

Clinical Trials: Task Sheet 4

Notes & Solutions

(Questions 1–12 are additional exercises on sample size calculations)

(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

- 1) Look at the solutions to Task sheet 3 and repeat the analyses given there (if you have not already done so).

Trust you have done this by now

- 2) 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.



- 3) 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 30 subjects.

- 4) Do you have an explanation for why the total numbers in Q4 and Q5 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.

- 5) 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.

- 6) *How many subjects are need 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.

- 7) *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



8) 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 Q5

9) 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.

10) 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.

11) Questions 5, 6, 7, 8, 9 and 10 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 $(r/n(1-r/n)/n)^{1/2}$ which is a maximum when $r/n=1/2$, i.e. proportions closer to 0.5 require a greater sample size for a specified precision than those further from 0.5.

12) 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 Q9).



13) 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 200 μ g salbutamol (a well established bronchodilator) and 12 μ g 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.

Model: usual one in notes. It is a good idea to plot the means for each group for each period (not shown 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: $t=0.17$ [$=(643.6-629.2)/(114.3^2/7+174^2/6)^{-1/2}$] $p \gg 0.05$, so can proceed with treatment & period tests:

Treatment: $t=4.22$ [$=(30.7-(-62.6))/(33.0^2/7+44.7^2/6)^{-1/2}$] 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.



14) 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.

<i>patient</i>	<i>period 1</i>	<i>period 2</i>
1	A	9
2	B	11
3	B	7
4	B	12
5	A	8
6	A	11
7	A	4
8	B	3
9	A	13
10	B	7
11	A	1
12	A	13
13	A	6
14	B	5
15	B	6
16	B	3

- Calculate the mean for each treatment in each period and display the results graphically.
- Assess the carryover effect.
- If appropriate, assess the treatment and period effects.

(NB These data are available in R, Minitab and S-PLUS forms on the course web pages)

Given below is a transcript of R performing all the required calculations using the command `t.test(.)`.



The relevant values and key steps needed to answer the questions above have been highlighted in the transcript below. Note the *slick trick* used to change the signs of the group 2 differences. This is not something you actually need to be able to do yourself, just recognise it later.

```
> hourssleep
  PERIOD1 PERIOD2 GROUP  sum diff
1         9         0     1  4.5   9
2        11        14     2 12.5  -3
3         7         3     2  5.0   4
4        12         8     2 10.0   4
5         8         8     1  8.0   0
6        11         1     1  6.0  10
7         4         4     1  4.0   0
8         3         4     2  3.5  -1
9        13         2     1  7.5  11
10        7         3     2  5.0   4
11         1         2     1  1.5  -1
12        13         1     1  7.0  12
13         6         3     1  4.5   3
14         5         6     2  5.5  -1
15         6         8     2  7.0  -2
16         3         7     2  5.0  -4
> attach(hourssleep)
> t.test(sum[GROUP==1],sum[GROUP==2])

Welch Two Sample t-test

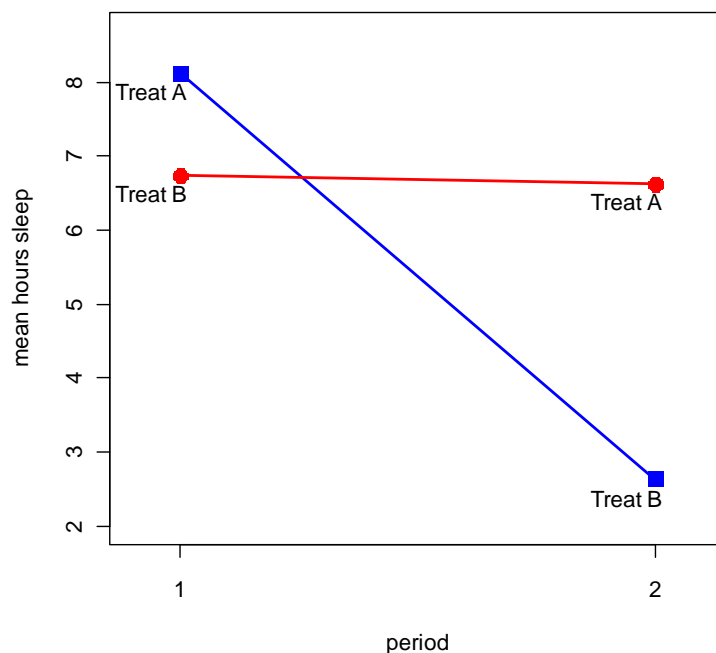
data:  sum[GROUP == 1] and sum[GROUP == 2]
t = -0.9929, df = 12.64, p-value = 0.3394
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -4.176408  1.551408
sample estimates:
mean of x mean of y
 5.3750   6.6875
```



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))
```

Plot of mean responses against periods



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 an 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.

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



Clinical Trials: Task Sheet 5

Notes & Solutions

- 1) 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:

Clinic	Ointment A		Ointment B	
	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 §8.3.

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

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

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

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
```

x	y	
	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
```

x	y	
	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
```

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

```
, , z = Clinic 2
```

x	y	
	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
```

```
>
```

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

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

	Cancer	No cancer
Drinker	366	2300
Non-Drinker	98	1856

a) Calculate the ratio 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>

