when the kurtosis and skewness values are manipulated. They found that, compared to normal distribution, the test is less robust for log-normal distributions, but that there is no signficant 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.

Chapter 3?

2.6 Goals of this study:

In this study, we aim to expand on previous ~~simulations~~ work, evaluating how methods for evaluated ing fixed effects perform under different nonnormal distributions and sample sizes. 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 no imposed covariance structure. Since most mixed models use unstructured covariance structure, it would be beneficial to see how these methods perform without considering covariance structure as a factor.]

structure is technically imposed by specification of random effects.

2.7 Simulation Set up:

→ lead by explaining values reflect data from longitudinal study.

2.7.1 Generating data: Sample size

one factor (trt)
one continuous (time)

time was treated as continuous in your fitted model even though the values were discrete. Re-write to clarify.

In this study, we consider a linear mixed effects model with two discrete covariates: time and treatment. 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.

Might be useful to write out the two LMMs:

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

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

Then use notation throughout.

2.7.2 Generating data: Fixed Effects

We have three fixed effects: the intercept value and the covariates time and treatment. The intercept, an arbitrary value, is set at $\beta_0 =$ 3.1. The coefficients for Time and treatment have a value of 0 $(\beta_1 = \beta_2 = 0)$ ~~, and~~ for simplicity and so we can evaluate the Type I error rates of treatment. ~~will be evaluated~~.

2.7.3 Generating data: Random effects

Talk about 2 models considered (intercept only, intercept + slope)

~~In order to generate a continuous response variable that is nonnormal,~~ W~~w~~e generate our 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, Stahel, & Abbt, 2001). More specifically many follow lognormal distributions, such as age of onset of Alzheimer's disease (Horner, 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, lambda values of .2, and

are you including random error in this?

use notation

package names: bold, normal font

functions have parentheses at the end

would be useful

.9 were used, (DO I NEED TO INSERT GRAPH?). For lognormal distribution, mean and standard deviation parameter combinations were (0,.25), and (1,.5). use notation

cite

Using the SimMultiCorrData package, ~~we derive kurtosis and skewness values based on the distributions specified above. The table below shows the range of skewness and kurtosis values for the Lognormal distribution~~. In the intercept only model, only one non-normal continuous variable is generated for the random effect, so the function ~~SimMultiCorrData::~~nonnormvar1() is used. 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. citation

describe Fleishman process once up front for MVN

Kurtosis and skew values for the distributions used in this simulation are shown below. 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. We use a correlation value of -.38 to generate the random effects, which is based off the correlation observed when fitting a linear mixed effects model from the dataset used in the application portion of this study. What about ratio of variances?

graph is more useful

mean	sd	skew	kurtosis	fifth	sixth
1.03	0.262	0.778	1.1	2.3	6.48
3.08	1.642	1.750	5.9	31.4	240.00
5.00	5.000	2.000	6.0	24.0	120.00
1.11	1.111	2.000	6.0	24.0	120.00

***** FIX THIS ***** Lastly, to account for measurement/sampling error, we assume that the error is random and drawn from a $\mathcal{N} \sim (0, .2)$. The standard deviation

$e_{ij} \overset{iid}{\sim} N(0, \boxed{})$

varied depending on model (specify how)

value was chosen to minimize the variation of the errors in relation to the random effects of the intercept and the covariate.

2.7.4 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-only model, where 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.

bold, cite

We use the `lmerTest` package to fit the linear mixed effects model, and evaluate the significance of the covariates? in the model. To evaluate significance, we compare

4 different Wald-type tests... (3 t, one t-as-Z)

discuss these here.

both the KR and Satterthwaite method for adjusting denominator degrees of freedom and its resulting p-value. Because the value of the covariate in our model is fixed at 0 in order to identify Type I error, we expect to see that the p-value for the covariate time to not be significant ($p > .05$) in an ideal scenario.

2.8 Evaluating and Results

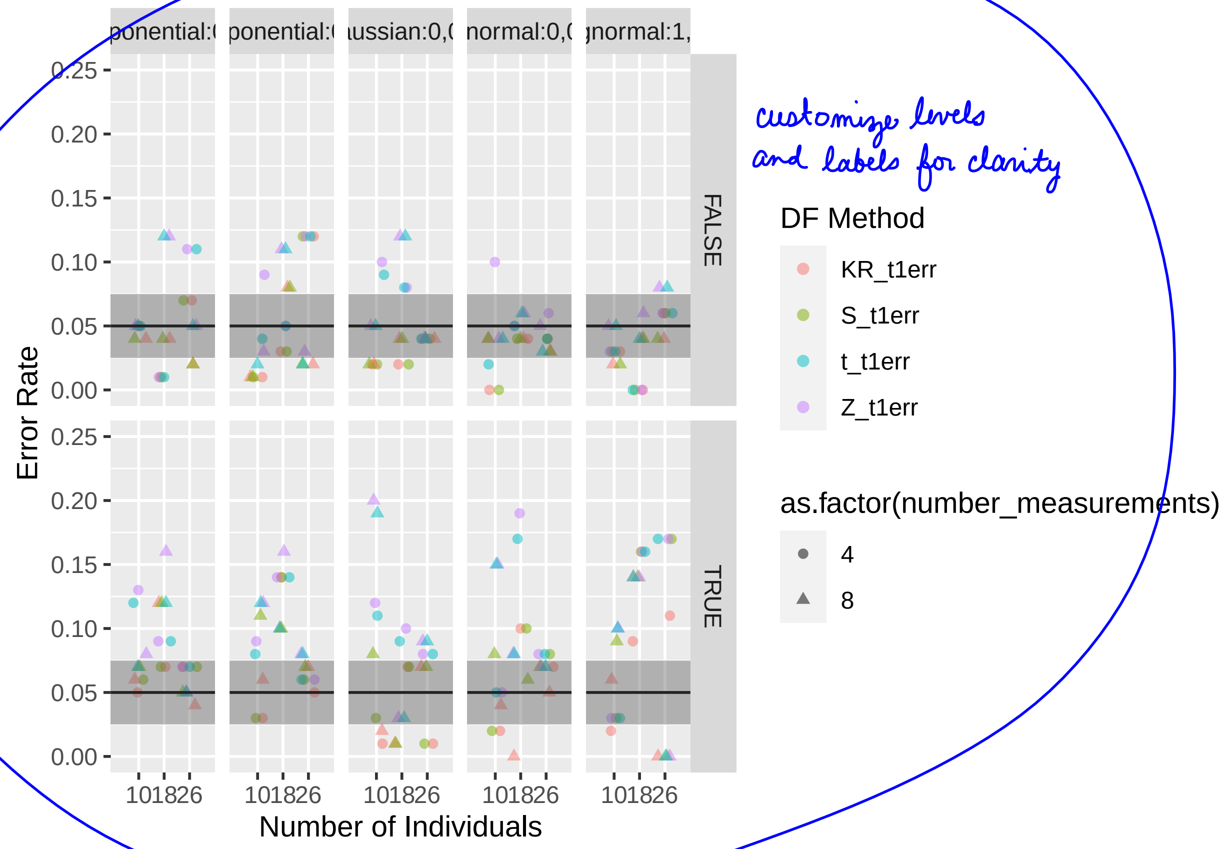
do you have 1000 reps? last I saw you only had 100? If you can get to at least 300 that would be ideal (Bradley's criterion could be used) (at 100, the MoE for your results is $1.96\sqrt{\frac{0.05(0.95)}{100}} = 0.0427$ so the simulation error is too high to talk about the results.

After performing 1,000 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. In the following section, we will

citation?

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. 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.

customizing plot sizes.



customize levels and labels for clarity

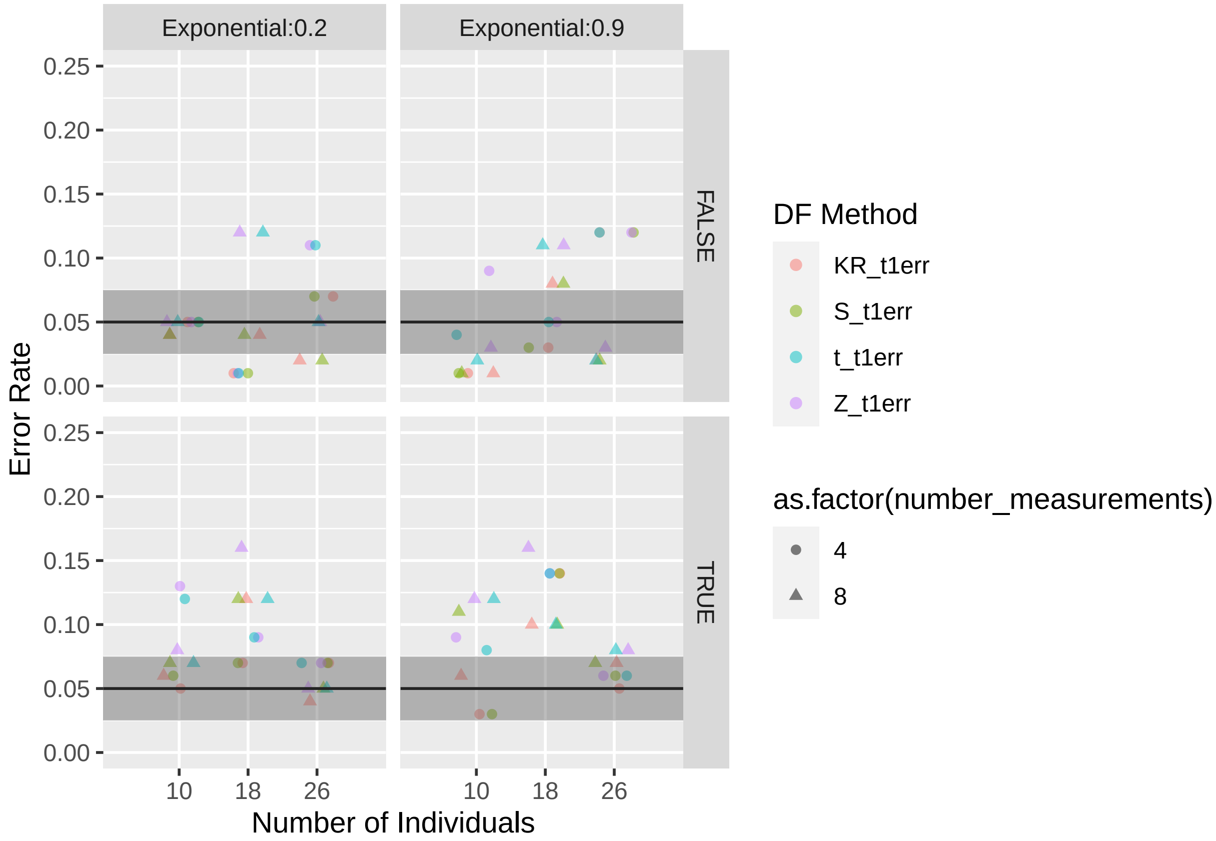
FIGURE 1 displays empirical Type I error rates from all 4 degrees of freedom methods by distribution, parameters, complexity of random model, number of measurements, and number of samples. The shaded region indicates error rates that are considered robust by Bradley's criterion. It is evident that there are varying patterns of performance by distribution. The common conception that larger sample sizes or large number of measurements can improve robustness is not necessarily evident across all distributions, ~~for example in the case of the exponential distribution.~~ While the random intercept model, a more structurally simple model, yields more robust error rates in the lognormal and G gaussian distribution, that is not the case in the exponential distribution.

make these reproducible

However, there is one trend that we can identify. 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 (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 before attempting to compare across them.

2.8.1 Exponential Distribution



```
# A tibble: 8 x 4
# Groups:   params, rslope [4]
  params rslope number_measurements robustness
  <chr>  <lgl>                <dbl>      <dbl>
```

```
1 0.2    FALSE                        4      0.5
2 0.2    FALSE                        8      0.667
3 0.2    TRUE                         4      0.667
4 0.2    TRUE                         8      0.583
5 0.9    FALSE                        4      0.417
6 0.9    FALSE                        8      0.167
7 0.9    TRUE                         4      0.5
8 0.9    TRUE                         8      0.25
```

Our simulation results contain two exponential distributions one with $\lambda = .9$ and $\lambda = .2$. ~~In FIGURE? we identified that in both exponential distributions, random slope models tended to yield more robust error rates, however, we continue to disaggregate by complexity of model to identify how it may work alongside other parameters.~~ At $\lambda = .2$, we can see that in random intercept models the error rates are less anti-conservative at sample sizes of 10 and 18 compared to the random slope model. However, at sample size 26 the random slope model performs significantly better. It is difficult to discern whether increasing the number of measurements reduces robustness overall.

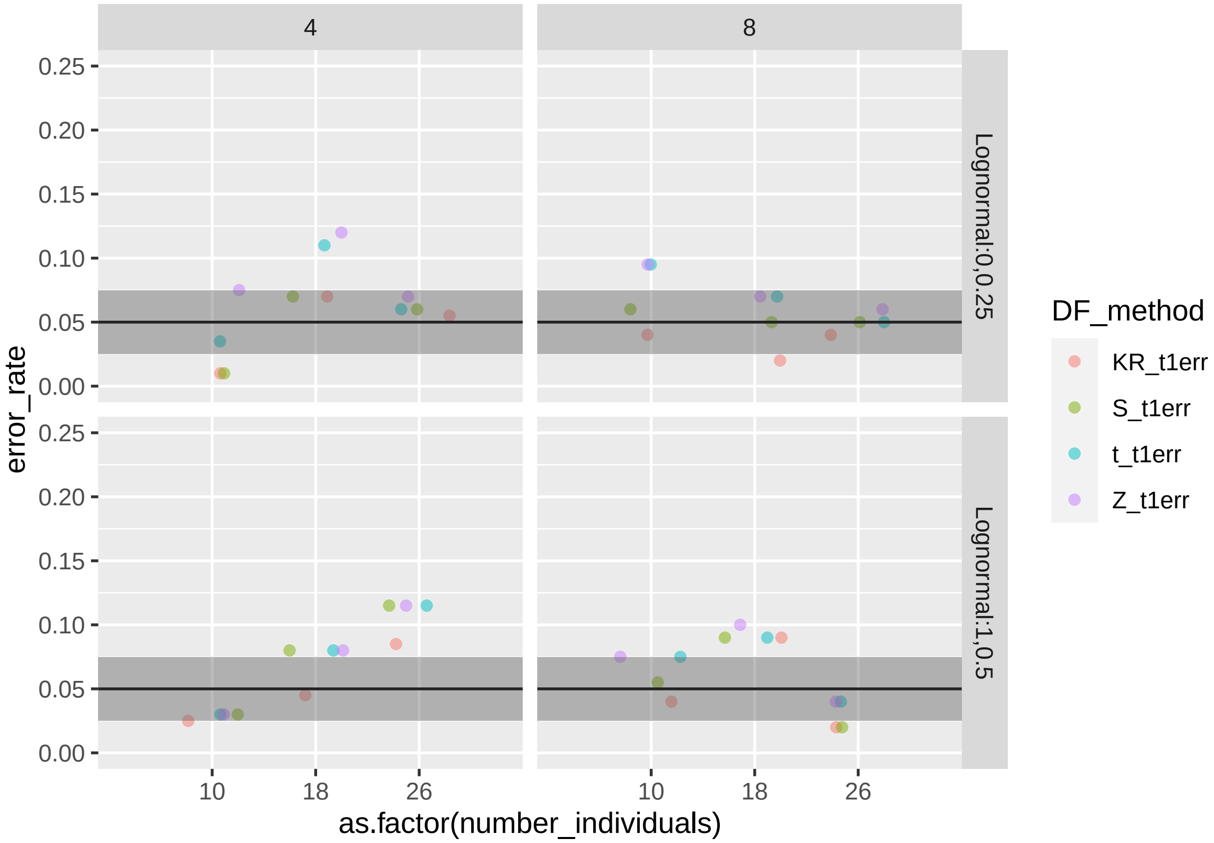
At $\lambda = .9$ and smaller sample sizes, the methods used on the random intercept model do produce more conservative, although not more robust, Type I error rates. However, similar to what was found in the earlier distribution, the random slope model at sample size 26 performs better than the random intercept, and at any size. Increasing the number of measurements to 8 seems to decrease anti-conservatism in DF methods in the random intercept model, but not the random slopes model.

We have to be careful here; we are comparing methods of computing DF, not fitted models. Language should be focused on how the methods perform under different circumstances.

Despite being sampled from the same type of distribution, ~~the application of~~ these different methods of computing DF ~~DF methods~~ to these two exponential distributions produce different trends in error rates. One trend that appears overall is that if the sample size is large enough, fitting a random slopes model will produce more robust error rates compared to a random intercept model in both distributions. Although these two distributions have the same variability is also changing!

skewness and kurtosis values, the difficulty in parsing these trends suggests that there are other aspects of this distribution that are affecting the performance of the DF methods, and generalizing to other exponential distributions is not recommended.

2.9 Lognormal



As seen in the first figure, across the lognormal distributions, random intercept models had consistently more robust error rates in comparison to random slope. For ease of interpretability, FIGURE ? displays error rates without disaggregating by complexity of model.

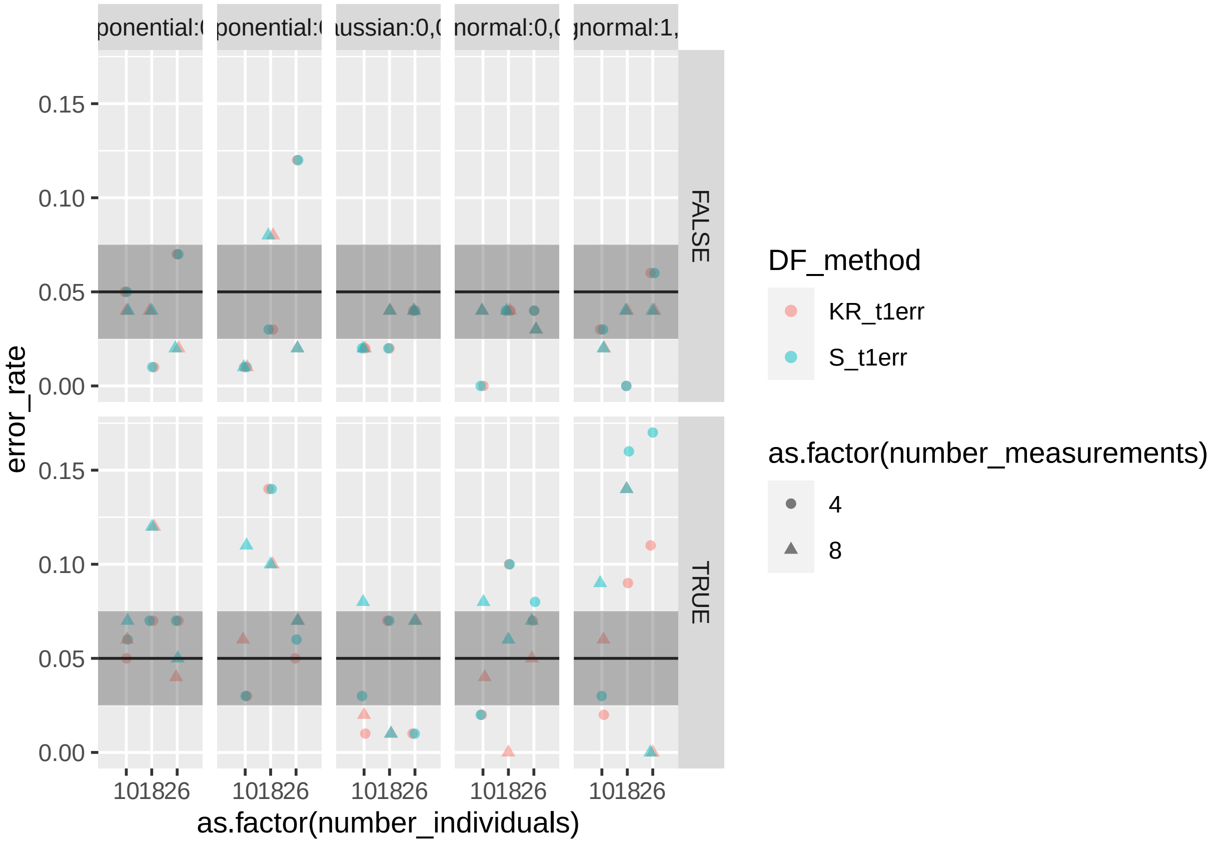
Our first lognormal distribution with parameters (0, .25) has lower values of kurtosis and skewness. At 4 measurements, performance of the 4 DF methods is more variable, but does converge and become more robust once the sample size increases. At sample size 26, all 4 methods yield robust error rates. In contrast, in 8 measurements, the performance of the methods is relatively stable, with t-as-z and standard DF method being slightly anti-conservative at smaller sample sizes. It appears that once the number of measurements has increased, the effect of increasing sample size is still positive, but less significant in this distribution.

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 is much different. At 4 measurements, the robustness of the 4 measures seems to decrease as sample size increases. At a sample of size 10, the performance is very robust and conservative across all methods. However, at size 26 all methods are anti-conservative. While increasing to 8 measurements increases overall robustness, the difference is not extreme. On a similar note, while increasing the sample size does not lead to worse performance in DB methods as in the 4 measurements condition, it is not signficantly better compared to smaller sample sizes. At sample size 26, the KR and Satterthwaite methods are too conservative to be considered robust, while at sample size 10 they are robust. It appears that in this distribution, sample size and number of measurements appear not to have a strong effect on Type I error rates.

2.9.1 KR vs Satterthwaite

Comparing performance across all 4 methods has yielded signficant 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. The following figure aims to narrow in on differences in performance between the two methods. Looking specifically at the effects of kurtosis and skewness, both KR and Satterthwaite methods tend to produce more anti-conservative error rates in nonnormal distributions

			Sample Size				
			10		18		
			Random Intercept	Random Slope	Random Intercept	Random Slope	Ra
skew	kurtosis	DF_method	random_intercept_10	random_slope_10	random_intercept_18	random_slope_18	rand
Normal							
0.000	0.0	KR_t1err	0.020	0.015	0.03	0.040	
0.000	0.0	S_t1err	0.020	0.055	0.03	0.040	
Lognormal							
0.778	1.1	KR_t1err	0.020	0.030	0.04	0.050	
0.778	1.1	S_t1err	0.020	0.050	0.04	0.080	
1.750	5.9	KR_t1err	0.025	0.040	0.02	0.115	
1.750	5.9	S_t1err	0.025	0.060	0.02	0.150	
Exponential							
2.000	6.0	KR_t1err	0.028	0.050	0.04	0.108	
2.000	6.0	S_t1err	0.028	0.068	0.04	0.108	



2.9.2 KR Only

While KR method appears to be the most robust adjustment, (TABLE 5?) depicts its relatively variable performance across different conditions. Careful consideration must be used when conducting inference, and if possible, an increase in both sample size and number of measurements appears to ensure more robust results.

				Sample Size			
				10		18	
				Random Intercept	Random Slope	Random Intercept	Random Slope
params	number_measurements	skew	kurtosis	10_FALSE	10_TRUE	18_FALSE	18_TRUE
Exponential							
0.2	4	2.000	6.0	0.05	0.05	0.01	0.07
0.2	8	2.000	6.0	0.04	0.06	0.04	0.12
0.9	4	2.000	6.0	0.01	0.03	0.03	0.14
0.9	8	2.000	6.0	0.01	0.06	0.08	0.10
Normal							
0,0.2	4	0.000	0.0	0.02	0.01	0.02	0.07
0,0.2	8	0.000	0.0	0.02	0.02	0.04	0.01
Lognormal							
0,0.25	4	0.778	1.1	0.00	0.02	0.04	0.10
0,0.25	8	0.778	1.1	0.04	0.04	0.04	0.00
1,0.5	4	1.750	5.9	0.03	0.02	0.00	0.09
1,0.5	8	1.750	5.9	0.02	0.06	0.04	0.14

2.10 Discussion

Can I compare across distributions?

Ultimately, these results strongly support using either KR or Satterthwaite degrees of freedom adjustments as opposed to methods aimed towards larger sample sizes.