F Test vs. Likelihood Ratio tests

Comparing Results when the Number of Replications Differ

Carly Lundgreen

Introduction

When attempting to discern whether levels of particular treatments (or comparison groups) have significant effects, both the F test and the Likelihood Ratio test are valid approaches. Both approaches allow researchers to obtain test statistics for hypothesis tests. The null and alternative hypotheses (H_0 and H_a respectively) depend on the inference question at hand. When attempting to evaluate whether levels of a treatment have a significant effect on the response, it is useful to consider 'contrasts' of the β, the vector of effects that has dimensions px1, where p is the number of unique treatment combinations or groups. The idea behind the Likelihood Ratio test is to compare two models, one that includes the effect(s) of interest and one that does not. These approaches will be discussed more in the **Methods** section.

We are interested in discovering whether changing the number of replications per group/combination of treatments has an effect on the results of the F test or the Likelihood ratio test when evaluating whether treatments, and their interaction, have a significant effect on the response.

Data

We will conduct a simulation study to answer our research question. The data to be generated will assume the form of a 2x2 factorial experiment with treatments A and B, each with 2 levels. On each block, there will be 2 replications per treatment combination, for 8 observations per block. We will assume that there are 3 blocks, for a total of 24 observations.

Now, we will repeat the same scenario (2x2 factorial experiment with 3 blocks) for 10, 30, and 50 replications per treatment combination. There will be four datasets of 24, 120, 360, and 600 observations each. Let the true treatment means for the six experimental combinations be: $\mu_{A1B1} = 92$, $\mu_{A1B2} = 93$, $\mu_{A2B1} = 98$, $\mu_{A2B2} = 99$. Let $\sigma^2_{error} = 16$ and $\sigma^2_{block} = 20$.

After generating these 4 preliminary datasets, we will generate 50 datasets under each of the 4 different replication numbers and evaluate these simulated datasets using the methods described in the following section.

Methods

The response, y, an nx1 vector, is distributed as follows:

$$y = X\beta + Zu + \epsilon, \epsilon \sim N(0, R),\ u \sim N(0, G)$$

Where $R = \sigma^2_{error} * I_n$, where n is the number of total observations, and $G = \sigma^2_{block} I_r$, where r is the number of blocks/random effects. Note that R is an nxn matrix, G is an rxr matrix, u is an rx1 vector, and ϵ is an nx1 vector. The treatments are considered fixed efffects, and we only desire to make inference to these two treatments. The blocks are considered the random effects, and we desire to make inference to any possible block that could come into the data in the future.

We will calculate the F values for the generated datasets to test the null hypotheses of no treatment A effect, no treatment B effect, and no interaction between treatment A and B. This is accomplished by assigning contrast matrices that corresponded to the three previously stated null hypothesis. Contrasts are linear combinations of β that correspond to the inference questions and can be used in the calculation of an F statistic. The null hypothesis of no treatment A effect implies that the means/effects for the two levels of A are not significantly different from one another. In other words, H_0 was that $0.5(\beta_{A1B1} + \beta_{A1B2}) - 0.5(\beta_{A2B1} + \beta_{A2B2}) = 0$. The corresponding result holds for the levels of B.

The respective contrasts to test the null hypotheses of no treatment A effect, no treatment B effect, and no interaction (the product of the two), are as follows: $C_A = [1/2, -1/2, 1/2, -1/2]$, $C_B = [1/2, 1/2, -1/2, -1/2]$, and $C_I = [1, -1, -1, 1]$. These contrasts are used in the following formula to calculate the F statistic (and the p-value calculated from the cdf value at that test statistic):

$$F = \frac{1}{r}(C\hat{\beta})'[\hat{V}C\hat{\beta})]^{-1}(C\hat{\beta})$$

After calculating the F statistics and corresponding p-values, we will perform likelihood ratio tests (LRT) for each of the main effects and the interaction term. The same null hypotheses will apply. The idea behind the LRT is to fit a 'reduced' model that does not include the factors/effects that we want to test and a 'full' model that includes all factors. If H_0 is true, then we would expect that the likelihood function of the reduced model will not differ significantly from the likelihood function at the full model. The full model is estimating parameters that do not improve the fit to the data (assuming H_0 is true.) After fitting these two models, we calculate $-2loglik(reduced) - (-2loglik(full))$, or the difference between residual and null deviance. This value follows a χ^2 distribution with $df_1 = r$ and $df_2 = n - r$. Using this result, we calculated p-values from the LRT. Note that maximum likelihood (ML) was used, rather than restricted maximum likelihood (REML), to calculate the estimates of β_i. This was to keep the analysis simple. A further analysis comparing REML to ML may be interesting.

Results

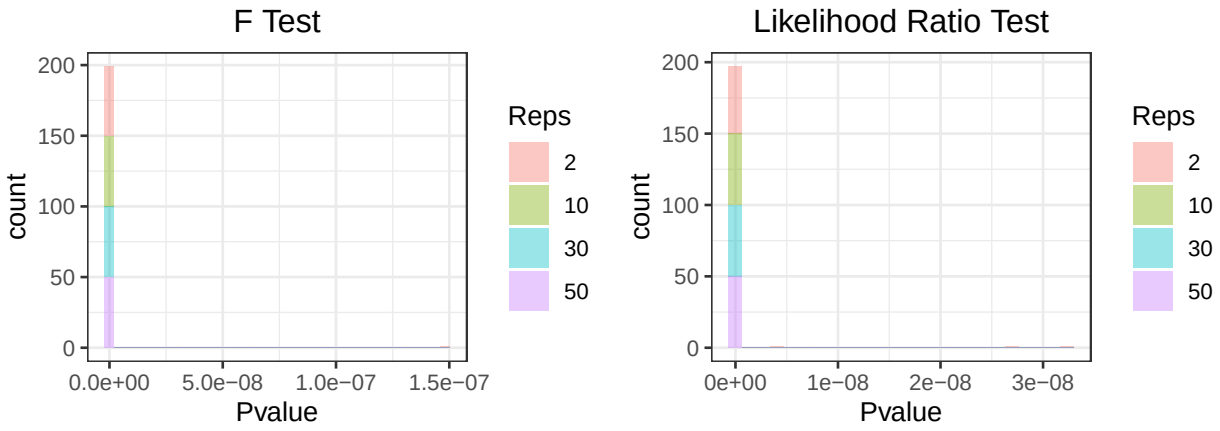
The following table shows the resulting p-values from F tests using 2, 10, 30, and 50 replications per treatment combination. Out of interest, we included 100 replications as a comparison. For the test of no treatment A effect, there was a significant treatment A effect detected regardless of the number of replications. A treatment B effect was detected at 2 replications, and then the result was not significant for 10 and 30 replications. Interestingly, the effect was found to be significant again at 50 replications (and 100 as well). A significant interaction effect was never found with the F test.

Repetitions	A	B	Interaction
2	0	0.0250	0.4362
10	0	0.2607	0.8858
30	0	0.1585	0.9190
50	0	0.0000	0.9636
100	0	0.0000	0.1421

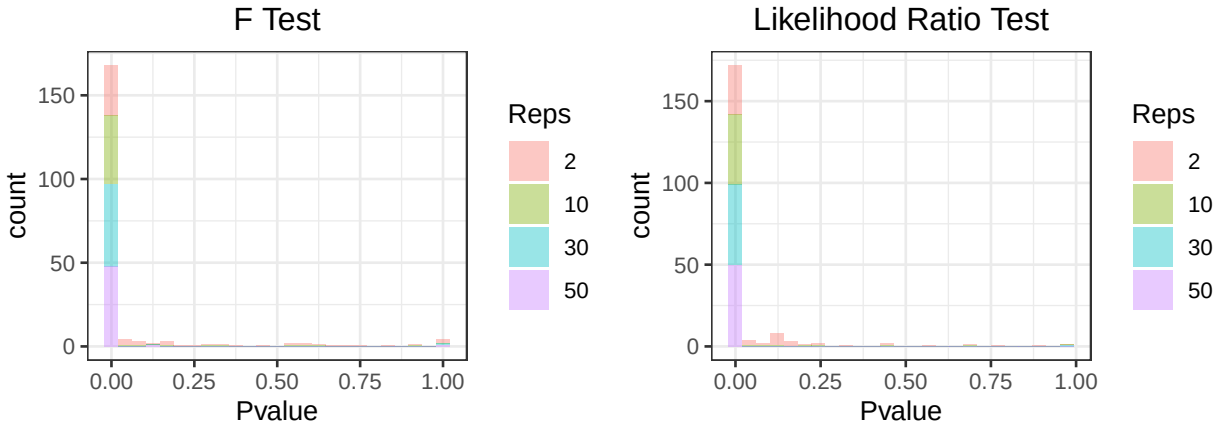
The next table shows similar results from the LRT method. There was not a significant effect of treatment A using only 2 replications per treatment, but for all other replications, the effect of A was significant. The pattern seen in treatment B is nearly the same as was observed with the F test. The interaction effect was significant for 2 replications per treatment combination, but not for higher values.

Repetitions	A	B	Interaction
2	0.268	0.0052	0.0002
10	0.000	0.4132	0.5529
30	0.000	0.3871	0.7434
50	0.000	0.0091	0.8069
100	0.000	0.0021	0.1923

The following histograms show the p-values from 50 simulated datasets using both the F and LRT methods to analyze whether there was a significant treatment A effect (i.e. the response variable depended upon the level of treatment A). Clearly, nearly all p-values from these simulations were essentially zero. The F test and LRT behave very similarly in this scenario no matter the number of replications.

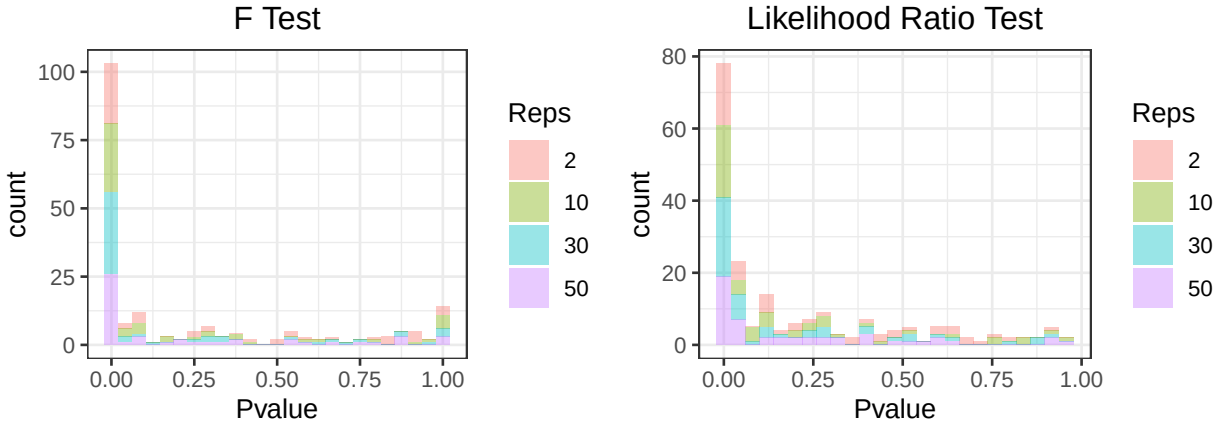


The next set of histograms draw the same comparison between F and the LRT, only for the test of the null hypothesis of no treatment B effect. Both methods start to have some p-values creeping up above the significance level of 0.05, but the majority are still close to zero. Again, the F and LRT behave similarly here. Interestingly, it seems that the larger p-values are mostly coming from datasets that used 2 replications per treatment combination.



The final set of plots compares F and LRT for the null hypothesis of no interaction effect. It appears that both the F and LRT had a bit more difficult time detecting a significant interaction effect, with some p-values

even approaching 1. We saw this in the results from the 4 preliminary datasets with larger p-values for the test of the interaction. It appears that the number of replications are all distributed relatively evenly among these higher p-values, so it appears that the number of replications doesn't have a very large effect on the results of the F test or LRT.



Discussion

Overall, it does not seem that changing the number of replications changes the results of the LRT or the F test. They seem to perform similarly in all cases, with the exception of detecting significant effects of treatment B for 2 replications per treatment combination. The treatments assigned in this simulation study, however, were completely arbitrary, as were the true effects and the variances. In order to determine whether the effect of the small replication number is not dependent upon the design of our experiment, it would be helpful to expand the analysis to multiple treatment designs. Another interesting point of further analysis may be to compare REML results to the ML results we observed in this analysis. Because the variance estimator given by ML is biased, this may have an effect on the results of the LRT (or the F test). This analysis only considered a few replication numbers, and only up to 50, because of the heavy computation required to simulate many large datasets and fit linear mixed effects models on these datasets.

The limitations described above are a caveat to the results of the simulation study. However, the general results indicate that the F test and the Likelihood Ratio test do not behave differently from each other for various replication numbers. In other words, the number of observations in each block will likely not change the result if you perform an F test or a likelihood ratio test to evaluate the effect of a treatment on the response.