Workshop: APS Annual Meeting, 2016

Generalized Linear Mixed Models for Data Analysis in Plant Pathology

Larry Madden — The Ohio State University, College of Food, Agricultural, and Environmental Sciences

Alissa Kriss — syngenta®

Case study 1

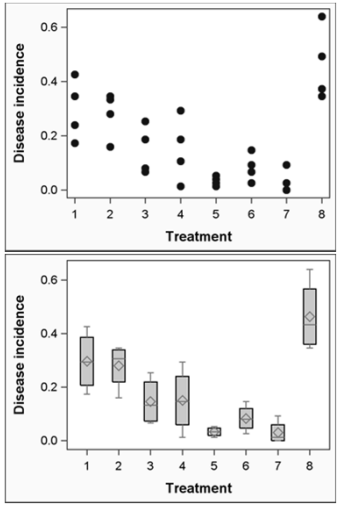
- Effect of fungicide treatment on Phomopsis leaf blight of strawberry (from Nita, Madden & Ellis)
  - Pathogen: *Phomopsis obscurans*
- Randomized complete block design (RCBD)
  - Four blocks (j = 1, …, 4)
  - Eight treatments (i = 1, …, 8) randomized within each block
- Response variable (Y): **leaflet disease incidence**
  - Number of diseased leaflets out of n = 75 leaflets in each experimental unit (=plot)
- Some questions:
  - Does treatment effect the probability of a leaflet being diseased (p)?
  - Which treatments are different from the others in terms of p?
- Analysis: linear mixed model (traditional approach) and especially generalized linear mixed model (GLMM) (contemporary approach)

Outline

- **Background and motivation**
  - Case study 1 (disease incidence and a randomized complete block)
  - Some concepts:
    - Models
    - Experimental layouts
    - Fixed vs. Random effects
- **From the Linear Mixed Model (LMM) to the Generalized Linear Mixed Model (GLMM)**
  - Binomial and other non-normal distributions
  - Conditional vs Marginal distributions
  - Conditional vs. Marginal models
- **Case study 1 in detail**
  - SAS GLIMMIX code and output
  - Dealing with data clustering and resulting over-dispersion
- **Technical issue:**
  - Different approaches to fitting GLMMs based on the likelihood principle
- **Overlooked challenge:**
  - The "*all zero problem*"
- **Case study 2: split plot**
  - SAS GLIMMIX code and output
- **Case study 3: Sub-sampling (time permitting)**

Case study 1: Phomopsis leaf blight data

blk	trt	y	n	prop
1	1	13	75	0.17333
1	2	12	75	0.16000
1	3	19	75	0.25333
1	4	22	75	0.29333
1	5	3	75	0.04000
1	6	11	75	0.14667
1	7	0	75	0.00000
1	8	48	75	0.64000
2	1	32	75	0.42667
2	2	21	75	0.28000
2	3	5	75	0.06667
2	4	14	75	0.18667
2	5	1	75	0.01333
2	6	7	75	0.09333
2	7	2	75	0.02667
2	8	26	75	0.34667
3	1	18	75	0.24000
3	2	26	75	0.34667
3	3	14	75	0.18667
3	4	8	75	0.10667
3	5	2	75	0.02667
3	6	2	75	0.02667
3	7	0	75	0.00000
3	8	37	75	0.49333
4	1	26	75	0.34667
4	2	25	75	0.33333
4	3	6	75	0.08000
4	4	1	75	0.01333
4	5	4	75	0.05333
4	6	5	75	0.06667
4	7	7	75	0.09333
4	8	28	75	0.37333



Conceptual Background

- **Response variable** (dependent variable) (Y):
  - A random variable that is measured or observed
  - **Continuous** (statistical distribution: **normal**, gamma, beta, etc.), or
  - **Discrete** (statistical distribution: Poisson or negative binomial [counts], **binomial** or beta-binomial [incidence], multinomial [ordinal categorical])
- *Investigations are carried out to determine if one or more factors or covariables affect the expected value (mean or location parameter) of the distribution of the response variable*
- **Factor**:
  - An explanatory variable that may affect the response variable, which is often "manipulated" by the investigator in an experiment
    - May be called a predictor variable or "independent variable"
  - Often, "factor" is used for a ***classification* or *class* variable**, consisting of two or more *levels* (*treat. 1, treat. 2, …; group 1 group 2, …; level 1, level 2, …*)
- **Covariable** (in contrast to a factor):
  - A *continuous* explanatory variable that is either manipulated or measured by the investigator (e.g., temperature; not considered in this workshop)

Models: Fixed- *versus* Random-Effects

- Models consist of *fixed-effect* variables and *random-effect* variables
- **Fixed-Effects Variables** (deterministic/structural component):
  - Levels (i.e., groups) in the study represent all possible levels of the factor, or all levels of interest by the investigator
    - **e.g., fungicide treatment, biocontrol treatment, inoculum dose, temperature, cultivar, etc.**
- **Random-Effects Variables** (stochastic/random component):
  - Levels (groups) in the study represent only a *random sample* of a larger set of levels (*a sample from a distribution of effects*)
    - **e.g., block, location (environment), plot, experimental repetition, …**
  - **Accounting for random effects results in appropriate estimates of fixed effects (e.g., treatment effects, means) and their standard errors**
  - We mostly consider here random effects that are a **consequence of the experimental design** (*splitting, sub-sampling, repeated measures*)
    - e.g., "**splitting" [blocking]** – randomly assigning levels of one factor (treatment) within levels of another factor
  - With random effects, the data are said to be "**clustered**"

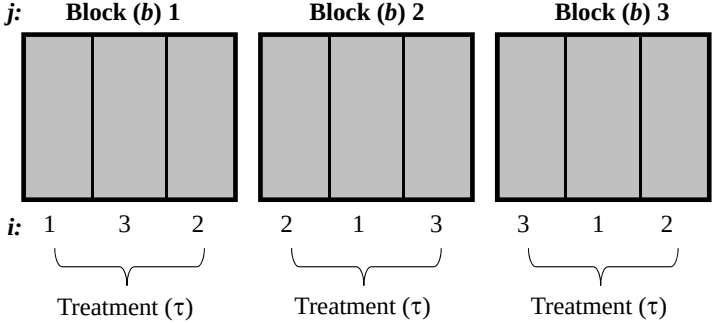
Conceptual background, *continued*

- Data analysis is key to determining if factors or covariables affect the expected value of the response variable
- In contemporary statistical science, data analysis is performed by fitting a model to data and interpreting the results
- **Model:**
  - *Abstraction of a real phenomenon or process that emphasizes those aspects relevant to the objectives of the user*
- **Statistical model:**
  - *"Plausible description of the process that gave rise to the observations"* – Stroup (2013)
  - *"Mathematical descriptions of how data conceivably can be produced"* – Littell et al. (2006)
  - Model with *stochastic* (**random-effect**) components and *deterministic* (**fixed-effect**) components, containing unknown constants (i.e., **parameters**) to be estimated
    -- Schabenberger & Pierce (2002)

Experimental Layout (example):

Randomized Complete Block Design (RCBD or RCB)

The ij combination (i-th treatment and j-th block) defines the plot (*experimental unit* or unit of replication) for a RCBD



b is the effect of block, τ is the effect of treatment

Model: Randomized Complete Block (RCB)

Response variable for *i-th* treatment of the *j-th* block

Constant, can be the **expected value** (mean) across all treatments and blocks

Effect of the *i*-th treatment *(i = 1, …, T)* on the response variable. Often of primary interest (expected value of the deviation of the response variable from the global average)

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}$$

If block is random, there are *two* random effects in model:
$b_j \sim N(0,\sigma_b^2)$
$e_{ij} \sim N(0, \sigma_e^2)$

Effect of the *j*-th block (*j* = 1, …, *N*) on the response variable

Usually considered a **random** effect

Residual (error term): Represents *variability* in response variable *not* accounted for by the block and treatment effects (***random***), With normal assumption: $e_{ij} \sim N(0, \sigma_e^2)$

Unique effect of each experimental unit (plot), identified by *ij* combination, on the response.

Equivalent to an interaction of treatment and block.

Model is for case where experimental unit (plot) is same as sampling unit.

Alternative model formulations

Traditional approach for **normal** data

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}$$
$$b_j \sim N(0,\sigma_b^2)$$
$$e_{ij} \sim N(0, \sigma_e^2)$$

Cannot be used for non-normal data—there is *no* sensible distributional concept of an additive non-normal residual. e.g., even if Y is binomial, the remainder (a fraction) cannot be binomial or be defined.

Equivalent (identical) for normal

Linear predictor (η) is the sum of all the terms affecting the response *except* the residual

Expectation-based approach for normal

$$\eta_{ij} = \theta + \tau_i + b_j$$
$$\mu_{ij} = \eta_{ij}$$
$$Y_{ij} \mid b_j \sim N(\mu_{ij}, \sigma_e^2)$$
$$b_j \sim N(0, \sigma_b^2)$$

Link the linear predictor to the so-called **conditional expected value** of the distribution (i.e., mean conditional on the random effects)

Define the so-called **conditional distribution of *Y*** for the experimental unit as a function of the **conditional expected value** (mean) of Y for treatment i and block j (μ_{ij}).
The residual variance (σ_e^2) becomes the conditional distribution variance.

See Supplemental Slides for a derivation

Alternative model formulations (RCB)

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}, \quad b_j \sim N(0,\sigma_b^2), e_{ij} \sim N(0, \sigma_e^2)$$

$$\eta_{ij} = \theta + \tau_i + b_j$$
$$Y_{ij} = \eta_{ij} + e_{ij}$$
$$b_j \sim N(0,\sigma_b^2), e_{ij} \sim N(0, \sigma_e^2)$$

Define the **linear predictor** (η; "eta") as the sum of all the terms affecting the response variable *except* the residual.

Response variable (Y) is the sum of the linear predictor and the residual (i.e., unique effect of each experimental unit (plot), identified by *ij* combination).

For normal distributions, the linear predictor (η_{ij}) defines the expected (or mean) value of Y for the *i*-th treatment and *j*-th block, $\mu_{ij} = \eta_{ij}$:

$$\mu_{ij} = \eta_{ij} = \theta + \tau_i + b_j$$

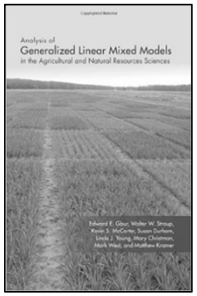
When there are random effects in the model, then μ_{ij} is known as the expected value *conditional* on the random effects (i.e., the treatment mean value for the *j*-th block): **Conditional expected value**

The conditional value of the response variable for the experimental unit is then written as $Y_{ij} \mid b_j$.

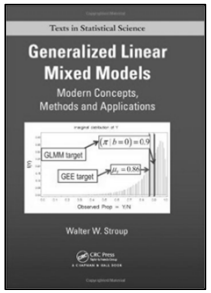
Modeling concepts

PROC GLIMMIX in SAS can be used for all of these models

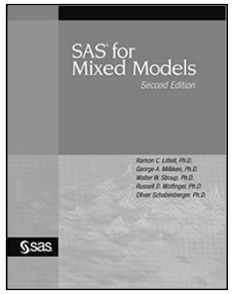
- Linear model (**LM**): Classical ANOVA, t-tests, etc.
  - Normal Y, all variables are fixed-effects terms (except for the residual)
- Linear mixed model (**LMM**): ANOVA with random effects, etc.
  - Normal Y, variables are either fixed- or random-effects terms
  - In particular, at least one random-effect term in model in addition to the residual (conditional distribution variance term)
  - Observations within a level of a random-effect variable are correlated
- Generalized linear model (**GLM**):
  - Non-normal Y (distribution in the exponential family), all variables are fixed-effects terms, except for the cond. dist. variance term, if there is a cond. dist. variance term to estimate (see later)
    - e.g., Poisson, binomial, gamma, beta, etc., for Y
    - (be careful: "PROC GLM" in SAS is for a LM, *not* for a GLM)
- Generalized linear mixed model (**GLMM**):
  - Non-normal Y, variables are either fixed- or random-effects terms
  - In particular, at least one random-effect term in model (in addition to a cond. dist. variance term, if there is a cond. dist. var. term to estimate)



Gbur et al.
Most accessible book.
Recommended



Stroup
Advanced (for stat grad students), but great stuff



Littell et al. (chap. 14)
All mixed models.
Great way to learn SAS

Rethinking the Analysis of Non-Normal Data in Plant and Soil Science

Walter W. Stroup*

***Agronomy Journal* 107(2): 811-827 (2015)**
Recommended [but be careful (don't use his beta distribution method for incidence)]

Binomial distribution: *Bin*(*p*,*n*)

Y: Number of individuals with a trait (e.g., disease) in an experimental unit or sampling unit (e.g., plot, plant) – response

n: Number of individuals observed for the trait (e.g., plants)

p: A location parameter: probability of trait, such as disease (e.g. probability that a leaf, plant, etc., is diseased) (analogous to μ of normal)

For a single simple random sample of *n* plants, disease incidence (as a proportion) is an estimate of *p*

***p* may be a function of treatment and block (or a function of any factor and random effects): the <u>conditional probability of disease</u>, p_{ij}**

Analogous to normal distribution μ_{ij}, with a random block effect,

Variance of conditional distribution of ***Y*** is ***np*(1-*p*)**, fully defined by ***n*** and ***p*** (<u>no separate variance parameter to estimate</u>)

- As *n* becomes large, ***Bin*(*p*,*n*)** is approximated by a normal distribution in a single sample, with mean ***np***, and variance ***np*(1-*p*)**
- ***However, with random effects, the distribution of Y (or the proportion Y/n) <u>across</u> the random effects will <u>not</u> be binomial, normal, or symmetric! (we will get to this soon…)***

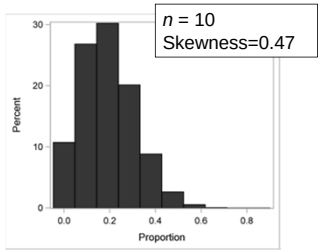
Non-normal distributions

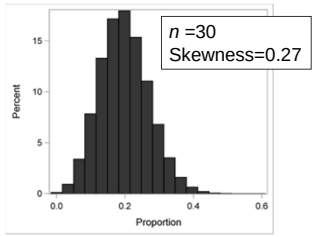
- Many response variables (Y) of relevance in plant pathology are not normally distributed
  - **<u>Discrete data</u>:**
    - Poisson or negative binomial distribution (**<u>counts</u>** <u>with no upper bound</u>)
      - **Lesions per plant, nematodes per volume of soil, oospores per gram of soil, …**
    - Binomial or beta-binomial distribution (**proportions**, or counts with an upper limit)
      - **Plant disease incidence, leaf disease incidence, ….**
        - » <u>Proportion</u> of plants diseased, proportion of leaves diseased
        - » <u>Number</u> of plants diseased out of *n* observed plants
  - **<u>Skewed (positive) continuous data</u>:**
    - Gamma, exponential, or beta distribution
      - **Example: Time to event**: Time to germinate, time to sporulate

Emphasis on discrete (**<u>binomial</u>**) data in workshop

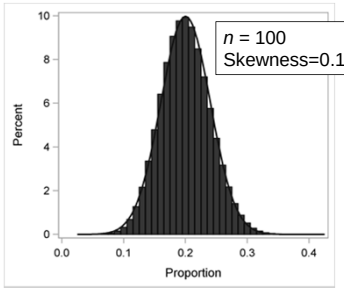
------Distributions in the exponential family------

NOTE: It is *not* adequate to simply look at a histogram of observations across *all* the treatments and blocks to determine the distribution of Y *for the individual experimental units*





<u>Binomial distribution</u>
(=probability mass function) at different values of *n*.
***p* = 0.20.**



Simulation.
No random effects. Becomes normal-like as *n* increases.

Data Analysis

- The old (and lingering) dogma:
  - Linear models or linear mixed models are *reasonably* robust to *moderate* departures from normality
  - Many discrete distributions (e.g., binomial, Poisson) can be well approximated by a normal (continuous) distribution when there is a large number of observations (n) or when counts are large
- In fact, many discrete distributions **may** be approximated by a normal distribution only under **some** circumstances!
  - **Despite "common wisdom", based on *pre-mixed-model thinking* (considering only fixed effects), these discrete distributions may never be approximated by a normal when there are random effects**
  - Alternatives to linear and linear mixed models are needed in analysis that do not rely on normality: **Generalized Linear Mixed Models**
- A property of non-normal distributions, in general, is that the variance of the distribution is a function of the mean
  - Property exists even when the distribution is approximated by a normal dist.
  - This means that factor levels (treatments) will have different variances, a violation of the basic (normal) linear mixed model
    - It is not possible for: $Y_{ij} \mid b_j \sim N(\mu_{ij}, \sigma_e^2)$ (i.e., constant variance)

Response (Y): *proportion diseased*
Fit with a Linear Mixed Model (assume normal)

Case study 1

$\eta_{ij} = \theta + \tau_i + b_j$

$\mu_{ij} = \eta_{ij}$

$Y_{ij} \mid b_j \sim N(\mu_{ij}, \sigma_e^2)$

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

Solutions for Fixed Effects

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		0.4633	0.04494	3	10.31	0.0019
trt	1	-0.1667	0.06356	21	-2.62	0.0159
trt	2	-0.1833	0.06356	21	-2.88	0.0089
trt	3	-0.3167	0.06356	21	-4.98	<.0001
trt	4	-0.3133	0.06356	21	-4.93	<.0001
trt	5	-0.4300	0.06356	21	-6.77	<.0001
trt	6	-0.3800	0.06356	21	-5.98	<.0001
trt	7	-0.4333	0.06356	21	-6.82	<.0001
trt	8	0	.	.	.	.

Estimated θ / Estimated τ_i

SAS (preview):

```
proc glimmix ;
class trt blk;
model prop = trt / s;
random  blk;
lsmeans trt / cl;
run;
```

$\hat{\mu}_i = \hat{\theta} + \hat{\tau}_i$

$\hat{\mu}_1 = \hat{\theta} + \hat{\tau}_1 = 0.4633 - 0.1667 = 0.2967$

...

$\hat{\mu}_8 = \hat{\theta} + \hat{\tau}_8 = 0.4633 + 0 = 0.4633$

$H_0: \tau_1 = \tau_2 = \ldots \tau_T$

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trt	7	21	11.25	<.0001

Data Analysis

- Unequal variances (a property of non-normal distributions)
  - In the pre-mixed-model days (when random effects were not really treated as random), it was shown by Bartlett, and many others, that a transformation of Y (Y^*) could have an *approximately* constant variance (i.e., a variance that was not a function of the mean)
  - Binomial, when Y is a proportion:
    - $Y^* = \arcsin(\sqrt{Y}) = \sin^{-1}(\sqrt{Y})$ (*angular transformation*)
  - Poisson, when Y is a count without an upper bound:
    - $Y^* = \sqrt{Y}$ (*square-root transformation*)
  - Transformations may also cause other problems:
    - One transformation may be appropriate to stabilize variances but a different transformation to obtain a linear (straight-line) relation between the response and a predictor; another transformation to obtain a symmetrical distribution
  - **One is *changing* the distribution to *force* it into a linear mixed model**

Response:
Proportion diseased (assume normal), *continued*

- Statistical theory shows that the estimated mean proportions (even with unrealistic normality assumption) are **unbiased** estimates of the true proportions for each treatment *across all blocks*.
- However, **the estimated standard errors (SEs) are all incorrect** (by definition)!
  - The SEs must be functions of the mean for binomial data [must be proportional to $p(1-p)$]
  - Incorrect SEs will give incorrect tests of significance for treatment effects and lead to incorrect conclusions

Depending on the level of incidence, one could:
falsely declare differences as significant when they are not;
or
fail to declare true differences as being significantly different!

$\hat{\mu}_i = \hat{\theta} + \hat{\tau}_i$

trt Least Squares Means

trt	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
1	0.2967	0.04494	21	6.60	<.0001	0.05	0.2032	0.3901
2	0.2800	0.04494	21	6.23	<.0001	0.05	0.1865	0.3735
3	0.1467	0.04494	21	3.26	0.0037	0.05	0.05320	0.2401
4	0.1500	0.04494	21	3.34	0.0031	0.05	0.05654	0.2435
5	0.03333	0.04494	21	0.74	0.4665	0.05	-0.06013	0.1268
6	0.08333	0.04494	21	1.85	0.0778	0.05	-0.01013	0.1768
7	0.03000	0.04494	21	0.67	0.5117	0.05	-0.06346	0.1235
8	0.4633	0.04494	21	10.31	<.0001	0.05	0.3699	0.5568

Response: **Angular transformation (assume normal)**

- Angular transformation ($Y^* = \arcsin(\sqrt{proportion})$) results in a distribution with an *approximately* constant variance at different levels of the mean
  - Thus, the constant SEs are reasonable for Y^*
- However, the estimated means are for the angular transformation
  - With significant block effect, back-transformation (inverse transformation) of the logit mean is not equal to the estimated mean proportion for the treatment across the blocks

Hypothetical (with $\sigma_b^2 = 1.5$)

True mean prop. = 0.15

Mean of angular = 0.365

Back-transform of mean = 0.127 (not 0.15)

trt Least Squares Means (*angular transformation*)

trt	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
1	0.5707	0.06422	21	8.89	<.00			
2	0.5535	0.06422	21	8.62	<.00			
3	0.3805	0.06422	21	5.93	<.00			
4	0.3669	0.06422	21	5.71	<.00			
5	0.1785	0.06422	21	2.78	0.01			
6	0.2822	0.06422	21	4.39	0.00			
7	0.1186	0.06422	21	1.85	0.07			
8	0.7482	0.06422	21	11.65	<.0001	0.05	0.6147	0.8818

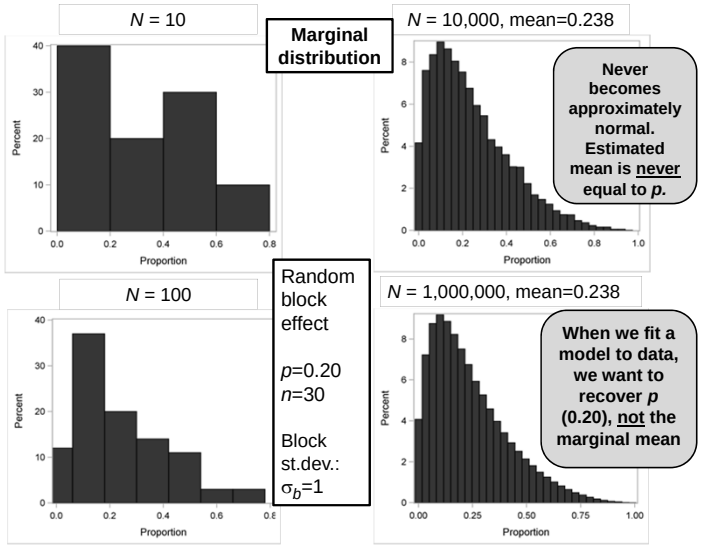
```
data p; set p;
angular =
arsin(sqrt(prop));

proc glimmix ;
class trt blk;
model angular = trt ;
random  blk;
lsmeans trt / cl;
run;
```

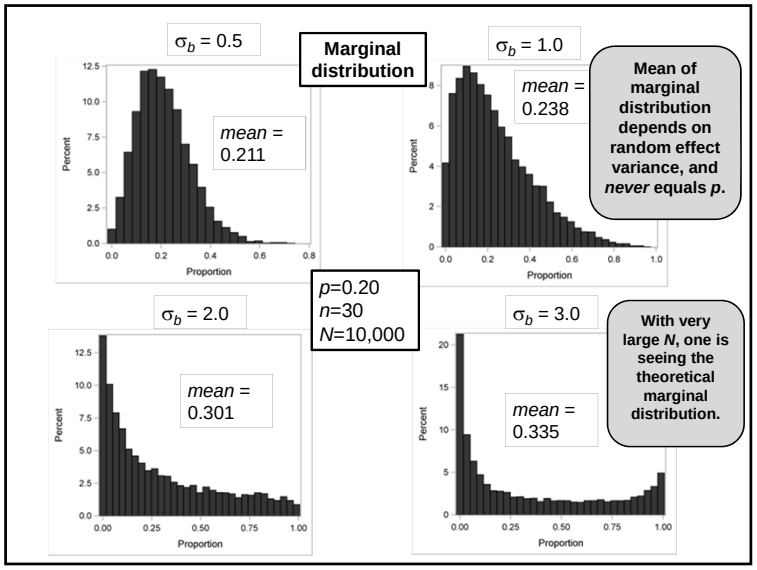
Simulation demonstration

- Consider a binomial distribution for Y, with p = 0.20 and n = 30, for an experimental unit (e.g., plot)
  - e.g., probability of disease is p = 0.20, with n = 30 plants per plot
  - So-called **conditional distribution: "$Y \mid plot \sim Bin(p,n)$"**
  - This conditional distribution could be for one of the treatments
- Consider the impact of random effects (blocks)
  - With one treatment, each plot is a block
  - **Consider that p is randomly perturbed by the block**, so that in some blocks p is higher than 0.20, and in other blocks p is lower than 0.20
    - On average, the perturbation is 0
- What is the distribution of Y (number), or Y/n (proportion), across all blocks for different degrees of perturbation of p (different variances) and different numbers of blocks (N; $j = 1, \ldots, N$)
  - This distribution is the so-called **marginal distribution**
- What is the mean proportion for the marginal distribution, and how does it compare to p? (in an analysis, we want to recover p)

Binomial data (some summary points)

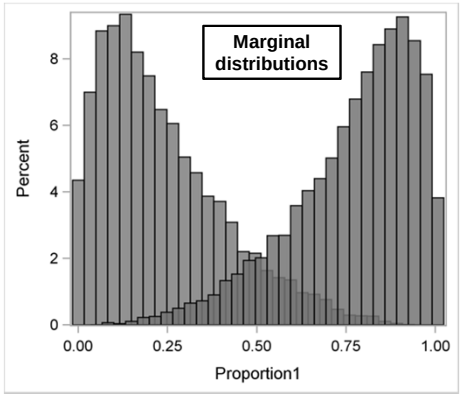
- A linear (i.e., *normal*) mixed model fitted to binomial data can result in unbiased estimates of the mean proportions (incidence) *across all blocks* for each treatment
  - However, all measures of variability (e.g., SEs) are wrong, which means that tests of hypotheses, confidence intervals, etc., are incorrect!
  - Moreover, the mean proportion diseased for a treatment across the blocks is *not* the estimated probability of disease for an experimental unit receiving a treatment (*explained soon*)
    - **One wants the latter: the conditional probability of disease, p_{ij}, for the *ij*-th experimental unit (plot)**
- A transformation of the proportions *attempts* to force the data to be normal with variance independent of the mean
  - Analysis (with back-transformation) does **not** give the mean proportion diseased for a treatment across all blocks, and does **not** give the conditional probability of disease for the *ij*-th experimental unit
- Neither of the above approaches is satisfactory in terms of Type I and Type II errors, power, confidence interval coverage, etc.





Establishing some concepts

- **Marginal distribution**: **"*Y* ~ *Distribution*(•)"**
  - The distribution of the observations (across all the levels of the random effects); may not be easily defined mathematically
  - *The distribution that we see*:
    - **Outcome of the (conditional) distribution of *Y* at the basic experimental unit (e.g., plot) coupled with distribution(s) of the random effects**
- **Conditional distribution: "*Y* | *random effects* ~ *Distribution*(•)"**
  - Distribution of *Y* conditional on the *specific* level of the random effects
    - For example, the distribution of the number of diseased plants in a plot receiving a given treatment in the first (or second, third, etc.) block
    - Typical distributions: normal, binomial, Poisson, gamma, beta
  - **Conditional distribution represents the stochastic process generating the data at the experimental unit (e.g., plot) level (thus interest should be parameters of this distribution)**
- **The marginal distribution is normal only if the conditional distribution is normal (in general, marginal distribution is asymmetric)**
- **The mean of the marginal distribution does not equal the mean of the conditional distribution, except for the normal distribution**



Two treatments
N = 10,000
(*n*=30 each trt)

σ_b = 1.0
p_1 = 0.2, p_2 = 0.8

p_2 - p_1 = 0.6 (true)

mean1 = 0.239
mean2 = 0.762

$mean_2$ - $mean_1$= 0.523

Means are closer together for marginal distribution compared with the true probabilities.

?? **"For binomial data, ANOVA with or without transformation should be considered unacceptable for scientific publication"**
--Stroup (2015) **??**

- A very strong recommendation *for* GLMMs and *against* LMMs (ANOVA is a special case of a LMM)! Based on:
  1. Some poor statistical results from LMMs (normality assumption) fitted to proportion data or to a transformation of proportion, and
  2. When one is using a LMM (normality assumption), one is *targeting* the mean of the marginal distribution (which depends on random-effect variances) and not the probability of disease in an experimental unit, the so-called **conditional probability**
     - The so-called **target of inference** is not *p* with an LMM analysis, but the target of inference is *p* with a GLMM analysis
- We do not fully agree with such a strong blanket recommendation
  - There are circumstances where use of GLMM is overly challenging
- However, we agree that GLMMs have much to offer

GLMM: From Normal to Binomial *Conditional* Distribution (demonstrated for a RCB)

- **Normal distribution** (two parameters for conditional dist.): μ and σ_e^2
  - $Y_{ij} \mid b_j \sim N(\mu_{ij}, \sigma_e^2)$
- **Binomial distribution** (one parameter for conditional dist.): p
  - p: Probability of trait, such as disease (e.g. probability that a leaf, plant, etc., is diseased) – parameter (analogous to μ of normal)
  - $Y_{ij} \mid b_j \sim Bin(p_{ij}, n)$

    Note: Variance of Y: $np(1-p)$, fully defined by n and p [$np_{ij}(1-p_{ij})$]

    There is no separate variance parameter! (*more on this later*)

    (For Poisson [count] data, there also is no variance parameter).

$\eta_{ij} = \theta + \tau_i + b_j$ → $\eta_{ij} = \theta + \tau_i + b_j$

$\mu_{ij} = \eta_{ij}$ ⇢ p_{ij} = ???

$Y_{ij} \mid b_j \sim N(\mu_{ij}, \sigma_e^2)$ → $Y_{ij} \mid b_j \sim Bin(p_{ij}, n)$

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

Generalized Linear Mixed Model (GLMM) for binomial data (a *conditional model*)

Link function, $g(p)$:

logit(p)**, probit**(p), complementary log-log function [**CLL**(p)], etc.

We will mostly stick with **logit**: **logit**$(p) = \ln[p/(1-p)]$

Linear predictor for RCB:

logit$(p_{ij}) = \eta_{ij} = \theta + \tau_i + b_j$

Model:

$\eta_{ij} = \theta + \tau_i + b_j$

logit$(p_{ij}) = \eta_{ij}$

$Y_{ij} \mid b_j \sim Bin(p_{ij}, n)$

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

Obtaining p from η:

$p_{ij} = 1/(1+ \exp(-\eta_{ij}))$

Or more generically as:

$p_{ij} = \text{logit}^{-1}(\eta_{ij})$

Or even more generically as the **inverse-link** function:

$p_{ij} = g^{-1}(\eta_{ij})$

Binomial: Linking the linear predictor (η) to p

One *could* directly link the two ($p_{ij} = \eta_{ij}$), so that $p_{ij} = \eta_{ij} = \theta + \tau_i + b_j$

However, this usually is **not** a good idea:

p is bound by 0 and 1, but predictions based on τ_i and b_j could include values less than 0 or greater than 1

There is usually a nonlinear relationship, with small changes in p near the limits (0 and 1) being on the same scale as large changes in p near the center of the probability scale (0.25-0.75)

Statistical theory based on the form of the distribution (not covered here) indicates that η is linked in a linear manner to a function of p: $g(p) = \eta$

$g(p_{ij}) = \eta_{ij} = \theta + \tau_i + b_j$

$g(p)$ is known as the **link function**, because it *links* η to p

The link function is a transformation of a parameter, *not* a transformation of a random variable (do not confuse these two)

Case study 1: Phomopsis leaf blight

$\eta_{ij} = \theta + \tau_i + b_j$

logit$(p_{ij}) = \eta_{ij}$

$Y_{ij} \mid b_j \sim Bin(p_{ij}, n)$

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

Note: $\theta + \tau_i$ is the logit of the conditional probability of disease for i-th treatment (at the typical value of the block effect (0))

Y_{ij}: Number of diseased leaflets for experimental unit -- treatment i and block j (binomial)

n: Number of observed leaflets in experimental unit

p_{ij}: Conditional probability of disease for treatment i and block j

θ: constant ("intercept")

τ_i: Effect of the i-th level of treatment on logit of p

b_j: Effect of the j-th level of block on logit of p

Model naturally and automatically takes into account that $\text{var}(Y_{ij}) = n\, p_{ij}(1-p_{ij})$

Note:
$p_i = \text{logit}^{-1}(\theta + \tau_i) = g^{-1}(\theta + \tau_i)$
is the conditional probability of disease for a plot (experimental unit) receiving treatment i, at the typical (average) value of the block effect (**0**).

This is **not** the average proportion across blocks!

$\eta_{ij} = \theta + \tau_i + b_j$

$\text{logit}(p_{ij}) = \eta_{ij}$

$Y_{ij} \mid b_j \sim Bin(p_{ij}, n)$

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

Model fitting is based on a version of maximum likelihood (ML) Discussed later.

```
proc glimmix;
class block treat;
model Y/n = treat / dist=binomial link=logit;
random block;
run;
```

Model statement gives response variable (*Y*; number diseased here) and number observed (*n*) in experimental unit.

Model statement also gives all the **fixed-effect** terms in the linear predictor (η), and specifies the **distribution** and **link** (the intercept [θ] is there by default)

One or more **random** statements are used to specify the random effects in the linear predictor (η) and the distributional properties of the random effects (by default)

GLMM: Dealing with effects "clustering"

Normal:

$\eta_{ij} = \theta + \tau_i + b_j$

$\mu_{ij} = \eta_{ij}$

$Y_{ij} \mid b_j \sim N(\mu_{ij}, \sigma_e^2)$

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

With normal distribution, there is a *residual* variance term to account for the unique (random) contribution of each experimental unit on *Y* (variation not accounted for by main effects of block and treatment).

Binomial:

$\eta_{ij} = \theta + \tau_i + b_j$

$\text{logit}(p_{ij}) = \eta_{ij}$

$Y_{ij} \mid b_j \sim Bin(p_{ij}, n)$

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

With binomial distribution, there is **no** residual variance term. Variance at the experimental-unit scale is fully defined as being exactly $np_{ij}(1\text{-}p_{ij})$.

- This **naïve approach** might make sense if the experimental unit in each block was the leaflet and *not* the plot.
- In reality, there is clustering of observations (all *n* leaflets clustered within the actual experimental unit)
- Each plot may have a unique (random) effect on the probability of a leaflet being diseased, even after accounting for treatment and block. Or, equivalently, the treatment effect varies randomly with plot.
- **This needs to be taken into account.**
  - There are several ways (we focus on **two**)

CaseStudy1.sas

Adjustment 1: Conditional GLMM

- One (modern) way to account for data clustering and unique contributions of experimental units (*ij* combinations [plots]) is to add a random effect to the linear predictor, $v_{ij} \sim N(0,\sigma_v^2)$
  - Equivalent to a **(random) interaction of block and treatment** (unique *ij*)
- The expanded GLMM: fully conditional model (new terms in black)
  - *Y* is conditionally binomial (conditional on *all* the random effects)
    - **That is, *Y* has a true conditional distribution (a plausible stochastic model for the generation of the observed data)**

$\eta_{ij} = \theta + \tau_i + b_j + v_{ij}$

$\text{logit}(p_{ij}) = \eta_{ij}$

$Y_{ij} \mid b_j, v_{ij} \sim Bin(p_{ij}, n)$

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

$v_{ij} \sim N(0,\sigma_v^2)$

- Note that the model now has *one more* random-effect term, analogous to the residual (e_{ij}) with a normal distribution
- The *ij* subscript identifies that each plot (experimental unit) has a unique effect

```
proc glimmix; /* minimal SAS code */
class block treat;
model Y/n = treat / dist=binomial link=logit;
random block block*treat;
run;
```

Adjustment 2: Quasi-likelihood

- One consequence of a unique (random) effect of each experimental unit (ij), or of *clustering*, in general, is that the variance will be *larger* than specified by the nominal conditional distribution
  - $Var(Y_{ij} \mid b_j) > np_{ij}(1-p_{ij})$
    - Known as **over-dispersion** (or **extra-binomial heterogeneity**)
- One can accommodate this over-dispersion by specifying (*defining*):
  - $Var(Y_{ij} \mid b_j) = \phi np_{ij}(1-p_{ij})$
    - ϕ: over-dispersion scale parameter, analogous to a form of residual variance with normal distribution (generally, ≥ 1)
- However, the conditional distribution is no longer binomial
  - In fact, the it is possible that **no** legitimate statistical distribution can be defined that has the properties of the binomial (for probability of Y=0, 1, etc.) **and** has the larger variance
    - We can think of the new "distribution" as a *quasi-distribution*
  - Thus, one uses a so-called **quasi-likelihood** instead of a (true) likelihood to fit the model
    - Quasi-likelihood defined in terms of expected values (means) and variances only

CaseStudy1.sas (continued)

Adjustment 2: Quasi-likelihood, *continued*

- Clustering accounted for by use of over-dispersion scale parameter
  - $Var(Y_{ij} \mid b_j) = \phi np_{ij}(1-p_{ij})$
  - ϕ: over-dispersion scale parameter
- One is not targeting the conditional p for the basic experimental unit (e.g., plot), but is targeting p for a higher level in the hierarchy. This is sort of a hybrid approach (once very common before modern GLMMs).

$$\eta_{ij} = \theta + \tau_i + b_j$$
$$\text{logit}(p_{ij}) = \eta_{ij}$$
$$Y_{ij} \mid b_j \sim \textit{quasi-Bin}(p_{ij}, n; \phi)$$
$$b_j \sim N(0, \sigma_b^2)$$

```
proc glimmix; /* minimal SAS code */
class block treat;
model Y/n = treat / dist=binomial link=logit;
random block;
random _residual_;
run;
```

There are other possible adjustments not covered in this workshop. See the supplemental slide at the end for some guidance.

Labels – Labels – Labels: *Be Careful*

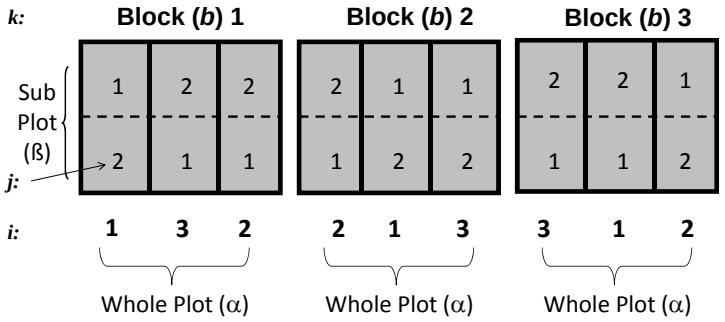
- In general, there is a great deal of inconsistency in the labels given to GLMMs. Different authors may use different names.
- For instance, two of the co-developers for GLMM methods (Breslow & Clayton) used the term "Maximum Quasi-likelihood" (MQL) for the situation where ϕ is *not* estimated (i.e., where $\phi = 1$) (there is a reason which we do not cover here)
  - We follow Stroup (2013), the authority on GLMMs, and use the label *quasi-likelihood* for the situation when one *estimates* ϕ (when one no longer has a true likelihood).
- Despite labeling confusion or inconsistency, one needs to distinguish:
  1. True conditional GLMM (true likelihood or true conditional distribution)
     Model targets p (location parameter) for *conditional* distribution (e.g., for the basic experimental unit (plot) when this term is in the model)
  2. Quasi-likelihood (over-dispersion), where one does not have a true likelihood or true conditional distribution
     Model does not target the conditional p for the basic experimental unit (such as the plot). Rather, it is more of a hybrid approach.

Random Effects, *revisited*

- Random effects arise from:
  - Random selection of the levels of factor studied (when the factor level effects come from a distribution of effects)
  - **Clustering of data**
    - **Cluster (or subject):** *collection of observations that are somehow stochastically related* (*correlated)*
    - Experimental design and type of data collection "create" (induce) the "clustering"
    - **Mechanisms for clustering**:
      - **Splitting: randomly assigning levels of one factor *within* levels of another factor**
      - Sampling and sub-sampling (*nesting* of sampling units)
        - » Where the experimental unit or unit of replication is 'larger' than the sampling unit (the unit of observation)
      - Repeated observations over time (or space)

Splitting of experimental units:
Split Plot (with blocks)

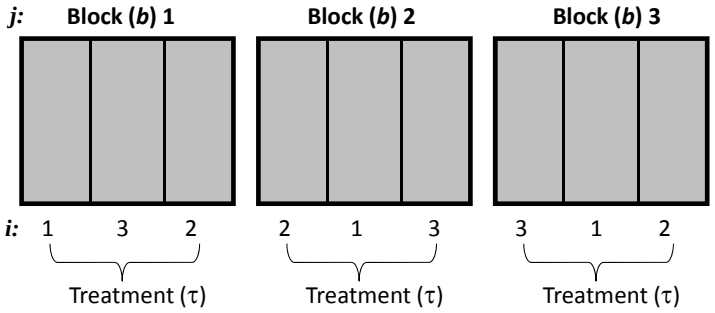
Variation among blocks (k), among whole plots (large experimental units), and among sub-plots. (Whole plot: ik combination. Sub-plot: ijk combination).



Each block is a cluster (with random assignment of "whole plot" levels), and each whole plot is a cluster (with random assignment of "sub-plot" levels)

Splitting of experimental units (continued): Randomized Complete Block

Variation among blocks (i) and among plots (ij) (the experimental units)



Each block is a cluster (randomly assigning treatments within each cluster)

Case study 2: Split plot with blocks

- Effects of wheat cultivar (variety) and fungicide-timing treatment on incidence of Fusarium head blight (FHB)
  - from D'Angelo et al. (2014 Plant Disease 98: 1387-1397)
- Split plot with blocks
  - Three blocks (b_k) (k = 1, …, 3)
  - Whole plot factor: wheat cultivar (α_i) (i = 1, 2), randomized within each block
  - Sub-plot factor: fungicide timing relative to anthesis (β_j) (j = 1, …,5), randomized within each whole plot
    - Control, spray at anthesis (flowering), or 2, 4, or 6 days after anthesis
- Response variable: spike disease incidence
  - Number of diseased spikes (Y) out of n observed spikes in each experimental unit (plot: ijk combination)
    - n varies from 70 to 100
- Analysis: generalized linear mixed model (naïve; conditional GLMM; quasi-likelihood)

Split Plot Design (with blocking)

Normal *Y* (LMM)

$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + v_{ik} + e_{ijk}$$
$$b_k \sim N(0, \sigma_b^2)$$
$$v_{ik} \sim N(0, \sigma_v^2)$$
$$e_{ijk} \sim N(0, \sigma_e^2)$$

Normal *Y* (LMM)

$$\eta_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + v_{ik}$$
$$\mu_{ijk} = \eta_{ijk}$$
$$Y_{ijk} \mid b_k, v_{ik} \sim N(\mu_{ijk}, \sigma_e^2)$$
$$b_k \sim N(0, \sigma_b^2)$$
$$v_{ik} \sim N(0, \sigma_v^2)$$

Y_{ijk}: response (dependent variable) for the *i-th* level of whole-plot factor, *j-th* level of sub-plot factor, and *k-th* block
θ: constant ("intercept")
α_i: Effect of the *i*-th level of whole-plot factor
β_j: Effect of the *j*-th level of sub-plot factor
$(\alpha\beta)_{ij}$ Interaction effect (effect of *i*-th whole plot and *j*-th subplot)

b_k: Random effect of the *k*-th level of block
v_{ik}: Random effect of whole-plot experimental unit (*ik*-th experimental unit), equivalent to interaction of block and whole-plot
e_{ijk}: Random effect of sub-plot (*ijk*-th) experimental unit [**residual or conditional dist. variance**] **equivalent to interaction of block, whole-plot and sub-plot**

Split Plot Design (with blocking)

Binomial *Y* (naïve GLMM)

$$\eta_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + v_{ik}$$
$$\text{logit}(p_{ijk}) = \eta_{ijk}$$
$$Y_{ijk} \mid b_k, v_{ik} \sim Bin(p_{ijk}, n)$$
$$b_k \sim N(0, \sigma_b^2)$$
$$v_{ik} \sim N(0, \sigma_v^2)$$

Binomial *Y* (condional GLMM)

$$\eta_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + v_{ik} + u_{ijk}$$
$$\text{logit}(p_{ijk}) = \eta_{ijk}$$
$$Y_{ijk} \mid b_k, v_{ik}, u_{ijk} \sim Bin(p_{ijk}, n)$$
$$b_j \sim N(0, \sigma_b^2)$$
$$v_{ik} \sim N(0, \sigma_v^2)$$
$$u_{ijk} \sim N(0, \sigma_u^2)$$

Binomial *Y* (quasi-likelihood)

$$\eta_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + v_{ik}$$
$$\text{logit}(p_{ijk}) = \eta_{ijk}$$
$$Y_{ijk} \mid b_k, v_{ik} \sim \textit{quasi-Bin}(p_{ijk}, n; \phi)$$
$$b_k \sim N(0, \sigma_b^2)$$
$$v_{ik} \sim N(0, \sigma_v^2)$$

Conditional variance $= \phi n p_{ijk}(1-p_{ijk})$

Split Plot Design (with blocking)

Binomial *Y* (naïve GLMM)

$$\eta_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + v_{ik}$$
$$\text{logit}(p_{ijk}) = \eta_{ijk}$$
$$Y_{ijk} \mid b_k, v_{ik} \sim Bin(p_{ijk}, n)$$
$$b_k \sim N(0, \sigma_b^2)$$
$$v_{ik} \sim N(0, \sigma_v^2)$$

Binomial *Y* (GLMM)

$$\eta_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + v_{ik} + u_{ijk}$$
$$\text{logit}(p_{ijk}) = \eta_{ijk}$$
$$Y_{ijk} \mid b_k, v_{ik}, u_{ijk} \sim Bin(p_{ijk}, n)$$
$$b_j \sim N(0, \sigma_b^2)$$
$$v_{ik} \sim N(0, \sigma_v^2)$$
$$u_{ijk} \sim N(0, \sigma_u^2)$$

Y_{ijk}: response (dependent variable) for the *i-th* level of whole-plot factor, *j-th* level of sub-plot factor, and *k-th* block
θ: constant ("intercept")
α_i: Effect of the *i*-th level of whole-plot factor
β_j: Effect of the *j*-th level of sub-plot factor
$(\alpha\beta)_{ij}$ Interaction effect (effect of *i*-th whole plot and *j*-th subplot)
b_k: Random effect of the *k*-th level of block
v_{ik}: Random effect of whole-plot experimental unit (*ik*-th experimental unit)
u_{ijk}: **Random effect of sub-plot experimental unit [analogous to a normal residual]**

Naïve GLMM

```
proc glimmix data=sp;
class blk var trt ;
model diseased/n = var|trt / dist=binomial link=logit;
random blk blk*var;
run;
```

Conditional GLMM (with unit-level variation)

```
proc glimmix data=sp;
class blk var trt ;
model diseased/n = var|trt / dist=binomial link=logit;
random blk blk*var blk*var*trt;
run;
```

Quasi-likelihood for over-dispersion

```
proc glimmix data=sp;
class blk var trt ;
model diseased/n = var|trt / dist=binomial link=logit;
random blk blk*var;
random _residual_;
run;
```

CaseStudy2.sas

Fitting GLMMs to data

- GLMMs are complex, and there is no universally accepted best method to fit GLMMs to data, although all methods rely on the principles of **maximum likelihood**
- Challenge: there is no analytical solution (equation) for the marginal distribution (likelihood) of the data (i.e., the distribution that one sees)

$$f(\mathbf{Y}) = \int f(\mathbf{Y}\,|\,\mathbf{b}) f(\mathbf{b}) d\mathbf{b}$$

Marginal distribution (**matrix notation**); e.g., $\mathbf{Y}_{ij}$ ~ ***"Distribution"***

Conditional distribution (e.g., binomial). Conditional on random effects (**b**); e.g., $\mathbf{Y}_{ij} \,|\, \mathbf{b}_j \sim Bin(p_{ij}, n)$

Distribution of random effects (**b**; e.g., block effect, block x treatment, etc.), usually normal; e.g., $\mathbf{b}_j \sim N(\mathbf{0}, \boldsymbol{\sigma}_b^2)$

- Two broad approaches based on the likelihood principle, plus Bayesian approaches (latter not covered here):
  1. **Approximate the GLMM model (known as "linearization" or "pseudo-likelihood")**
  2. **Approximate the marginal likelihood (known as Laplace approximation or Quadrature)**

All estimation methods are iterative (singly or doubly).

Fitting GLMMs: First Approach

- **Linearization** or **pseudo-likelihood** (do not confuse with quasi-likelihood)
  - Construct *pseudo-response-variable* ("*pseudo-data*") based on the model and the properties of the assumed conditional distribution (e.g., binomial)
  - Iteratively estimate model terms (e.g., τ_i, σ_b^2) by *maximizing the marginal likelihood of the pseudo-data*, until convergence (see Supplemental Slides)
- Can be used to fit true GLMMs (true conditional distributions) *and* quasi-likelihood (over-dispersion; ϕ) models
  - A very flexible approach with good convergence properties (and fast)!
    - Can handle large (or very small) data sets, and complex models
- Default in GLIMMIX (**`method=rspl`**): for true GLMMs and quasi-likelihood
- Can be done *only* for quasi-likelihood (estimate ϕ) with **`PQLglmm`** {**MASS**} in **`R`**
  - That is, a ϕ parameter is *always* estimated, whether one wants it or not!
  - WARNING: If you want to include a unit-level random effect in model (e.g., u_{ij} for RCBD), algorithm attempts to estimate this random effect and the ϕ residual scale parameter (giving **nonsensical results**—*very bad practice*!)
  - Many other restrictions on the models that can be fitted with **`R`**
- *Originally* thought to produce overly biased parameter estimates (especially with small *n*), new research since 2010 shows that linearization performs well, in general, and is a good choice for the default method

Fitting GLMMs: Second Approach

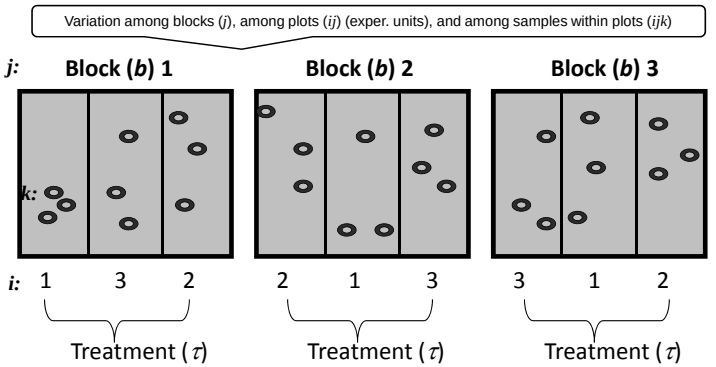
- Maximize the *approximate* marginal likelihood of Y (of the actual data, not of the pseudo-data) with the **Laplace** function or with **Quadrature** (use Laplace or quadrature to approximate the likelihood)
  - Quadrature can be slow, or ***impossibly slow,*** for moderate-to-large datasets and multiple factors, or can require too much computer memory
  - Laplace is almost as accurate as Quadrature, and is much faster and requires far less computational time
  - **Can only be used for *true* GLMMs, i.e., those with actual likelihoods and actual distributions (*not* for quasi-likelihood; cannot estimate ϕ)**
  - Only method to truly evaluate goodness of fit of a model (**can compare models with different random effects** using AIC statistics)
  - Best method to evaluate confidence intervals for variances (when this is an objective) -- **Must remove a random effect if its estimate is zero!**
- Obtained with **`method=laplace`** or **`method=quad`** in GLIMMIX of SAS
- Default with **`glmer`** {**`lme4`**} in **`R`**
  - Thus, `lme4` cannot be used for any quasi-likelihood models (one cannot estimate ϕ), and is restricted to generally *small* GLMM problems (can become impossibly slow)

Go to `CaseStudy1.sas` to see Laplace and quadrature

An overlooked challenge

- With binomial data, GLMM software properly handles Y values equal to 0 and n (remember, the link function [e.g., logit] is for the parameters [e.g., p], not for the response variable)
  - With Poisson data, GLMM software properly handles Y values equal to 0
- There is a problem when **all** the observations (*all* the replicate or block values) for a treatment equal 0 (or all equal n)
  - The estimate of the linear predictor (e.g., logit) for this treatment "tries to go" to minus infinity (-∞), with a corresponding standard error (SE) "trying to go" to positive infinity (+∞): this gives an estimated inverse-link (p) very close to 0
  - This results in *meaningless* hypothesis tests and confidence intervals!
- Call this the "**all-zero problem**"
- No easy solution

Sub-sampling within experimental units: Randomized Complete Block



Each block and each plot (unique ij combination) comprises a cluster, with **nesting** of sampling units within plots (unique ijk combination)

An overlooked challenge, *continued*

- No easy fix for the "*all-zero problem*" based on theory
  - If one knew the treatment must give all zeros, you could remove it
    - Gets tricky with factorials
  - More likely, one needs an *ad hoc* adjustment
    1. Add a small positive number (say, c = 0.5) to the 0 count value for *one* of the reps of just the treatments with this problem (with a corresponding increase in n)
    2. Add a small positive number to all observations (including the non-zeros), with corresponding increase in n
    3. Add a small positive number (0.5) to all zeros in the dataset (and increase in n)
  - Use y' and n' (for one of the above choices)

$$y' = y + c = y + 0.5$$
$$n' = n + 2c = n + 1$$

Go to `CaseStudy4.sas`

Case study 3: RCB plus sub-samples

- Same experiment as case study #1: Effect of fungicide treatment on Phomopsis leaf blight of strawberry
  - Explicitly account for sub-sampling within plots (ij combinations)
- Randomized complete block design (RCBD)
  - Four blocks (j = 1, …, 4), with eight treatments (i = 1, …, 8) randomized within each block
  - **Five sub-samples of n = 15 leaflets each in each experimental unit (each plot; each ij combination); k = 1, …, 5**
    - Previously considered the total of n = 75 leaflets
  - Block, experimental unit (plot), and sampling unit (within plot) are considered random effects, and treatment a fixed effect
- Response variable: **leaflet disease incidence**
  - Number of diseased leaflets (Y) out of n = 15 leaflets in each sub-sample in each experimental unit (plot)
- Analysis: generalized linear mixed model

Probably will not have time to cover

Randomized Complete Block with sub-sampling

Normal *Y* (LMM)

$Y_{ijk} = \theta + \tau_i + b_j + v_{ij} + e_{ijk}$

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

$v_{ij} \sim N(0, \sigma_v^2)$

$e_{ijk} \sim N(0, \sigma_e^2)$

Normal *Y* (LMM)

$\eta_{ij} = \theta + \tau_i + b_j + v_{ij}$

$\mu_{ij} = \eta_{ij}$

$Y_{ijk} \mid b_j, v_{ij} \sim N(\mu_{ij}, \sigma_e^2)$

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

$v_{ij} \sim N(0, \sigma_v^2)$

Y_{ijk}: response (dependent variable) for the k-th observation (sub-sample) within treatment i and block j

θ: constant ("intercept")

τ_i: Effect of the i-th level of treatment on Y

b_j: Random effect of the j-th level of block on Y

v_{ij}: Random effect of the ij-th experimental unit (plot) on Y [same as **block*treatment** or residual when no sub-sampling]

e_{ijk}: Error associated with k-th observation in treatment i of block j [residual] – random **sampling error** (same as block*treatment*sampling-unit interaction)

Randomized Complete Block with sub-sampling

Binomial *Y* (GLMM 1)

$\eta_{ij} = \theta + \tau_i + b_j + v_{ij}$

$\text{logit}(p_{ij}) = \eta_{ij}$

$Y_{ijk} \mid b_j, v_{ij} \sim Bin(p_{ij}, n)$

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

$v_{ij} \sim N(0, \sigma_v^2)$

Naïve: Duplicates results from GLMM of pooled data (case study #1)

Binomial *Y* (quasi-likelihood)

$\eta_{ij} = \theta + \tau_i + b_j + v_{ij}$

$\text{logit}(p_{ij}) = \eta_{ij}$

$Y_{ijk} \mid b_j, v_{ij} \sim quasi\text{-}Bin(p_{ij}, n; \phi)$

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

$v_{ij} \sim N(0, \sigma_v^2)$

Conditional variance = $\phi np(1-p)$

Binomial *Y* (GLMM 2)

$\eta_{ijk} = \theta + \tau_i + b_j + v_{ij} + u_{ijk}$

$\text{logit}(p_{ijk}) = \eta_{ijk}$

$Y_{ijk} \mid b_j, v_{ij}, u_{ijk} \sim Bin(p_{ijk}, n)$

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

$v_{ij} \sim N(0, \sigma_v^2)$

$u_{ijk} \sim N(0, \sigma_u^2)$

Randomized Complete Block with sub-sampling

Normal *Y* (LMM)

$\eta_{ij} = \theta + \tau_i + b_j + v_{ij}$

$\mu_{ij} = \eta_{ij}$

$Y_{ijk} \mid b_j, v_{ij} \sim N(\mu_{ij}, \sigma_e^2)$

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

$v_{ij} \sim N(0, \sigma_v^2)$

Binomial *Y* (GLMM 1: Naïve)

$\eta_{ij} = \theta + \tau_i + b_j + v_{ij}$

$\text{logit}(p_{ij}) = \eta_{ij}$

$Y_{ijk} \mid b_j, v_{ij} \sim Bin(p_{ij}, n)$

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

$v_{ij} \sim N(0, \sigma_v^2)$

Naïve: Duplicates results from GLMM of pooled data (case study #1). No sub-sampling results

`CaseStudy3.sas`

Summary

- Generalized linear mixed models (GLMMs) can accommodate observations from a large number of (conditional) statistical distributions, collected using a wide range of experimental and treatment designs
  - GLMMs can be fitted with "linearization" (pseudo-likelihood) methods or with Laplace/quadrature likelihood methods (the latter for true cond. distributions)
- GLMMs can target the mean or location parameter (e.g., p) of the conditional distribution, which is a major advantage over linear mixed models, which target the mean of the marginal distribution across all levels of the random effects (the latter depending on the variances of the random effects with non-normal data)
- Careful consideration of experimental and treatment design will indicate the proper GLMM to fit to the data
  - For data that nominally have a binomial (or Poisson) discrete conditional distribution, investigators should consider adding an experimental unit-level term in the link function (for a true conditional GLMM), or adding a scale-parameter (ϕ) for a quasi-likelihood analysis
- With GLMMs, one must pay special attention to conditional versus marginal distributions, data- versus model-scale inference, (e.g., logit vs. p), and the approach used for fitting models to data

Other references

- In addition to the books and the Stroup (2015) paper shown earlier, see:
  - Madden, Turechek, & Nita (2002). Evaluation of generalized linear mixed models for analyzing disease incidence data obtained in designed experiments. *Plant Disease* 86: 316-325.
    - Sub-sampling problem (see Case study 3). Article is now a bit dated, and the model labels are not quite the same as used here. This was written before PROC GLIMMIX was developed; the analyses (either for true conditional GLMM or for quasi-likelihood over-dispersion, and even for more complex analyses not covered in workshop) were conducted using a specialized macro in SAS.
    - The GLIMMIX procedure replaced the macro, and added a *great* deal more functionality and model-fitting methods.
    - In 2002, we were slightly in favor of quasi-likelihood. *Now* we slightly favor true conditional GLMMs (based on the most recent statistical research).
  - Kriss, Paul, & Madden, L. V. (2012). Characterizing heterogeneity of disease incidence in a spatial hierarchy: A case study from a decade of observations of Fusarium head blight of wheat. *Phytopathology* 102: 867-877.
    - For observational (survey data). Emphasis on variances and BLUPs.
  - Bolker, et al. (2008). Generalised linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution* 24(3): 127-135.
    - Good article, but the authors are too critical of penalized- and quasi-likelihood estimation methods based on recent statistical research. Best methods depends on circumstances.

A few supplemental slides follow

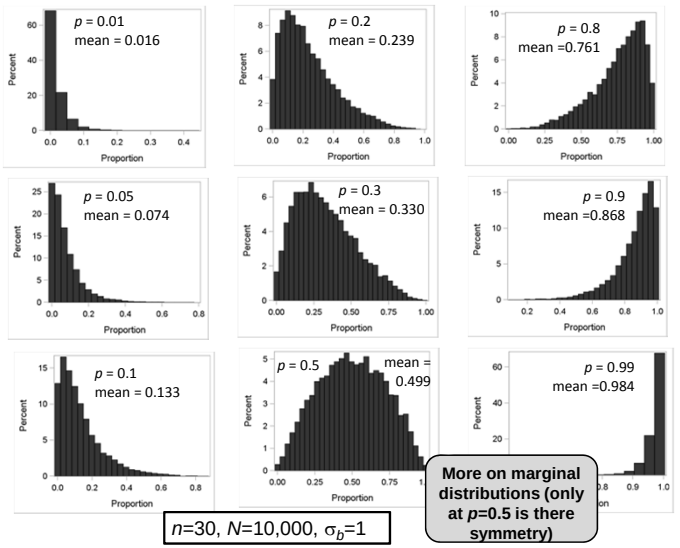
Background on binomial distribution

$$Bin(p,n) \ : \ \Pr(Y = y) = f(y) = \binom{n}{y} p^{n}(1-p)^{n-y}$$

- ***Y* ~*Bin*(*p*,*n*)**
  - ***Y*:** Number of individuals with a trait (e.g., disease) in an experimental unit or sampling unit (e.g., plot, plant) – response
    y: specific value of the random variable ***Y*** **(1, 2, …, *n*)**
  - ***n***: Number of individuals observed for the trait (e.g., plants)
  - ***p***: Parameter: probability of trait, such as disease (e.g. probability that a leaf, plant, etc., is diseased) – parameter (analogous to μ of normal)
  - Variance of ***Y*** : ***np*(1-*p*)**, fully defined by ***n*** and ***p*** (no separate parameter)

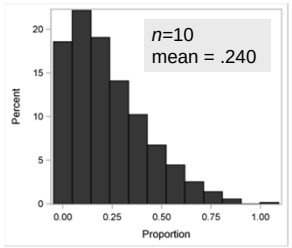
After some algebraic rearrangement, the log of the distribution function (or log of the likelihood when data are substituted) can be written as the following (note the logit within the formula):

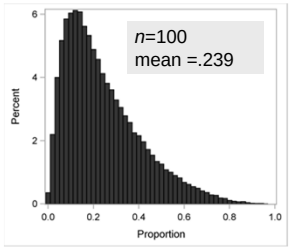
$$\ln(f(y)) = \ln\binom{n}{y} + y\ln\left[\frac{p}{1-p}\right] + n\ln[1-p]$$



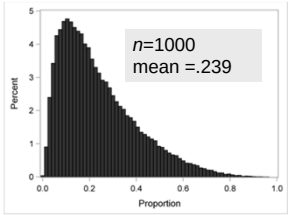
n=30, *N*=10,000, σ_b=1

More on marginal distributions (only at *p*=0.5 is there symmetry)

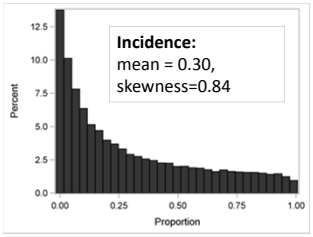


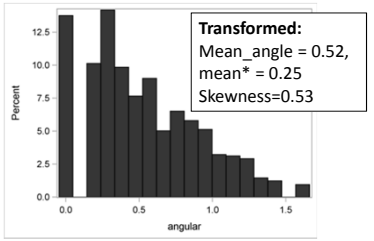


N = 100,000
p = 0.2
σ_b = 1



More on marginal distributions: Increasing *n* (in plot) does not lead to normal distribution





Does transformation of *Y* help? *NO*!

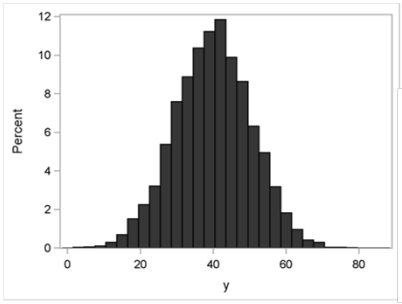
Binomial, with normal **random subject** effect (σ_b **=2**). ***p*=0.2**, ***n*=30**. *N*=100,000 blocks.

*mean** is estimated back-transformation (inverse transformation), which does not equal *p* and does not equal mean of marginal distribution of *Y/n*

Skewness of transformed is not 0.

Compromise (a marginal effect).

Marginal distribution (with normal conditional distribution for plot)



N = 10,000
Block st.dev.: σ_b=1

Plot conditional μ = 40,
Plot conditional st.dev.: σ_e=1

Estimated mean= 40

- Only with Normal conditional distribution (say, for the plot [experimental unit]) does the conditional mean equal the marginal mean
- Only for the Normal conditional distribution does the marginal distribution have the same form as the conditional distribution.

Expected values (means), etc.: RCB

Derivation in next several slides is applicable when all variables have *normal* distributions

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}\,, \qquad b_j \sim N(0,\sigma_b^2),\ e_{ij} \sim N(0,\sigma_e^2)$$

Look at expectations (means), $E(\bullet)$, and variances, $Var(\bullet)$:

$$E(Y_{ij}) = E(\theta) + E(\tau_i) + E(b_j) + E(e_{ij}) = \theta + \tau_i + 0 + 0 = \theta + \tau_i$$

$$\mathrm{Var}(Y_{ij}) = 0 + 0 + \mathrm{Var}(b_j) + Var(e_{ij}) = \sigma_b^2 + \sigma_e^2$$

Note: the expected value for the random effects is 0, by definition
Note: the variance of a fixed effect is 0, by definition

Marginal distribution result: $Y_{ij} \sim N(\theta+\tau_i\,,\ \sigma_b^2+\sigma_e^2)$

Now consider the expectation of Y for an individual block (for the j-th block). This is the *conditional* expected value or conditional mean.

Different model forms (preview)

Traditional approach for normal data

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}$$
$$b_j \sim N(0,\sigma_b^2),$$
$$e_{ij} \sim N(0,\sigma_e^2)$$

Above **cannot** be used for non-normal data (there is no sensible concept of an additive non-normal residual)

Equivalent for normal

Expectation-based approach for normal

$$\eta_{ij} = \theta + \tau_i + b_j$$
$$\mu_{ij} = \eta_{ij}$$
$$Y_{ij} \mid b_j \sim N(\mu_{ij}\,,\ \sigma_e^2)$$
$$b_j \sim N(0,\sigma_b^2)$$

In the next several slides, we derive the model structure that will work with non-normal data (especially binomial). **Be patient!**

$$\eta_{ij} = \theta + \tau_i + b_j$$
$$\mathrm{logit}(p_{ij}) = \eta_{ij}$$
$$Y_{ij} \mid b_j \sim Bin(p_{ij}\,,\ n)$$
$$b_j \sim N(0,\sigma_b^2)$$

Only the expectation approach provides an avenue to model non-normal data (such as the conditional binomial shown here).

Conditional model (conditional means): RCB

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}\,, \qquad b_j \sim N(0,\sigma_b^2),\ e_{ij} \sim N(0,\sigma_e^2)$$

Now: the expectation (mean) *conditional* on random block effect (the conditional expectation), $E(Y_{ij} \mid b_j)$.
That is, consider the expected value of Y for a *specific* block (e.g., j-th block). Then, the random block effect is *locked in* at the specific level.

$$E(Y_{ij} \mid b_j) = E(\theta) + E(\tau_i) + E(b_j) + E(e_{ij}) = \theta + \tau_i + b_j + 0 = \theta + \tau_i + b_j$$

$$\mathrm{Var}(Y_{ij} \mid b_j) = 0 + 0 + 0 + Var(e_{ij}) = \sigma_e^2$$

So, $Y_{ij} \mid b_j \sim N(\theta+\tau_i+b_j\,,\ \sigma_e^2)$ and $b_j \sim N(0,\sigma_b^2)$

Conditional distribution result, or **Conditional** model result

Conditional model, *continued*

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}, \qquad b_j \sim N(0,\sigma_b^2),\ e_{ij} \sim N(0, \sigma_e^2)$$

$$E(Y_{ij} | b_j) = \theta + \tau_i + b_j \qquad Y_{ij} | b_j \sim N(\theta+\tau_i+b_j, \sigma_e^2),\ b_j \sim N(0,\sigma_b^2)$$

- Note: *Conditional mean includes everything except the residual!*
- *The additive residual variance becomes the variance of the conditional distribution.*

New symbol: Define μ_{ij} as the **expectation** (mean) of the response variable for the *i*-th treatment and *j*-th block (a function of the treatment and block); this is the *conditional* expectation.

$$\mu_{ij} = E(Y_{ij} | b_j)$$

Now, write the RCB model as:

$$\mu_{ij} = \theta + \tau_i + b_j, \qquad Y_{ij} | b_j \sim N(\mu_{ij}, \sigma_e^2), \quad b_j \sim N(0, \sigma_b^2)$$

There are now two equivalent ways of writing the same model (for normally distributed data).

Model formulations for normal data

$$\mu_{ij} = \theta + \tau_i + b_j$$
$$Y_{ij} | b_j \sim N(\mu_{ij}, \sigma_e^2)$$
$$b_j \sim N(0, \sigma_b^2)$$

Need to add *another* symbol (*sorry*)!

The right-hand side of the model (without the residual term $[\theta+\tau_i+b_j]$) is known as the **linear predictor**. Based on treatment design and experimental design.

It meets the statistical criterion of being linear

It describes how predictor variables (covariates, factors, …) determine the mean *or a function of the mean*

Excessive notation for normal, but not for non-normal distributions

$$\eta_{ij} = \theta + \tau_i + b_j$$
$$\mu_{ij} = \eta_{ij}$$
$$Y_{ij} | b_j \sim N(\mu_{ij}, \sigma_e^2)$$
$$b_j \sim N(0, \sigma_b^2)$$

A general symbol for the linear predictor is η ("*eta*")

$$\eta_{ij} = \theta + \tau_i + b_j$$

or

$$\mu_{ij} = \eta_{ij} = \theta + \tau_i + b_j$$

Model formulations for normal data

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}$$
$$b_j \sim N(0,\sigma_b^2)$$
$$e_{ij} \sim N(0, \sigma_e^2)$$

Describe Y with additive error, obtain expected values (means) from the equation for Y (and definitions of random effects)

$$\mu_{ij} = \theta + \tau_i + b_j$$
$$Y_{ij} | b_j \sim N(\mu_{ij}, \sigma_e^2)$$
$$b_j \sim N(0, \sigma_b^2)$$

Describe the conditional expected value for the experimental unit, and the distribution of Y for the experimental unit (conditional on any random effects).
Then define the distribution of any random effects.

Two formulations are equivalent (*identical*) here. But only the second approach for defining the model provides the framework for non-normal data.
An additive residual does not work with non-normal distributions (the residual does not have an easy-to-define distribution when the conditional distribution of *Y* is not normal).

Model formulations for normal data

$$\mu_{ij} = \theta + \tau_i + b_j$$
$$Y_{ij} | b_j \sim N(\mu_{ij}, \sigma_e^2)$$
$$b_j \sim N(0, \sigma_b^2)$$

Need to add another symbol!

The right-hand side of the model without the residual term $(\theta+\tau_i+b_j)$ is known as the **linear predictor**.

It meets the statistical criterion of being linear

It describes how predictor variables (covariates, factors, …) determine the mean *or a function of the mean*

$$\eta_{ij} = \theta + \tau_i + b_j$$
$$\mu_{ij} = \eta_{ij}$$
$$Y_{ij} | b_j \sim N(\mu_{ij}, \sigma_e^2)$$
$$b_j \sim N(0, \sigma_b^2)$$

A general symbol for the linear predictor is η

$$\eta_{ij} = \theta + \tau_i + b_j$$

Seems excessive for normal data, but is *required* for non-normal data (as we shall see)

Adjustments (summary)

- True GLMM (using likelihood principles)
  - **Add** a random effect for the basic experimental unit [analogous to a residual term (v_{ij}), but is a true random effect]
- Hybrid approach using **quasi-likelihood**
  - Start with a true GLMM (e.g., fixed effects of treatment and random effects of block), *and* specify a different (larger) variance for the conditional distribution (not allowed by the distribution) – *no longer a true distribution or true GLMM*
- Use a true GLMM (e.g., fixed effects of treatment and random effects of block), but with a *different* conditional distribution (& no v_{ij} unit-level term)
  - i.e., start with naïve GLMM and then switch to a different conditional distribution
  - e.g., **beta-binomial** conditional dist. (for over-dispersion) instead of binomial
  - GLIMMIX does not allow for this conditional distribution, but Stroup (2015) gives a *trick* to achieve almost the same result (using a **beta** conditional distribution with disease proportion as the response variable)
    - Trick only works correctly if there are no 0s or 1s (**all 0s and 1s are converted to missing values!**). Stroup does not mention this – BE CAREFUL!
- Generalized estimating equations (**GEE**) approach: **quasi-likelihood**
  - Remove (i.e., "correct for") the random effects of block (b_j) and plot (v_{ij}), and fit model with quasi-likelihood – *all fixed effects, with over-dispersion terms*
    - Popular in some fields from the pre-true GLMM days (see Stroup 2015)

Not covered

Fitting GLMMs to data, *continued*

- **Second approach:** maximize the approximate marginal likelihood of Y (of the actual data, not the likelihood of pseudo-data) with the **Laplace** function or with **Quadrature**
  - *Singly iterative*, but can be *very slow*
    - Quadrature can be ***impossibly slow*** for moderate-to-large datasets and multiple factors, or require too much computer memory
  - Quadrature is technically more accurate, but Laplace is quite accurate
  - **Can only be used for true GLMMs, i.e., those with actual likelihoods and actual distributions (*not* for quasi-likelihoods) (cannot estimate ϕ)**
  - Only method to truly evaluate goodness of fit of a model (**can compare models with different random effects**)
  - Best method to evaluate confidence intervals for variances (when this is an objective)
    - **Must remove a random effect if its estimate is zero! Very important.**
  - Obtained with **`method=laplace`** or **`method=quad`** in GLIMMIX
  - Default with **`glmer`** {**`lme4`**} in R
    - Thus, `lme4` cannot be used for any quasi-likelihood models (one cannot estimate ϕ), and is restricted to generally *small* GLMM problems

Fitting GLMMs to data, *continued*

- **First approach**: So-called **linearization** or **pseudo-likelihood** (not to be confused with quasi-likelihood)
  - "Doubly iterative approach"(LMM fitting is iterative, in general)
  - In outer iteration step, construct *pseudo-response-variable* (pseudo-data) based on the model (fixed and random effects) and the properties of the assumed conditional distribution (e.g., binomial)
  - In inner iterations, fit weighted LMM to the pseudo-data
  - At (inner) convergence, update the pseudo-data in an outer iteration step (based on the GLMM and conditional distribution)
  - Continue with new inner iterations, and then outer iterations, and so on, until there is convergence with the outer iterations
- Can be used to fit true GLMMs (true conditional distributions) and quasi-likelihood models
  - A very flexible approach with good convergence properties (and fast)!
  - Default in GLIMMIX (**`method=rspl`**): for true GLMMs and quasi-likelihood
  - Can be done *only* for quasi-likelihood models with **`PQLglmm`** {**MASS**} in **`R`**
    - WARNING: If you want to include a unit-level random effect in model (e.g., u_{ij}), you estimate this random effect and the ϕ residual scale parameter (giving **nonsensical results**—*bad statistical practice*!)

Fitting GLMMs to data, *summary*

- "Early" statistical research (mid-1990s) suggested that the linearization method could lead to biased parameter estimates (such as treatment effects) when n (number of observations in a cluster for binomial data) was small
  - For a while, there were many recommendations against this approach
  - However, detailed simulation studies by Stroup and others since 2010 show that the linearization approach holds up well, in general
    - **Linearization is much faster, and can handle *much* larger problems**
  - The Laplace/Quadrature approach is preferred when n is very small (< 10) and when one is trying to compare models for goodness of fit

In SAS/GLIMMIX (default = linearization: **`method=rspl`**)
- Linearization (pseudo-likelihood): true GLMM and quasi-likelihood
- Laplace/Quadrature: true GLMM only (no quasi-likelihood)

Many restrictions in **`R`** (with current software, which is always changing)
- Can *only* fit quasi-likelihood models using linearization/pseudo-likelihood (no true conditional GLMMs can be fitted): **`PQLglmm`** {**MASS**}
- Can fit true GLMMs *only* using quadrature/Laplace (which means relatively small data sets): **`glmer`** {**`lme4`**}

Mixed models (LMM, GLMM)

- A very flexible and general approach to deal with fixed and random effects, and, in particular, account for the effects of experimental design (clustering, in general) on response variables
- Mixed models can accommodate:
  - Data correlated over time or over space
    - Temporal or spatial repeated measures
  - unequal variances,
  - missing values,
  - Lack of “balance”; that is:
    - not all treatments need to be in each block,
    - number of samples (observations) in each experimental unit can be different
  - multiple sources of variation (i.e., multiple random effects),
  - complex experimental designs (including restrictions of randomization),
  - covariance analysis (both factors and continuous co-variables)
  - BLUPs (prediction) and many other features…