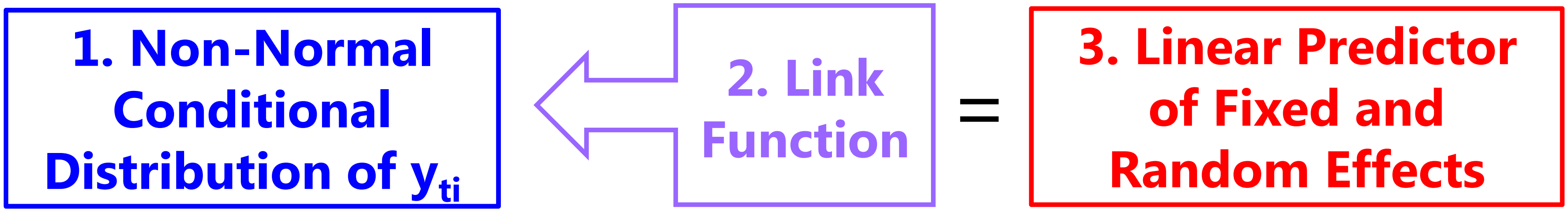
Generalized Linear Models for Count, Skewed, and "If and How Much" Outcomes

- Today's Class:
  - Review of 3 parts of a generalized model
  - Models for discrete count or continuous skewed outcomes
  - Models for two-part discrete or continuous outcomes

3 Parts of Generalized (Multilevel) Models



1. Non-Normal Conditional Distribution of y_{ti}

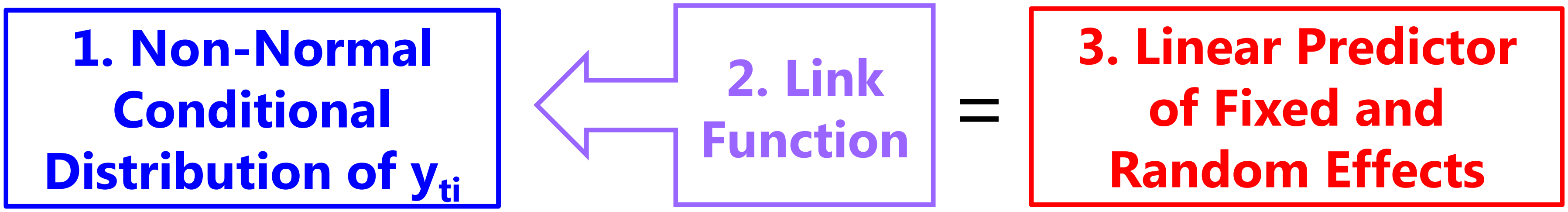
2. Link Function

=

3. Linear Predictor of Fixed and Random Effects

1. **Non-normal conditional distribution of y_{ti}:**
   - General MLM uses a *normal* conditional distribution to describe the y_{ti} variance remaining after fixed + random effects → we called this the level-1 residual variance, which is estimated separately and usually assumed constant across observations (unless modeled otherwise)
   - Other distributions will be more plausible for bounded/skewed $y_{ti,}$ so the ML function maximizes the likelihood using those instead
   - **Why?** To get the most correct **standard errors** for fixed effects
   - Although you can still think of this as *model for the variance*, not all conditional distributions will actually have a separately estimated residual variance (e.g., binary → Bernoulli, count → Poisson)

3 Parts of Generalized (Multilevel) Models



1. Non-Normal Conditional Distribution of y_{ti}

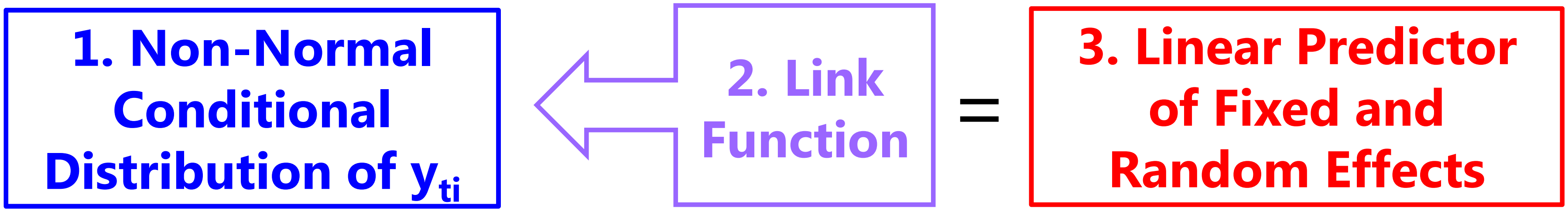
2. Link Function

=

3. Linear Predictor of Fixed and Random Effects

2. **Link Function = $g(\cdot)$:** How the conditional mean to be predicted is transformed so that the model predicts an **unbounded** outcome instead
   - **Inverse link** $g^{-1}(\cdot)$ = how to go back to conditional mean in y_{ti} scale
   - Predicted outcomes (found via inverse link) will then stay within bounds
   - e.g., binary outcome: conditional mean to be predicted is probability of a 1, so the model predicts a linked version (when inverse-linked, the predicted outcome will stay between a probability of 0 and 1)
   - e.g., count outcome: conditional mean is expected count, so the log of the expected count is predicted so that the expected count stays > 0
   - e.g., for normal outcome: an "identity" link function (y_{ti} * 1) is used given that the conditional mean to be predicted is already unbounded…

3 Parts of Generalized (Multilevel) Models



1. Non-Normal Conditional Distribution of y_{ti}

2. Link Function

=

3. Linear Predictor of Fixed and Random Effects

3. **Linear Predictor:** How the fixed and random effects of predictors combine additively to predict a link-transformed conditional mean

- This works the same as usual, except the linear predictor model **directly predicts the link-transformed conditional mean**, which we then convert (via inverse link) back into the original conditional mean
- That way we can still use the familiar "one-unit change" language to describe effects of model predictors (on the linked conditional mean)
- You can think of this as "model for the means" still, but it also includes the level-2 random effects for dependency of level-1 observations
- Fixed effects are no longer determined: they now have to be found through the ML algorithm, the same as the variance parameters

A Taxonomy of Not-Normal Outcomes

- **"Discrete" outcomes**—all responses are **whole** numbers
  - **Categorical variables** in which **values are labels**, not amounts
    - Binomial (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 +∞
    - Poisson or Negative Binomial distributions (usually)
    - 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?
    - Normal-ish? Skewed? Cut off? Mixture of different distributions?

A Revised Taxonomy

- Rather than just separating into discrete vs. continuous, think about models based on their shape AND kinds of data they fit
  - Note: You can use continuous models for discrete data (that only have integers), but not discrete models for continuous data (non-integers)

1. Skewed-looking distributions
   - Discrete: Poisson, Generalized Poisson, Negative Binomial (NB)
   - Continuous: Log-Normal, Beta, Gamma

2. Skewed with a pile of 0's: Becomes **If 0** and **How Much**
   - These models will differ in how they define the "If 0" part
   - Discrete: Zero-Inflated Poisson or NB, Hurdle Poisson or NB
   - Continuous: Two-Part (with normal or lognormal for the how much part)

Models for Count Outcomes

- Counts: non-negative integer unbounded responses
  - e.g., how many cigarettes did you smoke this week?
  - Traditionally uses natural log link so that predicted outcomes stay ≥ 0

- $\mathrm{g}(\bullet)$ $\quad \mathrm{Log}[\mathrm{E}(\mathrm{y_i})] = \mathrm{Log}(\mu_\mathrm{i}) = \mathrm{model}$ → predicts mean of $\mathrm{y_i}$
- $\mathrm{g}^{-1}(\bullet)$ $\mathrm{E}(\mathrm{y_i}) = \exp(\mathrm{model})$ → to un-log it, use $\exp(\mathrm{model})$
  - e.g., if $\mathrm{Log}(\mu_\mathrm{i}) = \mathrm{model}$ provides predicted $\mathrm{Log}(\mu_\mathrm{i}) = 1.098$, that translates to an actual predicted count of $\exp(1.098) = 3$
  - e.g., if $\mathrm{Log}(\mu_\mathrm{i}) = \mathrm{model}$ provides predicted $\mathrm{Log}(\mu_\mathrm{i}) = -5$, that translates to an actual predicted count of $\exp(-5) = 0.006738$

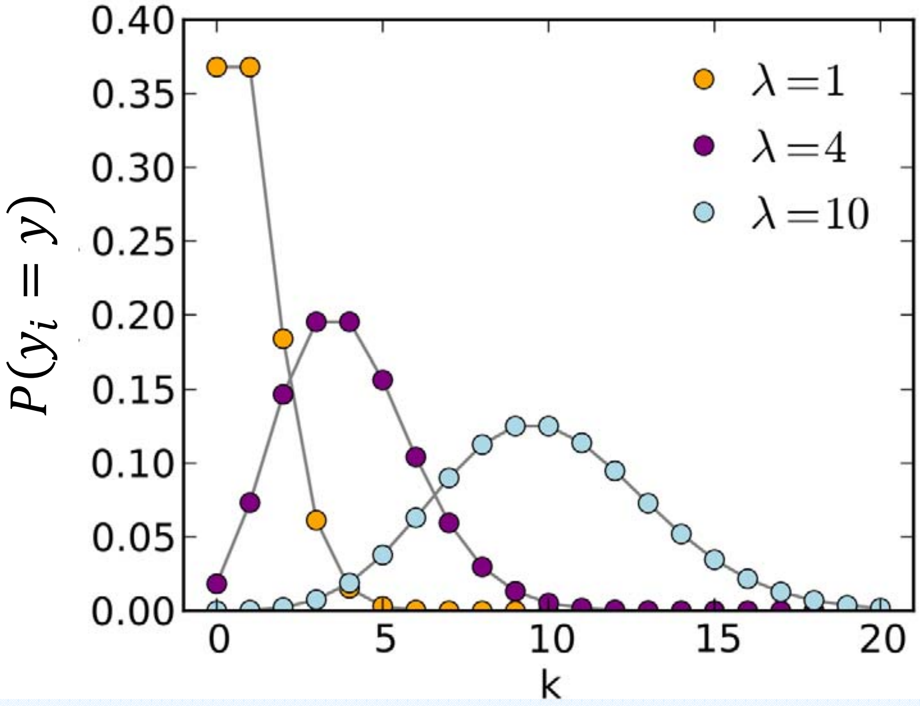
- So that’s how linear model predicts μ_i, the conditional mean for $\mathrm{y_i}$, but what about residual variance?

Poisson Conditional Distribution

- Poisson distribution has one parameter, λ, which is **both its mean and its variance** (so λ = mean = variance in Poisson)
- $f(\mathrm{y_i}|\lambda) = \mathrm{Prob}(\mathrm{y_i} = \mathrm{y}) = \frac{\lambda^{\mathrm{y}} * \exp(-\lambda)}{\mathrm{y}!}$
- PDF: $\mathrm{Prob}(\mathrm{y_i} = \mathrm{y}|\beta_0, \beta_1, \beta_2) = \frac{\mu_i^{\mathrm{y}} * \exp(-\mu_i)}{\mathrm{y}!}$

y! is factorial of y

DIST = POISSON in SAS;
MEPOISSON in STATA



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.

3 potential problems for 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; need to adjust the model)
- **Problem #3: When there are *too many* 0 values**
  - Some 0 values are expected from the Poisson and Negative Binomial models 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 another parameter that allows the variance to exceed the mean… becomes a **Negative Binomial** distribution
  - Says residuals are a mixture of Poisson and gamma distributions, such that λ itself is a random variable with a gamma distribution
  - So expected mean is still given by λ, but the variance will differ from Poisson

- Model: $\mathrm{Log}[\mathrm{E}(\mathrm{y_i})] = \mathrm{Log}(\mu_\mathrm{i}) = \beta_0 + \beta_1\mathrm{X_i} + \beta_2\mathrm{Z_i} + \mathrm{e_i^G}$
- Negative Binomial PDF with a new k dispersion parameter is now:
  - $\mathrm{Prob}(\mathrm{y_i} = \mathrm{y}|\beta_0, \beta_1, \beta_2) = \frac{\Gamma\left(\mathrm{y}+\frac{1}{k}\right)}{\Gamma(\mathrm{y}+1)*\Gamma\left(\frac{1}{k}\right)} * \frac{(k\mu_\mathrm{i})^\mathrm{y}}{(1+k\mu_\mathrm{i})^{\mathrm{y}+\frac{1}{k}}}$
  - $\boldsymbol{k}$ is dispersion, such that $\mathrm{Var}(\mathrm{y_i}) = \mu_\mathrm{i} + \boldsymbol{k}\mu_\mathrm{i}^2$
  - Can test whether $k > 0$ via −2LL test, although LL for $k = 0$ is undefined

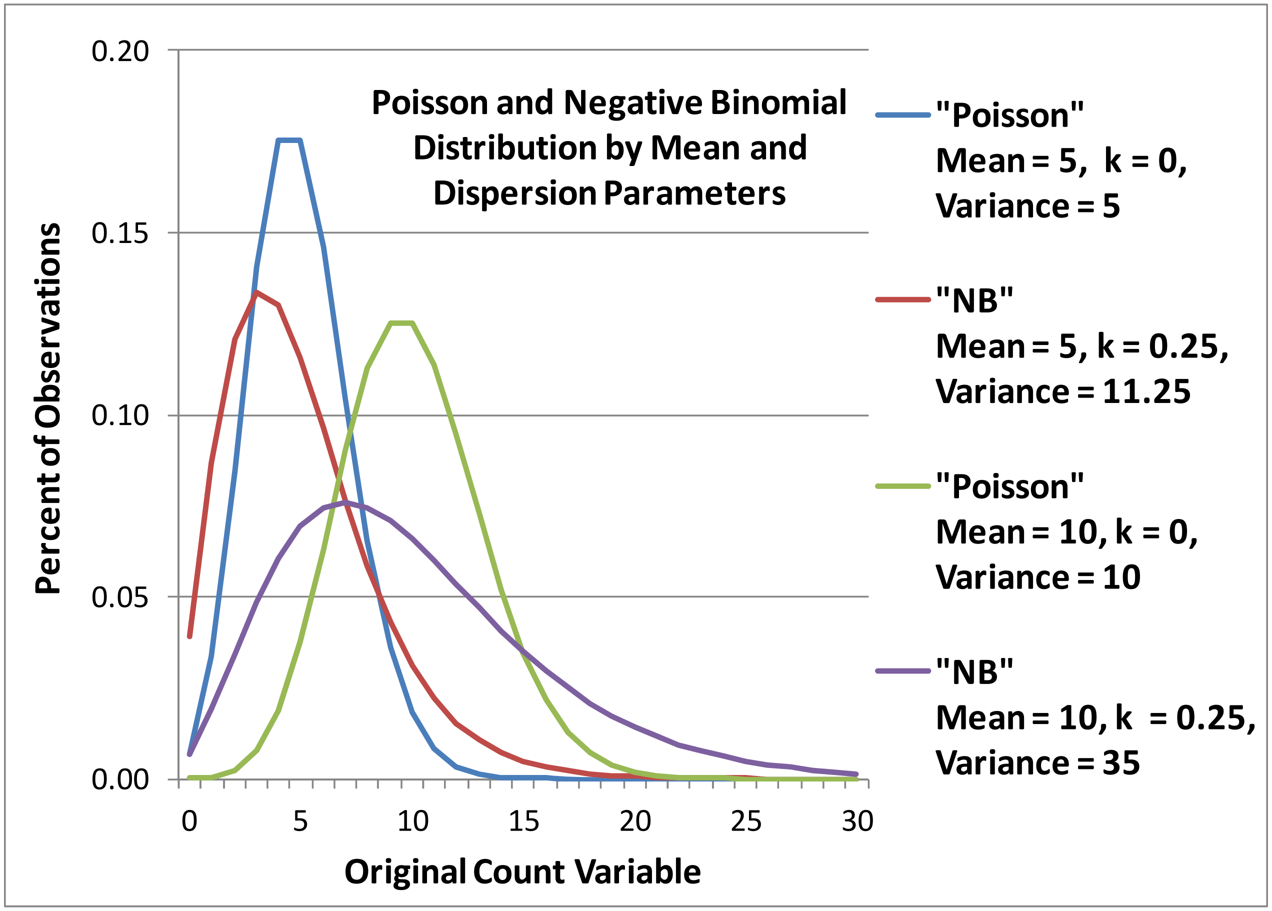
DIST = NEGBIN in SAS;
MENBREG in STATA

So ≈ Poisson if $k = 0$

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

GPOISSON
in STATA

Negative Binomial (NB) = "Stretchy" Poisson…



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

$$\text{Var}(y_i) = \lambda + k\lambda^2$$

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 $-2\left(LL_{Poisson} - LL_{NegBin}\right) > 3.84$ for $df = 1$? Then $p < .05$, keep NB

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

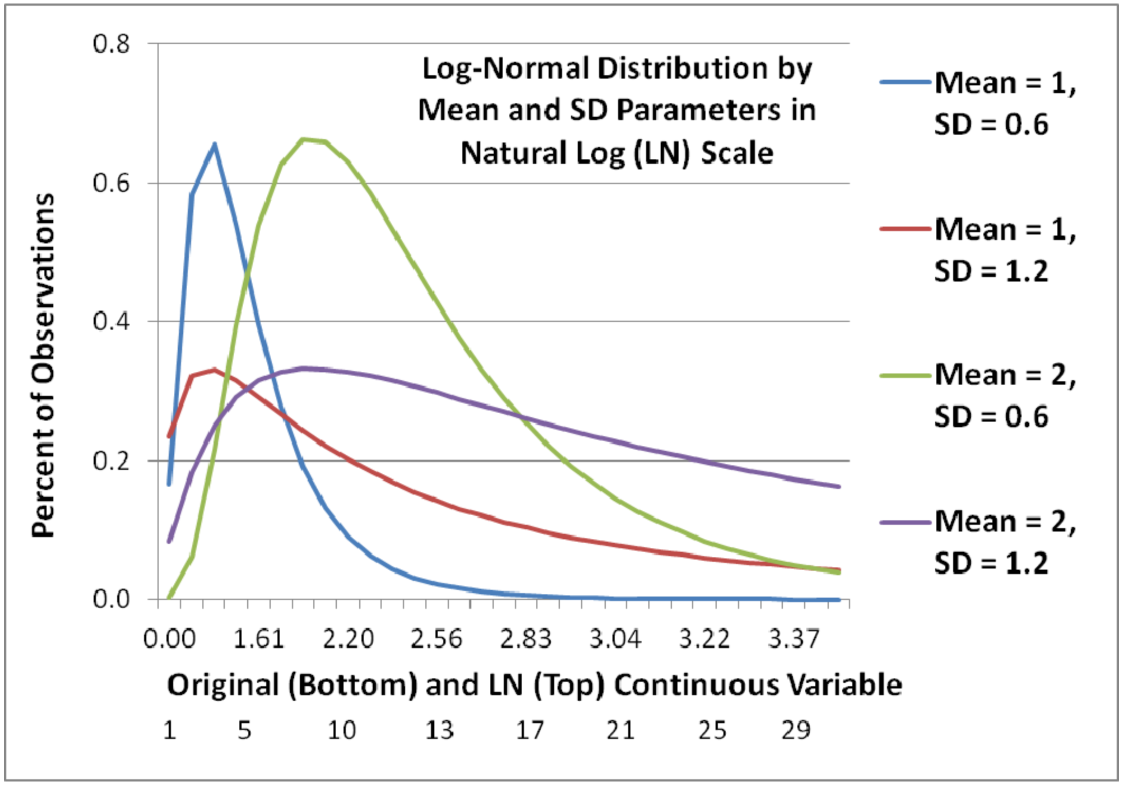
- Poisson PDF was: $\mathrm{Prob}(y_i = y|\mu_i\,) = \frac{\mu_i^y * \exp(-\mu_i)}{y!}$
- Zero-Truncated Poisson PDF is:
  - $\mathrm{Prob}\big(y_i = y|\mu_{i,} y_i > 0\big) = \frac{\mu_i^y * \exp(-\mu_i)}{y![1 - \exp(-\mu_i)]}$
  - $\mathrm{Prob}(y_i = 0) = \exp(-\mu_i)$, so $\mathrm{Prob}(y_i > 0) = 1 - \exp(-\mu_i)$
  - Divides by probability of non-0 outcomes so probability still sums to 1

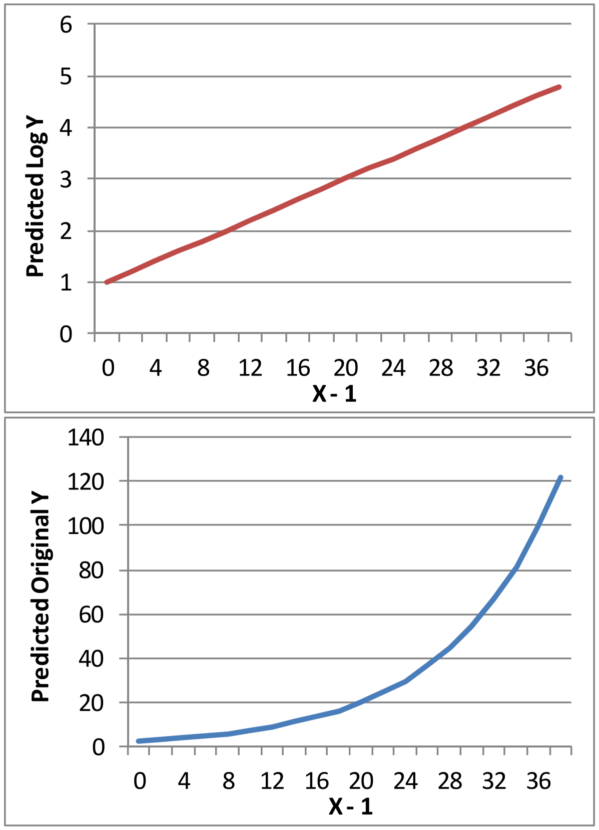
Software for Discrete Outcomes

- 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 in SAS:**
  - 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, multivariate generalized models in SAS 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

- **Up next: models for skewed continuous outcomes…**

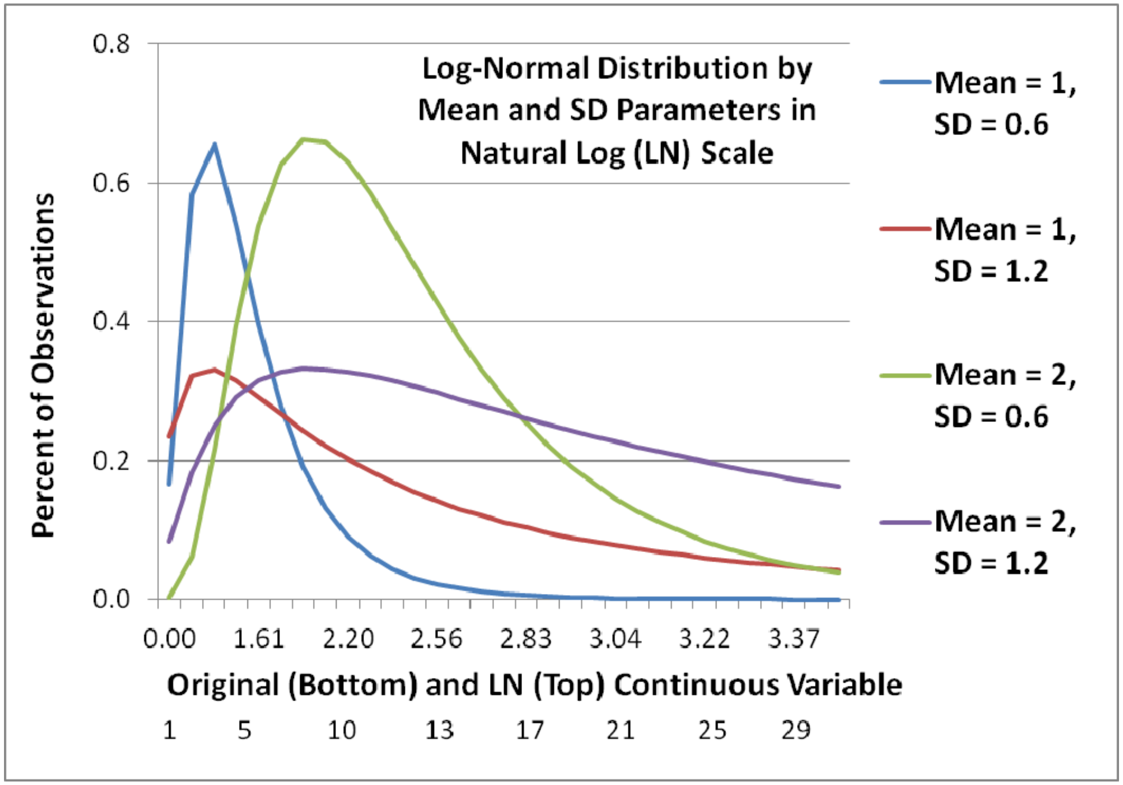
Log-Normal Distribution (Link=Identity)

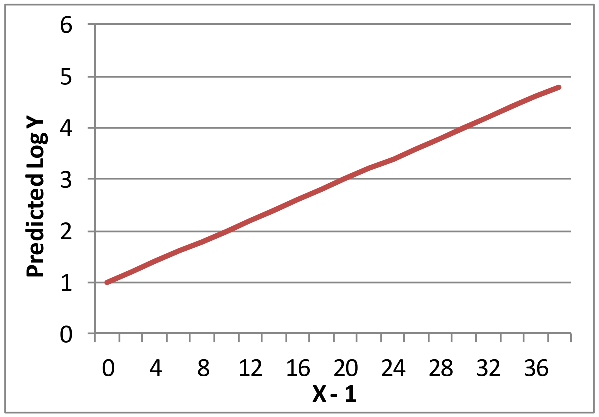


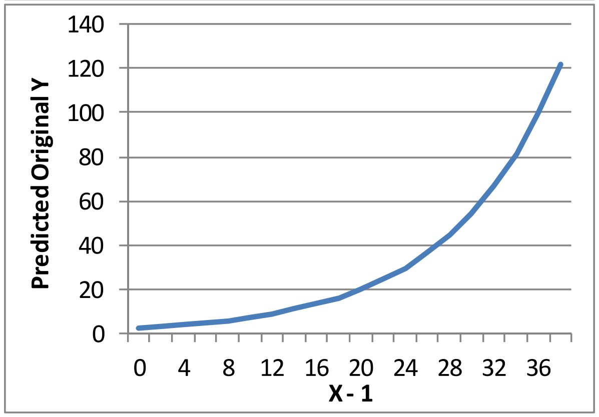


- $e_i \sim \text{LogNormal}(0, \sigma_e^2)$ → **log** of residuals is normal
  - Is same as log-transforming your outcome in this case…
  - The log link keeps the predicted values positive, but slopes then have an exponential (not linear) relation with original outcome

Log-Normal Distribution (Link=Identity)

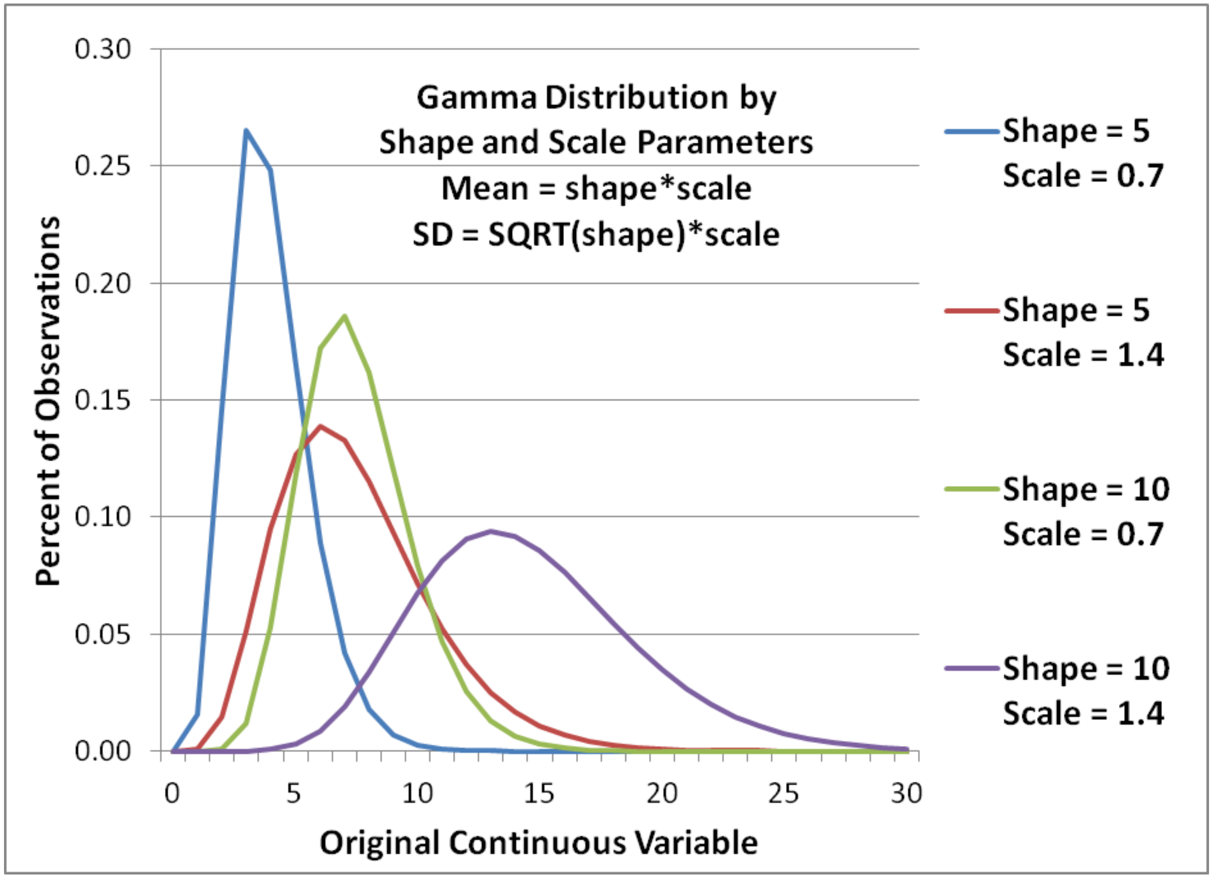


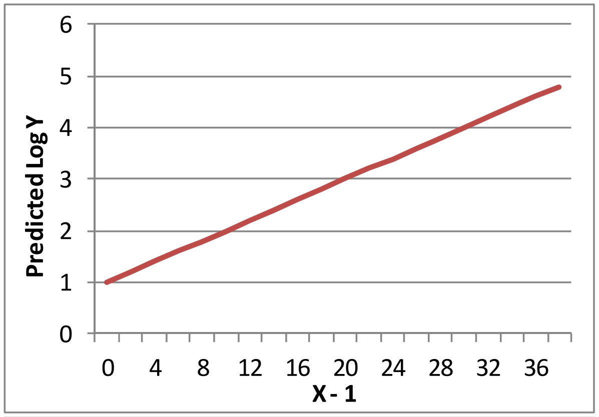


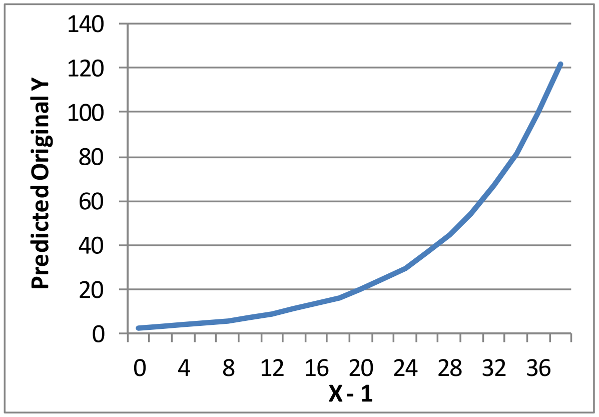


- GLIMMIX parameterization gives μ (= intercept) and $scale$ = (variance) to convert back into original data as follows:
  - $\text{Mean(Y)} = \exp(\mu) * \sqrt{\exp(scale)}$
  - $\text{Variance(Y)} = \exp(2\mu) * \exp(scale) * [\exp(scale) - 1]$

Gamma Response Distribution







- GLIMMIX parameterization with LINK=LOG gives μ (= intercept) and $scale$ = (dispersion) to convert into original data as follows:
  - $\text{Mean(Y)} = \exp(\mu) \approx (\text{shape*scale})$
  - $\text{Variance(Y)} = \exp(\mu)^2 * dispersion \approx (\text{shape} * \text{scale}^2)$

Software for Continuous Outcomes

- There are many choices for modeling not-normal ***continuous*** outcomes (that can include non-integer values); most use either an identity or log link
- **Single-level, univariate generalized models in SAS (not in Mplus):**
  - GENMOD: DIST= (and default link): Gamma (Inverse), Geometric (Log), Inverse Gaussian (Inverse^2), Normal (Identity)
  - FMM: DIST= (and default link): Beta (Logit), Betabinomial (Logit), Exponential (Log), Gamma (Log), Normal (Identity), Geometric (Log), Inverse Gaussian (Inverse^2), LogNormal (Identity), TCentral (Identity), Weibull (Log)
- **GLM in STATA** has gamma and inverse Gaussian distributions
- **Multilevel or multivariate generalized models in SAS via GLIMMIX:**
  - Beta (Logit), Exponential (Log), Gamma (Log), Geometric (Log), Inverse Gaussian (Inverse^2), Normal (Identity), LogNormal (Identity), TCentral (Identity)
  - BYOBS, which allows multivariate models by which you specify DV-specific link functions and distributions estimated simultaneously (e.g., two-part)
- NLMIXED can also be used to fit any user-defined model

Modeling Not-Normal Outcomes

- Previously we examined models for skewed distributions
  - Discrete: Poisson, Generalized Poisson, Negative Binomial (NB)
  - Continuous: Log-Normal, Gamma (also Beta from before)

- Now we will see additions to these models when the outcome also has a pile of 0's: Model becomes **If 0** and **How Much**
  - These models will differ in how they define the "If 0" part
  - Discrete → Zero-Inflated: Poisson, Generalized Poisson, or NB;
    Hurdle: Poisson, Generalized Poisson, or NB
  - Continuous → Two-Part (with normal or lognormal for how much)
  - Many of these can be estimated directly in Mplus or SAS GLIMMIX, but some will need to be programmed in SAS GLIMMIX or NLMIXED
  - More options for single-level data in SAS PROC FMM and in STATA

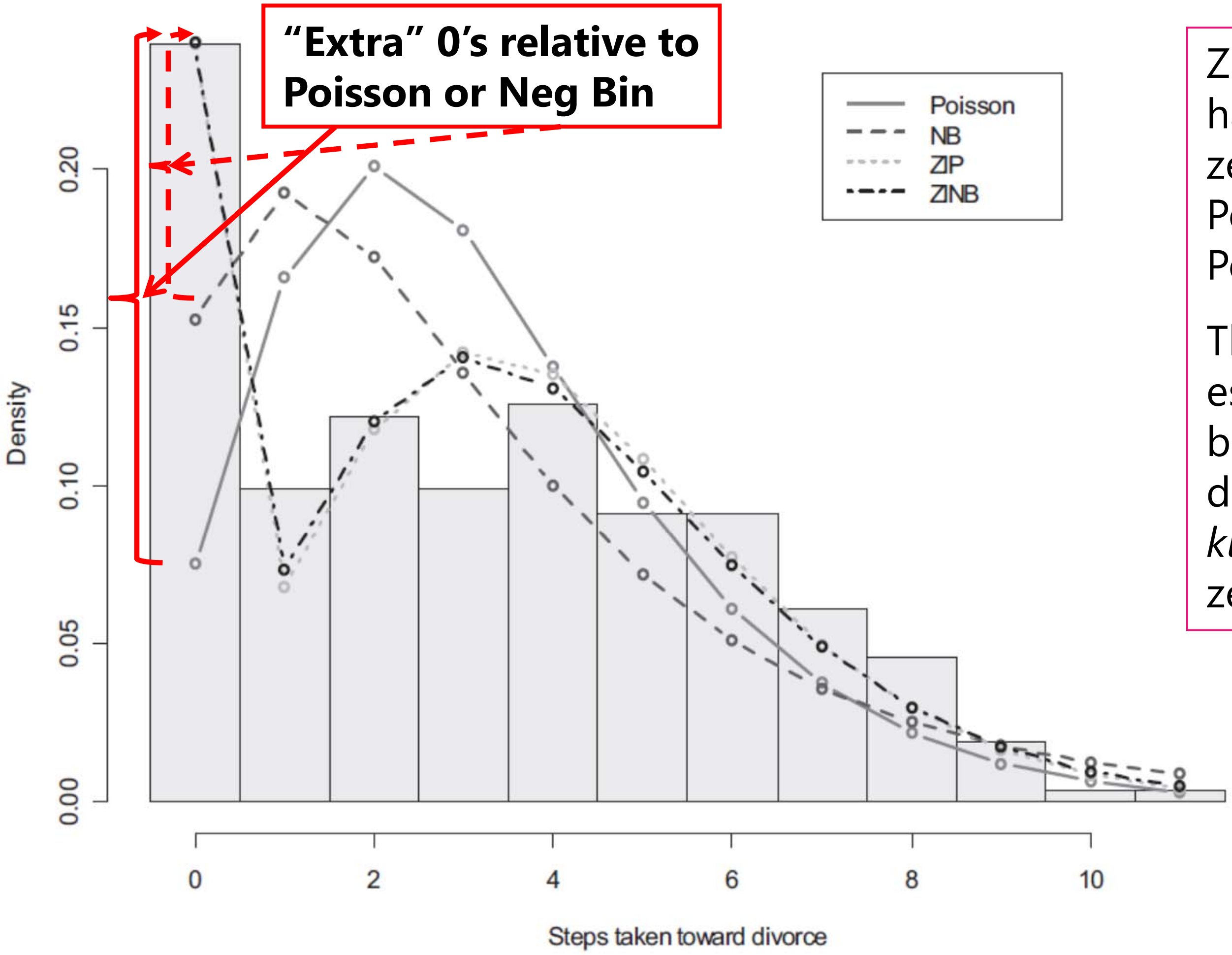
Problem #3: Too many 0 values, Option #1

- "**Zero-Inflated**" Poisson (DIST=ZIP) or NB(DIST=ZINB) in SAS GENMOD or Mplus; ZIP/ZI Generalized Poisson (ZIGP) in STATA
  - Distinguishes **two kinds of 0 values**: **expected** and **inflated** ("structural") through a mixture of distributions (Bernoulli + Poisson/NB)
  - 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: $\mathrm{Logit}[p(\mathrm{y_i} = \mathrm{extra}\ 0)] = \ \beta_{01} + \beta_{11}\mathrm{X_i} + \beta_{21}\mathrm{Z_i}$
  - Predict being an extra 0 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: $\mathrm{Log}[\mathrm{E}(\mathrm{y_i})] = \beta_{02} + \beta_{12}\mathrm{X_i} + \beta_{22}\mathrm{Z_i}$
  - Predict rest of counts <u>(including 0's)</u> using Link = Log, Distribution = Poisson

Example of Zero-Inflated Outcomes



"Extra" 0's relative to Poisson or Neg Bin

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 have extra "structural zeros" not expected from Poisson or NB ("stretched Poisson") distributions.

This can be tricky to estimate and interpret because the model distinguishes between *kinds of zeros* rather than zero or not…

Image borrowed from Atkins & Gallop, 2007

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

- The Zero-Inflated models get put back together as follows:
  - $\mathrm{\omega_i}$ is the predicted probability of being an extra 0, from:
  $$\mathrm{\omega_i} = \frac{\exp[\mathrm{Logit}[p(\mathrm{y_i} = \mathrm{extra}\ 0)]]}{1 + \exp[\mathrm{Logit}[p(\mathrm{y_i} = \mathrm{extra}\ 0)]]}$$
  - $\mathrm{\mu_i}$ is the predicted count for the rest of the distribution, from:
  $$\mathrm{\mu_i} = \exp[\mathrm{Log}(\mathrm{y_i})]$$
  - ZIP: $\mathrm{Mean\ (original\ y_i)} = (1 - \mathrm{\omega_i})\mathrm{\mu_i}$
  - ZIP: $\mathrm{Variance(original\ y_i)} = \mathrm{\mu_i} + \frac{\mathrm{\omega_i}}{(1-\mathrm{\omega_i})}\mathrm{\mu_i^2}$
  - ZINB: $\mathrm{Mean\ (original\ y_i)} = (1 - \mathrm{\omega_i})\mathrm{\mu_i}$
  - ZINB: $\mathrm{Variance(original\ y_i)} = \mathrm{\mu_i} + \left[\frac{\mathrm{\omega_i}}{(1-\mathrm{\omega_i})} + \frac{\mathrm{k}}{1-\mathrm{\omega_i}}\right]\mathrm{\mu_i^2}$

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 a mixture of 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: $\mathrm{Logit}[p(\mathrm{y_i} = 0)] = \beta_{01} + \beta_{11}\mathrm{X_i} + \beta_{21}\mathrm{Z_i}$
  - Predict being **<u>any 0</u>** using Link = Logit, Distribution = Bernoulli
  - Don't have to specify predictors for this part, can simply allow it to exist

- Submodel 2: $\mathrm{Log}[\mathrm{E}(\mathrm{y_i})|\mathrm{y_i} > 0] = \beta_{02} + \beta_{12}\mathrm{X_i} + \beta_{22}\mathrm{Z_i}$
  - Predict rest of **<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 into each submodel and estimate separately (in STATA)

Two-Part Models for Continuous Outcomes

- A two-part model is an analog to hurdle models for zero-inflated count outcomes (and could be used with count outcomes, too)
  - Explicitly **separates 0 from non-0 values** through a mixture of distributions (Bernoulli + Normal or LogNormal)
  - Creates two submodels to predict "if any not 0" and "if not 0, how much"?
    - Easier to think about in terms of prediction (in my opinion)

- Submodel 1: $\mathrm{Logit}[p(\mathrm{y_i} > 0)] = \beta_{01} + \beta_{11}\mathrm{X_i} + \beta_{21}\mathrm{Z_i}$
  - Predict being **any not 0** using Link = Logit, Distribution = Bernoulli
  - Usually do specify predictors for this part

- Submodel 2: $(\mathrm{y_i}|\mathrm{y_i} > 0) = \beta_{02} + \beta_{11}\mathrm{X_i} + \beta_{21}\mathrm{Z_i}$
  - Predict rest of **positive amount** using Link = Identity, Distribution = Normal or Log-Normal (often rest of distribution is skewed, so log works better)

- Two-part is in Mplus, but parts can be estimated separately in SAS/STATA
  - Logit of 0/1 for "if part" + log-transformed DV for "how much" part
  - Is related to "tobit" models for censored outcomes (for floor/ceiling effects)

Pile of 0's Taxonomy

- What kind of **amount** do you want to predict?
  - Discrete: Count → Poisson
  - Stretchy Count → Generalized Poisson or Negative Binomial
  - Continuous: Normal, Log-Normal, Gamma

- What kind of **If 0** do you want to predict?
  - Discrete: Extra 0 beyond predicted by amount?
    → Zero-inflated Poisson or Zero-inflated Negative Binomial
  - Discrete: Any 0 at all?
    → Hurdle Poisson or Hurdle Negative Binomial
  - Continuous: Any 0 at all?
    → Two-Part with Continuous Amount (see above)
  - Note: Given the same amount distribution, these alternative ways of predicting 0 will result in the same empty model fit

Comparing Generalized Models

- Whether or not a dispersion 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—see SAS macro online)
  - 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
  - You can compute predicted values under different models to see how reasonably they approximate the data for some unofficial guidance

Generalized Models: Summary

- There are many options for "amount" variables whose residuals may not be normally distributed
  - Discrete: Poisson, Negative Binomial
  - Continuous: Lognormal, Gamma, Beta
  - Too many 0's: Zero-inflated or hurdle for discrete; two-part

- Multivariate and multilevel versions of all the generalized models we covered *can* be estimated…
  - But it's harder to do and takes longer due to numeric integration (trying on all combinations of random effects at each iteration)
  - But there are fewer ready-made options for modeling differential variance/covariance across DVs (no easy R matrix structures in true ML)

- Program documentation will always be your friend to determine exactly what a given model is doing!