Overdispersion

A common phenomenon in situations where the obvious first-choice glm is binomial or Poisson.

Data may inherently be more variable than predicted by the model's mean-variance relationship.

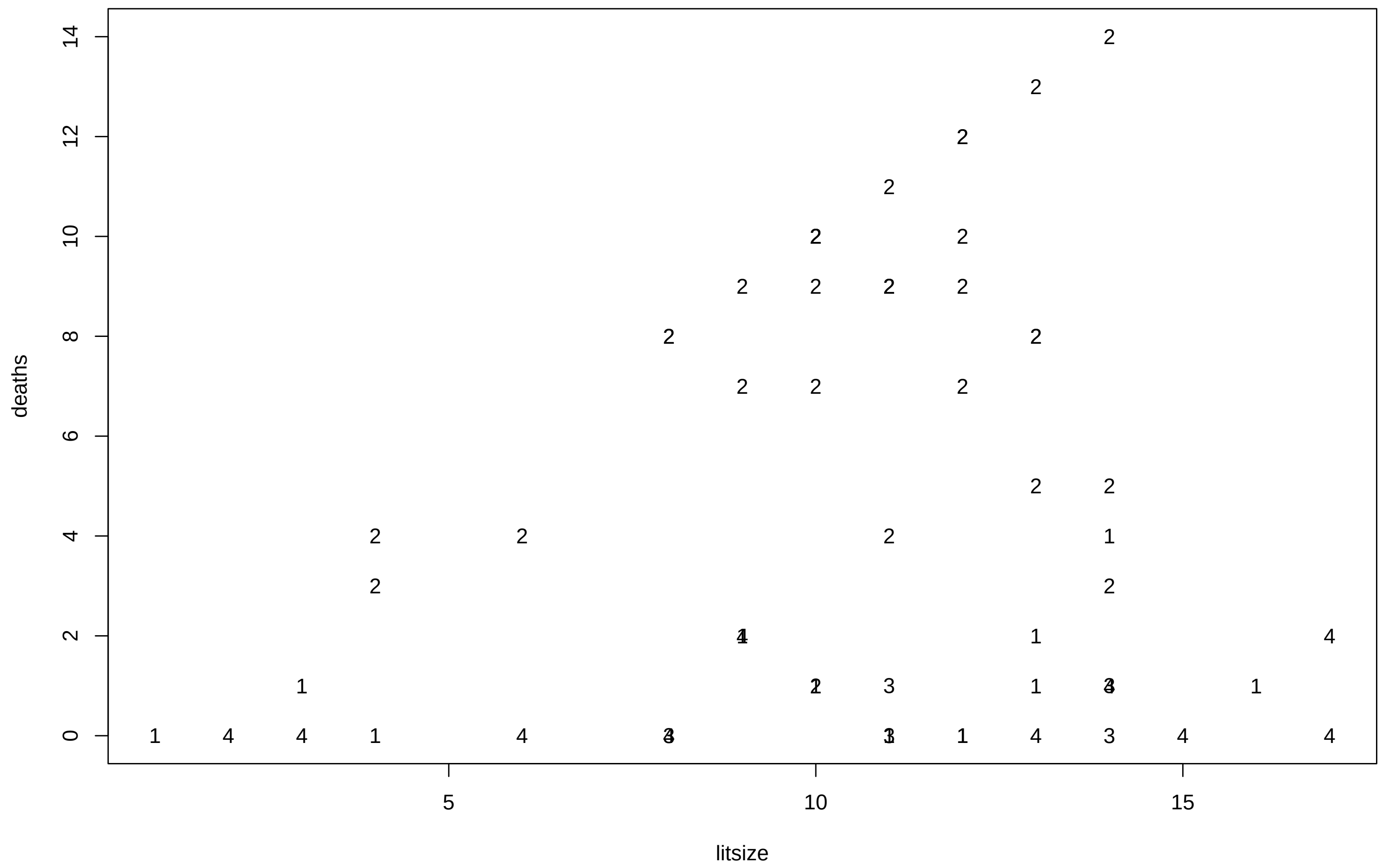
Yields large deviance.

Standard errors too small if problem ignored?

```
### data from Agresti book, available at www.stat.ufl.edu
### 58 pregnant female rats on iron-deficient diets
### mortality per litter in offspring
### treatment 2:none
###           1: iron supplement days 7 and 10
###           3: iron supplement days 0 and 7
###           4: iron supplement weekly

> dat <- read.table("rats.txt", header=F, row.names=1,
      col.names=c("", "trtmnt", "litsize", "deaths") )
> dat$trtmnt <- factor(dat$trtmnt, levels=c(2,1,3,4),
            labels=c("7/10", "none", "0/7", "weekly"))
> attach(dat)

plot(litsize, deaths,
     pch=as.character(as.numeric(trtmnt)))
```


deaths
litsize
0
2
4
6
8
10
12
14
5
10
15

Quasi-likelihood: Replace known scale parameter ϕ with estimated $\hat{\phi} = (n-p)^{-1}D$ or a better estimator.

Won't affect $\hat{\beta}$, but will boost standard errors by factor of $\sqrt{\hat{\phi}}$.

Recognition of additional uncertainty because of extra variability in data.

```
Call:glm(formula =cbind(deaths, litsize - deaths) ~ trtmnt,
         family = binomial)

              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.1785     0.3046  -7.153 8.51e-13 ***
trtmntnone      3.3225     0.3308  10.043  < 2e-16 ***
trtmnt0/7      -1.1537     0.7814  -1.476    0.140
trtmntweekly   -0.8071     0.5503  -1.467    0.142

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 509.43  on 57  degrees of freedom
Residual deviance: 173.45  on 54  degrees of freedom
AIC: 252.92
```

```
Call: glm(formula= cbind(deaths, litsize-deaths) ~ trtmnt,
          family = quasibinomial)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.1785     0.5155  -4.226 9.23e-05 ***
trtmntnone     3.3225     0.5600   5.933 2.18e-07 ***
trtmnt0/7     -1.1537     1.3227  -0.872    0.387
trtmntweekly  -0.8071     0.9315  -0.867    0.390

(Dispersion parameter for quasibinomial family
 taken to be 2.864945)
    Null deviance: 509.43  on 57  degrees of freedom
Residual deviance: 173.45  on 54  degrees of freedom
AIC: NA
```

Comfortable doing inference without a real model?

Say Y is a random variable taking values in $\{0, \ldots, n\}$ with mean np and variance $\phi np(1-p)$. Can show $\phi < n$.

Note that in the Rats data ex., $\hat{\phi} = 2.86$, while $n_i = 1, 2$ for some litters.

What about using a real model with $v(\mu) > \mu(1-\mu)$ for "binomial" data, or $v(\mu) > \mu$ for "Poisson" data.

Negative-Binomial Distribution

This family of distributions can be parameterized as $E(Y) = \mu$, $Var(Y) = \mu + \mu^2/\theta$.

So have a GLM, but with an unknown parameter in the variance function (does complicte fitting algorithm).

Connection to usual parameterization, $Y \sim$ number of failures in sequence of independent trials performed until α successes are seen, where p is success probability for each trial?

More relevant representation of NB distribution as a **mixture**.

```
> ### simulate neg-binom data
> set.seed(13)
> n <- 100
> x1 <- rnorm(n); x2 <- .8*x1 + sqrt(1-.8^2)*rnorm(n)
> mu <- exp(1+.25*x1)
> mu2 <- mu*(rgamma(n,3)/3)
> y <- rpois(n,mu2)

> fit3 <- glm(y~cbind(x1,x2),family=poisson)
> fit4 <- glm(y~cbind(x1,x2),family=quasipoisson)
> library("MASS")
> fit5 <- glm.nb(y~cbind(x1,x2))
```

```
Call: glm(formula = y ~ cbind(x1, x2), family = poisson)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-3.0262  -1.3048  -0.4312   0.7770   4.4132

                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.11110    0.05781  19.218  < 2e-16 ***
cbind(x1, x2)x1  0.29669    0.08992   3.300 0.000969 ***
cbind(x1, x2)x2 -0.07716    0.08815  -0.875 0.381410

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 240.42  on 99  degrees of freedom
Residual deviance: 224.11  on 97  degrees of freedom
AIC: 483.19
```

```
Call: glm(formula = y ~ cbind(x1, x2),
          family = quasipoisson)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.11110    0.08679  12.802   <2e-16 ***
cbind(x1, x2)x1  0.29669    0.13499   2.198   0.0303 *
cbind(x1, x2)x2 -0.07716    0.13234  -0.583   0.5612

(Dispersion parameter for quasipoisson family
 taken to be 2.253698)
    Null deviance: 240.42  on 99  degrees of freedom
Residual deviance: 224.11  on 97  degrees of freedom
AIC: NA
```

```
Call: glm.nb(formula = y ~ cbind(x1, x2),
             link = log)

                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.11061    0.08611  12.898   <2e-16 ***
cbind(x1, x2)x1  0.28608    0.13655   2.095   0.0362 *
cbind(x1, x2)x2 -0.05304    0.13282  -0.399   0.6897

(Dispersion parameter for Negative Binomial(2.4708)
 family taken to be 1)

    Null deviance: 117.46  on 99  degrees of freedom
Residual deviance: 110.11  on 97  degrees of freedom
AIC: 444.15
```

```
Correlation of Coefficients:
                (Intercept) cbind(x1, x2)x1
cbind(x1, x2)x1 -0.02
cbind(x1, x2)x2 -0.01       -0.74

              Theta:  2.471
          Std. Err.:  0.673

 2 x log-likelihood:  -436.149
```