Conference on Applied Statistics in Agriculture 2013 - 25th Annual Conference Proceedings

FIVE THINGS I WISH MY MOTHER HAD TOLD ME, ABOUT STATISTICS THAT IS

Philip M. Dixon

FIVE THINGS I WISH MY MOTHER HAD TOLD ME, ABOUT STATISTICS THAT IS

Philip M. Dixon, Department of Statistics, Iowa State University, Ames IA, 50011-1210

Abstract:
I present five short stories, each describing something I wish I had known and appreciated earlier in my statistical life. The five are Simpson's paradox is everywhere, numerical optimization algorithms can be deceived, you can't always trust the Satterthwaite approximation, BLUP's are wonderful things, and It's good to know Reverend Bayes.

1 Introduction

My statistical training didn't end in graduate school. A lot of what I learnt after grad. school was the details of statistical methods for unusual problems. Along the way, I also learnt some general "big" ideas. In retrospect, I wish I had known about these "big" ideas earlier in my statistical career. These include:

1. Simpson's paradox is everywhere,
2. Numerical optimization algorithms can be deceived,
3. You can't always trust the Satterthwaite approximation,
4. BLUP's are wonderful things, and
5. It's good to know Reverend Bayes.

The rest of this paper elaborates on these points.

2 Simpson's paradox is everywhere

I was introduced to Simpson's paradox as a consequence of aggregation of contingency tables. I wish I had known earlier that analogous issues arise in many other contexts. A common contingency table example is the UC Berkeley sex discrimination case (Bickel et al., 1975). The aggregated admissions rate to graduate school in Fall 1973 strongly suggest bias against women (Table 1).

	total # of applicants	% admitted
Men	8442	44%
Women	4321	35%

Table 1: Graduate admissions rates for men and women at UC Berkeley, Fall 1973

The higher admission rate for men essentially disappears when you disaggregate data from the six largest departments on campus. Four of the six departments have a higher admission rate for women, and considerably so in one department (Table 2).

	Men		Women	
Department	total # of applicants	% admitted	total # of applicants	% admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Table 2: Graduate admissions rates for men and women applying to the six largest departments at UC Berkeley, Fall 1973

Bickel et al.'s explanation of the paradox is that departments differ in both proportion of women applicants and overall proportion of admission. Women tended to apply to competitive departments with low rates of admission for both men and women while men tended to apply to less competitive departments with high rates of admission for both sexes. Simpson's paradox arises from the confounding of proportion of women applicants and overall proportion of admission.

Two conditions are necessary for Simpson's paradox to occur (Hsu 1989):

1. a confounding variable has a strong association with the response variable
2. that confounding variable is unequally distributed in the aggregated groups.

In the UC Berkeley example, the confounding variable is the competitiveness of the department. But it is clear that these conditions are not restricted to contingency tables.

Consider a study of two species, the Earwing and the perfumed Honeytailed Snouters, in the mammalian order of Snouter, *Rhinogradentia* (Stümpke 1967). The entire order is (or was) restricted to the Hi-yi-yi Islands in the Pacific, where they provide an incredible illustration of adaptive radiation on isolated islands. In the case of the Snouters, the noses have evolved unusual uses (Figure 1). The definitive (and only) monograph on the order is that by

Stümpke (1967). Unfortunately for further study, the entire group of islands disappeared during atomic bomb testing in the Pacific. An unusually large test 125 miles away triggered tectonic activity and the entire island group sank beneath the sea (Stümpke 1967).

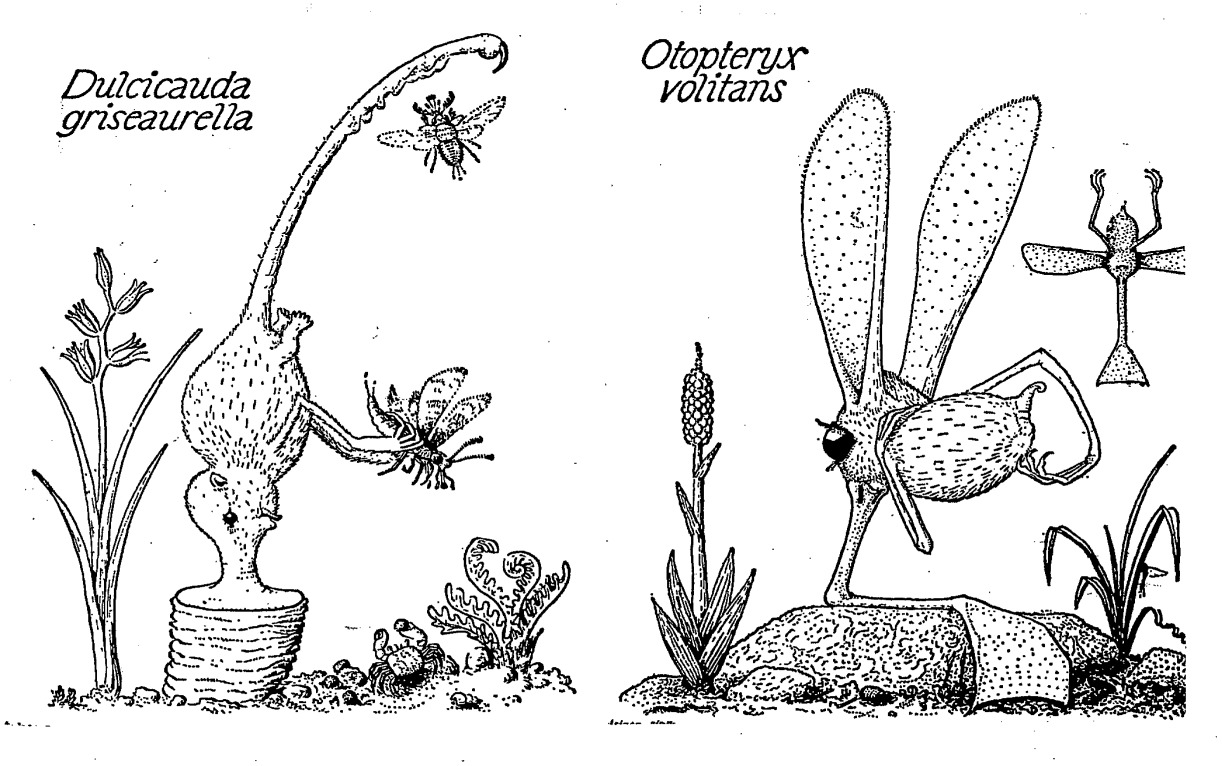

Figure 1: The perfumed Honeytailed Snouter (left) and Earwing (right).

I discovered a collection of specimens of the Earwing and perfumed Honeytailed Snouter hidden in a museum in Germany (where Stümpke worked). I measured the body mass of a random sample of specimens. I was very surprised to find that the mean body mass of the Earwing was slightly larger than that of the Honeytailed Snouter (Table 3). Because the Earwing can fly, while the Honeytailed Snouter has a huge heavy nose, I expected the opposite.

Species	Mean bodymass (gm)
Earwing	20.5
Honeytailed Snouter	19.5

Table 3: Mean body mass of two species of Snouter

It turns out that I measured a mix of juvenile and adult specimens. The Earwing sample is 70% adults while the Honeytailed Snouter is only 30% adults. When age is considered, the Earwing is considerably lighter (Table 4). Again, Simpson's paradox arises because age is strongly associated with body mass and in this sample, age is confounded with species.

	Mean body mass (gm)	
Species	Juvenile	Adut
Earwing	10	25
Honeytailed Snouter	15	30

Table 4: Mean body mass of two species of Snouter, by age.

I meet a third example of Simpson's paradox surprisingly often: examining correlations when the data are structured. Imagine measuring two variables, X and Y, in a series of groups of subjects. The groups could be different fields, different breeds of cattle, different age groups, different treatments in a designed experiment, or anything else that provides additional structure to a set of observations. Many folks calculate and test a correlation coefficient as if the structured observations were a simple random sample from one population, as is assumed by the usual test of correlation = 0. They are not! This problem is especially severe when observations from different treatments in a designed experiment are considered a single sample. If the treatment influences both the X and Y variables, the correlation between X and Y is a consequence of which treatments are used in the study.

When the data are structured, Simpson's paradox arises in many different forms (Figure 2). In one form, X and Y are independent within groups, but the group means differ, so the "ignoring groups" correlation is non-zero (positive in Figure 2a). In a second form, X and Y are negatively correlated within groups, but the "ignoring groups" correlation is positive (Figure 2b). In a third form, X and Y are correlated within each group (Figure 2c shows a positive correlation) but the "ignoring groups" correlation is close to zero. In all three cases, the pooled "within group" correlation is quite different from the "ignoring groups" correlation. Again, the problem underlying Simpson's paradox is confounding of a nuisance variable, in this case group id, with the responses.

The appropriate description of correlations in structured data is more complicated than just desribed. There are three correlations that could be estimated. The third one is the correlation among the group means, illustrated in Figure 3 for the three situations shown in Figure 2. Most statisticians are familiar with decomposing the variability in an observation into sources of variability, where the magnitude of each source is described by the variance component for that source. Similarly, a covariance matrix for two or more variables can be decomposed into "covariance components". These covariance components, and their observation-level sum, describe all three possible correlations.

3 Numerical optimization routines can be deceived

As data analysis has gotten more sophisticated, it has increasingly relied on numerical optimization. Analysis of a linear mixed model by method-of-moments (ANOVA) can be done without numerical optimization. Analysis of that same linear mixed model by REML usually

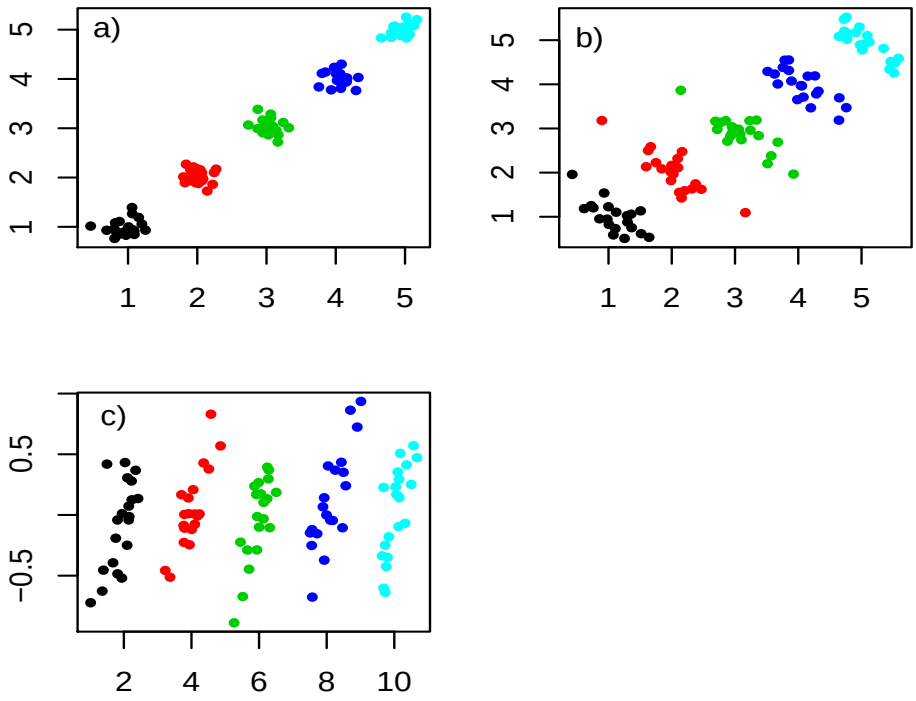

Figure 2: Three illustrations of Simpson's paradox for correlations. Each plot shows 20 data points in each of 5 groups. In all three cases, the 'account for group' correlation is quite different from the 'ignore group' correlation.

requires numerical optimization. Analysis of a generalized linear mixed model often requires two levels of numerical optimization (Gbur 2012). Software to perform this numerical optimization usually includes some form of "convergence check". The goal of this check is to verify that in fact the maximum (or minimum) has in fact been attained. The software decisions are commonly based on one (or more) of three criteria:

1. that the optimization algorithm is making no progress. That is, the change in parameters from one iteration to an another is very small.
2. that the optimization algorithm is making little progress (version 2). that is the value of the objective function (e.g. log-likelihood evaluated at the current parameter estimates) is not changing much.
3. that gradient of the objective function with respect to all parameters is close to zero.

Obviously, we should be concerned when software reports back that there is a problem with convergence. Usually, we do not care when it appears that the reported solution is valid.

A recent experience with a consulting problem taught me to be more sceptical about reported "convergence". A graduate student in plant pathology asked us for help analyzing a data

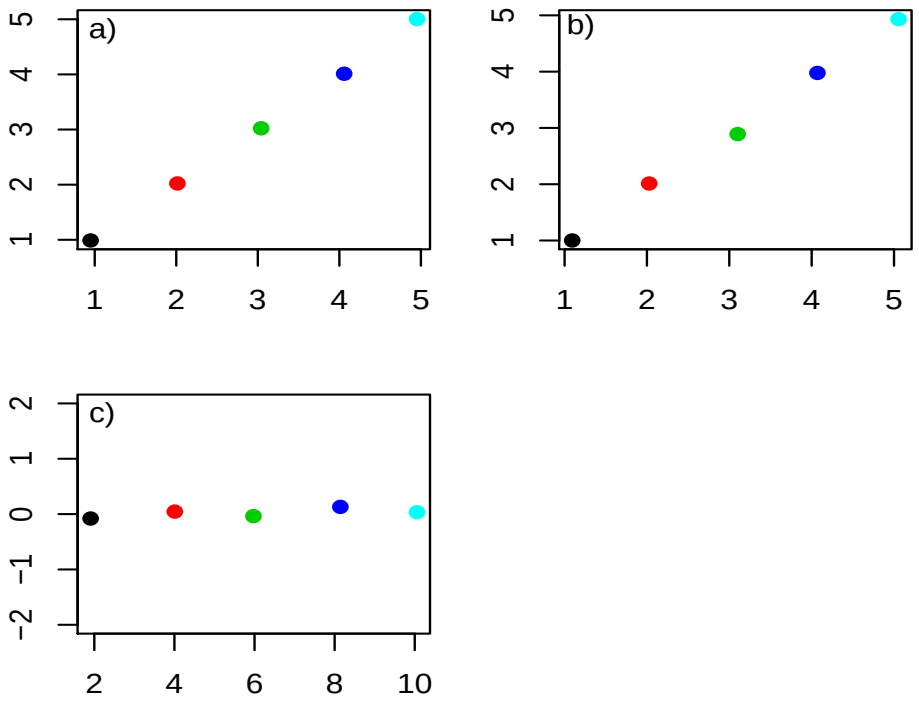

Figure 3: Plots of group means for each example in Figure 2.

set on pathogenicity of isolates of a pathogen on three melon *Curcurbit* species. An isolate is a collection of the fungus from one location. While all isolates are classified within the same species of fungus, they are collected from different locations and are all genetically different.

The student had grown seedlings of three melon species, then randomly chosen seedlings to be inoculated with one of 15 different isolates of a particular fungus. The student's questions concerned the size of the variability among isolates, the mean infectivity on each of the three melon species, and the magnitude of the species*isolate interaction. If the response had been continuous, I would have used a linear mixed model, but it was not.

The response was 'time to first symptom', which was measured by observing plants daily over 14 days and recording the day they first showed symptoms. A value of 14 means that the plant developed symptoms between the observation on day 13 and that on day 14. These are interval censored data. Because of practical considerations, the experiment was terminated at 21 days, even if plants not yet developed symptoms. These are left censored data. The exact response is "fuzzed" because of the data recording process. All "continuous" responses are fuzzed to some extent because of rounding of measurements, but such "fuzzing" is usually ignored. That is appropriate when the number of discrete possible values is large, but I have found that when the number of possible categories is small, e.g. ≤ 10, it is better to treat the data as interval censored, rather than exact.

SAS PROC LIFEREG fits regression models to data containing any combination of left, interval, and right censored observations. We converted an ANOVA model into a regression model the usual way, by constructing indicator variables and omitting the last level to obtain a full rank X matrix. LIFEREG fit this model without any apparent problem. The output includes estimated coefficients and their variance-covariance matrix. But, the student wanted to report marginal means, e.g. for each *Curcurbit* species averaged over isolates and each isolate averaged over *Curcurbit* species. Also, because the 15 isolates came from three different groups within the species, she wanted averages for each group for each melon. And, various tests associated with each estimate. How to do this statistically is well-understood: construct the appropriate C vector or matrix, then estimate $C\beta$ or test hypotheses concerning $C\beta$. To answer the student's questions, we needed a lot of C vectors, each of which had 45 elements.

Karl Pazdernik, one of my consulting students, was helping with the analysis. We decided the most efficient way to get the needed answers was to do the estimation in SAS, write the coefficients and variance-covariance matrix to .csv files, then use R for all the $C\beta$ manipulations. It all seemed straightforward until Karl stopped by my office to tell me that the variance of one of the $C\beta$ estimates was negative. Not good! I told him to check his C vectors. He did - no problems. I checked - no problems. After much hair pulling, it occured to me to check that the variance-covariance matrix was positive-definite. It was not. There were negative eigenvalues. LIFEREG had not converged to the mle's. No sign of problems in the SAS output. It believed it had converged. All coefficients had positive standard errors. It was only when we computed linear combinations of coefficients that we discovered there was a problem. We tightened the convergence criteria in LIFEREG, basically telling SAS to try harder to find the maximum of the likelihood equation, and very soon got a positive-definite variance-covariance matrix.

The modern world of data analysis requires that we trust computers to give us what we ask for. The moral of this story is that sometimes doesn't happen and it can be very difficult to identify those cases.

4 You can't always trust the C-S approximation

Applied statisticians often have to use tests that a theorist considers messy. A common example is the F test for main plot treatments in a split-plot study with unequal sample sizes. Another is most F tests in the analysis of an experiment repeated in location and time, where treatment*location, treatment*time, and treatment*location*time are considered random effects. In both examples, the denominator for the F test has to be constructed. The ANOVA table does not provide a direct estimate of the appropriate denominator mean square.

The customary solution is to use a linear combination of mean squares as the denominator

for an F test, i.e.

$$MS_{denom} = a_1\, MS_1 + a_2\, MS_2 + \ldots.$$

The degrees of freedom (d.f.) associated with MS_1 and MS_2 are known, but what are the d.f. for the linear combination, MS_{denom}? Statistical practice has been to use one of two approaches:

1. "Conservative": use the smaller of the two d.f., or if the linear combination involves more than two MS, use the smallest d.f. for any MS in the linear combination.
2. use the Cochran-Satterthwaite (C-S) approximation: The linear combination $\Sigma_{i=1}^{k} a_i MS_i$ is proportional to an approximate Chi-square distribution with ν_{CS} d.f., where

$$\frac{1}{\nu_{cs}} = \frac{1}{(\sum_{i=1}^{k} a_i MS_i)^2} \sum_{i=1}^{k} a_i^2 MS_i^2 \frac{1}{df_i}. \tag{1}$$

   Hence, the denominator d.f. for the F test is ν_{CS}.

One issue that is often overlooked: Satterthwaite's derivation of the C-S approximation stipulated that all the a_i were positive. But when was the last time you checked that all coefficients were positive? I confess I have regularly used the C-S approximation for any set of coefficients.

Negative coefficients in the linear combination of Mean Squares have some important consequences. When all coefficients are positive, the supports of both the linear combination and a Chi-square distribution are the positive half-line. Not so when one or more coefficient is negative. The support of the linear combination is now the entire real line. Because there is positive probability of a negative estimate, the mean of the linear combination may be small. This means that the approximate Chi-square distribution has a small d.f., and the "Conservative" approach may no longer be conservative.

Consider the C-S calculated d.f. for the linear combinations $1.5MS_1 + 0.5MS_2$ and $1.5MS_1 - 0.5\, MS_2$ for a variety of values for the Mean Squares and their d.f. (Table 5). When the coefficients of the linear combination are positive, the C-S d.f. are between the d.f. for the two mean squares. Not so when one coefficient is negative! The C-S d.f. are always less than the smaller of the two d.f. (Table 5), so the "Conservative" rule is no longer conservative! As the mean square associated with the negative coefficient increases, relative to the other mean square, the C-S d.f. decreases.

As an aside: I now have an explanation for something that has troubled me (and consulting clients) for years, why are C-S d.f. sometimes very small? Because there is a negative coefficient associated with a relatively large mean square.

However, the real question is whether the C-S approximation works even when some coefficients are negative. This can be evaluated in two ways:

				$1.5MS_1 + 0.5MS_2$	$1.5MS_1 - 0.5MS_2$
MS_1	df_1	MS_2	df_2	$\hat{\nu}_{cs}$	$\hat{\nu}_{cs}$
1	5	0	20	5	5
4	5	1	20	5.86	4.19
1	5	1	20	8.65	2.16
1	5	1	200	8.86	2.21
1	5	4	20	18.86	0.38
0	5	1	20	20	–

Table 5: Degrees of freedom calculated using the Cochran-Satterthwaite approximation for two linear combinations of Mean Squares, for various values of the Mean Squares and their degrees of freedom

1. Does the C-S approximate Chi-square distribution have quantiles that match the quantiles of the true distribution of the linear combination of mean squares? A reasonable match is crucial if you want to use the C-S approximation to calculate a confidence interval for or test hypotheses about MS_{denom}.

2. Does using the C-S approximation in an F test provide (approximately) correct type I error rates? A reasonable type-I error rate is crucial if you want to construct appropriate F tests of fixed or random effects.

I estimated the distribution of a linear combination of mean squares by simulating Chi-square random variables with specified d.f. and computing the linear combination. Quantiles of this distribution were computed from 5000 samples and compared to quantiles of the C-S approximate Chi-square distribution with d.f. calculated from the population values of the mean squares. This was done for two cases with all positive coefficients, and a two cases with a negative coefficient. The lower quantiles of the C-S approximating distribution are below the empirical quantiles when both coefficients are positive (cases a and b in Table 6) and above the empirical quantiles when one coefficient is negative (cases c and d). When all coefficients are positive (cases a and b in Table 6), the upper quantiles of the C-S approximating distribution are quite close to the quantiles of the empirical distribution (Table 6). But, they are not very close when one coefficient is negative (cases c and d in Table 6).

But, the performance of the C-S approximation in F-tests of fixed effects (or potentially random effects) is more important to users than are the details of the distribution of a random variable. I evaluated the performance of the C-S approximation in a two group comparison of means, when there is subsampling. The number of subsamples for each experimental unit varied, so there is no exact F test available as a ratio of mean squares in the ANOVA table. The null hypothesis that $\mu_1 = \mu_2$ can be tested in two ways: by an F test comparing the mean square for treatment to a linear combination of mean squares, or as a $C\beta$ t-test comparing the estimated difference between group means to its standard error. This was repeated for 5000 data sets for which the null hypothesis is true. The empirical type-I error rate was estimated as the proportion of tests with p-values less than a specified cutoff.

Case	a_1	a_2	C-S d.f.	Distribution	Quantile: 0.01	0.05	0.10	0.90	0.95	0.99
a)	1.5	0.5	20.8	Simulation	15.25	16.32	17.07	25.08	25.71	25.98
				C-S approx. χ^2	8.77	11.44	13.08	29.37	32.42	38.66
b)	0.8	0.2	19.1	Simulation	14.69	15.50	16.09	23.62	34.81	25.91
				C-S Approx. χ^2	7.71	10.20	11.74	27.34	30.29	36.35
c)	1.5	-0.5	5.2	Simulation	0.01	0.27	0.93	8.92	9.60	10.70
				C-S Approx. χ^2	0.61	1.14	1.72	9.52	11.38	15.44
d)	1.224	-0.224	8.4	Simulation	0.78	2.96	3.76	9.18	10.53	13.35
				C-S Approx. χ^2	1.82	2.96	3.76	13.94	16.13	20.79

Table 6: Comparison of quantiles of the simulated distribution of a linear combination of mean squares to the C-S approximating distribution

In the two cases evaluated here, the empirical type-I error rate is close to the nominal rate both when all coefficients are positive and when one coefficient is negative (Table 7). For practical purposes, the appropriate behavior of hypothesis tests using the C-S approximation in two scenarios is reassuring. One possible explanation for the different conclusions for quantiles and for p-values is that the quantile only evaluates MS_{denom}, the linear combination of mean squares. The F test is a ratio of a numerator to a denominator. The numerator of the F test and the GLS estimates of the difference in groups means depend on the same estimated mean squares used in the denominator. Perhaps some form of compensation occurs, perhaps because of a correlation between the estimated numerator and estimated denominator that results in appropriate behavior of the test.

a_1	a_2	d.f.	Test	Nominal p-value 0.100	0.050	0.010
0.8	0.2	17.6	ANOVA F	0.100	0.051	0.012
			$C\beta$ t	0.099	0.053	0.011
1.224	-0.224	8.4	ANOVA F	0.099	0.048	0.010
			$C\beta$ t	0.094	0.046	0.012

Table 7: Empirical type-I error rates for ANOVA F and $C\beta$ tests with error d.f. calculated using the Cochran-Satterthwaite approximation

The Cochran-Satterthwaite approximation, as given by equation (5), is only usuable when a mean square can be written as a linear combination of mean squares in an ANOVA table. Many mixed model data analyses require computing degrees of freedom for estimated variances that can not be written as linear combinations of mean squares. Degrees of freedom for these more general cases can be estimated using approximations proposed by Giesbrecht and

Burns (1985) and its generalization by Fai and Cornelius (1996) or by Kenward and Roger (1997). In McBride's simulation evaluation of the Fai / Cornelius and Kenward / Roger approximation for a variety of covariance structures for repeated measures data, both approximations provided appropriate a type I error for compound-symmetric variance-covariance structures. The Fai / Cornelius approximation did not maintain appropriate control of the type I error rate when the variance covariance structure was more complicated (Schaalje et al., 2001).

5 BLUP's are wonderful things

The context for this section and the next is modeling the loss of germinability for seeds in long-term storage (Trapp et al., 2012). The USDA maintains germplans storage facilities to preserve germplasm in case it is needed for future breeding programs. The regional facility in Ames, IA preserves over 6,000 lots of maize seeds. The problem is that over time, seeds lose their ability to germinate. The rate of loss differs between seed lots. If the germination rate gets near a critical threshold, curators will grow plants from existing seed to produce new seed. Determining the current germination rate is a destructive test, so lots are tested at haphazard intervals. If the loss of germinability can be modeled, then the time to reach the critical threshold can be estimated. If they trust the model, the seed biologists can wait until shortly before the time to cross the critical threshold to test the seed. This has the potential to considerably reduce the number of germination tests and save a considerably amount of staff time.

In maize seed, germination over time follows one of three general patterns (Figure 4). The traditional pattern is 100% (or almost 100%) germinability at harvest (age 0) followed by a decline in germinability. The high-viability pattern is one where the loss of germinability is very slow. The after-ripening pattern is one where initial germination is low, increases over time, then decreases. All of these patterns can be modeled, at least approximately, using a quadratic polynomial in time.

If we had one seed lot, we would estimate coefficients of the quadratic polynomial using ordinary least squares, or perhaps a model with correlated errors, and then predict germinability at a given age using the BLUE's of the parameters. Our data set included 2833 seed lots that had 3 or more germination values. Each seed lot has something in common (being a maize variety) but the coefficients of its germinability curve are expected to differ from those for other seed lots. Hence, it would make sense to use a random coefficient regression model (equation 2) and use the BLUP's for each seed lot to predict germination at a specific age. Which model makes more precise predictions when averaged over seed lots?

$$Y_{ij} = (\beta_0 + \alpha_{0i}) + (\beta_1 + \alpha_{1i})\, X_i + (\beta_2 + \alpha_{2i})\, X_i^2 + \varepsilon_{ij} \tag{2}$$

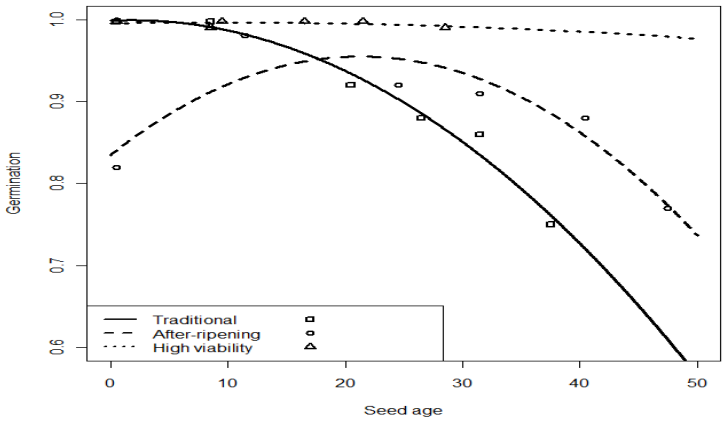

Figure 4: Examples of the three general patterns of maize seed germinability over time

I answered this question by a simulation that was modeled on the real data set. I simulated 100 seed lots, with 1/3 of them represented by 3 observations, 1/3 by 4 observations, and 1/3 by 5 observations. Coefficients for each seed lot were simulated from a multivariate normal distribution (Equation 3). Observations were made at random times between 0 and 26 years. I fit a fixed effects model with separate coefficients for each seed lot and a random coefficients model with a multivariate normal distribution for the coefficients (Equation 3). Deviations from the seed-lot specific germinability curve were modeled as normally distributed with constant variance. This was the same model used to fit the real data set.

After estimating coefficients for the fixed effects model and predicting coefficients in the random coefficient regression, we predicted germination at 30 years (slightly beyond the oldest observed value) and 40 years and calculated the root-Mean-Square-Error of prediction, rMSEP, separately for seed lots with 3 observations, 4 observations, and 5 observations, and for all seed lots.

$$\begin{bmatrix} \alpha_{0i} \\ \alpha_{1i} \\ \alpha_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{bmatrix} \right) \tag{3}$$

Averaged over seed lots, predictions using the BLUP's are considerably more accurate than those using the BLUE's, especially for the seed lots with 3 observations (Table 8). A quadratic model with 3 coefficients fit to a data set with three observations will exactly fit the three observations, the model is fitting the error in each observation. Hence, predictions are poor. The BLUP's use information from other seed lots to help separate error from the mean curve.

		# of years			
		3	4	5	all
30 years:	BLUE	84.4	57.8	18.8	60.0
	BLUP	9.5	7.3	6.2	7.9
40 years:	BLUE	169.5	122.8	44.0	123.5
	BLUP	17.4	14.2	12.6	14.9

Table 8: Root-Mean-Square-Errors of prediction estimated using BLUE's of quadratic coefficients from a fixed-effects model or BLUP's from a random coefficient regression model. The response is percent germination, so predictions range from 0 to 100.

6 Reverend Bayes can be your friend

The previous section evaluated predictions of germinability at a specific age. The seed biologists really wanted to predict when germination will cross the critical threshold, e.g. 50% in Figure 5. That time is a non-linar function of the seed-lot-specific coefficients in the quadratic model. Uncertainty in the prediction at a specific age translates into uncertainty in the estimated time-to-threshold (Figure 5).

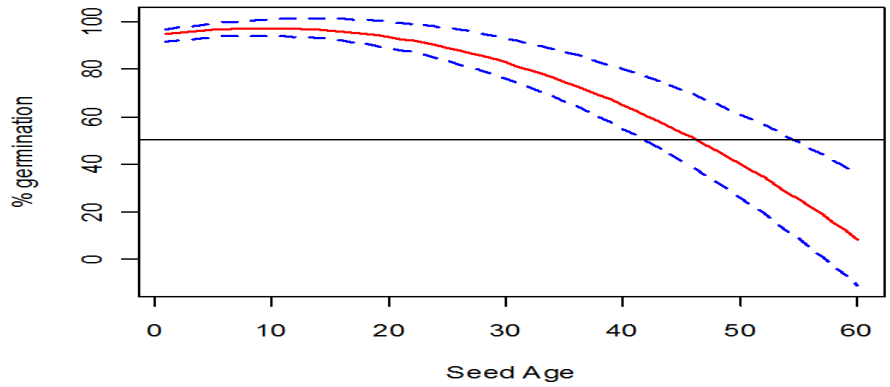

Figure 5: Typical estimated germination curve over time, with a critical threshold at 50% germination. Uncertainty in the germination at a specific age, shown by the dashed blue lines, translates into uncertainty in the estimated time to cross the threshold

A bad prediction of time-to-threshold introduces a cost to the seed biologists. If the prediction is too small (less than the true time-to-threshold), the seed lot is tested before it needed to be. If the prediction is too large (more than the true time-to-threshold), the seed lot will not be tested until after it has crossed the treshold. The costs of these two events are

not equal. Premature testing has a small cost (the staff time for one germination test); late testing has a large cost (loss of genetic variability due to genetic drift and potentially complete loss of the seed lot). To minimize total cost, the prediction should be a lower quantile (e.g. the 0.10 or 0.20 quantile) of the prediction distribution. So, we need to determine the prediction distribution of time-to-threshold for each of the 2833 seed lots (Figure 6).

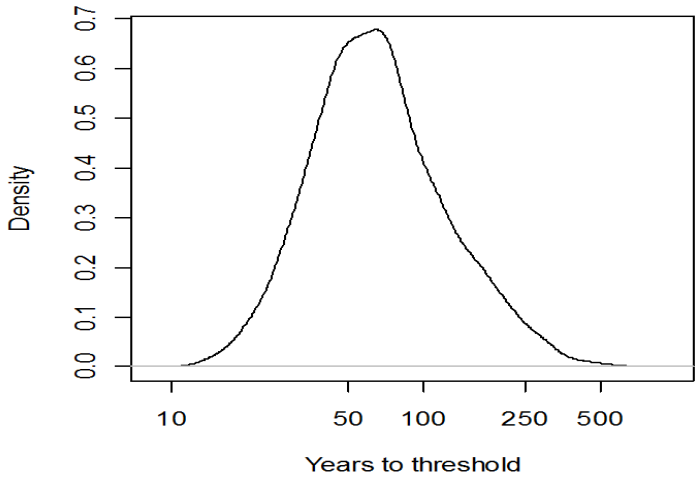

Figure 6: Distribution of predictions for time-to-threshold for one specific seed lot.

Our first choice of method to estimate these prediction distributions was a parametric bootstrap. In this bootstrap, coefficients for a seed lot were simulated from the estimated multivariate normal distribution of coefficients, then observations at the appropriate seed ages were predicted using the coefficients and the estimated error distribution. This was repeated for all seed lots, then the random coefficients model was fit, and the time-to-threshold for each seed lot was predicted using the predicted coefficients. The entire bootstrap was repeated 1000 times. We used a parametric bootstrap, instead of resampling observations (a non-parametric bootstrap), because we needed to maintain the hierarchical structure (observations within seed lots) of the data.

The parametric bootstrap was very slow. It required in excess of 48 hours of computing. When I gave a seminar on this work, Richard Barker, one of the audience members, suggested I use Bayesian methods to estimate the prediction distributions. Given realizations from the joint posterior distribution of the parameters, the distribution of predicted time-to-threshold is trivial to estimate numerically. We used MCMC methods in WinBUGS to estimate the joint posterior distribution. The results passed the common checks for model adequacy: the Gelman-Rubin statistic for multiple chains was very close to 1 and the results were not sensitive to changes in the choice of prior distributions. Results were very similar to those from the parametric bootstrap, but fitting this model required approximately 20 minutes of computing. We are planning to develop a web interface so other seed biologists can use this

method for their data. This is practical when the computing time is circa 20 minutes, but not when it is more than 48 hours.

As promised, I've shared five short stories that I wish I had heard and really understood much earlier in my career. I hope one (or more) will make your (statistical) life a bit easier.

7 References

Bickel, P.J., Hammel, E.A., and O'Connell, J.W. 1975. Sex Bias in Graduate Admissions: Data From Berkeley. Science 187: 398−404

Fai, A.H.T. and Cornelius, P.L. 1996. Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analysis of unbalanced split-plot experiments. Journal of Statistical Computing and Simulation 54:363-378.

Giesbrecht, F.G. and Burns, J.C. 1985. Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. Biometrics 41:477-486.

Hsu LM. 1989. Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. Journal of Consulting and Clinical Psychology. 57:131137

Kenward, M. G. and Roger, J.H. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983-997.

Stümpke, H. 1967. *The Snouters: Form and Life of the Rhinogrades*. Chicago: University of Chicago Press.

Trapp, A. Dixon, P.M., Widrlechner, M.P. and Kovach, D. 2012. Scheduling viability tests for seeds in long-term storage based on a Bayesian multi-level model. Journal of Agricultural, Biological and Ecological Statistics. 17:192−208.