Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 8 Statistical Inference with the General Linear Model

Elementary statistics courses for biologists tend to lead to the use of a stereotyped set of tests:
1 without critical attention to the underlying model involved;
2 without due regard to the precise distribution of sampling errors;
3 with little concern for the scale of measurement;
4 careless of dimensional homogeneity;
5 without considering the ideal transformation;
6 without any attempt at model simplification;
7 with too much emphasis on hypothesis testing and too little emphasis on parameter estimation.
M.J. Crawley. 1993. GLIM for Ecologists. (London, Blackwell)

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5, 6, 7)
ReCap Part III
8.1 Introduction
8.2 Component concepts
8.3 Generic Recipe

on chalk board

Experimentation with order of presentation.

1994	Components concepts Lec13 ANOVA example Lec14
1995	Component concepts Lec13 regression example Lec14.
1996	Component concepts Lec13 (in 20 minutes) then regression example Lec14
1997	Mon: Concepts L13 + ex L14 Wed: revisit L13 + ex L15 Went well.
1998	Same as 1997. Lec 13 in 15 minutes. Went well
2000	General material Lec 13 in 5 minutes. Components of GLM in 15 min Then to Lec 14. Went well
2002	Lec 13 General Intro and components in 10 minutes, then Lec14
2018	Add likelihood

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of fly heterozygosity, which combined models (what is the relation of fly heterozygosity to altitude?) with statistics (how certain can we be?)

ReCap Part II (Chapters 5,6,7)
Data equations summarize pattern in data as a series of parameters (means, slopes).
Frequency distributions, a key concept in statistics, are used to quantify uncertainty.
Hypothesis testing uses the logic of the null hypothesis to make a decision about an unknown population parameter.
Estimation is concerned with the specific value of an unknown population parameter.

We have now concluded the first third of course.
We move on to the second third, the General Linear Model

Today: Introduction to the General Linear Model
Begin with brief introduction to component concepts in a generic recipe.
Then work through an example, using the generic recipe.

Wrap-up
The general linear model has many advantages over learning a series of tests.
It lends itself naturally to a problem solving approach based on biological concepts.
We have already covered most of the component concepts.We will use a generic recipe that once learned, permits us to undertake a wide variety of analyses.

8.1 Introduction

Statistics are routinely presented as a collection of recipes in courses for scientists. The basic ingredients are null and alternative hypotheses, a statistic (F, t, or chisquare), a p-value, and the declaration of a decision. The recipes focus on the logic of the null hypothesis rather than on the biological relevance of the model. The recipes focus on p-values rather than the interpretation of parameters or the degree of uncertainty associated with a parameter estimate. The recipes often ignore diagnosis of assumptions or evaluation of the sample relative to the population. The recipe collection is huge. Widely used texts typically cover the following tests: one-sample hypotheses, two sample hypotheses, paired sample hypotheses, one-way ANOVA, multiple comparisons, two-way ANOVA, hierarchical ANOVA, multiway ANOVA, regression, multiple regression, analysis of covariance (ANCOVA), polynomial regression, logistic regression, goodness of fit tests, and contingency tests. The menus of widely used statistical packages (Minitab, SPSS, SAS, Systat) contain even longer lists of tests. Choosing from such a long list is daunting, and as it turns out, unnecessary.

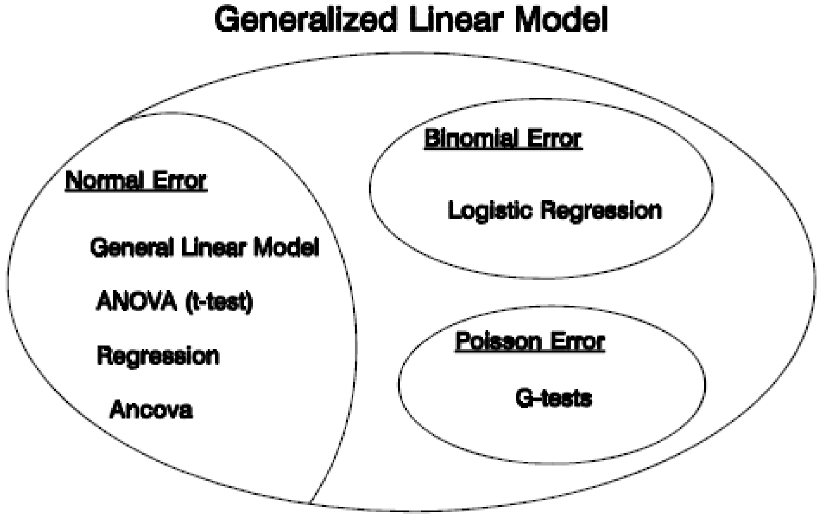


Figure 8.1. Named statistical tests that are special cases of the general linear model and the generalized linear model.

Analysis of data in science will usually entail some form of functional relation: How does some quantity *Q* vary as a function of another set of quantities *X1, X2... etc* ? For these problems we can employ ***model-based statistics***, which focus on a response variable in relation to one or more explanatory variables. We will use the generalized linear model (Nelder and Wedderburn 1972, McCullagh and Nelder 1989), one of the major developments in statistics in the last quarter of the 20th century. It allows analysis based on any of several error distributions. We'll begin with the general linear model, which assumes a normal error (Figure 8.1). The general linear model (GLM) has been available in the SAS software package since at least 1980, and is now available in any reputable stat package. The generalized linear model (GzLM) has been available in SAS since the first decade of this century, and is now widely available in code based (SAS, R) as well as menu based (SPSS) software. This development of software allows the generalized linear model to be presented in introductory courses in statistics at the undergraduate level.

Advantages and disadvantages of learning model based statistics.
The GLM is a way of thinking in quantitative terms, using a simple model structure that relate one quantity to another.

The GLM has many advantages

Unifying concepts instead of lots of special cases.
We evaluate the relation of one test to another
Special procedures are no longer needed for each named test.
For example, ANCOVA no longer needs to be learned as separate tests:
-one that uses a covariate (on a ratio scale) to control for another variable
-one that compares two regression lines

More useful. Far more can be accomplished with this general approach than by using a series of named tests. Our science questions and data sets do not always fit easily into a list of named tests. Each test has its own list of assumptions. But in fact many assumptions are common to a wide variety of procedures. It is more effective to learn a general procedure to evaluate assumptions.

Easier to learn in the long-run. Learning to use a generic recipe is not like learning to apply an equation. It is not like learning a script for making calculations. It is not like memorizing basic terms in a particular branch of science. For most students it is a new way of thinking and it takes work. But in the end a generic recipe is less work to learn than a whole collection of test procedures.

Disadvantages
Software in the past was formidably difficult to use. One of the first packages, GLIM package was far from "user-friendly." This changed when general and generalized linear model routines became available in packages such SAS, Minitab, and SPSS. The appearance of R steepened the learning curve. The appearance of Rstudio reduced the curve. For a fee you can rent a copy of SPlus, a graphics interface for R. For free (with no guarantees) you can use Rstudio. There is no free lunch.

Summary of advantages. First, students learn unifying concepts rather than a sequence of apparently unrelated procedures. Students can see the relation of one test to another, rather than having to learn special procedures for each topic. For example, ANCOVA can be presented as two applications of the same model, rather than as two separate procedures, one for comparing slopes and one for statistical control of a regression variable. Remedies for recurring problems (*e.g.*, heterogeneous variances) are presented once, rather than several times in different guises. The approach means that remedies can be learned, instead of memorizing specific remedies for each test. The mechanics of analysis are presented once, rather than different procedures for each test. Students are able to accomplish more with the general linear model than by learning statistics as a set of named procedures. For example, with this approach students can set up and execute the analysis of a response variable in relation to two categorical and a single regression variable. There is no name for this analysis, and hence it is outside any list of tests. This greater flexibility leads to better quantitative work in science. The GLM is a way of thinking in quantitative terms, using formal models that relate a quantity of interest to one or more explanatory variables.

The material in the next two weeks will use the same generic recipe, applying it to special cases such as regression, ANOVA, t-tests.
The GLM will become familiar through repetition.

From over 20 years experience (B4605/B7220 at MUN) 3rd and 4th year undergraduate biology majors readily grasped and executed the model-based approach. The presentation below begins by assembling the components learned so far – quantities, data equations, computing the fit of the model to the data, and computing the improvement in fit due to an explanatory variable, either categorical or regression. After a discussion of assumptions for computing p-values in an ANOVA table (which tracks the improvement in fit due to explanatory variables), the presentation moves to a generic recipe that will be applied first to regression (Chapter 9), then ANOVA (Chapters 10, 11), and then ANCOVA (Chapter 14).

Another look at 8.1

Biologists agree that the list of current bird species is finite and rapidly approaching completion. Do you think that a list of statistical tests is finite or could ever be complete ? Why or why not ?

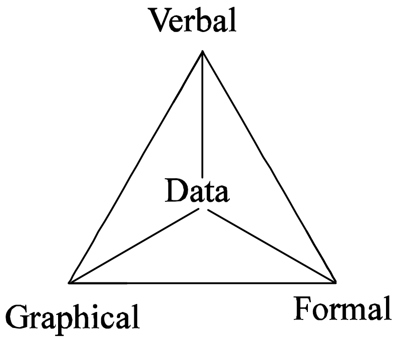
> Elementary statistics courses for biologists tend to lead to the use of a stereotyped set of tests:
> 1 without critical attention to the underlying model involved;
> 2 without due regard to the precise distribution of sampling errors;
> 3 with little concern for the scale of measurement;
> 4 careless of dimensional homogeneity;
> 5 without considering the ideal transformation;
> 6 without any attempt at model simplification;
> 7 with too much emphasis on hypothesis testing and too little emphasis on parameter estimation.
>
> M.J. Crawley. 1993. GLIM for Ecologists. (London, Blackwell)

8.2 Component Concepts

The General Linear Model is a sophisticated concept that substantially improves the quality of statistical analysis by students in science. We will be using component concepts already been covered in previous chapters.

Model based Statistics

The data at the center are summarized at the three apices of the triangle. The GLM summarizes data as a formal model. We begin with verbal model, often in the form of a question. We can use graphical display to express our model and as aid in constructing the formal model. We use the formal model (GLM) to undertake the statistical analysis, which leads to a measure of evidence and measures of uncertainty.



Quantity

Recall that a well-defined quantity has 5 parts:

procedural statement, name, symbol, values, and units.

The GLM relates one variable quantity (the response variable) to one or more other quantities (explanatory variables).

Example:	Plant growth	is	a function of nutrients and sunlight
	Growth Rate	=	f (nutrients, sunlight)
	G	=	f (N, PAR)

Each symbol stands for a variable, which can take on several values via direct measurement, via calculation from direct measurements, or (in the case of an explanatory variable) via categories values fixed by the scientific question. For example with an experiment we are interested in the evidence for differences between treatments and controls.

Variance of a quantity

Recall that the variance is one of several measures of dispersion in a quantity Q. the variance is the mean squared deviation from the average value of the quantity. The true value of the variance of a quantity is often unknown, so an estimate is made. The estimate is:

$$\mathrm{Var}(Q) = (n-1)^{-1}\,\Sigma(Q-\mathrm{mean}(Q)\,)^2$$

In this formula, you will recognize the sum of the squared deviations

$$SS = \Sigma(Q-\mathrm{mean}(Q)\,)^2$$

This is the fit to the simplest of all models, mean(Q)

Model components and data equations

The general linear model has three components: a ***response variable*** Y, a ***structural model*** consisting of one or more explanatory variables X1, X2, etc., and an ***error term***. Table 8.1 shows equivalent expressions. Each term in the model (response variable, explanatory variables, error) represents a vertical string (vector) of numbers. Consequently the symbolic expressions in Table 8.1 represents a series of ***data equations*** (see Chapter 5).

Table 8.1 Equivalent expressions of the general linear model.

Data	=	Model	+	residual
Observed	=	Expected	+	residual
Response	=	f(explanatory variables)	+	residual
Y	=	f(X1 + X2 +...)	+	error
Y	=	$\Sigma\ \beta i$ Xi	+	ε

The explanatory variables can be on a nominal type of scale (ANOVA), on a ratio type of scale (regression), or both (ANCOVA). The residuals are distributed normally.

Data Equations – The simplest model is the mean.

A data equation is written for each value of the response variable. The model in Table 8.2 is the simplest possible: the variable of interest *M* is equal to parameter β_o the true value of the average for the population. Of course, we do not know the true value of β_o except in the case of a full census of a population. Our best estimate of β_o is the mean $\hat{\beta}_o$ = 59 g, computed from the data we have.

Table 8.2 Data equations for measurement of the mass of 3 juvenile cod *Gadus morhua*.

Data	=	Fitted values	+	Residual
M	=	mean(*M*)	+	ε
55 g	=	59 g	+	–4 g
60 g	=	59 g	+	+1 g
62 g	=	59 g	+	+3 g
		H_o		

Data Equations –Comparison of models. Null and Alternative model

As scientists we are often interested in comparing models. For example, does juvenile cod biomass differ in vegetated and unvegetated habitats? Chapter 5 showed another example - does genetic variability decrease with altitude? The research hypothesis—decrease in heterozygosity—is called the “alternative” model H_A. It is compared to a “null” model, which summarizes the information we have based only on the mean only. Equivalently, the null model is that there is no change in heterozygosity with change in altitude. Based on the science that motivated the research, the alternative model was that heterozygosity *H* decreases as a function of elevation *E.*

$$\begin{array}{lclcll}
H & = & \mathrm{E}(H) & + & \varepsilon & < \text{- -population} \\
H & = & \hat{H} & + & \mathit{residual} & < \text{- - sample} \\
H & = & \mathit{Intercept} + \hat{\beta}_E \cdot E & + & \mathit{residual} & \\
H & = & 0.63 \; –0.1298 \cdot E & + & \mathit{residual} &
\end{array}$$

When we ran the numbers, the alternative hypothesis was 450 times more likely than the null hypothesis.

Chapter 5 showed another example, oat yield in two groups X (treated versus untreated).

$$\mathrm{Y} = \beta_0 + \beta_x \cdot \mathrm{X} + \varepsilon$$

Before continuing, apply what you learned in the Equations Lab.
State each symbol in words with units. For example: Y = oat yield in bushels/acres.
Then state the equation above in words.

The explanatory variable X is on nominal scale. It consists of categories.
The null model is no difference between the means for treatments with and without a chemical seed treatment Panogen.
Write the three-term null model ___________________________________
The alternative model was a difference between the means. In Chapter 5 we found that the alternative model ($\beta_x \neq 0$) was only 7 times more likely than the null model ($\beta_x = 0$).

Estimates of parameters

We have already encountered the concept of estimation.
It can refer to informal estimates, such as an order magnitude estimate of mass of an elephant 10^3 kg ? 10^4 kg?
It can refer to an estimate from a formula, such as calculating a mean, a variance, or a standard deviation.
It can refer to an iterative procedure, such as the maximum likelihood estimate we saw in Lab 3.
The most commonly estimated parameters are means, slopes, and odds ratios.
In statistical analysis with the GLM, we will be using the formal machinery of estimation (least squares) to calculate the "best" estimates of means and slopes, according to a least squares criterion. For non-normal error structures (GzLM) we let the statistical package do the work of making iteratively reweighted least squares or maximum likelihood estimates of parameters.

Evaluation of residuals. We have seen that to use a statistical distribution (*t, F*, chisquare, normal) to calculate Type I error (the p-value) we need to make assumptions. We will rely on graphical displays to evaluate the assumptions (Chatfield 1998; Gelman, Pasarica & Dodhia 2002). The statistical literature warns against statistical tests to evaluate assumptions and advocates graphical tools (Montgomery & Peck 1992; Draper & Smith 1998, Quinn & Keough 2002). La¨a¨ ra¨ (2009) gives several reasons for not applying preliminary tests for normality, including: most statistical techniques based on normal errors are robust against violation; for larger data sets the central limit theory implies approximate normality; for small samples the power of the tests is low; and for larger data sets the tests are sensitive to small deviations. In particular we will not adopt the mistaken practice of checking the response variable for normality. Instead we will obtain residuals to evaluate assumptions (e.g. Zuur *et al*. 2010).

Units, Dimensions, and Model Interpretation

Units and dimensions are typically not considered in the statistical analysis of data. They should be. The parameters (means and slopes) that result from statistical analyses are quantities with units and dimensions . They are not simply numbers, which is how they are almost always reported. A glance at the set of the three data equations for cod weights (Table 8.2) will reveal that the mean has the same units and dimensions as the response variable, which appears on the left side of the equality sign. In a regression equation ($Y = \alpha + \beta_x X +$ residual) the intercept α must have the same units and dimensions as the response variable Y. The residual term must also have the same units and dimensions as the response variable Y. The regression coefficient β_x will have the same units and dimensions as the ratio Y/X, in order for the equation to be dimensionally consistent. In the heterozygosity example (Box 8.1), the slope β_x quantifies the altitudinal gradient in genetic variability in units of %/km.

There are several reasons why GLM parameters should be recognized as scaled quantities, rather than treated as simply numbers. First, the rules for operations on scaled quantities, which differ from those for numbers, apply to parameters. Two means can be added only if they have the same units. The rules for rigid and elastic rescaling apply to parameters, a fact that is not evident if parameters are treated as mere numbers. Erroneous calculations result if a parameter is treated as a number when it isn't. A regression coefficient that is an estimate of a spatial gradient at a scale of 100 m cannot be used to calculate a gradient at another scale, unless that coefficient is rescaled according to its units and dimensions.

From likelihood ratios to hypothesis testing

Once we have a null and alternative model consistent with the data we can compare them as a likelihood ratio. Statistical practice in some areas of science is moving toward reports of a likelihood ratio with an effect size, such as -12.7%/km heterozygosity gradient with likelihood 450 times that of no gradient. Likelihood inference (Edwards 1972, Royall 1977) is common in genetics and some areas of ecology (Burnham and Anderson 1998). In some areas of science a measure of uncertainty is required. An example would be high costs of Type I error compared to Type II error. Another example is the use of preset rates of Type I and Type II error to design experiments. In cases where a population can be defined we can make an estimate of uncertainty either as a single number (p-values) or as a confidence interval. The evidential support measured by a likelihood ratio is used to calculate these measures of uncertainty. P-values are not measures of evidence. So we will follow recommendations (ASA 2019) against using p-values as measures of evidence.

Hypotheses tests use the likelihood ratio to calculate a p-value that is routinely used in the Neyman-Pearson sense (decision for are against the null) and rarely if ever in the Fisher sense of a flexible guide for discarding the null. So when should we use hypothesis testing? There are two reasons. The first is when we need to consider Type II as well as Type I error, and thus, the balance between the two in designing an experiment. The second reason is prevailing practice ***if*** we can justify the use of hypothesis testing. The justification is whether we can define a population to which we are inferring (Fisher) or whether we can define chance from a repeatable measurement procedure (Hacking 1965). In experimental work we have a repeatable measurements from a defined protocol repeatable by others. In observational studies where the number of uncontrolled variables is large we may well choose likelihood inference rather than try to defend our measurement protocol as repeatable or as a sample from a population of infinite repeats of the study. In this course we will make that choice in each example.

Hypothesis testing

In this course we will be using a generic recipe for GLM based statistical analysis. This recipe incorporates the generic recipe for hypothesis testing. Where a p-value is warranted the test statistic will be the F-statistic, the ratio of two variances. These variances will be obtained by partitioning the response variable *Q* into two components, that due to the model, and that remaining (the residual)

$$\text{Data} \quad = \quad \text{Model} \quad + \quad \text{Residual}$$
$$\text{Var(data)} \quad = \quad \text{Var(Model)} + \quad \text{Var(Residual)}$$

$$F = \text{var(model)} / \text{var(residual)}$$

We will use the F distribution to calculate the long run probability of any value of the F-ratio, calculated from the degrees of freedom in the numerator and denominator models.

Confidence Limits

Hypothesis testing against a fixed value of 5% has become the prevalent mode of statistical analysis in many areas of natural and social science. We will avoid this practice unless warranted. One reason (among many) is that the null hypothesis can be irrelevant. For example, in examining the relation of metabolic rate to body size, the null hypothesis is biologically irrelevant. We are more interested in excluding a naïve 1:1 scaling to body size (Swift 1726) or perhaps a surface to volume ratio scaling (Sarrus and Rameaux 1838).

Confidence limits, like p-values, are based on a preset probability range. Like p-values they rely on long-run probabilities. However, they are more informative than p-values. They allow exclusion of multiple hypotheses, not just the null hypothesis. Bayes (1764) was the first to propose using set probability limits as a measure of uncertainty (Bayes' Rule). Bayes used a binomial distribution, the only distribution known at the time. Probability limits based on binomial limits were re-invented 200 years later (Clopper and Pearson 1934).

8.3 A Generic Recipe for Applying the General Linear Model

The general linear model is not part of the traditional undergraduate curriculum for biology students. However, it is readily grasped by third and fourth year undergraduates in biology when presented as a analyutic procedure rather than a set of formulas to memorize. Students with limited backgrounds in mathematics and statistics can successfully apply the following generic recipe (Table 8.3) to novel data sets and to their own data.

Table 8.3 Generic Recipe for Statistical Inference with the General Linear Model.

1. Construct model. Begin with verbal and graphical model.
   Distinguish response from explanatory variables
   Assign symbols, state units and type of measurement scale for each.
   Write out statistical model.
2. Execute model   Place data in model format, code model statement.
   Compute fitted values from parameter estimates.
   Compute residuals and plot against fitted values.
3. Evaluate the model, using residuals.
   If straight line inappropriate, revise the model (back to step 1).
   If errors not homogeneous, consider using generalized linear model (step 1)
   If n small, evaluate assumptions for using chisquare, t, or F distribution.
   residuals homogeneous ? (residual versus fit plot)
   residuals independent ? (plot residuals versus residuals at lag 1)
   residuals normal ? (histogram of residuals, quantile or normal score plot)
   If not met, empirical distribution (by randomization) may be necessary
4. Partition df and SS according to model. Calculate likelihood ratio for omnibus model.
   State the full (null) and reduced (alternative) model pair
5. State population and whether the sample is representative
   Decide on mode of inference. Is hypothesis testing appropriate?
   If yes step 6, otherwise, calculate and report the likelihood ratio with parameter estimates.
6. State test statistic, its distribution (*t* or *F*), and tolerance of Type I error.
7. ANOVA: Table Source, SS, df, MS, F-ratio.
   Obtain Type I error (p-value) from distribution (*F* or t).
8. Recompute p-value if necessary.
   If assumptions not met compute better p-value by randomization if:
   sample small ($n < 30$) and if p near α.
9. Declare decision about model terms: If $p < \alpha$ then reject H_o in favor of H_A
   If $p \geq \alpha$ then can't reject H_o
   Report conclusion with evidence: Either the ANOVA table or
   F-ratio (df1,df2) or t-statistics (df) and p-value (not α) for terms of interest.
10. Report and interpret parameters of biological interest (means, slopes)
    along with one measure of uncertainty (st. error, st. dev., or conf. intervals).
    Use appropriate distribution (step 8) to compute confidence limits.

The next chapters work through the generic recipe step by step for commonly used analyses in biology

Exercises

1. List of key concepts for review and future reference.

____model-based statistics
____response variable
____data equations
____true (population) value
____null model
____goodness of fit
____degrees of freedom
____hypothesis testing
____randomized p-value
____ANCOVA
____general linear model
____structural model
____expected value
____estimate
____alternative model
____analysis of variance
____p-value of a variance ratio
____assumptions for p-values
____generalized linear model
____link functions

2. References

Bayes, T. 1764,

Chatfield 1998

C. J. Clopper, E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, 26, 404-413.

Draper and Smith

Eisenhart, C. 1947. Biometrics 3:1-21

Gelman, Pasarica & Dodhia 2002.

Hacking, I. 1965. Logic of Statistical Inference.

Läärä, E. 2009. Statistics: reasoning on uncertainty, and the insignificance of testing null. — Ann. Zool. Fennici 46: 138–157.

Montgomery and Peck

Neter, JW, MH Wasserman, MH Kutner.1983. Applied linear regression models.
Homewood Illinois, Richard D. Irwin, Inc.

Quinn, G and MJ Keough. 2002. Experimental Design and Data Analysis for Biologists. p 110, 280

Sarrus, F and J Rameaux, (1838). Rapport sur une mémoire adressé á l'Académic royale de Médecine. *Bull Acad R Med*, Paris 3 :1094–1100.

Seber, GAF. 1966. The Linear Hypothesis: A General Theory. London, Griffin.

Swift, J. 1726. Gulliver's Travels. London, Benjamin Motte

Tacha, TC, WD Warde, KP Burnham. 1982. Use and interpretation of statistics in wildlife journals. Wildlife Society Bulletin 10: 355-362.

Zuur et al 2010.