SAMPLE SIZE AND DATA ANALYSIS:

ARE WE CHARACTERIZING AND COMPARING DIET PROPERLY?

Lara A. Ferry
Department of Ecology and Evolutionary Biology
University of California
Irvine, CA 92717-2525 USA
phone: (714) 824-4332
fax: (714) 856-2181
e-mail: LAFERRY@UCLEDU

Gregor M. Cailliet
Moss Landing Marine Laboratories
P.O. Box 450
Moss Landing, CA 95039-0450 USA

Abstract

We reviewed over 200 papers that compared primarily fish diets between species, sites, and sample dates. Precise descriptions of each diet are necessary in order to make comparisons between them. Yet none of the studies we reviewed provided any estimates of precision. We advocate the use of cumulative prey curves (performance tests) for determining whether enough samples have been collected to describe diet precisely and for subsequent comparisons. The studies we reviewed used indices, correlations, and parametric statistics to compare diets, and many claimed to find no difference between diets. We used a power analysis technique to determine that a majority of those papers would have found a difference between diets had they increased their sample size only slightly. We also advocate using individual guts as a sample unit, thereby providing for the estimate of standard deviation for a given index, thus increasing its utility in determining if differences exist, especially if traditional parametric statistics cannot be applied. This does not, however, make indices strong inference tools. Particularly shocking was our observation that many studies used only indices to compare diet, without statistical analysis, and explicitly stated that “significant” differences existed between diets.

Introduction

In studies of fish feeding habits, it is often useful to ask whether the diets of two species differ, how much they differ, and whether these differences are statistically significant. Similarly, these questions can be extended to comparisons within a species between or among different locations, times, sexes, or age classes of fish. In order to make such comparisons, however, it is necessary to be sure that each diet being compared is described adequately, or precisely. Once comparisons have been made, studies often claim that no differences between diets exist. However, often the researchers have not collected enough samples to detect any differences, if a difference really exists. Non-significant results may actually have been significant had only a few more samples been collected. This error was made because the test performed lacked power.

A Basic Review of Power Analysis

The power of a test is defined as the probability of correctly rejecting a null hypothesis, and is mathematically equal to 1-β, where β is the probability of not rejecting a false null hypothesis (Type II error). Furthermore β is related to α, the probability of incorrectly rejecting a null hypothesis (Type I error). Mostly by convention, we generally set α at 0.05, meaning we are
willing to reject a null hypothesis that is not false 5% of the time (see Zar, 1984; Sokal and Rohlf, 1995). The extent to which \( \beta \) errors occur, however, is often ignored.

Traditionally, \textit{a posteriori} power analysis techniques are used to determine the power, or 1-\( \beta \), level of a test after it has been performed. To determine the expected power of a test that has yet to be performed, \textit{a priori} techniques are employed. In both cases, \( \beta \) will depend on several factors: variability of the data (precision), effect size (the difference to be detected), sample size, and \( \alpha \). Given a certain sample size, effect size, and precision, as one changes \( \alpha \), \( \beta \) will vary inversely. The equations for power can be manipulated so that one can solve for any one of the above factors. In doing so, the \( \alpha \) and \( \beta \) levels can be set by the researcher, and the researcher can then determine, for example, how many samples are necessary to detect a difference (effect size) of \( d \). Power analysis is infinitely useful for planning experiments and sampling designs, as well as determining the validity of a studies results.

It is the purpose of this review to elucidate problems in the current literature regarding the lack of precision in dietary descriptions and power in dietary comparisons. We provide a detailed description of a method for evaluating the precision of the data sets to be compared. We determine the proportion of recently published studies that were supported by a sufficient sample size for the comparisons made and conclusions drawn. We present \textit{a priori} power techniques for assessing the number of samples needed to detect a difference between diets given the stated effect size, and compare those values to the actual number of samples (i.e. guts) collected. Consideration is given to effect size, which is important for determining when these predictions of sample size are relevant. Finally, we describe techniques for increasing the inferential strength of diet indices, typically used to compare diets.

\textbf{Methods}

Published laboratory and field studies that compared diets between fish species, size classes, sites studied, and experimental groups were used to evaluate precision and sample size used for detecting differences in diet composition. Studies were chosen from published papers located using a standard literature data base (Aquatic Sciences and Fisheries Abstracts). Studies were targeted in the data base using the key words diet, food, feeding, or analysis, and selected based on the material contained in the abstract. We were looking specifically for mention of the type of analysis used in the study, and an indication of how much of the data were presented. More than 200 papers were subsequently scrutinized in the course of this study.

\textbf{Precision in Diet Description}

None of the original \( \approx 200 \) studies utilized any technique for determining if an adequate number of samples had been collected to precisely describe diet prior to performing comparisons. Therefore, we felt compelled to re-describe a technique for performing such evaluations. We recommend the use of cumulative prey curves (or performance curves; see also Elliott, 1971; Hurturbia, 1973; Cailliet, 1977; Hoffman, 1979; Cailliet et al., 1986). These are based on the fact that as sample sizes increase, variation (and species richness) tends to decrease, and thus the curve reaches an asymptote as new prey types are being introduced into the diet only rarely.

Cumulative prey curves are created by plotting the cumulative number of prey types (i.e. unique items) against the cumulative number of guts analyzed. Or, as mathematically described:

\[
S_n f(n),
\]

where \( S \) is the number of prey types found for a given number of guts analyzed \( n \). Variations of this technique can be found in Hurturbia (1973) and Hoffman (1979), where diversity indices \( H \) are recalculated as new guts are added to the sample size \( n \), and each subsequent \( H_n \) is plotted against \( n \). Curves can be plotted in the original order of gut analysis, or by randomizing. Guts are generally opened and the contents quantified in a haphazard order (usually the order of collection, which is often correlated with some other feature of the sampling design). It is
important to randomize the order of analysis to prevent bias. Numerous randomizations can be performed, from which a mean number of new prey types can be determined for each consecutive gut, and plotted. Standard deviations can be calculated providing a measure of the variation in number of new prey types throughout the analysis. This allows one to see not only if the curve reaches an asymptote, but the variability of the asymptotic region.

As an example of the technique, we constructed cumulative prey curves from three sample data sets (collected by LAF and GMC). Curves were plotted using both original and randomized gut orders, allowing for comparison of the asymptotic nature of the curves.

**Assessing Sample Size Sufficiency in Making Comparisons**

We used an *a priori* power analysis technique described by Cohen (1988) to determine if adequate sample sizes were being used to ascertain diet differences between study groups. The technique (described herein) requires that diet data be presented as either actual numerical values or as proportions of total diet for each variable quantified (i.e., actual numbers of a diet item found in fish from one site versus another). Because such raw data are rarely provided, and few papers supplied graphs that allowed the extraction of such values, only 55 studies of the original =200 presented enough data to be analyzed with the power analysis technique. Of these, it was necessary for the author(s) to have stated implicitly that no difference was detected between diets studied (the corollary being that if a difference in diet was detected, clearly the sample size used was large enough). Thus, for this analysis, 41 separate, independent comparisons were evaluated.

In these 41 studies, data were more often presented as proportions, rather than actual values. Proportions were expressed as either percent number (%N; quantity of a prey type relative to the total quantity of prey collected), percent frequency of occurrence (%FO; number of predators in which a prey type appeared in relative to the total number of predators), or a combination, such as Index of Relative Importance (IRI; [(%N + %V) x %FO], Pinkas et al., 1971). These values were compared using a variety of standard diet indices (i.e., Percent Similarity Index, Hutchinson’s Niche Breadth, Morisita’s Index of Similarity; see Silver, 1975, Cailliet and Barry, 1978, Krebs, 1989). Occasionally, proportions and actual data values were also compared using parametric t-tests and ANOVA, as well as correlation (usually Spearman’s ranked correlation; see Zar, 1984).

To our knowledge, the *a priori* power analysis techniques used in our study have not previously been used to determine sample size sufficiency for food habits studies. As previously mentioned, such techniques are generally used to determine the power of a test to be performed. We use *a priori* techniques here to evaluate tests already performed, as they are the only type possible given the data obtained from the studies we reviewed (this is discussed further in the next paragraphs). According to convention, we set $\alpha=0.05$ and $\beta=0.20$ (see Zar, 1984). Given these values and the effect sizes measured in the studies being critiqued, we calculated the sample size necessary ($n$) to detect a significant difference according to the formulas outlined here.

For diet comparisons using indices:

$$\hat{n} = \frac{1570}{100 \times h^2},$$

(Eq. 1)

where the value 1570 comes from Cohen’s (1988) tables for the $\alpha$ and $\beta$ levels set. The tables are specific to this particular equation. The symbol $h$ represents the difference between the arcsine-transformed proportions being compared by the index. When multiple values were compared in a study (i.e., proportions of several different diet items across sites), we selected the largest difference (effect size) presented, assuming this would be the easiest difference to detect and the most forgiving test of sample size sufficiency. To determine the validity of using Equation 1, we used control data sets (collected by the authors, some of which are the same data as used in the cumulative prey curves, see Results) and compared $\hat{n}$ calculated Equation 1 to $n$
calculated by a more traditional formula that incorporates variance (Cohen, 1988; Krebs, 1989):

\[
\hat{n} = \frac{(Z_\alpha + Z_\beta)^2 \times s^2}{d^2},
\]

(Eq. 2)

where \( s^2 \) is the variance, \( d \) is the untransformed difference between the proportions being compared, and \( Z_\alpha \) and \( Z_\beta \) are the \( z \)-scores for the \( \alpha \) and \( \beta \) values set (for \( \alpha = 0.05, Z_\alpha = 1.9600 \); for \( \beta = 0.20, Z_\beta = 0.8500 \)). Equation 2 was otherwise impossible for us to utilize since variance is not usually provided in studies using indices to compare diet. For the control datasets, \( %N \) of a diet item was the measure compared between diets.

This same method for estimating \( \hat{n} \) (Eq. 2; see also Winer, 1971; Sokal and Rohlf, 1995) was used to evaluate a second category of diet comparisons; those using parametric tests (t-test or ANOVA; for the latter, choosing one pair of values within the ANOVA according to the largest difference criterion used above). The type of analysis used (t-test or ANOVA) was simply a category to separate the original 200+ papers and for choosing the best power analysis technique from those available. (Note: our technique does not calculate the power of the statistical test per se since we are not using any value obtained from the statistical test. We are using the means and standard deviations provided by the authors to evaluate the sufficiency of their dataset. Although \( a \) \textit{posteriori} power analysis techniques can be used to determine virtually the same information, we chose \( a \) \textit{priori} techniques for their simplicity and to be consistent with the evaluation technique used for index-based studies.) Data sets that violate the assumptions of parametric statistics to a "moderate" degree do not affect the validity of the sample size predictions (Cohen, 1988), so the test can also be used to evaluate studies that used non-parametric (ranked) tests if desired (assuming that non-parametric tests were chosen because the assumptions of parametric statistics were violated).

A second method described by Cohen (1988) was used to determine if an adequate sample size had been used in a third category of studies, those relying on correlation tests to determine if diet was different. It is necessary to understand at this point that if two diets are significantly correlated, they are significantly \textit{similar}. However, finding significance, as mentioned above, necessarily implies that the sample size used in the test was sufficient. Thus, for correlations, we scrutinized studies that failed to find a significant correlation given the sample size used. These studies, therefore, imply that there are differences between the measures of diet being compared. It should be noted that only the null hypotheses have changed, the concept of needing a sufficient sample size to perform a given test, remains the same. The formula is:

\[
\hat{n} = \frac{1573 - 3}{100 \times z^2} + 3.
\]

(Eq. 3)

As in the previous formula, 1573 is from Cohen's table for the \( a \) and \( b \) levels set, and \( z \) is the Fisher transformed correlation coefficient (\( r \)). This test was designed for use with the Pearson product-moment correlation, but violating the assumptions of homogeneity or homoscedasticity to a "moderate" degree does not affect the validity of the power estimates (Cohen, 1988). Therefore, we used the test to evaluate sample size in studies using both Spearman's rank correlation and Kendall's tau.

In each case, because the equations provided are designed to estimate sample size needed for a given comparison, the value \( \hat{n} \) is the number of samples needed at each site, of each species, or each age class. (e.g. to compare two sites, one must have \( \hat{n} \) samples from site one, and the same number of samples, \( \hat{n} \), from site two). Many studies, however, did not obtain the same number of samples for each diet. Therefore, estimated sample sizes were contrasted with the actual sample sizes of the two statistical populations being compared (i.e. \( n_1 \) and \( n_2 \)), resulting in two contrasts per index, parametric statistic, or correlation (\( \hat{n} \) vs. \( n_1 \) and \( \hat{n} \) vs. \( n_2 \)).
Results and Discussion

Precision in Diet Description

When plotted in the order in which guts were analyzed (Fig. 1a, c, e), the cumulative prey curves for all three sample data sets leveled off, indicating that no new prey types are being found in the diet. However, notice that when gut order was randomized, two of the three data sets took many more guts to approach an asymptote (Fig. 1d, f), and in one case, the standard deviation remains quite large (Fig. 1d). This suggests that sample order may indeed cause a bias and its removal by randomizing may provide for a more reliable estimate of the sample sizes needed to describe diet precisely. Although clearly more conservative, randomizing can often provide an additional advantage, as new prey types can suddenly pop up in the diet after the quantification of tens or even hundreds of gut contents. Although not shown, randomization can smooth the terminal end of the curve, as these items, although appearing in the diet at the end of the sampling period, are actually quite rare overall.

![Cumulative Prey Curves](image)

Since a key to high power in a statistical or other type of comparison is low variability, such techniques are important precursors for the statistical techniques described in this paper. However, the body of fish feeding literature seems to be entirely free of such verification. As stated earlier, none of the original ~200 studies utilized cumulative prey curves, or any technique for determining if an adequate number of samples had been collected to precisely describe diet and perform any subsequent comparisons. Cumulative prey curves, however, do not replace the need to assess sample size sufficiency for performing subsequent diet comparisons.

Assessing Sample Size Sufficiency in Making Comparisons

Surprisingly, few of the studies critiqued here had collected enough samples to truly determine that diets were not different. Only one out of 13 indices (Fig. 2) used to indicate no difference between diets had a sample size (n) exceeding that estimated by power analysis (n; Eq. 1) for both diets being compared (n1 and n2; Fig. 2). Thus, for only one index had enough samples been collected to detect the stated difference (effect size), and a difference had not, in fact, been found by the researchers. This finding is validated by \( \hat{n} \) calculated using our second method (Eq. 2).
for our control data set (Table 1), which shows that the prediction of \( \hat{n} \) is consistent with that predicted by more traditional methods when the necessary information (standard deviation) is available. Though the latter method appears to be less conservative, for both comparisons evaluated, actual \( n \) values were still much smaller than \( \hat{n} \) (Table 1).

Figure 2. Plots comparing actual sample size \((n_1 \text{ as } 
abla; n_2 \text{ as } \bullet)\) to predicted sample size \((\hat{n})\) as calculated using Equation 1. In each case, the actual values shown were the sample sizes for two dietary datasets compared using standard indices (points for a single comparison are paired vertically on the graph because \( \hat{n} \) estimated for each dataset is the same for a given comparison; 2 datasets per comparison). The 45° line shown indicates where on the graph, actual sample size matches the predicted sample size necessary to detect a difference in diet. Studies with sufficient sample size have both points above the line shown. Breaks in the axes indicate where a series of numbers has been skipped to facilitate viewing all \( \hat{n} \) simultaneously.

Table 1. Predicted sample sizes \((\hat{n})\) given by Equations 1 and 2 as compared to actual sample sizes collected for the two diets being compared in each study \((n_1 \text{ and } n_2)\). The data are from the control dataset and give only a limited picture of how well the two methods compare in their predictions of \( \hat{n} \).

<table>
<thead>
<tr>
<th>Eq. 1</th>
<th>Eq. 2</th>
<th>( n_1 )</th>
<th>( n_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>13314</td>
<td>3787</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>3054</td>
<td>2201</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>430</td>
<td>253</td>
<td>25</td>
<td>48</td>
</tr>
<tr>
<td>368</td>
<td>197</td>
<td>39</td>
<td>35 * 2</td>
</tr>
<tr>
<td>368</td>
<td>213</td>
<td>35</td>
<td>41 * 3</td>
</tr>
</tbody>
</table>

* not shown in Figure 2

1 first data set shown in Figure 1
2 second data set shown in Figure 1, used to test effect size of 3.7% (a difference not thought to be significant in the original study) in making comparisons with the first data set. Since Equation 1 does not incorporate variance, it will provide the same \( \hat{n} \) for each data set given the effect size of 3.7%.
3 third data set shown in Figure 1, used here to make diet comparisons with the second data set, tested using the same effect size (3.7%, thus giving the same value for Eq. 1), but with the variation (\( s^2 \)) more typical of that data set (Eq. 2). Notice that more samples are needed to make a comparison with the same effect size when using data sets 2 and 3 (Eq. 2; \( \hat{n}=213 \)), which both had only slightly asymptotic cumulative prey curves, than 1 and 2 (Eq. 2; \( \hat{n}=197 \)).
Similarly, two out of 14 studies using t-test/ANOVA to indicate no significant difference between diets (n₁ and n₂) used the sample sizes necessary to detect such a difference (Fig. 3). This means that a difference may have existed, but the analysis simply lacked the statistical power to detect the difference due to low sample sizes. The difference was called non-significant due to this lack of power.

![Graph showing actual sample size compared to predicted sample size](image)

Figure 3. Plots comparing actual sample size (n₁ as ■, n₂ as ●) to predicted sample size (n̂) as calculated using Equation 2. In each case, the actual values shown were the sample sizes for two dietary datasets compared using t-tests or ANOVA (points for a single comparison are paired vertically on the graph because n̂ estimated for each dataset is the same for a given comparison; 2 datasets per comparison). Details are as in Figure 2.

Only three of 14 non-significant correlations used sample sizes that exceeded n̂ predicted by Cohen's second method (Eq. 3; Fig. 4). As above, a significant correlation may have existed, the analysis simply lacked the power to detect it. Only when sample sizes approach n̂ can a powerful correlation be performed and confidence be placed in the conclusions drawn.

![Graph showing actual sample size compared to predicted sample size](image)

Figure 4. Comparisons of actual sample size (n₁ as ■, n₂ as ●) to predicted sample size (n̂) as calculated using Equation 3. In each case, the actual values shown were the sample sizes for two dietary datasets compared using correlation (points for a single comparison are paired vertically on the graph because n̂ estimated for each dataset is the same for a given comparison; 2 datasets per comparison). Details are as in Figure 2.

It should be noted that for most of the studies that we evaluated, the n̂ estimates were only slightly larger than the actual number of samples collected. Thus, had only a few more samples been collected, a difference might have been found. Had no difference been detected with that sample size (n̂), the power of the test would have been extremely high (approaching 100%), and great confidence could have been placed in the results. This result is encouraging since it
suggests that dietary studies are amenable to powerful statistical comparisons at reasonable sample sizes.

In a few of the studies, however, the \( \hat{n} \) estimates were unreasonably high (over 10,000, see Fig. 2). In these cases, the sample size required (\( \hat{n} \)) is considered unachievable and a difference is likely to never be detected. Certainly most studies are undertaken with finite resources, the most limiting of which is usually time. In such cases, to collect the number of samples predicted by power analysis is close to impossible and perhaps ridiculous. The desired effect size is usually extremely small, and this small effect size is strongly influencing the power of the test. Under such circumstances, we maintain that the only recourse is to understand the limitations of the data and state any conclusions with the proper degree of confidence. Johnson (1995) points out that nearly null hypotheses actually are false, it only takes enough samples to find this statistically. The biological relevance of such small differences must be considered (see Yoccoz, 1991). It is necessary to evaluate what such differences between diets would mean for the conclusions of the study relative to its objective and desired interpretations and conclusions. If the effect size is close to or lower than the natural variability in diet found within one or both of the populations, time and samples could be better allocated to different questions.

**Data Analysis**

Finally, since most of the original =200 studies utilized indices, we would like to stress the importance of quantifying the variation around any given index. This is achieved by treating individual guts as sample units (or individual net tows, whatever level is most appropriate in maintaining independence among units), and calculating a mean value for %N or %FO of a prey type across all guts collected in a site, season, or species (whatever the unit of comparison). A similar technique is proposed by Smith (1985), whereby variance is estimated for overlap indices. A second multinomial technique has been described for estimating variance associated with proportions used in indices, however, the assumption of independently and identically distributed datasets must be met (i.e. it is only valid if data are not clustered; see Smith, 1985). The advantage of estimating either variance or standard deviation is there is a clearer picture of the distribution of data, allowing for clearer interpretations.

In selecting which index to choose, consider McDonald and Green's (1983) finding that there may be problems with indices that incorporate several measures of diet, like the IRI. Values like %N and %FO were highly correlated in their study of benthic soft-bottom and demersal fish predators. They argued that compound indices are difficult to interpret and analyze statistically, and require extra, redundant work. Multiple measures may be necessary, however, when prey items differ in size, as the choice of either abundance or weight, for example, may bias the estimate of one prey type's contribution to the diet (Cailliet et al., 1986; McDonald and Green, 1983).

In spite of the increased inferential power of an index with variation, it is not a statistic based on probability, and inferences per se cannot be made (see Platt, 1964). That is, **significant** differences cannot be inferred. Of the original =200 papers reviewed, a shocking proportion lacked inferential statistics of any kind (these generally refer to parametric statistics but non-parametric and non-distributional based statistics have some inferential power), yet stated that diets differed **significantly**.

Some studies did compliment the use of indices with parametric statistics, which is highly recommended if possible. Parametric statistics lend rigor to any study, and when accompanied by traditional methods of power analysis (see Zar, 1988), the level of confidence to place in the conclusions is easily determined. Because of the particular problems associated with proportional diet data (data values expressed as proportions), multivariate techniques have been recommended (see Crow, 1979; Ellison, 1979). Data sets, however, are likely to be heavily weighted with zeros, leading to problems of heterogeneity of variances among statistical populations, and seriously skewed distributions (see Cailliet and Barry, 1977; Underwood, 1981; Clarke and Warwick, 1994). Recently, non-parametric techniques have been described (see Crow, 1979 for
non-parametric multivariate statistics; or Clarke and Warwick, 1994 for ANOSIM), and may prove useful for making comparisons (note: see Potvin and Roff, 1993; Johnson, 1995; Garson and Moser, 1995; Smith, 1995 for a discussion of non-parametric versus parametric statistics).

With these stipulations in mind, it is not surprising that the majority of studies we reviewed relied on indices alone for making comparisons. Indices, as such, have no stated requirements of independent data sets, or distributional nature. However, the conclusions one draws are only as good as the data on which they are based. Thus, we reemphasize the need for precision in the data set, and sufficient sample size for powerful comparisons. In Cailliet and Barry’s (1979) paper in which indices were finally collected into one paper and their use in diet analysis described, they paraphrased Horn (1966) when saying “Indices...are only appropriate in situations in which there is implicit confidence that the proportions of items in each category are adequately characterized” Fifteen years later, we re-emphasize this condition.

There is no substitute for enough samples in reaching conclusions about the available data, whether you are using indices or statistics to make comparisons. In the last decade, power analysis has given us the techniques to evaluate the strength of a statistical test, usually determined directly by sampling effort. Those techniques have now provided us with a means of evaluating non-statistically based comparisons (indices). We strongly recommend their use in determining the strength of conclusions drawn from comparisons of data sets. In addition, we re-emphasize the necessity of using cumulative prey curve technique for determining when enough samples have been collected for precisely describing the data sets to be compared. It is only through such techniques that we will strengthen this area of comparative ecology.

Acknowledgments

We would like to thank M. Foster, J. Harvey, M. Graham, and A. Gibb for assistance with the manuscript and technical advice.

References


Pinkas, L, MS Oliphant, and ILR Iverson 1971 Food habits of albacore, bluefin tuna, and bonito in California waters. Fish Bulletin 152, California Department of Fish and Game, Sacramento, CA.


