Journal of Data Science **7**(2009), 537-550

Getting the Most from Data — Maximizing Information and Power by Using Appropriate and Modern Statistical Methods

Timothy E. O'Brien and Martin B. Berg
Loyola University Chicago

Abstract: Through a series of carefully chosen illustrations from biometry and biomedicine, this note underscores the importance of using appropriate analytical techniques to increase power in statistical modeling and testing. These examples also serve to highlight some of the important recent developments in applied statistics of use to practitioners.

Key words: Chi-square test, interval-censored data analysis, logistic regression, maximum likelihood estimation, McNemar's test, paired data, nonlinear regression, proportional odds model.

1. Introduction

Scientists who work in applied research may have learned introductory statistical methods in the days when emphasis was placed on fitting data analysis problems into the paradigms of t-tests, ANOVA, linear repeated measures analysis, or linear regression, and have found it challenging to stay current with important subsequent developments — even those over the past 30 or so years. One might then become aware of these newer developments when a reviewer returns a submitted manuscript requesting a more-appropriate analysis. For example, somewhat dated textbooks emphasize modifying non-normal, proportion or nonlinear data using an appropriately chosen transformation; they also lead us to believe that a choice has to be made between parametric tests such as a t-test or ANOVA on the one hand and a nonparametric test on the other.

With an eye to researchers and undergraduate and graduate students, this note underscores some of the common shortcomings observed in current data analysis. We do so using a series of illustrative examples; our list is in no way exhaustive but it is quite typical. Additional key references include Lange *et al.* (1994), Krzanowski (1998), Lindsey (2004), and Good and Hardin (2006). Given the fast pace at which new statistical methods are being developed, our aim is to

also update practitioners and students with some of the recent developments in techniques useful for biometric and biomedical research.

For those familiar with the Minitab and SAS packages, we provide programs and output in the Appendices posted at our website[1].

To aid in exposition here, we reference results given in these Appendices in this paper.

2. Some Important Illustrations

The following five examples have been chosen to demonstrate the advantages of statistical modeling, of examining original data instead of collapsing or pooling them, of incorporating pairing into data analysis, and of choosing the statistical technique that maximizes power and whose assumptions match the analysis.

Example 1 . Misapplying the usual χ^2 test; choosing a good statistical model.

When asked to examine the count data in Table 1 comparing two drugs in terms of pain relief, some researchers might incorrectly find no significant difference between the two treatments using either the generic chi-square test ($p = 0.1422$) or Fisher's Exact test ($p = 0.1618$).

Table 1: Count data relating Pain Relief and Drug factors.

	Pain Relief				Total
Drug	Complete (1)	Substantial (2)	Some (3)	None (4)	
A	5 (20%)	10 (40%)	7 (28%)	3 (12%)	25
B	2 (8%)	5 (20%)	11 (44%)	7 (28%)	25

These two tests are inappropriate in this example because they fail to take advantage of the ordinal nature of the outcome factor (pain relief) and thus lack statistical power. That is, for these data, the usual chi-square test and the Fisher Exact test (FET) are more prone to fail to see a significant difference between the two drugs when such a difference actually exists; stated differently, these two tests have higher probabilities of type II error rates for ordinal data. We may draw some comfort from typical software output (such as in Appendix 1a) in that the generic chi-square test is highlighted as dubious for these data due to the low expected cell count in 25% of the cells (leading novice data analysts to use the FET instead), but unfortunately there is no warning of the lack of power for the FET. Further details regarding these two tests are given in Chapter 2 of Agresti (2007).

[1]See: http://www.math.luc.edu/~tobrien/research/Appendices.pdf

Because the response variable here is indeed ordinal (in that 'complete' exceeds 'substantial' in terms of pain relief and so on), a more powerful and appropriate modelling approach and test is the Mantel-Haenszel (MH) chi-square test for a linear association; see Agresti (2007:41). For these data, the MH test yields a p-value of 0.0389, indicating a significant difference between the two drugs in terms of pain relief. Note that whereas the generic chi-square test is associated with three degrees of freedom, the MH test is associated with just one degree of freedom. The analogy to consider is that the first chi-square test is similar to comparing four treatments and the MH test is just looking at a linear contrast, which often is appropriate when the treatment is a quantitative factor. This highlights the fact that the MH test is more powerful to detect a linear trend in the 'pain relief' factor, but is predicated on the assumption that the trend is linear and not quadratic or cubic. Note that this linear trend is masked when using the generic chi-square test, which in a sense averages the linear, quadratic and cubic effects over the three degrees of freedom.

Another approach to analyzing these data is the use of the proportional odds (PO) model, and is one of the more preferred models for data of this sort, as proposed in categorical data analysis books such as Dobson (2002), Agresti (2002) and Agresti (2007). For these data, the PO model takes the form of the following three-equation expressions:

$$\log\left(\frac{\pi_1}{\pi_2+\pi_3+\pi_4}\right) = \alpha_1+\beta x \tag{2.1}$$

$$\log\left(\frac{\pi_1+\pi_2}{\pi_3+\pi_4}\right) = \alpha_2+\beta x \tag{2.2}$$

$$\log\left(\frac{\pi_1+\pi_2+\pi_3}{\pi_4}\right) = \alpha_3+\beta x \tag{2.3}$$

In these expressions, π_1 corresponds to the probability of a "Complete" pain relief response, π_2 corresponds to the probability of a "Substantial" pain relief response, and so on. The chosen numbering is shown at the top of Table 1 (and in the computer program given in Appendix 1a). Also, 'x' in this expression is an indicator variable that is chosen to equal one for drug A and zero for drug B. As shown in the output given in Appendix 1b, the PO model indicates a significant difference between the two drugs (p-value = 0.0241). The output of the PO model goes one step beyond the above chi-square tests by providing the following important interpretation: since the β parameter in the model is estimated to be 1.2205, the odds ratio is estimated to be $e^{1.2205} = 3.389$. This indicates that as an individual shifts from drug B to drug A, the odds of a response in the more complete pain relief direction increases by a factor of approximately 3.389; further details can be found in Agresti (2007, chap.6). The SAS output given in

Appendix 1b also confirms that this model is a good fit to these data indicating that the proportionality assumption is met ($p = 0.8745$). Had this condition not been met for our data, we would have chosen a richer model from those given in Agresti (2002, 2007).

This example demonstrates the range of choices of models and analyses that can be associated with a set of count data — each predicated upon different underlying assumptions which may or may not fit a specific case. This challenges researchers to think clearly about the appropriate assumptions which best match ones data. The MH chi-square and the PO models are preferred here because the response variable is ordinal, and both these methods suggest a significant difference between the drugs. As indicated, the PO model may well be preferred for these data because (1) it may not be straightforward to verify the linearity requirement for the MH method, (2) the proportionality requirement of the PO model is met for these data, and (3) the PO model is easier to interpret. More importantly, in either of these cases, a significant difference is suggested between these two drugs based on this study.

Although the above example comes from a biostatistical application involving the comparison of two drugs, these methods are equally important for other biological and ecological applications as well. The models used here are appropriate for ordinal count data, and analogous methods to the PO model for nominal count data have been developed as well and are discussed in Agresti (2007). The next example comes from the biological anthropology literature and highlights other important caveats.

Example 2. Collapsing data; testing for goodness of fit in Logistic regression.

Data are reported in Grauer (1993) based on recovered skeletal remains from York, England. These data, given in Table 2a, include the estimated age at death and the percentage of the individuals that showed evidence of porotic hyperostosis (PH), which has been shown to be associated with anemia.

Table 2a: Average age of individuals and percentage exhibiting evidence of porotic hyperostosis at death in recovered skulls in York, England.

Age	0.5	1.5	2.5	3.5	4.5	7.5	12.5	17.5	22.5	30	40	50	60
Percent	8	43	17	33	50	41	59	53	75	66	68	72	59

In Grauer (1993), these count data are subsequently reduced to Table 2b and a significant association is declared between age and presence of PH ($\chi^2_1 = 27.6, p < 0.0001$).

Table 2b: Collapsed data derived from original count data leading to percentages in Table 2a.

	Porotic Hyperostosis		Total
Age	Present (%)	Absent (%)	
0 - 20	79 (43.2%)	104 (56.8%)	183
20 - 65	188 (67.9%)	89 (32.1%)	277

While this conclusion is indeed correct, collapsing the original count data and ignoring the quantitative nature of the 'age' variable yields an important loss of information and statistical power. A more informative approach would be to fit the binary logistic model. This model holds that of the skulls recovered and for the individuals of the given age 'x', the probability (π) that the skull exhibits evidence of PH is given by the equation:

$$\pi = \frac{e^{\theta_3(x-\theta 2)}}{1 + e^{\theta_3(x-\theta_2)}} \tag{2.4}$$

This expression - which entails the estimation of two model parameters (θ_2 and θ_3) — is equivalently written in the following manner found in most textbooks,

$$\log\left(\frac{\pi}{1-\pi}\right) = \theta_3(x - \theta_2) \tag{2.5}$$

A comparison of the PO model in Equations (2.1) – (2.3) and the Logistic model in Equation (2.5) underscores the fact that the latter model is just a special case of the former one where the outcome variable - incidence of PH - has only two values (presence or absence). Unfortunately, the Logistic model does not fit these data as indicated in the following Minitab partial output:

```
Goodness-of-Fit Tests
Method           Chi-Square  DF      P
Pearson             19.2194  11  0.057
Deviance            20.4138  11  0.040
Hosmer-Lemeshow     12.6478   5  0.027
```

Specifically, the hypothesis being tested in each of these "goodness-of-fit" tests is that the logistic model in Equation (2.5) fits the data, and since two of these tests indicate lack-of-fit (p-values below 0.05), the logistic model provides a dubious fit.

It turns out that the problem here is that we should have chosen the log-scale - so we should use log(age) - instead of the original scale; this is indicated in the output given in Appendix 2a. Indeed the log-scale is indicated in this output since

the 95% confidence interval for the 'gamma' scale parameter contains zero (which corresponds to the log-scale) and not one (which corresponds to the original scale). Additional details related to this three-parameter 'scaled logistic model' are given in O'Brien *et al.* (2009a).

In this log-age case (again with x = age), the counterpart of Equation (2.4) is

$$\pi = \frac{(x/\theta_2)^{\theta_3}}{1 + (x/\theta_2)^{\theta_3}} \tag{2.6}$$

In the expression, θ_3 is the slope parameter and θ_2 (loosely called the LD_{50} here) is the parameter with the key interpretation: it is the value of x (age) for which $\pi = 1/2$. For these data, θ_2 is estimated to be 10.97 (and with a 95% confidence interval from 6.3 to 15.6). Therefore, this model predicts that for a (similar) individual who died at greater than approximately 11 years old, the probability that this individual died with PH exceeds 50%. Thus, this example points out that since the interpretation of the LD50 parameter is indeed paramount for these data, the author's decision to collapse the data resulted in the inability to estimate this key parameter.

This and other benefits associated with logistic regression notwithstanding, we hasten to add that the analysis performed here does not take into account the fact that the ages of the individuals at death were estimated for these data. It is very likely that if the actual ages at death were known and used here, then the estimated value of θ_2 would not change appreciably but that the 95% confidence interval would be wider; additional details of so-called *measurement error* — where an independent variable is estimated in addition to the dependent variable - are given and illustrated in Section 3.4 of Draper & Smith (1998) and in Stefanski (2002).

Many introductory texts and courses in applied statistics demonstrate the importance of taking account of pairing (or blocking) when data are so related. For example, Samuels and Witmer (2003: 351) provides an example in which the correct paired t-test shows a significant results but the incorrect two-sample t-test shows no such significant effect. Similarly, the following example demonstrates that the same can be said for paired count data as well.

Example 3. Accounting for the correlations inherent in paired count data.

Kaitz (1992) examines the results of a study measuring recognition by touch of a person familiar to oneself. In this study, 36 male-female couples were assessed in turn as to whether each blindfolded individual could correctly choose the respective partner out of a group of three similar individuals based on touching either (a) the person's face (forehead) or (b) the person's hand. The data presented and analyzed in the paper are given in Tables 3a and 3b below.

Table 3a: Face recognition data as given and examined in Kaitz (1992).

FACE	Correct	Incorrect	Total
Male	18 (50.0%)	18	36
Female	24 (66.7%)	12	36
Total	42	30	72

Table 3b: Hand recognition data as given and examined in Kaitz (1992).

HAND	Correct	Incorrect	Total
Male	16 (44.4%)	20	36
Female	25 (69.4%)	11	36
Total	41	31	72

The goal of this research was to see if there is a statistically significant difference between the two genders in terms of recognition of one's partner, so two chi-square tests were performed. For these data, the results are $\chi^2_1 = 2.057(p = 0.1515)$ for face data and $\chi^2_1 = 4.589(p = 0.0322)$ for hand data. Thus, Kaitz (1992) concludes that although no significant difference is observed for face recognition, the female participants fared significantly better than their male counterparts in terms of hand recognition.

As highlighted above, however, these data are inherently paired (dependent) - with a pair or so-called 'block' corresponding to a couple - and this pairing has been ignored in the above analysis (which assumes independence). Although the full datasets needed to do the paired analysis are unavailable, these data can be envisaged by examining the following two tables.

Table 3c: Desired count data for face recognition for 'a' an integer between 6 and 18.

FACE		Female		Total
		Correct	Incorrect	
Male	Correct	a	$18 - a$	18
	Incorrect	$24 - a$	$a - 6$	18
Total		24	12	36

Table 3d: Desired count data for hand recognition for 'b' an integer between 5 and 16.

HAND		Female		Total
		Correct	Incorrect	
Male	Correct	b	$16-b$	16
	Incorrect	$25-b$	$b-5$	20
Total		25	11	36

In these tables, each entry (count) corresponds to a couple; thus in Table 3c, the 'a' corresponds to the number of couples where both the male and female members of the couple correctly identified the corresponding partner by touching the face. The marginal totals in Tables 3c and 3d correspond to the data given in Tables 3a and 3b and examined above. Although the integers 'a' in Table 3c and 'b' in Table 3d are unknown (i.e., not provided in the original manuscript), constraints on the marginal totals necessitate that 'a' must lie somewhere between 6 and 18 and that 'b' lies somewhere between 5 and 16.

Because the data in Tables 3c and 3d are paired count data, treating and analyzing them as paired data will provide more powerful results than the analysis in Kaitz (1992). Further, as paired count data, they can best be analyzed using McNemar's (chi-square) test. This test compares the off-diagonal counts in Tables 3c and 3d, and is discussed in Agresti (2007:245). Interestingly, depending on the values of 'a' and 'b', the data can yield results either supporting or refuting the original analyses. For example, for the face variable, although non-significance was declared above, McNemar's test gives the test statistic $\chi^2_1 = 18/(21-a)$, and so a **significant** difference is indicated for the values $a = 17(p = 0.0339)$ and for $a = 18(p = 0.0143)$. Similarly, although significance was declared in Kaitz (1992) for the hand data, in this case McNemar's test gives the test statistic

$\chi^2_1 = 81/(41-2b)$ and a **non-significant** difference for $b = 5$ through 9. To illustrate, the results of this test are provided in Appendix 3 for the case $b = 6$.

As is often the case, we also could go one step further in our analysis of these paired count data: because the face and hand measurements for a couple are correlated, we might be wise to perform a multivariate test which takes account of this potential correlation. Indeed, a multivariate extension of McNemar's test is given in Klingenberg & Agresti (2006). Thus, as with the previous examples, this illustration underscores the fact that more appropriate and powerful statistical analysis can provide conclusions that disagree from those given by inappropriate or less-powerful analyses.

Researchers are often reminded to check underlying assumptions before performing any statistical test, and the following example highlights why this is important. It also provides a useful method for testing when certain key as-

sumptions may not be met and when nonparametric methods may be inappropriate or lack sufficient power. As mentioned above, before widespread access to personal computers and sophisticated statistical software packages, practitioners and applied statisticians were encouraged to transform data so as to meet key assumptions and requirements. For example, it was often suggested that count data be transformed using the log-transformation and that proportion data be transformed using the arcsine-square-root-transformation. Nowadays, PCs and software can handle the original data and perform the more-appropriate analysis as is demonstrated here.

Example 4. Analyzing ANOVA data with unequal variances using the 'LR χ^2 test'.

The environmental data used here correspond to effluent yield data for three types of wastewater treatment; these data are summarized in the Minitab output given in Appendix 4a and graphed in Figure 1.

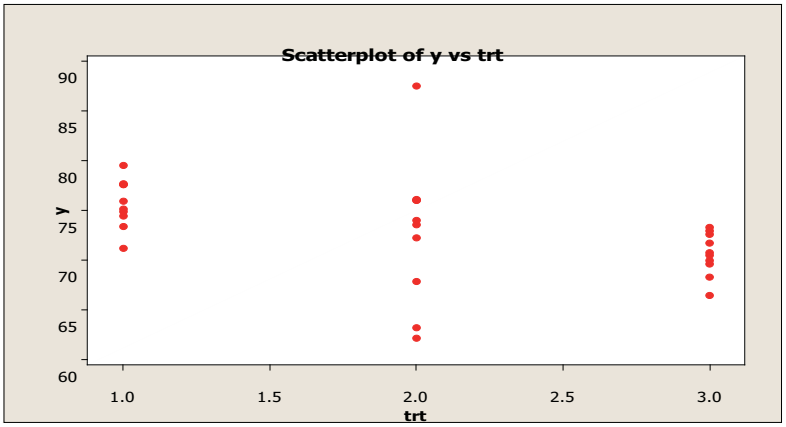

Figure 1: Effluent yield data for three wastewater treatments.

Environmental researchers, interested in deciding whether there is a significant difference in the average effluent yields for the three treatments, might be tempted to use the one-way ANOVA test performed in Appendix 4a. Because the reported p-value of 0.061 exceeds the 5% cut-off, some such researchers would declare no significant difference between these means. Although no warning is given in the Minitab computer output, Figure 1 clearly indicates that the constant-variance requirement is not met for these data.

Understandably, a more appropriate approach for testing is the likelihood ratio (LR) test; preference for likelihood methods over other tests is underscored in Pawitan (2001) and Agresti (2002, 2007). Because the normality assumption

appears met for these data (Figure 1), this LR χ^2 test is performed in Appendix 4b, and yields a strongly significant result ($\chi^2_2 = 16.7, p = 0.0002$). Briefly, this test is based on the logarithm of the likelihood under the null hypothesis over the likelihood under the alternative hypothesis and has a χ^2 distribution (Agresti, 2007, p.89). Thus, this more appropriate analysis here indicates that the average effluent yields are not the same, and the next step would be to determine which treatments differ using the likelihood analogue of a multiple comparison test. As for the original ANOVA analysis, this conclusion is based on the assumption of normality although here allowing for possibly unequal variances.

The LR program given in Appendix 4b can be easily adapted to distributions other than the normal distribution. For example, if the original data were count data, we might want to perform a LR test of means assuming a Poisson distribution. In this manner, we see that the heavy reliance on transformations performed in the past is largely unnecessary, especially since one is never sure of which transformation to use to render the data to be 'nearly normal'; an additional example illustrating the shortcomings of transformations is given in Samuels & Witmer (2003, p.289).

For these data, we could have done a Kruskal-Wallis nonparametric test, but in general nonparametric tests may lack power. Also, the term 'nonparametric' is really a misnomer since power calculations for nonparametric tests such as the Sign test, the Wilcoxon Signed-Rank test, and so on, are actually based upon the assumptions of symmetry, normality and constant variances. In addition, note that randomization methods would fail for analyzing these data since these methods inherently assume equal variances and this is clearly not the case.

Example 5. Modifying Logistic Regression for Interval-Censored Bioassay Data.

This illustration was originally motivated from work with a genetics graduate student interested in comparing several treatments in terms of survival profiles and is modified from the original project here. Similar examples can be found from the field of ecology in Muenchow (1986) and from the field of biomedicine in Collett (2003b).

We are interested here in comparing two treatments in terms of prolonging cancer survival rates. Twenty cancer patients are randomized to each treatment, and the number surviving at weeks 2, 4, 6 and 8 are recorded. These data and survival percentages are given in the following table. In addition, the data and fitted logistic curves are plotted in Figure 2, and the goal here is to quantify and test the degree of rightward shift in the treatment B curve relative to the treatment A one. This is very similar to the assessment of relative potency as discussed in Finney (1978) and Collett (2003a).

Table 4: Survival counts and probabilities of original 20 for each treatment at weeks 2, 4, 6, and 8.

	Number (and Percent) of Original 20 Surviving at Week			
Treatment	2	4	6	8
A	16 (80%)	10 (50%)	6 (30%)	2 (10%)
B	18 (90%)	12 (60%)	10 (50%)	6 (30%)

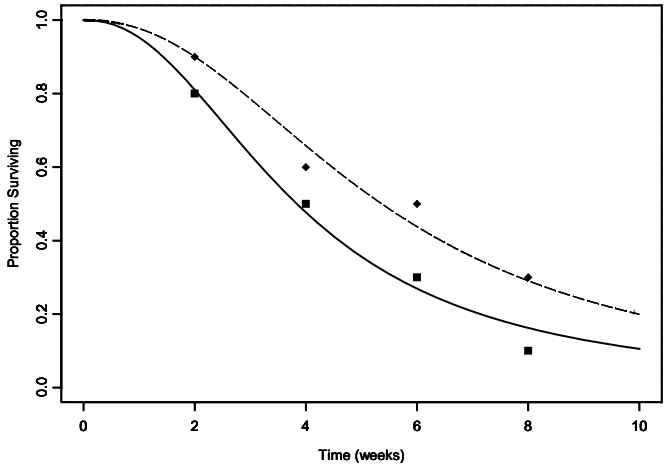

Figure 2. Plot of survival probabilities over time (in weeks) for treatment A (solid curve and squares) and treatment B (dashed curve and diamonds).

An incorrect analysis of these data would be to fit a logistic regression curve (of Example 2 above) to each of the two treatments since a requirement of logistic regression is that independent subjects be observed at each time point. Instead, for these data one observes the cohorts of twenty patients over time, and the measurements are therefore correlated. Nonetheless, we provide this incorrect analysis here to underscore the importance of this independent-subjects requirement and because this was the initial means of analysis performed by the graduate student; the student was in turn following the approach used in Legal *et al.* (1994).

The logistic curve in Equation (2.6) is incorrectly fit to each of the two treatments in Appendix 5a, and this output demonstrates - using the LR χ^2 test discussed in the previous example — that (from the results of the first and second runs of the NLMIXED procedure) parallelism of the curves ($\theta_{3A} = \theta_{3B}$) is

retained ($\chi^2_1 = 26.4 - 26.3 = 0.1, p = 0.7518$). Using the third NLMIXED procedure, the relative potency parameter, $\rho = \theta_{2B}/\theta_{2A}$, is estimated to be 1.4284, and the LR χ^2 test of equal potency of the two treatments ($H_0 : \rho = 1$) is rejected ($\chi^2_1 = 31.1 - 26.4 = 4.7, p = 0.0302$). As a side note, we point out that the 95% so-called Wald confidence interval for the relative potency parameter (ρ) provided in the third NLMIXED run, (0.8710 , 1.9857), does contain one thereby implying equal potency of these treatments. This apparent contradiction highlights the superiority of likelihood methods over Wald methods. To sum up: this incorrect logistic analysis implies a significant increase in survival rates when using treatment B because the hypothesis of equal potencies is rejected.

As indicated above, the logistic analysis in this example is erroneous because the same twenty individuals were observed over time. In truth, what is known is that four patients in the treatment A group died between time-points 0 and 2 weeks, six treatment A patients died between time-points 2 and 4 weeks, and so on. Thus, a more appropriate approach is to consider *interval-censored analysis methods*. The interval-censored methodology given in Chapter 9 of Collett (2003b) has been modified in Appendix 5b to use the log-logistic expression from Equation (2.6). These data again exhibit parallelism of the curves ($\chi^2_1 = 126.7 - 126.5 = 0.2, p = 0.6547$), but the test of equal potency is retained here ($\chi^2_1 = 128.4 - 126.7 = 1.7, p = 0.1923$). Our final conclusion regarding these data is that no significant difference is detected between the two treatments in terms of prolonging survival times in cancer patients — i.e., the opposite conclusion as that reached using the naïve logistic method.

In the above example, although we used the parametric approach of assuming the logistic curve fits these interval-censored data, we can draw comfort from the fact that the same conclusion is reached ($\chi^2_1 = 2.5, p = 0.1138$) using the non-parametric *proportional hazards* approach as discussed in Collett (2003b:282).

3. Discussion

In introductory statistics textbooks and courses, we are reminded to check the underlying assumptions before performing a statistical test, and the above examples illustrate what can go wrong when inappropriate techniques are used to analyze our data. Most notably, we sometimes reach the wrong conclusion or cannot answer our important research questions. Equally important is the issue of choosing an informative and useful experimental design; these design concerns are addressed for example in Kuehl (2000) for general problems and in O'Brien *et al.* (2009b) for dose-response bioassay models. Experiments are often too expensive and time consuming to squander by fitting dubious or incorrect statistical analyses.

It is often disconcerting to the practitioner and student that several statis-

tical analyses may be used to analyze the same dataset — depending on the assumptions that are made. This is demonstrated in Examples 1 and 5 above, and thankfully in these cases they produce the same conclusion. When they do not, one needs to better examine (and perhaps test) the underlying assumptions or conditions until we are convinced which analysis to use. Indeed, Box (1979) reminds us that "all models are wrong some models are useful", and this encourages us to be mindful of our need to test and validate our models and methods. Thus, the above illustrations underscore that applied statistical analysis is a field that is in a continuous state of change and improvement — due largely to better understanding, increased computing power and better statistical packages. We look forward to the subsequent developments yet to come.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition. Wiley.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd Edition. Wiley.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Robustness in Statistics (Edited by Launer, R.L. and Wilkinson, G.N.). Academic Press.

Collett, D. (2003a). *Modelling Binary Data*, 2nd Edition. Chapman and Hall.

Collett, D. (2003b). *Modelling Survival Data in Medical Research*, 2nd Edition. Chapman and Hall.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd Edition. Wiley.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd Edition. Chapman and Hall.

Finney, D. J. (1978). *Statistical Method in Biological Assay*, 3rd Edition. Charles Griffin.

Good, P. I. and Hardin, J. W. (2006). *Common Errors in Statistics (and How to Avoid Them)*, 2nd Edition. Wiley.

Grauer, A.L. (1993). Patterns of anemia and infection from medieval York, England, *Amer. J. Phys. Anthropology* **91**, 203-213.

Kaitz, M. (1992). Recognition of familiar individuals by touch. *Physiology and Behavior* **52**, 565-567.

Klingenberg, B. and Agresti, A. (2006). Multivariate extensions of McNemar's test. *Biometrics* **62**, 921-928.

Krzanowski, W. (1998). *An Introduction to Statistical Modelling*. Arnold.

Kuehl, R. O. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*, 2nd Edition. Duxbury.

Lange, N., Ryan, S., Billard, L., Brillinger, D., Conquest, L. and Greenhouse, J. (1994). *Case Studies in Biometry*. Wiley.

Legal, L., Chappe, B. and Jallon, J. M. (1994). Molecular basis of Morinda citrifolia (L.): Toxicity on Drosophila. *J. Chem. Ecology* **20**, 1931-1943.

Lindsey, J. K. (2001). *Nonlinear Models in Medical Statistics*. Oxford University Press.

Lindsey, J. K. (2004). *Introduction to Applied Statistics: A Modelling Approach*, 2nd Edition. Oxford University Press.

Muenchow, G. (1986). Ecological use of failure time analysis. *Ecology* **67**, 246-250.

O'Brien, T. E., Chooprateep, S. and Funk, G. (2009a). Encouraging students to think critically: Regression modelling and goodness-of-fit. *Journal of Data Science* **7**, 235-253.

O'Brien, T. E., Chooprateep, S. and Homkham, N. (2009b). Efficient geometric and uniform design strategies for Sigmoidal models. *South African Stat. J.* **43**, 49-83

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Clarendon Press.

Ratkowsky, D. A. (1983). *Nonlinear Regression Modeling: A Unified Practical Approach*. Dekker.

Ratkowsky, D. A. (1990). *Handbook of Nonlinear Regression Models*. Dekker.

Samuels, M. L. and Witmer, J. A. (2003). *Statistics for the Life Sciences*, 3rd Edition. Prentice Hall.

Stefanski, L. A. (2002). Measurement error models. In *Statistics in the 21st Century* (Edited by Raftery, A. E., Tanner, M.A. and Wells, M.T.). Chapman and Hall.

Received January 31, 2007; accepted March 7, 2008.

Timothy E. O'Brien
Department of Mathematics and Statistics
Loyola University Chicago
6525 N. Sheridan Road
Chicago, IL 60626 USA
TOBRIEL@LUC.EDU

Martin B. Berg
Department of Biology
Loyola University Chicago
6525 N. Sheridan Road
Chicago, IL 60626 USA
MBERG@LUC.EDU