1

Introduction to analysis of binary and proportion data

Binary data occur frequently in infection management (IM) and quality improvement (QI) work. Two examples are surgical site infections (SSIs) and patient mortality: an SSI occurs or doesn't occur, and similarly a patient does or does not survive. The number of such outcomes in a sample of patients, or over a period of time, is often described using a binomial distribution. The dominant assumptions underlying this distribution are that each outcome is independent and has the same probability of occurrence. Note that these assumptions may not always hold. For example, we describe data in Chapter 7 that appear to be binomial but tend to display too much or too little variability due to measurement error, repeated measurements on the same individual, clustering and so on. This has an impact on conclusions about the true level of infection rates and whether these have changed unexpectedly during a monitoring period or between groups. However, the binomial distribution is applicable, at least approximately, to many hospital IM and QI outcomes. In this chapter, we focus on the following problems that are commonly encountered in the analysis of grouped binary data.

```
1. Estimation and testing of a single proportion,
2. Estimation for a series of proportions,
3. Comparison of two proportions,
4. Evaluation of more than two proportions.
5. The analysis of stratified data.
```

1.1 Single proportion, samples and population

Consider the scenario in which 14 SSIs were observed in 74 consecutive major lower limb vascular operations. Based on these data, what can we say about the true population proportion

Statistical Methods for Hospital Monitoring with R, First Edition. Anthony Morton, Kerrie Mengersen, Michael Whitby and George Playford.

of infection (p) in the population of patients undergoing such operations, commonly called the infection rate? We consider three common approaches to estimating this value.

The first is a moment-based approach, in which the population infection rate is estimated by the sample proportion. The sample proportion, usually labeled $\hat{p}$, is calculated as the number of complications (X) divided by the number of operations (N), so that $\hat{p}$= X/N. In the above scenario, $\hat{p}$=14/74=0.19, so the sample percentage is 19%. (Note that in many texts, N is used to denote the population size and *n* is used to denote the sample size. Here we use N to indicate the sample size. Note also that later in these notes we may drop the notation where the context is obvious, or use other symbols where it is relevant, e.g. R=X/N.)

The second is a maximum likelihood approach. The likelihood is the chance of obtaining the observed data (e.g., 14 SSIs in 74 operations) given a certain value of p. The estimate of p is then taken to be the value that maximises this likelihood. For example, suppose that only two population rates were under consideration in the above example: p=0 or p=0.20. It is much more likely to observe 14 outcomes in 74 operations if the underlying population infection rate is 0.20 compared to the chance of observing these outcomes if the rate is 0, so the value of 0.20 would be chosen as the maximum likelihood estimate (MLE).

The third is a Bayesian approach, which estimates a (posterior) distribution for p based on the data (via the likelihood) and a prior distribution for p. This prior can be noninformative if nothing is known about p apart from the information provided in the data, or it can be based on information available about the population rate obtained from sources other than the data.

Regardless of the approach, the estimated value of p is just that: an estimate of the population rate based on a sample. The value of $\hat{p}$ is obviously dependent on the sample taken from the population. In particular, when samples are small, as they often are in QI and IM studies, there is a great deal of variation in the observed rates and corresponding uncertainty about the true population rate: for example, if p=0.5 then in a sample of four patients it would be equally probable to observe 1 out of 4 outcomes (moment-based estimated proportion $\hat{p}$=0.25) and 3 out of 4 outcomes ($\hat{p}$=0.75). Moreover, there is a question of representativeness in small samples: for example, if there were four patients and two had complications, it would be difficult to believe that these four patients with their observed 50% complication rate were representative of all such patients. If the next two patients studied did not have complications, the observed rate would fall from 50% to 30%.

The most common method of describing uncertainty is through some form of confidence interval (CI). A wider CI indicates greater uncertainty in the estimation of p. Under a moment-based approach, the CI is usually based on the variance of the binomial distribution, given by $\hat{p}(1\text{-}\hat{p})$/N. The width of the CI is thus determined by the magnitude of $\hat{p}$ and the sample size (N). Under a likelihood-based approach, the CI (usually called a supported range) tells us the range of possible values for p that are supported by the data ($\hat{p}$). Under a Bayesian approach, a range of values (a 95% credible interval) is found from the posterior distribution of p. These approaches are described in more detail in the following section. Under the frequentist idea of repeated sampling from a population, a 95% CI implies that on average, for every 100 samples drawn randomly from a population, 95 of the corresponding CIs computed from these samples will contain the true value of p (i.e. the true underlying complication rate). Under the Bayesian approach, a 95% credible interval is interpreted more directly, as a range of values that encompass the true value of p with 95% probability, given the observed data.

Another common objective is to use the observed data to test an hypothesis. For example, as we have described, we may know what SSI rate to expect based on published literature or from local data obtained when the vascular surgery SSI rate was considered satisfactory.

We wish then to determine whether the observed infection rate from a sample of patients is consistent with this value of p. If the data support a higher rate, we may need to examine our system of wound care in vascular surgery. The hypothesis test is usually evaluated using a form of P-value under a frequentist approach, likelihood ratio (LR) under a likelihood approach, or Bayes Factor (BF) under a Bayesian approach. As Goodman describes, these measures are inter-related. For example, a likelihood ratio of 1/7 coincides approximately with a 95% supported range that is numerically similar to a 95% CI. Although P-values that we discuss below and LRs are not strictly comparable, Goodman points out that, when testing certain hypotheses, a P-value of 0.05 (1/20) that corresponds to a 95% CI matches an LR of 1/7. The LR suggests that there is seven times as much support for the observed value $\hat{p}$ as there is for the population value p. If we are testing the hypothesis that the population rate is equal to a certain value (e.g., p=0) compared with an alternative (e.g., p=0.20), and we find that the P-value is small, the LR is large or the range of values supported by the data does not include p, it is reasonable to reject the hypothesis in favour of the alternative.

An important point to note at this stage is that although we use a 95% level of confidence in the above discussions, other confidence levels should also be considered, such as 68%, 95.5% or 99.7% (approximately equivalent to 1, 2 and 3 standard deviations (SDs) for normally distributed data). These may sometimes be more useful in surveillance and monitoring for detecting poor outcomes which can then be followed up through other means such as Morbidity and Mortality (M&M) meetings, as discussed in Chapter 8.

It is also important to understand the difference between CIs (often called precision limits) and control limits in control charts (often called prediction limits). The former surround an observed value such as $\hat{p}$ whereas the latter surround a mean value such as p and are thus analogous to significance tests that we describe for example in section 1.2 of this chapter. Often, if $\hat{p}$ does not differ greatly from p, precision and prediction limits do not differ greatly and are used interchangeably either because of availability or ease of calculation. Thus confidence limits are occasionally used to obtain approximate control chart limits. In addition, as Altman and Bland describe, confidence limits may, with symmetrical data, be employed to obtain approximate P-values (see below) and vice versa. Control charts are described in Chapters 3, 6 and 7.

1.1.1 Calculating the confidence interval

How do we calculate a CI for the unknown population rate p, based on binomial data? Under a frequentist approach, the most common approach is to assume that $\hat{p}$ is normally distributed and to calculate a CI as $\hat{p} \pm Z \times \sqrt{(\hat{p} \times (1\text{-}\hat{p})/N)}$, where N is the sample size and Z is the standard normal value corresponding to the confidence level (e.g. Z=1.96 for a 95% CI). For the example of 14 SSIs in 74 operations, the 95% CI obtained by this approach is 0.1 to 0.28). This approximation is suitable for large values of N and moderate values of p (i.e. not very small or very large). In many situations in IM and QI work, both p and N are small, so this common approximation is not sufficiently accurate.

An alternative method that is applicable to a much wider range of p and N values, and that avoids the normality approximation, is to describe p by a beta distribution; a common choice is p∼Beta(X+1,N-X+1), where ∼ means 'is distributed as'. We can then take the middle 95% of this probability distribution as the range for a 95% CI. This is consistent with both likelihood and Bayesian approaches. Moreover, the ready availability of beta distribution functions in computer software makes the use of the normal approximation unnecessary even

when samples are large. This calculation is easily performed in R. The required commands for a 95% CI are qbeta(0.025,X,N-X+1) and qbeta(0.975,X+1,N-X). Employing these formulas for the above scenario, qbeta(0.025,14,61)=0.107 or 10.7% and qbeta(0.975,15,60)=0.297 or 29.7%.

Another approach is to calculate so-called exact intervals. For the running example, results identical to those obtained above are found with binom.exact() in the R exactci library. However, these can be conservative for proportion and count data due to the discrete nature of the binomial distribution. (This also applies to hypergeometric and Poisson distributions.) Wilson score or exact mid-P intervals have been advocated to deal with this difficulty (Armitage and colleagues, Campbell and Swinscow, Altman and colleagues, Kirkwood and Sterne, Clayton and Hills). The Wilson formula is $(N/(N+Z^2))\times(X/N+Z^2/(2\times N)\pm Z\times((X\times(N-X)/N^3)+Z^2/(4\times N^2))^{.5})$, where for a 95% CI Z=1.96. It is available as scoreci() in the R PropCIs library. The mid-P method is slightly more complicated and is available as midPci() in the R PropCIs library.

```
# comparison of CI methods
# using the beta distribution
qbeta(0.025,14,61)  # [1] 0.107
qbeta(0.975,15,60)  # [1] 0.297
# exact intervals
library(exactci)
binom.exact(14,74)$conf.int
# [1] 0.107, 0.297
library(PropCIs)
# Wilson score intervals
scoreci(14,74,conflev=.95)
#  lowlim  uplim
#  0.116, 0.293
# mid-P exact intervals
midPci(14,74,alpha=.05)
#  lowlim  uplim
#  0.112, 0.290
```

For these data the limits based on the beta distribution are similar to the exact limits while the Wilson score and mid-P limits are similar and narrower than the exact and beta distribution limits. Following the recommendation of Armitage and colleagues, we have generally quoted both mid-P and exact intervals but give more emphasis to the former, particularly in functions that employ Newcombe's CI method. For example mid-P intervals are employed in section 1.4.1 of this chapter that deals with the CI for the difference between independent proportions.

1.1.2 Comparison with an expected rate

As discussed above, a frequent requirement is to compare an observed rate ($\hat{p}$) with a reference rate (p). For example, one hospital's SSI rate for a certain procedure may be compared with a rate obtained for a group of hospitals that is regarded as the achievable background rate for that procedure. Clearly a difference between $\hat{p}$ and p may be due to predictable (mostly random) variation in the single hospital's rate due to the particular sample taken and its size. We thus wish to determine whether the difference between $\hat{p}$ and p is so large that predictable variation is unlikely to be able to explain it. Note that the population rate, described above, may not necessarily be a reference rate. For example, the same approach may be used to

evaluate whether there has been a change in the previously stable process rate (p) within the hospital (e.g. in a control chart as described in Chapter 3). In all these cases, we can again use the beta distribution to assess the expected rate p.

To illustrate the method, suppose a rate of 10% was expected for the SSIs. We wish to determine whether the observed 19% rate (X=14, N=74) is so large that predictable variation is unlikely to have produced it. This is easily calculated by using pbeta(), the beta distribution function in R. If the observed rate exceeds the expected rate, as it does with the SSI data (19% observed, 10% expected), pbeta(p,X,N-X+1) is employed in R, where p is the expected 10% rate and pbeta(0.10,14,61) equals 0.014. This is analogous to a one tailed P-value (so it needs to be multiplied by 2 to be equivalent to the widely used conventional two-sided P-value of 0.028). When the observed proportion is less than expected, 1-pbeta(p,X+1,N-X) is used and the result must also be multiplied by 2 for a two tailed result (the differing positions of 1 in these formulas occur because of the discrete nature of the binomial distribution). The binom.exact() function mentioned above gives identical results. Approximate mid-P P-values can be calculated by getting the P-value for the next more extreme case, for example, X=15 in this case, and obtaining the average of that and the exact P-value for the observed X=14 (Vollset). These calculations are in the R function proportion() in rprogs that may also be accessed through IMenu().

```
# Proportion CI and P-value
library(exactci)
library(PropCIs)
#Enter numerator
x<-14
# Enter denominator
n<-74
p<-0 #if no reference proportion
# Enter reference proportion if available
p<-0.1
if (p!=0){jj<-"y"}else{jj<-"n"}
lo<-binom.exact(x,n)$conf.int[1]
up<-binom.exact(x,n)$conf.int[2]
midplo<-midPci(x,n,alpha=0.05)[1]
midpup<-midPci(x,n,alpha=0.05)[2]
cat("Proportion = ",round(x/n,3),", lower 95% limit =
",round(lo,3),", upper limit = ",round(up,3),"\nMid-p lower
95% limit = ",round(midplo,3),", upper limit =
",round(midpup,3),".\n",sep="")
#Proportion = 0.189, lower 95% limit = 0.107, upper limit = 0.297
#Mid-p lower 95% limit = 0.112, upper limit = 0.29.
if (jj=="y")
{
q<-binom.exact(x,n,p)$p.value
z<-qnorm(q/2)
lr0<-paste("1/",as.character(round(1/exp(-z^2/2),0)),sep="")
q1<-0
if (x!=0){
if (x/n>p){q1<-binom.exact(x+1,n,p)$p.value}else{q1<-
binom.exact(x-1,n,p)$p.value}
q1<-(q+q1)/2
z1<-qnorm(q1/2)
lr1<-paste("1/",as.character(round(1/exp(-z1^2/2),0)),sep="")
```

```
}
if (p!=x/n){
cat("P = ",round(q,3),", LR = ",lr0,sep="")
cat(".\n",sep="")
if (x!=0){
cat("Mid-P P = ",round(q1,3),", LR = ",lr1,sep="")
cat(".\n",sep="")
}
}
}
#P = 0.028, LR = 1/11.
#Mid-P P = 0.02, LR = 1/15.

#proportion function (also via IMenu(), select 1 and 1)
proportion()
```

```
Enter numerator 14
Enter denominator 74
Is a reference proportion available? y
Enter reference proportion .1
Proportion = 0.189, lower 95% limit = 0.107, upper limit = 0.297
Mid-P lower 95% limit = 0.112, upper limit = 0.29.
P = 0.028, LR = 1/11.
Mid-P P = 0.02, LR = 1/15.
```

1.2 Likelihood ratio (Bayes factor) & supported range

We have referred in the Introduction to the usefulness of the likelihood ratio (LR) concept. Goodman describes how an approximate LR can be obtained by converting the P-value (P) to a standard normal deviate. This can be accomplished in R by using Z<-abs(qnorm(P/2)) and then employing cat(paste("1/",as.character(round(1/exp(-Z^2/2),0)),sep=""),"\n"). For a two-sided P-value of 0.028, the LR is 1/11. (Although they are not strictly comparable, Goodman notes the difference implied by a P-value=0.028 or about 1/36 and a LR of 1/11.) This LR suggests that there is 11 times the support for the observed rate of 19% as there is for the expected rate of 10%, making predictable variation an unlikely explanation for the observed 19% rate. In addition, the supported range (CI) for R=14/74=19% suggests that the possible values of p supported by the data at the 95% level range from 10.7% (mid-P 11.2%) that are above the expected value of 10%. However, although predictable variation would be an unlikely cause for the observed difference, thus indicating for example, that an M&M audit should be performed, we cannot be absolutely certain since the LR indicates that there is some support for a 10% rate.

It is important to note that the above tests indicate whether a difference is unlikely to have arisen by chance; it does not tell whether the difference is of practical importance. In small samples, it may be difficult to be sure that large differences that may be of practical importance have not occurred by chance and in very large samples, differences of no practical importance may be conventionally statistically significant. It is useful to employ a combination of practical experience and the CI, P-value and LR to make informed decisions. In addition, the consequences of making a wrong decision are important: are they trivial or potentially serious? A practically important difference in a small sample that fails to reach conventional statistical significance requires further attention.

The Camp-Poulson approximation (Gebhardt), an accurate method for approximating the F-distribution that is related to the beta distribution, is a useful alternative for performing the significance test. If X/N>p, set X=X-1. Then F=(N-X)×p/((X+1)×(1-p)). A=$F^{(1/3)}$×(9-1/(N-X))+1/(X+1)-9. B=3×($F^{(2/3)}$×(1/(N-X))+1/(X+1))$^{.5}$, and Z=-A/B. It is also useful, with some adjustment, for a significance test for overdispersed count data analysed using the negative binomial distribution (section 4.10 of Chapter 4).

```
# Camp-Poulson approximation
X<-14;N<-74
X<-X-1
p<-.1
FF<-(N-X)*p/((X+1)*(1-p))
A<-FF^(1/3)*(9-1/(N-X))+1/(X+1)-9
B<-3*(FF^(2/3)*(1/(N-X))+1/(X+1))^.5
Z<--A/B
Z
#[1] 2.188263
#P-value
P<-2*(1-pnorm(Z))
P
#[1] 0.02865046
#
#mid-P
X<-15;N<-74
X<-X-1
p<-.1
FF<-(N-X)*p/((X+1)*(1-p))
A<-FF^(1/3)*(9-1/(N-X))+1/(X+1)-9
B<-3*(FF^(2/3)*(1/(N-X))+1/(X+1))^.5
Z1<--A/B
Z1
#[1] 2.50614
#P-value
P1<-2*(1-pnorm(Z1))
P1
#[1] 0.01220573
#
(P+P1)/2 #mid-p value
#[1]0.02042809
```

1.3 Confidence intervals for a series of proportions

Calculating confidence intervals for a series of proportions for a report is often a job for hospital scientists, for example, SSI rates for different procedures with differing expected rates. This can be time-consuming and prone to error. The file hosprops.csv has columns for procedure codes, procedure dates, risk-indexes and SSI codes for a group of hospitals. Our interest is in the in-hospital (Infection Type 1) SSI rates for each of the procedures.

```
# getting the hosprops.csv data with function g.d()
g.d()
```

```
Loading data.
Data from clipboard (C) or file (F) c
Do data column(s) have heading(s) (Y/N) y
Is column 1 a date column (Y/N) n
```

Tabulations are performed using tapply() and confidence interval calculations using binom.exact(). The tabulations can also be performed using the xtabs() function. Calculation of the overall rate and its confidence interval need to be performed separately.

The confidence interval calculations can be performed using the function multipleproportions(). The function multipleproportions() may also be accessed via IMenu().

```
# tabulate by procedure code using tapply()
Proc<-datain[,1]
SSI<-rep(0,length(datain[,1]))
SSI[datain[,4]==1]<-1
SSIs<-tapply(SSI,Proc,sum)
Procedures<-tapply(SSI,Proc,length)
ProcCode<-names(SSIs)
Table<-data.frame(ProcCode,SSIs,Procedures)
row.names(Table)<-1:length(SSIs)
Table
```

	ProcCode	SSIs	Procedures
1	1	20	508
2	3	78	4897
3	4	19	521
4	5	69	6269
5	6	8	348
6	7	44	702
7	8	65	12724
8	9	153	16654
9	10	90	8129
10	11	159	9479
11	12	6	568
12	13	8	673
13	14	3	589
14	15	52	3474

```
#calculate 95% & 99.7% confidence intervals
#mid-P values not calculated
library(exactci)
a<-Table[,2]
b<-Table[,3]
x<-length(a)
s<-0;l<-0;u<-0
ll<-0;uu<-0
for (i in 1:x){s[i]<-binom.exact(a[i],b[i])$estimate
l[i]<-binom.exact(a[i],b[i])$conf.int[1]
u[i]<-binom.exact(a[i],b[i])$conf.int[2]
ll[i]<-binom.exact(a[i],b[i],conf.level=.997)$conf.int[1]
uu[i]<-binom.exact(a[i],b[i],conf.level=.997)$conf.int[2]
}
```

```
Proc<-Table[,1]
ss<-data.frame(Proc,s,l,ll,u,uu)
# adding average rate and CIs
a0<-sum(Table[,2])
b0<-sum(Table[,3])
Mean<-a0/b0
l0<-binom.exact(a0,b0)$conf.int[1]
u0<-binom.exact(a0,b0)$conf.int[2]
ll0<-binom.exact(a0,b0,conf.level=.997)$conf.int[1]
uu0<-binom.exact(a0,b0,conf.level=.997)$conf.int[2]
ssall<-data.frame("All",Mean,l0,ll0,u0,uu0)
names(ssall)<-c("Proc","s","l","ll","u","uu")
ss<-rbind(ss,ssall)
#print(ss), see multipleproportions() below

#using the multipleproportions() function
multipleproportions(Table)
```

```
      Proc          s          l          ll          u        uu
1        1   0.039370   0.024211   0.0184405   0.060151   0.07203
2        3   0.015928   0.012610   0.0111286   0.019840   0.02199
3        4   0.036468   0.022096   0.0166802   0.056366   0.06778
4        5   0.011007   0.008574   0.0074982   0.013909   0.01551
5        6   0.022989   0.009976   0.0060866   0.044792   0.05792
6        7   0.062678   0.045908   0.0387491   0.083232   0.09457
7        8   0.005108   0.003945   0.0034334   0.006507   0.00728
8        9   0.009187   0.007794   0.0071453   0.010755   0.01160
9       10   0.011071   0.008912   0.0079378   0.013591   0.01497
10      11   0.016774   0.014285   0.0131226   0.019566   0.02107
11      12   0.010563   0.003886   0.0021197   0.022849   0.03045
12      13   0.011887   0.005146   0.0031365   0.023287   0.03021
13      14   0.005093   0.001052   0.0003735   0.014812   0.02113
14      15   0.014968   0.011199   0.0095739   0.019583   0.02215
15     All   0.011810   0.010997   0.0105966   0.012667   0.01312
[1]   "The results are is ss"
```

```
#l & u are equivalent to a 95% & ll & uu a 99.7% interval
#approximately equivalent to two and three standard
#deviations respectively
#producing a more attractive output
ss<-cbind(ss[,1],round(ss[,c(2:6)],4))
names(ss)<-c("Proc","SSI","L95","L99.7","U95","U99.7")
ss
```

```
   Proc    SSI    L95  L99.7    U95  U99.7
1     1 0.0394 0.0242 0.0184 0.0602 0.0720
2     3 0.0159 0.0126 0.0111 0.0198 0.0220
3     4 0.0365 0.0221 0.0167 0.0564 0.0678
4     5 0.0110 0.0086 0.0075 0.0139 0.0155
5     6 0.0230 0.0100 0.0061 0.0448 0.0579
6     7 0.0627 0.0459 0.0387 0.0832 0.0946
7     8 0.0051 0.0039 0.0034 0.0065 0.0073
8     9 0.0092 0.0078 0.0071 0.0108 0.0116
9    10 0.0111 0.0089 0.0079 0.0136 0.0150
```

```
10  11 0.0168 0.0143 0.0131 0.0196 0.0211
11  12 0.0106 0.0039 0.0021 0.0228 0.0304
12  13 0.0119 0.0051 0.0031 0.0233 0.0302
13  14 0.0051 0.0011 0.0004 0.0148 0.0211
14  15 0.0150 0.0112 0.0096 0.0196 0.0221
15 All 0.0118 0.0110 0.0106 0.0127 0.0131
```

```
s.d(ss) # if required, export to office program using the clipboard

# also use IMenu()
IMenu(Table)
```

```
Introductory Menu

1. Proportion data,
2. Count and rate data.
1
1. Single proportion,
2. Confidence intervals for a series of proportions,
3. Two proportions,
4. More than two proportions,
5. Weighted average of proportions.
2
```

These data refer to different procedures with differing expected SSI rates so the need for comparisons does not arise. The use of multiple confidence intervals for the same procedure within differing institutions, their comparison with a reference value and corrections for multiple testing are discussed when we deal with aggregated data obtained from several institutions in Chapter 2.

1.4 Difference between two proportions

In IM and QI studies one occasionally compares two independent proportions. The odds ratio, as described by Kirkwood and Sterne, is often used for this purpose. However, NNT, the number needed to treat to prevent one complication, also described by Kirkwood and Sterne, is useful in IM and QI work; this is the reciprocal of the risk difference. Thus, if a complication rate is 20% and a new process is able to reduce it to 10%, the difference would be 10%. By using the new method for the next 10 patients, one complication could potentially be averted; NNT is therefore 10 (i.e. the reciprocal of the risk difference is 1/0.1=10). Calculation of the CI for NNT involves the reciprocals of the confidence limits for the difference between proportions so the risk difference is often a better estimator for IM and QI work than the odds ratio. However, the odds ratio is mathematically a more natural estimator and we employ it, for example, via logistic regression. We also describe calculations for the ratio of two independent proportions (risk ratio).

1.4.1 Confidence intervals

Newcombe's method for the CI for the difference between two independent proportions is recommended by Altman and colleagues. First, one must obtain the confidence limits for

each proportion as described above. Let P_1 and P_2 be the two sample proportions, L_1 and L_2 their lower and U_1 and U_2 their upper confidence limits. Let D be the difference between the two proportions and U and L the required upper and lower difference confidence limits. Then the formulas are $U=D+\sqrt{[(P_2-L_2)^2+(U_1-P_1)^2]}$ and $L=D-\sqrt{[(P_1-L_1)^2+(U_2-P_2)^2]}$.

Nam describes an accurate score method for the calculation of the confidence interval for the risk ratio. It involves the solution of a complicated cubic equation. A difficulty occurs if there is a zero numerator; when this happens 0.5 can be added to each of the numerators and 1 to each of the denominators. The function riskscoreci() in the PropCIs library implements this method.

```
# Ratio (Nam)
z<-1.96
x1<-14
n1<-74
x2<-8
n2<-114
if (x2/n2>x1/n1){v<-x1;x1<-x2;x2<-v;v<-n1;n1<-n2;n2<-v}
if (x1/n1!=x2/n2)
{
if (x1==0 | x2==0){x1<-x1+0.5;x2<-x2+0.5;n1<-n1+1;n2<-n2+1}
xd<-x2+x1
nd<-n2+n1
a1<-n2*(n2*nd*x1+n1*(n2+x1)*z^2)
a2<--n2*(n2*n1*xd+2*nd*x2*x1+n1*(n2+x2+2*x1)*z^2)
a3<-2*n2*n1*x2*xd+nd*x2^2*x1+n2*n1*xd*z^2
a4<--n1*x2^2*xd
b1<-a2/a1
b2<-a3/a1
b3<-a4/a1
c1<-b2-b1^2/3
c2<-b3-b1*b2/3+2*b1^3/27
th<-acos(27^0.5*c2/(2*c1*(-c1)^0.5))
P<-pi
ca<-cos(P/3-th/3)
t1<--2*(-c1/3)^0.5*ca
cb<-cos(P/3+th/3)
t2<--2*(-c1/3)^0.5*cb
cc<-cos(th/3)
t3<-2*(-c1/3)^0.5*cc
t1<-t1-b1/3
t2<-t2-b1/3
t3<-t3-b1/3
ta<-t1;tb<-t2
if (t1>t2 & t1>t3){ta<-t2;tb<-t3}
if (t2>t1 & t2>t3){ta<-t1;tb<-t3}
upr<-(1-(n1-x1)*(1-ta)/(x2+n1-nd*ta))/ta
lor<-(1-(n1-x1)*(1-tb)/(x2+n1-nd*tb))/tb
if (lor>upr){y<-upr;upr<-lor;lor<-y}
if (x2>0){cat("Ratio ",round(x1*n2/(n1*x2),2),".\n",sep="")}
cat("Ratio 95% confidence limits are ",round(lor,2)," and
",round(upr,2),".\n",sep="")
}
#Ratio 2.7.
#Ratio 95% confidence limits are 1.22 and 6.
```

For the risk difference, Newcombe recommends that using Wilson score intervals for the two proportions gives better results than exact intervals as the latter for proportion and count data can be conservative due to the discrete nature of the binomial, hypergeometric and Poisson distributions. Wilson score or exact mid-P intervals have been advocated to deal with this difficulty (Armitage and colleagues, Campbell and Swinscow, Altman and colleagues). We have employed mid-P intervals using the R function midPci() in the PropCIs library.

Consider the vascular surgical unit SSIs. In 74 consecutive Class 1 operations there were 14 SSIs, a rate of 19%. Since this rate was considered unacceptably high, the processes of wound care were carefully revised and in the following 114 operations there were eight SSIs, giving a rate of 7%.

1.4.2 Hypothesis test

If there is indeed no real difference between the population rates in the two groups, what is the probability of observing the sample data? Although frequently used with case-control data, the Fisher Exact test may be employed to answer this question. The data are arranged in a 2×2 table. The test involves computing the exact probability of the observed data assuming no difference between the two groups and adding to this the probability of all possible less likely data arrangements, conditional on the row and column totals of the table remaining fixed. To illustrate the calculation for the observed data, let the four cell counts of the 2×2 table be a=14, b=8, c=74-14=60 and d=114-8=106. N=74+114=188 is the total and the marginal totals of the 2×2 table are m1=a+c=14+60=74, m2=b+d=8+106=114, m3=a+b=14+8=22 and m4=c+d=60+106=166.

The probability of the observed result if there is no difference between the groups is then P=(N!×a!×b!×c!×d!)/(m1!×m2!×m3!×m4!)=0.0094, where ! stands for factorial (because factorials rapidly become large numbers, logs are used and added and for large counts Stirling's approximation log(N!)≈log(N)*N-N+log($\sqrt{}$(2×π×N)) is employed). The exact two-sided P-value is obtained by adding to this the values for all less probable cell counts conditional upon the marginal totals m1, m2, m3, m4 of the resulting 2×2 table remaining fixed. It is considered to be conservative so may fail to indicate significance at the 5% level when the lower 95% confidence limit for the difference is greater than zero. The mid-P method has been advocated by Armitage and colleagues and Campbell and Swinscow to overcome this problem. To obtain the mid-P value, half the probability of the observed result is taken from the fisher.test P-value.

```
# calculating difference between and ratio of two proportions
library(PropCIs)
z<-1.96
x1<-8
n1<-114
x2<-14
n2<-74
y<-0
if (x2/n2>x1/n1){y<-x1;x1<-x2;x2<-y;y<-n1;n1<-n2;n2<-y}
q1<-n1-x1
q2<-n2-x2
diff<-x1/n1-x2/n2
lo1<-midPci(x1,n1,.05)[1]
up1<- midPci(x1,n1,.05)[2]
```

```
lo2<-midPci(x2,n2,.05)[1]
up2<- midPci(x2,n2,.05)[2]
l<-diff-((x1/n1-lo1)^2+(up2-x2/n2)^2)^.5
u<-diff+((x2/n2-lo2)^2+(up1-x1/n1)^2)^.5
diff
#[1] 0.1190138
l;u
#[1] 0.02228644
#[1] 0.2263485
#confidence limits for the risk difference of .119 are .022 & .226.
#
#Risk ratio
#using riskscoreci()
rr<-x1*n2/(x2*n1)
rr
#[1] 2.695946
library(PropCIs)
riskscoreci(x1,n1,x2,n2,.95)
#[1] 1.217218 5.995263
#confidence limits for the risk ratio of 2.7 are 1.2 & 6.

# Fisher Exact test
x1<-14
n1<-74
x2<-8
n2<-114
q1<-n1-x1
q2<-n2-x2
u1 <- c(x1, x2)
u2 <- c(q1, q2)
u3 <- data.frame(u1, u2)
ft <- fisher.test(u3)
ft1 <- ft$p.value
ft1 # Fisher test P-value
#[1] 0.01908641
sw <- T
if (x1 == 0 | x2 == 0) {
sw <- F
}
if (sw == T) {
a <- x1
a[2] <- q1
a[3] <- x2
a[4] <- q2
a[5] <- a[1] + a[2]
a[6] <- a[3] + a[4]
a[7] <- a[1] + a[3]
a[8] <- a[2] + a[4]
a[9] <- a[7] + a[8]
f <- 0
for (i in 1:9) {
if (a[i] > 100) {
f[i] <- a[i] * log(a[i]) - a[i] + log((2 * pi * a[i])^0.5)
}
else {
f[i] <- log(factorial(a[i]))
```

```
}
}
s1 <- sum(f[5:8])
s2 <- sum(c(f[1:4], f[9]))
d0 <- exp(s1 - s2)
fishermidp <- ft1 - d0/2
fishermidp # mid-P value
}
#[1] 0.01437578
```

1.4.3 The twoproportions function

The function twoproportions() incorporates the above calculations in a convenient format. Data may be entered at the keyboard or as twoproportions(data.frame(x_1,n_1,x_2,n_2)) for example, twoproportions(data.frame(14,74,8,114)).

```
#the twoproportions function (also via IMenu())
twoproportions()
```

```
Enter first numerator 14
Enter first denominator 74
Enter second numerator 8
Enter second denominator 114
First proportion 0.189, Second proportion 0.07
Difference between proportions 0.119.
Lower 95% limit 0.022, upper limit 0.226
Fisher Exact P-value = 0.019, LR = 1/16
Fisher Mid-P-value = 0.014, LR = 1/20
Ratio 2.7.
Ratio 95% confidence limits are 1.22 and 6.
```

```
d<-data.frame(14,74,8,114)
IMenu(d)
```

```
Introductory Menu

1. Proportion data,
2. Count and rate data.
1
1. Single proportion,
2. Confidence intervals for a series of proportions,
3. Two proportions,
4. More than two proportions,
5. Weighted average of proportions.
3
First proportion 0.189, Second proportion 0.07
Difference between proportions 0.119.
Lower 95% limit 0.022, upper limit 0.226
Fisher Exact P-value = 0.019, LR = 1/16
Fisher Mid-P-value = 0.014, LR = 1/20
Ratio 2.7.
Ratio 95% confidence limits are 1.22 and 6.
```

Occasionally the confidence intervals and the Mid-P P-value will be discordant, for example, the difference CI may not include zero or the ratio interval may not include one when the P-value is above 0.05. Alternatively, one or the other may not be reported, for example, in a published paper. Approximate CIs can be obtained from the P-value or the approximate P-value from the CIs using the methods described by Altman and Bland. Clearly, in the first instance (when there is discordance), one should adopt a consistent approach. A suitable one could be to use the riskscoreci() interval, that on a log scale is likely to be symmetrical, to obtain a concordant P-value. In the second instance (absence of one or other estimate in a published report), either one may be required, so we illustrate both. Since the interval for the logarithm of the risk ratio (RR) is usually closer to symmetrical than the risk difference (RD), we illustrate the Altman and Bland methods with the former. For the confidence interval $Z \approx 0.862 + (0.743 - 2.404 \times \log(\text{P-value}))^{.5}$, $SE \approx \text{abs}(\log(RR)/Z)$ and $CI \approx \exp(\log(RR) \pm 1.96^{*}SE)$. For the latter, $SE \approx (\log(U95) - \log(L95))/(2 \times 1.96)$, $Z \approx \log(RR)/SE$ and $\text{P.value} \approx \exp(-0.717 \times Z - 0.416 \times Z^2)$, where CI is the confidence interval and U95 and L95 are the upper and lower 95% confidence limits.

```
#adjusting the difference CI using the P-value (Bland and Altman)
P<-0.014
D<-0.119
Z<--0.862+(0.743-2.404*log(P))^.5
Z
#[1] 2.455371
SE<-abs(D/Z)
SE
#[1] 0.04846518
U<-D+1.96*SE
U
#[1] 0.2139918
L<-D-1.96*SE
L
#[1] 0.02400824

#adjusting the ratio CI using the P-value
P<-0.014
R<-2.7
Z<--0.862+(0.743-2.404*log(P))^.5
Z
#[1] 2.455371
SE<-abs(log(R)/Z)
SE
#[1] 0.4045221
U<-exp(log(R)+1.96*SE)
U
#[1] 5.966229
L<-exp(log(R)-1.96*SE)
L
#[1] 1.221877
#the Mid-P P-value and the ratio CIs are in good agreement for these data

#using the CI to recalculate the P-value
#calculating approximate P-value using difference (Altman and Bland)
#difference CIs must be approximately symmetric about the difference
SE<-(0.2263485-0.02228644)/(1.96*2)
```

```
Z<-0.1190138/SE
P.value<-exp(-0.717*Z-0.416*Z^2)
Z;P.value
#[1] 2.286236
#[1] 0.02206832
#this does not agree well with the Fisher Mid-P result
# difference between estimate and CIs
0.1190138-0.02228644
#[1] 0.09672731
0.2263485-0.1190138
#[1] 0.1073348
#there is some asymmetry about the difference

#calculating approximate P-value using ratio (Altman and Bland)
#logs used to obtain approximate symmetry about the risk ratio
SE<-(log(5.995263)-log(1.217218))/(2*1.96)
Z<-log(2.695946)/SE
P.value=exp(-0.717*Z-0.416*Z^2)
Z;P.value
#[1] 2.438317
#[1] 0.0146758
#good agreement with Fisher Mid-P result
```

1.5 Introducing a Bayesian approach

Although Bayesian methods are not the focus of these notes, they are increasingly being used for quality improvement in general and hospital surveillance in particular. There is a growing literature on useful approaches. These need not be complex; for example, as we have seen, a simple Beta distribution can be used to describe a population rate. Interested readers are referred, for example, to the texts by Woodworth and Spiegelhalter[F] and colleagues.

1.6 When the data are not just one or two independent samples

There are several issues when there are a number of samples. For example, is a comparison required or an average? The two common stratifying agents in QI and IM work are hospitals and years. The former are often exchangeable in the sense that their order is unimportant and they may be thought of as random selections from a group of hospitals unless there is prior reason to believe otherwise, for example, that hospital size and AE rates are related. However, with years it is frequently of interest to obtain a weighted average with stronger weighting for more recent years, or some QI activity may have been undertaken and a downwards trend or change point may have been anticipated. Years can therefore usually be thought of as being fixed agents, although it is possible that there may be times when years are exchangeable (although time series data are often not independent, when stratified by years, counts of AEs can usually be analysed as if independent). In the presence of inhomogeneity among the samples, due to excessive variation or interaction or a combination of both, analyses may differ. When this occurs and the stratifying agent is fixed, attention should be directed to the individual within stratum data but when it is random, a random effects analysis may

be possible. The issues of exchangeability and fixed and random effects are important, for example, when AEs from several hospitals are analysed in funnel plots in the presence of excessive variability (Spiegelhalter[B,C,D]), or when employing shrinkage analysis, and we return to it in later chapters.

1.6.1 More than two independent proportions

In some cases, there may be more than two proportions to study and compare and sometimes it may make sense to expect that a trend might exist among the proportions. These data may be analysed in a fixed analysis with the R function prop.test() that employs the chi-squared (χ^2) distribution. If a trend test is required, the R prop.trend.test() function is employed and departure from trend can be assessed by subtraction (Armitage and colleagues), that is, the prop.test χ^2 has degrees-of-freedom (DF) equal to the number of proportions minus one and the prop.trend.test χ^2 has one DF. Subtracting the latter χ^2 from the former gives a departure from trend χ^2 with DF equal to the number of proportions minus two. If the departure from trend χ^2 is large relative to its DF, we should examine the proportions individually to see which might be different; these are displayed by prop.test. Usually we will be alerted by the prop.test result whether or not a trend test is employed but occasionally there will be both a trend and a departure from the trend. Breslow and Day describe these χ^2 tests. Also, prop.test is unsuitable when samples are small.

For prop.test, let x_i be the numerator and n_i the denominator of the i^{th} of the I proportions, $p_i=x_i/n_i$, $p=\sum x_i/\sum n_i$ and q=1-p. Then $\chi^2=\sum((n_i\times(p_i-p)^2)/(p\times q))$ with DF=I-1. For prop.trend.test, also let $X=\sum x_i$, i=1..I and $N=\sum n_i$. Then $a=N\times(N\times\sum(x_i\times i)-X\times\sum(n_i\times i))^2$, $b=X\times(N-X)\times(N\times\sum(n_i\times i^2)-(\sum(n_i\times i))^2)$ and the trend $\chi^2=a/b$ with 1 DF.

It is useful to investigate those proportions that appear to differ. In cases where there is a reference rate or, as with the present data, risk adjustment values, these can be used for this purpose. For the present, we assume that the average rate for the data can be used to see which, if any, of the proportions differ, for example, by calculating Z-scores. We deal with risk adjusted data in Chapter 2.

1.6.2 Example 1, yearly data

The file ssi0106.csv has SSI data for a group of related surgical procedures from 2001 to 2006. SSI data from hospital G are selected to see if there were yearly variations in SSI rates and, if so, whether a trend existed.

```
#getting ssi0106.csv data via clipboard
g.d()
```

```
Loading data.
Data from clipboard (C) or file (F) c
Do data column(s) have heading(s) (Y/N) y
Is column 1 a date column (Y/N) n
```

```
options(digits=4) # reduce number of decimal places printed
# remember to change back to default (digits=7) if required
# prop.test()
G<-datain[datain$Hospital=="G",c(3,5)]
```

```
Y<-G$ProcedureDate
Y<-chron(as.character(Y),format="d-mmm-yy",out.format="dd-mmm-yyyy")
Yrs<-years(Y)
SSIs<-tapply(G$SSI,Yrs,sum)
Proc<-tapply(G$SSI,Yrs,length)
GYrs<-data.frame(SSIs,Proc)
#prop.test
pt<-prop.test(SSIs,Proc) # prop.test requires SSIs & Totals
pt
```

```
6-sample test for equality of proportions without continuity correction
data:  SSIs out of Proc
X-squared = 25.9, df = 5, p-value = 9.324e-05
alternative hypothesis: two.sided
sample estimates:
 prop 1  prop 2  prop 3  prop 4  prop 5  prop 6
0.03966 0.05476 0.04711 0.09456 0.03770 0.02778
```

The test suggests that the proportions are likely to differ and that 2004 may be an outlier.

```
# using prop.trend.test
ptt<-prop.trend.test(SSIs,Proc)
ptt
```

```
          Chi-squared Test for Trend in Proportions
data:  SSIs out of Proc ,
 using scores: 1 2 3 4 5 6
X-squared = 1.1741, df = 1, p-value = 0.2786
```

There is no evidence of a trend. The departure from trend is highly significant and we wish to find which proportions differ.

```
# departure from trend
Departure<-pt$statistic-ptt$statistic #departure from trend
DF<-pt$parameter-ptt$parameter # degrees of freedom
P.value<-as.numeric(1-pchisq(Departure,DF))
Departure;DF;P.value
#
#X-squared
#  24.7275
#df
# 4
# P-value
#[1] 5.707159e-05

# finding which proportions differ
# Z-scores obtained from the fixed analysis expected counts
# assumes average rate from the data
SSI<-GYrs[,1]
Totals<-GYrs[,2]
s1<-sum(SSI)
s2<-sum(Totals)
```

```
mn<-s1/s2
expected<-Totals*mn
Expecteds<-round(expected,2)
GYrs1<-data.frame(GYrs,Expecteds)
library(exactci)
rate<-GYrs1[,1]/GYrs1[,2]
d<-GYrs1[,3]/GYrs1[,2]
Z<-0;p<-0
for (i in 1:length(GYrs1[,1])){p[i]<-
binom.exact(GYrs1[i,1],GYrs1[i,2],d[i])$p.value;if(rate[i]>d[i])
{Z[i]<--qnorm(p[i]/2)}else{Z[i]<-qnorm(p[i]/2)}}
Z<-round(Z,2)
GYrs2<-data.frame(GYrs1,Z)
GYrs2
```

	SSIs	Proc	Expecteds	Z
2001	14	353	17.39	−0.69
2002	19	347	17.10	0.38
2003	22	467	23.01	−0.08
2004	40	423	20.84	3.78
2005	19	504	24.83	−1.11
2006	14	504	24.83	−2.26

Although not always strictly correct since the proportions may not be independent (e.g. as here by calculating the expected counts from the overall SSI rate), multiple confidence intervals or Z-scores are often used with these data to identify outliers. Here, we illustrate the use of Z-scores for this purpose. There are Z-scores greater |2| in 2004 and 2006. In the absence of risk adjustment or a reference rate, this analysis could be used to identify areas requiring further study, for example, using an M&M audit.

Although these data do not display a trend, inspection of GYrs3 suggests a trend during the latter three years and repeating prop.trend.test with GYrs[4:6,] confirms this. Had there been a QI response to the relatively high SSI rate in 2004, for example, an M&M audit with implementation of an evidence-based bundle, this could have been evidence of a subsequent downward trend.

```
prop.trend.test(SSIs[4:6],Proc[4:6])$p.value
#
#[1] 6.509e-06
```

The function manyproportions() can be used to perform the calculations using prop.test. It includes the option to calculate P-values by Monte Carlo simulation and Z-scores modified for multiple significance testing using the Benjamini-Hochberg procedure. Data may be entered manually using manyproportions() or from a data.frame.

```
SSI<-c(14,19,22,40,19,14)
Proc<-c(353,347,467,423,504,504)
manyproportions(data.frame(SSI,Proc))
```

```
6-sample test for equality of proportions without continuity correction.
Chisq = 25.9, DF = 5, P-value = 0.
P-value by Monte Carlo simulation = 0.
Is a trend expected (Y/N) y
```

```
Chi-squared Test for Trend in Proportions.
Chisq = 1.174, DF = 1, P-value = 0.279.
Departure from trend chisq = 24.73, DF = 4, P-value = 0.

    Successes   Totals   Proportions   Expecteds       Z    ZAdj
1          14      353          0.04       17.39   −0.69   −0.34
2          19      347          0.05       17.10    0.38    0.20
3          22      467          0.05       23.01   −0.08   −0.08
4          40      423          0.09       20.84    3.78    3.31
5          19      504          0.04       24.83   −1.11   −0.62
6          14      504          0.03       24.83   −2.26   −1.81
```

```
# via IMenu()
d<-data.frame(SSI,Proc)
IMenu(d)
```

```
Introductory Menu

1. Proportion data,
2. Count and rate data.
1
1. Single proportion,
2. Confidence intervals for a series of proportions,
3. Two proportions,
4. More than two proportions,
5. Weighted average of proportions.
4

6-sample test for equality of proportions without continuity correction.
Chisq = 25.9, DF = 5, P-value = 0.
P-value by Monte Carlo simulation = 0.
Is a trend expected (Y/N) y
Chi-squared Test for Trend in Proportions.
Chisq = 1.174, DF = 1, P-value = 0.279.
Departure from trend chisq = 24.73, DF = 4, P-value = 0.

    Successes   Totals   Proportions   Expecteds       Z    ZAdj
1          14      353          0.04       17.39   −0.69   −0.34
2          19      347          0.05       17.10    0.38    0.20
3          22      467          0.05       23.01   −0.08   −0.08
4          40      423          0.09       20.84    3.78    3.31
5          19      504          0.04       24.83   −1.11   −0.62
6          14      504          0.03       24.83   −2.26   −1.81
```

1.6.3 Example 2, hospital data

Orthopaedic SSI data from a group of hospitals are in the file orthcomp0106.csv. These data were collected between 2001 and 2006. We will examine the data for the earlier period of data collection 2001 to 2004 and the later period of 2005 and 2006. For the present we ignore risk stratification and restrict the analysis to hospitals reporting 10 or more SSIs in the later period. We assume that the hospitals are exchangeable.

```
g.d # orthcomp0106.csv
```

```
Loading data.
Data from clipboard (C) or file (F) c
Do data column(s) have heading(s) (Y/N) y
Is column 1 a date column (Y/N) n
```

```
head(datain)
```

```
    Hospital   ProcedureDate   SSI
1          A        1-Feb-01     0
2          A        5-Feb-01     0
3          A        7-Feb-01     0
4          A        8-Feb-01     0
5          A       12-Feb-01     0
6          A       14-Feb-01     0
```

```
# ordering by hospital
o<-order(datain[,1])
datain<-datain[o,]
# getting before and after end of 2004
d<-datain[,2]
library(chron)
d<-chron(as.character(d),format="d-mmm-yy",out.format="dd-mmm-yyyy")
Before<-datain[d<"01-Jan-2005",]
After<-datain[d>"31-Dec-2004",]
ssib<-tapply(Before[,3],Before[,1],sum)
procb<-tapply(Before[,3],Before[,1],length)
nameb<-names(ssib)
OrthSSIb<-data.frame(nameb,ssib,procb)
lb<-1:length(OrthSSIb[,1])
row.names(OrthSSIb)<-lb
ssia<-tapply(After[,3],After[,1],sum)
proca<-tapply(After[,3],After[,1],length)
namea<-names(ssia)
OrthSSIa<-data.frame(namea,ssia,proca)
la<-1:length(OrthSSIa[,1])
row.names(OrthSSIa)<-la
Group2<-OrthSSIa[OrthSSIa[,2]>10,]
Group1<-
OrthSSIb[OrthSSIb[,1]=="A"|OrthSSIb[,1]=="B"|OrthSSIb[,1]=="E"|
OrthSSIb[,1]=="F"|OrthSSIb[,1]=="G"|OrthSSIb[,1]=="I"|
OrthSSIb[,1]=="J"|OrthSSIb[,1]=="K"|OrthSSIb[,1]=="N"|OrthSSIb[,1]=="Q",]
options(digits=4)
prop.test(Group2$ssia,Group2$proca)
```

```
X-squared = 29.13, df = 9, p-value = 0.0006173
sample estimates:
 prop 1  prop 2  prop 3  prop 4  prop 5  prop 6  prop 7  prop 8  prop 9
0.08264 0.03325 0.05288 0.06180 0.03274 0.06849 0.04950 0.04898 0.06007
prop 10
0.02634
```

```
prop.test(Group1$ssib,Group1$procb)
```

```
X-squared = 58.35, df = 9, p-value = 2.783e-09
sample estimates:
 prop 1  prop 2  prop 3  prop 4  prop 5  prop 6  prop 7  prop 8
0.09631 0.03687 0.02926 0.05108 0.05975 0.06303 0.03650 0.07759
 prop 9 prop 10
0.06863 0.02609
```

These data are not homogeneous. This is not surprising as post-discharge SSI data collection was optional. We concentrate on the more recent data. In Chapter 2 we use expected numbers of outcomes derived from the NNIS risk index to determine which institutions differ from expected but here we obtain expected numbers from the unadjusted data.

```
# finding which proportions differ in Group2, 2005-2006 data
# assumes average rate from the data
SSI<-Group2$ssia
Totals<-Group2$proca
s1<-sum(SSI)
s2<-sum(Totals)
mn<-s1/s2
expected<-Totals*mn
Expecteds<-round(expected,2)
Group2a<-data.frame(Group2,Expecteds)
library(exactci)
rate<-Group2a[,2]/Group2a[,3]
d<-Group2a[,4]/Group2a[,3]
Z<-0;p<-0
for (i in 1:length(Group2a[,1])){p[i]<-binom.exact(Group2a
[i,2],Group2a[i,3],d[i])$p.value;if(rate[i]
>d[i]){Z[i]<--qnorm(p[i]/2)}else{Z[i]<-qnorm(p[i]/2)}}
Z<-round(Z,2)
Group2b<-data.frame(Group2a,Z)
Group2b
```

	namea	ssia	proca	Expecteds	Z
1	A	20	242	10.66	2.54
2	B	13	391	17.23	−0.91
5	E	22	416	18.33	0.77
6	F	11	178	7.84	0.98
7	G	33	1008	44.42	−1.72
9	I	25	365	16.08	2.04
10	J	15	303	13.35	0.36
11	K	12	245	10.80	0.26
14	N	17	283	12.47	1.16
17	Q	25	949	41.82	−2.75

```
# random effects analysis
# data as individual observations
Bbefore<-
Before[Before[,1]=="A"|Before[,1]=="B"|Before[,1]=="E"|Before[,1]=="F"|
Before[,1]=="G"|Before[,1]=="I"|Before[,1]=="J"|Before[,1]=="K"|
Before[,1]=="N"|Before[,1]=="Q",]
```

```
s.d(Bbefore)
#
#Use clipboard (c) or a data file (f) f
#Enter a file name d:/Before
#

Bafter<-
After[After[,1]=="A"|After[,1]=="B"|After[,1]=="E"|After[,1]==
"F"|After[,1]=="G"|After[,1]=="I"|After[,1]=="J"|After[,1]=="K"
|After[,1]=="N"|After[,1]=="Q",]

s.d(Bafter)
#
#Use clipboard (c) or a data file (f) f
#Enter a file name d:/After
#

g.d() # After.csv

library(hglm)
h<-
hglm(fixed=datain$SSI~1,random=~1|datain$Hospital,fix.disp=1,
family=binomial(link=logit))
Diffs<-data.frame(Group2b,round(h$ranef/h$SeRe,2))
names(Diffs)<-c("Name","SSI","Proc","Exp","Zfixed","Zrandom")
row.names(Diffs)<-1:length(Diffs[,1])
Diffs
```

	Name	SSI	Proc	Exp	Zfixed	Zrandom
1	A	20	242	10.66	2.54	1.72
2	B	13	391	17.23	−0.91	−1.05
3	E	22	416	18.33	0.77	0.33
4	F	11	178	7.84	0.98	0.56
5	G	33	1008	44.42	−1.72	−1.73
6	I	25	365	16.08	2.04	1.30
7	J	15	303	13.35	0.36	0.07
8	K	12	245	10.80	0.26	0.04
9	N	17	283	12.47	1.16	0.67
10	Q	25	949	41.82	−2.75	−2.44

We see, from this seemingly simple data set, how difficult such aggregated hospital AE data can be to analyse. When hospitals that are members of a group of hospitals and are exchangeable are compared, this difficulty may be compounded. It seems much better if institutions implement evidence-based bundles and monitor their own data sequentially (Chapters 3 and 6). Sometimes average rates for several years appear in reports. For the After data the average rate was $100\times\sum$SSIs/$\sum$Proc=4.41%. This differs from the rates for most of the hospitals (8.26, 3.32, 5.29, 6.18, 3.27, 6.85, 4.95, 4.90, 6.01 and 2.63 per cent respectively). It is difficult to see how, in this case, such an average could be interpretable. Also, we should bear in mind the advice of Mohammed and colleagues: look first for data or analysis error, then for problems with risk adjustment if used, then for system errors, and finally for problems involving staff. For example, was data collection different among the hospitals? We know that collection of post-discharge SSI data was optional at the time these data were collected and those who chose to collect these data would have higher rates. In

addition, we should consider variations related to regression to the mean; an institution's AE rate may vary randomly within predictable limits from year to year without any system change and this may be of particular concern with smaller institutions.

1.6.4 Prop test and small samples

When there are small expected values, the prop.test χ^2 can become artificially large, it provides a warning message and its result can be ignored. This will often be the case with SSI and similar adverse event (AE) data. When samples are small, prop.test() can be replaced with fisher.test(). We illustrate with similar data from a smaller hospital Q. Although it will usually be unnecessary, it is possible to obtain a 2 DF χ^2 value corresponding to the fisher.test() P-value (χ^2_2=-2×log(P-value)). The prop.trend.test() is less affected by small sample numbers than prop.test() and its χ^2_1 value could be compared with the χ^2_2 value from the fisher.test() to obtain an approximate departure from trend analysis. The Z-scores should remain useful. Although there is no suggestion of a significant difference or a trend among the yearly data from hospital Q, we use them to illustrate the calculations employing the Fisher exact test.

```
#getting ssi0106.csv data via clipboard
g.d()

#data from hospital Q
Q<-datain[datain$Hospital=="Q",c(3,5)]
Y<-Q$ProcedureDate
Y<-chron(as.character(Y),format="d-mmm-yy",out.format="dd-mmm-yyyy")
Yrs<-years(Y)
SSIs<-tapply(Q$SSI,Yrs,sum)
Proc<-tapply(Q$SSI,Yrs,length)
QYrs<-data.frame(SSIs,Proc)
QYrs<-QYrs[-1,] #remove single 2001 procedure
#
#Fisher test
ft<-fisher.test(data.frame(QYrs$SSIs,QYrs$Proc-QYrs$SSIs))
ft$p.value
#
#[1] 0.1596462
#
FT<--2*log(ft$p.value)
FT #fisher.test χ², DF=2
#
#[1] 3.66959
#
TT<-prop.trend.test(QYrs$SSIs,QYrs$Proc)
TT$statistic #prop.trend.test X2, DF=1
#X-squared
#0.08975511
as.numeric(1-pchisq(TT$statistic,1)) # trend test P-value
#0.7644887
DT<-FT-TT$statistic
DT #departure from trend X2, DF=1
#X-squared
#3.579835
as.numeric(1-pchisq(DT,1)) # departure from trend P-value
#0.05848496
```

Since the departure from trend χ^2 result of 0.06 is close to conventional statistical significance, further scrutiny of these data might be warranted. Comparing the first (2002) and second (2003) yearly data gives the following result.

```
twoproportions(data.frame(7,165,5,393))
```

```
First proportion 0.042, Second proportion 0.013
Difference between proportions 0.03.
Lower 95% limit 0.002, upper limit 0.07
Fisher Exact P-value = 0.048, LR = 1/7
Fisher Mid-P-value = 0.035, LR = 1/9
Ratio 3.33.
Ratio 95% confidence limits are 1.13 and 9.81.
```

A logistic regression analysis can be performed. The year 2003 appears to differ from the 2002 reference year but the overall test for the model gives a P-value of 0.17, similar to that of the Fisher exact test. The test involving just the 2002 and 2003 data is suggested by the data and fails to provide information about the overall group of years. Post-hoc analyses of this kind can mislead. It would only be of interest if, at the end of 2002, the test had been performed because of prior interest in these two years, for example, there may have been a change in the management of surgical wounds in 2003. Also, there were more procedures performed that year and this apparent increase in workload could have produced an interest in its influence on SSI rates at that time.

```
# logistic regression
YEARS<-c(rep(1,165),rep(2,393),rep(3,476),rep(4,477),rep(5,472))
SSIS<-c(rep(1,7),rep(0,165-7),rep(1,5),rep(0,393-5),rep(1,15),rep(0,476-
15),rep(1,10),rep(0,477-10),rep(1,15),rep(0,472-15))
g<-glm(SSIS~as.factor(YEARS),family=binomial)
summary(g)
```

```
Coefficients:
                   Estimate Std.    Error   z value   Pr(>|z|)
(Intercept)              −3.1167   0.3862   −8.069   7.08e-16   ***
as.factor(YEARS)2        −1.2349   0.5931   −2.082     0.0373   *
as.factor(YEARS)3        −0.3087   0.4669   −0.661     0.5086
as.factor(YEARS)4        −0.7271   0.5013   −1.450     0.1470
as.factor(YEARS)5        −0.2999   0.4669   −0.642     0.5206
    Null deviance: 481.30 on 1982 degrees of freedom
Residual deviance: 474.84 on 1978 degrees of freedom
1-pchisq(481.3-474.84,4) #deviance χ² test
[1] 0.1673282
```

1.7 Summarising stratified proportion data

Frequently, rather than seeking differences in annual AE rates, there is an interest in obtaining an overall average value. However, in doing so it often makes sense to give more weight to the most recent data, for example last year's data could receive half the weight of this year's data with each preceding year receiving half the weight of its successor. The result is a directly standardised rate. We illustrate the necessary calculations using Newcombe's square-and-add procedure.

The function stratifiedproportionsn() employs Newcombe's procedure as follows: X_i and N_i are the numerators and denominators for stratum i and W_i are the weights. The upper and lower limits u_i and l_i for each stratum are calculated, for example using midPci() in the PropCIs library as previously described. $p_i=X_i/N_i$ and $W_i=W_i/\sum W_i$ (so the weights sum to one). Then the weighted average is $A_W=\sum W_i p_i$. The lower (L) limit is $L=A_W-\sqrt{l}$ and the upper limit (U) is $U=A_W+\sqrt{u}$, where $l=\sum(((p_i-l_i)\times W_i)^2)$ and $u=\sum(((u_i-p_i)\times W_i)^2)$.

We now illustrate with data from the smaller hospital. Data may be entered into the R editor from the keyboard or as a data.frame. Once again, we should check that the proportions are reasonably homogeneous.

```
# check of homogeneity
x1<-c(7,5,15,10,15)
n1<-c(165,393,476,477,472)
fisher.test(data.frame(x1,n1-x1))$p.value
#
#[1] 0.1596 (suggests proportions are reasonably homogeneous
#stratified proportion data calculations
library(PropCIs)
I<-length(n1)
w<-1
for (i in 2:I){w[i]<-w[i-1]*2}
i1<-1:I
N<-rep(0,I)
up<-rep(0,I)
lo<-rep(0,I)
up1<-rep(0,I)
lo1<-rep(0,I)
s<-data.frame(i1,x1,n1,w)
#default weights geometrically decreasing, may be changed
s<-edit(s) # to R data editor for checking & editing data

#click on right top cross to return to the analysis
w<-s$w/sum(s$w)
u11<-0;l11<-0
for (i in 1:I){
l11[i]<-midPci(s$x1[i],s$n1[i],.05)[1]
u11[i]<-midPci(s$x1[i],s$n1[i],.05)[2]
}
p1<-s$x1/s$n1
a1<-sum(p1*w)
a0<-a1/sum(w)
a22<-sum(((p1-l11)*w)^2)
a33<-sum(((p1-u11)*w)^2)
l011<-a0-a22^.5;u011<-a0+a33^.5
cat("Weighted average = ",round(a0,3),".\nLower 95% limit =
",round(l011,3),", upper limit = ",round(u011,3),".\n",sep="")
#Weighted average = 0.028.
#Lower 95% limit = 0.021, upper limit = 0.039.

# using stratifiedproportionsn()
#to enter data from the keyboard, use stratifiedproportionsn()
#the R data editor will appear for manual data entry
#illustrating entry as a data.frame
x1<-c(7,5,15,10,15)
```

```
n1<-c(165,393,476,477,472)
w<-c(1,2,4,8,16)
s<-data.frame(x1,n1,w)
stratifiedproportionsn(s) #entry from data.frame

#check data in editor revise weights if required then click on close.
```

```
Weighted average = 0.028.
Lower 95% limit = 0.021, upper limit = 0.039.
```

The function stratifiedproportionsn() may also be accessed via IMenu().

```
x1<-c(7,5,15,10,15)
n1<-c(165,393,476,477,472)
w<-c(1,2,4,8,16)
s<-data.frame(x1,n1,w)
IMenu(s)
```

```
Introductory Menu

1. Proportion data,
2. Count and rate data.
1
1. Single proportion,
2. Confidence intervals for a series of proportions,
3. Two proportions,
4. More than two proportions,
5. Weighted average of proportions.
5
Weighted average = 0.028.
Lower 95% limit = 0.021, upper limit = 0.039.
```

1.8 Stratified proportion data, differences between rates

It may occasionally be necessary to assess the difference between two sets of stratified rates, for example, rates of SSIs for two units doing the same procedures. Once again we may need to distinguish between data that are stratified by fixed and random agents. Here we employ some hypothetical small-sample data to illustrate the former using Newcombe's method. When the rates and CIs for each of the sets of stratified rates have been determined, as in the previous section, Newcombe's method for the difference between them can be applied as described in the section describing the difference between two independent proportions. Data may be entered into the R editor from the keyboard or as a data.frame.

```
# differences between rates, hypothetical data
GroupA<-c(1,32,4,43,2,21);GroupB<-c(0,97,3,142,2,49)
ssi2<-data.frame(GroupA,GroupB)
row.names(ssi2)<-
c("Stratum1Failures","Stratum1Procedurers","Stratum2Failures",
"Stratum2Procedurers","Stratum3Failures","Stratum3Procedurers")
ssi2
```

	GroupA	GroupB
Stratum1Failures	1	0
Stratum1Procedurers	32	97
Stratum2Failures	4	3
Stratum2Procedurers	43	142
Stratum3Failures	2	2
Stratum3Procedurers	21	49

```
Group<-c(rep(1,3),rep(0,3))
Stratum<-gl(3,1,6)
AE<-c(1,4,2,0,3,2)
Total<-c(32,43,21,97,142,49)
Hypothdat<-data.frame(Group,Stratum,AE,Total)
x1<-Hypothdat[1:3,3]
n1<-Hypothdat[1:3,4]
x2<-Hypothdat[4:6,3]
n2<-Hypothdat[4:6,4]
I<-(length(Hypothdat[,1]))/2
w<-rep(1,I) #equal weights
i1<-1:I
s<-data.frame(i1,x1,n1,x2,n2,w)
s<-edit(s) #to R data editor for checking & editing data

# click on right top cross to return to the analysis
#stratified risk difference using Newcombe's method
library(PropCIs)
x1<-s$x1;x2<-s$x2;n1<-s$n1;n2<-s$n2;w<-s$w
s$y1<-s$n1+s$n2
w<-w/sum(w)
u11<-0;l11<-0
for (i in 1:I){
l11[i]<-midPci(x1[i],n1[i],.05)[1]
u11[i]<-midPci(x1[i],n1[i],.05)[2]
}
u22<-0;l22<-0
for (i in 1:I){
l22[i]<-midPci(x2[i],n2[i],.05)[1]
u22[i]<-midPci(x2[i],n2[i],.05)[2]}
p1<-x1/n1
a1<-sum(p1*w)
a0<-a1/sum(w)
a22<-sum(((p1-l11)*w)^2)
a33<-sum(((p1-u11)*w)^2)
l011<-a0-a22^.5;u011<-a0+a33^.5
p2<-x2/n2
b1<-sum(p2*w)
b0<-b1/sum(w)
b22<-sum(((p2-l22)*w)^2)
b33<-sum(((p2-u22)*w)^2)
l022<-b0-b22^.5;u022<-b0+b33^.5
d<-a0-b0
lomh1<-d-((a0-l011)^2+(u022-b0)^2)^.5
upmh1<-d+((b0-l022)^2+(u011-a0)^2)^.5
N1<-paste("\nNewcombe's method using mid-P binomial 95% confidence
limits.\n")
```

```
N2<-paste("Weighted average in first group ",round(a0,3),".",sep="")
N3<-paste("\nFirst group confidence limits ",round(l011,3),"
to ",round(u011,3),".",sep="")
N4<-paste("\nWeighted average in second group ",round(b0,3),".",sep="")
N5<-paste("\nSecond group confidence limits ",round(l022,3),"
to ",round(u022,3),".",sep="")
N6<-paste("\nWeighted difference ",round(d,3),".",sep="")
N7<-paste("\nDifference confidence limits ",round(lomh1,3),"
to ",round(upmh1,3),".",sep="")
cat(N1,N2,N3,N4,N5,N6,N7,"\n",sep="")
```

```
Newcombe's method using mid-P binomial 95% confidence limits.
Weighted average in first group 0.073.
First group confidence limits 0.038 to 0.155.
Weighted average in second group 0.021.
Second group confidence limits 0.008 to 0.054.
Weighted difference 0.053.
Difference confidence limits 0.005 to 0.135.
```

```
#stratified risk ratio using Newcombe's method
P1<-a0
P2<-b0
U1<-u011
L1<-l011
U2<-u022
L2<-l022
U=exp(log(P1)-log(P2)+((log(U1)-log(P1))^2+(log(P2)-log(L2))^2)^.5)
L=exp(log(P1)-log(P2)-((log(P1)-log(L1))^2+(log(U2)-log(P2))^2)^.5)
R1<-paste("\nRatio ",round(P1/P2,3),".",sep="")
R2<-paste("\nApproximate 95% ratio confidence limits
",round(L,3)," to ",round(U,3),".",sep="")
cat(R1,R2,"\n",sep="")
```

```
Ratio 3.544.
Approximate 95% ratio confidence limits 1.119 to 11.575.
```

1.8.1 Yearly data

The function stratified2proportionsn() performs the calculations described in section 1.8. Data may be entered via the R editor from the keyboard (leave the space between () blank and use the R data editor for data entry), or enter the data as a data.frame. The former method is tedious and can be prone to error so the latter is recommended. The analysis using stratified2proportionsn() is illustrated with yearly SSI data from two hospitals. Note that when the larger rates are in the second group, the groups are interchanged.

```
#using function stratified2proportionsn()
Year<-2002:2006
SSI1<-c(7,5,15,10,15)
Proc1<-c(165,393,476,477,472)
SSI2<-c(19,22,40,19,14)
Proc2<-c(347,467,423,504,504)
I<-length(Year)
```

```
w<-1:I # weights can be changed in the R Editor
TYrs<-data.frame(Year,SSI1,Proc1,SSI2,Proc2,w)
TYrs
```

```
  Year SSI1 Proc1 SSI2 Proc2 w
1 2002    7   165   19   347 1
2 2003    5   393   22   467 2
3 2004   15   476   40   423 3
4 2005   10   477   19   504 4
5 2006   15   472   14   504 5
```

```
stratified2proportionsn(TYrs[,-1]) #weights 1 to 5
#Enter required confidence level .95
```

```
The two groups have been interchanged so the larger is first.
Newcombe's method using mid-P binomial 95% confidence limits.
Weighted average in first group 0.04815626.
First group confidence limits 0.0403548 to 0.05852714.
Weighted average in second group 0.02701087.
Second group confidence limits 0.02098259 to 0.03603815.
Weighted difference 0.02114539.
Difference confidence limits 0.009214145 to 0.03314103.
Ratio 1.782847.
Ratio confidence limits 1.271269 to 2.452981.
```

An approximate P-value can be calculated using the method of Bland and Altman. As shown in section 1.9 that deals with heterogeneity, the CI may be too narrow and the P-value may be too small.

```
#calculating approximate P-value (Altman and Bland)
SE<-(log(2.452981)-log(1.271269))/(2*1.96)
Z<-log(1.782847)/SE
P.value=exp(-0.717*Z-0.416*Z^2)
Z;P.value
#[1] 3.448394
#[1] 0.0005995516

#calculating 3 SD equivalent limits using stratified2proportionsn()
stratified2proportionsn(TYrs[,-1])
#Enter required confidence level .997
```

```
The two groups have been interchanged so the larger is first.
Newcombe's method using mid-P binomial 99.7% #confidence limits.
Weighted average in first group 0.04815626.
First group confidence limits 0.03714774 to 0.06479125.
Weighted average in second group 0.02701087.
Second group confidence limits 0.01867971 to 0.0417506.
Weighted difference 0.02114539.
Difference confidence limits 0.002748458 to 0.03974999.
Ratio 1.782847.
Ratio confidence limits 1.073857 to 2.862108.
```

1.8.2 Hospital data

The previous section dealt with stratified yearly data where the stratifying agent, years, is fixed so that in the presence of heterogeneity among the strata, it will usually be best to study the within stratum rate differences and ratios individually. When the stratifying agent is hospitals, a random effects approach may be feasible and this is illustrated by Spiegelhalter[B,C,D] with funnel plots displaying data from groups of hospitals (Chapter 2). In this case, we are unlikely in advance to consider the hospitals as having special characteristics that would make them fixed agents. In the following section we discuss the role of Mantel-Haenszel, Homogeneity, Trend Tests and the derSimonian-Laird random effects method in the analysis of data stratified by years and by hospitals.

1.9 Mantel-Haenszel, homogeneity and trend tests

We are primarily interested in risk differences, and to a lesser extent risk ratios, because they seem to us to be the most useful for analysing routine surveillance data. The natural estimator for binomial data is the odds ratio (OR), and this is especially employed for case-control studies as well as being reported generally in analyses based on models such as logistic regression. While such studies are frequently performed in IM and QI departments for research, they are performed less frequently when analysing routine surveillance data (a relatively uncommon exception is the investigation of an outbreak, for example, of food poisoning, as described in Chapter 7). These notes are primarily for staff in IM and QI departments in hospitals as an aid to their analysis of routine surveillance data. Consequently, we have given less emphasis to the OR. However, it is employed in the following section.

We include an exact Mantel-Haenszel (MH) analysis (Rosner), an alternative to the Newcombe method for combining and comparing two sets of data in different strata that is based on the OR. The MH test is a generalisation of the Fisher exact test to data in several strata using weights derived from the data. The weights for the MH test for stratum i are $N_{1i} \times N_{2i}/(N_{1i}+N_{2i})$ where N_{1i} and N_{2i} are the denominators for that stratum (Armitage and colleagues). These weights have the desirable property of being inversely proportional to the stratum i variance. The MH method is more conservative than the Newcombe method described above as the latter is based on mid-P binomial intervals and the OR is based on the hypergeometric distribution. In addition, the ability to weight the data in the strata to obtain a directly standardised rate as described above no longer applies.

The stratification methods that employ the OR such as the MH method may be especially useful in routine IM and QI work when there may be heterogeneity (unusual variation) among the strata with the results in the individual strata differing. Tests for heterogeneity are often called homogeneity tests and they may include a test for a trend. In addition, there are random-effects modifications for dealing with heterogeneity such as the derSimonian-Laird method. The latter comes from meta-analysis where a number of related studies are analysed as a group and it is reasonable to regard them as random selections of such studies (i.e. they are in a sense exchangeable).We should ensure that it is reasonable to regard our stratifying agent as a random agent before using the derSimonian-Laird method; for example, this would not be the case if the stratifying agent were sex or, usually, years as more recent years should in general receive greater weight. Since the stratifying agent will often be hospitals among a group of hospitals or units within a hospital performing similar work, it will often be the case that it is reasonable to regard the stratifying agent as being random. Although the Newcombe and MH methods differ, if there is evidence of heterogeneity of the ORs among the strata, the

Newcombe analysis may be invalid and it may be preferable to concentrate on the differences within the individual strata if they are fixed. (The Newcombe analysis with weighting would not be suitable if the stratum agent were random.)

The homogeneity test seeks evidence of unusual variability among the strata; for example, the effect of the group variable may differ among them. In most cases the strata will not be ordered; for example, they could represent hospitals or the wards of a hospital that could be regarded as exchangeable and therefore as random agents. Some stratifying agents such as sex or years, where more recent ones should in general receive greater weight, would not be exchangeable and should therefore be treated as being fixed. Occasionally, they may be ordered, for example, patients' ages that have been placed in ordered groups, and a homogeneity trend test may then be required. Ordered strata would be fixed as they would not be exchangeable. When there is unusual variability among the strata and they may be regarded as being random, a random effects analysis may be appropriate. This is illustrated by Spiegelhalter[B,C,D] when analysing highly variable AE data among groups of hospitals using funnel plots. We illustrate these methods in Chapters 2 and 5. When there is heterogeneity among the strata and the stratifying agent is fixed, analysis of the data within the separate strata will usually be needed.

The MH CI method requires finding the fitted values in the first group cases or AEs given the MH OR; this is described by Breslow and Day. The homogeneity test requires large samples and typically has low power. An alternative although less powerful approach, on which the derSimonian and Laird method of moments random-effects analysis is often based, employs the Woolf log(OR). We illustrate the use of the log(OR) with the yearly data from the two hospitals. The analysis is further illustrated in the Outbreak Investigation in Chapter 7.

The Woolf and derSimonian-Laird procedures are described by Kirkwood and Sterne and Rosner. To illustrate the calculation of the former for the yearly SSI data from hospitals G and Q, the group 1 AEs in the first stratum (2002) were x1=7 (hospital Q) and in the second group there were x2=19 AEs (hospital G). The corresponding totals were N1=165 and N2=347. Therefore the four cell counts in the resulting stratum one 2×2 table are a=7, b=19, c=165-7=158 and d=347-19=328.

The marginal totals of the 2×2 table are m1=N1, m2=N2, m3=a+b=7+19=26 and m4=c+d=158+328=486. The data in the remaining strata are arranged in a similar manner. The odds ratio for stratum i (i=1..I) is $OR_i=(a_i \times d_i)/(b_i \times c_i)$. The approximate variance of $\log(OR_i)$ is $V_i=1/a_i+1/b_i+1/c_i+1/d_i$. To avoid division by zero when there are zero AEs, 0.5 is usually added to a_i, b_i, c_i and d_i. A weight $w_i=1/V_i$ is defined and the stratified log(OR) is then $\sum(w_i \times \log(OR_i))/\sum w_i$ with variance $V=1/\sum w_i$. The homogeneity of the stratum log(ORs) is assessed by $\chi^2=\sum(w_i \times (\log(OR_i)-\log(OR))^2)$ with DF=I-1. For a 1 DF trend test define x=1..I and $lor_i=\log(OR_i)$. Then $\chi^2{}_1=Lxy^2/Lxx$ where Lxy=a-b×d/e with $a=\sum(w_i \times x_i \times lor_i)$, $b=\sum(w_i \times lor_i)$, $d=\sum(x_i \times w_i)$ and $e=\sum w_i$; and Lxx=f-d/e, where $f=\sum(w_i \times x_i^2)$. The homogeneity test described by Breslow and Day mentioned above is more powerful but the Woolf log(OR) procedure is in common use. The former is available in the function epi.2by2() in the epiR library that we employ in Chapter 4.

If the homogeneity test suggests heterogeneity among the strata, the approximate additive method of moments random effects analysis described by derSimonian and Laird can then be employed provided it is reasonable to regard the levels of the stratifying agent as being exchangeable and therefore the stratum agent as being random. The within stratum variance V_i is widened to include the between stratum variance tau=max(0,(Q-I+1)/W) where Q is the homogeneity χ^2 and $W=\sum w_i-\sum w_i^2/\sum w_i$. Then $V^*{}_i=V_i+tau$, $\log(OR_{Adj})=\sum(w^*{}_i \times \log(OR_i))/\sum w^*{}_i$, and $V^*=1/\sum w^*{}_i$ where $w^*{}_i=1/V^*{}_i$.

1.9.1 Yearly data

The hospital G and hospital Q data for 2002 to 2006 are shown analysed by stratum (year), converted to individual observations and placed in a .csv file using s.d() (twosamplestratified.csv) and tabulated. We illustrate converting to individual observations here as they are used later in this chapter.

```
# converting data to individual observations
Group<-c(rep(1,5),rep(0,5))
Stratum<-gl(5,1,10)
AE<-c(19,22,40,19,14,7,5,15,10,15)
Total<-c(347,467,423,504,504,165,393,476,477,472)
NoAE<-Total-AE
dta<-data.frame(Group,Stratum,AE,NoAE)
I<-(length(dta[,1]))/2
dta1<-data.frame(Group,Stratum,AE,Total)
dta2<-cbind(dta1[1:I,],dta1[(I+1):(2*I),])
group<-c(rep(1,sum(dta2[,4])),rep(0,sum(dta2[,8])))
stratum<-0
for (i in 1:length(dta2[,1])){stratum<-c(stratum,rep(dta2[i,2],dta2[i,4]))}
stratum<-stratum[-1]
for (i in 1:length(dta2[,1])){stratum<-c(stratum,rep(dta2[i,6],dta2[i,8]))}
outcome<-0
for (i in 1:length(dta2[,1])){outcome<-c(outcome,rep(1,dta2[i,3]),
rep(0,dta2[i,4]-dta2[i,3]))}
outcome<-outcome[-1]
for (i in 1:length(dta2[,1])){outcome<-c(outcome,rep(1,dta2[i,7]),
rep(0,dta2[i,8]-dta2[i,7]))}
dta4<-data.frame(group,outcome,stratum)
```

Next, the Mantel-Haenszel analysis is illustrated. A logistic regression analysis shows that stratification by years is necessary. The stratified analysis is repeated using the Woolf method. Homogeneity and trend tests are illustrated and it can be seen that there is some evidence of heterogeneity among the strata but no evidence of a trend. Since years are here regarded as fixed, no random effects analysis is performed.

```
# logistic regression, is stratification necessary?
g1<-glm(outcome~group,family=binomial)
g<-glm(outcome~group+as.factor(stratum),family=binomial)
a<-anova(g1,g)
p<-1-pchisq(a$Deviance[2],a$Df[2])
a;cat("P-value=",round(p,4),"\n",sep="")
#the P-value is highly significant, stratification is required
```

```
Analysis of Deviance Table
Model 1: outcome ~ group
Model 2: outcome ~ group + as.factor(stratum)
  Resid. Df Resid. Dev Df Deviance
1      4226     1382.9
2      4222     1364.1  4   18.776
P-value=9e-04
```

```
#Mantel-Haenszel analysis
mant2<-mantelhaen.test(table(group,outcome,stratum),exact=T,alternative="t")
mant2$p.value
```

```
#
#[1] 2.731074e-05

# Woolf stratified analysis
dta3<-data.frame(dta[1:I,2:4],dta[(I+1):(2*I),3:4])
names(dta3)<-c("Stratum","Group1AE","Group1NoAE","Group2AE","Group2NoAE")
lor<-0;vlor<-0
for (i in 1:I){
x<-c(dta3$Group1AE[i],dta3$Group1NoAE[i])
y<-c(dta3$Group2AE[i],dta3$Group2NoAE[i])
ta<-data.frame(x,y)
ta1<-ta+.5
or1<-ta1[1,1]*ta1[2,2]/(ta1[1,2]*ta1[2,1])
lor[i]<-log(or1)
vlor[i]<-sum(1/ta1)
}
alor<-sum(lor/vlor)/sum(1/vlor)
valor<-1/sum(1/vlor)
OR<-exp(alor)
L95<-exp(alor-1.96*valor^.5)
U95<-exp(alor+1.96*valor^.5)
Z.score<-alor/valor^.5
P.value<-2*(1-pnorm(Z.score))
cat("OR=",round(OR,2),"\nL95=",round(L95,2),"\nU95=",round(U95,2),
"\nZ-score=",round(Z.score,2),"\nP-value=",round(P.value,5),"\n",sep="")
```

```
OR=1.89
L95=1.35
U95=2.65
Z-score=3.73
P-value=2e-04
```

```
# homogeneity and trend tests
ch<-sum((1/vlor)*(lor-alor)^2)
p<-1-pchisq(ch,(I-1))
H<-paste("\nHomogneity test chi-squared = ",round(ch,2))
P<-paste("\nP-value = ",round(p,3))
if (I>2){
xx<-1:I
w<-1/vlor
aa<-sum(w*xx*lor);bb<-sum(w*lor);dd<-sum(xx*w);ee<-sum(w);ff<-sum(w*xx^2)
Lxy<-aa-bb*dd/ee
Lxx<-ff-dd^2/ee
ch.sq<-Lxy^2/Lxx
pp<-1-pchisq(ch.sq,1)
TR<-paste("\nTrend test chi-squared = ",round(ch.sq,2))
PT<-paste("\nP-value = ",round(pp,3))
}
cat("\n")
cat(H,P,"\n")
if (I>2){cat(TR,PT,"\n")}
```

```
Homogneity test chi-squared =  9.73
P-value =  0.045
Trend test chi-squared =  1.73
P-value =  0.189
```

It may be of interest to compare logistic regression including group by stratum interaction in a fixed effects analysis. There is evidence of group by stratum interaction. For the standard logistic regression the OR is exp(0.7043)=2.02 and the P-value is 0.00004, similar to the MH values of OR=2.02 and P-value=0.00003 (the group effect is difficult to assess when an interaction term is included). Thus the most appropriate approach may be to analyse these data within the strata separately.

```
# logistic regression and interaction
G<-glm(outcome~group+as.factor(stratum),family=binomial)
#adding interaction term
G1<-glm(outcome~group*as.factor(stratum),family=binomial)
A<-anova(G,G1)
P<-1-pchisq(A$Deviance[2],A$Df[2])
A; cat("P-value=",round(P,4),"\n",sep="")
```

```
Analysis of Deviance Table
Model 1: outcome ~ group + as.factor(stratum)
Model 2: outcome ~ group * as.factor(stratum)
  Resid. Df Resid. Dev Df Deviance
1      4222     1364.1
2      4218     1354.1  4   10.033
P = 0.0399
```

```
# examining the individual strata
Group<-c(rep(1,5),rep(0,5))
Stratum<-gl(5,1,10)
AE<-c(19,22,40,19,14,7,5,15,10,15)
Total<-c(347,467,423,504,504,165,393,476,477,472)
NoAE<-Total-AE
dta<-data.frame(Group,Stratum,AE,NoAE)
I<-(length(dta[,1]))/2
dta1<-data.frame(Group,Stratum,AE,Total)
dta2<-cbind(dta1[1:I,],dta1[(I+1):(2*I),])
#individual strata using twoproportions
for (i in 1:I){cat("\n",i,"\n",sep="");twoproportions(dta2[i,c(3,4,7,8)])}
```

```
1. First proportion 0.055, Second proportion 0.042
Difference 0.012, L95 -0.032, U95 0.049
Fisher Exact P-value = 0.669, LR = 1/1
Fisher Mid-P-value = 0.595, LR = 1/1
Ratio 1.29, L95 0.57, U95 2.95
```

```
2. First proportion 0.047, Second proportion 0.013
Difference 0.034, L95 0.012, U95 0.058
Fisher Exact P-value = 0.005, LR = 1/51
Fisher Mid-P-value = 0.004, LR = 1/63
Ratio 3.7, L95 1.47, U95 9.39.
```

```
3. First proportion 0.095, Second proportion 0.032
Difference 0.063, L95 0.032, U95 0.096
Fisher Exact P-value = 0, LR = 1/2404
Fisher Mid-P-value = 0, LR = 1/3270
Ratio 3, L95 1.7, U95 5.32.
```

```
4. First proportion 0.038, Second proportion 0.021
Difference 0.017, L95 -0.004, U95 0.038
Fisher Exact P-value = 0.135, LR = 1/3
Fisher Mid-P-value = 0.111, LR = 1/4
Ratio 1.8, L95 0.86, U955 3.77.
```

```
5.First proportion 0.028, Second proportion 0.032
Difference -0.004, L95 -0.026, U95 0.017
Fisher Exact P-value = 0.851, LR = 1/1
Fisher Mid-P-value = 0.781, LR = 1/1
Ratio 0.87, L95 0.43, U95 1.77.
```

Lastly, we could employ a bootstrap analysis of the standardised risk difference and risk ratio. This also appears to give intervals that are too narrow and this is not corrected using bias and acceleration (BCa) adjustments. The outlier SSI rate for hospital G in 2004 has resulted in significant heterogeneity. The practice of making comparisons of these and similar data should be approached with great caution. We have repeatedly urged caution when comparisons are made between hospitals (see section 1.4). This is another example of the potential difficulties entailed in such comparisons.

1.9.2 Data stratified by hospital

We now return to the earlier data in the files Before.csv and After.csv. These may be stratified by hospital and compared to see if there was overall improvement in the later period. In this analysis we employ the random effects method.

```
g.d() # Before.csv

Before<-datain
Group<-rep(1,length(Before[,1]))
Before<-data.frame(Before,Group)

g.d() # After.csv

After<-datain
Group<-rep(2,length(After[,1]))
After<-data.frame(After,Group)

# amalgamating
All<-rbind(Before,After)
stratum<-All[,1]
group<-All[,4]
outcome<-All[,3]

# logistic regression, is stratification necessary?
# is interaction present?
g1<-glm(outcome~group,family=binomial)
g<-glm(outcome~group*as.factor(stratum),family=binomial)
g2<-glm(outcome~group+as.factor(stratum),family=binomial)
a<-anova(g1,g2) # stratification
p<-1-pchisq(a$Deviance[2],a$Df[2])
b<-anova(g2,g) # interaction
```

```
p1<-1-pchisq(b$Deviance[2],b$Df[2])
a;cat("P-value=",round(p,4),"\n",sep="")
#the P-value is highly significant, stratification is required
```

```
Model 1: outcome ~ group
Model 2: outcome ~ group + as.factor(stratum)
  Resid. Df Resid. Dev Df Deviance
1     10920     4266.9
2     10911     4196.7  9   70.262
P-value=0
```

```
b;cat("P-value=",round(p1,4),"\n",sep="")
# interaction P-value is of borderline significance
```

```
Model 1: outcome ~ group + as.factor(stratum)
Model 2: outcome ~ group * as.factor(stratum)
  Resid. Df Resid. Dev Df Deviance
1     10911     4196.7
2     10902     4181.2  9   15.463
P-value=0.079
```

```
#Mantel-Haenszel analysis
mant2<-
mantelhaen.test(table(group,outcome,stratum),exact=T,alternative="t")
mant2$estimate
#
#common odds ratio
#       0.8874502
# reciprocal
1/mant2$estimate
#
#common odds ratio
#        1.126824
mant2$p.value
#
#[1] 0.2133

# Woolf stratified analysis
AE1<-tapply(Before$SSI,Before$Hospital,sum)
AE2<-tapply(After$SSI,After$Hospital,sum)
To1<-tapply(Before$SSI,Before$Hospital,length)
To2<-tapply(After$SSI,After$Hospital,length)
No1<-To1-AE1
No2<-To2-AE2
I<-length(AE1)
AEA<-c(AE1,AE2)
AEA<-as.numeric(AEA)
TotalA<-c(To1,To2)
TotalA<-as.numeric(TotalA)
NoAEA<-TotalA-AEA
GroupA<-c(rep(1,I),rep(0,I))
StratumA<-gl(I,1,2*I)
dta<-data.frame(GroupA,StratumA,AEA,NoAEA)
dta3<-data.frame(dta[1:I,2:4],dta[(I+1):(2*I),3:4])
```

```
names(dta3)<-c("Stratum","Group1AE","Group1NoAE","Group2AE","Group2NoAE")
lor<-0;vlor<-0
for (i in 1:I){
x<-c(dta3$Group1AE[i],dta3$Group1NoAE[i])
y<-c(dta3$Group2AE[i],dta3$Group2NoAE[i])
ta<-data.frame(x,y)
ta1<-ta+.5
or1<-ta1[1,1]*ta1[2,2]/(ta1[1,2]*ta1[2,1])
lor[i]<-log(or1)
vlor[i]<-sum(1/ta1)
}
w<-1/vlor
alor<-sum(lor*w)/sum(w)
valor<-1/sum(w)
OR<-exp(alor)
L95<-exp(alor-1.96*valor^.5)
U95<-exp(alor+1.96*valor^.5)
Z.score<-alor/valor^.5
P.value<-2*(1-pnorm(Z.score))
cat("OR=",round(OR,2),"\nL95=",round(L95,2),"\nU95=",round(U95,2),"
\nZ-score=",round(Z.score,2),"\nP-value=",round(P.value,5),"\n",sep="")
```

```
OR=1.1
L95=0.92
U95=1.32
Z-score=1.03
P-value=0.3026
```

```
# homogeneity test
# trend test not indicated for exchangeable strata
ch<-sum(w*(lor-alor)^2)
p<-1-pchisq(ch,(I-1))
H<-paste("\nHomogeneity test chi-squared = ",round(ch,2))
P<-paste("\nP-value = ",round(p,3))
cat("\n")
cat(H,P,"\n")
```

```
Homogeneity test chi-squared =  14.98
P-value =  0.091
```

```
# derSimonian-Laird random Effects
W<-sum(w)-(sum(w^2)/sum(w))
Q<-max(0,ch-I+1)
Q/W # random effects variance
#[1] 0.05954
qw<-Q/W
rvlor<-vlor+qw
w1<-1/rvlor
rlor<-sum(w1*lor)/sum(w1)
vrvlor<-1/sum(w1)
L<-exp(rlor-1.96*vrvlor^.5)
U<-exp(rlor+1.96*vrvlor^.5)
Z<-rlor/vrvlor^.5
P<-2*(1-pnorm(rlor/vrvlor^.5))
```

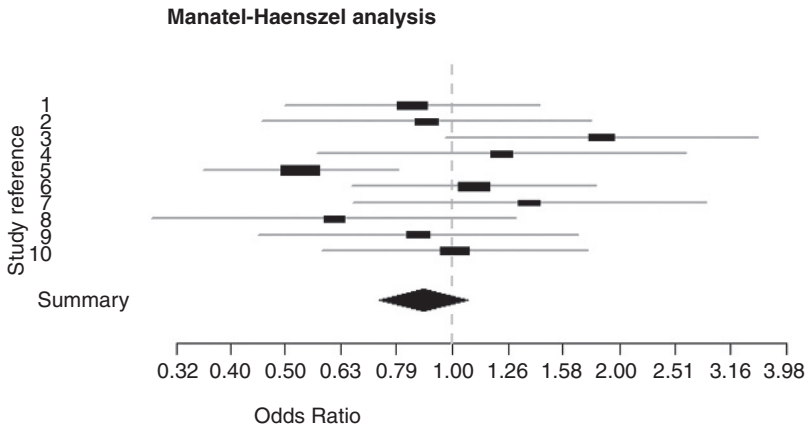

Figure 1.1 Stratified data, Mantel-Haenszel analysis.

```
cat("Adjusted OR=",round(exp(rlor),2),"\nL95=",round(L,2),"\nU95=",
round(U,2),"\nZ-score=",round(Z,2),"\nP-value=",round(P,5),"\n",sep="")
```

```
Adjusted OR=1.06
L95=0.83
U95=1.35
Z-score=0.46
P-value=0.6466
```

A derSimonian and Laird random-effects analysis was performed. This should be suitable as the hospitals could be considered as random agents. It may be convenient to use metaDSL() in the rmeta library to perform this analysis. This has the advantage of providing a plot of the confidence intervals, called a forest plot (Figure 1.1 and Figure 1.2). There is no evidence that, overall, the Before and After proportions differ.

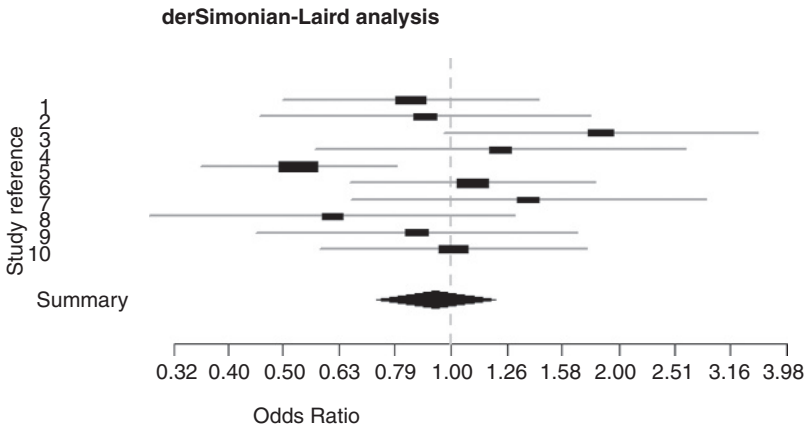

Figure 1.2 Stratified data, derSimonian-Laird analysis.

```
#using the rmeta library
library(rmeta)
SSI1<-AE1 #Before
Proc1<-To1
SSI2<-AE2 #After
Proc2<-To2
ORD<-meta.DSL(Proc2,Proc1,SSI2,SSI1,names=1:10)
ORM<-meta.MH(Proc2,Proc1,SSI2,SSI1,names=1:10)
summary(ORM)
```

```
Fixed effects ( Mantel-Haenszel ) meta-analysis
Call: meta.MH(ntrt = Proc2, nctrl = Proc1, ptrt = SSI2, pctrl = SSI1,
    names = 1:10)
------------------------------------
     OR (lower 95% upper)
1  0.85    0.50      1.44
2  0.90    0.45      1.78
3  1.85    0.97      3.53
4  1.22    0.57      2.63
5  0.53    0.36      0.80
6  1.09    0.66      1.81
7  1.38    0.66      2.86
8  0.61    0.29      1.30
9  0.87    0.45      1.68
10 1.01    0.58      1.75
------------------------------------
Mantel-Haenszel OR =0.89 95% CI ( 0.74,1.07 )
Test for heterogeneity: X^2( 9 ) = 15 ( p-value 0.0908 )
```

```
ORD
```

```
Random effects ( DerSimonian-Laird ) meta-analysis
SummaryOR= 0.94  95% CI ( 0.73,1.2 )
Test for heterogeneity: X^2( 9 ) = 14.98 ( p-value 0.0914 )
Estimated random effects variance: 0.06
```

```
plot(ORM,main="Mantel-Haenszel Analysis")

plot(ORD,main="derSimonian-Laird Analysis")
```

1.10 Stratified rates and overdispersion

One of the methods described by Spiegelhalter[B,C,D] for displaying overdispersed hospital AE data in funnel plots involves modifying prediction limits using the derSimonian-Laird random effects variance and the null variance of the AE rate. Since there is no difference between the Before and After data but considerable differences among the hospitals, we illustrate the methods of Spiegelhalter[B,C,D] and Laney (Mohammed and Laney[A]) using the combined data in the data.frame All. This issue is discussed further in Chapters 2 and 5.

First we obtain the rates for the hospitals, the overall mean rate and the null and non-null variances for the hospital rates. Z-scores and the derSinomian-Laird random-effects variance are then obtained.

$Z_i=(H_i\text{-}M)/V0_i^{.5}$, where H_i is the rate for hospital i, M is the mean rate and $V0_i$ is the null variance for that hospital ($V0_i=M\times(1\text{-}M)/N_i$, where N_i is the hospital i denominator). The random effects variance is (Q-(L-1))/W, where $Q=\sum(Z_i^2)$, L is the number of hospitals and $W=\sum w_i-\sum w_i^2/\sum w_i$, where w_i is the reciprocal of the non-null variance for hospital i ($V_i=H_i\times(1\text{-}H_i)/N_i$). The revised Z-score for hospital i is then $Z_i=(H_i\text{-}M)/(V0_i+Q)^{.5}$.

```
M<-mean(All$SSI)
S<-tapply(All$SSI,All$Hospital,sum) # observed counts
N<-tapply(All$SSI,All$Hospital,length)
E<-round(M*N,2) # expected counts
L<-length(N)
H<-S/N # observed rates
VH<-H*(1-H)/N
V0<-M*(1-M)/N
ZH<-(H-M)/VH^.5
Z0<-(H-M)/V0^.5
IZ<-sum(Z0^2)
w<-1/VH
W<-sum(w)-sum(w^2)/sum(w)
Q<-(IZ-(L-1))/W
ZHDSL<-(H-M)/(V0+Q)^.5
data.frame(S,N,E,H,Z0,ZHDSL)
```

	S	N	E	H	Z0	ZHDSL
A	80	865	42	0.092	5.928	2.309
B	38	1069	52	0.036	−2.035	−0.724
E	39	997	49	0.039	−1.443	−0.529
F	30	550	27	0.055	0.604	0.283
G	128	2598	127	0.049	0.067	0.016
I	70	1079	53	0.065	2.419	0.857
J	30	714	35	0.042	−0.863	−0.364
K	30	477	23	0.063	1.407	0.697
N	38	589	29	0.065	1.747	0.797
Q	52	1984	97	0.026	−4.700	−1.264

As well as the additive correction (Z0 is the unadjusted Z-score and ZHDSL is the Z-score with the additive adjustment described above), Spiegelhalter describes a multiplicative adjustment that employs the square root of the mean of the squared Z-scores (the Z0s) and Laney advocates using the standard deviation (sd) of the Z-scores for the adjustment (ZS and ZL in the box). For funnel plot control limits using, for example, Laney's correction $U,L=M\pm Z\times sd(Z0)\times\sqrt{V0_i}$, U and L are upper and lower control limits, and Z is for example, 3 for a 3 sd control limit.

```
(mean(Z0^2))^.5 # Spiegelhalter's multiplicative correction

#[1] 2.7

sd(Z0) # Laney's correction

#[1] 2.9

ZS<-Z0/(mean(Z0^2))^.5 # Z-score with Spiegelhalter's
# multiplicative correction
```

```
ZL<-Z0/sd(Z0) # Z-score with Laney's correction
data.frame(Z0,ZHDSL,ZS,ZL)
```

	Z0	ZHDSL	ZS	ZL
A	5.928	2.309	2.158	2.061
B	−2.035	−0.724	−0.741	−0.708
E	−1.443	−0.529	−0.525	−0.502
F	0.604	0.283	0.220	0.210
G	0.067	0.016	0.024	0.023
I	2.419	0.857	0.880	0.841
J	−0.863	−0.364	−0.314	−0.300
K	1.407	0.697	0.512	0.489
N	1.747	0.797	0.636	0.607
Q	−4.700	−1.264	−1.711	−1.634

These corrections can be excessive and Spiegelhalter recommends winsorising, for example, 10% of the outlier Z-scores. Clearly, this adjustment requires considerable expertise and will be beyond the scope of most hospital scientists. There are 10 hospitals in the example and to illustrate we winsorise the largest and smallest Z-scores (20% winsorising).

```
# Z-scores modified using 20% winsorising
Z1<-Z0
Z1[1]<-Z0[6]
Z1[10]<-Z0[2]
(mean(Z1^2))^.5 # Spiegelhalter's multiplicative correction

#[1] 1.7

sd(Z1) # Laney's correction

#[1] 1.8

ZHS1<-Z0/(mean(Z1^2))^.5
ZHL1<-Z0/sd(Z1)
data.frame(Z0,ZHS1,ZHL1)
```

	Z0	ZHS1	ZHL1
A	5.928	3.53	3.380
B	−2.035	−1.21	−1.160
E	−1.443	−0.86	−0.823
F	0.604	0.36	0.345
G	0.067	0.04	0.038
I	2.419	1.44	1.379
J	−0.863	−0.51	−0.492
K	1.407	0.84	0.802
N	1.747	1.04	0.996
Q	−4.700	−2.80	−2.679

These corrections are now more like those from a random effects logistic regression.

```
library(hglm)
h<-hglm(fixed=SSI~1,random=~1|Hospital,family=binomial(link=logit),
fix.disp=1,data=All)
```

```
ZH<-h$ranef/h$SeRe
H<-data.frame(Z0,ZHL1,ZH)
H # Z0 (Basic Z-scores), ZHL1 (Laney's correction with
winsorising), ZH (empirical Bayes random effects)
```

	Z0	ZHL1	ZH
A	5.92783301	3.37951606	3.6637585
B	−2.03541984	−1.16041292	−1.6827657
E	−1.44339214	−0.82289209	−1.2287119
F	0.60432864	0.34453372	0.3373963
G	0.06729646	0.03836637	−0.1533509
I	2.41850576	1.37881399	1.4696057
J	−0.86251894	−0.49173056	−0.7909855
K	1.40749594	0.80242731	0.9370932
N	1.74653449	0.99571654	1.1579940
Q	−4.69992412	−2.67947309	−3.5371427

Spiegelhalter[B,C] notes that with winsorising variances are underestimated but a correction factor does not appear to be used.