A Look at Likelihood Ratio Tests for Variance Components in Mixed Effects Models

Peter Norwood

April 2019

Department of Statistics, North Carolina State University

Table of contents

Introduction

Mixed Effects Models

$$y = X\beta + Z\alpha + e$$

$$\alpha \sim N(\mu, \sigma_1^2 V_1)$$

$$e \sim N(0, \sigma^2 V)$$

- Treat some covariates as fixed values, others as random variables
- Natural Question: is there evidence of a random effect?
- How can we test whether or not there is a random effect?

Main Question

$$H_0 : \sigma_1^2 = 0$$
$$H_a : \sigma_1^2 > 0$$

- Consider the Likelihood Ratio Test
- Asymptotically, the test statistic *should* converge to χ_v^2
- One major problem

Major Problem

$$\lambda(\mathrm{x}) = \frac{L(\hat{\theta}_0|\mathrm{x})}{L(\hat{\theta}|\mathrm{x})}$$

- The likelihood ratio test is based on the maximum likelihood estimation, which is restricted to the parameter space
- Parameter space: $\sigma_1^2 \geq 0$
- $H_0 : \sigma_1^2 = 0$
- Since we are testing on the boundary, what does the test statistic converge to? (Hint: not χ_v^2)

p-values in Mixed Models

"Perhaps I can try again to explain why I don't quote p-values or, more to the point, why I do not take the "obviously correct" approach of attempting to reproduce the results provided by SAS. Let me just say that, although there are those who feel that the purpose of the R Project - indeed the purpose of any statistical computing whatsoever - is to reproduce the p-values provided by SAS, I am not a member of that group." - Doug Bates, 2006

p-values in Mixed Models

"Perhaps I can try again to explain why I don't quote p-values or, more to the point, why I do not take the "obviously correct" approach of attempting to reproduce the results provided by SAS. Let me just say that, although there are those who feel that the purpose of the R Project - indeed the purpose of any statistical computing whatsoever - is to reproduce the p-values provided by SAS, I am not a member of that group." - Doug Bates, 2006

Research Questions

1. Under the null hypothesis, what does the sampling distribution of the likelihood ratio test statistic (LRT) look like? What known distributions can we compare it to?
2. When comparing the LRT to known distributions, what is the type I error rate of the test? How often are we claiming significance when none actually exists?
3. For the different distributions, how powerful is the likelihood ratio test? How much signal is needed to be confident the likelihood ratio test will pick it up?

Methods

Random Intercept, Random Slope Model

$$y_{ij} = \beta_1 x_i + \alpha_j + \gamma_{i(j)} x_i + e_{ij}$$

$$i = 0, .., 9, x_i = i$$

$$j = 1, .., b$$

$$\alpha_j \sim N(0, \sigma_1^2), \gamma_{i(j)} \sim N(0, \sigma_2^2)$$

$$e_{ij} \sim N(0, \sigma^2)$$

We do not specify or restrict any correlation structure between α_j and $\gamma_{i(j)}$. We assume α_j and $\gamma_{i(j)}$ are independent of e_{ij}. This is based off the *sleepstudy* dataset.

Split Plot Model

$$y_{ijk} = \mu + \alpha_i + \gamma_{k(i)} + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

$$i = 1, .., 3, j = 1, .., 4, k = 1, ., b$$

$$\gamma_{k(i)} \sim N(0, \sigma_1^2)$$

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

We assume e_{ijk} is uncorrected with $\gamma_{k(i)}$. Based of the *corn* dataset.

Simulation

Random Intercept, Random Slope

- Vary subjects: 2, 5, 10, 20, 50, 100, 200, 500, 1000
- Vary signal $\sigma_1 = \sigma_2$: 0 to 25 by 5

Split Plot

- Vary replicates: 1, 2, 4, 8, 16, 32, 64, 128, 256
- Vary signal σ_1: 0 to 2.5 by 5

10,000 simulated datasets for each model under each sample size, signal combination

Simulation

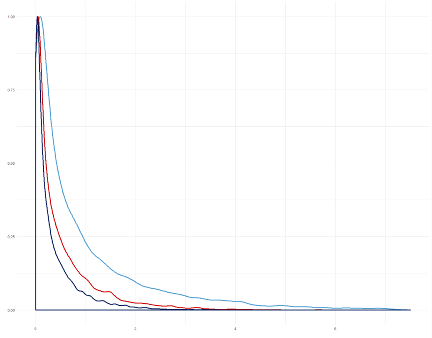
- *lme4* package does not give p-values or LRTs for random effects
- *lmerTest* gives p-values for fixed and random effects and the LRT for random effects

```
fit <- lmerTest::lmer(reaction ~ day + (day|subject),
                      data=dat)

fit <- lmerTest::lmer(yield~pesticide*method +
                      (1|field:pesticide), data=dat)
```

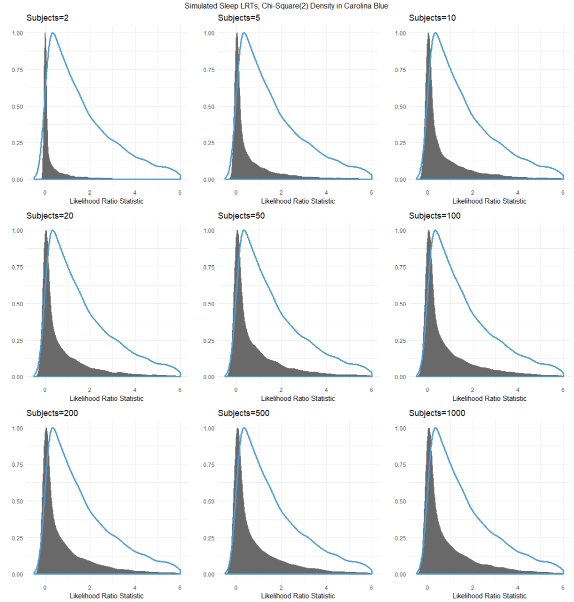
Distributions

- Basic chi-square distribution: χ^2_v
- Custom chi-square mixture distribution: $p\chi^2_0 + (1-p)\chi^2_v$
- 50/50 chi-square mixture distribution: $.5\chi^2_0 + .5\chi^2_v$
  Zhang, Daowen and Lin, Xihong (2008)

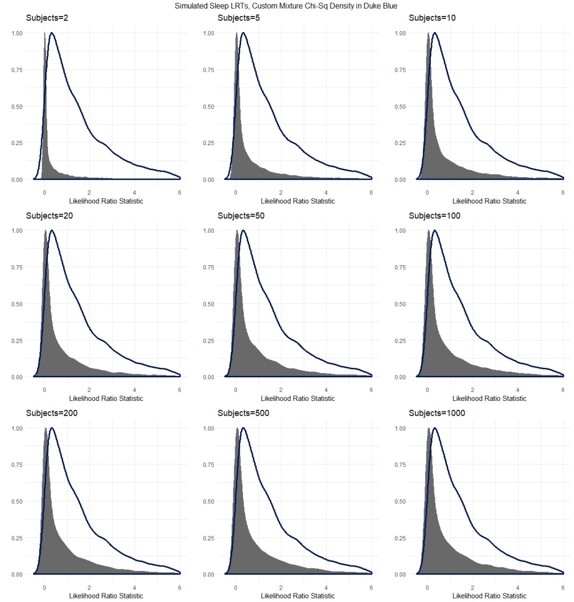


Results

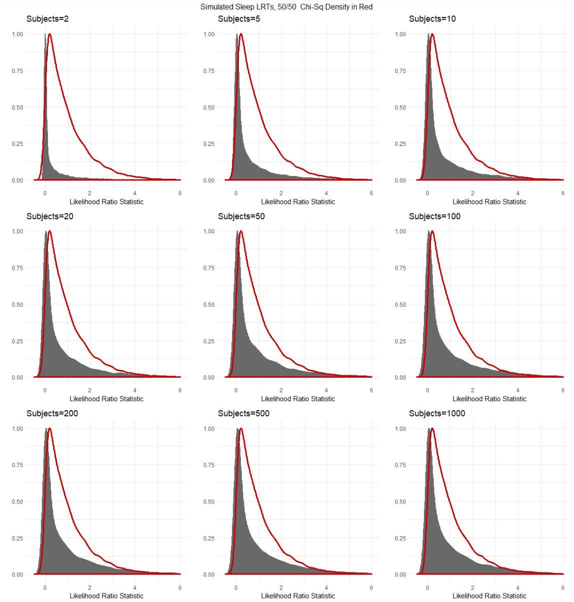
Random Intercept, Random Slope LRT Distribution



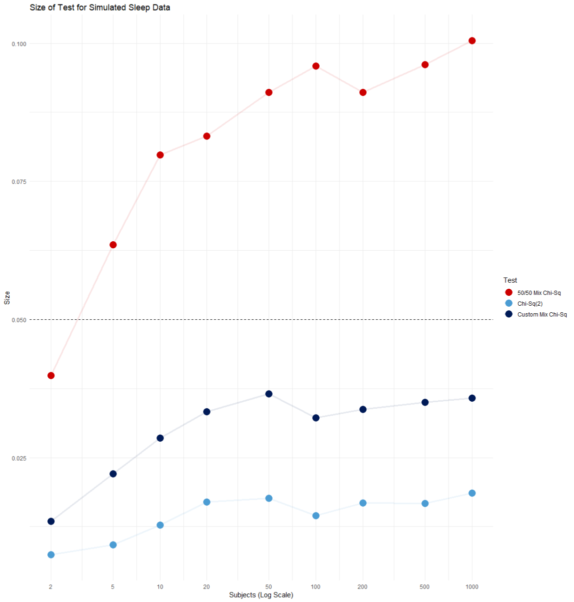
Random Intercept, Random Slope LRT Distribution



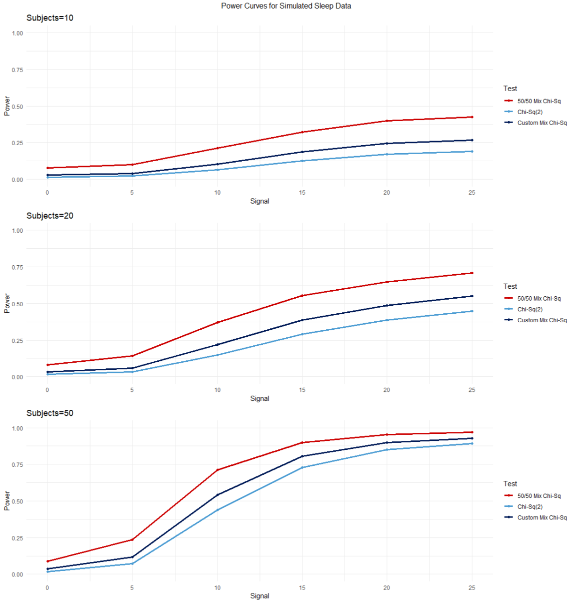
Random Intercept, Random Slope LRT Distribution



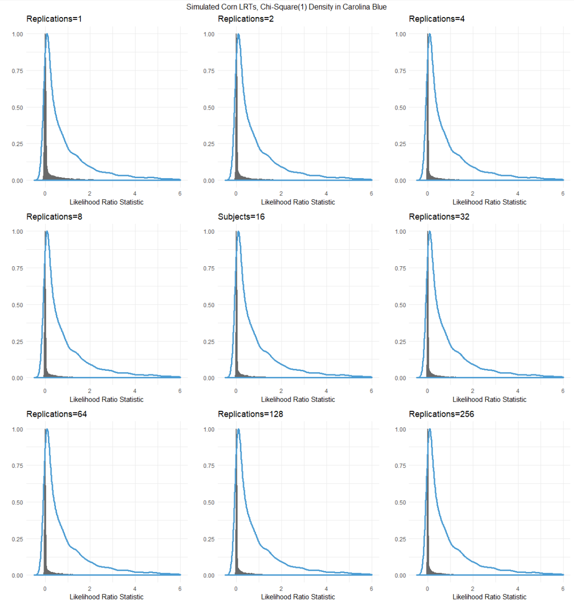
Random Intercept, Random Slope Type I Error



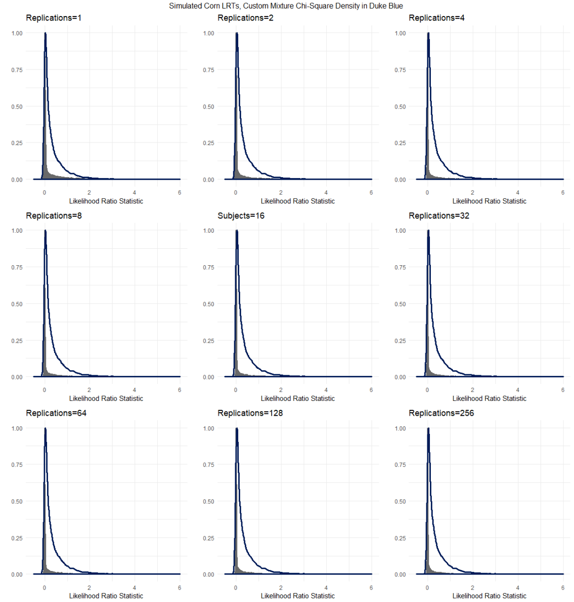
Random Intercept, Random Slope Power



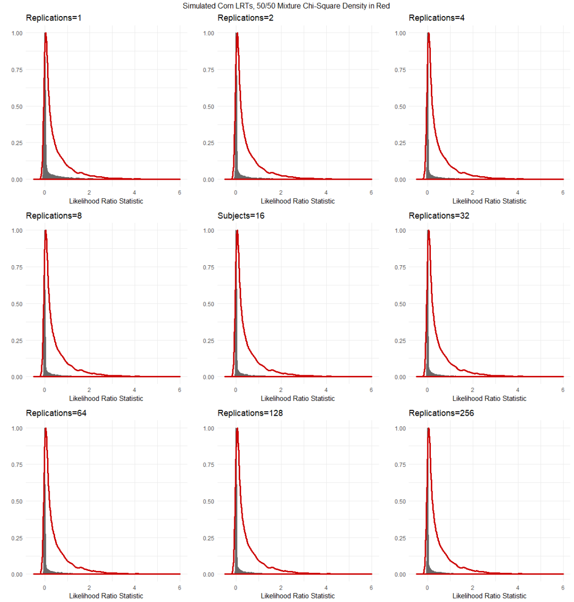
Split Plot LRT Distribution



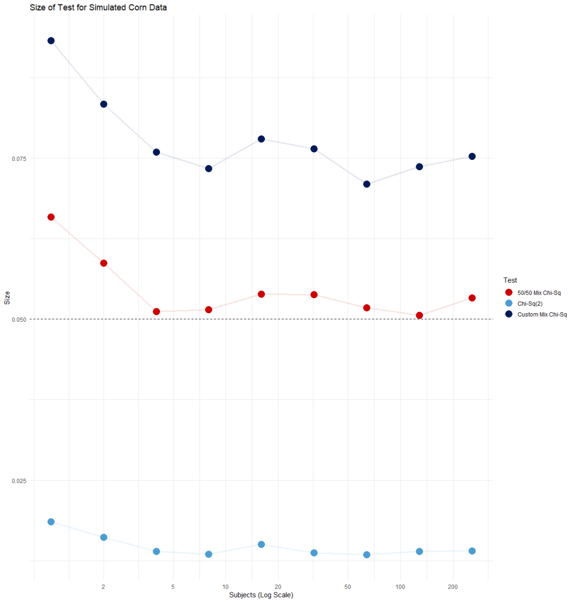
Split Plot LRT Distribution



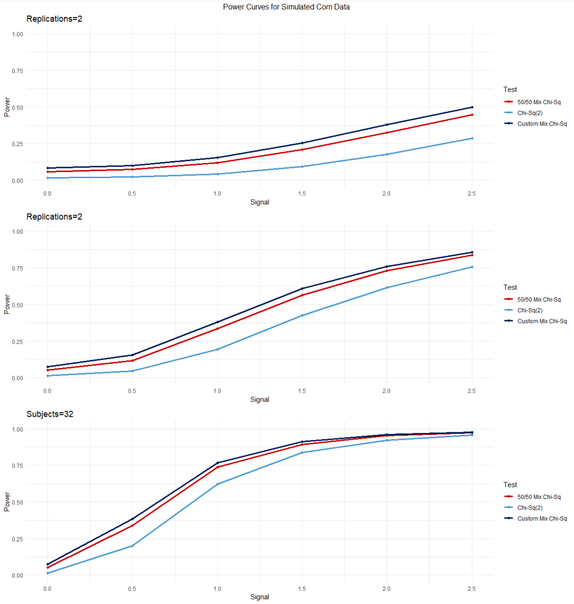
Split Plot LRT Distribution



Split Plot Type I Error



Split Plot Power



Discussion

Chi-Square Distribution

- Does not pass the eye test
- Conservative for both models
- Lower power compared to the mixture distributions

Custom Mixture Distribution

- Visually fits better than chi-squared
- Slightly conservative for random intercept, random slope
- Liberal for split plot model
- Lower power than 50/50 for random intercept, random slope, slightly better for split plot

50/50 Mixture Distribution

- Visually fits better than chi-squared
- Liberal for random intercept, random slope
- Close to correct size for split plot model
- Better power than chi-square, better or close to custom mixture

Main Takeaways

- No one distribution is the perfect fit
- When $p > .5$, 50/50 looked solid. When $p < .5$, 50/50 was too liberal.
- If p-values are necessary consider if type I error or type II error is more grave.
- Best to look at p-values from both distributions. If they both render the same decision, we can be comfortable with the conclusion.
- Bootstrap confidence intervals and other distribution-free evaluations can be better if the answer isn't clear from these two distributions.

Future Work

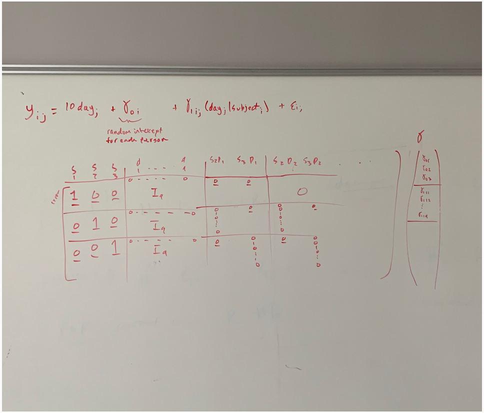
- Same study, more models
- Similar study, new tests. Zhang, Daowen and Lin, Xihong (2008) shows Score tests have appropriate size and solid power

Table 1: *Size and Power comparisons of the likelihood ratio tests and score tests for a single variance component based on 500 simulations under the logistic model (15)*

Method	Size	Power				
	$\psi = 0$	$\psi = 0.2$	$\psi = 0.4$	$\psi = 0.6$	$\psi = 0.8$	$\psi = 1.0$
LRT	0.034	0.370	0.790	0.922	0.990	1.000
Regular LRT	0.020	0.280	0.672	0.882	0.968	0.992
One-sided score test	0.054	0.416	0.834	0.938	0.996	1.000
Two-sided score test	0.050	0.336	0.736	0.910	0.980	0.998

Problems Encountered

- Simulating the data correctly
- R packages
- Knowing when to stop/settle
- Computational Time



Application

Sleep Study Dataset p-values

The likelihood ratio statistic is 42.84.

Chi-Sq(2)	Custom Chi-Sq Mixture	50/50 Chi-Sq Mixture
4.99e-10	0	0

Even with the problems previously discussed, it is clear random variation is coming from the day nested within subject.

Corn Dataset p-values

The likelihood ratio statistic is 7.30.

Chi-Sq(1)	Custom Chi-Sq Mixture	50/50 Chi-Sq Mixture
0.00691	0.00001	0.00014

Again, it is clear that the field nested within pesticide – the sub plot error – is introducing random variation.

Thank you

- Dr. Hughes-Oliver
- Dr. Maity

Questions?

References

1. Doug Bates. [r] lmer, p-values and all that.
2. Daowen Zhang and Xihong Lin. *Variance Component Testing in General-ized Linear Mixed Models for Longitudinal/Clustered Data and other RelatedTopics*, pages 19-35. 01 2008.