Statistics in Clinical Trials

Outline Solutions to Exercises

Dr Nick Fieller

Probability & Statistics, SoMaS

University of Sheffield

visiting

2012

(this page left blank for notes)

1) *Read the article referred to in §1.8, this can be accessed from the web address given there or from the link given in the course web pages. Use the facility on the BMJ web pages to find related articles both earlier and later.*

   Trust you have done this by now.

2) *Revision of t-tests and non-parametric tests.* And this also.

3) *Using your general knowledge compare the following two theories against the Bradford-Hill Criteria:*

   i) Smoking causes lung cancer

   Most of the criteria are satisfied. The weakest is whether or not there is a confounding factor that predisposes someone to smoke and that also increases the likelihood of developing lung cancer, possibly genetic. Establishing this criterion can be difficult in the absence of randomised controlled trials (out of the question with humans). The arguments against in this case are that there is evidence of passive smoking being harmful, clear evidence of links between smoking and other diseases (both other forms of cancer and non-cancer conditions such as heart disease), evidence of a link between chewing tobacco and cancers in site topically affected by tobacco juice (mouth and throat in particular).

   ii) The MMR (mumps, measles and rubella) vaccine given to young babies causes autism in later childhood.

This theory falls on several criteria. Firstly in terms of consistency, extensive studies in other countries have failed to find evidence of such a connection. In particular a very extensive study in Finland (I leave you to trace an account of this, try *googlescholar* and also Ben Goldacre's

Bad Science web page). Secondly, specificity is not easy to establish, thirdly no plausible biological mechanism explanation has been offered.

4) *For each of the proposed trials listed below, select the most appropriate study design, allocating onne design to onne trial. (Onne≡'one and only one'!)*

A→b
B→a
C→d
D→c

is the best allocation subject to the constraint of onne design used onnce. Some other design might be appropriate for the situation described, e.g. C→a

5) *In a recent radio programme an experiment was proposed to investigate whether common garden snails have a homing instinct and return to their 'home territory' if they are moved to some distance away.. The proposal is that you should collect a number of snails, mark them with a distinctly coloured nail varnish, and place all of them in your neighbour's garden. Your neighbour should do likewise (using a different colour) and place their snails in your garden. You and your neighbour should each observe how many snails returned to their own garden and how many stayed in their neighbour's. Full details are given at http://downloads.bbc.co.uk/radio4/so-you-want-to-be-a-scientist/Snail-Swapping-Experiment-Instructions.pdf*

(a) What flaws does the design of this experiment have?

(b) How could the design of the experiment be improved?

(Note: this question is open-ended and there are many possible acceptable answers to both parts. Discussion is intended)

This question was set in the context of the discussion in lectures of randomized double-blind controlled trials. So the first steps are to consider what the experimental and control groups and what is the 'intervention' (i.e. the action performed by the experimenter on the test subjects which might affect the measured outcome — the

intervention is performed on the experimental group but not on the control group). In this case the intervention is to move snails from their home territory and place them at some distance. The measured response is to see whether they return to their home territory. Examination of the design shows that ***there is no control group***. This is a major flaw in the design of the experiment. All of the snails caught in the owner's home garden are marked and placed in the neighbour's garden. Further, all of the snails marked by the neighbour in there garden are removed to the owner's garden. If the neighbour marked their snails and then released them back in their own garden then this would be a control group (since they would not have received the intervention). Without this control group you cannot rule out with any certainty whether snails always wander around quite a large territory covering adjacent gardens (remember the time scale is quite long – a week – between intervention and measurement of response).

A further, maybe less serious flaw, is that there is little randomization in the experiment. Presumably the snails that were captured and marked were not randomly selected from all of those in the garden but were those that were out and about and not hiding in obscure places. It is not realistic to catch all the snails in the garden and select a random sample to be exiled next door. However, a better design would be to catch say 2N snails in the owner's garden, randomly select N of them to be marked with one colour and then exiled next door, the other N would be marked with a different colour and allowed to stay at home. The neighbour could reciprocate with 2M snails, using two further colours. This would allow control of further potential explanatory factors such as whether snails naturally drift in one

direction along the road or whether one garden is particularly attractive to snails because of the presence of young green plants in only one of the gardens and these giving off aromatic signals detectable by snails. If snails equally migrate home in both directions and none of the control groups migrate then it does suggest that the homing instinct is because of homesickness rather than seeking food or some other attraction.

A further design issue is the question of blinding. It would be too easy to bias the results at the point of measurement of response towards a desired outcome by ['subconsciously' or otherwise] not collecting snails marked with the 'wrong' colour. Better would be for an independent third party who does not know the colour coding to collect all the marked snails they can find.

The results are given on http://www.bbc.co.uk/radio4/features/so-you-want-to-be-a-scientist/experiments/homing-snails/results/

Results

Totnes Garden

Returned to:

14H, 26A, 10m

Collected from:	H	A
Home	8	0
Away	0	9

Fisher's Exact Test: p = 0.00004

Cornwall Campus

Returned to:

54H, 74A, 30m

Collected from:	H	A
Home	8	1
Away	0	7

Fisher's Exact Test: p = 0.0014

10H, 8A, 8m

Collected from:	H	A
Home	8	0
Away	1	6

Fisher's Exact Test: p = 0.0014

81H, 68A, 8m

Collected from:	H	A
H	20	9
A	4	26

Fisher's Exact Test: p = 0.00001

Key: H,A = number of home and away snails; **m** = distance between bases; Fisher's test gives probability of getting our results by chance alone. Small p-values confirm homing instinct.

Findings in this experiment were complicated by a spell of exceptionally dry weather, during which many snails disappeared - presumably in shade and sealed up in their epiphragms. But in those instances where snails were recovered over short distances (up to 10 metres), there was again strong evidence of homing instinct. Over longer distances,

> particularly over 30 metres, results were inconclusive. This could have been due to the many variables: terrain, e.g. a wood; the type of barrier: e.g. road, building; the hot weather; or the actual distance itself.

This suggests the analysis presented was a Fisher's exact test (an alternative to a χ^2 test of independence) of a 2×2 contingency table, ignoring the fact that few of the snails marked were later found (especially in the 'Cornwall Campus'). A better analysis is invited from you.

The next three questions are designed to give practise at locating the available (if any) published evidence to support claims of medical connections. The most reliable evidence is provided by published (in peer-reviews journals) [double blind] randomized controlled clinical trials of adequate size.

6) *What evidence is there that grapefruit juice should never be taken by people receiving treatment for high cholesterol with statins?*

Even after looking in the standard resources (PubMed, clinicaltrials.gov etc) there is little evidence of any RCTs being conducted on realistic numbers subjects. The only one of any note is that below which reports on a trial with around 10 subjects in a 'treated and untreated group. The effect shewn was an increase in metabolism of the drug and so potentially adversely affecting health. However the amounts used in thetrial were substantially higher than those likely to be used in most people's daily diet.

Jari J. Lilja, Mikko Neuvonen & Pertti J. Neuvonen, 2004. *Effects of regular consumption of grapefruit juice on the pharmacokinetics of simvastatin*, British Journal of Clinical Pharmacology, 58:1, 56-60.

7) What evidence is there that taking fish oil helps schoolchildren concentrate

In summary the answer is very little evidence if any at all. A quick search on Ben Goldacre's page should lead you quickly to this article http://www.badscience.net/2010/06/the-return-of-a-2bn-fishy-friend/#more-1675 which tells much of the story. In short, this theory has been reported widely in many newspapers (including recently The Observer, a generally well-regarded Sunday Newspaper) as proven fact. Tracing the Observer article to its source reveals that the study referred to did not involve fish oil nor was it designed to test whether it helped schoolchildren concentrate. It is salutary reading.

8) On a recent BBC Radio programme (Front Row, Friday 03/10/08, http://www.bbc.co.uk/radio4/arts/frontrow/) there was an interview with Bettany Hughes, (http://www.bettanyhughes.co.uk/), a historian, who was talking about gold (in relation to an exhibition of a gold statue of Kate Moss in the British Museum). She made the surprising statement

"....ingesting gold can cure some forms of cancer."

I would only regard this as true if there has been a randomized controlled clinical trial where one of the treatments was gold taken by mouth and where the measured outcome was cure of a type of cancer. The task is to find a record of such a clinical trial or else find a plausible source that might explain this historian's rash statement.

The basis of this story seems to be reports that gold nano particles have been observed to bind to receptors on certain types of cancer cells. This is a long way from saying that gold *cures cancer.* Looking on clinicaltrials.gov and searching under 'gold' 'cancer' lists 80+ trials which include the two words 'gold' and 'cancer' somewhere in their protocols.

Several of these use 'gold' in the phrase 'gold standard' and don't involve administering actual gold. Others seem to involve studies where gold is not claimed to be the active agent but used as a delivery vehicle for some therapeutic agent bound to colloidal gold (gold pulverised to a very fine powder). I wasn't able to find details of a couple of Phase I trials (e.g. by Mayo Clinic) but no later phases and no links to publications were given.

Questions on Constructing Randomization Lists

Look at Martin Bland's guide to randomization software

http://www-users.york.ac.uk/~mb55/guide/randsery.htm

and see what is available and what descriptions there are available. He also has guides to many other sources for software of use in Medical Statistics.

9) *Patients are to be allocated randomly to 3 treatments. Construct a randomization list*

   i) for a simple, unrestricted random allocation of 24 patients

   ii) for a restricted allocation stratified on the following factors with 4 patients available in each factor combination:

   Sex: M or F Age: <30; 30≤&<50; ≥50.

i) e.g. take 1,2,3 → A; 4,5,6 → B; 7,8,9 → C; 0 → discard. Or in R:

```
> x<-c("A","B","C")
> y<-sample(x,24,replace=TRUE)
> y
 [1] "C" "B" "B" "A" "A" "A" "A" "A" "A" "A" "C" "B" "C" "C"
"A" "A" "C" "B" "A"
[20] "B" "C" "B" "A" "A"
```

iii) Would usually take 1→ABC; 2→ACB; 3→BAC; 4→BCA; 5→CAB; 6→CBA using randomly permuted blocks of size 3. However, there are only 4 patients available at each factor

combination. Possibilities are to choose 4th treatment for each factor combination (a) randomly or (b) selecting if one treatment is more important than the other 2 — then position that treatment randomly in the sequence (4 possible positions). Perhaps the best in the absence of other information is to take a random permutation of (A,B,C) for the three age groups for men and similarly for women.

More sophisticated in R is either:

```
> x<-LETTERS[1:3]
> lapply(rep(list(c(x,sample(x,1))),6),sample)
[[1]]
[1] "A" "C" "A" "B"
[[2]]
[1] "B" "A" "C" "A"
[[3]]
[1] "A" "B" "A" "C"
[[4]]
[1] "A" "A" "C" "B"
[[5]]
[1] "A" "A" "C" "B"
[[6]]
[1] "A" "B" "A" "C"
```

Which replicates a randomly chosen treatment for each of the six factor combinations or

```
> y<-matrix(x,3,6)
>
matrix(apply(y,2,sample),3,6);c(sample(x),sample(
x))
     [,1] [,2] [,3] [,4] [,5] [,6]
```

```
     [1,] "C"  "C"  "A"  "A"  "B"  "A"
     [2,] "B"  "A"  "C"  "C"  "A"  "C"
     [3,] "A"  "B"  "B"  "B"  "C"  "B"
     [1]  "C"  "A"  "B"  "A"  "C"  "B"
     >
```

where this is to be interpreted as giving the four treatments in the columns for the six factor combinations, preferably with the first three of one gender and the second three columns as the other.

10) Patients are to be randomly assigned to active and placebo treatments in the ratio 2:1. To ensure 'balance' a block size of 6 is to be used. Construct a randomisation list for a total sample size of 24.

There 15 (=6!/4!2!) blocks of size six of form AAAAPP. Note that a block size of 3 gives only 3 possibilities and so is unsatisfactory – too easy to crack. This can be done easily in R with `rep()` and `sample()`:

```
> sample(c(rep("A",4),rep("P",2)),6)
[1] "A" "A" "A" "P" "A" "P"
> sample(c(rep("A",4),rep("P",2)),6)
[1] "A" "A" "P" "P" "A" "A"
> sample(c(rep("A",4),rep("P",2)),6)
[1] "P" "A" "A" "A" "A" "P"
> sample(c(rep("A",4),rep("P",2)),6)
[1] "A" "A" "A" "P" "A" "P"
>
```

More sophisticated is

```
matrix(apply(matrix(c(rep("A",4),rep("P",2)),6,4),2,sample),1,6*4)
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10]
       [,11] [,12] [,13] [,14]
[1,] "A"   "A"   "A"   "P"   "P"   "A"   "A"   "P"   "A"   "P"    "A"
"A"   "A"   "P"
       [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23]
       [,24]
[1,] "A"   "A"   "A"   "P"   "A"   "A"   "P"   "A"   "A"   "P"
>
```

11) Patients are to be randomly assigned to active and placebo treatments in the ratio 3:2. To ensure 'balance' a block size of 5 is to be used. Construct a randomisation list for a total sample size of 30.

There are 10 (=5!/3!2!) blocks of size 5 of form AAAPP. Note that a block size of 10 of form AAAPPAAAPP would give 10!/6!4!=210 possibilities, perhaps too many (overkill), 10 possibilities with block size 5 is probably adequate and not easy to crack, or else take random subset of these of say 5 sets.

Either use repeatedly:

```
sample(c(rep("A",3),rep("P",2)),5)
[1] "A" "A" "P" "A" "P"
>
```

Or, more sophisticated

```
>
matrix(apply(matrix(c(rep("A",3),rep("P",2)),5,6),2,sample),1,5*6)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,] "A"  "A"  "A"  "P"  "P"  "P"  "A"  "A"  "P"  "A"   "A"   "P"   "A"   "P"
     [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26]
[1,] "A"   "A"   "P"   "A"   "A"   "P"   "A"   "P"   "A"   "P"   "A"   "A"
     [,27] [,28] [,29] [,30]
[1,] "P"   "A"   "P"   "A"
```

12) In the comparison of a new drug A with a standard drug B it is required that patients are assigned to drugs A and B in the proportions 3:1 respectively. Illustrate how this may be achieved for a group of 32 patients, and provide an appropriate randomization list. Comment on the rationale for selecting a greater proportion of patients for drug A.

(i) Need blocks of form AAAB (or of form AAAAAABB). There are 4 of form AAAB (and 28 of size 8). Using 1,2→AAAB; 3,4→AABA; 5,6→ABAA; 7,8→BAAA, 9,0→ignore, a sequence of random digits

7,1,4,2,0,1,8,1,2,4 gives

BAAA|AAAB|AABA|AAAB|AAAB|BAAA|AAAB|AAAB.

In **R**, to produce a random block of form AAAB do:

```
> sample(c(rep("A",3),"B"))
[1] "A" "A" "B" "A"
```

and then repeat as often as necessary or build into a loop. Alternatively, to get exact balance without blocks do:

```
> sample(c(rep("A",24),rep("B",8)))
 [1] "B" "B" "A" "A" "A" "A" "A" "B" "B" "A" "A" "A" "A" "A"
[15] "A" "A" "A" "A" "A" "A" "A" "A" "B" "A" "B" "A" "A" "B"
[29] "A" "B" "A" "A"
```

which is not entirely satisfactory even if there are exactly 32 subjects and the trial does not stop early because there could be imbalance with respect to time if there are moBs towards the beginning or end. Dependence on time might be one of the potential covariates that use of randomization protects against.

There could be economic reasons for using more As than Bs, but more likely if B is the standard then there will be interest in efficacy and safety of the new treatment but this is likely to be known for the standard, as would be drop out rates, standard deviations etc. Having more patients on the new treatment protects against uncertainty in drop-out rates (or side effects) and consistency of response. Further, there will be more interest and enthusiasm amongst both patients and investigators if there is a greater chance of receiving the new treatment and so easier to recruit centres and patients. This last reason is probably the most important in practice though not obviously 'statistical'.

13) The table below gives the age (≤55/>55), gender (M/F), disease stage (I/II/III) of subjects entering a randomized controlled clinical trial at various intervals and who are to be allocated to treatment or placebo in approximately equal proportions immediately on entry.

order of entry	Age	Gender	Stage
1	≤55	F	III
2	≤55	M	III
3	≤55	M	I
4	≤55	F	I
5	>55	F	II
6	≤55	F	III
7	>55	F	I
8	>55	M	III
9	≤55	M	III
10	>55	F	III
11	≤55	F	III
12	≤55	M	I
13	>55	F	I

i) Use a minimization method designed to achieve an overall balance between the factors to allocate these subjects in the order given to the two treatments and provide the resulting list of allocations.

				First Run		Second Run	
order of entry	Age	Gender	Stage	score for T	score for P	score for T	score for P
1	≤55	F	III	0★	0	0	0★
2	≤55	M	III	2	0★	0★	2
3	≤55	M	I	1★	2	2	1★
4	≤55	F	I	4	1★	1★	3
5	>55	F	II	1	1★	1	1★
6	≤55	F	III	4★	5	4★	5
7	>55	F	I	3★	4	3★	4
8	>55	M	III	4	3★	4	3★
9	≤55	M	III	6★	6	6	6★
10	>55	F	III	7	6★	6★	7
11	≤55	F	III	9	8★	10	8★
12	≤55	M	I	8	6★	6★	7
13	>55	F	I	6★	9	9	6★

The first subject has to be allocated randomly to T or P. The ★ indicates which of T or P is selected. Then for each subsequent subject it is easy to calculate the score for T and P as the total number of characteristics held in common between the new arrival and those subjects already allocated to that group. Two runs are presented above, one resulting from a choice of T for the first subject — this leads to a tied score for the 5th subject and P was [randomly] chosen, another tie for the 9th and T was [randomly] chosen. The second run with P selected first also leads to a tie on the 5th arrival and then the 9th.

ii) Cross-tabulate the treatment received with each [separate] factor.

Run 1:

	Age			Gender			Stage			
	≤55	>55	total	M	F	total	I	II	III	total
T	4	2	6	2	4	6	3	0	3	6
P	4	3	7	3	4	7	2	1	4	7
total	8	5	13	5	8	13	5	1	7	13

Run 2:

	Age			Gender			Stage			
	≤55	>55	total	M	F	total	I	II	III	total
T	4	2	6	2	4	6	3	0	3	6
P	4	3	7	3	4	7	2	1	4	7
total	8	5	13	5	8	13	5	1	7	13

Note that these are identical, as are essentially all possible runs (i.e. up to an interchange of T and P). Even with a different order of arrival of these patients the final allocations are not substantially different.

iii) Construct a list to allocate the subjects to treatment completely randomly without taking any account of any prognostic factor and compare the balance between treatment groups achieved on each of the factors.

In **R** the function `sample(.)` with the `replace=TRUE` option gives the same facility:

```
> sample(c("T","P"),13,replace=TRUE)
 [1] "T" "P" "T" "T" "T" "T" "T" "T" "P" "P" "T" "P" "T" "
```

	Age			Gender			Stage			
	≤55	>55	total	M	F	total	I	II	III	total
T	6	3	9	3	6	9	4	0	5	9
P	2	2	4	2	2	4	1	1	2	4
total	8	5	13	5	8	13	5	1	7	13

(Different randomisations will lead to different cross-tabulations.)

Sample size questions

(in all cases take the significance level as 0.05)

The commands in **R** for calculation of power, sample size etc are `power.t.test()` and `power.prop.test()`. Note that typing the ↑ recalls the last **R** command and use of Backspace and the ← key allows you to edit the command and run a new version

*14) How many subjects are needed to achieve a power of 80% when the standard deviation is 1.5 to detect a difference in two populations means of 0.8 using a two sample t-test? (Note that **R** gives the number needed in each group, i.e. total is twice number given)*

```
> power.t.test(sd=1.5,power=.8,delta=0.8)

     Two-sample t test power calculation
              n = 56.16413
          delta = 0.8
             sd = 1.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
 NOTE: n is number in *each* group
```

So we need 57 in each group (note we need to round fractional sample sizes **up** to nearest integer) and therefore 114 in total.

15) How many subjects are needed to achieve a power of 80% when the standard deviation is 1.5 to detect a difference in one population mean from a specified value of 0.8 using a one sample t–test?

```
>
power.t.test(sd=1.5,power=.8,delta=0.8,type="one.sample")

     One-sample t test power calculation
              n = 29.57195
          delta = 0.8
             sd = 1.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Thus we need 30 subjects.

16) *Do you have an explanation for why the total numbers in Q14 and Q15 are so different?*

Some people might think that if you need N for specified power and delta with a one sample test then you need 2N for a two sample test but in fact you will need about 4N. My personal 'explanation/visualisation' of what is happening is that with two samples each sample mean can be either above or below the target population mean – it is only when they are both as far away from the other population mean as possible that the strongest evidence of a difference in population means is provided. This is only one of the four possible combinations of whether the two sample means are above or below their population means. Perhaps a more technical explanation is that two variances have to be estimated rather than only one.

17) *How many subjects are needed to detect a change of 20% from a standard incidence rate of 50% using a two sample test of proportions with a power of 90%?*

```
> power.prop.test(power=.9,p1=.5,p2=.7)
     Two-sample    comparison    of    proportions    power
calculation
              n = 123.9986
             p1 = 0.5
             p2 = 0.7
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
 NOTE: n is number in *each* group

> power.prop.test(power=.9,p1=.5,p2=.3)

     Two-sample    comparison    of    proportions    power
calculation
              n = 123.9986
             p1 = 0.5
             p2 = 0.3
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

 NOTE: n is number in *each* group
```

Note that it does not matter whether the change from .5 is up or down. Rounding up we see we need 124 in each group so 248 in total.

18) How many subjects are needed to detect a change from 30% to 10% using a two sample test of proportions with a power of 90%?

```
power.prop.test(power=.9,p1=.1,p2=.3)

     Two-sample comparison of proportions power
calculation

              n = 81.96206
             p1 = 0.1
             p2 = 0.3
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

 NOTE: n is number in *each* group
```

So we need 164 in total.

19) How many subjects are needed to detect a change from 60% to 80% using a two sample test of proportions with a power of 90%?

```
> power.prop.test(power=0.9,p1=.6,p2=.8)

     Two-sample comparison of proportions power
calculation
              n = 108.2355
```

So we need 218 in total

20) How many subjects are needed to detect a change from 50% to 30% using a two sample test of proportions with a power of 90%?

You should have answered this in Q17

21) How many subjects are needed to detect a change from 75% to 55% using a two sample test of proportions with a power of 90%?

```
> power.prop.test(power=0.9,p1=.75,p2=.55)

     Two-sample comparison of proportions power
calculation
              n = 117.4307
```

So 236 in total.

22) How many subjects are needed to detect a change from 40% to 60% using a two sample test of proportions with a power of 90%?

```
> power.prop.test(power=0.9,p1=.4,p2=.6)

     Two-sample comparison of proportions power
calculation
              n = 129.2529
```

So 260 in total.

23) *Questions 17–21 all involve changes of 20% and a power of 90%. Why are the answers not all identical?*

It is because when estimating a proportion as the number of success r out of n trials the standard error of the estimate is $(r/n(1–r/n)/n)^{½}$ which is a maximum when r/n=½, i.e. proportions closer to 0.5 require a greater sample size for a specified precision than those further from 0.5.

*24) Without doing any calculations (neither by hand nor in **R**) write down the number of subjects needed to detect a change from 45% to 25% using a two sample test of proportions with a power of 90*

236 in total (same as Q21).

25) A trial for the relief of pain in patients with osteoarthritis of the knee is being planned on the basis of a pilot survey which gave a 25% placebo response rate against a 45% active treatment response rate.

a) *How many patients will be needed to be recruited to a trial which in a two-sided 5% level test will detect a difference of this order of magnitude with 90% power? (Calculate this first 'by hand' and then using a computer package and compare the answers).*

b) *With equal numbers in placebo and active groups, what active rates would be detected with power in the range 50% to 95% and group sizes 60 to 140? (Calculate for power in steps of 15% and group sizes in steps of 20).*

26) Woollard & Cooper (1983) Clinical Trials Journal, 20, 89-97, report a clinical trial comparing Moducren and Propranolol as initial therapies in essential hypertension. These authors propose to compare the change in initial blood pressure under the two drugs.

i) Given that they can recruit only 100 patients in total to the study, calculate the approximate power of the two-sided 5% level t-test which will detect a difference in mean values of 0.5σ, where σ is the common standard deviation.

```
> power.t.test(n=50,sd=1,delta=.5)

     Two-sample t test power calculation

              n = 50
          delta = 0.5
             sd = 1
      sig.level = 0.05
          power = 0.6968888
    alternative = two.sided

 NOTE: n is number in *each* group
```

Note that the sample size in each group is 50 (total 100). Also note that a CRD of ½σ means you enter the standard deviation as 1.0 and the CRD as ½.

The programme power.exe gives a value for the power of 69.69%. (The formula for the approximation may give a slightly different answer).

ii) How big a sample would be needed in each group if they required a power of 95%? (Calculate this first 'by hand' and then using a computer package and compare the answers).

```
> power.t.test(power=0.95,sd=1,delta=.5)

     Two-sample t test power calculation

              n = 104.9280
          delta = 0.5
             sd = 1
      sig.level = 0.05
          power = 0.95
    alternative = two.sided
NOTE: n is number in *each* group
```

Programme power.exe gives 105 in each group (210 in total).

Cross-Over Trials

*27) Senn and Auclair (Statistics in Medicine, 1990, **9**) report on the results of a clinical trial to compare the effects of single inhaled doses of 200μg salbutamol (a well established bronchodilator) and 12μg formoterol (a more recently developed bronchodilator) for children with moderate or severe asthma. A two-treatment, two-period crossover design was used with 13 children entering the trial, and the observations of the peak expiratory flow, a measure of lung function where large values are associated with good responses, were taken. The following summary of the data is provided.*

	Group 1: formoterol $\rightarrow$ salbutamol ($n_1 = 7$)			
	Period 1	*Period 2*	*Sum (1 + 2)*	*Difference(1 - 2)*
mean	*337.1*	*306.4*	*643.6*	*30.7*
s.d.	*53.8*	*64.7*	*114.3*	*33.0*
	Group 2: salbutamol $\rightarrow$ formoterol ($n_2 = 6$)			
	Period 1	*Period 2*	*Sum (1 + 2)*	*Difference(1 - 2)*
mean	*283.3*	*345.8*	*629.2*	*-62.6*
s.d.	*105.4*	*70.9*	*174.0*	*44.7*

a) Specify a model for peak expiratory flow which incorporates treatment, period and carryover effects.

a) Model: usual one in notes. It is a good idea to plot the means for each group for each period (not shewn here) and then see that it is suggestive that treatment 2 is superior, no obvious carryover nor period effects.

b) *Assess the carryover effect, and, if appropriate, investigate treatment differences. In each case specify the hypotheses of interest and illustrate the appropriateness of the test.*

Carryover: t=0.17 [=(643.6-629.2)/(114.3^2/7+174^2/6)$^{-½}$] p>>0.05, so can proceed with treatment & period tests:

Treatment: t=4.22 [=(30.7-(-62.6))/(33.0^2/7+44.7^2/6)$^{-½}$] on 6 d.f., p<0.01, so clear evidence of a difference between the treatments. Inspection of the means shews that formoterol is superior.

Period: t=–1.44 (on 6 df), p=0.2, no evidence of a systematic difference between periods.

(demonstrate appropriateness of tests by reference to model as in notes).

Conclude that there is strong evidence that formoterol gives a better response than salbutamol.

28) A and B are two hypnosis treatments given to insomniacs one week apart. The order of receiving the treatment is randomized between patients. The measured response is the number of hours sleep during the night. Data are given in the following table.

patient	***period 1***		***period 2***	
1	*A*	*9*	*B*	*0*
2	*B*	*11*	*A*	*14*
3	*B*	*7*	*A*	*3*
4	*B*	*12*	*A*	*8*
5	*A*	*8*	*B*	*8*
6	*A*	*11*	*B*	*1*
7	*A*	*4*	*B*	*4*
8	*B*	*3*	*A*	*4*
9	*A*	*13*	*B*	*2*
10	*B*	*7*	*A*	*3*
11	*A*	*1*	*B*	*2*
12	*A*	*13*	*B*	*1*
13	*A*	*6*	*B*	*3*
14	*B*	*5*	*A*	*6*
15	*B*	*6*	*A*	*8*
16	*B*	*3*	*A*	*7*

a) Calculate the mean for each treatment in each period and display the results graphically.

b) Assess the carryover effect.

c) If appropriate, assess the treatment and period effects.

(NB These data are available in on the course web pages)

```
> hourssleep
   period1 period2 group  sum period.diffs treat.diffs
1        9       0     1  4.5            9           9
2       11      14     2 12.5           -3           3
3        7       3     2  5.0            4          -4
4       12       8     2 10.0            4          -4
5        8       8     1  8.0            0           0
6       11       1     1  6.0           10          10
7        4       4     1  4.0            0           0
8        3       4     2  3.5           -1           1
9       13       2     1  7.5           11          11
10       7       3     2  5.0            4          -4
11       1       2     1  1.5           -1          -1
12      13       1     1  7.0           12          12
13       6       3     1  4.5            3           3
14       5       6     2  5.5           -1           1
15       6       8     2  7.0           -2           2
16       3       7     2  5.0           -4           4
> attach(hourssleep)
> t.test(sum~group)

        Welch Two Sample t-test

data:  sum by group
t = -0.9929, df = 12.64, p-value = 0.3394
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -4.176408  1.551408
sample estimates:
mean in group 1 mean in group 2
         5.3750          6.6875

> t.test(treat.diffs~group)

        Welch Two Sample t-test

data:  treat.diffs by group
t = 2.4597, df = 11.543, p-value = 0.03077
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  0.6203012 10.6296988
sample estimates:
mean in group 1 mean in group 2
          5.500          -0.125

> t.test(period.diffs~group)

        Welch Two Sample t-test

data:  period.diffs by group
t = 2.3503, df = 11.543, p-value = 0.03746
```

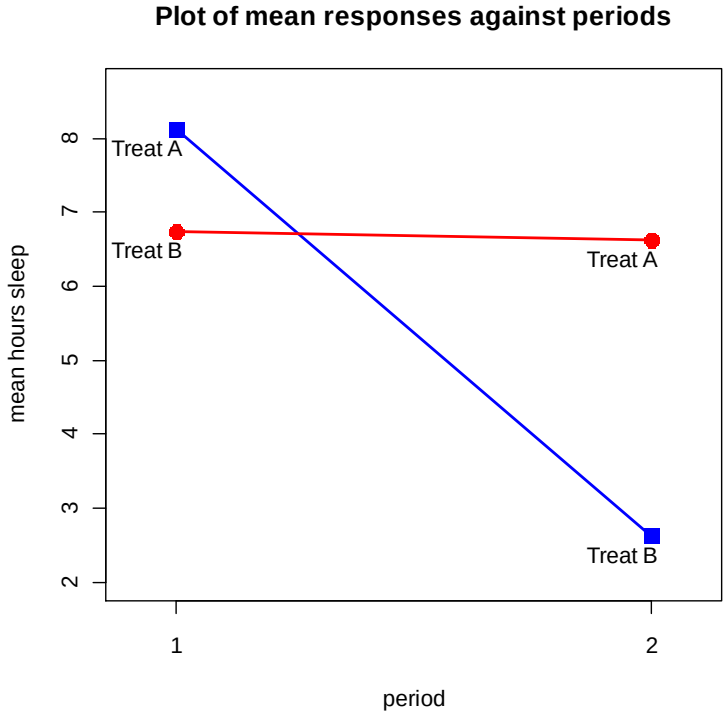

```
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  0.3703012 10.3796988
sample estimates:
mean in group 1 mean in group 2
          5.500           0.125

>
```

Group 1: A→ B (n_1 = 8)				
	Period 1	Period 2	Sum (1 + 2)	Difference(1 - 2)
mean	8.13	2.625	5.375	5.50
s.d.	4.29	2.50	2.16	5.53
Group 2: B→ A(n_2 = 8)				
	Period 1	Period 2	Sum (1 + 2)	Difference(1 - 2)
mean	6.75	6.69	6.69	0.13
s.d.	3.33	3.62	3.05	3.36

The following R code will produce a 'nice' plot of mean responses but it is probably sufficient in most routine cases to produce a quick one by hand.

```
> GP1PER1mean<-mean(PERIOD1[GROUP==1])
> GP1PER2mean<-mean(PERIOD2[GROUP==1])
> GP2PER1mean<-mean(PERIOD1[GROUP==2])
> GP2PER2mean<-mean(PERIOD2[GROUP==2])
> per<-c(1,2)
> gp1<-c(GP1PER1mean,GP1PER2mean)
> gp2<-c(GP2PER1mean,GP2PER2mean)
> ymax<-max(GP1PER1mean,GP1PER2mean,GP2PER1mean,GP2PER2mean)
> ymin<-min(GP1PER1mean,GP1PER2mean,GP2PER1mean,GP2PER2mean)
> ymax<-ymax+0.1*(ymax-ymin)
> ymin<-ymin-0.1*(ymax-ymin)
> plot(xlim<-c(0.9,2.1),ylim<-
c(ymin,ymax),type="n",xlab="period",
+ ylab="mean hours sleep",xaxt="n",
+ main="Plot of mean responses against periods")
> axis(1,at=c(1,2))
> points(per,gp1,pch=15,col="blue",cex=1.5)
> points(per,gp2,pch=16,col="red",cex=1.5)
> lines(per,gp1,col="blue",lwd=2)
> lines(per,gp2,col="red",lwd=2)
> gp1labels<-c("Treat A","Treat B")
> text(per,gp1,labels=gp1labels,adj=c(.9,1.4))
> gp2labels<-c("Treat B","Treat A")
> text(per,gp2,labels=gp2labels,adj=c(.9,1.4))
```



Note that plot suggests that A is better than B and that there is a period effect (the average results in period 2 are lower than those in period 1). Whether there is a carryover effect is a more difficult matter of judgement. If there is carryover then it is quite complex and not only is B persisting to depress the results on A for group 2 but A is interacting with B to produce substantially lower results in period 2 for group 1. It would be surprising that such and interaction would be so different for the two groups. A simpler explanation (i.e. use Occam's Razor) is that it is a combination of period and treatment effects. This is not contradicted by the formal statistical tests. These are (taking values from output — though you could do this from the summary statistics in the table above using the two sample t-test used in the first question, though with a conservative d.f. = 8 rather than **R'**s calculated values of 11 or 12).

Carryover: t = –0.99, d.f.=12, p=0.340, no evidence.

Period: t = 2.46, d.f.=11, p=0.032, good evidence of difference in periods with period 2 lower than the first period.

Treatment: t = 2.35, d.f.=11, p=0.038, good evidence that A is better than B.

29) Two ointments A and B have been widely used for the treatment of athlete's foot. In a recent report the following results were noted, where response indicated temporary relief from the outbreak.

	Response	*No Response*
Ointment A	*174*	*96*
Ointment B	*149*	*121*

a) Based on these results the report concluded that ointment A was more effective than ointment B. Use the Mantel-Haenszel test to verify this conclusion.

b) Further investigation into the source of the data revealed that the data had been pooled from two clinics. The results from individual clinics were:

	Ointment A		*Ointment B*	
Clinic	*Response*	*No response*	*Response*	*No response*
1	*129*	*71*	*113*	*87*
2	*45*	*25*	*36*	*34*

Reassess the evidence in the light of these additional facts.

Use the formulae in the noes,

Overall : $E[Y_1]=161.5$, $var(Y_1)=32.50$, $\chi^2_{MH}=4.8$; $p<0.05$

Clinic 1: $E[Y_1]=121.0$, $var(Y_1)=23.96$, $\chi^2_{MH}=2.67$; $p>0.05$

Clinic 2: $E[Y_1]= 40.5$, $var(Y_1)= 8.59$, $\chi^2_{MH}=2.36$; $p>0.05$

Conclude that there is very strong evidence that A is more effective. (response rates are 64.5%, and 64.3% — very close, so few doubts on validity of combining results.)

Below is a complete analysis in R:

```
> x<-factor(rep(c(1,2),c(200,200)),labels=c("Oint A","Oint B"))
> y<-factor(rep(c(1,2,1,2),c(129,71,113,87)),labels=c("Response","No
   Response"))
> z<-factor(rep(1,400),labels="Clinic 1")
> table(x,y,z)
, , z = Clinic 1

        y
x        Response No Response
  Oint A      129          71
  Oint B      113          87

> mantelhaen.test(x,y,z,correct=F)

        Mantel-Haenszel chi-squared test without continuity
   correction

data:  x and y and z
Mantel-Haenszel X-squared = 2.6714, df = 1, p-value = 0.1022
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.9353062 2.0921389
sample estimates:
common odds ratio
         1.398853

>
> x<-factor(rep(c(1,2),c(70,70)),labels=c("Oint A","Oint B"))
> y<-factor(rep(c(1,2,1,2),c(45,25,36,34)),labels=c("Response","No
   Response"))
> z<-factor(rep(1,140),labels="Clinic 2")
> table(x,y,z)
, , z = Clinic 2

        y
x        Response No Response
  Oint A       45          25
  Oint B       36          34

> mantelhaen.test(x,y,z,correct=F)

        Mantel-Haenszel chi-squared test without continuity
   correction

data:  x and y and z
Mantel-Haenszel X-squared = 2.3559, df = 1, p-value = 0.1248
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.8635901 3.3464951
sample estimates:
common odds ratio
             1.7

>
>
> x<-factor(rep(c(1,2,1,2),c(200,200,70,70)),
+ labels=c("Oint A","Oint B"))
```

```
> y<-factor(rep(c(1,2,1,2,1,2,1,2),
+ c(129,71,113,87,45,25,36,34)),
+ labels=c("Response","No Response"))
> z<-factor(rep(c(1,2),c(400,140)),
+ labels=c("Clinic 1" ,"Clinic 2"))
> table(x,y,z)
, , z = Clinic 1

         y
x         Response No Response
  Oint A      129          71
  Oint B      113          87

, , z = Clinic 2

         y
x         Response No Response
  Oint A       45          25
  Oint B       36          34

> mantelhaen.test(x,y,z,correct=F)

        Mantel-Haenszel chi-squared test without continuity
    correction

data:  x and y and z
Mantel-Haenszel X-squared = 4.7999, df = 1, p-value = 0.02846
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.041550 2.080194
sample estimates:
common odds ratio
         1.471946

>
```

30) (Artificial data from Ben Goldacre, 06/08/11).

Imagine a study was conducted to examine relationship between heavy drinking of alcohol and developing ling cancer, obtaing the following results:

	Cancer	*No cancer*
Drinker	*366*	*2300*
Non-Drinker	*98*	*1856*

a) Calculate the ration of the odds of developing cancer for drinkers to non-drinkers. What conclusions do you draw from this odds ratio?

The odds ratio is 3.01, suggesting that the odds for developing cancer are three times higher for drinkers than for non-drinkers. An approximate 95% confidence interval for the odds ratio is (2.38, 3.81)

b) It transpires that 330 of the drinkers developing cancer were smokers and 1100 of the drinkers who smoked did not, with corresponding figures for the non-drinkers of 47 and 156. Calculate the odds ratios separately for smokers and non-smokers. What conclusions do you draw?.

Both the odds ratios are 1.0, suggesting that the key difference in cancer rates is between smokers and non-smokers with no evidence of a difference between drinkers and non-drinkers. This effect is essentially the same as that observed in Simpson's paradox and illustrates the danger of post-hoc regrouping of tables. See the original article at

http://www.guardian.co.uk/commentisfree/2011/aug/05/bad-science-adjusting-figures

31) In a clinical trial of the use of a drug in twin pregnancies an obstetrician wishes to show a significant prolongation of pregnancy by use of the drug when compared to placebo. She assesses that the standard deviation of pregnancy length is 1.5 weeks, and considers a clinically significant increase in pregnancy length of 1 week to be appropriate.

i) How many pregnancies should be observed to detect such a difference in a test with a 5% significance level and with 80% power?

Require a two-sided two sample t-test. Formula gives 35.3 per group and **R,** Minitab and programme POWER give 37 in each group (S-PLUS gives 36) **so 74 (or 72) pregnancies in total need to be observed.**

```
> power.t.test(sd=1.5,delta=1,power=0.8)
     Two-sample t test power calculation
              n = 36.3058
          delta = 1
             sd = 1.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group
```

ii) It is thought that between 40 and 60 pregnancies will be observed to term during the course of the study. What range of increases in length of pregnancy will the study have a reasonable chance (i.e. between 70% and 90%) of detecting?

Note that "40 to 60 in total" means 20 to 30 in each group.

Results produced by programme POWER below:

```
Results
-------
Two Sample T test
 Table of CRD calculations
           Sample size group 1
         :       20 :       25 :       30 :
-----------------------------------------
      70 :  1.20670 :  1.07390 :  0.97708 :
      75 :  1.27967 :  1.13884 :  1.03617 :
      80 :  1.36103 :  1.21125 :  1.10205 :
      85 :  1.45595 :  1.29572 :  1.17890 :
      90 :  1.57545 :  1.40207 :  1.27566 :
-----------------------------------------
Rows are: power significance level = 0.05   standard deviation = 1.5
```

This will give an answer apparently accurate to about 6 seconds (since the working units are days and so they should be rounded to one (or at most two) decimal places.

In **R**, we have

```
> group<-seq(20,30,by=5)
> power<-seq(0.70,0.90,by=0.05)
> group
[1] 20 25 30
> power
[1] 0.70 0.75 0.80 0.85 0.90
> delta<-matrix(nrow=5,ncol=3)
>  for (i in 1:5) {
+   for (j in 1:3) (
+    delta[i,j]<-power.t.test(sd=1.5,power=power[i],
+    n=group[j])$delta
+   )
+  }
> options(digits=3)
> delta
     [,1] [,2]  [,3]
[1,] 1.21 1.08 0.978
[2,] 1.28 1.14 1.038
[3,] 1.36 1.21 1.103
[4,] 1.46 1.30 1.180
[5,] 1.58 1.40 1.277
```

There are some numerical differences in these but only of the order of about 10 minutes.

32) Given below is an edited extract from an SPSS session analysing the results of a two period crossover trial to investigate the effects of two treatments A (standard) and B (new) for cirrhosis of the liver. The figures represent the maximal rate of urea synthesis over a short period and high values are desirable. Patients were randomly allocated to two groups: the 8 subjects in group 1 received treatment A in period 1 and B in period 2. Group 2 (13 subjects) received the treatments in the opposite order.

i) Specify a suitable model for these data which incorporates treatment, period and carryover effects.

ii) Assess the evidence that there is a carryover effect from period 1 to period 2.

iii) Do the data provide evidence that there is a difference in average response between periods 1 and 2?

iv) Assess whether the treatments differ in effect, taking into account the results of your assessments of carryover and period effects.

v) Repeat the statistical analysis in R

Extract from SPSS Analysis of Crossover Trial on Liver Treatment

Summarize

Case Summaries(a)

Patnum	Group	Period1	Period2	Sum1+2	PeriodDiff	TreatDiff
1.00	1.00	48.00	51.00	99.00	-3.00	-3.00
2.00	1.00	43.00	47.00	90.00	-4.00	-4.00
3.00	1.00	60.00	66.00	126.00	-6.00	-6.00
4.00	1.00	35.00	40.00	75.00	-5.00	-5.00
5.00	1.00	36.00	39.00	75.00	-3.00	-3.00
6.00	1.00	43.00	46.00	89.00	-3.00	-3.00
7.00	1.00	46.00	52.00	98.00	-6.00	-6.00
8.00	1.00	54.00	42.00	96.00	12.00	12.00
9.00	2.00	31.00	34.00	65.00	-3.00	3.00
10.00	2.00	51.00	40.00	91.00	11.00	-11.00
11.00	2.00	31.00	34.00	65.00	-3.00	3.00
12.00	2.00	43.00	36.00	79.00	7.00	-7.00
13.00	2.00	47.00	38.00	85.00	9.00	-9.00
14.00	2.00	29.00	32.00	61.00	-3.00	3.00
15.00	2.00	35.00	44.00	79.00	-9.00	9.00
16.00	2.00	58.00	50.00	108.00	8.00	-8.00
17.00	2.00	60.00	60.00	120.00	.00	.00
18.00	2.00	82.00	63.00	145.00	19.00	-19.00
19.00	2.00	51.00	50.00	101.00	1.00	-1.00
20.00	2.00	49.00	42.00	91.00	7.00	-7.00
21.00	2.00	47.00	43.00	90.00	4.00	-4.00

T-Test

Independent Samples Test

	Mean Difference	Std. Error Difference	t	Df	Sig. (2-tailed)
Sum1+2	2.7308	8.7046	.314	18.683	.757
PeriodDiff	-5.9423	2.9429	-2.019	17.646	.059
TreatDiff	1.4423	2.9429	.490	17.646	.630

Summarize

Case Summaries(a)

				Summ1+2	PeriodDiff	TreatDiff
GROUP	1.00	1		99.00	-3.00	-3.00
		2		90.00	-4.00	-4.00
		3		126.00	-6.00	-6.00
		4		75.00	-5.00	-5.00
		5		75.00	-3.00	-3.00
		6		89.00	-3.00	-3.00
		7		98.00	-6.00	-6.00
		8		96.00	12.00	12.00
		Total	N	8	8	8
			Mean	93.5000	-2.2500	-2.2500
			Std. Deviation	16.1688	5.8979	5.8979
	2.00	1		65.00	-3.00	3.00
		2		91.00	11.00	-11.00
		3		65.00	-3.00	3.00
		4		79.00	7.00	-7.00
		5		85.00	9.00	-9.00
		6		61.00	-3.00	3.00
		7		79.00	-9.00	9.00
		8		108.00	8.00	-8.00
		9		120.00	.00	.00
		10		145.00	19.00	-19.00
		11		101.00	1.00	-1.00
		12		91.00	7.00	-7.00
		13		90.00	4.00	-4.00
		Total	N	13	13	13
			Mean	90.7692	3.6923	-3.6923
			Std. Deviation	23.6684	7.4876	7.4876
	Total	N		21	21	21
		Mean		91.8095	1.4286	-3.1429
		Std. Deviation		20.7235	7.3863	6.8065

A plot of mean responses (not shewn here, but always advisable) indicates that there looks to be a difference between the treatments (with B better) and little suggestion of period or carryover effects. This gives a useful guide to ensuring the t-tests are selected correctly.

i) Usual model from notes, including the identifiability constraints (i.e. sums = 0)

ii) No evidence of carryover (t = .314)

iii) Little evidence of difference in periods (t = 0.49, p = 0.63) (period 1 lower)

iv) Some evidence of treatment differences, t = –2.019, p = 0.059 (using both periods since no evidence of carryover (nor period) effect). mean response to B is higher than to A so some evidence that new treatment is better.

```
> attach(cirrhosis)
> cirrhosis[1:5,]
  Patnum Group Period1 Period2 Sum1.2 PeriodDiffs TreatDiffs
1      1     1      48      51     99          -3         -3
2      2     1      43      47     90          -4         -4
3      3     1      60      66    126          -6         -6
4      4     1      35      40     75          -5         -5
5      5     1      36      39     75          -3         -3
>
>
> t.test(Sum1.2 ~ Group)

        Welch Two Sample t-test

data:  Sum1.2 by Group
t = 0.3137, df = 18.683, p-value = 0.7572
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -15.50916  20.97070
sample estimates:
mean in group 1 mean in group 2
       93.50000        90.76923
```

```
> t.test(PeriodDiffs ~ Group)

        Welch Two Sample t-test

data:  PeriodDiffs by Group
t = -2.0192, df = 17.646, p-value = 0.05893
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -12.1340837   0.2494683
sample estimates:
mean in group 1 mean in group 2
      -2.250000        3.692308

> t.test(TreatDiffs ~ Group)

        Welch Two Sample t-test

data:  TreatDiffs by Group
t = 0.4901, df = 17.646, p-value = 0.6301
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -4.749468  7.634084
sample estimates:
mean in group 1 mean in group 2
      -2.250000       -3.692308

>
> t.test(PeriodDiffs)

        One Sample t-test

data:  PeriodDiffs
t = 0.8863, df = 20, p-value = 0.386
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.933623  4.790766
sample estimates:
mean of x
 1.428571
```

```
> t.test(TreatDiffs)

        One Sample t-test

data:  TreatDiffs
t = -2.116, df = 20, p-value = 0.04709
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -6.24114312 -0.04457117
sample estimates:
mean of x
-3.142857
```

33) Several studies have considered the relationship between elevated blood glucose levels and occurrence of heart problems. The results of two similar studies are summarized below.

	Study 1			***Study 2***		
	heart problems			*heart problems*		
glucose level	*yes*	*no*		*yes*	*no*	
elevated	*61*	*1284*	*1345*	*32*	*996*	*1028*
not elevated	*82*	*1930*	*2012*	*25*	*633*	*658*
	143	*3214*	*3357*	*57*	*1629*	*1686*

i) What can be concluded from these data regarding the influence of glucose on heart problems?

ii) Do you have any doubts on the validity of the form of analysis you have used?

Mantel-Haenszel tests:

Study 1: $E[Y_1]=1345\times143/3357=57.29$

$\text{var}(Y_1)=1345\times2012\times143\times3214/(3357^2\times3356)=32.89$

so $\chi^2_{MH}=0.417$, $p>>0.05$.

Study 2: $E[Y_2]=34.75$, $\text{var}(Y_2)=13.11$, $\chi^2_{MH}=0.579$, $p>>0.05$.

Combined gives $\chi^2_{MH}=0.02$.

Conclude that there is no evidence of influence of glucose on heart problems. Response rates in the two studies are 4.5% and 3.1%, not very different in absolute terms so few doubts as to validity of analysis,

and in any case the results are so far away from significance. Note that the Pearson χ^2 values are nearly identical to the Mantel-Haenszel ones.

Just for illustration, but beyond the scope of this question, here is an analysis using logistic regression: First set up the data as

```
> frequency<-c(61,82,1284,1930,32,25,996,633)
> problems<-c(rep(c(1,1,0,0),2))
> glucose<-c(rep(c(1,0),4))
> study<-c(rep(0,4),rep(1,4))
>
> heart.glm<-
glm(problems~glucose+study,weights=frequency,family=binomial)
>
> summary(heart.glm)

Call:
glm(formula = problems ~ glucose + study, family = binomial,
    weights = frequency)

Deviance Residuals:
      1         2         3         4         5         6         7
8
 19.585    22.779   -10.637   -12.910    14.706    13.037    -8.310
-6.558

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.12076    0.10426 -29.933   <2e-16 ***
glucose      0.02069    0.14737   0.140    0.888
study       -0.24457    0.16251  -1.505    0.132
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1682.9  on 7  degrees of freedom
Residual deviance: 1680.6  on 5  degrees of freedom
AIC: 1686.6

Number of Fisher Scoring iterations: 6

>
```

34) A randomized, parallel group, placebo controlled trial was undertaken to assess the effect on children of a cream in reducing the pain associated with venepuncture at the induction of anaesthesia. A binary response of Y=0 for 'did not hurt' and Y=1 for 'hurt' was recorded for each of the 40 children who entered the trial, together with the treatment given (x1) and two covariates, sex (x2) and age (x3), which were thought might affect pain levels. A logistic model was fitted and the following details are available.

Factor	**Reg. Coeff.**	**Standard Error of Coefficient**
Intercept	2.058	1.917
x1: treatment (0 = placebo, 1 = cream)	-1.543	0.665
x2: sex (0 = boy, 1 = girl)	0.609	0.872
x3: age (years)	-0.461	0.214

i) Interpret and assess the treatment effect and also the effects of sex and age.

ii) Estimate the relative risk of hurting with the cream compared to the placebo.

Factor	Coefficient	coefficient/s.e.	p-value
treatment	–1.543	–2.32	.0204
sex	0.609	0.698	.485
age	–0.461	–2.15	.032

Good evidence that treatment reduces the relative risk of hurting. Also good evidence that this risk decreases with age. No evidence of differences between sexes.

Estimate of relative risk using cream is $e^{-1.543} = 0.2137$ or 21.4%, with an approximate 95% CI of (5.7%, 80.8%). So the reduction in risk when using the cream is estimated as 79%, with 95% CI of (19%, 94%).