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

Peter Norwood

April 2019

Contents

1 Introduction **3**

2 Research Questions **4**

3 Methods **5**
3.1 Random Intercept, Random Slope Model 5
3.2 Split Plot Model 6
3.3 Distributions To Consider 6
3.4 Simulation 7
3.5 Evaluating the Tests 8

4 Results **10**
4.1 Random Intercept, Random Slope Model 10
4.2 Split Plot Model 15

5 Discussion **20**
5.1 Review of Results 20
5.2 Conclusion 21
5.3 Future Work 21

6 Application **22**

7 Code Appendix **23**

1 Introduction

Random effects and mixed effects models are used to model a response using covariates that may be treated as random variables. One may use a random effect to simply account for additional variation or to extend the model to the wider population which the observed covariates are a sample of.

Like fixed effects, these natural questions arise: is there evidence of this random effect? What does the random effect contribute to the model? How can we estimate the random effect, what is a range of reasonable values for the variation it provides?

This paper mainly concerns itself with the first question: does this random effect exist? This can be expressed as: does the random effect account for variation in the model? In terms of a testable hypothesis we have:

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

where σ_1^2 is the variance of the random effect.

Like many hypotheses, a likelihood ratio test is a reasonable place to start. Asymptotically the likelihood ratio test statistic (commonly referred to as LRT in this report) converges to a χ_v^2 distribution where v is the number of parameters we are testing. The critical issue here is the range of σ_1^2, which is a non-negative value. Since the likelihood ratio statistic is dependent on likelihood-based estimators, which are constrained to the parameter space, we end up testing on the boundary of the parameter space. Thus the distribution of the LRT under the null hypothesis may not converge to χ_v^2.

Due to this and other issues with hypothesis testing within mixed effects models, the writers of the *lme4* package in R – the most popular package for fitting mixed effects models – refuse to include p-values in their output. Procedures in SAS that fit mixed effects models do include p-values for a variety of asymptotic tests on variance components. Doug Bates (2006), a main developer of the *lme4* package, is skeptical to say the least:

"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." [1]

This report's main purpose is to investigate the issue of testing variance components in mixed effects models specifically using the likelihood ratio test. It looks at the convergence of the test statistic under the null hypothesis and uses different distributions to explore what may be an appropriate and well performing test.

2 Research Questions

The main research question is: are we able to responsibly and accurately test whether random effects exist in a certain model? We ask: for $H_0 : \sigma_1^2 = 0$ vs $H_a : \sigma_1^2 > 0$, is there a valid statistical test? If so, how does that test perform?

This report specifically considers whether the likelihood ratio test is able to test this hypothesis. To gauge whether or not the likelihood ratio rest can, we must consider:

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?

3 Methods

To answer our research questions, we perform Monte Carlo simulations for two different mixed effects models – the random intercept, random slope model and a split-plot model. We vary both the sample size and the amount of variation from the random effect to get a gauge of the distribution of the LRT under the null hypothesis, the size of the tests, and the power of the tests.

3.1 Random Intercept, Random Slope Model

The first model we consider is the random intercept, random slope model. This model is based on the *sleepstudy* dataset commonly used for teaching mixed effects models. The model is:

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

Here y_{ij} is the response, the average reaction time from sleep deprived patients in a study.

x_i is the day in the study $i = 0, .., 9$. β_1 is the increase in reaction time for every one increase in day, a fixed effect. For our study we set this to 10.

α_j is the random intercept effect from each patient in the study. Each patient has their own baseline for reaction time, but we are not interested in necessarily estimating these baselines, just accounting for them. Hence the effect is random. $\alpha_j \sim N(0, \sigma_1^2), j = 1, .., b$. For our different scenarios, we have $b = 2, 5, 10, 20, 50, 100, 200, 500, 1000$.

$\gamma_{i(j)}$ is the random slope effect from each day for each patient. This is a nested effect, nested within each patient. We are accounting for each patients' day-by-day development after sleep deprivation. $\gamma_{i(j)} \sim N(0, \sigma_2^2)$ For our simulations we set $\sigma_1 = \sigma_2$ and vary the standard deviations from 0 to 25 by 5. We allow the random effects to have covariance, but do not specify any structure.

e_{ij} is the random error after accounting for all other factors. $e_{ij} \sim N(0, \sigma^2)$. For our simulations we set $\sigma = 56$. Both random effects are uncorrelated with the error term.

3.2 Split Plot Model

The second model is a split model model based on the corn yields dataset used extensively in PhD level statistical methods courses at NC State University. The model is:

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

Here y_{ijk} is the corn yield from a certain field nested in a certain planting method receiving a certain pesticide.

α_i is the fixed effect for each pesticide. $i = 1, .., 3$. β_j is the fixed effect for each planting method, $j = 1, .., 4$. $(\alpha\beta)_{ij}$ is the pesticide-method fixed interaction term.

$\gamma_{k(i)}$ is the random effect coming from a certain field nested in each method within each pesticide. We simulate samples from $k = 1, 2, 4, 8, 16, 32, 64, 128, 256$. $\gamma_{k(i)} \sim N(0, \sigma_1^2)$. We vary σ_1 from 0 to 2.5 by .5.

e_{ijk} is the error after accounting for all other factors. $e_{ijk} \sim N(0, \sigma^2)$ and e_{ijk} is uncorrected with $\gamma_{k(i)}$. We set $\sigma = 5$ for all simulations.

The varying sample sizes and standard deviations for both models are based on exploratory work and computational considerations. The larger signal levels are where the power is close to 1 and the sample sizes are large enough that a convergence of the sampling distribution is seen.

3.3 Distributions To Consider

If we are testing $H_0 : \sigma_1^2 = c$ with c much larger than 0, then a χ_v^2 distribution where v is the appropriate degrees of freedom should be the asymptotic distribution of the LRT under null. We will use this central χ_v^2 distribution as our baseline. The degrees of freedom for the random intercept, random slope model is two, since we are testing two variance components. We have one degree of freedom for the split plot since we are testing one variance component. The critical region for these distributions is the 95th quantile of the distributions.

From exploratory simulations, we see many LRTs under null either zero or less than zero. Since χ_1^2, χ_2^2 are non-negative distributions, a mixture χ^2 distribution – a weighted sum of multiple χ^2 distributions – with a weight on χ_0^2 is potentially appropriate. Thus we have:

$$w_1\chi_0^2 + w_2\chi_v^2$$

where v is the appropriate degrees of freedom. The first approach for this, what we can call the naive approach, is to set w_1 equal to the sample proportion of the LRTs (p) that are less than or equal to zero. We code this for LRTs less than .01 to include LRTs that are practically zero. We use the sampling distribution from the largest sample size to find p. Then our mixture is:

$$p\chi_0^2 + (1-p)\chi_v^2$$

This approach has a clear weakness – one must sample thousands of datasets generated under the null hypothesis to come up with p. Hence it is our naive approach to be compared against more realistic alternatives.

The last distribution to consider is a result discussed by Zhang, Daowen and Lin, Xihong (2008) [2]. They cover many different cases of variance component testing for generalized linear mixed models. Our cases fall under their first case which suggests comparing the LRT to a 50/50 χ^2_v mixture. That is:

$$.5\chi^2_0 + .5\chi^2_v$$

Throughout this paper, we will refer to the first distribution simply as the chi-squared distribution, the second as the custom chi-squared mixture, and the third as the 50/50 chi-squared mixture.

3.4 Simulation

As previously discussed, the simulations vary both the amount of variation from the random effect and the sample size. Of course the sample size is measured differently, with more subjects in the random intercept, random slope model and through replications in the split plot design.

We use the *lme4* package to fit the mixed effects models and the *lmerTest* package to extract the LRTs from those models.

The response in the random intercept, random slope model is constructed as follows:

```
sim_sleep$reaction <- 10*sim_sleep$day + ## fixed effect of day
                      rep(rand_int,each=10) + ## random intercept
                      rnorm(subjects*day,mean=0,
                      sd=sigmaDAY)*sim_sleep$day + ## random effect of day|subject
                      rnorm(length(sim_sleep$day),mean=300,sd=sigma) ## error
```

and the model was fit as follows:

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

The response for the split-plot model is generated as follows:

```
## everything without a rnorm multiplier is a fixed effect
## everything with is a random effect
sim_corn$yield <- 63.6 -
                8.2*I(corn$pesticide==1) + 2.050*I(corn$pesticide==2) -
                6.95*I(corn$method==1) - 4.2*I(corn$method==2) - 3.8*I(corn$method==3) +
                1.5*I(corn$pesticide==1 & corn$method==1) -
                2.95*I(corn$pesticide==2 & corn$method==1) +
                1.25*I(corn$pesticide==1 & corn$method==2) -
                4.85*I(corn$pesticide==2 & corn$method==2) +
                2.10*I(corn$pesticide==1 & corn$method==3) -
                0.85*I(corn$pesticide==2 & corn$method==3) +
                rnorm(1,mean=0,sd=sigma2)*I(corn$field==1 & corn$pesticide==1) +
                rnorm(1,mean=0,sd=sigma2)*I(corn$field==1 & corn$pesticide==2) +
                rnorm(1,mean=0,sd=sigma2)*I(corn$field==1 & corn$pesticide==3) +
                rnorm(1,mean=0,sd=sigma2)*I(corn$field==2 & corn$pesticide==1) +
                rnorm(1,mean=0,sd=sigma2)*I(corn$field==2 & corn$pesticide==2) +
                rnorm(1,mean=0,sd=sigma2)*I(corn$field==2 & corn$pesticide==3) +
                rnorm(length(sim_corn$pesticide),mean=0,sd=sigma1)
```

and the model was fit as follows:

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

The fixed effects for both models are close to the estimates from the real data. Since this report is not concerned with fixed effects, no additional thought was put towards them.

We simulate 10,000 datasets from each sample size (nine sizes) and each signal level (six levels) for both models.

3.5 Evaluating the Tests

Since many LRTs, especially under the null case are below zero, we go ahead and set them all to zero.

To obtain an understand of what the distributions of the LRTs under null look like, we plot the sampling distribution of the LRTs from each sample size and also plot density curves from the chi-squared and mixture chi-squared distributions. We scale the y-axis for both the density curves and the histograms. These plots are organized by each model in a matrix to understand what the sampling distribution is converging to as the sample size increases.

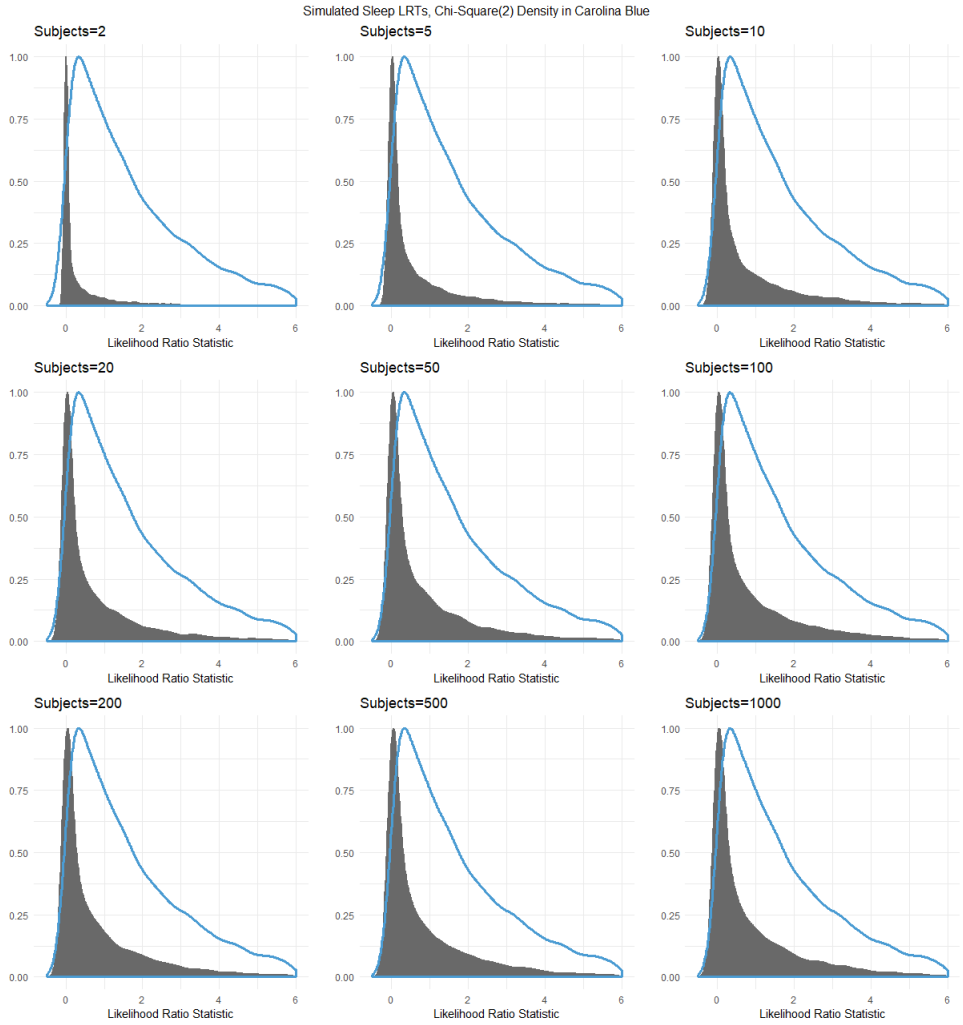
To measure the size of each of the three tests, we gather the LRTs from the null setting and calculate what percentage of them are above the 95th quantile for the respective distribution. For the chi-squared distribution, we simple grab the 95th percentile using *qchisq()*. For the mixture distributions, we generate samples of 10,000 observations and select the 95th quantile from the sample for the cutoff. We calculate the size of the test for all sample sizes for both models.

To evaluate power, we aggregate data from more reasonable sample sizes for each model. We select 10, 20, and 50 subjects for the random intercept, random slope model and we select 2, 8, and 32 replicates for the split-plot model. We then calculate the proportion of datasets where we reject the null hypothesis for each signal level.

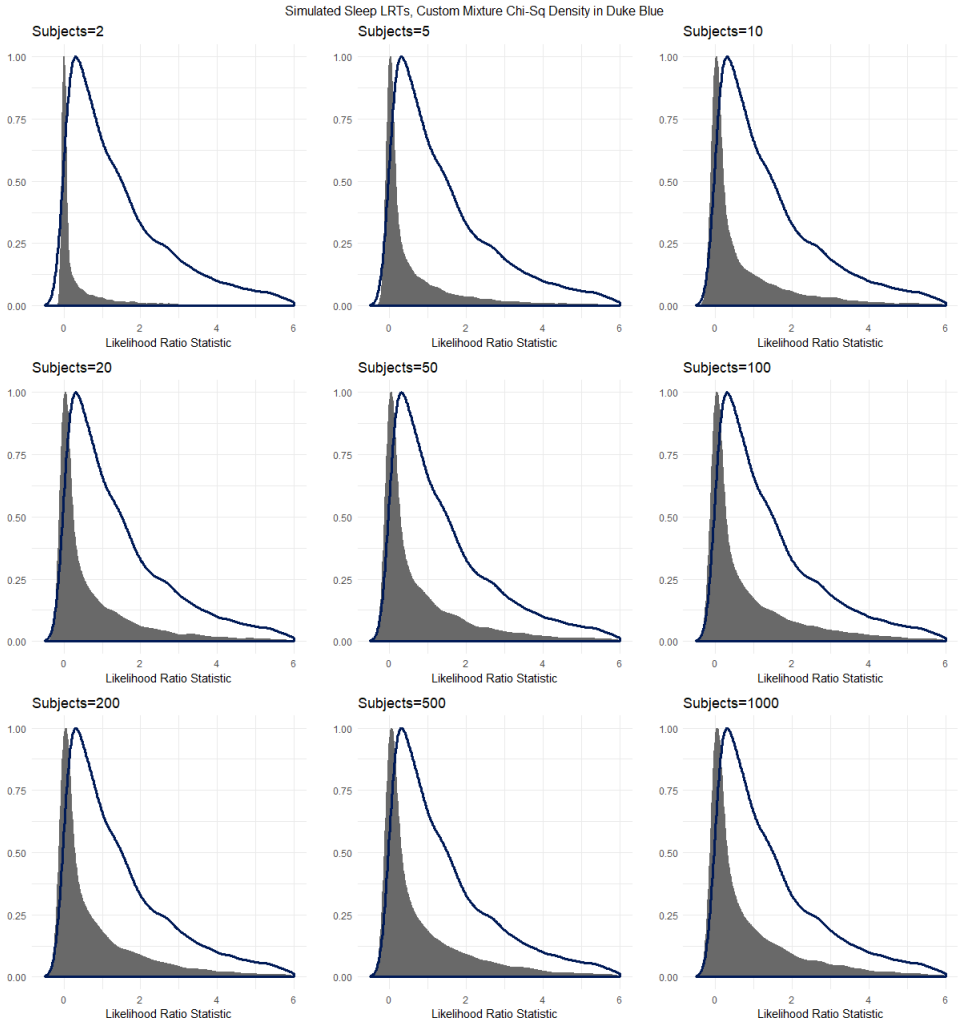
Packages in the *tidyverse*, specifically *dplyr* for data manipulation and *ggplot2* for plotting were extensively used.

4 Results

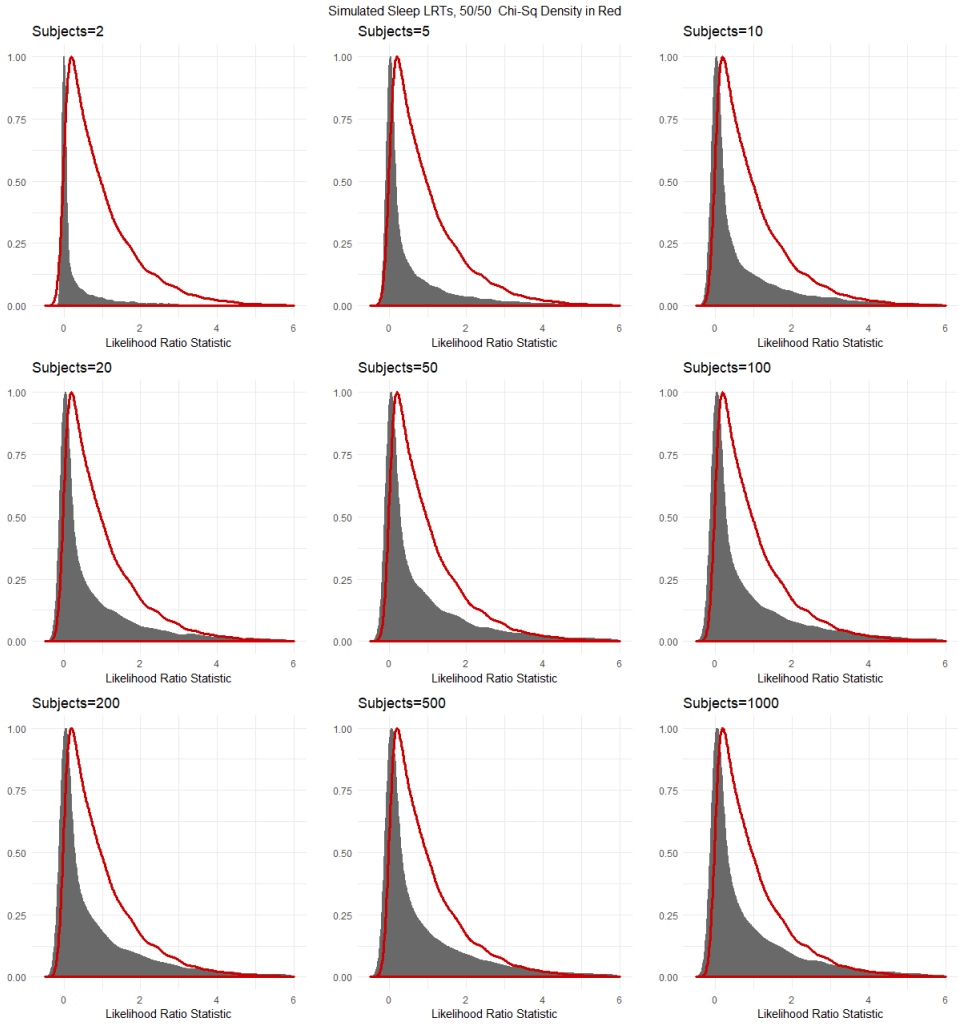
4.1 Random Intercept, Random Slope Model



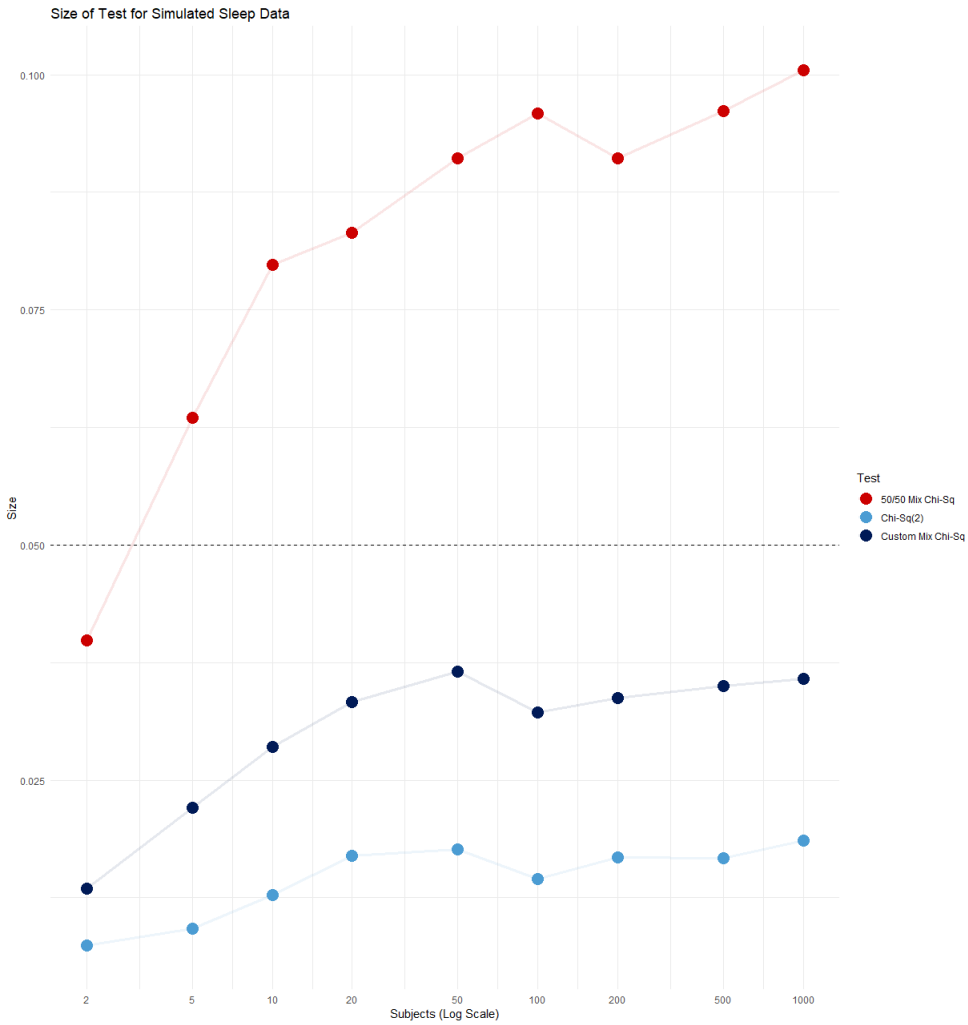
Here we can see how the distribution of the test statistic under null changes as the sample size increases. We can see that it tends to fill out and has the general form of a chi-square. When comparing it to a chi-square distribution with two degrees of freedom, we see it is still peakier and has a mode slightly to the left of the chi-square distribution.



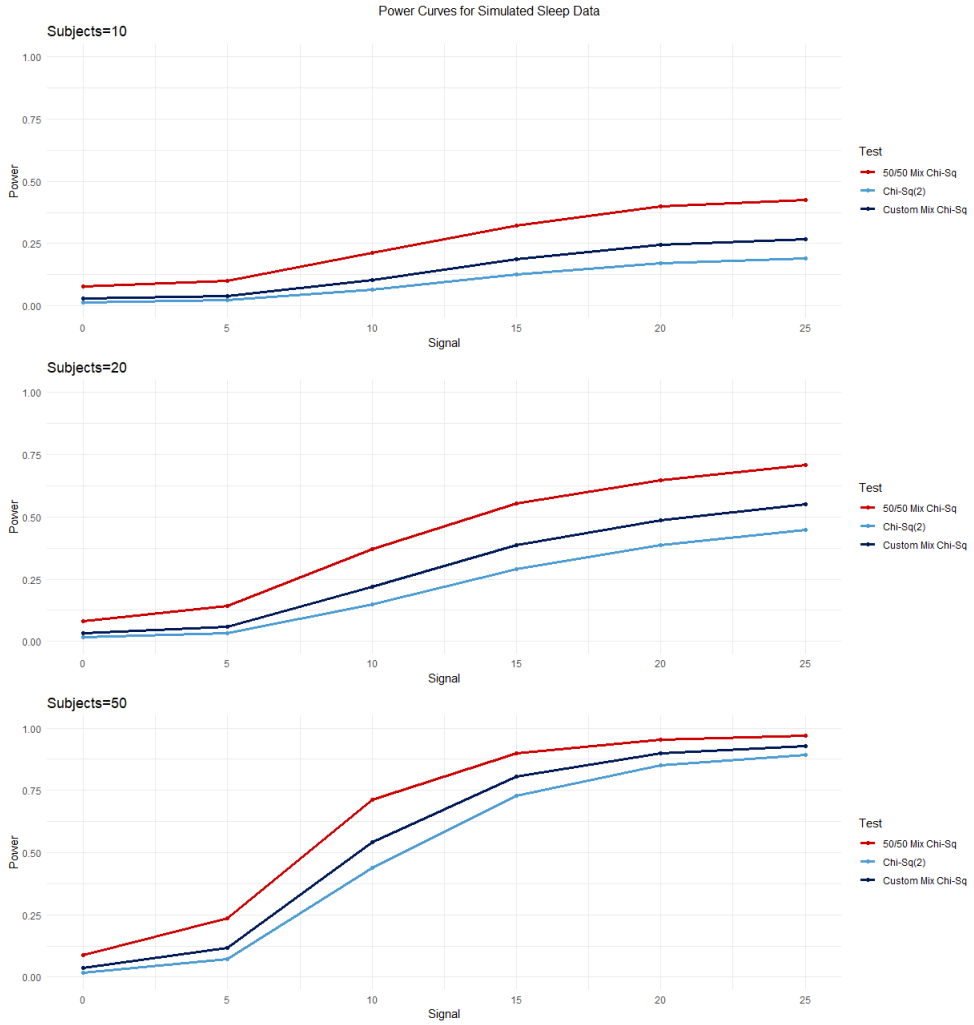
Next we compare the test statistic to the custom chi-square distribution. Here we find 20.25% of the LRTs from the largest sample size case are practically zero, hence our weight for the χ^2_0 part of the mixture is .2025. We can see that this fits the data better than the non-mixture chi-square distribution, but still has some of the same issues – more of a peak at zero, mode shifted to the left.



Finally we see that the 50/50 chi-square mixture is clearly the best distribution – of the three we are comparing – to compare to the test statistic visually. The test statistic distribution still has more of a peak, but compared to the other two distributions, the 50/50 chi-square mixture visually looks the best.

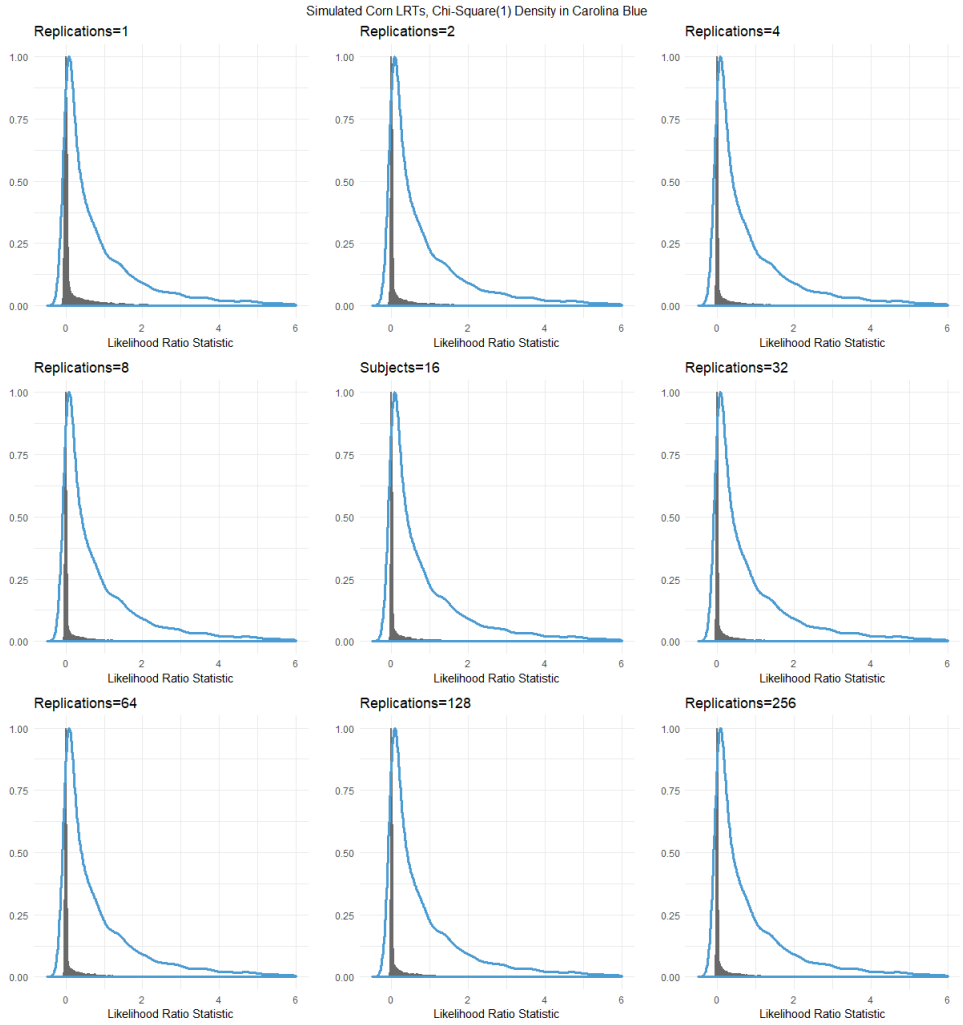


Here we evaluate how often we reject the null hypothesis – with a significance level of .05 – when there is no random effect. We can see the 50/50 chi-square mixture is the most liberal test, committing type I errors more often than 5% of the time in all but one of our sample sizes. Both the custom mixture and the true chi-square commit type I errors less than 5% of the time.

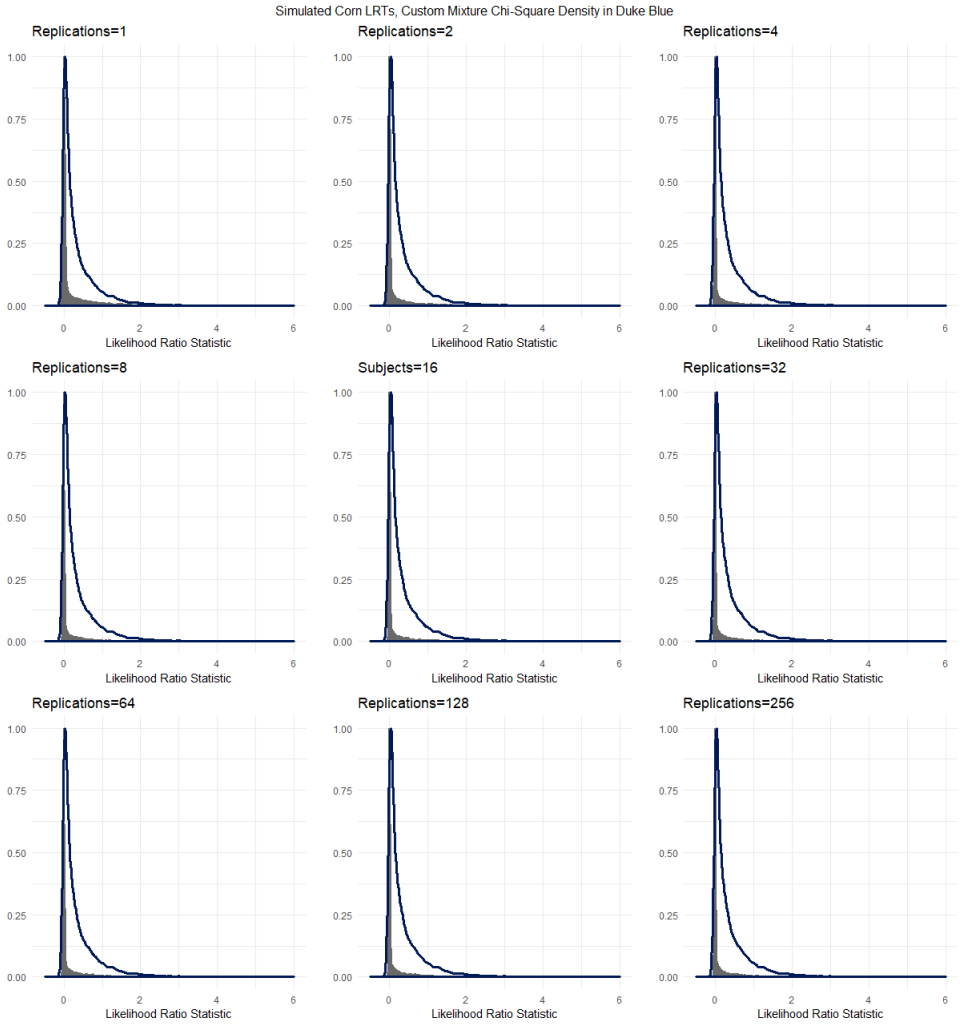


Next we look at the power of these tests as our standard deviations from both random effects rise to 25 ($\sigma_1 = \sigma_2$). The 50/50 chi-square mixture has the highest power for all levels and all sample sizes. The custom mixture then has better power than the chi-square distribution.

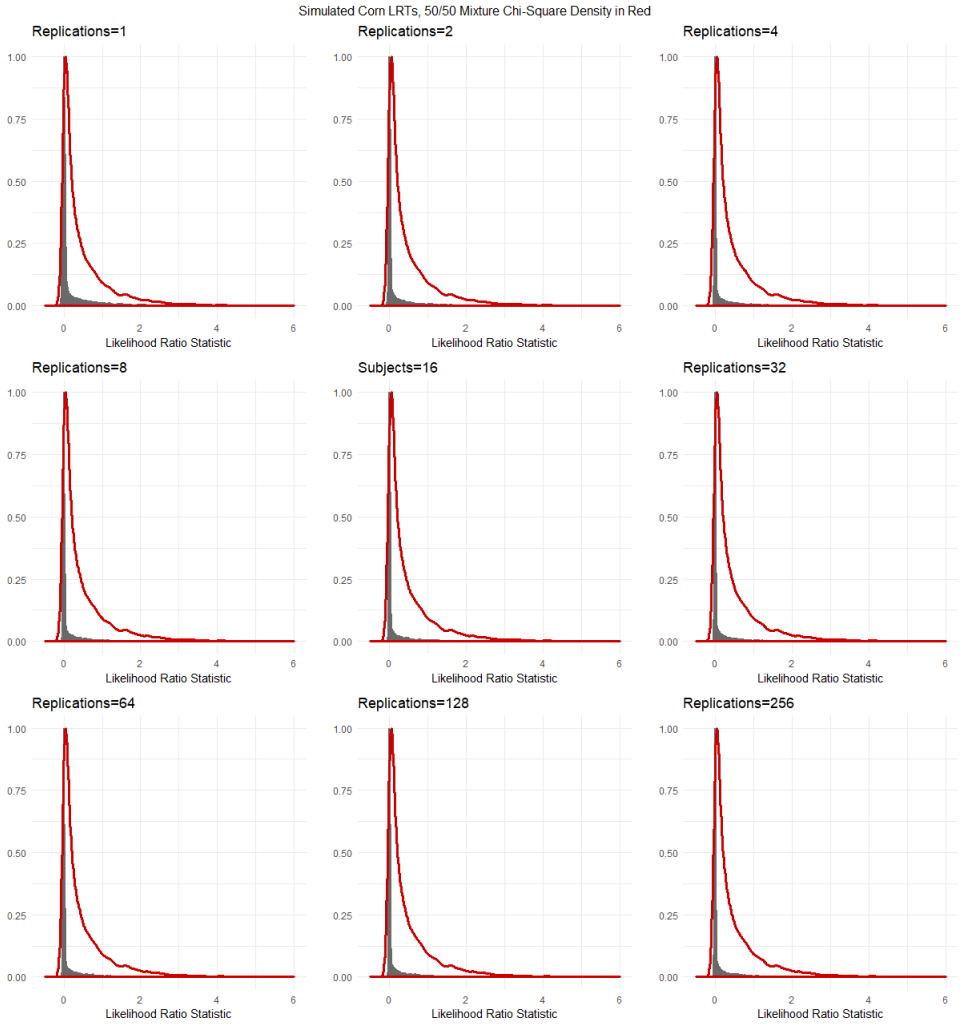
4.2 Split Plot Model



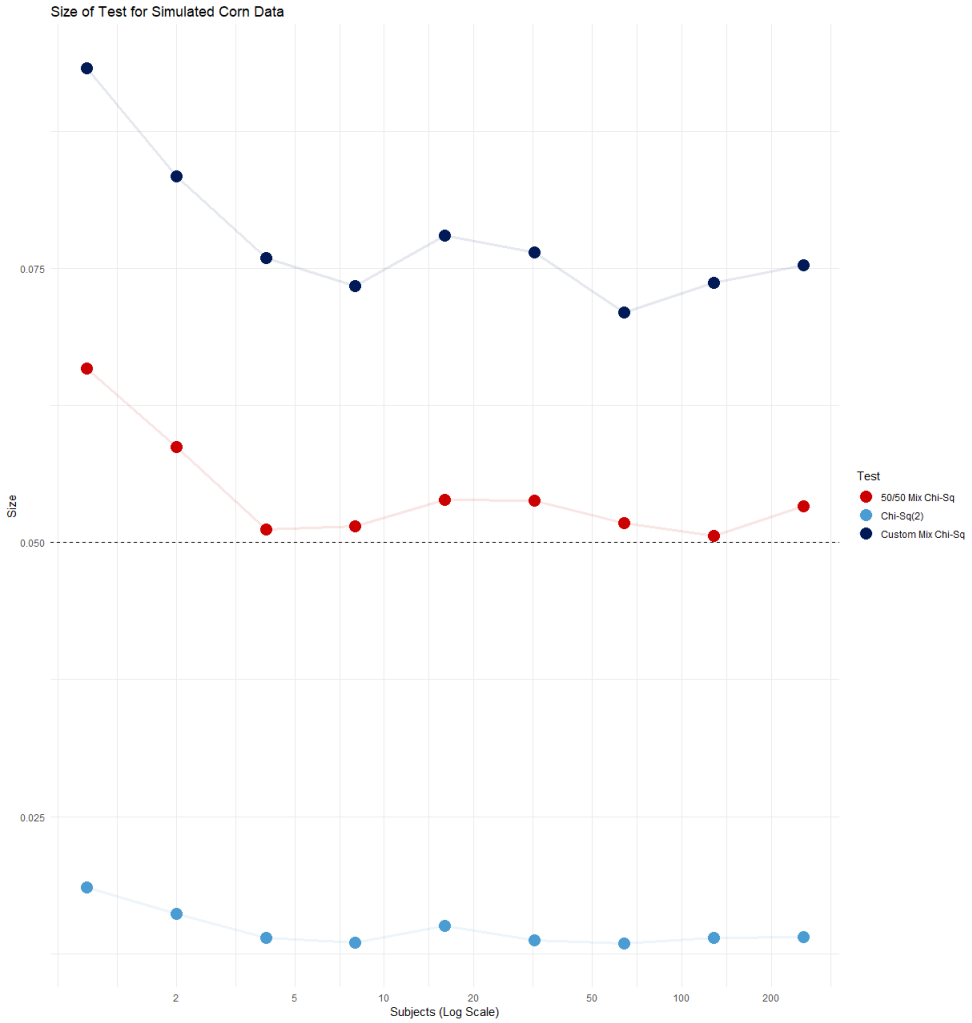
For the split plot model, the distribution of the LRT has a stronger peak than the random intercept, random slope model. A higher proportion of the test statistics, under null, are practically zero. We can see here the distribution also has the same general form of a chi-square, but the fit is less desirable than the previous model.



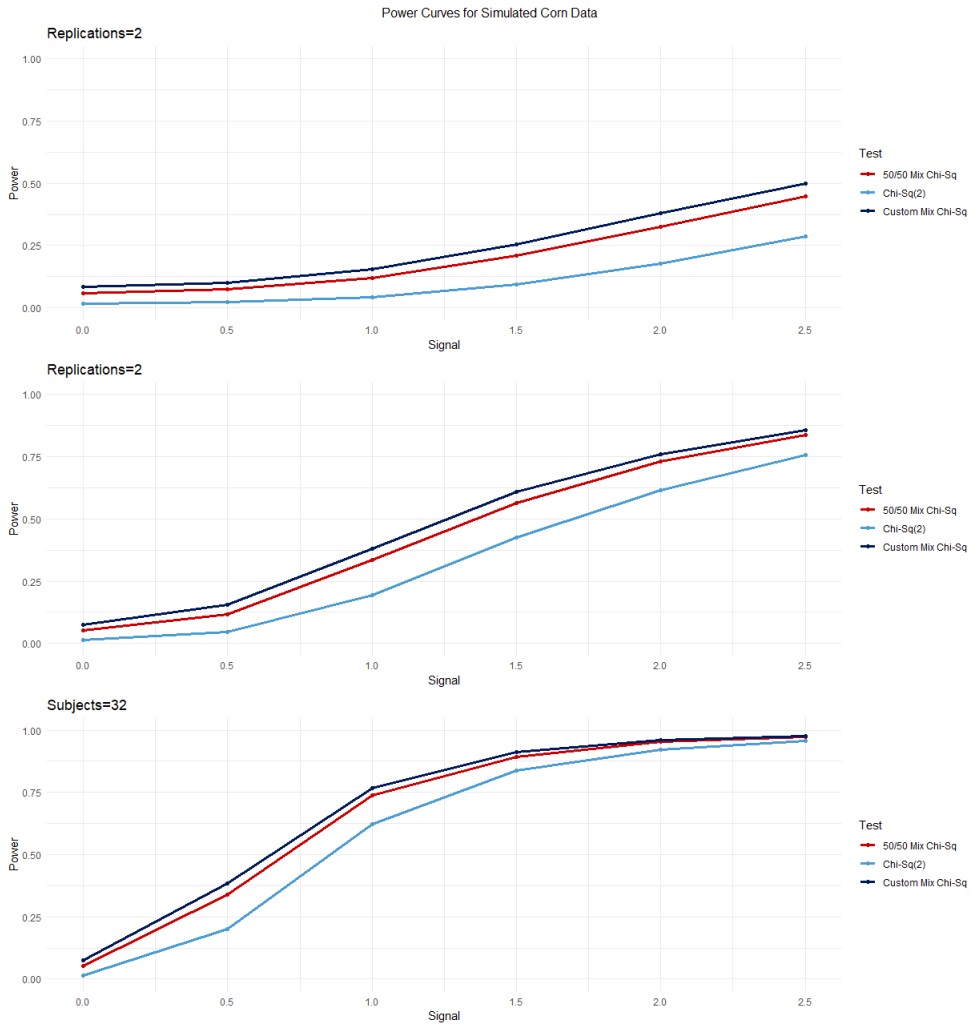
Here we definitely see improvement, the custom chi-square distribution is much more concentrated around 0. 64.2% of the test statistics at the largest sample size are practically zero, so the custom mixture is $.642\chi^2_0 + .358\chi^2_1$.



The 50/50 mixture is an improvement from the non-mixture distribution, but does not seem to fit as well as the custom. This is to be expected since more than 50% of the LRTs from the largest sample size are practically zero.



Similar to the first model, the distribution which visually fits the data best is more prone to type I error. Here the 50/50 mixture's size hovers right above .05, which is desirable. The non-mixture chi-square hardly ever rejects.



As expected, the custom mixture has the highest power, then the 50/50 mixture, then the non-mixture distribution. The two mixtures are quite close – closer than the previous model – so you do not gain much more power from using the custom mixture vs the 50/50.

5 Discussion

5.1 Review of Results

Not surprisingly, our results do not provide a clear, sure-fire way to test for random effects in mixed effects models. While the ultimate goal may be to find the best test and distribution to compare, a more realistic goal is to find a reasonable test and distribution and to understand its advantages and disadvantages. First, we consider the chi-square distribution with degrees of freedom equal to the number of variance components we are testing. We can see in both models that the sampling distribution of the LRT has the shape of a chi-square, but has many more instances at or close to 0. It is visually clear the chi-square distribution is not the correct fit.

The eye test is confirmed by the size and power calculations. Using the chi-square distribution to test for variance components is a conservative test. For both models and all sample sizes, the size of the test never reached .025, half of what is should be – .05. Both models show the power is uniformly lower than the other two distributions.

The advantages to using the chi-square distribution are that it is easy to use, there's no issue finding the cutoff value, and you are not at a high risk of committing type I error. If rejecting the null hypothesis that random effects exist has major consequences, but needs to be considered, perhaps this is the better way to test.

Next we consider the mixture distributions. Both the custom and 50/50 mixtures fit the data better visually. In the random intercept, random effects model – where the proportion of LRTs in the sampling distribution were practically zero was less than one half – the 50/50 mixture appeared to fit the data best. In the split plot model – with a much higher proportion of practically zero LRTs – the custom mixture, which accounted for this higher proportion, visually fit the data better.

This led to a similar result in size and power calculations. The 50/50 mixture has superior power, but high type I error for the random intercept, random slope model. For this model, the custom mixture had size mostly between .025 and .05 for the different sample sizes. The custom mixture is fairly lower than the 50/50 in regards to power.

Likewise, the custom mixture has better power, but high type I error for the split plot model. The 50/50 mixture has type I error just north of .5 for most sample sizes, which seems reasonable. The power drop off is minimal too, the 50/50 mixture is just shy of the custom for the split plot model.

If the 50/50 mixture has the same results for the random intercept, random slope model, we are more optimistic that it may be a better universal distribution to compare the LRT to. Unfortunately it does not. The 50/50 mixture has the advantage of being easy to calculate and showing strong power. It may be the case that when the sampling distribution of LRTs under null has a low percentage of practically zero values, the 50/50 mixture is too liberal. When the sampling distribution has a high percentage of practically zero values, the 50/50

mixture may be a solid option. This proportion of LRTs under null that are practically zero may be a funciton of the number of parameters we are testing as well.

5.2 Conclusion

Again, these results fail to provide a clear answer if one distribution is better than than another for testing variance components. Luckily, the results are clear about the advantages and disadvantages of the chi-square distribution and the 50/50 mixture chi-square distribution.

The chi-square distribution is going to be a conservative test, with small probability for type I error, but lower power. The 50/50 mixture is a liberal test. The type I error may well exceed your significance level, but in return you will receive higher power.

If p-values are necessary, one needs to consider the consequences of either type of error in their study. If type I is more grave, perhaps stick with the chi-square. If type II is, the 50/50 mixture is better. To bridge the gap, it is useful to look at both. If p-values based on both distributions yield significance or both fail to yield significance, one can be confident in their final decision. If the two distributions give contradictory decisions, we should be hesitant to conclude significance or not. More investigation is then necessary.

While the custom mixture is less realistic in practice, other options like bootstrap confidence intervals may give more insight. These methods likely will be more computationally expensive than the previous results.

5.3 Future Work

Two items, using similar research methods, can provide additional insight to this problem.

First, we want to consider more models. Simulation studies for many of the popular mixed effects models can be implemented fairly easily using the approach in this report. Using more models also allows us to look further at the percentage of the LRT sampling distribution which is practically different from 0. This can give us an understanding of which models the 50/50 mixture may be too liberal for and for which it has appropriate type I error.

Second, we can study other asymptotic tests. The Wald and Score tests may have better properties than the LRT and may have better convergence to certain distributions. Zhang, Daowen and Lin, Xihong (2008) [2] study this and find Score tests to have accurate size ($\approx .05$) and similar power to LRTs in one example.

6 Application

Using the real data from the sleepstudy dataset and corn dataset we can test for random effects. For the sleepstudy, we find the LRT to be 42.84. The corresponding p-values from each test are:

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

Here, the likelihood ratio statistic is massive and the p-values are all practically zero. Even with the issues with each test, these results strongly suggest random effects are present. It is safe to reject $H_0 : \sigma_1 = \sigma_2 = 0$. In the context of this dataset, we can say that the effects from either the random intercept for each subject, or the random slope from each day for each subject provide considerable variation to the response. Due to this, it is wise to account for those in the model.

Next, the likelihood ratio statistic for the split-plot model is 7.30. The corresponding p-values are:

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

While not as extreme as the sleepstudy data, we still have very low p-values for the corn dataset. At nearly all significance levels we reject $H_0 : \sigma_1 = 0$. It is safe to conclude that the field, nested in certain pesticide, is providing variation to the corn yields. It is important to account for this in the model.

Both of these examples unfortunately do not provide particularly interesting applications of the LRT. In both cases there is clearly a random effect that needs to be accounted for, hence why both are popular examples for teaching. For better demonstration, datasets which yield p-values between, say, .05 and .01, would provide that angle.

References

[1] Doug Bates. [r] lmer, p-values and all that.

[2] Daowen Zhang and Xihong Lin. *Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and other Related Topics*, pages 19–36. 01 2008.

7 Code Appendix

All R code can be found at github.com/peterpnorwood/LRTMixedModels.