Generalized Linear Models for Count Outcomes and Zero-Inflated Count Outcomes

- Topics:
  - Roadmap of models for non-normal outcomes
  - Generalized linear models for count outcomes
  - Model adjustments for misfit to Poisson distribution
  - SAS and STATA software routines for fitting count models

Generali*zed* Linear Models

- **Generali*zed* linear models:** link-transformed conditional mean is predicted by the linear model; ML estimator uses not-normal conditional distributions in the outcome data likelihood
  - **Btw, in multilevel models,** level-1 conditional model has some not-normal distribution, but level-2 random effects are usually multivariate normal

- **Two parts: Link function + other conditional distribution**
  - **Done: Categorical → Logit/Probit/Log-Log/C-Log-Log**
    - **Bernoulli for binary; multinomial for ordinal or nominal**
  - **Now: Counts → Log + some kind of Poisson or Negative Binomial**
    - **Zero-inflated counts → zero-inflated or hurdle variants**
  - **Later: Bounded → Logit + some kind of Binomial or Beta**
  - **Later: Skewed Continuous → Log + Log-Normal/Gamma**
    - **Zero-inflated continuous → hurdle variants**

A Taxonomy of Not-Normal Outcomes

- **"Discrete" outcomes**—all responses are **whole** numbers
  - **Categorical variables** in which **values are labels**, not amounts
    - Bernoulli (2 options) or multinomial (3+ options) distributions
    - Question: Are the values ordered → **which link?**
  - **Count of things that happened**, so values < 0 cannot exist
    - Sample space goes from 0 to +∞
    - Some kind of Poisson or Negative Binomial distribution
    - **Log link (usually) so predicted outcomes can't go below 0**
    - Question: Are there *extra* 0 values? What to do about them?

- **"Continuous" outcomes**—responses can be **any** number
  - Question: What does the residual distribution look like?
    - Symmetric or skewed? Unnatural boundary (censored)?

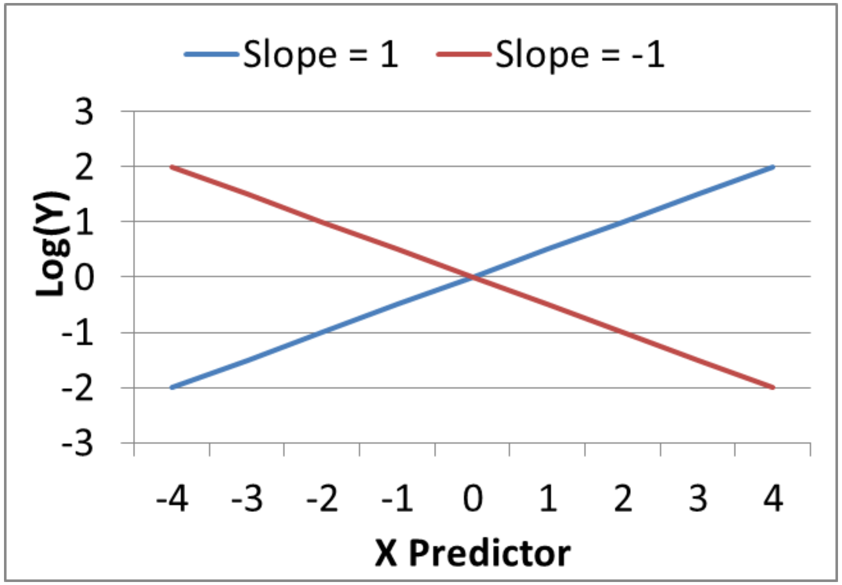
Log Link for Count Outcomes

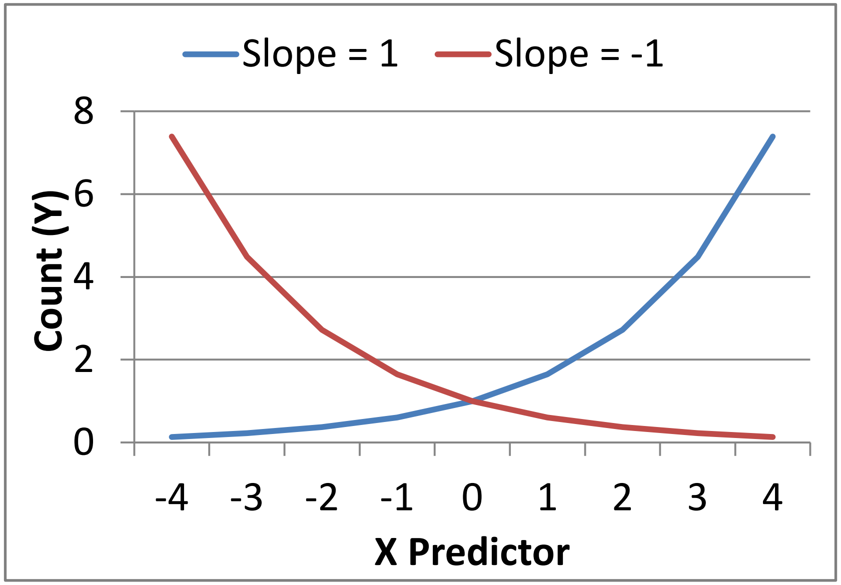
This is an ***unbounded* linear model** that predicts the Log of the Expected Count…

$$Log[E(y_i)] = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i})$$

…that becomes an expected count bounded at 0 via an inverse link of exp(log count):

$$[E(y_i)] = \boldsymbol{exp}[\boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i})]$$





Models for Count Outcomes

- **Counts**: non-negative integer responses (unbounded positive)
  - Link: g(•) $Log[E(y_i)] = Log(\mu_i) = [\text{model}]$ → predicts log of count as $\hat{y}_i$
  - Inverse Link: $\text{g}^{-1}(\bullet)$ $E(y_i) = \exp(\hat{y}_i)$→ to un-log $\hat{y}_i$ back to count
  - e.g., if the model-scale predict log count: *Log*$(\mu_i) = -1$, the data-scale predicted actual count is: $\exp(-1) = 0.368$
    - So even though counts are only integers, predicted counts are not!
  - Btw, you can control for differences in time measured via an **offset** (or **exposure**) log-transformed predictor variable whose slope is fixed =1

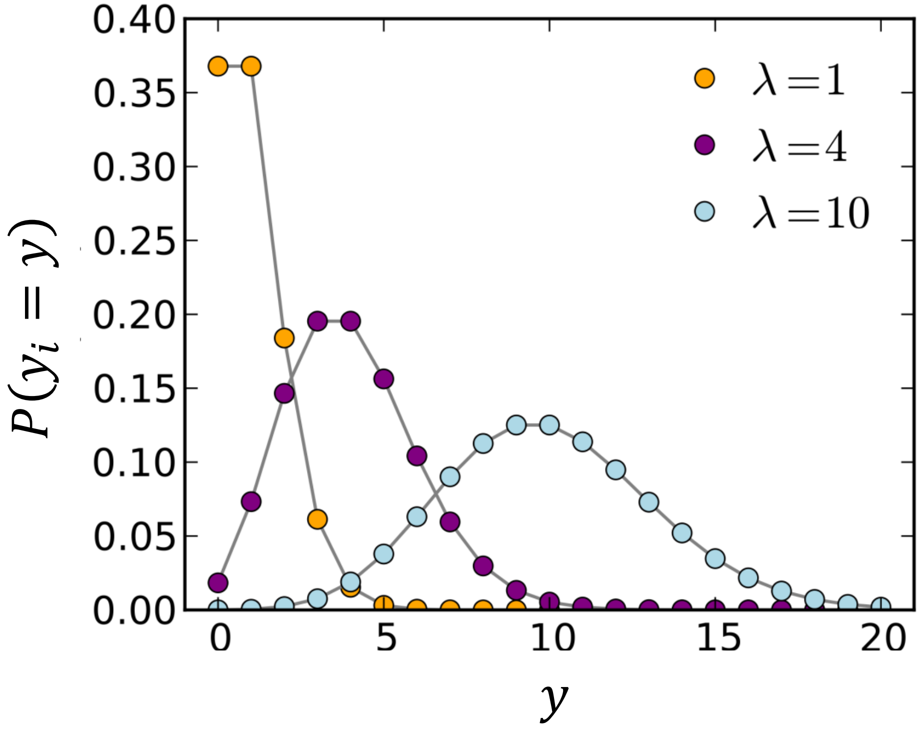
- $\mathbf{exp}(\boldsymbol{\beta}_x)$ gives an effect size called an "**incidence-rate ratio**" (**IRR**) that is on same scale as an odds ratio (IRR = 1 means no effect)
  - e.g., IRR = 1.25 for $x_i = 0$ or 1? $x_i = 1$ counts are "25% higher"
  - e.g., IRR = 0.75 for $x_i = 0$ or 1? $x_i = 1$ counts are "25% lower"
  - Stata also gives McFadden's **pseudo-R²** $= 1 - (LL_{model}/LL_{empty})$

- Choosing the "right" **conditional distribution** is the tricky part!

Poisson Conditional Distribution

- Poisson distribution has **one parameter, λ,** which is both its mean and its variance (so λ = mean μ = variance in Poisson)
- PDF: $f(y_i) = Prob(y_i = y) = \frac{\mu^y * exp(-\mu)}{y!}$

$y!$ = factorial of y = gamma function $\Gamma(y+1)$



The dots indicate that only integer values are observed.

Distributions with a small expected value (mean or λ) are predicted to have a lot of 0's.

Once $\lambda > 6$ or so, the shape of the distribution is close to a that of a normal distribution.

Image borrowed from: https://en.wikipedia.org/wiki/Poisson_distribution

3 potential problems with Poisson…

- The standard Poisson distribution is rarely sufficient, though

- **Problem #1: When mean ≠ variance**
  - If variance < mean, this leads to "under-dispersion" (not that likely)
  - If variance > mean, this leads to "over-dispersion" (happens frequently)

- **Problem #2: When there are *no* 0 values**
  - Some 0 values are expected from count models, but in some contexts $y_i > 0$ always (but subtracting 1 won't fix it correctly; need to adjust the model)

- **Problem #3: When there are *too many* 0 values**
  - Some amount of 0 values are expected from count distributions already, but many times there are even more 0 values observed than that
  - To fix it, there are two main options, depending on what you do to the 0's

- Each of these problems requires a model adjustment to fix it…

Problem #1: Variance > mean = over-dispersion

- To fix it, we must add a parameter that allows the variance to exceed the mean… it then can become a **Negative Binomial** distribution
  - Two types of extra variance: constant = NB1, quadratic = NB2 (better)

- **Negative Binomial (2)** PDF with **mean $\boldsymbol{\mu}$** and **dispersion scale $\boldsymbol{k}$**:
  - $\mathrm{Prob}(y_i = y\,) = \frac{\Gamma\left(y+\frac{1}{k}\right)}{\Gamma(y+1)*\Gamma\left(\frac{1}{k}\right)} * \frac{(k\mu)^y}{(1+k\mu)^{y+\frac{1}{k}}}$
  - $\boldsymbol{k}$ is a multiplier, such that $\mathrm{Var}(y) = \mu + \boldsymbol{k}\mu^2$
  - Can test whether $k > 0$ via $-$2LL test, although LL for $k = 0$ is undefined

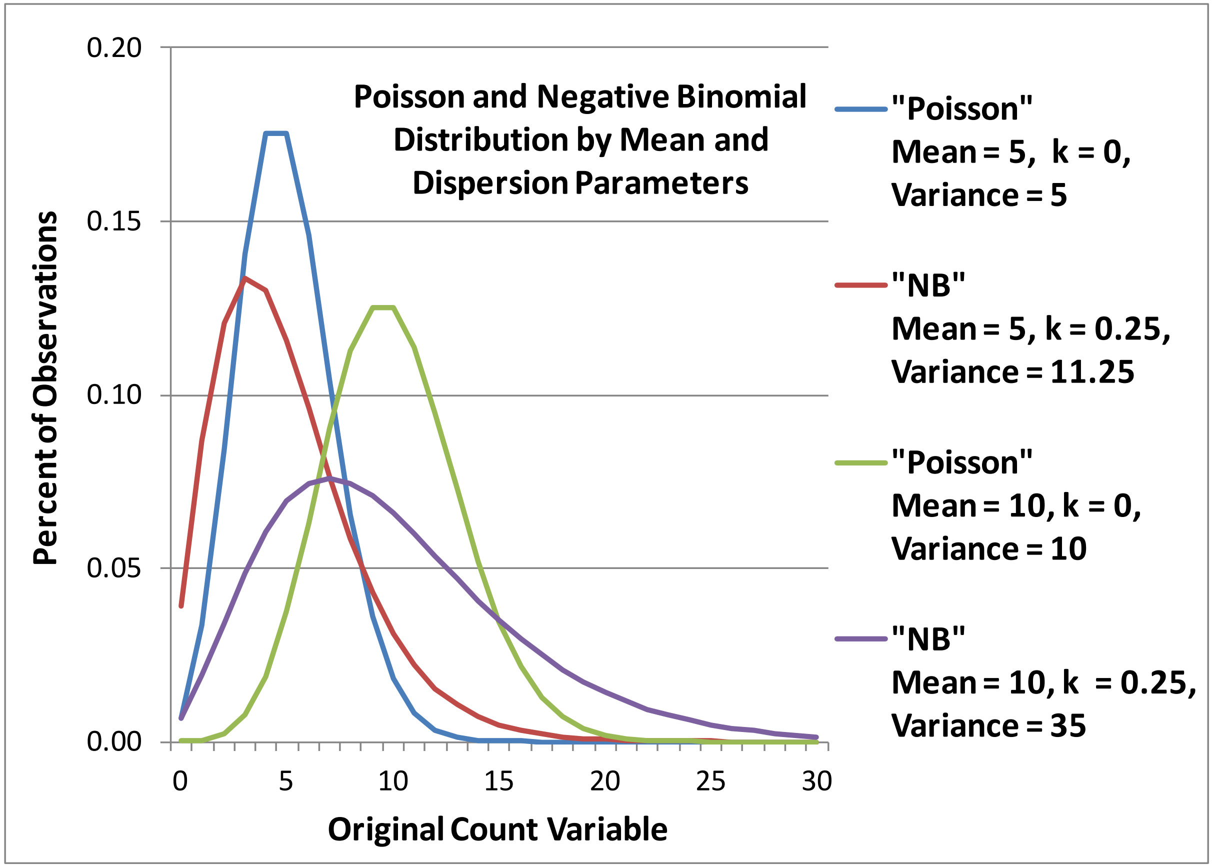
DIST = NEGBIN in SAS;
NBREG or **GLM** in STATA

So ≈ Poisson if $k = 0$

- An alternative model with the same idea is the **generalized Poisson**:
  - Mean: $\frac{\lambda}{1-k}$, Variance: $\frac{\mu}{(1-k)^2}$, so LL is defined for $k = 0$
  - In SAS FMM (and in GLIMMIX via user-defined functions)

GPOISSON
in STATA

Negative Binomial (NB) = “Stretchy” Poisson…



$\text{Mean} = \mu$
$\text{Dispersion} = k$

$$Var(y_i) = \mu + k\mu^2$$

A Negative Binomial model can be useful for count outcomes with extra skewness, but that otherwise follow a Poisson conditional distribution.

- Because its k dispersion parameter is fixed to 0, the Poisson model is nested within the Negative Binomial model—to test improvement in fit:
- Is $-2\left(LL_{Poisson} - LL_{NegBin}\right) > 3.84 \text{ for } df = 1$? Then $p < .05$, keep NB

Pause: Clarifying Terminology

The same words can mean lots of different things, such as:

- “**nonlinear**”—this adjective could refer to:
  - A category of model (that violates “slope*variable + slope*variable”)
  - An effect of a slope (e.g., that creates a quadratic relationship)
  - A type of regression (e.g., using link functions, as opposed to “linear”)

- “**Fit**”—this could mean:
  - Mistakenly used to refer to predictive quality (i.e., amount of variance explained)—this is NOT fit, it’s overall model effect size
  - In **multivariate** models, **fit usually refers to the match** between the model-predicted and real-data covariance matrices or cross-tabs, and has nothing to do with effect size in terms of model predictive quality
  - In generalized linear models, **fit can also refer to the match of the chosen conditional distribution to the observed outcome**… *this is the one I am talking about next!*

Absolute Conditional Distribution Fit

- In addition to comparing the **relative fit** of the Poisson and Negative Binomial (NB) distributions, we also need to examine the "**absolute fit**" of the conditional distribution for the observed outcome distribution
  - e.g., NB may be relatively better than Poisson, but is NB "good enough"?

- **Conditional distribution fit** can be examined using a statistic reported as **Pearson χ^2 / degrees of freedom** in which 1 = good fit
  - Sum over persons of $\left(\frac{y_i - \hat{y}_i}{SD\ at\ \hat{y}_i}\right)^2$, then divided by sample size as DF
  - = data-average / model-expected deviation from mean (should be same, 1)
  - Available for each distribution in SAS via GENMOD or GLIMMIX (possibly others); in STATA via GLM (possibly others)

- Btw, Hardin & Hilbe (2012) describe other "generalized" negative binomial models, including the "heterogeneous negative binomial" in which the dispersion scale factor itself can be predicted—cool!
  - Not available directly in SAS, but I found some NLMIXED code to do it
  - Absolute distribution fit will not be readily available in these types of models

Problem #2: There are no 0 values

- "**Zero-Altered**" or "**Zero-Truncated**" Poisson or Negative Binomial: ZAP/ZANB or ZTP/ZTNB (used in hurdle models)
  - Is usual count distribution, just not allowing any 0 values
  - Single-level models are in SAS PROC FMM using DIST=TRUNCPOISSON for ZTP or DIST=TRUNCNEGBIN for ZTNB
  - Single-level TPOISSON (for ZTP) and TNBREG (for ZTNB) in STATA
  - Multivariate versions could be fitted in SAS NLMIXED or Mplus, too

- e.g., Poisson PDF: $Prob(y_i = y) = \frac{\mu^y * exp(-\mu)}{y!}$
- e.g., Zero-Truncated Poisson PDF: $Prob(y_i = y \mid y_i > 0) = \frac{\mu^y * exp(-\mu)}{y![1 - exp(-\mu)]}$
  - $Prob(y_i = 0) = exp(-\mu)$, so $Prob(y_i > 0) = 1 - exp(-\mu)$
  - Divides by probability of non-0 outcomes so probability still sums to 1

Problem #3: Too many 0 values, Option #1

- "**Zero-Inflated**" Poisson (DIST=ZIP) or NB (DIST=ZINB) in SAS GENMOD or Mplus; ZIP/ZINB/ZIGP in STATA
  - Distinguishes **two kinds of 0 values**: **expected** and **inflated/structural** (extra) through a mixture of Bernoulli + Poisson/NB/GenPoisson)
  - Creates two submodels to predict "if *extra* 0" and "if not, how much"?
    - Does not readily map onto most hypotheses (in my opinion)
    - But a ZIP example would look like this… (ZINB would add *k* dispersion, too)

- Submodel 1: $Logit[p(y_i = extra\ 0)] = \beta_{0z} + \beta_{1z}(x_i)$
  - Predict **<u>being an extra 0</u>** using Link = Logit, Distribution = Bernoulli
  - Don't have to specify predictors for this part, can simply allow an intercept (but need ZEROMODEL option to include predictors in SAS GENMOD)

- Submodel 2: $Log[E(y_i)] = \beta_{0c} + \beta_{1c}(x_i)$
  - Predict **<u>rest of counts (including 0's)</u>** using Link = Log, Distribution = Poisson/Negative Binomial/Generalized Poisson

Zero-Inflated Models for Counts

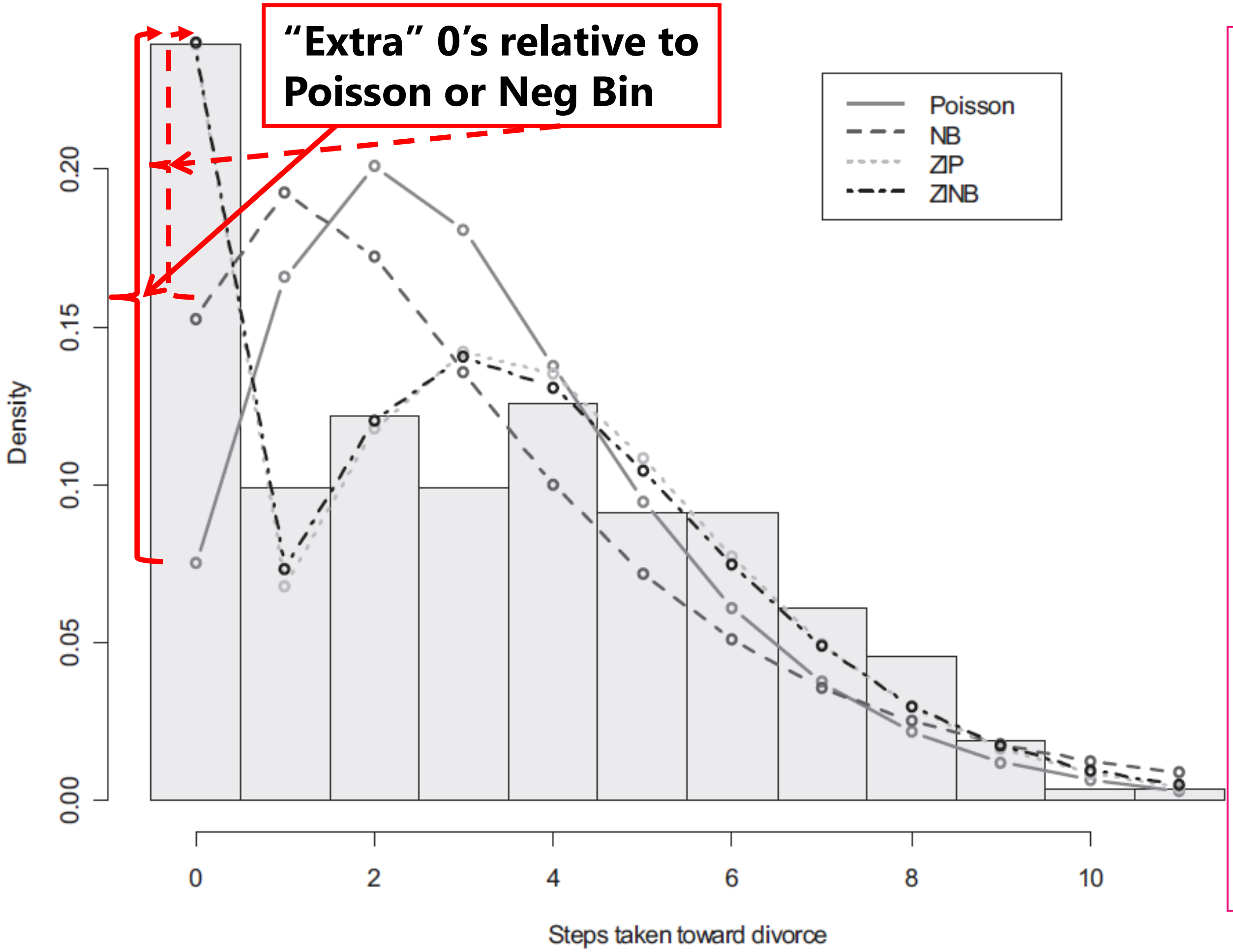


Figure 1. Histogram of Marital Status Inventory with predicted probabilities from regressions. NB = negative binomial; ZIP = zero-inflated Poisson; ZINB = zero-inflated negative binomial.

Zero-inflated distributions come in two flavors: Poisson (mean = variance) and Negative Binomial (variance exceeds mean).

When predictors have this type of distribution it can be helpful to think of them as **semi-continuous** in an "**if and how much**" model (my own terminology):

Pred1: =0 if x=0, 1 if x > 0
= Pred1 is binary
Pred2: =how much if x > 0
= Pred2 is quantitative

Image borrowed and doctored from: Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*(4), 726-735.

Problem #3: Too many 0 values, Option #1

- The Zero-Inflated models get put back together as follows:
  - ω_i ("omega") is the model-predicted probability of being an extra 0:

$$\omega_i = \frac{exp[Logit[p(y_i = extra\ 0)]]}{1 + exp[Logit[p(y_i = extra\ 0)]]}$$

  - μ_i is the model-predicted count for the rest of the distribution:

$$\mu_i = \exp(\hat{y}_i)$$

  - ZIP: Mean (original y_i) $= (1 - \omega_i)\mu_i$
  - ZIP: Variance (original y_i) $= \mu_i + \frac{\omega_i}{(1-\omega_i)}\mu_i^2$
  - ZINB: Mean (original y_i) $= (1 - \omega_i)\mu_i$
  - ZINB: Variance (original y_i) $= \mu_i + \left[\frac{\omega_i}{(1-\omega_i)} + \frac{k}{1-\omega_i}\right]\mu_i^2$

Problem #3: Too many 0 values, Option #2

- "**Hurdle**" models for Poisson or Negative Binomial
  - PH or NBH: Explicitly **separates 0 from non-0 values** through two distinct outcome distributions (Bernoulli + Zero-Altered Poisson/NB)
  - Creates two submodels to predict "if any 0" and "if not 0, how much"?
    - Easier to think about in terms of prediction (in my opinion)

- Submodel 1: $Logit[p(y_i = 0)] = \beta_{0z} + \beta_{1z}(x_i)$
  - Predict **<u>being any 0</u>** using Link = Logit, Distribution = Bernoulli

- Submodel 2: $Log[E(y_i)|y_i > 0] = \beta_{0c} + \beta_{1c}(x_i)$
  - Predict **<u>positive counts</u>** using Link = Log, Distribution = ZAP/ZANB

- These models are not readily available in SAS, but NBH is in Mplus
  - Could be fit in SAS NLMIXED (as could ZIP/ZINB)
  - Can also split DV explicitly and estimate each submodel separately, but you lose the ability for multivariate test of an effect across submodels

More on Comparing Count Models

- Whether or not a dispersion scale parameter is needed (to distinguish Poisson and NB) can be answered via a likelihood ratio test
  - For the most fair comparison, keep the linear predictor model the same
- Whether or not a zero-inflation model is needed should, in theory, also be answerable via a likelihood ratio test… But people disagree about this
  - Problem? Zero-inflation probability can't be negative, so is bounded at 0
  - Other tests have been proposed (e.g., Vuong test in SAS macro or STATA)
  - Can always check AIC and BIC (smaller is better)
- In general, models with the same distribution and different links can be compared via AIC and BIC, but one cannot use AIC and BIC to compare across alternative distributions (e.g., normal or not?)
  - Log-Likelihoods are not on the same scale due to using different PDFs
  - Pearson χ^2 / DF can provide guidance as to fit of conditional distribution

SAS for Discrete Outcomes

- There are many choices for modeling not-normal ***discrete*** outcomes (that include integer values only); most use either an identity or log link
- **Single-level, univariate generalized models by PROC:**
  - GENMOD: DIST= (and default link): Binomial (Logit), Poisson (Log), Zero-Inflated Poisson (Log), Negative Binomial (Log), Zero-Inflated Negative Binomial (Log)
  - FMM: DIST= (and default link): Binomial (Logit), Poisson (Log), Generalized Poisson (Log), Truncated Poisson (Log), Negative Binomial (Log), Uniform
- **Multilevel or multivariate generalized models through GLIMMIX:**
  - Binomial (Logit), Poisson (Log), Negative Binomial (Log)
  - BYOBS, which allows multivariate models by which you specify DV-specific link functions and distributions estimated simultaneously
  - User-defined variance functions for special cases (e.g., generalized Poisson)
  - NLMIXED can also be used to fit any user-defined model

STATA 16 for Discrete Outcomes

- There are many choices for modeling not-normal ***discrete*** outcomes (that include integer values only); most use either an identity or log link

- **Single-level, univariate generalized models:**
  - glm for multiple options, logit, probit, or cloglog for binary, ologit or oprobit for ordinal, poisson, nbreg, or gnbreg for counts, and many more options
  - Most of these allow cluster-corrected or robust standard errors (stay tuned)

- **Multilevel or multivariate generalized models:**
  - meglm for multiple options, melogit, meprobit, or mecloglog for binary, meologit or meoprobit for ordinal, mepoisson or menbreg for counts
  - menl can also be used to fit any user-defined model (haven’t tried that yet)

Summary: Predicting Counts

- A **count** is a discrete outcome that:
  - Is quantitative (numbers are really numbers)
  - Ranges from 0 (or 1) to positive infinity
    - Don't have any zeros? Need "zero-truncated/altered" distribution
  - Is predicted using a **log link function** to ensure predicted counts > 0

- Determining the "right" distribution for a count outcome is aided by examining **conditional distribution fit**: Pearson $\chi^2/DF \approx 1$
  - Counts often have more variance (because of positive skewness) than expected by Poisson (in which mean = variance)—this "over-dispersion" can be fixed by adding a "scale" parameter by which variance > mean
  - If you have more zero values than expected, may need to add a "zero-inflation" submodel or switch to a "hurdle" two-submodel variant
  - But both dispersion parameters and zero-inflation models are trying to accommodate for skewness, so you may not need both (check fit to see)