Generalized Linear Mixed Models

What, Why and How

Timeline for GLMMs

- Mixed model issues appear in literature in 1930s (Yates)
- Mixed models named as such appear in 1950s
- Widespread applicability of mixed models not fully appreciated until 1980s
- Viable mixed model software available in 1990s
- Generalized linear models appear in literature in 1970s
- Generalized + mixed linear model literature in 1990s
- Viable GLMM software appears mid-to-late 2000s
- Appreciation of widespread applicability of GLMM still a work in progress

Setting for Statistical Models

Classic Statistical Model Format:
response =
systematic + random/residual

		systematic/explanatory		random/residual	
	response variable	Categorical (ANOVA) $\mu + \tau_i$	continuous (regression) $\beta_0 + \beta_1 X$	model effects	Residual error structure (e.g. serial/spatial)
Gaussian (normal)					
Non-Gaussian	discrete proportion binomial multinomial				
	continuous proportion beta				
	count Poisson negative Binomial				
	time to event				
	etc....				

Setting for Statistical Models

Classic Statistical Model Format:
response =
systematic + random/residual

		systematic/explanatory		random/residual	
		Categorical	**continuous**		Residual error structure (e.g. serial/spatial)
response variable		$\mu + \tau_i$	$\beta_0 + \beta_1 X$	model effects	
Gaussian (normal)		**Linear Model classical ANOVA and regression**			
Non-Gaussian	**discrete proportion** **binomial** **multinomial**				
	continuous proportion **beta**				
	count **Poisson** **negative Binomial**				
	time to event				
	etc....				

Setting for Statistical Models

Classic Statistical Model Format:
response =
systematic + random/residual

		systematic/explanatory		random/residual	
		Categorical	continuous		Residual error structure
response variable		$\mu + \tau_i$	$\beta_0 + \beta_1 X$	model effects	(e.g. serial/spatial)
Gaussian (normal)		**Linear Mixed Model (LMM)**			
Non-Gaussian	**discrete proportion** binomial multinomial				
	continuous proportion beta				
	count Poisson negative Binomial				
	time to event				
	etc....				

Setting for Statistical Models

Classic Statistical Model Format:
response =
systematic + random/residual

		systematic/explanatory		random/residual	
	response variable	Categorical $\mu + \tau_i$	continuous $\beta_0 + \beta_1 X$	model effects	Residual error structure (e.g. serial/spatial)
Gaussian (normal)		**Generalized Linear Mixed Model (GLMM)**			
Non-Gaussian	**discrete proportion** binomial multinomial				
	continuous proportion beta				
	count Poisson negative Binomial				
	time to event				
	etc....				

Defining Elements of GLMM

- **Link function = linear predictor**
- $\boldsymbol{\eta = g(\mu) = X\beta + Zb}$
- $\boldsymbol{\beta}$ **denotes fixed (ANOVA or regression) effects**
- $\boldsymbol{b}$ **denotes random model effects, assume** $\boldsymbol{b \sim N(0, G)}$
- $\boldsymbol{y}$ **denotes observations**
- $\boldsymbol{y|b \sim \mathfrak{D}(\mu, \Sigma)}$
- **linear predictor is mixed model;** $\boldsymbol{\mathfrak{D}}$ **and link function accommodate non-Gaussian data**

Why GLMMs - Motivating Examples

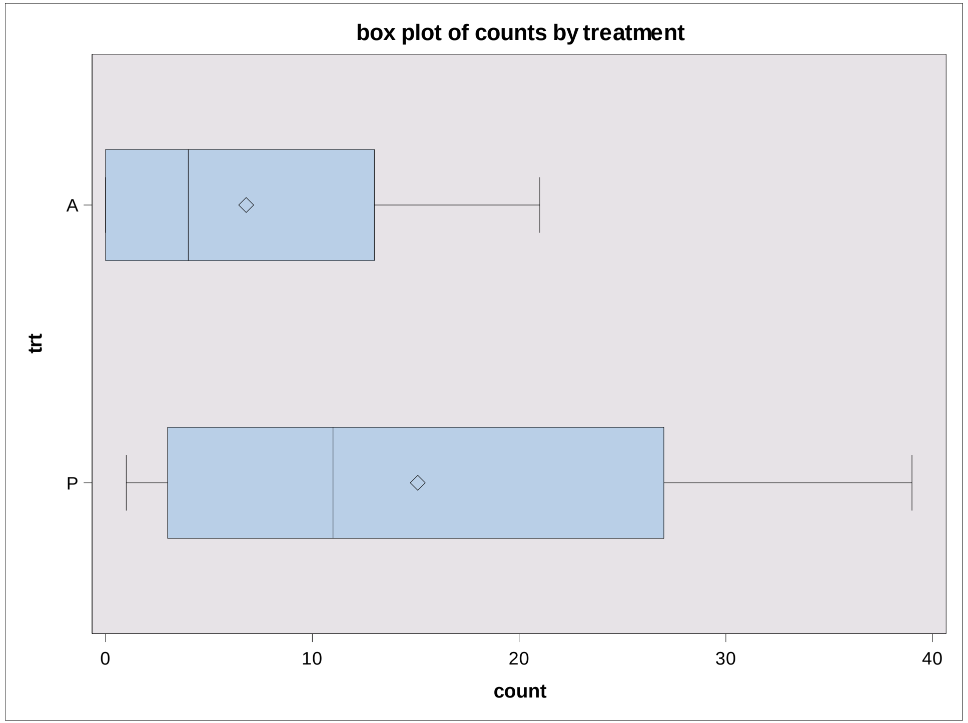
- Main issues are target of inference & accuracy of inferential statistics
- Example 1
- Paired Comparison Experiment:
  - a.k.a. Randomized Complete Block Design
  - 10 Pairs / Blocks
  - 2 Treatments – “Treatment 0”, “Treatment 1”

    e.g. **“control”** & **“test”**; **“A” & “P”**
  - Response: **count**

    e.g. number of events / claims / defects = 0,1,2,...

Example 1: The Data

Obs	clinic	trt	count
1	1	A	7
2	1	P	5
3	2	A	1
4	2	P	1
5	3	A	13
6	3	P	15
7	4	A	5
8	4	P	7
9	5	A	3
10	5	P	32
11	6	A	0
12	6	P	1
13	7	A	21
14	7	P	21
15	8	A	0
16	8	P	3
17	9	A	18
18	9	P	39
19	10	A	0
20	10	P	27

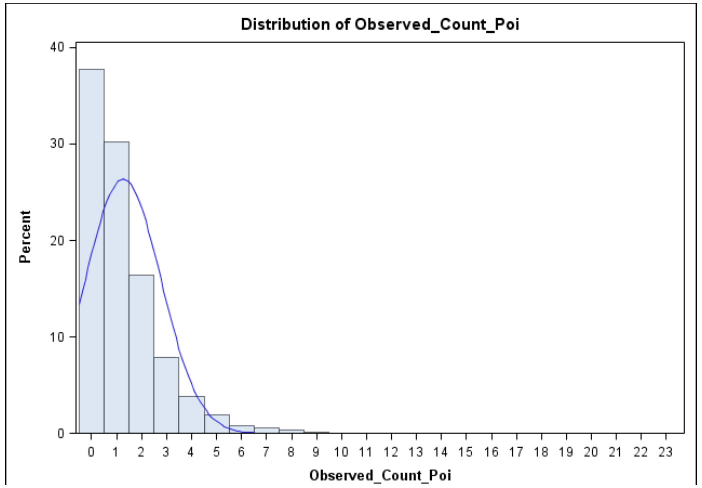


Treatment	Mean	Variance	Median
A	6.80	61.7	7.86
P	15.1	193.9	13.92

Count Distributions?

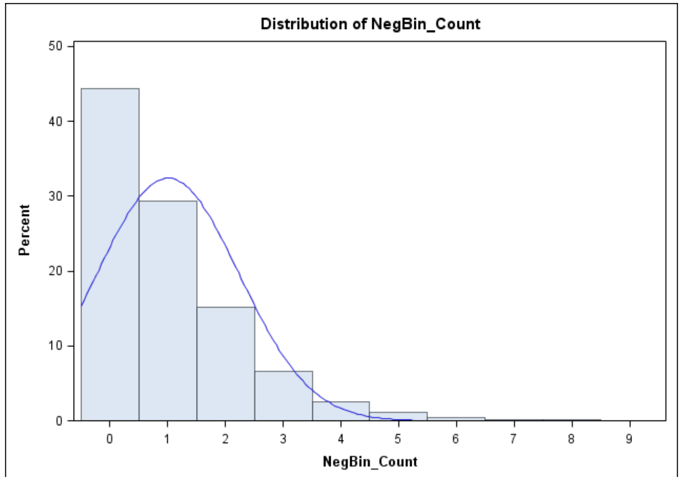
Poisson

$\lambda = 1$



Negative Binomial

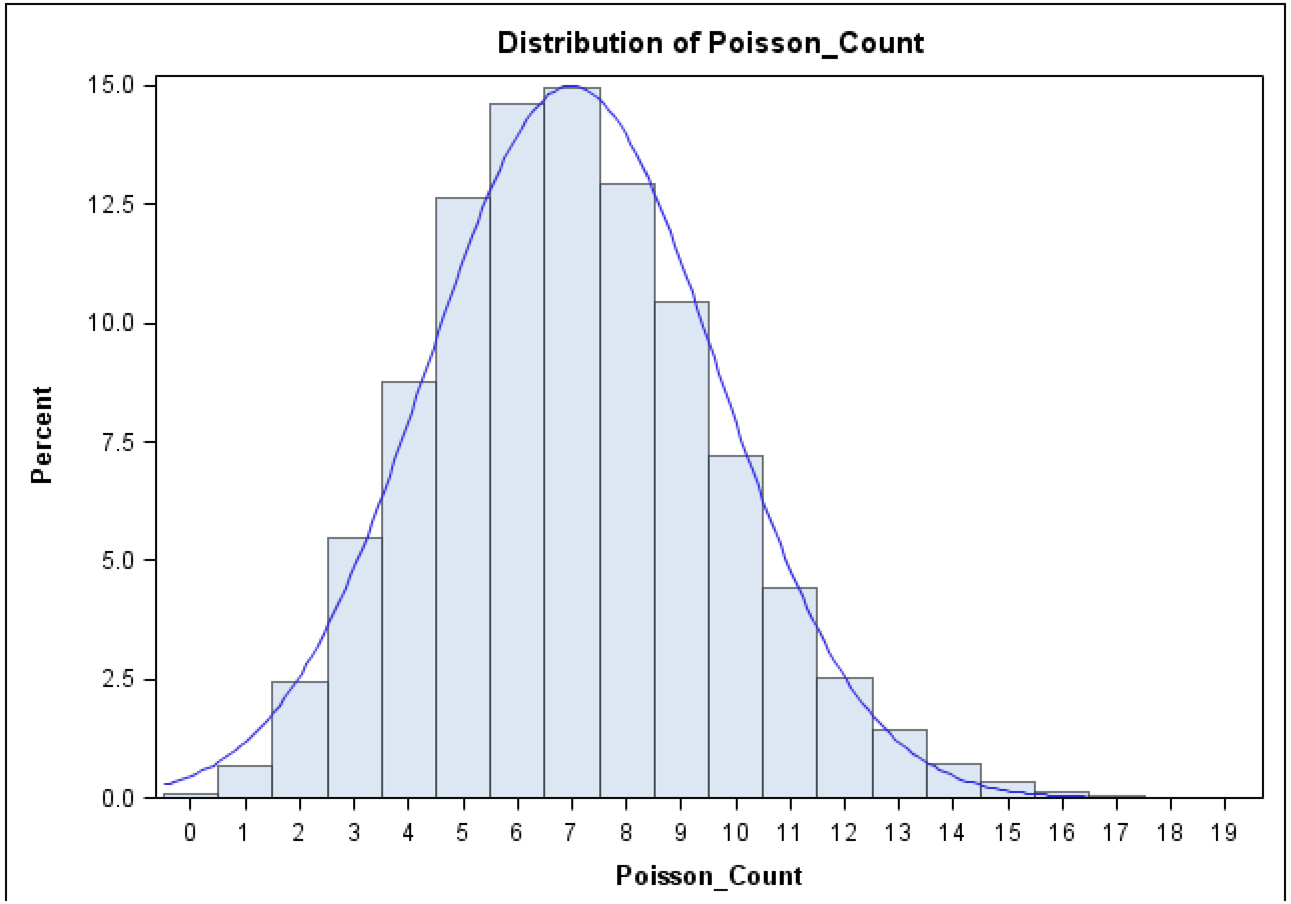
$\lambda = 1; \varphi = 0.5$



Count Distributions?

Poisson

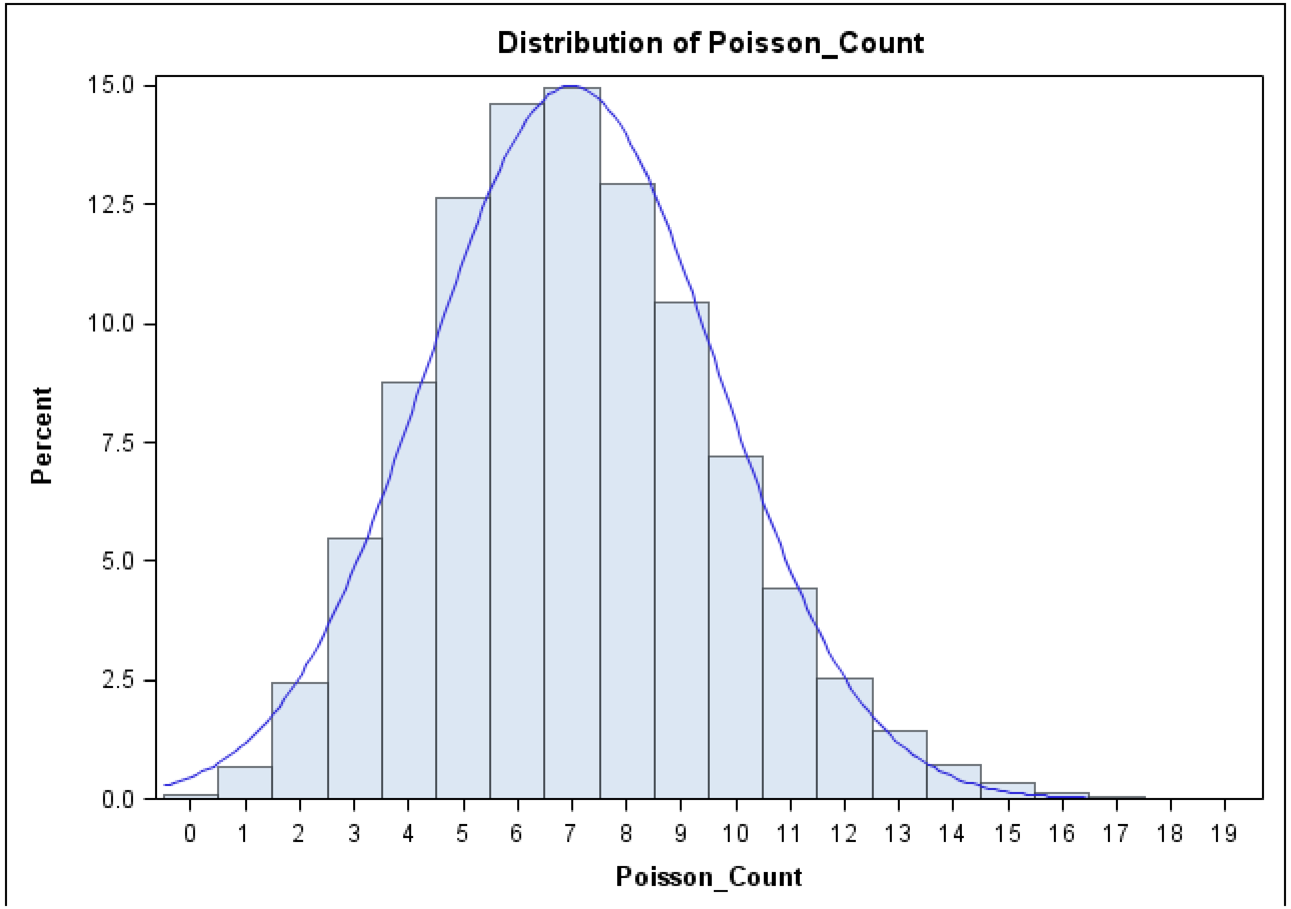
$\lambda = 7$



Count Distributions?

Poisson

$\lambda = 7$



Doesn’t this imply count ~ approx Normal $\Rightarrow$ **ANOVA** with **count** as response variable okay?

Linear Model for RCBD Count Data

- ❑ ANOVA ➔ “general” linear model for RCBD
- ❑ Model: count = intercept + trt + block + resid

$$count_{ij} = \mu + \tau_i + b_j + e_{ij}$$

- ❑ You can implement ANOVA using linear mixed model (LMM) software

Example 1 - Analysis 1 – ANOVA

Source	d.f.	MS	F	Pr>F
block	9	180.9		
treatment	1	344.5	4.61	0.0603
error	9	74.7		
Total	19			

Estimates of Treatment Means

Treatment	Estimate	Std Error
A	6.8	2.73
P	15.1	2.73

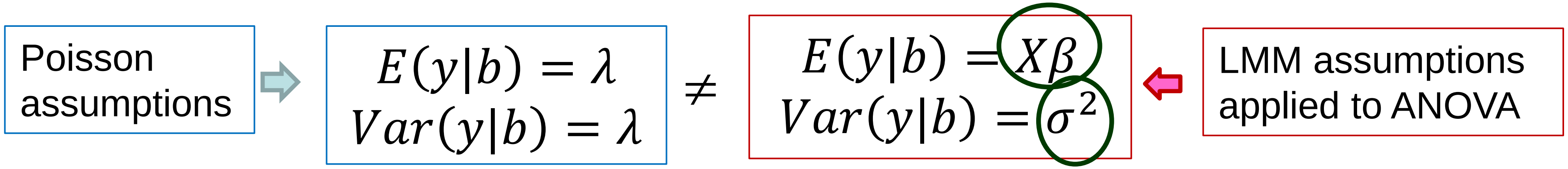
??Poisson → mean=variance??

Neg Bin → mean ∝ variance

Problems using LMM ($\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{e}$) for count data

- Assumes $\mathbf{X}\widehat{\boldsymbol{\beta}}$ estimates $E(y) = \lambda$
- $\hat{\lambda}$ must be ≥ 0
- No guarantee $\mathbf{X}\hat{\boldsymbol{\beta}} \geq 0$

  e.g. regression provides easy examples
- Logical issues



Poisson assumptions ⇨

$$E(y|b) = \lambda \quad Var(y|b) = \lambda$$

$\neq$

$$E(y|b) = X\beta \quad Var(y|b) = \sigma^2$$

⇦ LMM assumptions applied to ANOVA

- with count data, “Residual” has no meaning
- We need a better approach

Possible Models – Count Data

- Normal Approximation
  - $y_{ij} = \mu + \tau_i + b_j + e_{ij}; b_j \ iid \ N(0, \sigma_b{}^2); e_{ij} \ iid \ N(0, \sigma^2)$
  - standard ANOVA model
- Variance Stabilizing Transformation
  - $\log(y_{ij}) = \mu + \tau_i + b_j + e_{ij}$ -- $log(y_{ij} + 1)$ if there are zeros
  - $\sqrt{y_{ij} + {}^3/_8} = \mu + \tau_i + b_j + e_{ij}$
  - variance stabilizing transformation – standard pre-GLMM
- Poisson Generalized Linear Mixed Model (GLMM)
  - $y_{ij}|b_j{\sim}Poisson(\lambda_{ij}); b_j \ iid \ N(0, \sigma_b{}^2)$
  - $\eta_{ij} = log(\lambda_{ij})$=$\eta + \tau_i + b_j$ “Naive model”
  - $\eta_{ij} = log(\lambda_{ij})$=$\eta + \tau_i + b_j + u_{ij}; u_{ij} \ iid \ N(0, \sigma_u^2)$ better – we’ll see why

Determining an Appropriate Model

- Types of Blocked Designs
  - Multi-location, multi-center, multi-clinic
  - Matched pairs
  - Before and after on same subject
  - Field plots
  - etc.

motivating a better approach

- Visualization

Treatment design	
Trt 1	Trt 2

“Experiment” (study) design		
Block	Unit	
1		
2		
3		
4		
….		
10		

Full design		
Block	Unit	
1	trt 1	trt 2
2	trt 2	trt 1
3	trt 2	trt 1
4	trt 1	trt 2
….		
10	trt 2	trt 1

Repurposed ANOVA

Experiment		Treatment		Combined	
Source	d.f.	Source	d.f.	Source	d.f.
block	9			block	9
		trt	2-1=1	trt	1
unit(block)	10×(2-1)=10	"parallels" (Yates, Fisher, 1935)	18	unit(block) | trt a.k.a. "residual" a.k.a "block x trt"	10-1=9
Total	19	Total	19	Total	19

Repurposed ANOVA and Sensible Model

sensible model → one-to-one ANOVA effect – model parameter match

LMM Linear predictor for blocked design includes **block** and **treatment** effect
However, LMM also accounts for unit variation. If GLMM mimics LMM literally, **trouble**

combined		**model**		
Source	**d.f.**	**LMM**	**naive GLMM**	**GLMM accounting for unit effect**
block	9	b_j	b_j	b_j
treatment	1	τ_j	τ_j	τ_j
unit(block) block x trt “residual”	9	e_{ij} or σ^2	**→ overdispersion likely**	bt_{ij} (or u_{ij})
total	**19**			

overdispersion: model fails to adequately account for variation in data
usual consequence: confidence intervals too narrow; type I error rate ↑

Repurposed ANOVA to determine appropriate model

Combined	
Source	d.f.
block	9
trt	1
unit(block) | trt a.k.a. “residual” a.k.a “block x trt”	10-1=9
Total	19

$b_j \sim N\left(0, \sigma_b^2\right)$

$unit_{ij} \equiv bt_{ij} \sim N(0, \sigma_u^2)$

resulting linear predictor

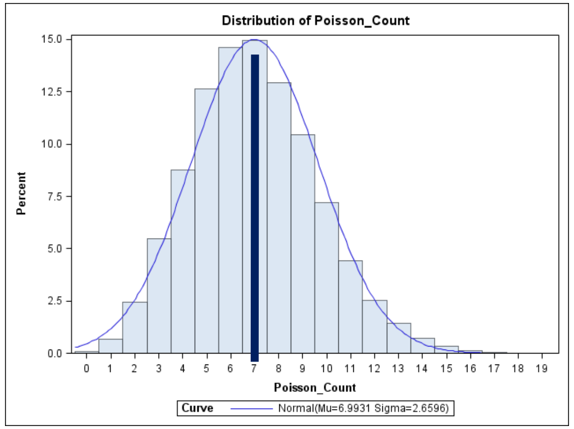
$$\eta_{ij} = log\left(\lambda_{ij}\right)$$
$$= \eta + \tau_i + b_j + bt_{ij}$$

resulting estimate of λ_i is $exp(\hat{\eta} + \hat{\tau}_i)$ (called **inverse link**)

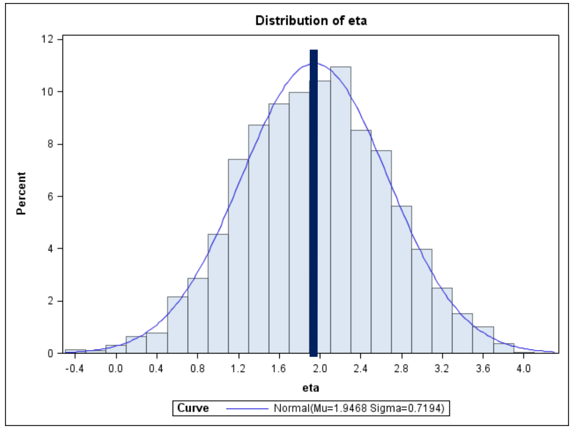
$$y_{ij} | b_j, bt_{ij} \sim Poisson\left(\lambda_{ij}\right)$$

Distributions imply Target of Inference

$$Y \mid b \sim$$
$$Poisson\left(\lambda = 7\right)$$



$$b \sim NI\left(0,\sigma_b^2\right)$$
$$g\left(\lambda \mid b\right) = \eta + \tau_i + b_j + \left(bt\right)_{ij}$$
$$g\left(\lambda \mid b\right) = \eta = \log\left(\lambda \mid b\right)$$
$$\lambda = 7 \Rightarrow \eta \sim N\left(1.95,\sigma_b^2\right)$$

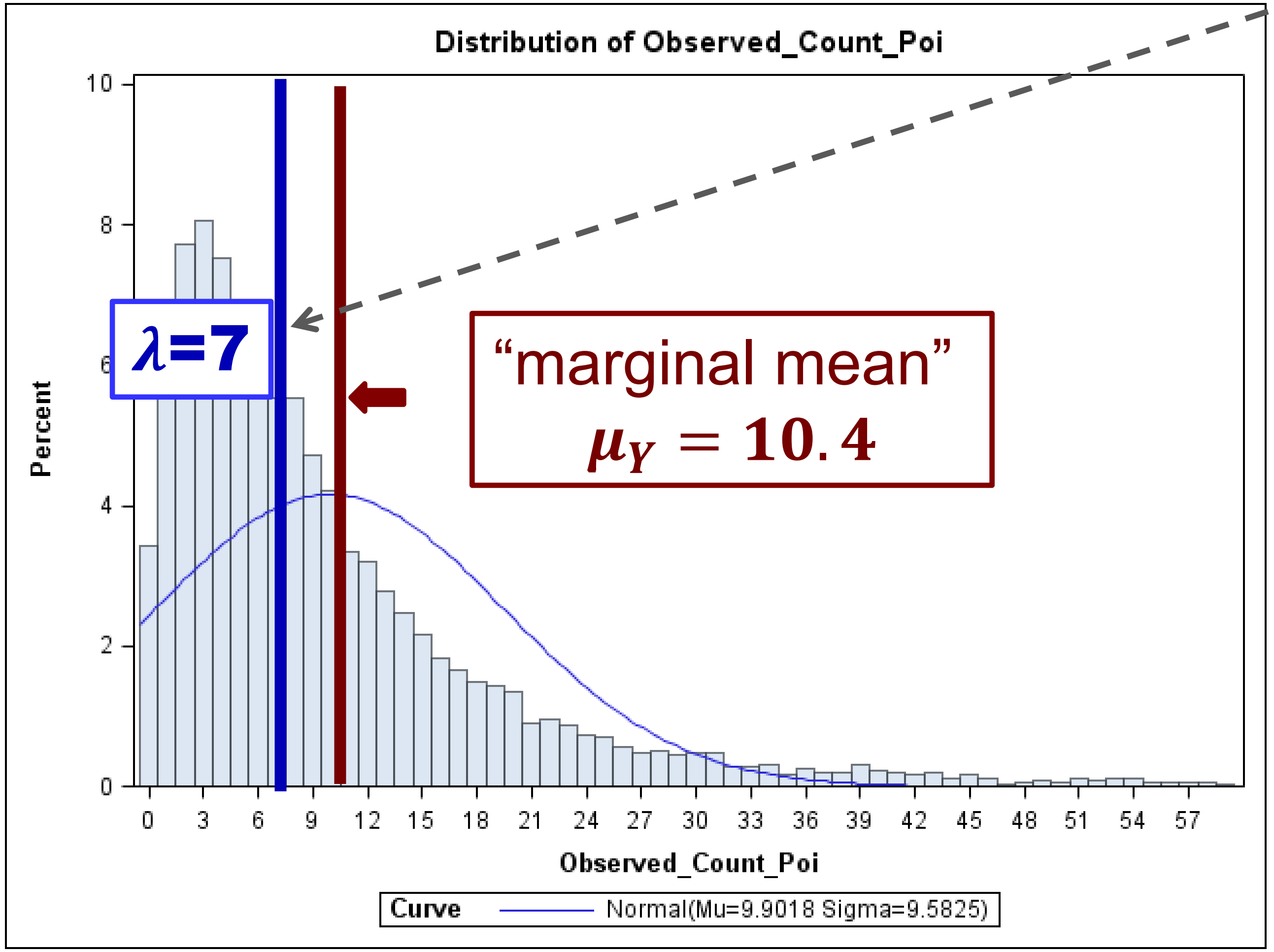


These distributions define the GLMM,

but you cannot observe either directly

Resulting Distribution of the Observations

$$block \sim N(0,0.8) \qquad y\,|\,block \sim Poisson(\lambda = 7)$$



λ=7

"marginal mean"

$\mu_Y = 10.4$

classical ANOVA will target μ_Y **overestimates λ**

GLMM will target **the true λ**

transformations

????

(somewhere in between?)

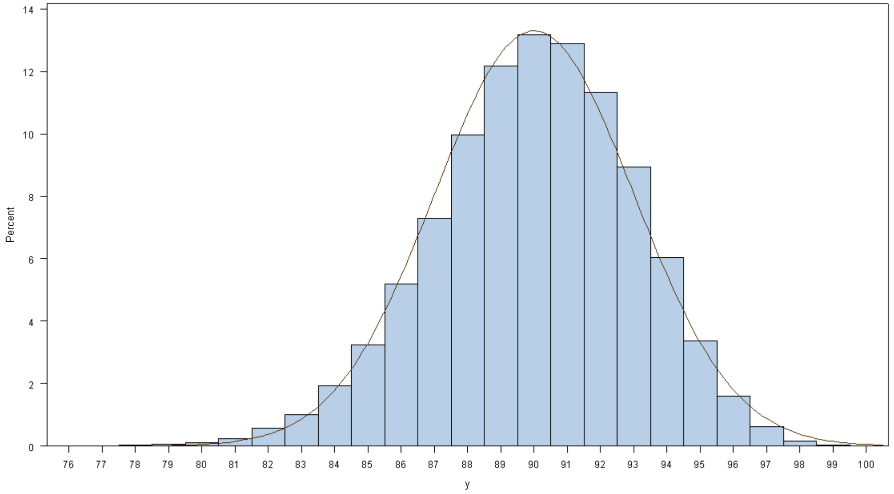
Same Issue Occurs with Binomial Data

- Example 2
- Paired Comparison, a.k.a. blocked design
- Two treatments, ***B*** blocks
- Response variable is *Y* “successes” out of *N* observations on given trt-block
- Hence $y_{ij}|b_j \sim \mathrm{Binomial}(N_{ij}, p_{ij})$ where *p* denotes probability of success

Normal Approximation – Isn’t this okay?

$\pi = 0.90$

N = 100



$$N = 100 \qquad \text{therefore} \quad Y_i \big/ N \sim \text{ approx } N\left(\pi, \frac{\pi(1-\pi)}{N}\right)$$

$$\text{ANOVA via LMM:} \quad Y_i \big/ N = p_i = \mu + \tau_i + b_j + e_i$$

Repurposed ANOVA to determine appropriate binomial GLMM

Combined	
Source	d.f.
block	9
trt	1
unit(block) | trt a.k.a. "residual" a.k.a "block x trt"	10-1=9
Total	19

$b_j \sim N(0, \sigma_b^2)$

$unit_{ij} \equiv bt_{ij} \sim N(0, \sigma_u^2)$

$$y_{ij} | b_j, bt_{ij} \sim Binomial(N_{ij}, p_{ij})$$

resulting linear predictor

$$\eta_{ij} = logit(p_{ij})$$
$$= \ log\left(\frac{p_{ij}}{1 - p_{ij}}\right)$$
$$= \eta + \tau_i + b_j + bt_{ij}$$

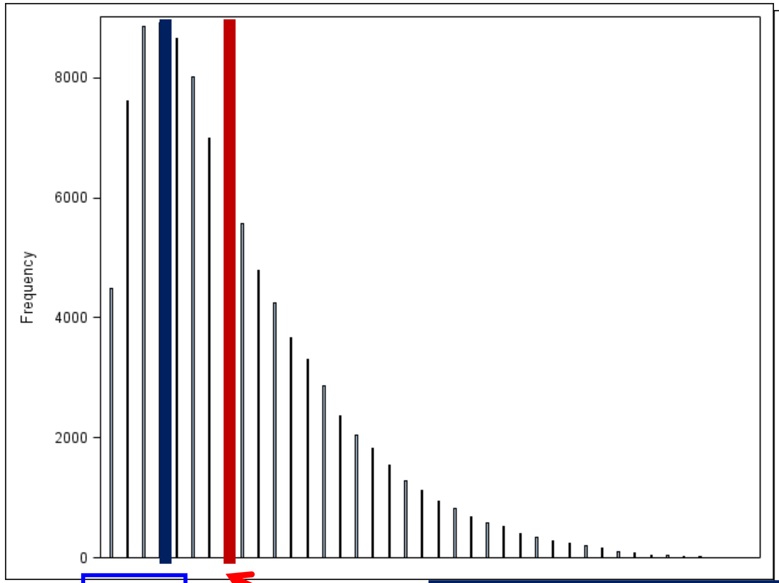
resulting estimate of p_i is

$$\frac{1}{1+exp[-(\hat{\eta}+\hat{\tau}_i)]}$$

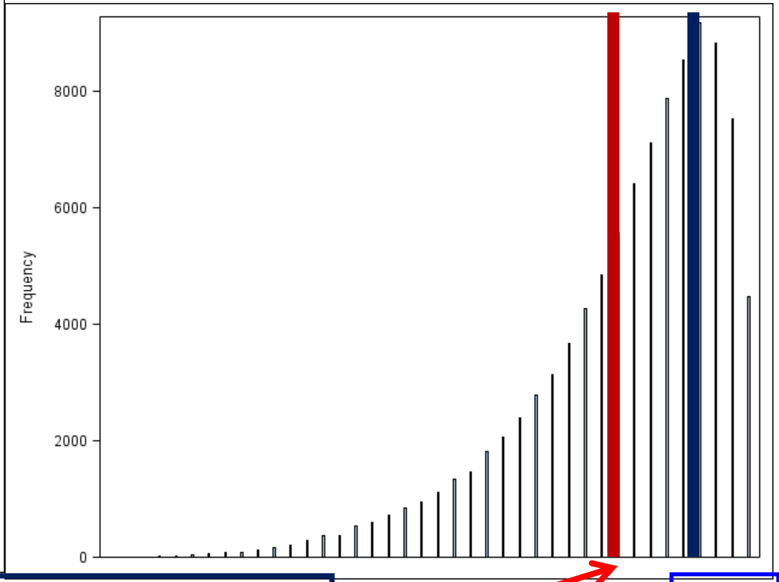
logistic inverse link

Consequence of Normal Approximation vs. GLMM

If *p<0.5*



If *p>0.5*



$\mathbf{p}_1$

normal appox (ANOVA)

estimates marginal $\boldsymbol{\mu}_{\mathbf{Y}}$

binomial GLMM

estimates $\mathbf{p}_{\mathbf{i}}$

$\mathbf{p}_2$

- **IMPORTANT:**
  what is your target & why?
- **IF you get this wrong:**
  - **inaccurate estimates of p_i & odds-ratio**
  - **loss of power**

List of examples

- ❑ **Elizabeth will cover**
  - **Example 11.5 in *SAS for Mixed Models* (Stroup, et al. 2018).** Multi-center clinical trial with binomial data (from Beitler and Landis, *Biometrics*, 1985)
  - **Example 12.3 in *SAS for Mixed Models.*** Manufacturing; data from multiple lots; response variable is number of micro-sites (discrete count). Random coefficient regression
- ❑ **Also in *SAS for Mixed Models (2018)***
  - **Exanple 11.6.** Estimate genetic parameters from count data
  - **Example 13.3.** Split-plot (multi-level mixed model) with count data
  - **Example 13.4.** Repeated measures (longitudinal data) with binomial data
- ❑ **Situations calling for GLMMs**
  - **designs:** blocks; matched pairs; multiple sources, batches, lots, locations, times; multi-level, split-plot; repeated measures; longitudinal **– all call for mixed model**
  - If, in addition, you have **non-Gaussian response variable(s), then GLMM**