Degrees of Freedom Methods and Simulation Study

In chapter 1, we outlined the basics of analyzing longitudinal data and introduced linear mixed models. Next, we will examine inference of linear mixed models, and how methods such as Kenward-Roger (KR) and Satterthwaite can be used in situations where standard procedures for inference may produce questionable results.

Inference

In statistical inference, the goal is to make conclusions about the underlying characteristics of a set of data and establish a relationship between certain variables. Hypothesis testing is one of the primary examples of inference, and is carried out in order to assess the true value of a population parameter. In linear models, the significance of a slope parameter, β_k, is often assessed, where the null hypothesis, H_0 is $\beta_k = 0$, and the alternative hypothesis H_a is $\beta_k \neq 0$. A test of the null hypothesis involves using a Wald statistic in the form

$$Z = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var(\hat{\beta}_k)}}},$$

which is then compared to the normal distribution, and a subsequent p-value is obtained.

Building on foundations of a general linear hypothesis test, given a matrix L of size $q \times p$, where q represents the number of estimable functions of β,

$$(L\hat{\beta} - L\beta)'[L(X'\widehat{Cov}(\hat{\beta})X)^{-}L']^{-1}(L\hat{\beta} - L\beta)$$

is approximately $\chi^2(q)$ [@lme_instatistics]. For a null hypothesis $H_0 : L\beta = 0$, the test statistic G is

$$(L\hat{\beta})'[L(X'\widehat{Cov}(\hat{\beta})X)^{-}L']^{-1}(L\hat{\beta}).$$

Likelihood ratio tests are another method to make inferences about β. While there are benefits to using the likelihood ratio test, the rest of this study will focus on method of using the Wald statistic.

Inference in Small Sample Sizes

One crucial assumption when conducting inference using the ML estimate for β is that the sample size is sufficient enough where it does not affect the estimate for Σ_i. However, what happens when the sample size is too small? This causes $\hat{\Sigma}_i$ to underestimate the true variance, which in turn causes $\widehat{\text{Cov}}(\hat{\beta})$ to be too small since it relies on covariance estimator. If $\widehat{\text{Cov}}(\hat{\beta})$ is too small, the denominator of the test statistic is inflated, leading to increased Type I error. One can see that the bias of the covariance estimator weakens the entire foundation of estimation and inference.

In very limited cases, where data are complete, balanced, and produce nonnegative values in REML estimation, is it possible to perform exact small-sample inferences. If $[L(X'\widehat{\text{Cov}}(\hat{\beta})X)^{-}L']^{-1}$ with g degrees of freedom can be rewritten such that

$$\frac{(L\hat{\beta})'Q(L\hat{\beta})}{g}\frac{w}{d} = \frac{(L\hat{\beta})'[L(X'\widehat{\text{Cov}}(\hat{\beta})X)^{-}L']^{-1}(L\hat{\beta})}{g},$$

where w is a chi-square random variable with d degrees of freedom, then this statistic is F-distributed.

However, in most scenarios, an approximate small-sample method must be used, in which the statistic

$$F = \frac{(L\hat{\beta})'[L(X'\widehat{\text{Cov}}(\hat{\beta})X)^{-}L']^{-1}(L\hat{\beta})}{g}$$

follows a distribution with numerator degrees of freedom g, and unknown denominator degrees of freedom (DDF). There are several ways to approximate the DDF.

Both Satterthwaite and KR are proposed methods of reductions to the DDF when conducting tests in order to account for the uncertainty of the covariance estimator. The KR method goes one step forward to also adjust the test statistic itself.

Satterthwaite

Sattherthwaite approximation was developed by @hrong-tai_fai_approximate_1996, with the F statistic following the form:

$$F = \frac{1}{l}\hat{\beta}' L' (L\widehat{\text{Cov}}(\hat{\beta})L')^{-1} L\hat{\beta}.$$

For the denominator degrees of freedom we perform spectral decomposition on $L'\widehat{\text{Cov}}(\hat{\beta})L = P'DP$, where D is a diagonal matrix of eigenvalues and P is an orthogonal matrix of eigenvectors. When r represents the r^{th} row of $P'L$, we have $v_r = \frac{2(d_r)^2}{g_r' W g_r}$, where g_r is a gradient vector, d_r is the r^{th} diagonal element of D, and W is the covariance matrix of $\hat{\sigma}^2$. The denominator degrees of freedom is calculated by:

$$\frac{2E}{E-l},$$

where $E = \sum_{r=1}^{l} \frac{v_r}{v_r - 2} I(v_r > 2)$ if $E > l$, otherwise $DF = 1$.

When $l = 1$ the KR and Satterthwaite approximation will produce the same denominator degrees of freedom. However, since the statistic used for the two methods are not the same, the results for inference will not be the same. It is important to note that both methods are only valid when using REML.

Kenward-Roger

In @kenward_small_1997, a Wald statistic is proposed in the form of:

$$F = 1/l(\hat{\beta} - \beta)^T L (L^T \hat{\Phi}_A L)^{-1} L^T (\hat{\beta} - \beta),$$

where l represents the number of linear combinations of the elements in β, L is a fixed matrix, and $\hat{\Phi}_A$ is the adjusted estimator for the covariance matrix of $\hat{\beta}$. As mentioned previously, $\widehat{\text{Cov}}(\hat{\beta})$ is a biased estimator of $\text{Cov}(\hat{\beta})$ when samples are small, and underestimates. This adjusted estimator is broken down into $\hat{\Phi}_A = \widehat{\text{Cov}}(\hat{\beta}) + 2\hat{\Lambda}$, where $\hat{\Lambda}$ accounts for the amount of variation that was underestimated by the original estimator of covariance of $\hat{\beta}$. The value Λ is approximated using a Taylor series expansion around σ, to be

$$\Lambda \text{Cov}(\hat{\beta})[\sum_{i=1}^{r}\sum_{j=1}^{r} W_{ij}(Q_{ij} - P_i \text{Cov}(\hat{\beta}) P_j)]\text{Cov}(\hat{\beta}),$$

where $P_i = X^T \frac{\partial \Sigma^{-1}}{\partial \sigma_i} X$ $Q_{ij} = X^T \frac{\partial \Sigma^{-1}}{\partial \sigma_i} \Sigma \frac{\partial \Sigma^{-1}}{\partial \sigma_j} X$, and W_{ij} is the (i,j)th element of W = $V[\hat{\sigma}]$.

This Wald statistic that uses the adjusted estimator is scaled in the form:

$$F^* = \frac{m}{m + l - 1}\lambda F,$$

where m is the denominator degrees of freedom, and λ is a scale factor. Using the expectation and variance of the Wald statistic, F Both m and λ need to be calculated from the data, such that:

$$m = 4 + \frac{l+2}{l\rho - 1},$$

where $\rho = \frac{V[F]}{2E[F]^2}$ and $\lambda = \frac{m}{E[F](m-2)}$. This statistic will ultimately follow an exact $F_{l,m}$ distribution.

Other Methods

Residual DDF: The DDF is calculated as $N - rank[X]$, where N is the total number of individuals in the dataset. This method is only suitable for data that are independent and identically distributed, so it is not typically used in linear mixed models.

Containment Method In the containment method, random effects that contain the fixed effect of interest are isolated. The smallest rank contribution to the $[XZ]$ matrix among these random effects becomes the DDF. If there are no effects found, then the DDF is equal to the residual DDF.

Between-Within Method @Schluchte propose a DDF method where residual DDF are calculated for both between-subject and within-subject subgroups. If there are changes in the fixed effect within subjects, then the within-subject DDF is used, otherwise the between-subject DDF is used.

Existing Literature

Both KR and Satterthwaite methods are frequently used and compared, and its performance is highly dependent on the structure of the data. A majority of studies focusing on DF method comparison in mixed models use split-plot design, as small sample sizes are more common in agricultural and biological fields. @schaalje_adequacy_2002 found that in comparison to other degrees of freedom-adjusting methods like Satterthaite, KR was the most suitable for small sample data. Using factors such as imbalance, covariance structure, and sample size, they demonstrated that the KR method produced simulated Type I error rates closest to target values. However, their focus was primarily on complexity of covariance structure, and they found that more complicated structures, such as ante-dependence, produced inflated error rates when coupled with small sample size. @arnau_analyzing_2009 found that KR produces more robust results compared to Satterthwaite and Between-Within approaches, especially in cases where larger sample size was paired with covariance matricies with larger values.

These studies are conducted with data drawn from normal distributions. However, real-world data used in fields such as psychometrics have distributions that are nonnormal. In their paper, @arnau_using_2012 extend their evaluation of KR for split-plot data that follow a log-normal or exponential distribution, and for when the kurtosis and skewness values are manipulated. They found that, compared to normal distribution, KR is less robust for log-normal distributions, but that there is no significant difference in performance between exponential and normal distributions. In addition, they suggest that skewness has a bigger effect on robustness of KR compared to kurtosis.

Existing research evaluating the performance of methods that reduce Type I error rate in small samples are thorough, however, the differences in simulation setup and structure of data used make generalizations difficult. Although the KR method has been shown as a viable option for analysis of small samples in many occasions, it should continue to evaluated against other methods. To date, there is no literature on the performance of Satterthwaite for nonnormal longitudinal data design. Given the prevalence of nonnormal and small data samples, it is important to continue exploring methods that ensure robust results.

Goals of This Study:

In this study, we aim to expand on previous work evaluating how methods for estimating fixed effects perform under different nonnormal distributions, sample sizes, number of measurements, and complexity. The aforementioned studies often use a split-plot design and impose a covariance structure, but the goal of this study will be to compare performance of KR and Satterthwaite methods for repeated measures longitudinal data fitted with a linear mixed effects model, and an unstructured covariance structure imposed by the random effects.

Simulation Set Up:

The overarching goal is to compare performance of DF methods across various conditions, with a specific goal of looking at scenarios with smaller sample sizes and number of measurements in order to see variability in Type I error rates. There is flexibility in how the parameters are chosen; in order to narrow the scope, we aim to design conditions that are relatively similar to an application data set that will be further explored in Chapter 3. In the following sections, the range of parameters used will be similar to the characteristics of a longitudinal study about Children's Health.

Generating Data: Sample Dize

In this study, we consider a linear mixed effects model with time as a continuous variable and treatment as a factor. The range of possible values that time takes on depends on how many number of measurements per individual, which can be 4 or 8. The treatment covariate takes on values of 0 or 1, and each assigned to half of the sample. The number of individuals take on possible values of 10, 18, and 26. These were chosen to reflect possible samples that would not hold under the common assumption that the sample size must be at least 30 for it to be considered sufficient enough for the Central Limit Theorem to hold.

Generating Data: Fixed Effects

We have three fixed effects: the intercept value, time and treatment. The intercept, an arbitrary value, is set at $B_0 = 3.1$. The coefficients for time and treatment have a value of $B_1 = B_2 = 0$ for simplicity and so we can evaluate Type 1 error rates.

Generating Data: Random effects

We generate our intercept and slope random effects values from nonnormal distributions, which are either exponential or lognormal. Previous research shows that many data used in social and health sciences follow nonnormal distributions [@limpert_log-normal_2001]. More specifically many follow lognormal distributions, such as age of onset of Alzheimer's disease [@horner_age_1987], or exponential distribution to model survival data. In order to cover a wide range of exponential and lognormal distributions, parameters were chosen to model distinct distributions. For exponential distributions, $X \sim Exp(\lambda)$, where $\lambda = .2$ or .9 Lognormal distributions were $X \sim Log(0, .25)$ and $X \sim Log(1, .5)$.

Using the *SimMultiCorrData* package, values are generated through Fleishman's method for simulating nonnormal data by matching moments using mean, variance, skew, and kurtosis and then transforming normally distributed values [@Simtest].

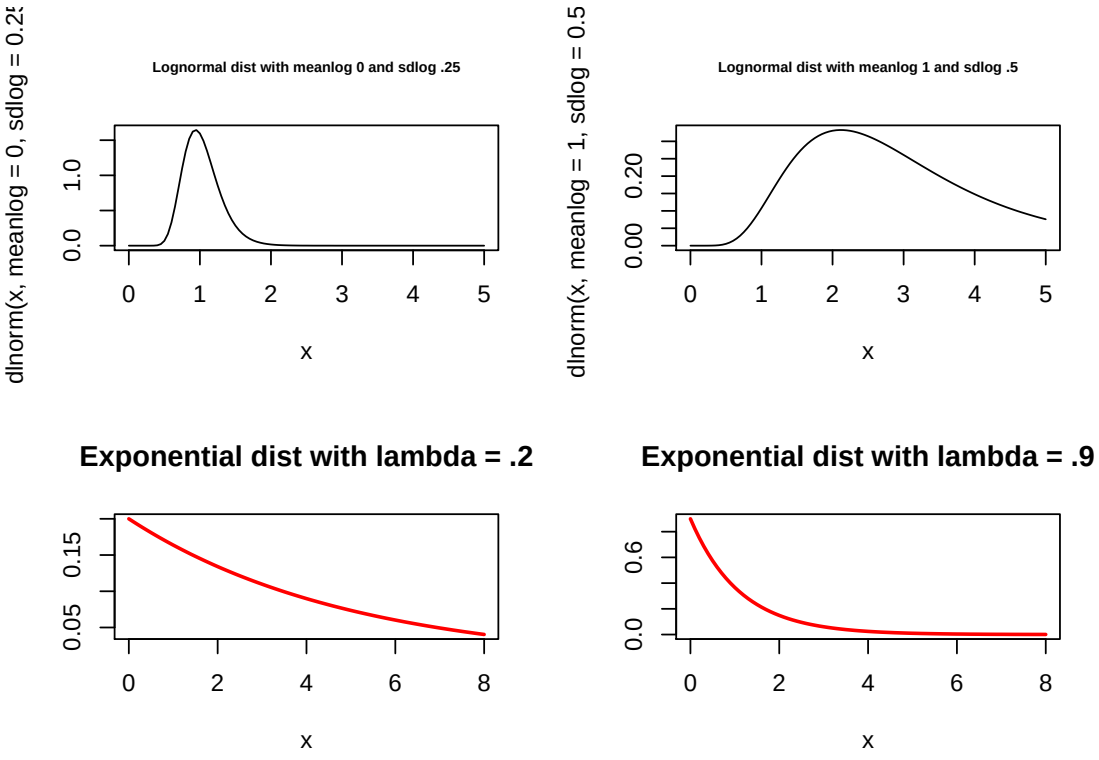
When simulation data from nonnormal distributions, we use the standard deviation produced by Fleishman's method as the parameter. In order to align the simulation data from the Children's Health study, we use the same ratio of variances between the residual variance and variance of the random effects of time and intercept derived from modeling linear mixed models on that data in our simulation.

In the intercept only model, only one non-normal continuous variable is generated for the random effect, so the function `nonnormvar1()` is used. In order to generate measurement error, we simulate values from a $X \sim N(0, \sqrt{4}SD$, where SD is the standard deviation from each distribution we want explore.

In the case of the linear model that has both random effects for intercept and slope, we want to generate random effects values that are correlated. Using `SimMultiCorrData::rcorrvar()`, we use a similar process for generating one nonnormal continuous variable, but extend it to generating variables from multivariate normal distribution that take in to account a specified correlation matrix, and are then transformed to be nonnormal. Also, in order to maintain the same ratio of variances to the application data, the variance of the distribution that the slope random effects are generated is equal to $\frac{1}{8}$Var, where Var refers to the variance of the distribution that the random effects for the intercept are generated from. On a similar

note, we use a correlation value of -.38 to generate the random effects for the intercept and slope, which is based off the correlation observed between the random effects of time and intercept when fitting a random slope model in Chapter 3. For random slope models, the measurement error is simulated for a distribution $X \sim N(0, \sqrt{.168}SD$.

Distributions used are depicted below.



Linear Mixed Effects Model

In a linear mixed effects model, the amount of random effects that will be modeled depends on the research question at hand. Here, we will examine both a random intercepts model, where only the intercept of the model is assumed to have a random effects structure, as well as a random intercept and slope model, where in addition to intercept, the covariate time will also have a random effects structure. The random intercept model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 Trt_i + \beta_2 Time_{ij} + b_{0j} + e_{ij},$$

and the random slope model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 Trt_i + \beta_2 Time_{ij} + b_{0j} + b_{1i} Time_{ij} + e_{ij}.$$

We use the *lmerTest* package to fit the linear mixed effects model and evaluate random and fixed effects. To evaluate significance of the treatment variable, we compare the performance and resulting p-values from 4 different Wald-type tests: KR, Satterthwaite, standard DF method, and t-as-z. t-as-z and standard DF formula are not adjustments to account for smaller sample sizes, and are used as comparison to Satterthwaite and KR, since they are expected to be anti-conservative. Because the value of the treatment in our model is fixed at 0 in order to identify Type I error, we expect to see that the p-value to not be significant ($p >$ .05) in an ideal scenario.

Evaluating and Results

After performing 400 replications of each condition at a significance level of .05, we evaluate robustness using Bradley's criterion, which considers a test to robust if the empirical error rate is between .025 and .075 [@bradley]. In the following section, we will compare Type I error rates produced from KR and Satterthwaite methods as well as *t*-as-*z* and using the standard DF formula, further stratified by distribution and other manipulated parameters.

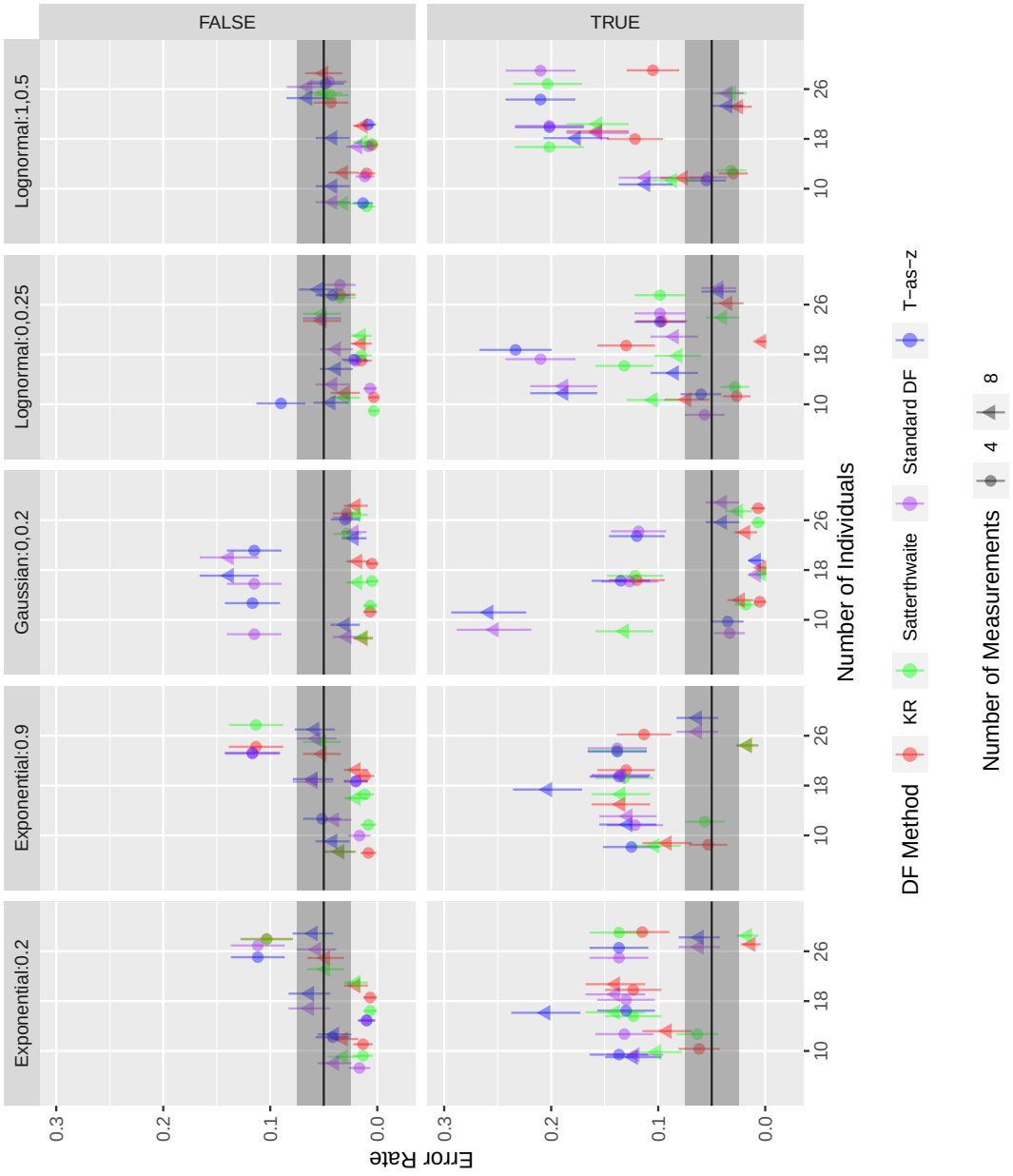


Figure 1: Type I Error Rates by DF method

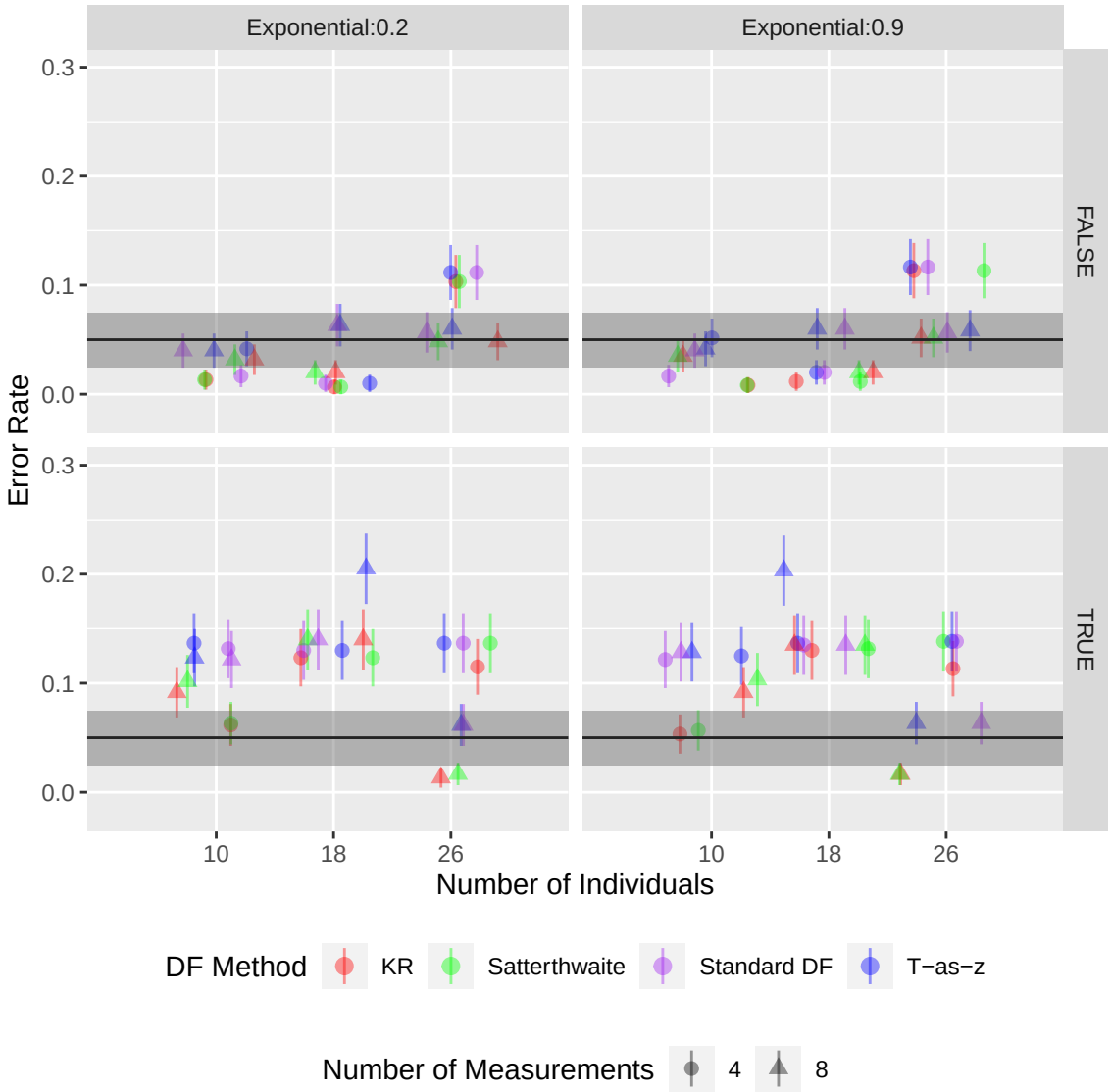
Figure @ref(fig:fig1) displays error rates from all 4 degrees of freedom methods by distribution, parameters, complexity of random model, number of measurements, and number of samples. The shaded region indicates error rates that are considered robust by @bradley's criterion. It is evident that there are varying patterns of performance by distribution. The common conception that larger sample sizes or large number of mea-

surements can improve robustness is not necessarily evident across all distributions, for example in the case of the exponential distribution. One trend that appears to be evident across all three distributions is that when the degrees of freedom methods are applied to a random intercept model, a more structurally simple model, they yield more robust error rates in comparison to an application to the random slopes model.

In addition, when looking at performance of the 4 methods overall, we can see that the t-as-z and standard DF approach produce significantly more anti-conservative results, regardless of the values of other parameters. These trends align closely with a previous study by @luke_evaluating_2017 examining only normal distributions.

In order to make more specific observations and identify trends, we will examine performance within each of the three distributions by sample size and number of measurements.

Exponential Distribution



Our simulation results explore two exponential distributions, one with $\lambda = .9$ and $\lambda = .2$. Looking at figure @ref(fig:fig2), at $\lambda = .2$, we can see that in random slope models, comparing sample size of 10 to 26 marginally improves the DF methods, but only on those that are applied to conditions with more repeated

measurements. The relationship between sample size and robustness is not linear, as increasing the sample size from 10 to 18 does not help DF methods achieve less anti-conservative error rates. On the other hand, in random intercept models, increasing the sample size did not improve the performance of DF methods, and in some cases were associated with more anti-conservative Type I error rates; in the case of sample size 26 and 4 measurements, Type I error rates performed significantly worse in comparison to smaller sample sizes, holding other conditions constant.

At $\lambda = .9$, we see virtually the same trends in terms of the effect of sample size, complexity of model, and number of measurements on the performance of the DF methods. Overall, across both distributions, increasing the number of repeated measures impacted the DF methods' performance and increased robustness, while the effect of sample size was hard to pinpoint. Additionally, Kenward-Roger and Satterthwaite methods tended to produce more conservative error rates.

Despite having different parameter values, the application of DF methods to these two exponential distributions produce similar different trends in error rates. Considering that the two distributions have the same skewness and kurtosis values, we hypothesize whether this similarity in values contributes to the performance of the DF methods. Next, we will examine the performance of DF methods in lognormal distributions.

Lognormal Distribution

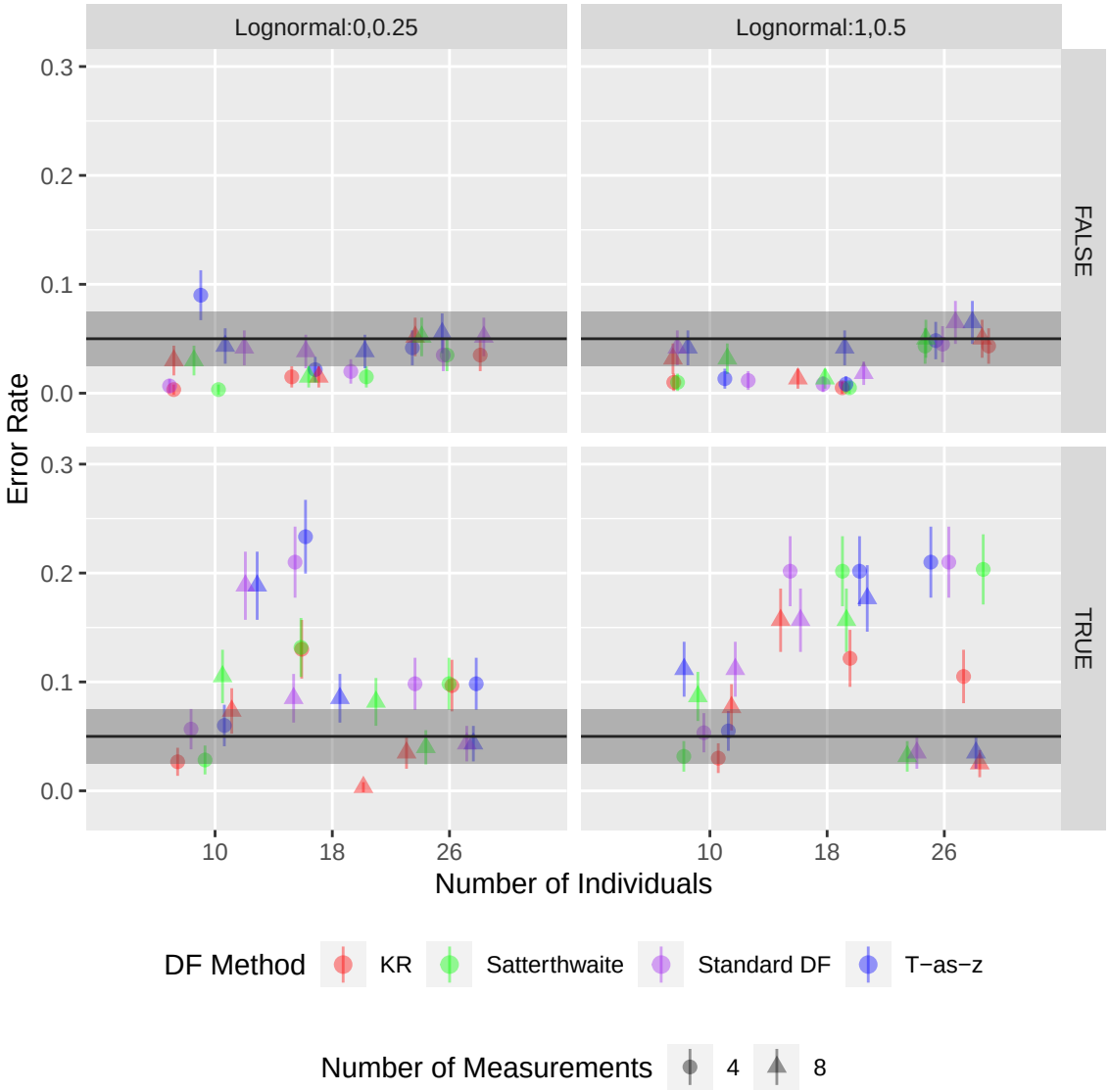


Figure @ref(fig:fig3) highlights Type I error rates produced by DF method for lognormal data. As seen in the exponential distribution, across the lognormal distributions, DF methods applied to random intercept models had consistently more robust error rates in comparison to random slope.

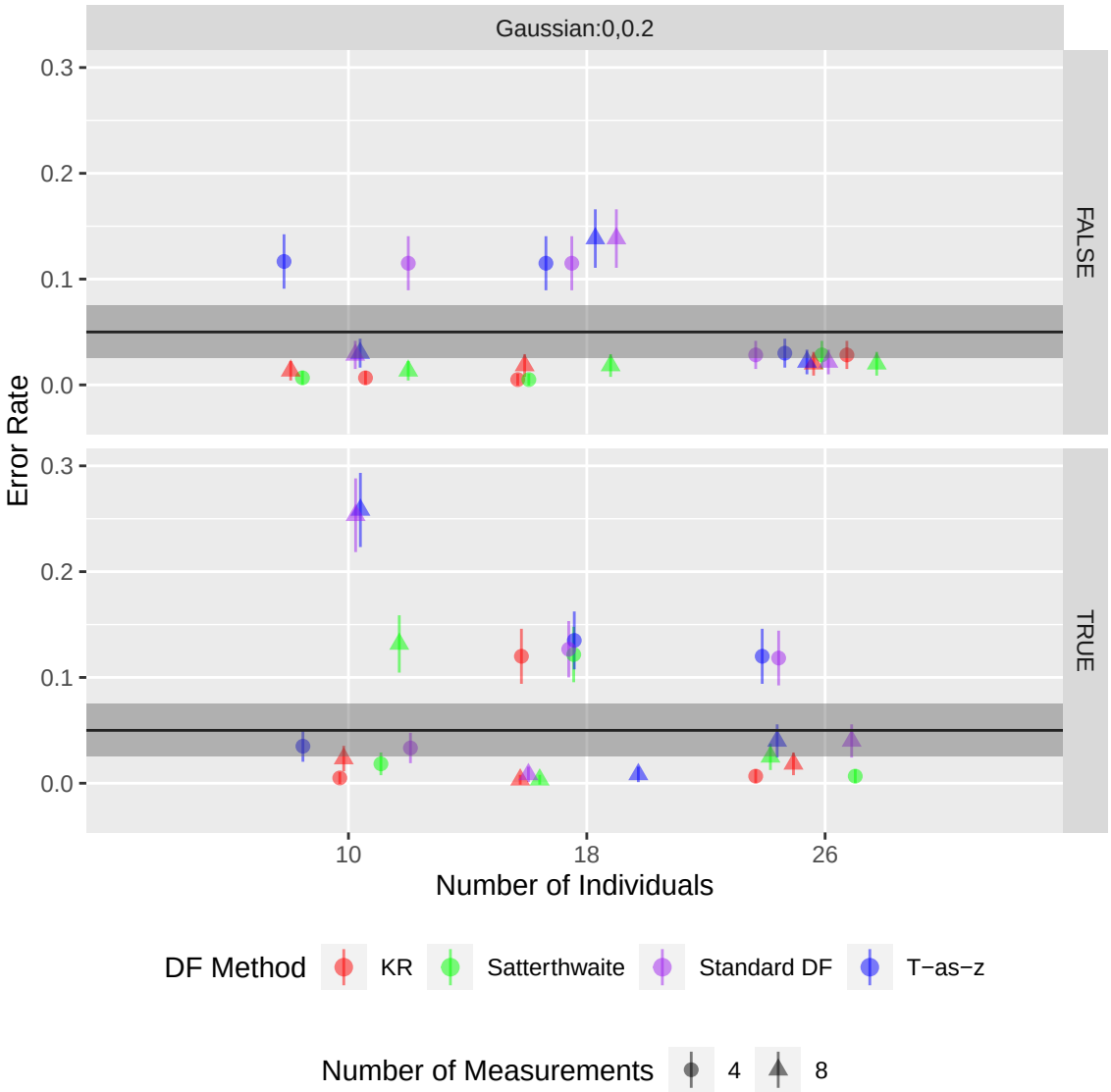
Our first lognormal distribution with parameters $(0, .25)$ has lower values of kurtosis and skewness. In the random slope model, increasing sample size from 10 to 18 only improves performance of the t-as-z and standard DF method when individuals have 8 measurements each, and in all other cases leads to less robustness. At sample size 26, DF methods are significantly more robust when applied to 8 measurements per individual rather than 4. In the random intercept model, we see a similar lack of consistent pattern of Type I error rates when sample size increases from 10 to 18, but at sample size 26 all methods regardless of number of measurements are robust.

On the other hand, with higher levels of skewness and kurtosis with a lognormal distribution with parameters $(1, .5)$, the effect of number of measurements and sample size are slightly different. Increasing the sample size from 10 to 18 has either no effect or an adverse effect on DF methods; in random intercept models the DF methods generally produce Type I error rates that are too conservative, and the opposite is seen in random slope models. Similar to the other lognormal distribution, DF methods applied to random intercept models

of size 26 are all robust. The same difference between performance of DF methods of 4 and 8 measurements at size 26 in random slope models are observed, but that difference is widened. All DF methods at 4 measurements aside from the KR method have Type I error rates that almost 4 times the error rates at 8 measurements.

Based on comparisons between the two distributions, our results suggest that there are only slight differences that arise, suggesting that to a certain extent, distributions with greater skew and kurtosis may not significantly impact the performance of DF methods. In both distributions, DF methods applied to models with size 26 and 8 measurements were the most robust. Increasing sample size from 10 to 18 yields different Type I error rates across the two distributions, but it is difficult to discern a definite pattern.

Normal Distribution



While nonnormal distributions are the focus of this study, comparing performance of DF methods to the normal distribution is important as a point of reference. As we can see in figure @ref(fig:fig4), it is interesting to note how robustness has decreased for DF methods applied to the normal distribution compared to lognormal and exponential distributions, but it is also important to acknowledge that the methods, especially KR and Satterthwaite, produce conservative error rates, rather than anti-conservative ones. In most scenarios, KR

and Satterthwaite remain unaffected when changing sample size. On the other hand, t-as-z and standard DF oftentimes produce the most anti-conservative error rates, but improvements arise when sample size is at 26.

Overall, in the normal distribution, there is a clearer distinction in performance of KR and Satterthwaite vs t-as-z and standard DF methods. Increasing the number of measurements is another factor that can affect Type I error rates produced by DF method within a particular method, rather than comparing across them.

KR vs Satterthwaite

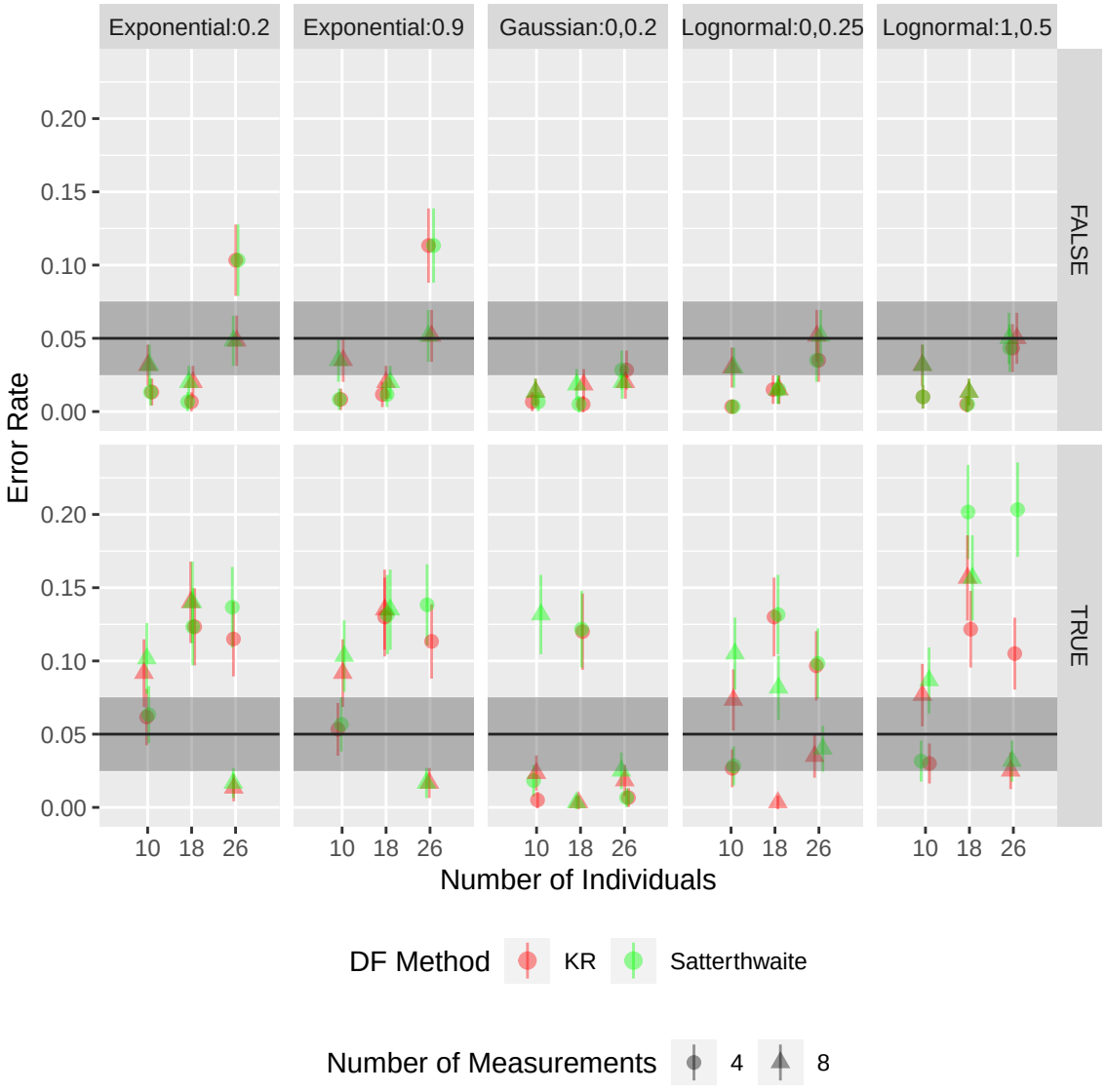


Figure 2: Type I Error Rates Produced by KR vs Satterthwaite

Comparing performance across all 4 methods has yielded significant evidence that KR and Satterthwaite are superior methods when using linear mixed models on small samples. @luke_evaluating_2017 suggests that both KR and Satterthwaite are comparable solutions to obtain adequate Type I error. Figure @ref(fig:fig5) aims to narrow in on differences in performance between the two methods. One can see that across random

intercept models, KR and Satterthwaite methods have identical performance. Looking more closely at random slope models, it appears that KR consistently produces more conservative error rates when holding distribution, sample size, and number of measurements constant. In two scenarios, of which models were of size 10 and 18, the KR method produced more conservative error rates, but were not robust, while the Satterthwaite method produced robust error rates.

While in a majority of cases KR is a viable solution when aiming to keep Type I error rates low, it is not necessarily the optimal DF choice in every scenario if the goal is to produce error rates closest to .05.

Tables for all error rates produced by 4 DF methods are located in the appendix @ref(app:pvalue-tables)

Discussion

After comparing the error rates produced by DF methods across three distributions, three sample sizes, and two numbers of measurements and complexity of models totaling to 96 unique conditions, we find that overall, KR and Satterthwaite DF methods yield the most robust error rates when constructing LMM with small sample sizes. In addition, random slope models are significantly more complicated in terms of using DF methods to produce robust error rates, and also challenge the superiority of the KR method over the Satterthwaite method when looking at robustness of Type I errors of random slope models of smaller sample sizes.

This stark contrast in performance between random slope and random intercept models across all distributions requires further investigation. @barr_random_2013 suggest that LMM with maximal effects are preferable and can reduce Type I error rates, as random slope models can account for between individual variation in slopes and reduce residual variance. However, in their study the distribution of the variable of interest is normal, and random effects were generated from a bivariate normal distribution. In this simulation, random effects for the intercept and time were drawn from two nonnormal distributions which can further complicate the distribution of the outcome variable and complicate the fit of the model.

When comparing and constrasting performance of Satterthwaite and KR methods, which have both been demonstrated to be suitable options for producing robust Type I error, it seems that complexity of the model as well as sample size are two parameters that can differentiate their performance. In random intercept models, performance between the two methods are identical. In random slope models, KR is more conservative, and in a few conditions where the sample size is less than 26, it is too conservative in comparison to Satterthwaite. Given that KR is a further adjustment to Satterthwaite, its more conservative performance is not surprising; however, when evaluating fixed effects, conservative Type I error rates are preferable to anti-conservative rates. Thus, we prefer KR as the DF approximation in small and nonnormally distributed samples. In terms of sample size and number of measurements, it appears that increasing the number of measurements increases the proportion of DF methods that produce robust Type I error rates in most conditions. Difference in DF method performance across measurement numbers is most evident in random slope models with samples of size 26. On the other hand, increasing sample size is not strongly associated with increased robustness of Type I error rates produced by DF methods. Increasing from sample size from 10 to 18 produces inconclusive patterns: sometimes DF methods produce even more anti-conservative error rates, and in other cases there is virtually no difference. In most cases, sample size of 26 is the most ideal for robustness, but in some nonnormal distributions robustness can only be achieved by also having more repeated measurements. It is plausible that the differences between a sample size of 10 and 18 are insignificant because both are too small in terms of a "large-enough" sample of size 30, so the DF methods are not impacted by this increase in sample. Models of sample size 26 may be ideal in some circumstances mainly because of its proximity to being considered sufficiently large enough.

While previous studies comparing DF methods focus primarily on normal distributions, the results from this study demonstrate that applying DF methods to LMM with nonnormal distributions and small sample sizes can still produce robust Type I error rates. However, careful consideration of the complexity of the model, number of measurements and individuals, as well as the skewness and kurtosis of the data must be considered, as they each have an effect on the DF method not only individually, but when interacting with

other parameters as well. Overall, these results also indicate that Type I error rates are closest to .05 when employing Kenward-Roger and Satterthwaite methods.