**Slide 1**

**Reference:** Agresti, Sections 7.4, 7.5

### Models for zero-inflated data

- One reason for overdispersion may be an excessive amount of zeros in the data.
- Data that are **zero-inflated** relative to data expected for a Poisson GLM are common when many subjects have a 0 response.

**Slide 2**

### Zero-inflated Poisson (ZIP) model

The ZIP model is based on a mixture of a Poisson distribution for $Y$ and a degenerate distribution at 0

$$y_i \sim \begin{cases} 0 & \text{with probability } 1-\phi_i \\ \text{Poisson}(\lambda_i) & \text{with probability } \phi_i \end{cases}$$

so that

$$f(y_i) = (1-\phi_i)\delta_0 + \phi_i f_p(y_i), \qquad y_i = 0, 1, 2, \ldots$$

which can be re-expressed as

$$\begin{aligned} P(y_i = 0) &= (1-\phi_i) + \phi_i e^{-\lambda_i} \\ P(y_i = j) &= \phi_i \tfrac{e^{-\lambda_i}\lambda_i^j}{j!}, \qquad j = 1, 2, \ldots \end{aligned}$$

**Slide 3**

A latent variable is introduced to fit this model, such that

$$\begin{cases} y_i = 0 & \text{if } z_i = 0 \\ y_i \sim \text{ Poisson}(\lambda_i) & \text{if } z_i = 1 \end{cases}$$

**Slide 4**

The mean and variance for the mixture model are given by:

$$\begin{aligned}
E[y_i] &= E[E[y_i|z_i]] = \sum_{z_i=0}^{1} E[y_i|z_i] \cdot p(z_i) \\
&= E[y_i|z_i = 0] \cdot P(z_i = 0) + E[y_i|z_i = 1] \cdot P(z_i = 1) \\
&= 0(1 - \phi_i) + \lambda_i \phi_i = \lambda_i \phi_i \\
Var(y_i) &= E[Var(y_i|z_i)] + Var(E[y_i|z_i]) \\
&= \sum_{z_i=0}^{1} Var[y_i|z_i] \cdot p(z_i) + E[(E[y_i|z_i] - E[E[y_i|z_i]])^2] \\
&= 0(1 - \phi_i) + \lambda_i \phi_i + (E[y_i|z_i = 0] - \lambda_i \phi_i)^2 (1 - \phi_i) + (E[y_i|z_i = 1] - \lambda_i \phi_i)^2 \phi_i \\
&= \lambda_i \phi_i + (0 - \lambda_i \phi_i)^2 (1 - \phi_i) + (\lambda_i - \lambda_i \phi_i)^2 \phi_i \\
&= \lambda_i \phi_i (1 + \lambda_i (1 - \phi_i))
\end{aligned}$$

Note that $E[Y] < Var(Y)$, so this model accounts for overdispersion relative to a Poisson model.

**Slide 5**

- The expectation of $y_i$ depends on $\lambda_i$ and $\phi_i$.
- In addition to modeling the mean $\lambda_i$ of the Poisson distribution, we also need a link function for $\phi_i$.
- Let $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ represent two subsets of covariates, which may overlap. Their effects on $\lambda_i$ and $\phi_i$ are modeled as

$$\begin{aligned}\log\left(\frac{\phi_i}{1-\phi_i}\right) &= \boldsymbol{x}_{1i}\boldsymbol{\beta}_1 \\ \log(\lambda_i) &= \boldsymbol{x}_{2i}\boldsymbol{\beta}_2\end{aligned}$$

The explanatory variables affecting $\phi_i$ need not be the same as those affecting $\lambda_i$.

**Slide 6**

### Inference for ZIP model

The likelihood and log-likelihood for the two parts of the model are given by

$$\begin{aligned}L(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) &= \prod_{i=1}^{N} I_{y_i=0}\left[(1-\phi_i)+\phi_i e^{-\lambda_i}\right]\cdot I_{y_i>0}\phi_i\frac{\lambda_i\lambda_i^{y_i}}{y_i!} \\ \log L(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) &= \sum_{y_i=0}\log\left\{(1-\phi_i)+\phi_i e^{-\lambda_i}\right\}+\sum_{y_i>0}\{\log\phi_i-\lambda_i+y_i\log\lambda_i-\log(y_i!)\} \\ &= \sum_{y_i=0}\log\left[1+e^{\boldsymbol{x}_{1i}\boldsymbol{\beta}_1}\exp(-e^{\boldsymbol{x}_{2i}\boldsymbol{\beta}_2})\right]-\sum_{i=1}^{N}\log(1+e^{\boldsymbol{x}_{1i}\boldsymbol{\beta}_1}) \\ &\quad+\sum_{y_i>0}\left[\boldsymbol{x}_{1i}\boldsymbol{\beta}_1-e^{\boldsymbol{x}_{2i}\boldsymbol{\beta}_2}+y_i\boldsymbol{x}_{2i}\boldsymbol{\beta}_2-\log(y_i!)\right]\end{aligned}$$

The MLE's can be obtained using optimization algorithms, such as the Newton-Raphson algorithm.

**Slide 7**

- In practice, overdispersion still occurs when we condition on $z_i = 1$ in the latent formulation of the ZIP model.
- A zero-inflated negative binomial model may then be more appropriate.

$$\begin{cases} y_i = 0 & \text{if } z_i = 0 \\ y_i \sim \text{ negative-binomial}(\lambda_i, \gamma) & \text{if } z_i = 1 \end{cases}$$

**Slide 8**

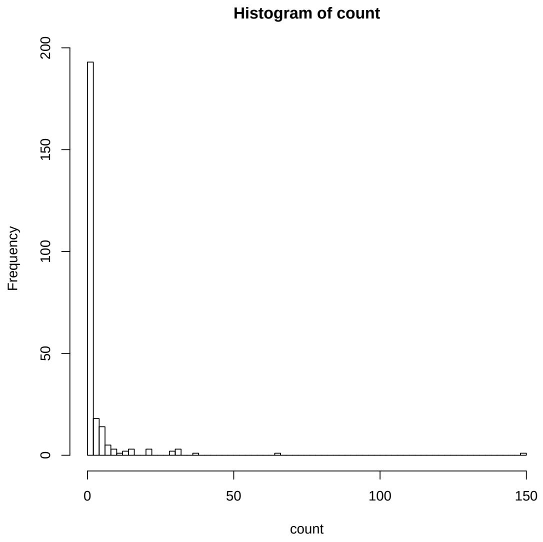
**Example**

Wildlife biologists want to model how many fish are being caught by fishermen at a state park. We have data on 250 groups who went to a park. Each group was questioned about how many fish they caught (`count`), how many children were in the group (`child`), how many people were in the group (`persons`), and whether or not they brought a camper on the park (`camper`).

Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish. So there are excess zeros in the data because of the people who did not fish. The data are saved in the file `fish.txt`.

**Slide 9**

```
hist(count, breaks=100)
```



**Slide 10**

Let us try a Poisson regression model.

```
fit.pois = glm(count ~ camper + child, family=poisson, data=fish)

> summary(fit.pois)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.91026    0.08119   11.21   <2e-16 ***
camper       1.05267    0.08871   11.87   <2e-16 ***
child       -1.23476    0.08029  -15.38   <2e-16 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2958.4  on 249  degrees of freedom
Residual deviance: 2380.1  on 247  degrees of freedom
AIC: 2723.2
```

**Slide 11**

Let us fit a zero-inflated Poisson model with the zero inflation component dependent on `persons`

```
library(pscl)
fit.zip = zeroinfl(count ~ camper+child | persons, data=fish)

> summary(fit.zip)

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.59789    0.08554  18.680   <2e-16 ***
camper       0.83402    0.09363   8.908   <2e-16 ***
child       -1.04284    0.09999 -10.430   <2e-16 ***


Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
```

**Slide 12**

```
(Intercept)   1.2974     0.3739   3.470 0.000520 ***
persons      -0.5643     0.1630  -3.463 0.000534 ***
---
Number of iterations in BFGS optimization: 12
Log-likelihood: -1032 on 5 Df
```

**Slide 13**

- The predictor of excess zeros, `persons`, is statistically significant. The log-odds of excess zeros is 0.56 lower for each additional person in the group.
- The expected change in log(count) for a one-unit increase in child is -1.043, adjusting for `camper` and `persons`.
- Groups with campers have an expected log count that is 0.834 higher than groups without campers, controlling for `child` and `persons`.
- The AIC for this model is

$$-2 \times \log L + 2 \times p = -2(-1032) + 2(5) = 2074$$

- The zero-inflated Poisson model fits the data better than a Poisson model.

**Slide 14**

We can use the Vuong non-nested test to compare the fit of two non-nested models to the same data.

```
> vuong(fit.pois, fit.zip)
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                   -3.574254 model2 > model1 0.00017561
AIC-corrected         -3.552392 model2 > model1 0.00019087
BIC-corrected         -3.513900 model2 > model1 0.00022079
```

**Slide 15**

How about a negative binomial model?

```
fit.nb = glm.nb(count ~ camper+child, data=fish)

> summary(fit.nb)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0727     0.2425   4.424 9.69e-06 ***
camper        0.9094     0.2836   3.206  0.00135 **
child        -1.3753     0.1958  -7.025 2.14e-12 ***
---
(Dispersion parameter for Negative Binomial(0.2553) family taken to be 1)

    Null deviance: 258.93  on 249  degrees of freedom
Residual deviance: 201.89  on 247  degrees of freedom
AIC: 887.42
```

**Slide 16**

```
              Theta:  0.2553
          Std. Err.:  0.0329

 2 x log-likelihood:  -879.4210
```

- The negative binomial dispersion parameter is estimated to be $\hat{\gamma} = 1/0.2553 = 3.917$ with a 95% CI $(3.13, 5.24)$.
- Based on the AIC, the negative binomial model fits the data better than the Poisson and the zero-inflated Poisson models.

**Slide 17**

Would a zero-inflated negative binomial model fit better?

```
fit.zipnb = zeroinfl(count ~ camper+child | persons, dist="negbin", data=fish)

> summary(fit.zipnb)

Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3710     0.2561   5.353 8.64e-08 ***
camper        0.8791     0.2693   3.265   0.0011 **
child        -1.5153     0.1956  -7.747 9.41e-15 ***
Log(theta)   -0.9854     0.1760  -5.600 2.14e-08 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6031     0.8365   1.916   0.0553 .
persons      -1.6666     0.6793  -2.453   0.0142 *
```

**Slide 18**

```
---
Theta = 0.3733
Number of iterations in BFGS optimization: 22
Log-likelihood: -432.9 on 6 Df
```

**Slide 19**

- The predictors `child` and `camper` in the part of the negative binomial regression model predicting number of fish caught (`count`) are both statistically significant.
- The predictor `person` in the part of the logit model predicting excess zeros is statistically significant.
- The expected count of fish caught is about 5 times smaller ($\exp(-1.515) = 0.22$) for each additional child in the group, holding `camper` and `person` constant.
- Groups with a camper have an expected count that is 2.4 times higher ($\exp(0.879) = 2.41$) than groups without a camper, holding `camper` and `person` constant.
- The log odds of having excess zeros would decrease by 1.67 for every additional person in the group. In other words, the more people in the group the less likely the zeros are due to not having gone fishing.

**Slide 20**

- The negative binomial dispersion parameter is estimated to be $\hat{\gamma} = 1/0.3733 = 2.679$.
- The AIC for this model is

$$-2 \times \log L + 2 \times p = -2(-432.9) + 2(6) = 877.8$$

- The data are overdispersed and a zero-inflated negative binomial model is more appropriate than a zero-inflated Poisson model.
- The Vuong test suggests that the zero-inflated negative binomial model may fit the data marginally better than a standard negative binomial model.

**Slide 21**

```
> vuong(fit.nb, fit.zipnb)
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A  p-value
Raw                   -1.7017116 model2 > model1 0.044405
AIC-corrected         -1.2026316 model2 > model1 0.114559
BIC-corrected         -0.3238863 model2 > model1 0.373012
```