POOLING OF VARIANCES: THE SKELETON IN THE MIXED MODEL CLOSET?

Philip M. Dixon
Department of Statistics,
Iowa State University,
Ames IA, 50011

Abstract:

I explore three related issues concerning pooling of error variances: when is it appropriate (or not) to pool, how best to evaluate equality of variances, and whether there is a cost to never pooling. I focus on pooling decisions in a combined analysis of a multi-site experiment. A-priori, sites should have different error variances. My primary question is whether an analysis that ignores unequal variances is wrong.

I find that ignoring heteroscedasticity between sites maintains, or provides slightly conservative, tests of average treatment effects and treatment-by-site interactions. Models with site-specific variances do provide more powerful tests when variances are different. Never pooling, i.e., using site-specific variances when variances are equal, also reduces power. In contrast to the relatively benign effects of pooling across sites, incorrectly pooling across treatments is much more serious.

AIC-based evaluations of variances are very sensitive to non-normality, with a strong tendency to indicate unequal variances when that is incorrect and the data are non-normal. While Levene's test is somewhat liberal when errors are skewed or heavy-tailed, it is much more robust than AIC.

I conclude that ignoring site-specific error variances is not wrong, but modeling that heterogeneity will increase power. If there is any possibility that errors are non-normal, I suggest that variance models be evaluated using Levene's test instead of AIC.

Keywords: heteroscedasticity, combined experiments, AIC model selection, Levene's test

1 Introduction

The combined analysis of repeated experiments extracts more information from a collection of related experiments than does a series of experiment-specific analyses. As just one example, Thompson et al. (1993), describe what was provided by a combined analysis using mixed models of 12 grazing studies: "The mixed models procedure permitted estimation of the fixed effects of treatments over a broad inference space of future years and different tall fescue pastures over a wide geographic range; detected relationships that had not been apparent in the individual studies, such as the interactions between clover presence and E+ infestation levels; and provided a more coherent body of information than did the results obtained from each discrete study."

Repeated experiments study the same question, usually applying the same treatments, in multiple environments. Because repetition of an experiment can occur in multiple ways, i.e., over years, or over sites, or both, for simplicity we will refer to each repetition as an environment. My primary focus is repetition over sites. Repeated experiments commonly use the same experimental design in each repetition, but this is not essential. Repeated experiments have various names, including repeated experiment or multi-environment trial in the agronomic literature and multi-center clinical trial in the biomedical literature. The combined analysis of data from a repeated experiment uses one model fit to all the observations. Alternate analyses, not discussed here, include meta-analysis (Koricheva et al., 2013) and two-stage analysis (Piepho et al., 2012)

A combined analysis raises issues that are not usually relevant for environment-specific analyses (Moore and Dixon, 2015; Dixon et al., 2020). These issues include how to model the environment-by-treatment interaction (as a fixed effect or as a random effect), whether and how to subdivide a random environment-by-treatment interaction, and whether or not to pool error variances. Dixon et al., 2020, discuss all three issues. This paper elaborates on the issue of pooling error variances. Specifically, I discuss the consequences of pooling when variances are not equal, the consequences of not pooling when variances are equal, and how to evaluate whether error variances are similar. I will use an repeated oat cultivar study to illustrate the issues and simulation to evaluate consequences.

1.1 Consequences of ignoring heteroscedasticity in simpler situations

There are four general approaches to pooling:
1. Assume that a treatment only shifts the population mean, so always assume equal variances.
2. Assume that a treatment may change both the population mean and population variance, so always assume unequal variances.
3. Use the data to decide whether to assume equal or unequal variances.
4. Model the variance using a function of the mean, as in a generalized linear model.

There is a large literature discussing these approaches in simpler situations such as comparison of means from two or more independent samples. Especially good summaries of this literature are in Miller (1986), Madansky (1988), and Keppel and Wickens (2004). The literature on pooling

shows many strong opinions and only a moderate amount of consensus, especially when applied practices are compared across fields, e.g., agronomy and psychology.

My sense of the prevailing opinion in agriculture and biology includes:
1. When there is a variance to mean relationship, transform the data to more equal variances or use a generalized linear model.
2. Two-sample t-tests and overall F tests in a one-way ANOVA are robust to unequal variances so long as sample sizes are equal.
3. Factor-specific F tests in a factorial ANOVA are sensitive to unequal variances across levels of that factor and robust to unequal variances across levels of the crossed factors (Box 1954).
4. Comparisons of pairs of treatments after an ANOVA are sensitive to unequal variances.
5. Likelihood-based tests of equal variances are very sensitive to non-normality.
The consequences of pooling have not been investigated for repeated experiments. Those consequences might differ because repeated experiments are more complicated than what has been previously studied. These additional complications include interactions between factors (not considered by Box 1954) and potentially a mixed model when some interactions are modeled as a random effect.

2 A model for data from a repeated experiment, with some variations

I will focus on data from a balanced randomized complete block design, repeated in multiple environments. One common model for such data is:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \alpha\beta_{ik} + \tau_j + \alpha\tau_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\sim N(0, \sigma^2), \end{aligned} \tag{1}$$

where α_i are the environment effects, $\alpha\beta_{ik}$ are block effects nested within environments, τ_j are the treatment effects, and $\alpha\tau_{ij}$ are the environment-by-treatment interaction effects, and ε_{ijk} are the observation-specific errors. In model (1), errors are assumed to come from a single distribution, so they all have the same variance. Later, this assumption will be modified.

For simplicity of exposition, the rest of this section describes properties for completely balanced data, with the same number of replicates for each treatment in each environment, and assumes all quantities are estimable. In general, all the statements below apply to estimable functions of model parameters but that will be left unsaid.

Model (1) can be varied many ways. Different choices of experimental design, e.g., completely randomized, split plot, or lattice, will remove or introduce additional terms to account for the restrictions on randomization (Casella 2008). Those experimental design terms may be modeled as fixed effects or as random effects. For example, block effects may be modeled as random by adding, $\alpha\beta_{ik} \sim \mathrm{N}(0, \sigma^2_{block})$ to model (1). The consequences of the choice of model for block effects is explored in Dixon (2016).

The most important modeling choice is whether the environment by treatment interactions, $\alpha\tau_{ij}$, are modeled as fixed effects or as random effects (Dixon et al. 2020). This choice always changes the interpretation of treatment effects and often has a large effect on the numerical results. When the interaction is considered a fixed effect, inferences about treatment effects, τ_j, describe averages over the specific environments used in the study. This is narrow-sense inference (McLean et al., 1991). When the interaction is considered a random effect, inferences about treatment effects describe averages over a large population of environments. Those environments used in the study are considered to be a simple random sample from that large population. This is broad-sense inference (McLean et al., 1991).

Practically, the choice of fixed or random interaction has large consequences on the results because the precision of treatment effects depends on that choice (Dixon et al., 2020). When the interaction is fixed, the variance of the difference (or linear contrast) among treatment means depends only on the mean square error. When the interaction is random, the variance of the difference (or linear contrast) among treatment means depends on the interaction mean square. Compared to the error mean square, the interaction mean square is generally larger with fewer degrees of freedom. Both characteristics reduce the precision of estimated treatment effects in broad- or intermediate-sense inference.

3 A repeated oat cultivar study

Issues associated with pooling will be illustrated with data from a repeated oat cultivar study, described briefly in Dixon et al., 2020. In this study, 10 oat cultivars were grown in a randomized complete block design. This was repeated at 3 locations (Ames, Kanawha and Washington, all in Iowa) and 2 years (1985, 1986), with 3 blocks per location. The response is harvest index (HI), the ratio of grain to total shoot biomass, expressed as a percentage. The data set is available in the supplemental material for Dixon et al, 2020.

Model (1) was fit to these data. The block effects, $\alpha\beta_{ik}$, were considered random; the block variance component was estimated by REML.

Figure 1 shows average HI for two cultivars in each of the 6 environments (all combinations of locations and years). The Don cultivar has a consistently larger HI than does Cherokee, but the difference between the two appears to vary across locations and years.

The estimated error variances, i.e., $\widehat{\mathrm{Var}}\,\varepsilon_{ijk}$, and block variances, i.e., $\widehat{\mathrm{Var}}\,\alpha\beta_{ik}$, for each environment are given in Table 1. The error variances are similar in 1985 and 1986 but are consistently about twice as large at Ames than at Kanawha.

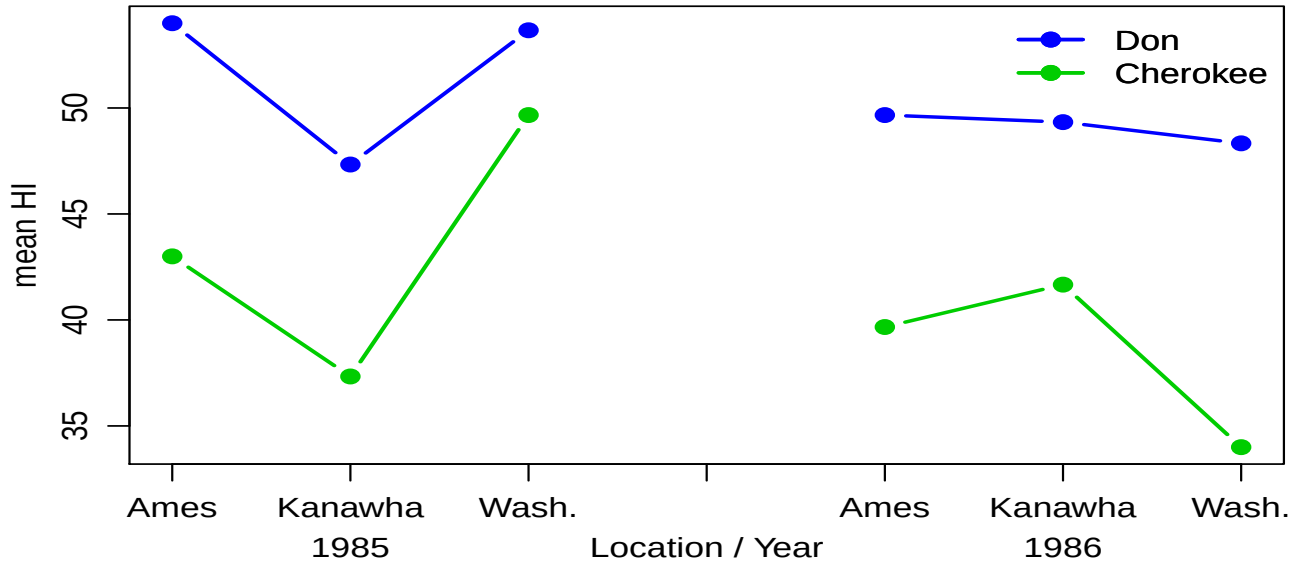

Figure 1: Average harvest index (HI) for 2 oat cultivars, grown in 2 years at 3 locations.

Location	Year	Observations $(\widehat{\text{Var}}\ \varepsilon_{ijk})$	Blocks $(\widehat{\text{Var}}\ \alpha\beta_{ik})$
Ames	1985	18.1	0
	1986	17.1	0
Kanawha	1985	8.8	0.47
	1986	7.5	0
Washington	1985	6.12	0
	1986	11.9	0

Table 1: Variance components for observations and blocks, for each location and year.

3.1 What results change when you change the variance model?

In a repeated experiment, error variances may depend on the treatment or the environment or both. I will focus on heterogeneity among environments. When environment includes both locations and years, as in the oat study, there are at least four variance models:
1. complete pooling, i.e., one variance for all locations and years,
2. pooling over years, i.e., one variance for each location, shared by all years,
3. pooling over sites, i.e. one variance for each year, shared by all locations, and
4. no pooling, i.e., a different variance for each combination of location and year.

To simplify the discussion, I suppress years and consider only the 1985 data. This is consistent with the larger source of heteroscedasticity in error variances (Table 1). I consider two variance models: location-specific error variances or a single pooled error variance.

When a narrow sense analysis is used, the choice of variance model has no effect on either the estimates or their standard errors. The estimated means and standard errors for two cultivars are the same for either variance model, both for this data set (Table 2) and in general (proof in Supplemental material). Intuitively, the proof is analogous to why the type III least-squares means in a factorial ANOVA do not depend on the number of replicates. Because the oat study has an equal number of replicates per location and cultivar and no missing data, the standard errors are also the same (Table 2).

Variety	Pooled	Location-specific
Cherokee	43.33 (1.11)	43.33 (1.11)
Don	51.67 (1.11)	51.67 (1.11)

Table 2: Narrow sense inference for two oat cultivars when variances are pooled and when variances are location-specific. Values are mean harvest index (standard error).

When broad-sense inference is used, the interaction (location*trt) effects are no longer columns of $\boldsymbol{X}$. The estimated cultivar means depend on the variance model (Table 3). The marginal means for each cultivar are a weighted average of the cell means (i.e., means for each treatment and location), with weights that depend on the variance model. With equal sample sizes, under the pooled variance model, each location contributes equally to the marginal mean. Under the location-specific variance model, sites with larger error variances contribute less, in the sense of having a smaller weight, to the marginal mean. Using the Cherokee cultivar as an example, the three location-specific averages are 43.0 for Ames, 37.3 for Kanawha, and 49.67 for Washington. The Cherokee marginal mean with location-specific variances is larger than that with pooled variances because the cell mean for Washington has a smaller variance and hence a larger contribution to the marginal mean.

Variety	Pooled	Site-specific
Cherokee	43.33 (1.65)	43.55 (1.71)
Don	51.67 (1.65)	51.48 (1.71)

Table 3: Broad sense inference for two oat cultivars when variances are pooled and when variances are location-specific. Values are mean harvest index (standard error).

3.2 Which variance model is more appropriate?

Under broad-sense inference, the results for the two variance models are not identical. So which analysis should be reported? Different data-based evaluations provide different answers. AIC, small-sample corrected AIC (AICc) or BIC all suggest that variances are location-specific, but the support for that model is not overwhelming (Table 7). Applying Levene's test to the residuals provides no evidence of unequal variances. Potential reasons for the discrepancy between these results are explored in Section 5.

	Variance model			
Criterion	Pooled	Location	p-value	Decision
AIC	456.0	455.1		location-specific (weakly)
AICc	456.4	455.9		location-specific (weakly)
BIC	456.6	456.1		location-specific (weakly)
-2 log L	326.3	320.7	0.061	location-specific (weakly)
Levene's			0.47	pooled

Table 4: Evaluation of pooled and location-specific variance models. Model selection statistics (AIC, AICc, and BIC) and log likelihoods are for the model with random block(location) and cultivar*location interactions. Levene's test is computed from absolute values of residuals from the same model. Results are similar when based on the model with fixed effects of location, block(location), cultivar and cultivar*location.

4 Consequences of pooling when location-specific variances are unequal

I used simulation to better understand the consequences of pooling when locations have different variances. The hypothetical study is a repeated completely randomized design, with 3 locations and 10 treatments. The simulation scenarios consider different location-specific variances and both equal or unequal numbers of replicates per location. The analysis uses broad sense infer-

ence, so the location*treatment interaction is modeled as a random effect. Because broad-sense inferences about treatment effects depend on the magnitude of the location*treatment variance component, the simulation scenarios also consider a range of location*treatment variances. Details of the simulation scenarious are given in the Location*Trt, Variance and Sample size columns of Table 5.

2500 data sets were simulated and analyzed using SAS PROC MIXED with the Kenward-Rogers adjustment. We focus on inference about differences in treatment means. Some pairs of treatments had the same mean; these were used to estimate the empirical type-I error rate. Other pairs of treatments had different means; these were used to estimate power.

Table 5 shows empirical type-1 error rates for nominal 5% tests in 8 simulation scenarios. The estimated standard error for all estimated error rates is circa 0.3%. Pooling when variances are as much as 10 fold different leads to conservative analyses (type-1 error rate less than nominal), especially when sample sizes are the same at each location (Table 5). For example, when the location-specific variances are 6, 9, and 60 with the same sample size at each location (3/3/3 in Table 5), a nominal 5% test using the pooled error variance has an estimated empirical type-1 error rate of 2.8%. Inferences based on the location-specific variance model for the same data sets are also conservative, with an estimated empirical type-1 error rate of 3.5% (Table 5). As expected, increasing the location*treatment interaction variance component increases the type-1 error rate towards the nominal 5%. Moderately unequal sample sizes do not change the basic conclusion; the empirical type-1 error rates with a pooled variance are conservative or close to the nominal 5%, even when the location with the smallest variance has the largest sample size.

Scenario			When variances are:	
Location*Trt	Variance	Sample size	Pooled	Location-specific
0	6 / 9 / 18	3 / 3 / 3	0.031	0.037
0	6 / 9 / 60	3 / 3 / 3	0.028	0.035
1	6 / 9 / 60	3 / 3 / 3	0.035	0.044
1	6 / 9 / 18	3 / 3 / 3	0.048	0.049
3	6 / 9 / 60	3 / 3 / 3	0.054	0.054
5	6 / 9 / 60	3 / 3 / 3	0.052	0.054
0	6 / 9 / 60	6 / 4 / 2	0.046	0.038
0	6 / 9 / 60	2 / 4 / 6	0.052	0.043

Table 5: Empirical type-I error rates for nominal 5% tests for different combinations of location*treatment interaction variance, location-specific error variances, and number of replicates per location and treatment. Study design mimics the oat experiment, with 10 treatments and 3 locations. The type I error rate is computed from 2500 simulated data sets analyzed by both the pooled variance model and the location-specific variance model.

In repeated experiments, I do not find the large inflation of type-1 error that is seen in simpler

designs when the group with the smallest variance has the largest sample size. Two possible reasons for this are:
1) The focus is on treatment differences, but variances differ among locations. Treatment differences are averaged over locations, so the variance of a treatment mean is the same for all treatments.
2) The usual model for a repeated experiment (1) includes a location*treatment interaction that may absorb unanticipated variability in cell (i.e., location- and treatment-specific) means.

Pooling when variances are unequal does reduce the approximate power of the comparison between treatment means (Table 6). The rejection probabilities reported in Table 6 are "User's power"; they are the probability that a nominal 5% test rejects the null hypothesis. Because the empirical type-1 error rates are not 5%, the values only approximate the power of an $\alpha = 5\%$ test. Even so, the differences between the pooled and location-specific variance models are substantial, especially when the location*treatment interaction variance is small (Table 6). For example, when the location*treatment variance is 0, using a location-specific variance model rejects the null hypothesis of no difference in 80% of the data sets, while the pooled variance model rejects only in 51% of the data sets. It is when the location*treatment interaction variance is small that the cell means, i.e. location- and treatment-specific means, have the most unequal variances. As the location*treatment interaction variance component increases, the power difference diminishes, as expected because the variances of cell means are more similar.

Scenario			When variances are:	
Location*Trt	Variance	Sample size	Pooled	Location-specific
0	6 / 9 / 60	3 / 3 / 3	0.51	0.80
1	6 / 9 / 60	3 / 3 / 3	0.48	0.69
3	6 / 9 / 60	3 / 3 / 3	0.30	0.33
5	6 / 9 / 60	3 / 3 / 3	0.20	0.20
0	6 / 9 / 60	2 / 4 / 6	0.91	0.96
0	6 / 9 / 60	6 / 4 / 2	0.55	0.93

Table 6: Empirical rejection rates for nominal 5% tests when the true treatment difference = 2, for different combinations of location*treatment interaction variance, location-specific error variances, and number of replicates per location and treatment. Study design mimics the oat experiment, with 10 treatments and 3 locations. Rejection rates are computed from 2500 simulated data sets analyzed by both the pooled variance model and the location-specific variance model.

5 Performance of AIC-based variance model selection when errors are non-normal

Model selection statistics such as the Akaike Information Criterion (AIC), small-sample corrected AIC (AICc), or Bayesian Information Criterion (BIC) are the standard approach to choose a model for the random effects in a linear mixed model (Diggle et al. 2002). Hu et al. (2014) illustrates using AIC to choose whether or not to pool variances. However, current knowledge about testing equality of variances suggests that AIC, AICc, and BIC will be very sensitive to the assumption of normality. Both the likelihood-ratio test and Bartlett's test of equal variances are known to be very sensitive to non-normality (Box 1953). Both of these tests are based on the log-likelihood. AIC, AICc, and BIC are also based on the log-likelihood, but their robustness to non-normality has never been evaluated.

Non-normality may be an issue with the oat cultivar study. A normal quantile plot of the residuals from the pooled variance, broad sense inference model (equ. 1) shows weak evidence of heavy tailed residuals (Figure 2).

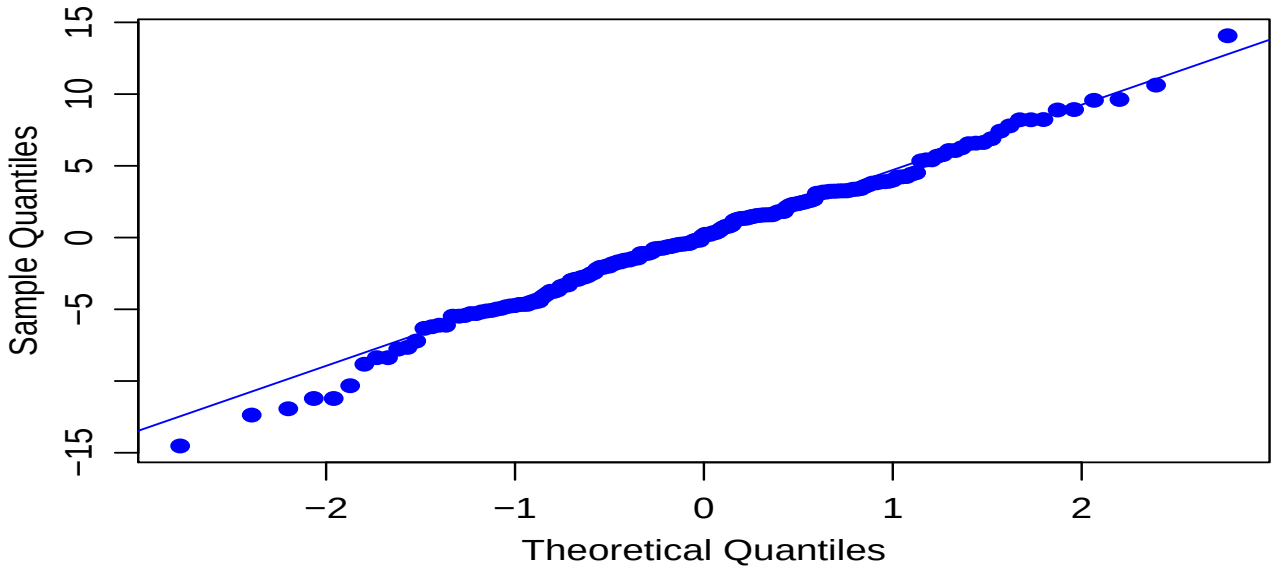

Figure 2: Normal quantile-quantile plot of the residuals from the broad-sense analysis of the oat cultivar study.

I use Tukey g-h distributions (Tukey 1960, Hoaglin 1983) to generate data sets with different amounts of skewness and kurtosis. The Tukey g-h family of distributions has a probability density function with 4 parameters. Two parameters, A, and B control the location and spread. The g parameter controls the skewness. A symmetric distribution has $g = 0$. The h parameter controls the kurtosis. When $h = 0$, the kurtosis = 3, the same as a normal distribution. To simulate a

value from a Tukey g-h distribution, generate $Z \sim N(0,1)$, then transform Z by:

$$\begin{array}{ll} A + B\, e^{h/2Z^2}\, Z & \text{when } g = 0 \\ A + B\, e^{h/2Z^2}\, (e^{gZ} - 1) & \text{when } g \neq 0 \end{array}$$

Figure 3 shows the probability density functions for a normal distribution, $g = 0$ and $h = 0$, 2 non-zero values of g with $h = 0$, 4 non-zero values of h with $g = 0$, and one instance with both skewness and kurtosis ($g = 0.25$, $h = 0.1$).

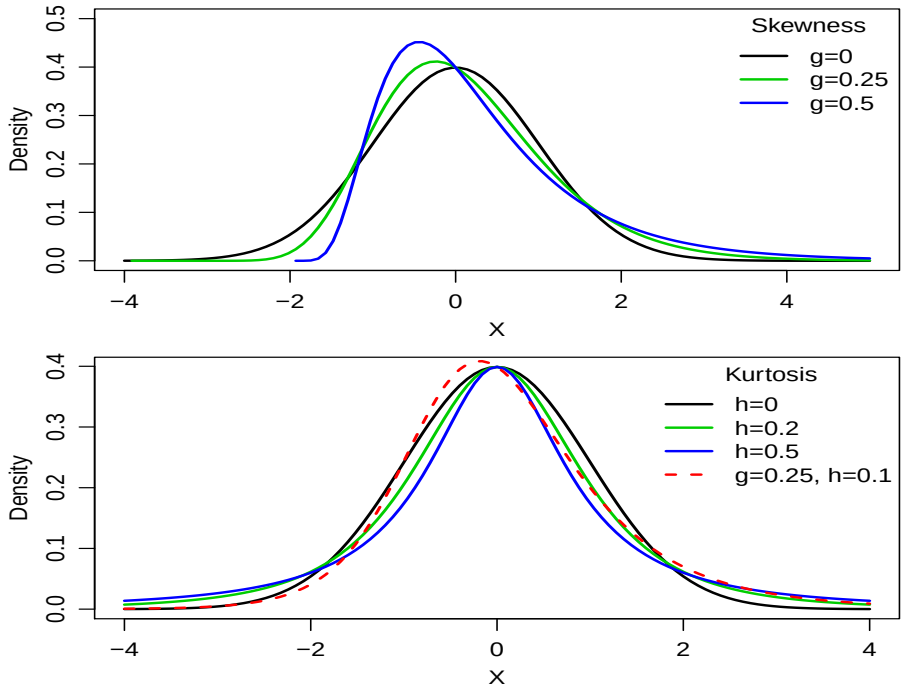

Figure 3: Plots of density functions for distributions in the Tukey g-h family with different values of g controlling the skewness (top panel) and h controlling the kurtosis (bottom panel).

I ran two sets of simulations. One evaluated the performance of AIC and BIC model selection in a one-way ANOVA with all combinations of k=3 or k =10 groups and # replicates per group, N, of 3, 10, 25, or 100. The second set evaluated performance in repeated experiments like the oat cultivar study. That is with 3 locations, 10 treatments, and 3 replicates per location and treatment.

In both sets of simulations, observation errors were generated from distributions with equal variances and the specified values of g and h. The location and scale parameters were set to 0 and 1. In the repeated experiment, the random location*treatment interactions were generated from normal distributions with a variance of 0.2. AIC, AICc, and BIC statistics were computed for the pooled and the unequal variance models. The model with the smaller model selection statistic was recorded for each criterion. This was repeated for 2500 data sets.

None of the model selection statistics are robust to non-normality, in either the one-way ANOVA or the repeated experiment. When observations are simulated from normal distributions (g=0,

h=0), AIC selects the equal variance model, i.e., the model used to generate the data, between 63.8% and 99.9% of the time for the one-way ANOVA and 85.6% of the time for the repeated experiment (Table 7). Increasing either the skewness or the kurtosis reduces the probability of selecting the correct model (Table 7). Kurtosis is a more serious issue than skewness. When $h = 0.5$, AIC selects the correct model between 0.9% and 18.6% of the time for the one-way ANOVA and 18.7% of the time for the repeated experiment.

Increasing the number of replicates per group from $N = 3$ or $N = 10$ per group to $N = 100$ per group is good when the populations are slightly non-normal and bad when the populations are more severely non-normal (Figure 7). For example, when $h = 0.2$, the correct model is chosen 54.7% of the time for $N = 10$ and 68.7% of the time for $N = 100$. But, when $h = 0.5$, the correct model is chosen less often when $N = 100$. Increasing the number of groups from $k = 3$ to $k = 10$ decreases the probability of choosing the correct model unless the populations are close to normal. This is especially so when the populations are heavy tailed. With $k = 10$ groups of $N = 10$ observations with $h = 0.5$, AIC almost always chooses the wrong model (Table 7).

g	h	k=3 N=3	k=3 N=10	k=3 N=100	k=10 N = 10	repeated experiment
0.00	0.0	63.8	91.6	99.9	99.0	85.6
0.25	0.0	57.2	84.4	98.3	92.8	75.7
0.50	0.0	47.5	64.9	82.1	51.8	56.2
0.00	0.1	46.4	74.9	93.5	78.6	68.8
0.00	0.2	34.9	54.7	68.7	37.2	50.5
0.00	0.3	30.8	36.7	39.6	12.6	36.1
0.00	0.4	22.2	23.6	20.2	3.2	25.8
0.00	0.5	18.6	15.4	10.2	0.9	18.7
0.25	0.1	45.3	66.4	80.4	56.2	59.1

Table 7: Probability of AIC choosing the equal variance model when observation errors are from normal and non-normal distributions in the Tukey g-h family. Results for $k = 3$ groups with $N = 3$, $N = 10$, and $N = 100$ observations per group, $k = 10$ groups with $N = 10$ observation per group, and a repeated experiment with 3 locations, 10 treatments, and 3 observations per group.

While the performance of Levene's test, using $\mid Y_{ij} - \hat{Y}_{ij} \mid$, is far from ideal (Table 8), it is much better than that using model selection. For the one-way ANOVA model with normal and non-normal errors, a nominal 5% Levene's test has an empirical type-1 error rate of up to $\approx 14.5\%$. That is for 10 replicates from the population with the largest kurtosis ($h = 0.5$). The performance of Levene's test consistently improves with a larger sample size. For example, for $N = 100$ observations per group, the empirical type-1 error for $h = 0.5$ drops to 8.2%.

There are three different ways to conduct Levene's test for data from a repeated experiment.

g	h	N=10	N=100
0.00	0.0	6.5	4.8
0.25	0.0	8.4	6.4
0.50	0.0	13.0	12.8
0.00	0.1	6.1	6.0
0.00	0.2	8.0	4.5
0.00	0.3	9.4	5.1
0.00	0.4	11.9	6.6
0.00	0.5	14.3	8.2
0.25	0.1	8.7	6.6

Table 8: Empirical type-1 error rates for nominal 5% Levene's tests applied to data from Tukey g-h distributions. g controls the skewness and h controls the kurtosis. Data sets have $k = 3$ groups with $N = 10$ or $N = 100$ observations per group.

The residuals that are the starting point for Levene's test could be estimated from a fixed-effects model with location, treatment, and their interaction, or they could be estimated from a mixed-model where the location*treatment interaction is modeled as a random effect. The model fit to the absolute values of the residuals could include only location and treatment effects or it could additionally include the interaction. I considered three combinations:
narrow/main: fixed effect residuals with location and treatment in the analysis model
narrow/interaction: fixed effect residuals with location, treatment and their interaction in the analysis model
broad/main: mixed model residuals with location and treatment in the analysis model.
2500 data sets were simulated for each of the 9 combinations of g and h and analyzed using the R lm() and lme() functions. The residuals from lme() are the difference between the observed value and the sum of the estimated fixed effect and the BLUP of the random effects.

The empirical type-1 error rates for nominal 5% tests are shown in Table 9. These suggest that the fixed effect residuals should not be used for studies of this size. Levene's tests using the fixed effect residuals have unacceptable type-1 error rates, even for normally distributed data ($g = 0$, $h = 0$). Fitting an analysis model with an interaction is even worse. The performance using the mixed model residuals is much better. The empirical type-1 error rates are above 10% only for the most skewed data sets ($g = 0.5$).

I suspect the poor performance with the narrow-sense residuals occurs because there are only 3 observations for each fitted mean. Each group of 3 residuals sums to zero, which induces a very large negative correlation and distorts Levene's test. The broad-sense residuals do not sum to zero, so their correlation is much smaller. This hypothesis remains to be investigated.

Because the error rates are lower and consistently improve with increasing sample size, I sug-

g	h	Narrow Main	Narrow Interaction	Broad Main
0.00	0.0	14.1	18.5	5.6
0.25	0.0	18.2	25.5	9.1
0.50	0.0	22.1	37.7	12.0
0.00	0.1	18.4	27.8	6.6
0.00	0.2	23.6	38.6	7.2
0.00	0.3	29.3	51.5	7.2
0.00	0.4	32.4	60.0	7.2
0.00	0.5	36.7	68.6	7.3
0.25	0.1	19.2	30.2	7.7

Table 9: Empirical type-1 error rates for nominal 5% Levene's tests applied to data from Tukey g-h distributions from a repeated experiment. g controls the skewness and h controls the kurtosis. Data sets have 3 locations, 10 groups, and $N = 3$ observations per location and group. Narrow and Broad indicate how residuals were estimated, from a fixed effect or mixed model, respectively. Main and Interaction indicate the model used to conduct Levene's test, with location and treatment only, or additionally with their interaction.

gest using Levene's test instead of model selection statistics to assess a variance model. When applied to repeated experiments with relatively few replicates (e.g., 3) per location and treatment, I suggest calculating residuals from a mixed model with a random location by treatment interaction.

6 Why not always fit a location-specific variance model?

Instead of using the data to choose a variance model, one could decide to always use the location-specific variance model. This is the second approach to pooling in section 1.1. What are the consequences of always fitting location-specific variances? This evaluation is ongoing, so the conclusions are preliminary.

Intuitively, these consequences of always fitting location-specific variances will be largest when the equal variance model is actually the correct model. Hence, I simulate data sets from repeated experiments with a relatively small location*treatment interaction variance component. I consider three study designs: one modeled on the oat cultivar study with 3 locations and 10 treatments, one with 10 locations and 3 treatments, and one with 10 locations and 2 treatments. In each, there are three replicates of each combination of treatment and location, all locations have the same error variance, and the location*treatment variance component is 20% of the error variance. All random variables are drawn from independent normal distributions. I simulated 2500 data sets for each of the three study designs and analyzed them using SAS PROC MIXED.

Degrees of freedom were calculated using the Kenward-Rogers approximation.

I focus on inferences about the difference between two treatments from two models: one with pooled error variances (here, the correct model) and the other with location-specific variances. For each variance model and study design, I calculate the empirical type-1 error rate, the variance of the estimated differences, and the average degrees of freedom for the variance of the difference (Table 10). I expect that inferences using the location-specific model (the incorrect model) will have fewer error degrees of freedom and more variable estimates.

# Locations	# trts	variance model	error rate	ave. d.f.	empirical var. diff
3	10	pool	0.045	42.8	0.252
3	10	locations	0.048	40	0.266
10	3	pool	0.044	42.3	0.074
10	3	locations	0.0492	33.3	0.089
10	2	pool	0.037	27.1	0.073
10	2	locations	0.053	19.6	0.096

Table 10: Inferences about the difference of two treatments using pooled and location-specific variance models in replicated experiments with three different combinations of locations and treatments. The error rate is the empirical type-1 error rate of a nominal 5% test, ave. d.f. is the average Kenward-Rogers error degrees of freedom, ave. and var. diff. is the sample variance of the estimated differences.

The type-1 error rates for the location-specific variance models are close to the nominal 5%, while those for the pooled variance model are slightly conservative, i.e., $< 5\%$. This mirrors the results in Section 4. With 10 locations and 2 treatments per location, the error degrees of freedom using the location-specific variance model is substantially smaller than that using the pooled variance model, but both degrees of freedom are large enough that there is only a 2% difference in the 0.975 quantiles of the respective T distributions. However, the estimated differences from the pooled variance model are much less variable than those from the location-specific variance model. The pooled model has a relative efficiency of 1.31 (= 0.096 / 0.073) relative to the location-specific variance model in studies with 2 treatments and 10 locations. Increasing the number of treatments to 3 reduces the differences between the two variance models. The error d.f. for both models are larger and the relative efficiency of the pooled variance model drops to 1.20. With 3 locations and 10 treatments, there is little difference between the two variance models. In summary, when the design has many locations and few treatments per location, as is common in on-farm studies, there is a moderate cost to always using a location-specific model. When there are fewer locations and more information about each location's variance, as with more treatments per location, the cost of always using a location-specific variance model is quite small.

7 Extensions and Recommendations

Discussions of pooling in models for repeated experiments, where there many components, can get very complicated. I have chosen to focus on the conclusions a user will make about treatment main effects. This focus ignores issues such as the estimation of the location*treatment interaction variance component or predictions of location-specific treatment effects. Either could be the topic of another study.

Every random effect in a model reflects a decision about pooling, although this decision is often made by default. For example, if blocks within environments are random, the common default model assumes that variability among block means is the same in each environment. Even when error variances are assumed to be location-specific, I rarely see analyses with location-specific block variances. A data-based decision about block variances will be hard because there are fewer degrees of freedom for blocks than for errors.

I have focused on location-specific error variances in studies with a factorial structure for treatments and locations. Error variances could also vary between treatments. This could arise in an agronomic study when one cultivar is more sensitive than another to random variation in plot characteristics. Box (1954) evaluated the consequences of heteroscedasticity in two-way factorial designs. Applied to locations and treatments, his results imply that treatment-specific error variances have large consequences for conclusions about treatment effects and minimal consequences for conclusions about location effects. Hence, if there is concern about unequal variances in a repeated experiment, I recommend that the priority be to evaluate treatment-specific variances.

My focus has been on estimation of treatment means and their differences. In variety trials, cultivar is often modeled as a random effect, because this provides more accurate predictions of performance at new locations or in future years. Those predictions are functions of the error variance. When error variances are location- or environment-specific, it is unclear how to make predictions of random treatment effects. Intuitively, pooling provides an estimated error variance that is an average over environments and could be used to make predictions for new environments. The properties of such an approach remain to be studied.

I have shown that there are clear advantages to using the correct model for error variances. When variances are equal, pooling gives more precise estimates of treatment means and their differences. When variances are not equal, using location-specific variances gives more powerful tests of treatment differences. However, it can be very difficult to determine the correct model, especially in studies with few replicates per treatment and location. AIC-based model selection is very sensitive to non-normality of the residuals, so I recommend using Levene's test, which is more robust. For repeated experiments, residuals from broad-sense inference provide the best calibrated Levene's test. When the correct model is unclear, there is no harm in pooling error variances, but fitting location-specific variances will increase the power.

8 Acknowledgements

I thank Walt Stroup for his insightful comments on the talk that was the precursor to this paper. In particular, his question, "why not always fit a location-specific variance model?", led to that section of this paper. I also thank the statistical consulting group at Iowa State for their many thoughtful questions and comments. Any errors and misconceptions are mine alone.

9 References

Box, G.E.P. 1953. Non-normality and tests on variances. Biometrika 40:318-335.

Box, G.E.P. 1954. Some theorems on quadratic forms applied in the study of analysis of variance problems. 2. Effects of inequality of variance and of correlation between errors in the 2-way classification. Annals of Mathematical Statistics 25:484-498.

Casella, G. 2008. *Study Design* Springer, New York.

Diggle, P.J., Heagerty, P., Liang, K-Y, and Zeger, S.L. 2002. *Analysis of Longitudinal Data*, 2nd ed. Oxford University Press, New York.

Dixon, P.M. 2016. Should blocks be fixed or random. Conference on Applied Statistics in Agriculture. available online: https://doi.org/10.4148/2475-7772.1474

Dixon, P.M., Moore, K.J., and van Santen, E., 2020. The Analysis of Combined Experiments. Chapter 8 in Glaz, B. and Yeater, K.M. (eds.) Applied Statistics in Agricultural, Biological, and Environmental Sciences. Agronomy Society of America.

Hoaglin, D.C. 1983. g-and-h distributions, In S. Kotz and N. L. Johnson (Eds.), Encyclopedia of Statistical Sciences, Vol. 3, pp. 298-301, Wiley, New York, 1983.

Hu, X., Yan, S., and Li, S. 2014. The influence of error variance variation on analysis of genotype stability in multi-environment trials. Field Crops Research 156:84-90.

Keppel, G. and Wickens, T.D. 2004. *Design and Analysis: A Researcher's Handbook.* Pearson, Upper Saddle River, NJ.

Koricheva, J., Gurevitch, J., and Mengersen, K. (eds.) 2013. *Handbook of Meta-analysis in Ecology and Evolution.* Princeton University Press, Princeton NJ.

McLean, R.A., Sanders, W.L., and Stroup, W.W. 1991. A unified approach to mixed linear models. American Statistician 45(1):54-64.

Miller, R.G. Jr. 1986. *Beyond Anova, basics of applied statistics*. Wiley, New York.

Moore, K.J. and Dixon, P.M. 2015. Analysis of Combined Experiments Revisited. Agronomy Journal 107:763–771

Muller and Zhao. 1995. On a semi parametric variance function model and a test for heteroscedasticity. The Annals of Statistics. 23 (3):946–967

Piepho, H-P, Möhring, J., Schulz-Streeck, T., and Ogutu, J.O. 2012. A stage-wise approach for the analysis of multi-environment trials. Biometrical Journal 54:844-860.

Thompson, R.W., Fribourg, H.A.. Waller, J.C., Sanders, W.L., Reynolds, J.H., Phillips, J.M., Schmidt, S.P., Crawford, R.J., Allen, V.G., and Faulkner, D.B. 1993. Combined analysis of tall fescue steer grazing studies in the Eastern United States. Journal of Animal Science 71: 1940-1946.

Tukey, J. W. 1960. The Practical Relationship between the Common Transformations of Counts of Amounts. Technical Report 36, Princeton University Statistical Techniques Research Group, Princeton.

10 Supplemental material

Comparison of treatment means estimated by narrow-sense inference with pooled- and location-specific variances.

For any study, write the treatment design matrix as $\boldsymbol{X}$, not necessarily of full column rank. Under the pooled variance model, estimates of any estimable function of the treatment effects, $\boldsymbol{C\beta}$ can be written as:

$$\widehat{\boldsymbol{C\beta}} = \boldsymbol{CX}(\boldsymbol{X'X})^{-}\boldsymbol{X'Y} = \boldsymbol{C}\widehat{\boldsymbol{Y}}_{ols}. \tag{2}$$

Under the unequal variance model, the estimates are:

$$\widehat{\boldsymbol{C\beta}} = \boldsymbol{CX}(\boldsymbol{X'WX})^{-}\boldsymbol{X'WY} = \boldsymbol{C}\widehat{\boldsymbol{Y}}_{wls}, \tag{3}$$

where $\boldsymbol{W}$ is a diagonal matrix with the reciprocal variances for each observation along the diagonal. These are equal iff $\boldsymbol{CX}(\boldsymbol{X}^{'}\boldsymbol{X})^{-}\boldsymbol{X}^{'} = \boldsymbol{CX}(\boldsymbol{X}^{'}\boldsymbol{WX})^{-}\boldsymbol{X}^{'}\boldsymbol{W}$.

Index the elements of $\boldsymbol{Y}$ by three indices, i, j, and k. i includes all treatment effects that have different variances. j includes all treatment effects that have the same variance, and k indexes the replicates. For example, consider a 3 x 4 factorial treatment design in a randomized complete block experiment design with 5 blocks. The variance depends on the combination of treatment factors, so i would have 12 values, one for each combination of treatment factors. j would have 1 value, because all treatments have different variances, and k would have 5 values, one for each

block. As a second example, consider 4 treatments evaluated at 3 locations, with 5 replicates in a completely randomized design at each location. The variance depends on the location but not the treatment. i has 3 values, one for each location, j has 4 values, one for each treatment, and k has 5 values.

It can be proven that $\boldsymbol{CX}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}' = \boldsymbol{CX}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{W}$ when:
1) The treatment design is saturated, so $\widehat{Y}_{ij} = \overline{Y}_{ij.}$ for all i, j, and
2) The weights do not depend on k.

The weighted least squares estimates of $\widehat{\boldsymbol{Y}}_{wls}$ minimize $\Sigma_{ijk} w_{ijk}(Y_{ijk} - \widehat{Y}_{ijk})^2$. Since w_{ijk} depends only on i,

$$\begin{aligned}
\Sigma_{ijk} w_{ijk}(Y_{ijk} - \widehat{Y}_{ijk})^2 &= \Sigma_{ij} w_i \, \Sigma_k (Y_{ijk} - \widehat{Y}_{ijk})^2 \\
&= \Sigma_{ij} w_i \, \Sigma_k \left[(Y_{ijk} - \overline{Y}_{ij.}) - (\overline{Y}_{ij.} - \widehat{Y}_{ijk}) \right]^2 \\
&= \Sigma_{ij} w_i \, \Sigma_k (Y_{ijk} - \overline{Y}_{ij.})^2 + \Sigma_{ij} w_i \, \Sigma_k (\overline{Y}_{ij.} - \widehat{Y}_{ijk})^2 \\
&\quad + 2\Sigma_{ij} w_i \, \Sigma_k (Y_{ijk} - \overline{Y}_{ij.})(\overline{Y}_{ij.} - \widehat{Y}_{ijk})
\end{aligned}$$

The last term in the sum is zero because $\Sigma_k (Y_{ijk} - \overline{Y}_{ij.})(\overline{Y}_{ij.} - \widehat{Y}_{ijk}) = 0$ for all i, j. The first term in the sum is a positive constant. The second term is a weighted sum of non-negative values. This is minimized when $\widehat{Y}_{ij} = \overline{Y}_{ij}$ so long as $\overline{Y}_{ij}$ is in the column space of $\boldsymbol{X}$, which is always the case when the $\boldsymbol{X}$ matrix specifies a separate mean for each combination of location and treatment. Hence,

$$\begin{aligned}
\widehat{Y}_{ijk} &= \overline{Y}_{ij}, \text{for all patterns of } w_i, \text{ including } w_i = 1 \text{ for all } i, \text{ so} \\
\widehat{\boldsymbol{Y}}_{wls} &= \widehat{\boldsymbol{Y}}_{ols}, \text{ so:} \\
\boldsymbol{CX}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y} &= \boldsymbol{CX}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}, \text{ for all } Y, \text{ so:} \\
\boldsymbol{CX}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}' &= \boldsymbol{CX}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{W}
\end{aligned}$$

In general, the estimated variance of $\boldsymbol{C}\widehat{\boldsymbol{Y}}_{ols}$ is not the same as the estimated variance of $\boldsymbol{C}\widehat{\boldsymbol{Y}}_{wls}$. One exception is when all treatments have the same variance, but locations have different variances, and there are the same number of replicates of each combination of treatment and location.

Using BIC to choose whether or not to pool variances:

Here I provide details on the performance of BIC as the criterion to decide whether variances are equal or not. These results are based on the same data sets and model fits described in Section 5.

g	h	k=3 N=3	k=3 N=10	k=3 N=100	k=10 N = 10	repeated experiment
0.00	0.0	32.0	96.6	100.0	100.0	98.4
0.25	0.0	28.2	92.6	99.6	99.9	95.4
0.50	0.0	25.3	78.0	94.0	94.2	83.2
0.00	0.1	22.2	86.5	98.4	99.0	91.5
0.00	0.2	15.7	68.3	84.6	84.2	76.4
0.00	0.3	13.3	52.1	57.4	55.1	60.4
0.00	0.4	9.8	37.5	34.6	28.2	45.0
0.00	0.5	8.7	26.7	20.5	13.7	35.6
0.25	0.1	21.5	79.4	92.9	92.8	84.2

Table 11: Probability of BIC choosing the equal variance model over the unequal variance model when observation errors are from normal (g=0, h=0) and non-normal distributions in the Tukey g-h family. Results shown for $k = 3$ groups with $N = 3$, $N = 10$, and $N = 100$ observations per group, $k = 10$ groups with $N = 10$ observation per group, and a repeated experiment with 3 locations, 10 treatments, and 3 observations per group.