

# *Random denominators and the analysis of ratio data*

MARTIN LIERMANN,<sup>1\*</sup> ASHLEY STEEL,<sup>1</sup>  
MICHAEL ROSING<sup>2</sup> and PETER GUTTORP<sup>3</sup>

<sup>1</sup>*Watershed Program, NW Fisheries Science Center, 2725 Montlake Blvd. East, Seattle, WA 98112*  
*E-mail: martin.liermann@noaa.gov*

<sup>2</sup>*Greenland Institute of Natural Resources*

<sup>3</sup>*National Research Center for Statistics and the Environment, Box 354323, University of Washington, Seattle, WA 98195*

---

Ratio data, observations in which one random value is divided by another random value, present unique analytical challenges. The best statistical technique varies depending on the unit on which the inference is based. We present three environmental case studies where ratios are used to compare two groups, and we provide three parametric models from which to simulate ratio data. The models describe situations in which (1) the numerator variance and mean are proportional to the denominator, (2) the numerator mean is proportional to the denominator but its variance is proportional to a quadratic function of the denominator and (3) the numerator and denominator are independent. We compared standard approaches for drawing inference about differences between two distributions of ratios: *t*-tests, *t*-tests with transformations, permutation tests, the Wilcoxon rank test, and ANCOVA-based tests. Comparisons between tests were based both on achieving the specified alpha-level and on statistical power. The tests performed comparably with a few notable exceptions. We developed simple guidelines for choosing a test based on the unit of inference and relationship between the numerator and denominator.

**Keywords:** ANCOVA, catch per-unit effort, fish density, indices, per-capita, per-unit, randomization tests, *t*-tests, waste composition

1352-8505 © 2004  Kluwer Academic Publishers

---

## 1. Introduction

Ratios are used extensively in fields such as environmental science, social science, economics, and zoology. Quantities that are otherwise not comparable can be presented together in a concise and easy to interpret way. Examples of the use of ratios include catch per-unit effort, morphometric indices, per-capita economic indices, and nutrient concentrations. If we define ratio data as any data in which the quantity of interest is one random variable divided by a second random number, the examples are endless. Analyzing data as ratios can lead to a wide variety of problems including the violation of standard statistical

\*Corresponding author.

1352-8505 © 2004  Kluwer Academic Publishers

assumptions and, potentially, misinterpretation of results. Yet, ratios are a convenient and intuitive way of summarizing information. They are and will continue to be used in many disciplines. The purpose of this paper is to evaluate the effectiveness of common statistical techniques when used for comparing two distributions of ratios.

Ratios are often used to standardize samples. Catch per unit effort is an example of such a situation. To compare the number of animals under two different conditions, it is necessary to control for the area searched, the time searched, or the intensity of the search effort. If the samples are drawn from a population in which the animals follow a uniform distribution in space or time, then any one observation can be thought of as contributing a group of smaller, equally sized observations. We might think of an observation such as 12 animals in 200 m<sup>2</sup> as being equivalent to 200 1 m<sup>2</sup> observations, each with an expected value of 12/200 animals. In this situation, it is appropriate to combine the observations in each treatment with a weight proportional to the denominator (area, time, or intensity). We call this “per-unit inference”. Per-unit inference is appropriate where the denominator can be broken into discrete units that are similar across observations.

We define the other extreme as “whole object inference”. In this case, the researcher does not consider the denominator to be composed of multiple, smaller, identical units. Each observation would then deserve equal weight, no matter the size of the denominator. Whole object inference is more common in applications such as morphometric indices where each ratio reflects the value for one individual. For example, in fish ecology, the ratio of body weight to length is often used as an indicator of fish condition. In many research examples, the scale of inference is clear. In other cases, the scale of inference is a subtle issue. The research question could be framed in either way. The most appropriate analytical methods will depend on which question the researcher wants to answer. The first data example below is an example of such a situation.

A second analytical concern with ratio data is whether the expected relationship between the numerator and the denominator is isometric (linear through the origin). This is often true; one would not expect to observe animals in 0 km<sup>2</sup>. In these cases, there is often an implied dependence and one variable is clearly the denominator. In other cases, the ratio itself has ecological meaning and the relationship between the numerator and denominator may not go through the origin. For example, a pond may have to be greater than a certain size for fish to be present. In this case, the relationship between the number of fish and pond area will not go through the origin. These types of ratio models have been less frequently considered in the statistical literature. Investigation of these situations requires graphical analysis in addition to statistical testing because of the multiple ways in which two populations may differ if not constrained to go through the origin. We hypothesize that the best analytical approach will differ between situations that reflect per-unit versus whole object inference and will also depend on the presumed relationship between the numerator and the denominator.

Two problems with ratio data have been well treated in the statistical literature already and are not the subject of our analyzes. First, several authors have pointed out that interpreting results of analyzes based on ratio data can be non-intuitive, potentially leading to unintended inference and incorrect conclusions (Atchley *et al.*, 1976; Beaupre and Dunham, 1995; Jasienski and Fakhri, 1999; Raubenheimer and Simpson, 1992). They advocate reformulating the hypothesis in terms of the numerator, using the denominator as an explanatory variable. While we agree that inference based on the comparison of ratios requires caution, we argue that ratio-based inference allows for simple yet robust statistical

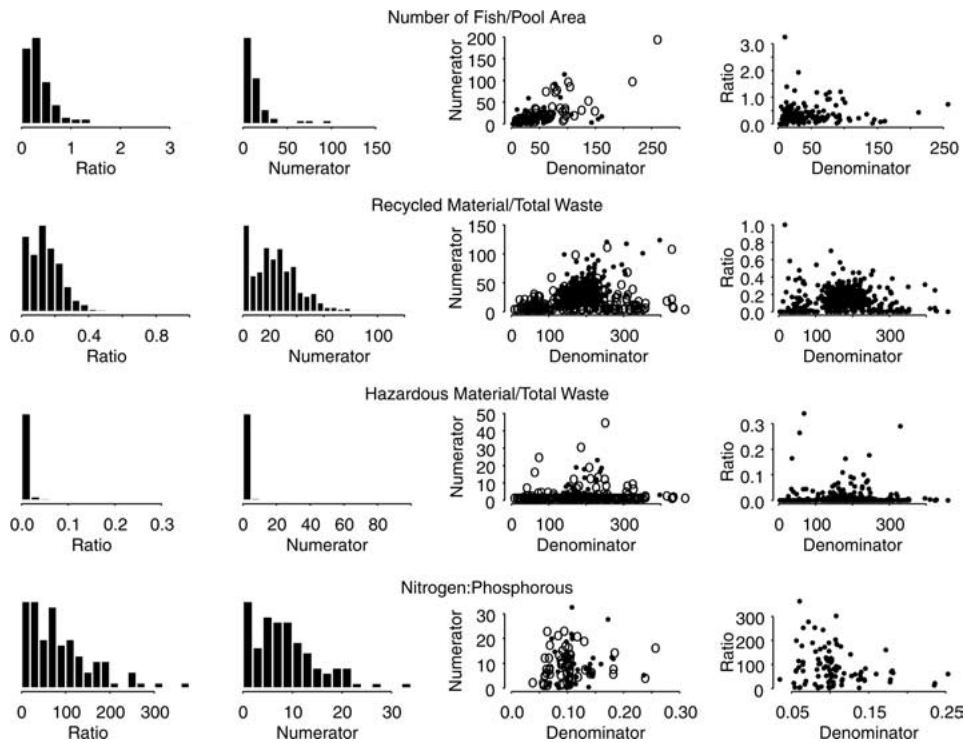
techniques, while use and interpretation of analysis of covariance (ANCOVA) requires a greater degree of statistical sophistication and is vulnerable to more violations of assumptions (Good and Hardin, 2003). In this paper, we focus on ratio-based approaches but include simple ANCOVA methods for comparison. The other well-documented statistical issue is that of spurious correlation (Atchley *et al.*, 1976; Kronmal, 1993; Pearson, 1897). Pearson drew attention to the problem of spurious correlation in 1897 demonstrating that two uncorrelated variables may appear highly correlated when divided by the same covariate. We restrict our discussion in this paper to those situations in which the researcher is simply interested in comparing two populations on the basis of the distributions of  $X_1/Y_1$  and  $X_2/Y_2$ .

In this paper, we differentiate between types of ratio problems and consider which of the common methods of inference might be best in each case. We emphasize those techniques that can be easily applied by both statisticians and non-statisticians alike and do not explore multi-step procedures, the role of diagnostics, nor techniques such as likelihood ratio tests that are not available in standard software packages. We apply the approaches to a diverse set of empirical and simulated data and compare the performance of statistical approaches based on both observed versus expected alpha levels and statistical power. We begin by presenting the empirical data sets and the models from which the simulated data are generated. A series of statistical approaches to analyzing ratios are then described. Finally, the empirical and simulated data are used to compare statistical approaches across data sets. We conclude by providing suggestions for appropriate analysis of ratio data.

## 2. Data

We motivate this paper with three examples in which inference about ratio data is desired. The first problem investigates the effectiveness of stream restoration techniques (Roni and Quinn, 2001). Stream restoration is a common tool in the effort to recover endangered salmonids in the Pacific Northwest. The data we use in this paper come from a series of restoration projects (the placement of logs to create pool habitat and cover in streams) in western Oregon and Washington states. The question of interest is whether fish density is different between restored and control streams. Fish numbers are collected in the field by electro-fishing an entire habitat unit (either a pool or a riffle), which might range in size from a few  $\text{m}^2$  to over  $100 \text{m}^2$  (Fig. 1). Logistically, it is impossible to control for pool size. Fish density, calculated as the number of fish per  $\text{m}^2$ , is therefore a ratio with a random denominator. Whether per-unit or whole object inference is most appropriate is a matter of biological importance. Per-unit inference would assume that all  $\text{m}^2$  are similar. Statistical analysis would consider differences in fish per  $\text{m}^2$  between treated and control streams. Whole object inference, on the other hand, should be used if all  $\text{m}^2$  of pool were not equivalent. Inference would then consider differences between densities calculated at the pool scale. How the analyses for these two types of inference might differ is described below.

The second example concerns municipal solid waste. Researchers want to answer questions about the distribution of various materials in the wastestream. For example, they might be interested in whether hazardous materials compose a larger fraction of the material in the dry versus the wet season. Often, they are also concerned with detecting the effect of a particular program, for example, a new hazardous materials disposal station or



**Figure 1.** Data from the empirical examples. Each row describes data from one of the examples. Histograms of the ratio and the numerator are in the left two columns. The next column presents the correlation between the numerator and the denominator for each example. Points and circles differentiate the two treatments. The far right scatterplot examines the relationship between the ratio and the denominator for each example.

the distribution of recycling bins. Much of the waste arrives at the transfer station as self-hauled material and therefore arrives in various amounts from a single bag of garbage to a flatbed truck loaded with trash. In most cases, this material is unevenly distributed and very difficult to mix. For example, it might include a refrigerator, several steel I-beams, and a large volume of demolition material. As a result, the denominator, the total volume or weight of material sorted, cannot be controlled. Usually, the fraction of potentially recyclable materials is fairly large and there are few zero observations. Hazardous materials, however, tend to be rare and their distribution includes a large spike at zero (Fig. 1). Per-unit inference is most appropriate for both the fraction of recyclable materials and the fraction of hazardous materials.

Our final example looks at nutrient cycling in subsurface stream water. These nutrients regulate bacterial growth in the subsurface water and affect the quantity and type of nutrients returned to surface water. Nutrient ratios in the surface water can accelerate or limit algae growth, invertebrate biomass, and therefore, fish growth and survival. The ratio of nitrogen to phosphorus determines which of these nutrients is limiting growth. The data we use in this study are from a series of wells on a riparian terrace on the Queets River in western Washington State (Clinton, 2001). The data were collected in February and

October. At the onset of fall rains in October, it was predicted that nitrogen concentrations would be high as rainwater moved through soils in which alder forests had been sequestering nitrogen. Nitrogen concentrations are expected to be lower in February because much of the nitrogen is already leached out of the soil. This data set is particularly interesting because, unlike many common applications of ratio data, there is no reason why nitrogen and phosphorous would be correlated in natural systems. The source of phosphorous is geologic, from weathered rocks, while the source of nitrogen is atmospheric. In this case, the denominator is not used to standardize the numerator; it is the ratio itself that is of interest to researchers and the expected relationship does not go through the origin (Fig. 1). Whole object inference is appropriate for this question.

### 3. Models

The appropriate statistical treatment of ratio data depends on the joint distribution of the numerator and the denominator. Although the examples above give us confidence that the results have relevance to real environmental problems, we do not know the exact mechanisms by which they were generated, for example, how the numerator and denominator are related. To provide us with this more detailed information we develop a series of models motivated by the major issues described in the introduction (scale of inference and expected relationship of the numerator and the denominator) and the environmental examples. Our models describe different relationships between the denominator ( $D$ ), and the mean and variance of the numerator ( $E(N)$  and  $V(N)$ ). The specific parameterizations of the models used to generate data are chosen to provide additional contrast to the empirical data (Table 1).

#### 3.1 Model 1— $E(N)$ and $V(N)$ are proportional to $D$

Assume that each ratio describes a group of sub-samples for which the denominator units are equal (e.g.,  $6/4 \leftarrow 1/1, 3/1, 1/1, 0/1$ ), and the distribution of the sub-samples is independently and identically distributed (iid) across all ratios in the sample. In this case, the variance of the numerator for a given denominator size is proportional to the

**Table 1.** The parameterizations used when generating data with the three models. The two numbers in parentheses are parameter values for the two groups of data that are compared in the tests.

<i>Model</i>	<i>Parameterization</i>
Model 1 (Poisson)	$d = (1, 1.1), \mu_D = 50, cv_D = 0.5$
Model 2 (negativebinomial)	$d = (1, 2.5), \delta = 2, \mu_D = 50, cv_D = 0.5$
Model 3 (ratio of independent gammas)	$\mu_N = (1, 2), cv_N = 0.5, \mu_D = 1, cv_D = 0.5$

denominator, and the variance of the ratio for a given denominator is proportional to the reciprocal of the denominator.

$$\text{Var}\left(\frac{N}{D} \middle| D\right) = \text{Var}\left(\frac{\sum n_i}{\sum u}\right) = \frac{m\text{Var}(n)}{(mu)^2} \propto \frac{1}{m} \propto \frac{1}{D}.$$

Here  $u$  is the denominator of the sub-samples and  $m$  is the number of sub-samples.

A biological example of this is the number of fish in a habitat unit divided by the area of the habitat unit (as described in the first data example above). If the fish are distributed approximately randomly through all of the square meters of habitat, the conditions above are met and the variance of a density estimate for a pool is proportional to the reciprocal of the pool size. In other words, larger pools contain more information about fish density. This model describes a situation where per-unit inference is appropriate. We simulate this process using a Poisson distribution for the numerator and a gamma distribution for the denominator, where the mean of the numerator is proportional to the denominator and the fish per-unit area is  $d$ . The Gamma distribution is parameterized with mean,  $\mu$ , and coefficient of variation,  $cv$ .

Model 1:

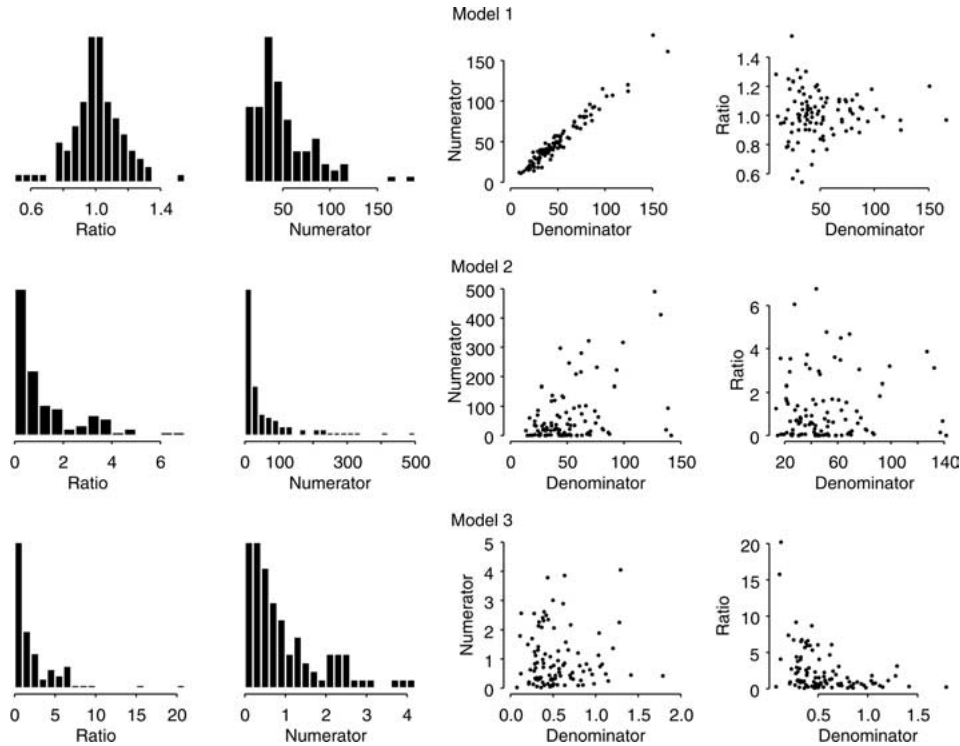
$$\begin{aligned} N_i | D_i &\sim \text{Poisson}(dD_i) \\ D_i &\sim \text{Gamma}(\mu, cv) \end{aligned}$$

The distribution of ratios generated from this model is approximately normal. There is a strong correlation between the numerator and the denominator and relatively little correlation between the ratio and the denominator (Fig. 2). This model captures many of the features of the empirical data set. The modeled data display a similar pattern to fish habitat data although the ratios from the fish habitat data set appear to follow a skewed distribution (Figs 1 and 2).

### 3.2 Model 2— $E(N)$ is proportional to $D$ and $V(N)$ is proportional to $dD + \delta(dD)^2$

In Model 1, the equally sized sub-samples (from which the ratios come) have the same distribution and are iid. It is easy to imagine a situation in which the means of the sub-samples vary between ratios. Returning to the fish per habitat unit example, we might expect that the number of fish is affected not only by the area of the unit but also by other physical conditions that may differ between units such as temperature, number of refugia from high velocity flows, and quality of food supply. This violates the assumptions of Model 1, as the mean density varies between habitat units.

This process can be modeled by assuming that the density ( $d$  in the model above) is no longer constant but instead is a random variable with mean  $d$  (Fig. 2). If the random variable follows a gamma distribution (with standard deviation proportional to the mean) then the distribution of the numerator, conditioned on the denominator, follows the negative binomial distribution with mean  $dD_i$  and variance  $D_i + \delta(dD_i)^2$  (see, for example, Cameron and Trivedi, 1990).



**Figure 2.** Data simulated from the three models of ratio data. Each row describes data from one of the models. Histograms of the ratio and the numerator are in the left two columns. The next column presents the correlation between the numerator and the denominator for each model. The far right scatterplot examines the relationship between the ratio and the denominator for each model. The parameter values used to generate the data are: Model 1:  $d = 1, \delta = 0, \mu_D = 50, cv_D = 0.5$ ; Model 2:  $d = 1, \delta = 2, \mu_D = 50, cv_D = 0.5$ ; Model 3:  $\mu_N = 1, cv_N = 0.5, \mu_D = 1, cv_D = 0.5$ .

Model 2:

$$N_i|D_i \sim \text{Negativebinomial}(dD_i, \delta)$$

$$D_i \sim \text{Gamma}(\mu, cv)$$

Here, the parameter  $\delta$  determines the degree of over-dispersion. If  $\delta$  is 0 the model reduces to Model 1 above. For large  $\delta$ , the variance of the density is independent of the habitat unit area ( $D_i$ ), and the information content for each unit is equal. In this case, whole object inference is appropriate.

In Model 2, the distribution of the ratios is highly skewed. While there is a relationship between the numerator and the denominator, it is not as strong as for Model 1. Ratios generated from this model are not correlated with the denominator (Fig. 2). The distribution of the ratios from the fish habitat data is somewhat better represented by this model than Model 1 (Figs 1 and 2).

Another mechanism that can lead to the negative binomial distribution is a birth process where the probability of a birth is a positive linear function of the current population size (e.g., Dennis, 1989). This can be translated to our fish density example by assuming the

probability of additional fish inhabiting a habitat unit is positively dependent on the number of fish currently in the unit.

### 3.3 Model 3— $N$ and $D$ are independent

Typically when the denominator is being used to standardize the numerator,  $E(N)$  is assumed to be proportional to  $D$ , as in the models above. However, there are many examples where this is not true and the expected relationship does not go through the origin. For example, in nitrogen to phosphorous ratios (described above), there is no reason why a very low value of nitrogen would dictate a low value of phosphorous. We model this type of process as the ratio of two independent gamma random variables.

Model 3:

$$\begin{aligned} N_i &\sim \text{Gamma}(\mu_N, cv_N) \\ D_i &\sim \text{Gamma}(\mu_D, cv_D) \end{aligned}$$

As expected, the ratio is not normally distributed and there is no correlation between the numerator and denominator in the simulated data (Fig. 2). The nitrogen to phosphorus ratios display a similar pattern (Fig. 1). Many other examples exist in the environmental literature of important ratios in which  $N$  and  $D$  may be independent. The ratio of hatchery to wild fish in a particular stream, for example, is used to examine potential genetic and behavioral impacts of hatchery fish. The number of wild fish provides no information about the number of hatchery fish. Likewise, air quality metrics may involve the ratio of unrelated contaminants.

## 4. Approaches to the analysis of ratio data

In this section, we describe methods used to analyze ratio data. We begin by discussing transformations that can be used to reduce heteroscedasticity and non-normality in ratio data, allowing for standard statistical approaches. We then demonstrate one way to account for differing information content between ratios by using inverse variance weighting. Finally, we use this groundwork and suggestions from the literature to motivate several tests for comparing two populations of ratios.

### 4.1 Transformations

A common approach used in the analysis of ratio data is transformation of the data followed by standard statistical analysis. Motivations for transforming the data include correcting for non-normality, variance stabilization, and converting multiplicative relationships to additive relationships. Often there is no single transformation that satisfies all of these goals. For example, for the Poisson distribution, the transformation that stabilizes the variance,  $\sqrt{X}$ , overcompensates for non-normality (which only requires

$X^{2/3}$ ). However, the variance stabilizing transformation tends to produce data that is approximately normal (Efron, 1982).

When the variance of the ratio is proportional to the expected value, as in Model 1, the variance stabilizing transformation is  $\sqrt{X}$ . If the variance is proportional to the expected value raised to a power between 1 and 2, as in Model 2, then the inverse hyperbolic sine square root transform,  $\text{arc sinh}(\sqrt{\delta X})$ , is variance stabilizing. Here  $\delta$  is the degree of overdispersion. For variance proportional to the square of the mean, the variance stabilizing function is the logarithm,  $\log(X)$ . To improve performance for small values Bartlett (1947) suggested  $\log(X + 1)$ . This relationship between the variance and expected value occurs for Model 2 as  $\delta$  becomes large. Also, if the numerator and denominator of a ratio are independent, and have variances proportional to their means, as in Model 3, then the logarithm is variance stabilizing for the ratio.

Another common transformation used for data summary and inference is ranking. Statistics and tests based on ranks are less sensitive to many of the problematic distributional characteristics of ratios such as skew and inequality of variances.

## 4.2 Accounting for ratios with differing information content

Ratios generated from Model 1 have variance inversely proportional to the size of the denominator. The size of a ratio's denominator in this case determines the number of effective sub-samples; i.e., the denominator determines the amount of information contained in the ratio. One way of accounting for the different information content in the different ratios is to use inverse variance weighting when calculating the mean. In the case of Model 1 this suggests weights equal to the reciprocal of the denominator. Applying this weighting to the mean produces the ratio estimator:

$$\hat{R} = \frac{D_1(N_1/D_1) + D_2(N_2/D_2) + \cdots + D_n(N_n/D_n)}{D_1 + D_2 + \cdots + D_n} = \frac{\sum N_i}{\sum D_i}. \quad (1)$$

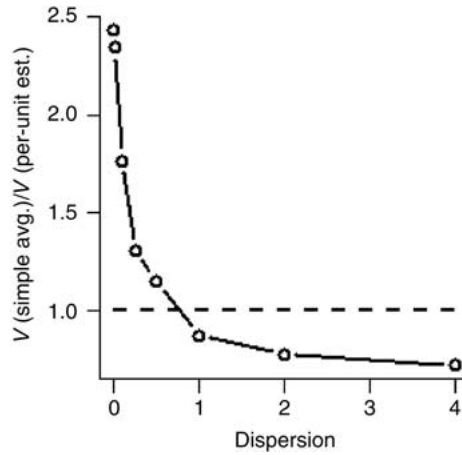
This ratio estimator has been used in the context of clusters of discrete elements where inference is about the per-unit mean. Cochran (1977, p. 31) uses the example of estimating suits per male member of a household. In this case, for each household we know the number of males and the number of suits. The number of suits per male can then be calculated as above. This can also be viewed as an estimator derived from stratified random sampling where only a small proportion of the strata are sampled but those that are sampled are sampled completely.

For Model 2, the variance of the ratio (conditional on the denominator) is:

$$V(R_i | D_i) = \frac{d}{D_i} + \delta d^2.$$

Here, part of the variance is determined by the denominator (as in Model 1) while the other part is independent of the denominator. As  $\delta$  gets large, this reduces to a constant with respect to  $D_i$ . Using inverse variance weighting, this reduces to the average of the ratios.

$$\hat{R} = \frac{c(N_1/D_1) + c(N_2/D_2) + \cdots + c(N_n/D_n)}{c + c + \cdots + c} = \frac{\sum R_i}{n}. \quad (2)$$



**Figure 3.** Plot of ratio of variance (simple average) / variance (per-unit estimate) against the dispersion parameter.

So as the habitat units become more unique, the relative information about average pool density becomes less dependent on the size of the unit (i.e., the size of the denominator).

If the assumptions of Model 2 are met, both the per-unit estimator (1) and the simple average of the ratios (2) are unbiased estimators of  $E(R)$ . The variance of the  $E(N)/E(D)$  (per-unit estimator) as approximated using Taylor's series expansion is:

$$\left(\frac{\bar{N}}{\bar{D}}\right)^2 \left( \frac{\text{Var}(\bar{N})}{\bar{N}^2} + \frac{\text{Var}(\bar{D})}{\bar{D}^2} - 2 \frac{\text{Cov}(\bar{N}, \bar{D})}{\bar{N}\bar{D}} \right) \approx \left(\frac{\bar{N}}{n\bar{D}}\right)^2 \left( \frac{\text{Var}(N)}{\bar{N}^2} + \frac{\text{Var}(D)}{\bar{D}^2} - 2 \frac{\text{Cov}(N, D)}{\bar{N}\bar{D}} \right).$$

This estimator of the variance should be applied with caution since even for moderately large samples it tends to substantially underestimate the true variance (Cochran, 1977, p. 163).

The amount of variation between pool densities (captured by the dispersion parameter  $\delta$ ) will determine whether weighting by the denominator provides a more efficient estimator than the simple average. This is illustrated by comparing the variances of the two estimators for data generated from Model 2 with  $\delta$  increasing from zero (Fig. 3). For small values of the dispersion parameter, Model 2 is close to Model 1 and the  $E(N)/E(D)$  estimator is more efficient than the  $E(N/D)$  estimator. As the between unit variability increases (i.e., dispersion increases), the  $E(N/D)$  estimator becomes more efficient. For a thorough review of the relative merit of these two estimators in the context of fixed-area quadrat sampling, see Williams (2001).

In the case of Model 3, the numerator and denominator are independent so the inverse variance weighting is,

$$\frac{\sum D_i^2 R_i}{\sum D_i^2} = \frac{\sum N_i D_i}{\sum D_i^2}.$$

This is the regression estimator for a regression of  $N$  on  $D$  through the origin. Because  $N$

and  $D$  are independent,  $R$  and  $D^2$  are dependent, and this statistic is a biased estimate of  $E(R)$ . For finite  $E(1/D)$  the estimator,

$$\frac{\sum N_i}{n} \frac{\sum 1/D_i}{n},$$

is an unbiased estimate of  $E(R)$ , has smaller variance than the simple average estimator, and is asymptotically normal (Welsh *et al.*, 1988).

The mean of the ratio of two independent random variables is often poorly defined since small values of the denominator are not necessarily paired with small numerators and therefore can produce very large ratios. Basing inference on the mean may therefore be ill advised in this case. Alternatives to estimating the mean include rank-based metrics such as the median and a generalized shift model described by Doksum and Sievers (1976).

### 4.3 Tests

The results above suggest many potential tests for detecting differences in the distribution of  $N/D$  between populations. Suggested tests based on transformations of the data include  $t$ -tests on log transformed data, square root transformed data, and arc sinh transformed data. The arc sin square root transform, suggested by statistical texts for proportions (e.g., Zar, 1996), is variance stabilizing for proportions generated from binomial data. However, it is not appropriate for ratios that may be larger than 1. Differences in appropriate estimators for per-unit and whole object inference suggest two permutation tests: one test based on the per-unit estimator and one test based on the average of the ratios. This allows evaluation of the two estimators using the same testing methodology.

Although the ANCOVA approach does not treat the ratio as the unit of data, we include it in our analysis because many authors promote it as a superior alternative to the analysis of ratios (e.g., see Jackson *et al.*, 1990; Raubenheimer and Simpson, 1992). The two ANCOVA models we consider predict the log transformed numerator using a linear function of the log transformed denominator:

$$\log(N_{i,j}) = \alpha_j + \beta_j \log(D_{i,j}) + \varepsilon_{i,j}.$$

## 5. Comparing the tests using simulations

We constructed seven models (Table 2) to represent a wide range of ratio data and then simulated from these models to assess the performance of the testing approaches. For simplicity we limit our analyses to the case in which we are comparing two populations. However, the results of these simulations readily generalize to comparisons of more than two populations. The models used were the empirical distributions of the four data sets (sampling with replacement) (Fig. 1) and the three constructed models discussed in the models section above. For each model, 10,000 sample data sets were generated. Each sample data set consisted of two samples of size 20 drawn from the same distribution, and one sample of size 20 drawn from a distribution that differed from the first by some effect size. For each sample data set, each of the tests was run twice, once comparing the two samples drawn from the same distribution to assess the true alpha level, and a second time

**Table 2.** The seven models and the methods used to generate the data.

	<i>Fish</i>	<i>Haz</i>	<i>Recyc</i>	<i>N to P</i>	<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>
<i>t</i> -test	0.045	0.032	0.678	0.127	0.456	0.390	0.321
sqrt <i>t</i> -test	0.049	0.053	0.886	0.147	0.455	0.421	0.387
log <i>t</i> -test	0.049	0.071	0.982	0.147	0.455	0.297	0.349
Wilcoxon	0.052	0.491	0.932	0.159	0.478	0.332	0.368
log ANCOVA w/int	0.126	0.023	0.003	0.092	0.076	0.027	0.071
log ANCOVA	0.070	0.094	0.996	0.158	0.441	0.316	0.363
permutation $E[R]$	0.060	0.080	0.878	0.108	0.462	0.466	0.392
permutation $E[N]/E[D]$	0.058	0.037	0.702	0.132	0.560	0.409	0.435

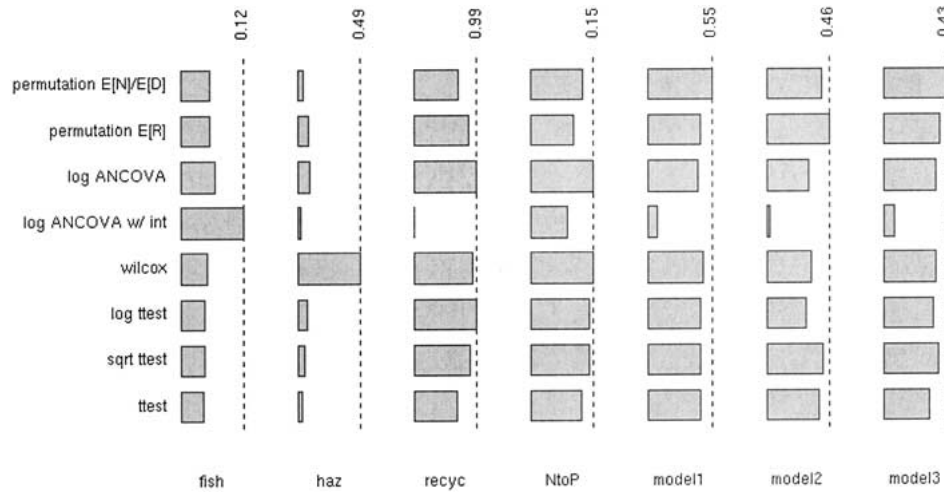
using samples drawn from the two different distributions to assess the power. For the four empirical data sets, the effect sizes were determined for us by the two different groups of data provided with each data set. The effect sizes for the constructed Models 1, 2 and 3 were chosen to produce power of approximately 0.5 (Table 1). All of the tests were two-tailed with level of 0.05.

The two sample tests that we assessed can be broken into four groups: *t*-tests on raw and transformed data, permutation tests based on different estimators, a non-parametric test, and ANCOVA-based tests. The *t*-tests were performed on untransformed data, square root transformed data and log transformed data. Welch's *t*-test for unequal variances was used to account for likely departures from the equal variance assumption. Two permutation tests were constructed, one based on the simple average of the ratios and the other on the per-unit estimator. For a test statistic, we used the squared difference between the estimated means for the two groups. The Wilcoxon two-sample rank test represents the nonparametric approach. Two ANCOVA's on the log transformed data are included. The first ANCOVA assumes the relationship between  $\log(N)$  and  $\log(D)$  is linear through the origin. In the second test, this assumption is relaxed allowing non-zero intercepts. Because this method can detect differences in the slopes or intercepts, a rejection is considered to occur when either rejects at the 0.025 level. Because the Bonferroni correction is conservative, the nominal alpha-level may be substantially less than 0.05.

## 6. Results

The estimated alpha levels for the tests were all approximately correct or conservative (in that they did not exceed the stated 0.05 level) except for the log ANCOVA with the fish data (0.064) and the log ANCOVA w/ intercept with the data from Model 1 (0.068) (Fig. 4 and/or Table 3). The log ANCOVA with intercept and *t*-test were both conservative for a number of data sets. This is understandable, given the rejection procedure described above for the log ANCOVA with intercept and the degree to which the *t*-test assumptions are violated. For  $p = 0.05$  (the probability of rejection) and  $n = 10,000$ , the standard deviation due to simulation is 0.0022. For values close to 0.05, this leads to an approximate 95% CI of  $(p - 0.0043, p + 0.0043)$ .

We also compared the power of the tests for each model (Fig. 5 and/or Table 4). The differences in power between tests were not large with two exceptions. The log ANCOVA



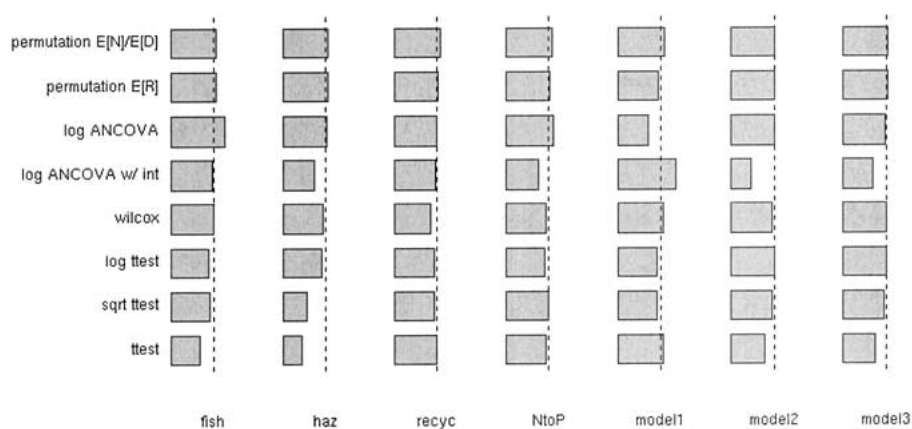
**Figure 4.** The estimated alpha levels of the different tests for each model. Dashed lines represent the 0.05 level (which was the specified level for all of the tests). If a bar extends past the dotted line, the nominal level is greater than 0.05.

**Table 3.** The estimated alpha levels of the different tests for each model.

	<i>Fish</i>	<i>Haz</i>	<i>Recyc</i>	<i>N to P</i>	<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>
<i>t</i> -test	0.034	0.022	0.050	0.047	0.052	0.039	0.038
sqrt <i>t</i> -test	0.046	0.029	0.048	0.050	0.045	0.049	0.048
log <i>t</i> -test	0.045	0.046	0.048	0.046	0.045	0.051	0.050
Wilcoxon	0.050	0.047	0.043	0.047	0.053	0.048	0.050
log ANCOVA w/int	0.048	0.037	0.049	0.038	0.068	0.022	0.034
log ANCOVA	0.064	0.052	0.050	0.055	0.035	0.051	0.050
permutation $E[R]$	0.053	0.053	0.052	0.052	0.047	0.051	0.052
permutation $E[N]/E[D]$	0.053	0.053	0.054	0.054	0.054	0.051	0.053

w/ intercept was almost twice as powerful as other tests for the model generated from the fish data. For the remaining models and data sets, however, it was generally the least powerful of the tests. Both ANCOVA tests were very sensitive to changes in the constant used with the log transform (e.g.,  $\log(X + 0.01)$  vs.  $\log(X + 1)$ ). The Wilcoxon two sample rank test also did much better (five times) than all other tests for the model based on the hazardous waste data. This can be explained by a large number of zeros in one of the groups. The group with fewer zeros had many small values as well so the means of the two groups were comparable. However, the many zeros in the first group resulted in drastically different rank statistics. Thus the rank-based test found a much larger difference between the groups than the mean-based tests.

The permutation tests based on  $E(N)/E(D)$  and  $E(N/D)$  both performed relatively well across all the data sets. As expected, the  $E(N)/E(D)$  based test performed better for Model 1, while the  $E(N/D)$  based test did better for data generated from Model 2. The



**Figure 5.** The estimated power of the different tests for each model. The dashed line indicates the highest achieved power for each model. For example, the Wilcoxon rank test (wilcox) was the most powerful of the tests when analyzing the hazardous waste data (haz). It had a power of 0.49.

**Table 4.** The estimated power of the different tests for each model.

	<i>Fish</i>	<i>Haz</i>	<i>Recyc</i>	<i>N to P</i>	<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>
<i>t</i> -test	0.034	0.022	0.050	0.047	0.052	0.039	0.038
Sqrt <i>t</i> -test	0.046	0.029	0.048	0.050	0.045	0.049	0.048
Log <i>t</i> -test	0.045	0.046	0.048	0.046	0.045	0.051	0.050
Wilcoxon	0.050	0.047	0.043	0.047	0.053	0.048	0.050
Log ANCOVA w/int	0.048	0.037	0.049	0.038	0.068	0.022	0.034
Log ANCOVA	0.064	0.052	0.050	0.055	0.035	0.051	0.050
Randomization $E[R]$	0.053	0.053	0.052	0.052	0.047	0.051	0.052
Randomization $E[N]/E[D]$	0.053	0.053	0.054	0.054	0.054	0.051	0.053

$E(N)/E(D)$  based test had highest power of all the tests for data generated from Model 3, where the numerator and denominator are independent. The *t*-tests were generally comparable in power to the other tests. Although the log and sqrt transformed *t*-tests both tended to do better than the untransformed *t*-test, the *t*-test on the untransformed data also did relatively well for even highly skewed data.

## 7. Discussion

Ratio data are extremely common and provide a useful method for summarizing information. Analysis of ratio data requires consideration of the scale of inference and the expected relationship between the numerator and the denominator. We presented empirical data sets in which multiple scales of inference were possible and data sets in which per-unit and whole object inference were clearly the best choice. We also provided an example of aquatic chemistry data in which the expected relationship between the numerator and the denominator did not go through the origin. To complement the empirical data and better understand the effects of specific assumptions, we developed

three mathematical models to simulate ratio data. These models emphasize the importance of both scale of inference and whether or not the relationship between the numerator and denominator is linear through the origin. While we did not attempt to simulate the empirical data exactly, the simulated data sets replicated many of the key patterns found in the empirical case studies (Figs 1 and 2). The fish density data patterns fell in-between those observed for Models 1 and 2 and the aquatic chemistry data like model 3 (which did not go through the origin).

Our simulation study demonstrated that most tests were safe tests—the observed alpha was less than or equal to the specified value. The only situations in which the test rejected more often than expected involved the ANCOVA-based tests. Error rates ANCOVA-based tests might be reduced by improving model specification through alternate transformations, residual diagnostics, or formulation of alternative error distributions (e.g., negative binomial, or quasipoisson) (see, Neter *et al.*, 1990). However, properly transforming the data and choosing an appropriate error distribution are frequently non-trivial, especially when dealing with the highly skewed data often observed in environmental examples of ratio data. At a minimum, our results demonstrate that the success of the ANCOVA approach is very sensitive to model formulation. For example, alternate formations of the log transformation substantially altered the observed type I error rate. In addition the log ANCOVA w/ intercept had the lowest power of all tests for six out of the seven models.

The general hypothesis being tested is “is there a difference in central tendency between the two groups of ratios?”. However, the statistical procedures we examined test this general hypothesis in different ways. The choice of estimator and the transformations of the data affect the precise statement of the hypothesis. Also, the ANCOVA model tests a slightly different hypothesis. It tests whether there is a common linear relationship between the numerator and denominator for the two treatments. Differences in the hypothesis being tested may explain observed differences in power between approaches. For example, although the Wilcoxon test had high power when comparing ranks between the two groups in the toxic waste example, none of the mean-based tests had high power. This is explained by many zeros in one of the groups. When selecting example data sets for inclusion in this paper, we frequently encountered highly skewed data with a large number of zeros similar to the toxic waste data. This suggests that this sensitivity of test results to test choice may be relatively common when analyzing environmental ratio data.

There was often little difference in the power between tests for a given model, suggesting that in situations where the data are not highly skewed the choice of test should be based on an understanding of the underlying data model. We offer recommendations for analysis of ratio data based on the intended scale of inference and the expected relationship between the numerator and the denominator. In all cases, we strongly encourage graphical analysis before selecting a statistical test. If the intended scale of inference is at the per-unit scale, the permutation test based on  $E(N)/E(D)$  is preferable to the other approaches. Our other candidate approaches consider the ratio as a unit and do not use an estimator of location that reflects the differing information content of observations with different denominators.

Where whole object inference is required, more of our tests reflect the correct underlying model. Where the expected relationship between the numerator and denominator is linear through the origin, we recommend one of the  $t$ -tests because of their simplicity and robustness. Asymptotically the power of the  $t$ -test is insensitive to violations of the normality assumption (Lehmann, 1986). This assures good performance for the  $t$ -tests where  $n$  is large. If the expected relationship is not linear through the origin,

there are three additional considerations. First, extremely small denominators will produce ratios that tend toward infinity. These extremely high values may obscure overall patterns. Second, there are multiple ways in which the distributions may differ between treatments. The slope might be different, the intercept might be different, or both. And, finally, if there is no reason to use one value or the other as the denominator, inference may vary according to the choice of denominator. For whole object inference and relationships that are not linear through the origin, we recommend the Wilcoxon test. The non-parametric test performed well for all our example data models and has the advantage that observations with small denominators are converted to ranks and so do not have a disproportionate effect on the test statistic.

As in all simulation studies, we could only include a small number of possible scenarios. Instead of focusing on the extreme cases where the analyses were most likely to falter, we used what we thought was a varied yet realistic group of models (based on data and constructed mechanisms), under common testing conditions. Care should therefore be taken when extrapolating these results to more extreme situations. Our use of equal sample sizes likely reduced the negative effects of inequality of variance, and the decision to use two sided as opposed to one sided tests made our *t*-tests more robust to skewed data. With unequal sample sizes or for one-sided tests, we recommend using the Wilcoxon rank-based test or one of the permutation tests.

## Acknowledgments

The authors thank Phil Roni, Watershed Program, NW Fisheries Science Center, Seattle, WA, Sandra Clinton, Center for Streamside Studies, University of Washington, Seattle, WA, and Cascadia Consulting Group, Seattle, WA for providing data samples. Emily Silverman, Rich Zabel, Owen Hamel, Phil Roni, Phillip Good, and two anonymous reviewers provided many useful comments when reviewing earlier drafts of this paper.

## References

- Atchley, W.R., Gaskins, C.T., and Anderson, D. (1976) Statistical properties of ratios. I. Empirical results. *Systematic Zoology*, **25**, 137–48.
- Bartlett, M.S. (1947) The use of transformations. *Biometrics*, **3**, 39–52.
- Beaupre, S.J. and Dunham, A.E. (1995) A comparison of ratio-based and covariance analyses of a nutritional data set. *Functional Ecology*, **9**, 876–80.
- Cameron, A. and Trivedi, P. (1990) Regression based tests for overdispersion in the Poisson model. *Journal of Econometrics*, **46**, 347–64.
- Clinton, S. (2001) *Microbial Metabolism, Enzyme Activity and Production in the Hyporheic Zone of a Floodplain River*, Seattle, University of Washington.
- Cochran, W.G. (1977) *Sampling Techniques*, John Wiley & Sons, New York.
- Dennis, B. (1989) Allee effects: population growth, critical density, and the chance of extinction. *Natural Resource Modeling*, **3**, 481–538.
- Doksum, K.A. and Sievers, G.L. (1976) Plotting with confidence: graphical comparisons of two populations. *Biometrika*, **63**, 421–34.
- Efron, B. (1982) Transformation theory: how normal is a family of distributions? *The Annals of Statistics*, **10**, 323–39.

- Good, P.I. and Hardin, J.W. (2003) *Common errors in statistics (and how to avoid them)*, Wiley-Interscience, Hoboken, NJ.
- Jackson, D.A., Harvey, H.H., and Somers, K.M. (1990) Ratios in aquatic sciences: statistical shortcomings with mean depth and the morphoedaphic index. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 1788–95.
- Jasienski, M. and Fakhri, B.A. (1999) The fallacy of ratios and the testability of models in biology. *Oikos*, **84**, 312–26.
- Kronmal, R.A. (1993) Spurious correlation and the fallacy of the ratio standard. *Journal of the Royal Statistical Society A*, **156**, 379–92.
- Neter, J., Wasserman, W., and Kutner, M.H. (1990) *Applied Linear Statistical Models Regression, Analysis of Variance, and Experimental Designs*, Irwin, Homewood, IL.
- Pearson, K. (1897) On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, **60**, 489–502.
- Raubenheimer, D. and Simpson, S.J. (1992) Analysis of covariance: an alternative to nutritional indices. *Entomologia Experimentalis Et Applicata*, **62**, 221–31.
- Roni, P. and Quinn, T.P. (2001) Density and size of juvenile salmonids in response to placement of large woody debris in western Oregon and Washington streams. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 282–92.
- Welsh, A.H., Peterson, A.T., and Altmann, S.A. (1988) The fallacy of averages. *Am. Nat.*, **132**, 277–88.
- Williams, M.S. (2001) Performance of two fixed-area (quadrat) sampling estimators in ecological surveys. *Environmetrics*, **12**, 421–36.
- Zar, J.H. (1996) *Biostatistical Analysis*, Prentice-Hall, Upper Saddle River, New Jersey.

## Biographical sketches

Dr Martin Liermann is a statistician with the Watershed Program at the NW Fisheries Science Center where he leads statistical research projects and provides consultations on a wide variety of fish and fish habitat research projects. He also participates in the field collection of fish density, stream restoration, and habitat survey data. Dr Liermann has a Ph.D. in Quantitative Ecology and Resource Management from the University of Washington and an M.S. in Environmental Systems from Humboldt State University.

Dr Ashley Steel is a quantitative ecologist with the Watershed Program at the NW Fisheries Science Center where she leads a series of landscape-scale projects for salmon recovery planning. She also conducts research on the effects of water quality on fish movements and survival. Dr Steel has a Ph.D. in Quantitative Ecology and Resource Management, an M.S. in Statistics, and an M.S. in River Ecology from the University of Washington.

Michael Rosing is database manager and in-house statistical consultant for the Greenland Institute of Natural Resources where he has the opportunity to work on several different species both terrestrial and marine. His main contributions lie in the field of estimating gear selectivity parameters. He has an M.S. in Quantitative Ecology and Resource Management from the University of Washington.

Dr Peter Guttorp is Professor of Statistics at the University of Washington. He received his bachelor's degree from the University of Lund, Sweden, in 1974, and his Ph.D. from the University of California at Berkeley in 1980. His research interests are in stochastic modeling of scientific data, particularly in environmental and earth sciences.