected.

Exercise 12:

20000 - 10024 = 9976 which will be approximately 1.64 standard deviations if the values are normally distributed (since 5% of the values would then be more than 1.64 SD greater than the mean).

So, the standard deviation would be approximately equal to 6082.83

(= 9976/1.64).

Approximately 95% of the data values would be expected to lie within the range (10024 ± 2(6082.83)) = (10024 ± 12165.86)

= (-2141.86, 22189.86)

Exercise 13:

i) 28.8 cm is (28.8 - 26.4)/3.1 = 0.77 standard deviations away from the mean

Referring to the normal distribution table, we find that 0.441 or 44.1% of the values lie more than 0.77 standard deviations away from the mean: Almost 22% in either direction i.e. 22% will be greater than 28.8cm.

ii) Similarly, (26.4 - 25.4)/3.1 = 0.32, and approximately 75% are further away from the mean i.e. 37.5% in either direction and hence 37.5% will be smaller than 25.4

Would expect 100-22-37.5 = 40.5% of the 28-30 weeks gestation babies to receive a maturity score of 1.

Exercise 14:

Taking a range (mean ± 2sd) for many of the variables gives an unfeasible range (e.g. negative duration of diabetes). This suggests that the mean and sd are not ideal summaries for the data, the p-values were also probably inappropriately calculated.

Exercise 15:

From the figures it appears that all of the distributions shown apart from total serum cholesterol amongst the mothers are upwardly skew. In the table the authors have logged the serum triglyceride values prior to summarising these. Given the upward skew of the data this is reasonable. Means and standard deviations are not useful summaries of skew data in its' raw form. The ranges given for the serum triglycerides (mean ± 1SD) would be expected to contain approximately 68% of the data values.

The total serum cholesterol values have not been transformed prior to summarising as the mean and standard deviation. These summaries are clearly inappropriate for the neonates. If the data were normally distributed, the range mean ± 2SD would be expected to contain approximately 95% of the data values. In the case of the neonate values this range would be from -0.33 to 3.71 and negative total serum cholesterols are impossible.

Exercise 16:

The raw scores for the ill high IQ group are clearly upwardly skew. There is a similar tendency amongst the ill low IQ and less so for the healthy low IQ group. The healthy high IQ group have a non-normal distribution but this is not upwardly skew. Taking logs of the scores will tend to normalise all of the groups except for the healthy high IQ group for which the effect may be to introduce downward skewing. Since there is not a single transformation the will normalise all of the subgroups simultaneously it may be safer to use non-parametric statistical tests to make comparisons between the scores for the 4 groups.

Exercise 17:

For all age groups the mean and median seem to be very close with small deviations. This would make us conclude that the normality assumption would be reasonable and probably lead to robust results.

Exercise 18:

The odds of being in level III of the maternity unit for pregnancy follow up is:

$$\frac{135}{(790\text{-}135)} = 0.206$$

And the odds of being in level I-II of the maternity ward of birth is:

$$\frac{163}{(790\text{-}163)} = 0.26$$

The odds of **not** receiving the therapy if in level III of the maternity unit for pregnancy follow up is:

$$\frac{17.8}{82.2} = 0.217$$

And the odds of **not** receiving therapy if in level I-II of the maternity ward of birth is:

$$\frac{60.1}{39.9} = 1.51$$

Exercise 19:

The normality assumption for numeric variables can be investigated by calculating the approximate 95% ranges (mean±2s.d.). E.g. if ages were normally distributed, we would expect about 95% of mothers in the calcium group to be aged between (20.8 ± 2x4.7 =) 11.4 and 30.2 which seems implausible. This indicates ages are not normally distributed and the median and IQR should have been used instead. The same issue exists for BMI and potentially age. For variables that have a reasonable 95% range (such as blood pressures), the mean and SD would be appropriate. Proportions and percentages are reasonable for categorical variables.

The odds of a mother to be smoking during pregnancy if she is in the calcium group is: $\frac{41}{210} = 0.195$ and the odds in the placebo group is $\frac{55}{203} = 0.27$.

Chapter 4

Exercise 1:

Odds of **not** receiving antenatal therapy in any of following categories:

Level of the maternity unit for pregnancy follow-up III $\frac{17.8}{82.2}$ = 0.217

Level of the maternity unit for pregnancy follow-up I-II $\frac{19.7}{80.3}$ = 0.245

Level of the maternity ward of birth III $\frac{8.8}{91.2}$ = 0.096

Level of the maternity ward of birth I-II $\frac{60.1}{39.9}$ = 1.506

Odds ratios of **not** receiving antenatal therapy are given at the last column of the following table:

		Antenatal Therapy	
	N (790)	No (n = 153) %	Yes (n = 637) %
Level of the maternity unit for pregnancy follow-up			
III	(135)	17.8	82.2
I-II	(655)	19.7	80.3
Antepartum transfer			
Follow-up and delivery in level III	(135)	17.8	82.2
Follow-up in level I-II, delivery in level III	(492)	6.3	93.7
Follow-up and delivery in levels I-II	(163)	60.1	39.9
Level of the maternity ward of birth			
III	(627)	8.8	91.2
I-II	(163)	60.1	39.9

Relative risk of **not** receiving antenatal therapy between each of the following cases:

Level of maternity unit for pregnancy follow-up III **and** I-II $\frac{19.7}{17.8}$ = 1.11

(reference category: level III)

Level of the maternity ward of birth III **and** I-II $\frac{60.1}{8.8}$ = 6.83

(reference category: level III)

Exercise 2:

i) Heights of normal and diseased children: difference in mean height
ii) Chances of obesity in UK and USA: difference in proportions of prevalence of obesity in the two countries (odds ratios)
iii) Brain weight of small and big sized animals: difference in mean weight
iv) Number of people that use public transport in UK and Germany: difference in proportions/percentages
v) Number of people that go on holidays within their countries or abroad: difference in proportions/percentages

Chapter 5

Exercise 1:

Standard error = Standard deviation/ Sqrt(sample size)

Hence, standard deviation = standard error x sqrt(sample size)

For the 28-30 weeks gestation babies, standard error= 0.5, sample size=38. So, standard deviation = 0.5 x sqrt(38) = 0.5 x 6.2 = 3.1.

For the 34-36 weeks gestation babies, standard error=0.3, sample size=41. So, standard deviation for this group = 0.3 x sqrt(41) = 0.3 x 6.4 = 1.92.

For the 37-39 weeks gestation babies, standard error=0.3, sample size=308. So, standard deviation for this group = 0.3 x sqrt(308) = 0.3 x 17.55 = 5.26.

The standard deviations tell us how variable the actual observations are. Approximately 95% of the head circumferences will lie within the range mean±2 standard errors.

The standard errors tell us how precisely the sample mean quantifies the population value. The standard errors are smaller for larger sample sizes. If the sample size is large then we are more sure that the estimate of the population mean head circumference (for that gestational age range) obtained from that sample will be accurate (or precise).

The standard deviations calculated above show that the head circumferences were much more variable amongst the 37-39 weeks gestation group. The standard errors for the 34-36 weeks and the 37-39 weeks groups show that the means were estimated with the same precision (despite the larger sample size in the 37-39 week group this is offset by the greater variability in that group).

Exercise 2:

From the 56 babies born to the cocaine only group:

sample mean birthweight=2648g, standard deviation=597g.

Hence, standard error = 597/sqrt(56) = 597/7.48 = 79.81

99% confidence interval for the population mean is given by:

$$2648 \pm 2.58(79.81) = 2648 \pm 206 = (2442, 2854g)$$

From the 27 babies born to the cocaine only group:

sample mean birthweight=2358g, standard deviation=698g.

Hence, standard error = 698/sqrt(27) = 698/5.2 = 134.23

99% confidence interval for the population mean is given by:

$$2358 \pm 2.58(134.23) = 2358 \pm 346 = (2012, 2704g)$$

The intervals do overlap (between 2442g and 2704g lies within both of the confidence intervals).

The confidence intervals give the range of population mean birthweights that the samples are compatible with. The fact that the intervals overlap shows that it is theoretically possible for the two samples to have come from populations with the same average birthweight.

Exercise 3:

Standard error = standard deviation/sqrt(sample size)

The standard deviation of the gestational ages = 4.84 within the sample of 24 infants, hence:

$$\text{Standard error} = 4.84/\sqrt{24} = 4.84/4.9 = 0.99$$

95% confidence interval for the population mean gestational age is given by:

$$32.04 \pm 1.96(0.99) = 32.04 \pm 1.94 = (30.1, 33.98 \text{ weeks})$$

(since the sample mean gestational age was 32.04 weeks).

The confidence interval gives the range of values within which we are reasonably confident (95%) that the population mean gestational age lies (based on this sample of 24 infants).

40 weeks does not lie within this interval, and is in fact a long way outside. Hence, we do not believe that the population mean gestational age is 40 weeks but it is not impossible. The population mean could be 40 weeks but it is very unlikely given the evidence from the current sample.

80% confidence interval is given by:

$$32.04 \pm 1.28(0.99) = 32.04 \pm 1.27 = (30.77, 33.31 \text{ weeks})$$

This interval is not as wide as the 95% confidence interval and we are less confident that it contains the population mean (80% as opposed to 95% confident). For 80% of samples the 80% confidence interval will contain the population mean, whereas the

95% confidence interval will contain the population mean for an additional 15% of samples.

Exercise 4:

Mean NAA/Ch (of the 21 patients) = 1.56, standard deviation 0.36

Hence, standard error = 0.36/sqrt(21) = 0.36/4.58 = 0.079 and this is what is used to construct the confidence intervals.

- 80% confidence interval = 1.56 ± 1.28(0.079) = 1.56±0.10 = (1.46, 1.66)
- 95% confidence interval = 1.56 ± 1.96(0.079) = 1.56±0.15 = (1.41, 1.71)
- 99% confidence interval = 1.56 ± 2.58(0.079) = 1.56±0.20 = (1.36, 1.76)

Exercise 5:

The percentage drug-screen positive = 9.09%

The sample is relatively small and the percentage reasonably extreme. Hence using the method for small samples and/or extreme proportions gives:

95% confidence interval (2.53, 27.81%)

99% confidence interval (1.80, 35.34%)

Exercise 6:

The preschool group were recruited from children attending their two year check up or one of four mother and toddler groups. These are unlikely to form a random sample of all preschool children. Children who attend their 2 year check up may be biased in some way, perhaps their parents are more health conscious. We are not told how the mother and toddler groups were chosen. Even if the groups were randomly selected and covered the range of social classes, the children who attend may not be representative of all preschoolers in the area the groups serve. The parents of children attending toddler groups may be less likely to be in full time employment for example.

The infant school group may be more representative as almost all normally developing children would be expected to attend school. However, uptake is low - only 66 children from 6 classes, and there is no guarantee that the children who responded were not biased in some way.

For the preschool group, standard error = Sqrt {(71.8(100-71.8))/39} = 7.21%

a) 80% confidence interval = (71.8±1.28(7.21)) = (62.6, 81.0%)
b) 95% confidence interval = (71.8±1.96(7.21)) = (57.7, 85.9%)
c) 99% confidence interval = (71.8±2.58(7.21)) = (53.2, 90.4%)

For the infant school group, standard error = Sqrt {(50.0(100-50.0))/66} = 6.15%

- 80% confidence interval = (50±1.28(6.15)) = (42.1, 57.9%)
- 95% confidence interval = (50±1.96(6.15)) = (37.9, 62.1%)
- 99% confidence interval = (50±2.58(6.15)) = (34.1, 65.9%)

The intervals give the range of values within which there is a certain confidence the population percentage lies. For greater confidence, the intervals are wider (i.e. the 99% confidence intervals are widest and the 80% narrowest).

Exercise 7:

According to the table, 13 patients had colonic cancer and this is 13/89 = 14.6% of the total. The 14 patients stated in the text is probably a mis-type as 14 would be nearer to 16% rather than the 15% given in the text. If we take the percentage to be 15, the standard error of the sample estimate is given by

$$\text{sqrt}((15(100-15))/89) = 3.78\%$$

A 95% confidence interval can be constructed:

$$15 \pm (1.96*3.78) = 15 \pm 7.41 = (7.59, 22.41\%)$$

If the population percentage is 11%, then the standard error of the sample estimate is given by:

sqrt((11(100-11))/89) = 3.32%

The following part of the solution refers to the next chapter, Significance testing.

The observed percentage (15) is (15-11)/3.32 = 1.20 standard errors away from 11 (the hypothesised percentage). By referring to the **Normal table (page 85 of notes)**, we see that the p-value is between 0.4 and 0.2, i.e. $0.2 < p < 0.4$.

Exercise 8:

Using the appropriate spreadsheet gives:

- 95% confidence interval: (0, 11.03%)
- 99% confidence interval: (0, 17.63%)

Zero adverse events from 31 infants does not imply that the drug is safe. We would not be very surprised to find none in 31 if as many as 11% (more than 1 in 10) of the babies given the drug had severe reactions.

Exercise 9:

a) The confidence interval of the odds ratio between age groups 28-29 and 30-31 (reference) is of interest. The odds ratio is 1.1 and the natural logarithm of it is 0.117.

The standard error of the natural logarithm of the odds ratio is given by:

se(log(e) of odds ratio)

$$= \sqrt{\frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}}$$

$$= \sqrt{\frac{1}{364 \cdot 0.181} + \frac{1}{364 \cdot (1-0.181)} + \frac{1}{231 \cdot 0.199} + \frac{1}{231(1-0.199)}}$$

$$= \sqrt{\frac{1}{65} + \frac{1}{298} + \frac{1}{46} + \frac{1}{185}}$$

$= \sqrt{0.046}$

$= 0.214$

95% confidence interval for the ln(OR) is

$$0.117 \pm 1.96\ (0.214) = (-0.302, 0.536)$$

In order to find the 95% confidence interval for the odds ratio (1.1) we need to exponentiate the two limits above.

95% confidence interval for the OR is

$$(\exp(-0.302), \exp(0.536))$$

$$(0.739, 1.709)$$

b) se(average daily intake) $\frac{sd}{\sqrt{n}} = \frac{0.735}{\sqrt{237}} = 0.05$

95% confidence interval for average daily intake (2.15)

$$2.15 \pm 1.96\ (0.05) = (2.052, 2.248)$$

c) se(difference between average daily intake between the calcium and placebo groups)

$$= \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \left(\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}\right)}$$

$$= \sqrt{\left(\frac{1}{237} + \frac{1}{237}\right) \times \left(\frac{0.735^2\ (237\text{-}1) + 0.711^2\ (237\text{-}1)}{237+237-2}\right)}$$

$$= \sqrt{\left(\frac{2}{237}\right) \times \left(\frac{127.5+119.3}{472}\right)}$$

$= 0.0664$

95% confidence interval for difference in average daily intake (2.15-2.30=0.15) between calcium and placebo groups:

$$0.15 \pm 1.96\ (0.0664) = (0.0198, 0.280)$$

d) se(% that complied with ultrasound scans at week 28 in the calcium group) =

$$\text{se} = \sqrt{\frac{81\ (100\text{-}81)}{237}} = 2.55$$

95% confidence interval for the percentage (81.0%)

$$81 \pm 1.96\ (2.55) = (76, 86)\ \%$$

e) se(for the difference in %s that complied with ultrasound scans at week 28 between the calcium and placebo groups) =

$$\sqrt{\frac{p_{1\%}(100 - p_{1\%})}{n_1} + \frac{p_{2\%}(100 - p_{2\%})}{n_2}} = \sqrt{\frac{81(100 - 81)}{237} + \frac{86(100 - 86)}{236}} = 3.4$$

95% confidence interval for the percentage (86-81=5%)

$$5 \pm 1.96\ (3.4) = (-1.7, 11.7)\%$$

f) Standard error of transformed correlation coefficient $\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$

$$\text{se} = \frac{1}{\sqrt{n\text{-}3}} = \frac{1}{\sqrt{61\text{-}3}} = 0.13$$

Transformed correlation coefficient:

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.82}{1-0.82}\right) = 1.16$$

95% confidence interval for transformed correlation coefficient:

$$1.16 \pm 1.96\ (0.13) = (0.90, 1.415)$$

Transform the above limits back to the correlation coefficient scale (r) by taking $\frac{e^{2t} - 1}{e^{2t} + 1}$ of each limit. This gives us the 95% confidence interval for the original correlation coefficient:

$$\left(\frac{e^{2(0.90)} - 1}{e^{2(0.90)} + 1}, \frac{e^{2(1.415)} - 1}{e^{2(1.415)} + 1} \right) = (0.71, 0.88)$$

Chapter 6

Exercise 1:

The difference between the average height found in the sample of 20 Jamaican boys and the hypothesised value of 86.5 is 86.5-84.25 = 2.25cm.

To determine whether this is a difference unlikely to have occurred by chance, we need to express it in terms of the number of standard errors.

Standard error = 3.16/sqrt(20) = 3.16/4.5 = 0.7 cm

Number of standard errors, therefore = 2.25/0.71 = 3.21 and referring this to the table of the normal distribution yields a p-value of 0.001, i.e. if 20 boys were taken from a sample with a mean height of 86.5 cm then this sample would have a mean height 2.25cm or more away from 86.5cm only 1 time in 1000.

This is a very small chance and it suggests that the Jamaican boys do not have a population mean height the same as that of the UK boys. Possible reasons might be:

- sickle cell is associated with reduced height
- Jamaican boys are on average shorter than UK boys
- measurements were made in some systematically different way in the two groups (e.g. the Jamaican boys may have been measured with a different device that systematically tended to give lower readings than the device used to measure the height of the UK boys.

It is impossible to separate the potential effects of ethnicity and presence of disease. We cannot be sure that the boys were measured in the same way and this could be avoided by having a concurrent comparison group.

Exercise 2:

A 95% confidence interval = 84.25±1.96(0.7) = 84.25±1.37 = (82.88, 85.62 cm)

A 99% confidence interval = 84.25±2.58(0.7) = 84.25±1.81 = (82.44, 86.06 cm)

These give the ranges within which we are 95 and 99% confident (respectively) that the population mean height of Jamaican boys with sickle cell disease lies (assuming that the 20 heights given are from a random sample of boys from this population). The 99% confidence interval is wider since we are more confident that it contains the population mean.

To estimate the population mean height more precisely we could want to decrease the width of the confidence interval. Since the confidence interval is constructed by taking multiples of the standard error either side of the sample mean, this decrease could be achieved by making the standard error smaller. The size of the standard error is directly inversely related to the sample size. Taking a larger sample will therefore reduce the size of the standard error and hence decrease the width of the confidence interval.

Exercise 3:

Standard error = 0.99

Sample mean = 32.04 weeks, which is 40-32.04 = 7.96 weeks away from the hypothesised value of 40.

7.96 weeks = 7.96/0.99 = 8.04 standard errors, and referring this value to the table of the normal distribution gives a p-value of less than 0.0005, i.e. the observed sample was highly unlikely to have occurred if the population mean gestational age (from which the sample was drawn) was 40 weeks.

The hypothesised value of 40 weeks was outside the 95% confidence interval of (30.1, 33.98 weeks), so, as expected, the p value was less than 0.05. The value of 40 weeks was a long way outside and the p value is much less than 0.05.

Exercise 4:

Standard error = 0.079

The sample mean of 1.56 is (1.67-1.56)/0.079 = 0.11/0.079 = 1.39 standard errors away from the hypothesised value of 1.67.

Referring the 1.39 standard errors to the table of the normal distribution gives a p-value of 0.165.

The 80% confidence interval was calculated to be (1.46, 1.66). Since 1.67 is outside this interval, a significance test would be expected to yield a p-value<0.2.

The 95% confidence interval was calculated to be (1.41, 1.71). Since 1.67 lies within this interval, a significance test would be expected to yield a p-value>0.05.

The results of the significance test are as expected given the previously calculated confidence intervals.

Exercise 5:

The confidence intervals and the p-values match as expected.

Difference in average daily intake: 95% c.i.: (0.085,0.21) and p-value=0.031. Both indicate that it is quite unlikely that the two groups have the same average daily intake.

Similarly for the difference in percentage that complied with ultrasound scans; 95% c.i.: (-2.3,12.3)% and p-value=0.173. Both indicate that it is quite likely that the calcium and placebo groups have the same percentage of people that comply with ultrasound scans.

Exercise 6:

The 2-sample t-test is only valid if:

- the samples need to be large enough to give reasonably accurate estimates of the population standard deviation
- the standard deviations of the two samples are not very different
- the values are approximately normally distributed in each of the samples

For both IgA sizes 1 and 2, there were very few rats (5) in each group. It is usually recommended that there are at least 20 per sample to ensure a reasonable estimate of population standard deviation.

For the IgA size 2 rats, the standard deviations of the two groups (35.8 and 226.0) are very different.

With the given information, it is not possible to comment on the normality of the values.

The use of the student's t-test is not valid for the IgA size 2 samples and may not be valid for the IgA size 1 samples.

Exercise 7:

The numbers used to calculate the standard error of the difference to be 147.88 are the sample sizes (56 and 27) and the sample standard deviations (597 and 698g).

We can obtain a 95% confidence interval for the mean difference of 290g of (0.16, 579.84g) and p=0.05 for the test of the null hypothesis of no difference in average birthweight between the groups.

The authors give a p-value of NS which means Non-Significant with p>0.05. Re-doing the calculations shows borderline significance with p=0.05 and the 95% confidence interval having a lower limit just above zero. The samples are compatible with a reduction in average birthweight amongst the cocaine plus group of anywhere between 0.16 and 579.84g. The lower limit of 0.16g is almost certainly of no clinical

significance, whereas the upper limit of 579.84g is probably of some importance. A larger sample size would be needed to be more precise about the difference associated with additional drug usage.

Exercise 8:

Average difference = 480.5 - 358.8 = 121.7 ng/ml.

Given that the sICAM-1 concentrations are normally distributed and that there are enough in each sample to give reasonable estimates of the population standard deviation, then the 2-sample t-test is appropriate to compare the means. (Note that the 3rd requirement for validity - not very different sample standard deviations - holds).

The null hypothesis is that there is zero average difference in sICAM concentrations between the two groups. The observed difference on 121.7 is 121.7/38.79 = 3.137 standard errors away from the hypothesised value of zero. This gives a p-value of 0.002.

A 95% confidence interval for the difference is given by 121.7 ± 1.96(38.79) = (45.7, 197.7 ng/ml).

Exercise 9:

The standard errors are based on the sample sizes (106 and 119) and the standard deviations of the measurements (16.7 and 15.6 for intelligence, 4.4 and 4.0 for self-esteem).

Intelligence: Short children have IQ on average 6 points lower (95% confidence interval 1.78, 10.22; p=0.005).

Self-esteem: Short children sore on average 0.8 points lower (95% confidence interval -0.3, 1.9; p=0.153).

The short children have significantly lower IQs, although the difference may actually be of a relatively unimportant amount (1.78 IQ points). This is an observational study

and hence any differences found may be explained by confounders. For example, the short children may be of a different social class, or have a different ethnic mix, and it is these features that are associated with IQ rather than height per se.

Exercise 10:

There are reasonable numbers of observations in each of the samples (55 and 110). Examination of the means and standard deviations suggests that DM duration may be skew (taking intervals mean ± 2sd would give negative lower limits and it is impossible to have a negative duration). Otherwise, there is nothing to suggest that the measures are not normally distributed. The largest difference in relative standard deviation estimates is for diabetic knowledge score, where the estimate for those with foot lcers is just over twice as large as that for the group without foot ulcers. For this outcome, the validity of a 2-sample t-test may therefore be questionable.

ABI: Mean difference -0.044 (95% ci (-0.12, 0.04)), p=0.284 (note discrepancy with published figure).

Diabetic Knowledge: Mean difference -1.84 (-2.92, -0.76), p=0.001.

Foot-care: Mean difference -1.24 (-2.11, -0.37), p=0.005 (again discrepancy but not so large).

Exercise 11:

The times were upwardly skew and hence log-transformation may normalise the data.

Deputising doctors took on average between 1.19 and 1.64 times longer to arrive than the practice doctors (i.e. there was an increase in the time waiting of between 19 and 64%).

Exercise 12:

Standard error = sqrt ((50(100-50))/200) = 3.54%

Difference between observed (50) and hypothesised (75) percentages = 25%.

Expressing this as a number of standard errors = 25/3.54 = 7.06, which when referred to normal tables yields $p < 0.0005$.

To calculate confidence intervals, the standard error based on the sample estimate is used (i.e. 3.54%).

- 80% confidence interval = (50 ± 1.28(3.54)) = (45.47, 54.53%)
- 95% confidence interval = (50 ± 1.96(3.54)) = (43.06, 56.94%)
- 99% confidence interval = (50 ± 2.58(3.54)) = (40.87, 59.13%)

The greater the confidence, the wider the interval. This is because the higher the confidence level, the more chance there is that the true population value lies within that range i.e. the range increases as the confidence becomes more.

Exercise 13:

	Autistic	**Non-autistic**	
Fragile X	18	52	70
No fragile X	126	202	328
	144	254	398

Research question: Is there an association between fragile X and autism?

Null hypothesis: The prevalence of fragile X syndrome is the same amongst autistics as non-autistics.

Percentage fragile X amongst autistics = 12.5, percentage fragile X amongst non-autistics = 20.5

Difference in percentages = 8%, standard error for the difference = 3.7%.

Testing the null hypothesis of zero difference between the percentages amongst autistic and non-autistic, p=0.03.

A 95% confidence interval for the difference is given by (0.6, 15.4)

The p-value of 0.03 shows that the observed difference would occur less than 4 times in 100 by chance if the null hypothesis were true. I.e. there is some evidence to suggest a difference over and above that expected by chance.

The 95% confidence interval shows that the sample is compatible with differences of anywhere between 0.6% and 15.4% more of one group being affected. The interval does not contain zero as anticipated given the $p<0.05$.

The p-value is not especially small (0.03) and the confidence interval is very close to zero at one limit (0.6%) and hence 'substantial' may be a bit exaggerated. Furthermore, it appears that a positive association is assumed (i.e. more fragile X syndrome patients amongst the autistics) and this paper was quoted in future papers as supporting this hypothesis. However, the data here are actually compatible with an excess of fragile X syndrome individuals amongst the **non**-autistics. I.e. the observed association is negative.

In interpreting the results it is important to determine how the samples of autistic and non-autistic individuals were chosen: Using samples from the general population would yield different results to taking samples from psychiatric units.

The reduction in sample size would lead to an increase in the standard error (less precision), which will lead to a wider confidence interval and larger p-value.

Exercise 14:

Percentages willing to have similar anaesthetic again:

- Music group = 80%
- Control group = 52%

Difference = 28%, standard error = 12.8%, p=0.029; 95% confidence interval (2.9, 53.1%)

The study provides some evidence that playing music changed the women's willingness. On average between a quarter and a third (28%) extra of the women played music were willing to have a repeat. The 95% confidence interval shows that

the data are compatible with that excess being somewhere in the region of 2.9 to 53.1%. The p-value shows that the difference was unlikely to have occurred by chance if the music playing had no effect.

Exercise 15:

	Peppermint Oil	**No peppermint oil**	
Relief	13 (68%)	6 (26%)	**19**
No relief	6	17	**23**
	19	**23**	**42**

Using the appropriate spreadsheet, gives p=0.003, 95% confidence interval for difference in percentages obtaining relief (12, 63.5%)

There is no difference in the percentages of individuals who did or did not take peppermint oil reporting relief.

The difference in percentages obtaining relief in the samples taken was very unlikely to have occurred by chance (3 chances in 1000).

The confidence interval shows that the data are compatible with anywhere between an additional 12 and an additional 63.5% of the group who took peppermint oil reporting relief.

To interpret the results, we need to know how the patients were divided into the groups who did and did not take peppermint oil. There may be confounders and these are more likely to be present if the study was observational. We also need to know how the patients were chosen and whether they are representative of individuals with irritable bowel syndrome.

Additionally, we would like to know whether the group not given the oil were given a placebo and whether the groups and the assessors were blind to treatment group.

How was 'relief' measured? Is the measurement known to be reliable and valid?

Exercise 16:

For the first set, the nurse managed 24.1%.

Standard error = sqrt ((24.1(100-24.1))/29) = 7.94% and

95% confidence interval = 24.1 ± 1.96(7.94) =

24.1 ± 15.56 = (8.54, 39.66%)

For the second set, the nurse managed 51.9%.

Standard error = sqrt((51.9(100-51.9))/27) = 9.62% and

95% confidence interval = 51.9 ± 1.96(9.62) =

51.9 ± 18.86 = (33.04, 70.76%)

Null hypothesis: There is no difference in ability to manage calls over time. Testing this gives p=0.026.

95% confidence interval for the difference in percentage of calls managed in the two sets of sessions is given by (3.3, 52.2%) more in the second compared to the first.

The data suggest an increase in the number of calls managed over and above any random fluctuations that could be expected.

Exercise 17:

Testing the null hypothesis that the number of repeats is not affected by warming the heel gives p=0.337 using 2 sample test of the difference in proportions/percentages.

The authors state that there were no significant differences and this is upheld by the p-value of 0.337. The results in the table are a little confusing in that p> 0.05 is shown with an asterisk and the conventional means would be to denote p<0.05 by asterisks.

The confidence interval shows that even though the data is compatible with no differences in the number of repeats attributable to heel warming (which corroborates the non-significant p-value), the data are also compatible with an increase of 23.7% repeats in the unwarmed group. Such a difference is likely to be of clinical importance. A large study is needed to determine whether there really is a difference due to heel warming or not. We cannot yet discount heel warming as an important factor in capillary blood sampling.

Exercise 18:

Some of the women do not seem to have responded (4 from the group allocated to RSG and 3 from the group allocated to CG).

Of the 30 RSG, 17 recalled at least one (16+4-3) = 56.7%. Of the 31 CG, 16 recalled at least one (14+3-1) = 51.6%.

Testing the null hypothesis of zero difference, p=0.692; 95% confidence interval for the difference (-19.9, 30.0%)

There is no evidence that the percentages differ in the 2 groups. However, the confidence interval is wide and fairly large differences cannot be discounted. To gain a more precise estimate of the difference between the groups, larger samples need to be tested.

Exercise 19:

4/60 = 6.7%, 7/39 = 17.95%; difference in percentage of depressed = 17.95-6.7 = 11.28%. Standard error for the difference = 6.94%

The difference is non-significant (p=0.104). A 95% confidence interval for the difference is given by (-2.3, 24.9%).

- 80% confidence interval = 11.28±1.28(6.94) = (2.4, 20.2%)
- 95% confidence interval = 11.28±1.96(6.94) = (-2.3, 24.9%)
- 99% confidence interval = 11.28±2.58(6.94) = (-6.6, 29.2%)

4/60 = 6.7%, 4/39 = 10.3%; difference in percentage who experienced a fall or syncopal episode = 3.6%. Standard error for the difference = 5.84%

- 80% confidence interval = 3.6±1.28(5.84) = (-3.9, 11.1%)
- 95% confidence interval = 3.6±1.96(5.84) = (-7.8, 15.1%)
- 99% confidence interval = 3.6±2.58(5.84) = (-11.5, 18.7%)

p=0.535; The percentage who experience a fall or syncopal episode is the same for the AD and CLB groups.

The difference observed in the percentages (3.6% more of the CLB group) was quite likely to have occurred by chance if the null hypothesis were true.

It is always important to consider the method of data collection when interpreting the results of analyses. What is of importance here particularly is whether the under-reporting is likely to differ between the comparison groups (AD and CLB).

Exercise 20:

All patients: p=0.101, 95% confidence interval for difference (-39.9, 3.6%)

PI>10: p=0.005, 95% confidence interval for difference (-68.8, -12.0%)

The confidence intervals are wide. In particular, when all patients are taken together, large differences cannot be excluded even though the result is non-significant.

Subgroup analyses should always be treated with caution. How was the cut-off PI>10 decided on? Was it a post-hoc decision? If PI is important, then the analysis should include this factor as a continuous covariate. Dichotomising PI is wasteful of the data and may obscure any trend with increasing PI values.

40% less patients were hospitalised in the prednisolone group. Hence 100/40 = 2.5 patients would need to be treated for one extra to benefit.

A 95% confidence interval for the percentage difference was (12.0, 68.8), hence a confidence interval for the number needed to treat goes from 100/68.8 to 100/12 i.e. (1.45, 8.33 patients).

Exercise 21:

Using the formulae and spreadsheet for small samples and/or extreme proportions:

9/150 = 6%, 95% confidence interval (3.2, 11.0%)

0/139 = 0%, 95% confidence interval (0, 2.7%)

Difference = 6%, 95% confidence interval for the difference = (2.1, 11.0%)

Chapter 7

Exercise 1:

Before therapy there does not seem to be anything 'odd' or distinctive about the distribution of VOCs/yr. The values are not obviously non-normally distributed.

During therapy, half of the patients have no recorded VOCs and the distribution of the numbers per patient is j-shaped (i.e. non-normally distributed).

The changes in VOCs for each patient before and during therapy are:

Patient	Change (during-before)
1	-3.1 (=0.8-3.9)
2	-2.9
3	-1.9
4	-1.7
5	-4.2
6	0.8
7	-0.2
8	-0.6
9	-3.0

10	-0.8
11	-1.8
12	-1.7
13	-0.3
14	0

The distribution of the differences is not obviously non-normal despite the fact that one of the components (the 'during therapy' values) are.

Note: It is the distribution of the within-pair differences that is important when considering the validity of the paired t-test. It is not uncommon to find distinctly non-normal distributions yielding normally distributed within pair differences.

The automatic standard deviation calculator can be used to obtain the mean difference of -1.53. Standard error = 1.42/sqrt(14) = 0.38.

The observed difference (-1.53) is 1.53/0.38 = 4.026 standard errors away from the hypothesised value of zero (i.e. no change on average during therapy when compared to 'before' values). This gives a p-value (using the normal distribution table) of < 0.0005. The low p-value shows that the differences seen during therapy were unlikely to have occurred by chance if there really were no change on average over this time period.

The 95% confidence interval is given by:

-1.53 ± 1.96(0.38) = (-2.27, -0.79), which shows that the sample of 14 individuals observed are compatible with an average population shift of a reduction of anywhere between 0.79 and 2.27 VOCs/yr.

However, we do not know whether this shift is due to the therapy as there is **no control group**. The ‘before’ measurements from the same individuals **do not form an adequate control** for the treatment period. The individuals were selected at a time when their disease was being particularly troublesome and it may be that left alone the individuals would improve (i.e. it could have been a temporary fluctuation in their illness which would have subsequently improved regardless). To establish whether treatment is causally related to changes in the frequency of VOCs would

require a randomised controlled trial with a control group consisting on untreated (or placebo) individuals.

Exercise 2:

The standard error of the difference = 27.05/sqrt(10) = 8.55

If we assume normality of the differences and that there are enough values (10) to give a reasonable estimate of the population spread, then a paired t-test can be performed. The number of standard errors which the observed average difference of 12.8 is away from the hypothesised value of zero difference (i.e. assuming population difference on average is zero - indicating no difference between the treatments) is 12.8/8.55 = 1.5, which when referred to a table of the normal distribution gives a p-value of 0.134. So, there is no evidence of a difference of the effects of the two treatments on systolic blood pressure.

A 95% confidence interval is given by (12.8 ± 1.96(8.55)) = (-3.96, 29.56 mmHg)

Although there is no significant difference in the effects of the two treatments (as shown by p=0.134), the confidence interval is wide and indicates that the data are compatible with as much as a 29 point reduction in blood pressure in favour of the new drug. A larger sample size is needed to make a more precise statement about the relative effects of the two treatments.

Exercise 3:

The study is concerned with the longitudinal analyses of 21 patients. Of interest are the within person changes over time and it would probably be more relevant to report the average changes over time (i.e. the average of the within person changes) rather than the average in 1976 for the entire group and the average in 1985 for the entire group. The pairing (within person) has been lost in the presentation of the results. If independent 2-sample t-tests were used to compare the averages, then the pairing within person would also be lost in the analysis. A paired t-test would be the preferable means of comparing changes over time within individual.

The authors argue that since the assays used at the different time points are comparable then the results can be compared directly. It is stated that the differences seen are probably attributable to the initiation of DF treatment in 1981. However, there were probably lots of other things that changed over the time period in question and it is impossible to say what the causal factor was. There is no evidence that the change is due to DF treatment initiation.

Exercise 4:

95% ci is given by $1.4 \pm 1.96(0.525) = (0.37, 2.43)$

If the pairing between individuals is ignored:

95% ci $= 1.4 \pm 1.96(0.69) = (0.05, 2.75)$, $p=0.042$

The standard error is larger, and hence the confidence interval wider, when the pairing is ignored. The pairing of the individuals has increased the precision with which differences between the groups are estimated as it removes the potential for age, sex and height to confound the comparison.

Exercise 5:

Values are paired before and after therapy within child but this pairing is lost in the displays of the data. Scatterplots showing how each child's initial measurement relates to their later one would be more informative.

For some of the measurements, there is a marked skewing of the values after treatment and this is a pattern not atypical whereby treatment has a pronounced effect in some, but not all, individuals. These differences in distributions are reflected in the table on examination of the differing standard deviation before and after therapy.

It is unclear whether the significance tests that were performed are paired. They should be. Even though there is some non-normality and very different standard deviations, the paired t-test may still be valid as, despite these problems with the separate distributions, the within child differences may be normally distributed.

There does not appear to be a control group of children who were not given the therapy. Any differences that occurred may be attributable to other causes than treatment if we do not have a control comparison group measured over the same period.

Exercise 6:

	Post treatment: Present	**Absent**	
Pre-treatment: Present	186	85	**271**
Absent	59	140	**199**
	245	**225**	**470**

Research question: Does Ivermection affect itching?

Null hypothesis: Ivermectin has no effect on itching.

Difference in percentage with itching pre and post treatment = 5.5% (57.6% pre, 52.1% post), standard error = 2.5%.

5.5/2.5 = 2.2, leading to p=0.028

95% confidence interval for paired difference in percentages with itching = (-10.5, -0.6%).

There is some evidence to suggest that Ivermectin reduces itching, but the confidence interval shows that the effect may be quite small (0.6% of patients benefit).

Chapter 8

Exercise 1:

Null hypothesis: There is no difference in the average CD4 counts between healthy and HIV-positive men (i.e. Healthy and HIV-positive men have the same average CD4 count.)

There are 20 values (10 per group) and the sum of all the ranks will equal 210 (=1+2+3+4+...+19+20). If the null hypothesis were true then we would expect the ranks to be evenly distributed between the 2 groups (healthy and HIV-positive) and the sum of the ranks in each group to be approximately equal i.e. in each group the sum would be expected to be about 105 (= half of 210).

The Mann-Whitney U-test could be used to determine whether there was an association between CD4 count and positivity. This hypothesis test is the non-parametric equivalent of the 2-sample t-test and is used to test the null hypothesis that the medians of 2 groups are equal.

The p-value is determined by considering how likely the observed spit of the rankings between the groups is if the null hypothesis were true. The split observed in the healthy and HIV-positive groups, 151 and 59, is quite different from that expected under the null hypothesis (i.e. 105 per group). Hence a small p-value would be expected (i.e. the split observed has a low probability of occurring by chance if the null hypothesis is true).

A p-value of less than 0.001 would mean that there was less than 1 chance in 1000 of observing such a split by chance if the null hypothesis were true. It would indicate and association between HIV-positivity and CD4 count but NOT that the relationship were necessarily causal. I.e. we cannot say that HIV-positivity AFFECTS CD4 count as this was an observational study and it may be that some confounder is the causal agent.

This was the highest ranked patient (rank 20) and if their value were 3504 they would still be the highest ranked. Since non-parametric tests are based on the ranks rather than the actual data values, the results would be unaffected by this change in the value (since the individuals' rank remains the same).

Exercise 2:

The ranks for cases 1 to 21 are in order:

4, 5, 9.5, 3, 1, 6.5, 9.5, 2, 6.5, 21, 12, 16, 8, 13.5, 11, 17, 20, 15, 13.5, 19, 18

Average for the 9 boys under 5 years = (4+5+9.5+3+1+6.5+9.5+2+6.5)/9 = 47/9 =5.22

Average for the 12 boys older than 5 years of age = (21+12+16+...+19+18)/12 = 184/12 = 15.33

Although the average rank is much higher in those over 5 years of age, this may be due to the use of a different measuring tool. The comparison of IQ with age is completely confounded by the change of measurement scale at 5 years of age. If there were some overlap between age and measurement scale, it may be possible to separate the effects of age and scale.

It is impossible to say from the information given. Validity implies that the assessment made truly reflects the persons' IQ.

Exercise 3:

At each time point (24-36 hours and 7 days) comparisons are made between 2 groups (buccal and oral) according to 6 different measurements (nausea, vomiting and sickness each according to severity and frequency). The outcomes being compared (severity and frequency) are rated on an ordinal scale, hence the chi-square for trend would be appropriate. A total of 12 tests for the 2 groups (buccal or oral); 3 outcomes (nausea, dizziness, vomiting) by 2 characteristics of each outcome (frequency and severity) measured at 2 time points (24-26 hours and 7 days).

The table gives the p-values if these are small (presumably all values <0.1 although this is not clear). For larger p-values, the actual values are not given and the table simply states that they are 'NS' (non-significant). It is bad practice not to give the actual p-values; - the interpretation of the results may differ according to whether p=0.1001 or p=0.96.

It would be better to give some estimate of the size of any differences found between the treatments - perhaps the percentage in each rating level with confidence interval. This would be much more informative than merely giving the p-values.

Chapter 9

Exercise 1

Some variation between sample estimates as expected, since each is based on different random samples. Differences are however small since the number of samples is large (10,000 for each) and so averages are similar.

In practice, would only do one set of bootstrap samples and use the summary statistics for this. Rather than 3 sets of 10,000; could do a single set of 30,000 to give potentially improved estimation.

The results from the 100,000 bootstrap estimates are similar to those from 10,000.

The results when only 20 bootstrap samples are taken, are more variable as would be expected.

All of the bootstrap confidence intervals contain the true population average (3263.57g).

In this instance the distribution of bootstrap means appears normally distributed and so the mean and median would be similar. For other bootstrap distributions this may not be the case. The median is always going to be a good measure of the centre of the distribution and denotes the point at which half the estimates are higher and half lower.

Exercise 2

Proportion = 7/57 = 0.123; n=57; standard error = sqrt((0.123 x (1-0.123))/57) = sqrt((0.123 x 0.877)/57) = sqrt (0.107871/57) = 0.0435.

95% ci = (0.123 ± 1.96 (0.0435)) = (0.123 ± 0.085) = (0.038, 0.208)

Chapter 10

Exercise 1

i. The one-way ANOVA was chosen as they wanted to compare a continuous variable (tests score) between three groups (learning environment). The assumptions of this test were:

- n > 20 for all groups
- normal distribution for all groups
- similar SD for all groups

All groups have over 20 participants and no group has a SD more than twice as large as another. To test the normality we can estimate the 95% range of scores using the mean and SD given in the table. For example, if the IACE scores for the blended learning group were normally distributed, then approximately 95% of the children in this group scored between 45.8 ± 2*34.6= -23.4, 115. A negative score or a score above 100% is impossible indicating this group is not normally distributed.

This comparison should have been carried out using the non-parametric equivalent of the one-way ANOVA, the Kruskal-Wallis ANOVA.

ii. This column is showing the results of ‘post-hoc’ tests, multiple testing has been carried out and the p-values have been adjusted to take account of this. This is used to identify where the significant difference between groups is arising from.

Exercise 2

i. Only ‘significant’ associations displayed, no p-values for ‘non-significant’ predictors – should always give full p-values rather than just ‘non-significant’ as could be as small as 0.051 or 0.9. p-values of 0 are impossible, should give p-values < a certain value. No confidence intervals given but standard error can be used to estimate these ourselves from the confidence interval formula.

ii. Regression coefficients are estimates of effects after adjusting for other predictors in the model. When the predictors included in the model change, so will the effect size.

iii. For every year increase in age, the outcome is expected to decrease by 4.523 after adjusting for education, gender, ethnicity, HADS-Total and SLE status. This decrease is significant ($p<0.005$).

iv. $[-4.523 \pm 1.96*0.586] = [-5.672, -3.374]$

This confidence interval does not contain the null hypothesis, 0 (no association), so the association between age and outcome is significant at the 5% level after adjusting for other predictors in the model. This agrees with the p-value.

Chapter 11

For the practicals in this Chapter, a selection of comments for each paper is listed below and these represent our views on the pros and cons of each table/figure. Some of the points made are based on subjective opinions and personal preference hence you might not agree with all and/or you might notice things that we might have overlooked.

Exercise 1

- Efficient display of observed percentages via the side by side chart.
- The 3-dimensions of the bars slightly hinders the clarity and message of the graph.
- Table 2 supplements Figure 1 as it contains the exact p-values calculated for the comparison made in Fig 1.
- Confidence intervals are shown in Table 1 which is definitely a positive.
- Both adjusted and unadjusted odds ratios given in Table 2 are a benefit to show both results before and after adjusting for confounders in the analysis (these results are likely to have been derived from logistic regression not covered in this course).

Exercise 2

- Successful representation of the associations between two numerical variables (scatterplot)
- As well as the display of the regression line, equation formula and R squared value for quick interpretation for the goodness of fit of the model.
- Adding the sample size on the graph would have assisted further interpretation due to the collection of points on the left hand side of Fig 3, i.e. to convince the reader that there was not a whole lot of zeros (on the second graph) that could possibly make the model not fit the data well.
- Similarly, different symbol design (such as hollow circles, or crosses) might improve the display of any overlapping observations.

Exercise 3

- The side-by-side bar charts display 4 different variables of interest across time.
- As the same variable has been measured at several days, a line plot would have been a better display for this kind of data, to show the progression of each pup.
- The legend does not specify if the bars shown represent the CI, the SE or the reference range.
- The mean along with the bars could be displayed on the left and right hand side of each graph and/or at the top or bottom of each day ‘column’ of points.
- Interpretation would be easier if a visual key was included to show what the solid grey and lined bars represent (this information is currently more difficult to pick out from the caption). Remember, the reader should be able to pick out the main message of the graph as easily as possible (even if only slight improvements can be made).

Exercise 4

- This graph tries to summarise the relationship between parental smoking and intellectual performance of sons via a numerical quantity called Effect

Size. We have not talked about this specific measurement during this module but it's obvious that neither of the variables mentioned above are shown on this graph – and this is the reason we don't advocate the use of the term 'effect size'.

- Ideally we would like to see a dot plot with the intellectual measurement on the y axis and the 2 columns of smoking or not on the x axis with different symbols used for the related and unrelated sons.
- Then the results of the adjusted analysis could be displayed in a table with exact p-values and CIs.

Exercise 5

- Overall the right type of graph has been chosen for each presentation, bar plots and line plots. Some might argue that graphs A and C are redundant as the two percentages shown there could had been simply included in the text alone.
- We'd expect exact p-values and CIs to be presented in the paper as the Figure is already very busy.
- Where possible, it is better to use the same scale when presenting comparable information. For example, the two graphs in D could be presented with the same y-scale. In graph B, this would be more difficult (the right-hand graph would have all the lines bunched up together if it was on a scale of 0 to 9 like the left-hand side graph).

Exercise 6

- A good visual display of the results that is perhaps appropriate for a poster to be presented at a scientific conference. You could argue that the display is not the best use of space for a journal article.
- It is questionable whether graphs A and C could had been replaced by line diagrams to keep to connection between the time points?

Exercise 7

- Exact p-values are presented, and the sample sizes are also included in the legend. However, full explanations of abbreviations should be included in a graph/table footer. For example, something like: “L-PAM= …”
- We would like to know whether correction for multiple tests has been performed (such as the Bonferroni correction).
- It would also be useful to have the CIs reported in the paper.