Deviance Distributions and Aggregation for Count Data

The residual deviance in fitting a generalised linear model has an approximate chi-square distribution under some circumstances.

The main requirement is that the components of the deviance are individually reasonably close to chi-square distributions. For models such as glms with Binomial or Poisson or the multinomial models, the components (this would require a lot of maths to justify rigorously) tend to behave like the squares of the contingency table chi-square statistic ie $\frac{o-e}{\sqrt{e}}$. This tends to be approximately normal if e is large. For example, if $X \sim \texttt{Pois}(\lambda)$, then $\frac{X-\lambda}{\sqrt{\lambda}}$ tends to standard normal as $\lambda \to \infty$.

So the end result is that the residual deviance tends to be approximately chi-square distributed as long as the expectations (of the individual counts that were used to obtain it) are large - and 5 is the rule of thumb.

This is illustrated in the following R function and the implementations of it for various choices of n, N, λ that follow.

The figures show a poor fit to the chi-square distribution with 99 degrees of freedom for the histogram of 1000 residual deviances, when fitting a Poisson model to 100 observations from a Poisson distribution with mean $\lambda = 1, 2$, a better fit with $\lambda = 5$ and a good fit with $\lambda = 10$.

```r
ShowDevianceDist <- function(n,N,lambda){
numsim <- n*N
RandPois <- rpois(numsim,lambda=lambda)
DeviPois <- rep(0,N)
for (i in 1:N) {
j = (i-1)*n + 1
k = i*n
DeviPois[i] <- glm(RandPois[j:k]~1,family="poisson")$deviance
}
histout <- hist(DeviPois, breaks = 40,freq=FALSE)
lower <- min(qchisq(0.001,df=n-1),min(histout$mids))
upper <- max(qchisq(0.999,df=n-1),max(histout$mids))
x <- seq(from=lower,to=upper,by=0.2)
y <- dchisq(x,df=n-1)
lines(x,y)
}
```

Figure 1: Histogram of 1000 Deviances for independent groups of 100 Poisson(1) fitted by glm

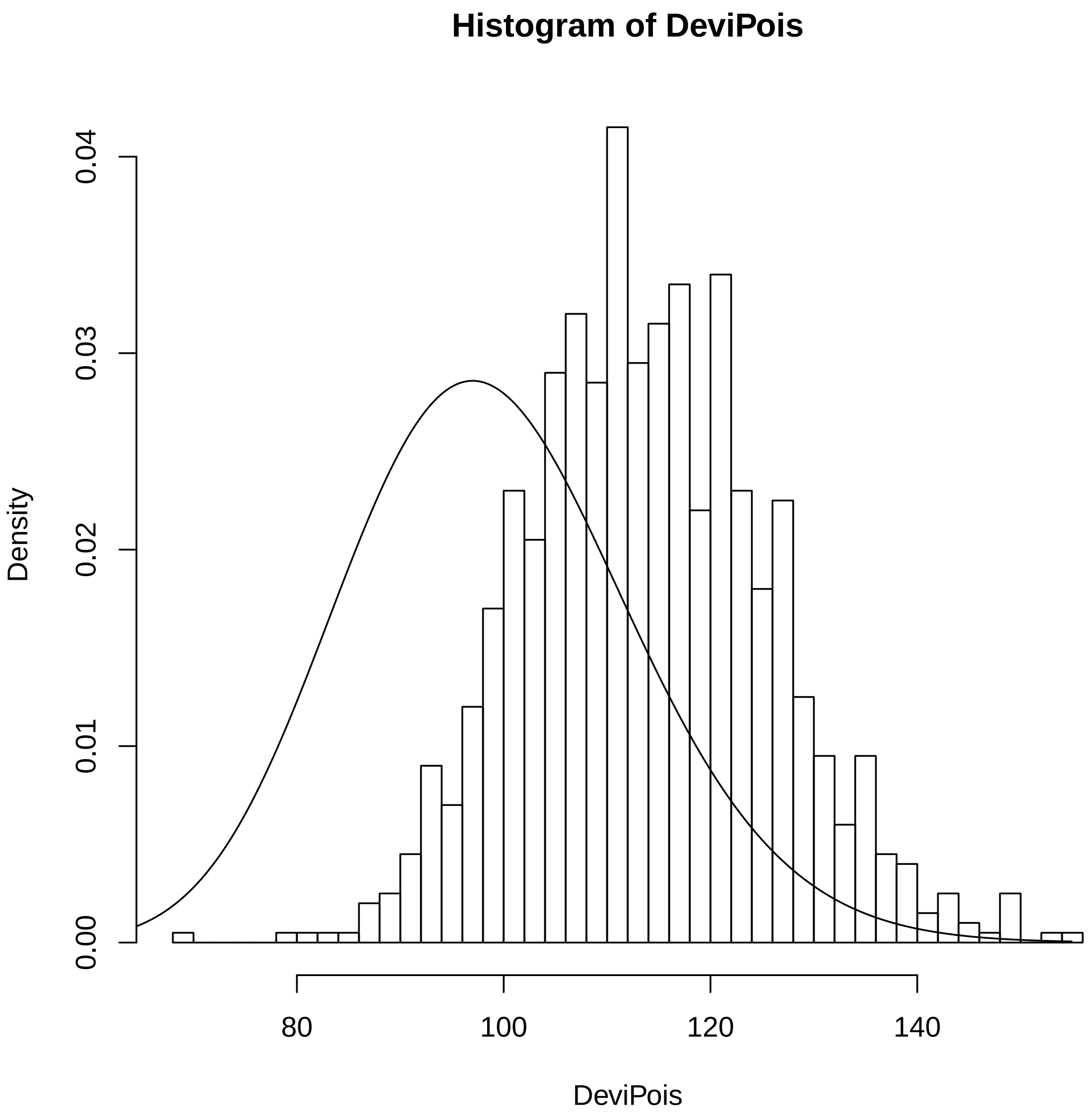


Figure 2: Histogram of 1000 Deviances for independent groups of 100 Poisson(1) fitted by glm

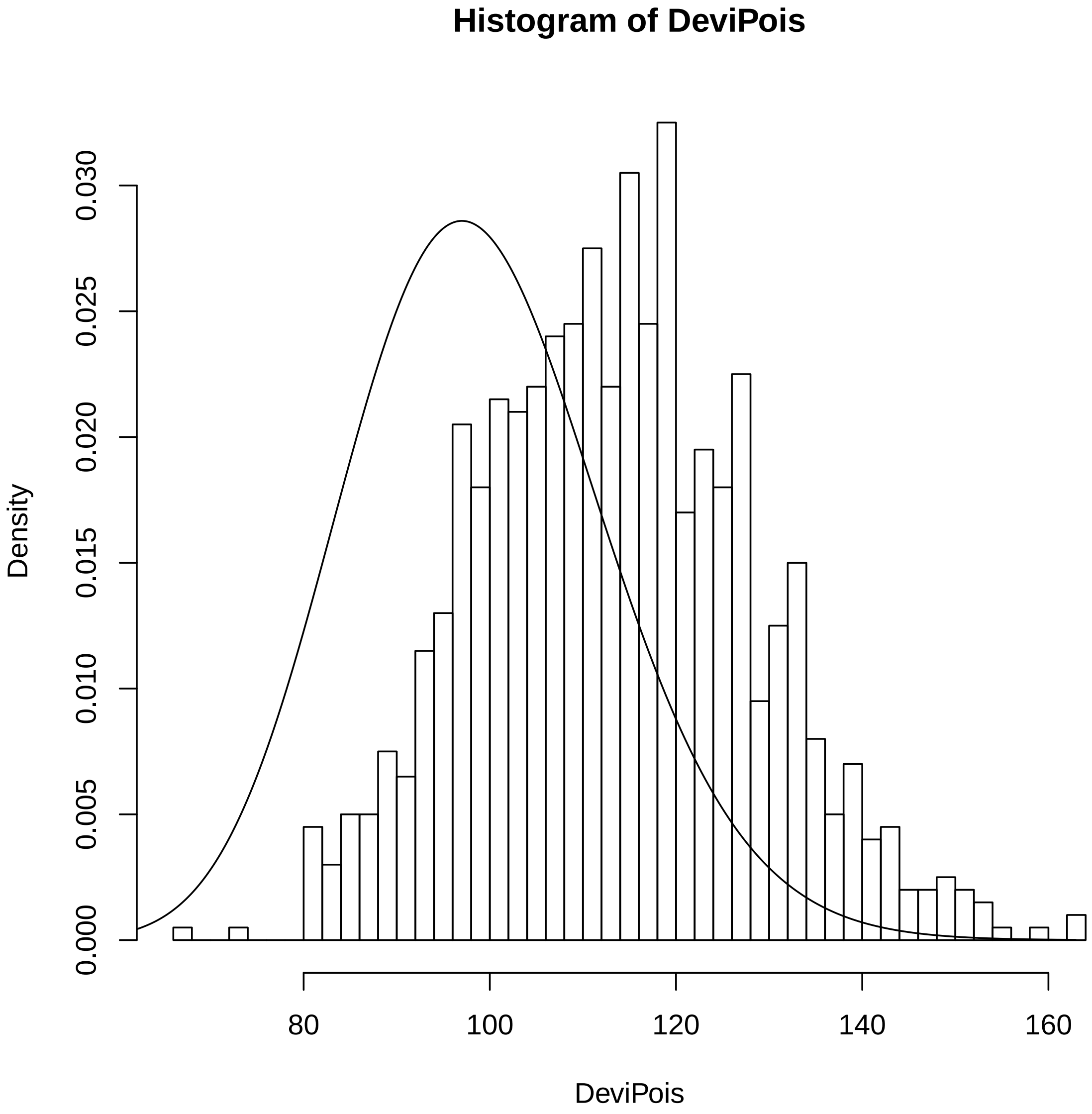


Figure 3: Histogram of 1000 Deviances for independent groups of 100 Poisson(1) fitted by glm

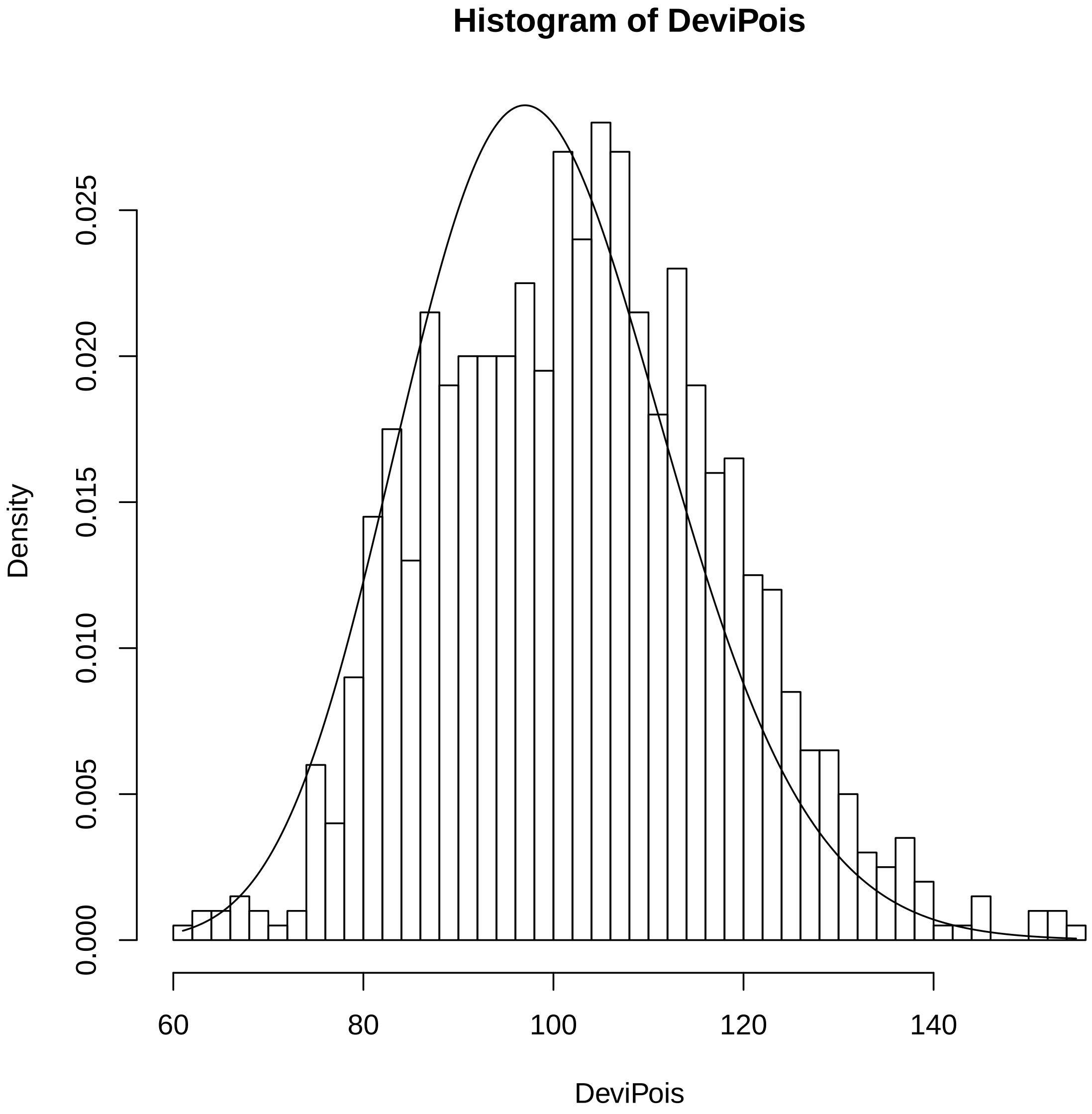
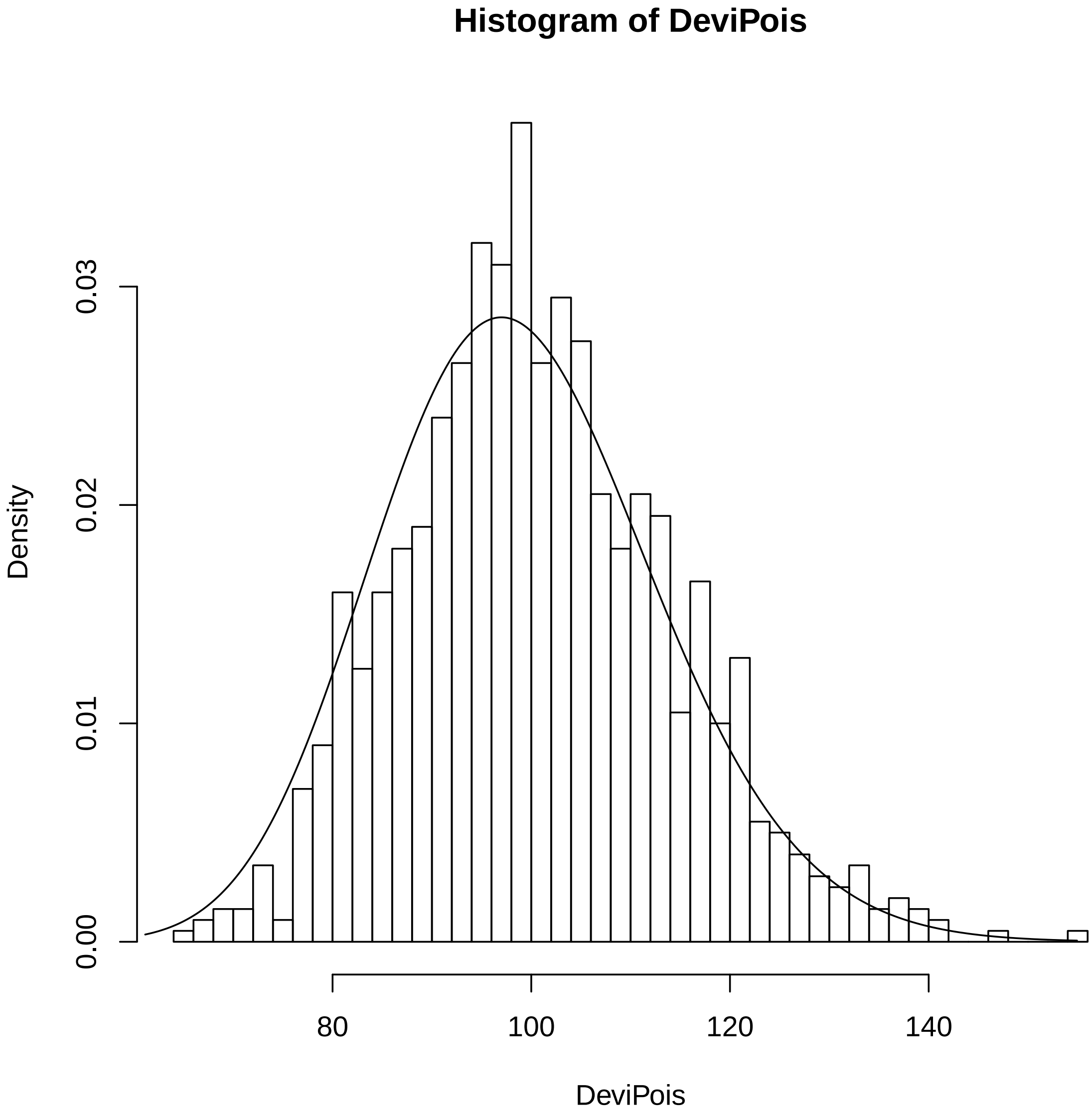


Figure 4: Histogram of 1000 Deviances for independent groups of 100 Poisson(1) fitted by glm



The implication of this is that it best to *aggregate* count data to the highest level possible for fitting the needed models, in order to have believable resdiual deviances. So if there were 100 patients in two groups of size 50 one of whom got a treatment and the other group did not, rather than treat this as 100 independent observations on Bernoulli random variables with a classifying factor for the two groups. It is best to treat this as two observations on independent Binomials, each with $n = 50$ with a classifying factor.

Note that in the Binomial case, the degrees of freedom for the residual deviance when one probability is fitted to all 100 observations are 1 (2 observations with 1 parameter fitted) and they are 0 for fitting two probabilities (2 observations with 2 parameters fitted) , whereas for two groups of 50 Bernoulli the degrees of freedom are 99 (100 observations with 1 parameter fitted) and 98 (100 observations with 2 parameters fitted) respectively.

The desirability of aggregation is also true for multinomial models (for example, where the 100 patients independently each had a random groups chosen for them) or Poisson models (where the overall number of patients was random).

However, whether you aggregate or not, does not change the testing for a treatment effect. Although the *residual deviances* will have a better chi-square fit for the aggregated data and a smaller number of degrees of freedom , the *difference of residual deviances* is exactly the same for these two models.

This is also true for the different models for contingency tables (for example, multinomial and product multinomial).

This is illustrated in the following R output which gives random outcomes in the scenario above with probablity of success, 0.5, for group 1 and probability of success, 0.8, for group 2. The data in z is the aggregated values for groups 1 and 2.

```
y <- c(rbinom(50,size=1,prob=0.5),rbinom(50,size=1, prob=0.8))
group <- factor(c(rep(1,50),rep(2,50)))
model1 <- glm(cbind(y,1-y) ~ 1, family = "binomial")
model2 <- glm(cbind(y,1-y) ~ group, family = "binomial")
anova(model1,model2)

## Analysis of Deviance Table
##
## Model 1: cbind(y, 1 - y) ~ 1
## Model 2: cbind(y, 1 - y) ~ group
##   Resid. Df Resid. Dev Df Deviance
## 1        99     131.79
## 2        98     121.92  1   9.8656

z <- c(sum(y[1:50]),sum(y[51:100]))
groupz <- factor(c(1,2))
model3 <- glm(cbind(z,50-z) ~ 1, family = "binomial")
model4 <- glm(cbind(z,50-z) ~ groupz, family = "binomial")
anova(model3,model4)

## Analysis of Deviance Table
##
## Model 1: cbind(z, 50 - z) ~ 1
## Model 2: cbind(z, 50 - z) ~ groupz
##   Resid. Df Resid. Dev Df Deviance
## 1         1     9.8656
## 2         0     0.0000  1   9.8656
```