Overdispersion in count data.

P.M.E.Altham, Statistical Laboratory,
University of Cambridge
Centre for the Mathematical Sciences,
Wilberforce Road, Cambridge CB3 OWB,
UK, Fax no 01223-337956.

P.Altham@statslab.cam.ac.uk

Seminar given at MRC Biostatistics Unit,
November 14, 2000.

Introduction

Statistical analysis for discrete data, particularly for probability models such as the binomial, Poisson and multinomial, is by now very well understood, with a wealth of suitable software.

Such software typically exploits the connexion between these models and *generalized linear modelling* (glm), so that for example, it is very easy to do log-linear regression for the Poisson and the multinomial, and logistic or probit regression for the binomial.

It can happen that the standard glm software is not completely appropriate, since *over-dispersion* is present, relative to the standard distributions such as the Poisson or the binomial.
What exactly do we mean by over-dispersion? One way of answering this question is to note that both the binomial and the Poisson make a very strong assumption about the structure of the *variance*. For example, if Y has a Poisson distribution, then if we assume that $E(Y) = \mu$, then $var(Y)$ is forced to be μ also,
although in practice we may suspect that

$$var(Y) > \mu.$$

It is possible to take account of this over-dispersion by modelling Y as negative-binomial, which corresponds to assuming that the distribution of Y conditional on the parameter μ is Poisson, but μ itself is a random variable with a gamma distribution. Suitable software is available via the library(MASS) suite of functions compiled by Venables and Ripley. See, for example, the

```
glm.nb()
```

function, which fits the negative-binomial distribution.

If we assume that Y is binomial, with parameters n, p, then
$E(Y) = np$ and $var(Y) = np(1-p)$.
If in fact we have over-dispersion relative to the binomial, then we will find that

$$var(Y) > np(1-p).$$

Failure to take account of this over-dispersion, for example in fitting a model such as

$$log(p/(1-p)) = \alpha + \beta x$$

(where the covariate x is the *dose*) will mean that our estimates of β will be less precise than the binomial-based formula gives us. Thus for example, with no correction for the extra-binomial variation, we will be quoting confidence intervals for β that are
$\boxed{\text{too narrow}}$.

One way of coping with this problem is to use a probability model which is more general than the binomial, and one such model, which we can easily fit in S-Plus, is the beta-binomial. Today I will discuss beta-binomial modelling, firstly in relation to the interesting data set of Spiegelhalter and Marshall (1998) on success rates of 52 *in vitro* fertilisation clinics in the UK. E.C.Marshall and D.J.Spiegelhalter present and analyse the data we shall discuss today. To quote from E.C.Marshall's unpublished PhD thesis, which also includes these data, 'In July 1996 the Human Fertilisation and Embryology Authority reported on 25730 *in vitro* fertilisation treatments carried out in 52 clinics over the period from 1 April 1994 to 31 March 1995. An overall adjusted live birth rate of 14.5 % was found.'

The full dataset is given in Marshall's thesis, and is not reproduced here. If we denote by r the number of live births, and let n be the number of fertilisations, then the figures for r, n and r/n range from the least successful of
Withington $(r, n, r/n = 7, 147, 0.047)$,
Manchester Fertility $(r, n, r/n = 41, 506, .081)$,
Fazakerley $(r, n, r/n = 20, 240, .083)$,
......
to
St James's $(r, n, r/n = 121, 537, .225)$,
Birmingham Women's $(r, n, r/n = 60, 267, .225)$,
and finally, the most successful,
NURTURE, Nottingham $(r, n, r/n = 204, 861, .237)$.

The full dataset shows that there is not only substantial variation in r/n, the proportion of successful attempts, but also in n, the total number of attempts. First we will fit the binomial with constant probability p to these data, namely

$$r_i \sim \text{independent} \quad Bi(n_i, p),\ 1 \le i \le 52.$$

This is easily achieved within S-Plus by

```
data _ read.table("hospitals.data", header=T)
attach(data)
first.glm _ glm(r/n ~ 1, binomial, weights=n)
summary(first.glm)
```

which shows a deviance of 390.76, with $df = 51$. Refer this to the corresponding χ^2, to see that we have substantial overdispersion with respect to the model of constant binomial parameter p: the model of constant probability p of success fails to fit.

We will compute the binomial residuals, for comparison later with the betabinomial residuals.

```
p _ first.glm$fitted.values ; q _ 1-p
res _ (r-n*p)/sqrt(n*p*q)
sum(res^2) # as a check
chisq.test(cbind(r,n-r)) # as another check
# sqrt(n) * resid(first.glm)
would give us the deviance residuals instead
```

Our next step is to allow one extra parameter: we assume that

$$r_i|p_i \sim Bi(n_i, p_i)$$

and assume further that p_i has the beta distribution, parameters θ, ϕ.

This has the consequence that each r_i then has a beta-binomial distribution, parameters n_i, θ, ϕ.

Again assume that all the r_i's are independent.

We pause to derive the frequency function for the beta-binomial, and also its mean and variance. Now

$$f(r|p) = \binom{n}{r} p^r (1-p)^{n-r}, \ for \ r = 0, \cdots, n$$

where p has density $g(p)$ say, where

$$g(p) = \frac{\Gamma(\theta + \phi)}{\Gamma(\theta)\Gamma(\phi)} \ \ p^{\theta-1}(1-p)^{\phi-1}, \text{ for } 0 \leq p \leq 1.$$

Thus, integrating with respect to p, we find that the frequency function for r is

$$f(r) = \int f(r|p)g(p)dp$$

$$= \binom{n}{r} \frac{\Gamma(\theta+\phi)}{\Gamma(\theta)\Gamma(\phi)} \frac{\Gamma(\theta+r)\Gamma(\phi+n-r)}{\Gamma(\theta+\phi+n)}.$$

It is easy to see that

$E(r) = E(E(r|p)) = nE(p) = n\ \theta/(\theta{+}\phi) = np'$,

say. Similarly

$$var(r) = E(var(r|p)) + var(E(r|p)),$$

or, alternatively, if we denote $X_1, \cdots, X_n$ as the responses (1 or 0), of the $1st, 2nd, \cdots, nth$ member of the set of n individuals which make up the response for a given hospital, we see that

$$var(r) = var(X_1 + \cdots + X_n)$$

$$= n\ var(X_1) + n(n-1)cov(X_1, X_2),$$

giving

$$var(r) = np'q' + n(n-1)\rho p'q'$$

where $p' = \theta/(\theta + \phi)$ as above
and $\rho = 1/(\theta + \phi + 1) = corr(X_1, X_2)$.

In the S-Plus commands below, we compute

$$-\Sigma_i log f(r_i|\theta,\phi)$$

as MINUS the loglikelihood function, and then minimise it to find the maximum likelihood estimates of θ, ϕ. ‘General optimization and maximum likelihood estimation’ is given as Chapter 8 in Venables and Ripley (1999).

```
lbetabin _ function(p)
{
th <- p[1]
phi <- p[2]
sum ( - lgamma(th+r)-lgamma(phi+n-r)
+ lgamma(th + phi + n) +
lgamma(th)+lgamma(phi)-lgamma(th+phi))
}
p _ c(.15,.85)
```

These are our initial estimates of theta, phi, taken from the binomial fit, and setting theta + phi =1. One way to proceed is as follows

```
fit.first _ nlmin(lbetabin,p,print.level=1)
# this does not quite converge, and
fit.first$converged
```

shows that we have not yet reached convergence, but

```
fit.first$x    # shows that we have
#  estimates theta =10.76 , phi=63.25.
```

So we use these as starting values, thus

```
p _  fit.first$x
fit.next _  nlmin(lbetabin,p,print.level=1)
# now quickly converges, giving
# the following estimates
fit.next$x
 10.92 63.23  # for theta, phi
 # Now we try a different minimisation function
p _ c(.15,.85) # same starting values
fit.betabin _ nlminb(start = p,objective = lbetabin
 # which gives
fit.betabin #whose contents include  the following
$parameters:
[1] 10.92643 63.25428
$objective:
[1] 10184.99
$message:
[1] "RELATIVE FUNCTION CONVERGENCE"
```

(We edit the output to save space here.)

```
library(MASS)
```

```
vcov.nlminb(fit.betabin)
```

gives us the approximate covariance matrix for these parameter estimates, as
6.36 36.71
36.71 222.26.

It is interesting that we find

$$\widehat{\theta} = 10.93(se = 2.52), \widehat{\phi} = 63.25(se = 14.91)$$

which corresponds to a beta-density for p which is quite sharply peaked. The plot is given in Figure 1, and is obtained as follows:

```
th _ 10.93; phi _ 63.25
p _ (1:100)/100
f _ dbeta(p,th,phi)
plot(p,f,type="l")
```

We can use the parameter estimates to compute the correct estimated variance for r_i, and hence compute a χ^2 goodness of fit statistic for the model.

```
 th _ 10.93; phi _ 63.25; pi _ th/(th + phi)
betabin.resid  _
(r-n*pi)/sqrt(n*pi*(1-pi)*(1+(n-1)/(th + phi+1)))
plot(res,betabin.resid)
```

```
betabin.chi2 _ sum(betabin.resid^2)
```

This finds the χ^2 statistic as 50.41, with 50 df, showing that the inclusion of just 1 extra parameter gives a model that satisfactorily accounts for the ‘over-dispersion’ relative to the ordinary binomial.

Here are the ordered binomial residuals.

```
round(sort(res),2)
```

This shows us 'best' and 'worst' on crude 1-parameter binomial model: the residuals are from
King'sColl ManchesterFS Ninewells Hull Withington Cromwell Walsgrave
-6.85 -4.36 -4.16 -3.63 -3.48 -3.4 -3.11
-2.9 -2.65 -2.51 -2.15 -2.1 -2.04 -1.8 -1.36
-1.21 -1.14 -1.08 -0.98 -0.97 -0.82 -0.66
-0.66 -0.51 -0.42 -0.41 -0.27 -0.16 -0.09
0.01 0.18 0.29 0.41 0.47 0.67 0.81
0.93 1.19 1.2 1.22 1.65 1.67
1.76 1.77 2.09 3.41 4.0 4.27
and finally
RMHBelfast StJames's Lister NURTURE
4.75 4.87 6.59 7.12.
Here are the ordered beta-binomial residuals,

which can also be compared to the standard normal

```
round(sort(betabin.resid),2)
```

and thus the betabinomial residuals are

```
Withington ManchesterFS King'sColl Ninewells
Hull Fazakerley Cromwell
-1.99 -1.47 -1.46 -1.45 -1.41 -1.37 -1.26
-1.2 -1.11 -1.09 -0.79 -0.74 -0.72 -0.65
-0.61 -0.53 -0.47 -0.47 -0.38 -0.35 -0.32
-0.25 -0.23 -0.22 -0.16 -0.11 -0.09 -0.02
0.03 0.07 0.12 0.15 0.26 0.28 0.30
0.31 0.41 0.5 0.6 0.69 0.84 0.86
0.93 1.15 1.16 1.28 1.50 1.67
```

and finally

```
RMHBelfast Lister StJames's NURTURE
1.73 1.74 1.79 2.10.
```

Discussion and Conclusions

- Allowing for over-dispersion via the beta-binomial model shows us that in terms of this model, there is only one hospital with a large and negative residual (Withington) and only four with large and positive residuals. It seems more sensible to compare the 52 institutions via their beta-binomial residuals rather than their binomial residuals, since we know that the model of a constant binomial parameter p is such a poor fit.
- Note that the estimates of θ, ϕ obtained above give a very small estimate for ρ, the correlation between individual responses at the same hospital, namely $\widehat{\rho} = 1/(1 + 10.93 + 63.25) = 0.013$. But this very small positive correlation 'magnifies' the variance of r/n relative to that of the true binomial because of the large values of n that are involved: half of these are between 210 and 641.

• The sample correlation matrix for $\widehat{\theta}, \widehat{\phi}$ suggests that from the point of view of the function-minimisation problem, we could find a much 'better' parametrisation, in which the two parameters are closer to being orthogonal. It is worth experimenting with the parametrisation $\pi = \theta/(\theta + \phi)$, $\psi = \theta + \phi$.

• One of the objectives of Marshall and Spiegelhalter in looking at this table was to produce a 'reliable' ranking of the hospitals, since a ranking based on the crude success rate can be quite misleading. How do we address this question with the benefit of our beta-binomial model?

Of course, whether we use a binomial or a beta-binomial distribution, our statistical 'comparison' of clinics will be extremely simplistic, since it will fail to take account of what must be relevant background information, such as the ages of the women trying for conceptions, and so forth. However, even a simplistic analysis make be useful, in that it will prompt us to ask, say ‘What is it about NURTURE that makes it so much more successful than the others?’ and to call for more data than just the bare figures (r, n) given here.

- The betabinomial and other models for binomial overdispersion are discussed in the paper by Lindsey and Altham (1998), which includes an analysis of sex-ratio data.
- The newest version of S-Plus, ie Splus5, gives just slightly different parameter estimates etc from the ones quoted above; these very slight differences do not affect the argument of the paper.

References

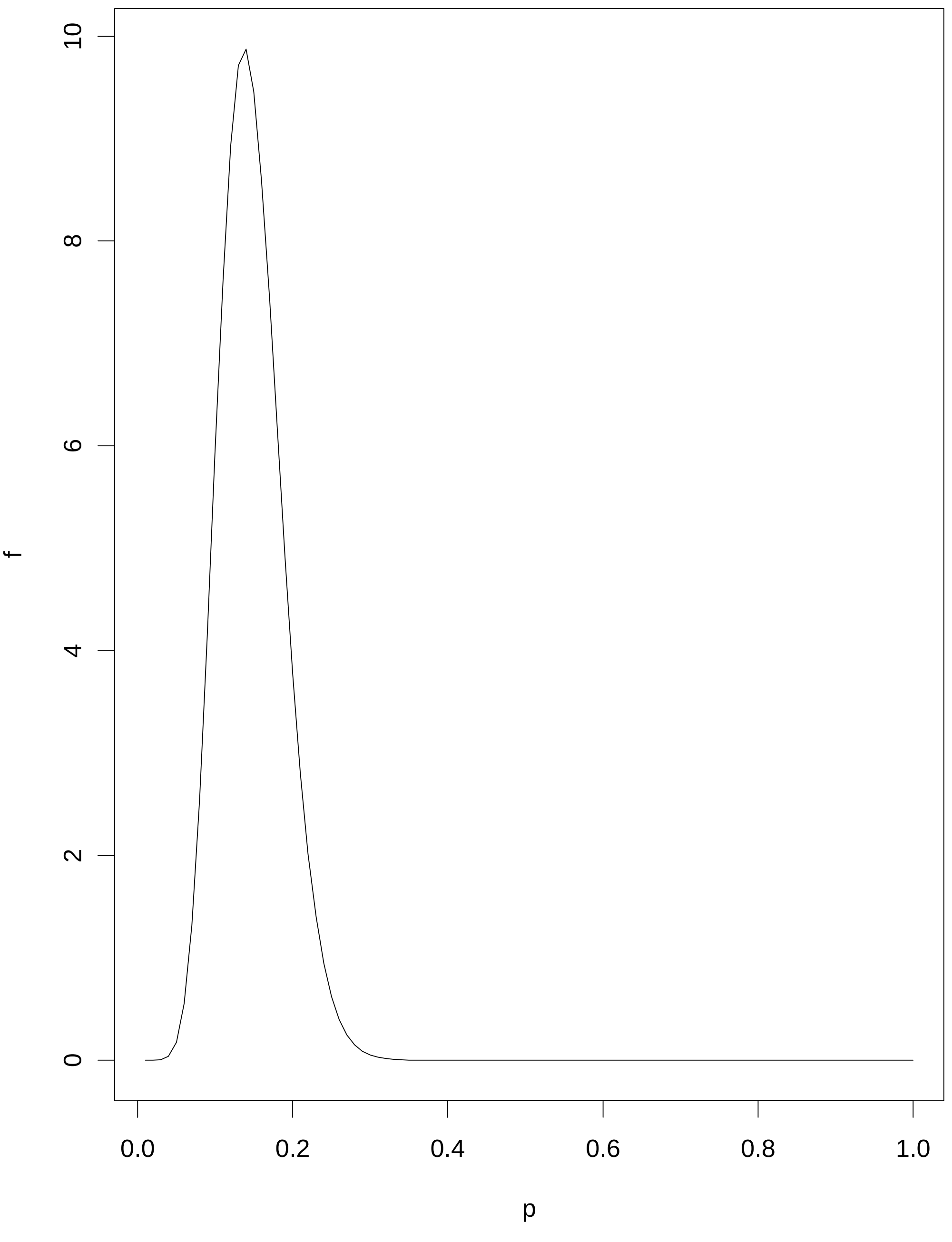
Lindsey, JK and Altham, PME: Analysis of the human sex ratio using overdispersion models. App Statist 1998;47:149–157.

Marshall, EC: Statistical methods for Institutional Comparisons. PhD thesis, University of Cambridge, 1999.

Marshall, EC and Spiegelhalter, DJ: Reliability of league tables of *in vitro* fertilisation clinics: retrospective analysis of live birth rates. British Medical Journal 1998;316:1701–4.

Venables WN and Ripley BD: Modern Applied Statistics with S-Plus. New York, Springer, 1999.

Figure 1



Another data set: comparison of educational institutions Here is the data, kindly sent to me by Peter Tompkins in September 1999. Peter used this to construct the 'Tompkins table' for Cambridge colleges in 1999. The final column, sc, is the $score$ for that college, computed as

$$(5 * n1 + 3 * n2 + 2.5 * n3 + 2 * n4 + 1 * n5)/n,$$

and is used to present in 'The Independent' a rank order of the colleges.

```
 Data for Tompkins table.
                 First X2.1 X2nd X2.2 Thd Alw sc
        Christ's   106  157    1   73   7 2 3.34
       Churchill    81  168    3  109  22 4 2.99
           Clare    77  168    4  100  12 2 3.06
  Corpus_Christi    50   97    2   47   7 0 3.19
         Downing    61  189    2   90   7 1 3.04
        Emmanuel    99  193    1   88  12 0 3.22
     Fitzwilliam    61  200    0   91  16 1 2.99
          Girton    61  222    0  124  21 0 2.90
  Gonville&Caius   105  189    3  105  16 0 3.17
           Jesus    72  212    0   92  12 1 3.06
          King's    64  177    0   78  11 1 3.08
       Magdalene    37  130    1   92  10 0 2.86
        New_Hall    45  135    2   96  12 0 2.89
         Newnham    41  194    1  104  13 0 2.86
        Pembroke    69  154    1   92  13 1 3.05
      Peterhouse    42   97    2   41   9 3 3.08
         Queens'   114  197    3   90  16 1 3.24
        Robinson    64  170    0   73   9 0 3.12
  St.Catherine's    81  191    1   81  13 1 3.14
```

St.John's	118	226	1	111	11	2	3.21
Selwyn	62	153	1	66	9	2	3.11
Sidney_Sussex	50	148	2	60	10	1	3.06
Trinity	168	254	3	114	17	0	3.34
Trinity_Hall	53	151	4	65	17	2	3.00
Lucy_Cavendish	2	23	0	17	4	0	2.54
Wolfson	6	19	0	19	5	0	2.65
Hughes_Hall	9	0	5	3	0	0	3.74
StEdmund's	6	37	0	20	3	0	2.79

How do the colleges compare if we model $r = n1 + n2$ as Binomial (n, p), and allow p to vary as a Beta, with parameters θ, ϕ, as we did for the hospitals data above?

We find

$$\Sigma(r)/\Sigma(n) = .7059,$$

and

```
first.glm _ glm(r/n ~ 1, binomial, weights=n)
```

has deviance 80.74 on 27 df, showing that a model of $r \sim B(n,p)$, (constant p) fails to fit, as we might expect. What is the best fitting beta-binomial?
It turns out to correspond to $p \sim Beta(\theta, \phi)$, with

$$\theta = 118.59, \phi = 50.68,$$

corresponding to a between-unit correlation of $\rho = .0059$ (ie correlation between students at the same college).
We compute the residuals under this beta-binomial model:

re is the binomial residual, be is the betabinomial residual, n is the size of the cohort.

```
                 sc   n    re    be
      Christ's 3.34 346  2.21  1.39
     Churchill 2.99 387 -2.70 -1.36
         Clare 3.06 363 -1.29 -0.60
Corpus_Christi 3.19 203  0.57  0.50
       Downing 3.04 350  0.35  0.32
      Emmanuel 3.22 393  1.62  1.01
   Fitzwilliam 2.99 369  0.06  0.16
        Girton 2.90 428 -2.03 -0.95
Gonville&Caius 3.17 418 -0.11  0.07
         Jesus 3.06 389  1.05  0.70
        King's 3.08 331  0.89  0.64
     Magdalene 2.86 270 -3.15 -1.83
      New_Hall 2.89 290 -3.18 -1.81
       Newnham 2.86 353 -1.66 -0.82
      Pembroke 3.05 330 -1.20 -0.58
    Peterhouse 3.08 194  0.32  0.33
```

```
       Queens' 3.24 421  1.48  0.92
      Robinson 3.12 316  1.35  0.92
St.Catherine's 3.14 368  1.40  0.91
     St.John's 3.21 469  1.31  0.80
        Selwyn 3.11 293  1.05  0.75
 Sidney_Sussex 3.06 271  0.89  0.67
       Trinity 3.34 556  2.75  1.46
  Trinity_Hall 3.00 292 -0.27 -0.04
Lucy_Cavendish 2.54  46 -2.42 -2.07
       Wolfson 2.65  49 -3.01 -2.57
   Hughes_Hall 3.74  17 -1.60 -1.47
    StEdmund's 2.79  66 -0.97 -0.74
```

We can compare the 2 sets of residuals graphically via

```
qqnorm(re) ; qqline(re)
qqnorm(be) ; qqline(be)
```

and this would be improved by arranging that the (x,y) axes are the same in the two graphs.
The next step?
Let n_1, n_2, n_3 be the number of First, II(i)'s and 'the rest', respectively.
Here is the 3-cell generalization of the beta-binomial.
Assume that, conditional on the vector p, where $p_1 + p_2 + p_3 = 1$, that (n_1, n_2, n_3) has the trinomial distribution $Mn(n, p_1, p_2, p_3)$, and then give p the Dirichlet distribution, with unknown parameters $(\theta_1, \theta_2, \theta_3)$ say. Then the unconditional frequency function is

$$pr(n|\theta) \propto c(\theta) \Pi_i [\Gamma(n_i + \theta_i)/\Gamma(\theta_i)]$$

where

$$c(\theta) = \frac{\Gamma(\theta_1 + \theta_2 + \theta_3)}{\Gamma(n + \theta_1 + \theta_2 + \theta_3)}$$

Thus, if we find the product of this over the 28 observations, we can maximise the corresponding log-likelihood with respect to $(\theta_1, \theta_2, \theta_3,)$

and then pick out colleges that are ‘outliers’. The corresponding maximum likelihood estimates are

$$(\widehat{\theta}) = (23.37, 56.90, 34.63).$$