14 EXPERIMENTAL DESIGN AND ANALYSIS IN AQUACULTURE

Christopher F. Knud-Hansen

INTRODUCTION

Although aquaculture as a farming practice dates back thousands of years, during the last three decades several simultaneous occurrences have stimulated scientific research of shellfish and finfish cultivation. First, per capita consumption of fish, long appreciated as an excellent source of dietary protein, is increasing across the globe. Second, in countries with rapidly expanding human populations, natural waters no longer meet the growing demand for fish due to overfishing and water quality degradation from poor watershed and waste disposal management (Edwards, 1991). And third, technological advances, such as hormonally induced spawning, sophisticated recirculating systems, and pelleted feeds, have moved the production of commercial species (e.g., the tiger prawn and channel catfish) into large-scale operations.

As aquaculture research rapidly expands in all directions, new species are constantly being considered for grow-out and market potential. Egg production and fry rearing strategies are improving with experimentation. Investigations into semi-intensive and integrated farming management often relate inputs (e.g., manures, cassava leaves, and urea) to water quality, primary production, and fish yield. Identification of nutritional requirements has helped develop more efficient formulated feeds for intensive fish culture systems. Progress has been made. But as the irony of science would have it, from each question answered springs forth more questions posed.

Identifying precisely what questions to ask is only slightly easier than devising the proper methodologies to answer them. A well-designed experiment will yield observations or measurements under controlled conditions and hopefully will expand existing knowledge of how the system works. With limited resources (human, material, time, and capital), it is essential for any investigation to determine not only what data to collect but what data not to collect. Objectives, hypotheses, treatments, sources of experimental error, and the types, methods, and frequency of measurements must all be clearly defined and understood in order to optimize research grants and to be in a position to reasonably request more. The limitation of available funds for research necessitates focused, cost-efficient experimental designs.

Knowledge of statistics, the foundation of experimental design, benefits the researcher in three very important ways. First, it gives the researcher the analytical skills to effectively test hypotheses, question assumptions, and tease apart relationships. Science is more than collecting data; it is knowing which data need to be gathered to effectively answer a specific problem. Too often piles of data are passed on to a statistician with the desperate plea "analyze this and tell me if anything is significant!" Sometimes the only "significant" observations are the sizes of piles and the anxiety over what to do with them, now that so much time, effort, and money had been spent in their creation. There must be thousands of diel temperature, dissolved oxygen, alkalinity and pH measurements just waiting, … and waiting.

0-56670-274-7/97/$0.00+$.50

Second, statistics enable the researcher to have an objective measure of confidence in her or his results and interpretations. Some people believe something is true because they see it; others see it because they believe it is true. If you feel 99% sure that using a brand-named high-priced feed does not give any better fish yields than the bargain variety feed, it is easier to convince others of your conclusions when you let the numbers do the talking.

Third, statistical knowledge gives the scientist the power to critically review the literature, whether it be to discover design flaws and thereby alter stated conclusions, to enhance existing analyses, or to confirm in one's own mind the appropriateness of a given interpretation.

Unfortunately, most biologists' first (and often last) introduction to statistics took place in a large lecture hall with a professor at the lectern droning in uninspiring Greek. There was little joy memorizing formulas and grunting through calculations, which were difficult to relate to and never quite made much sense. Surviving such threats as pooled variances and mean square errors was the main objective. That the only mathematical talent one needs to gain sufficient statistical knowledge to do quality science is the ability to add, subtract, divide, and multiply is rarely made clear at the onset of introductory statistics courses.

This chapter attempts to further develop the design and analytical skills of aquaculturists who, on the average, tend to have more fear than background in mathematics/statistics. Major concepts of descriptive and inferential statistics are covered without plodding along the traditional well-worn "cookbook" paths found in most statistics books. Experimental design and data analysis are two sides of the same coin, and many aquaculture researchers focus on data analysis without ever appreciating the connection. Two analogies may help illustrate the importance of experimental design in research. First, the field of statistics can be thought of as a language. Methods of data analysis are like the words, but it is knowledge of experimental design that allows the scientist to use the words efficiently and to their fullest potential. Researchers who memorized words (i.e, statistical formulas) in statistics courses rarely learned how to put the words together. Scientific eloquence requires a knowledge of both data analysis (vocabulary) and experimental design (grammar). A second conceptual analogy is the toolbox. Data analyses are like tools, but experimental designs represent the knowledge of how, when, and where these tools should be used for maximum efficiency and utility. As there are already enough "cookbooks" and software packages to do the mechanics of data analyses, this chapter emphasizes the concept and philosophy of experimental designs. Except as foundational requirements, little emphasis is placed on mathematical computations of statistical theory. Where calculations are presented, emphasis will be on when, where, and why, and not how.

The practioner's approach to controlled experimentation used in this chapter is meant to complement the reader's own more theoretical statistical literature. Sections on "Basic Concepts" and "Scientific Approach to Aquaculture Research" provide basic statistical concepts and discuss hypothesis testing, respectively. Sections on "Treatments" through "System Characterization and Monitoring" show how to choose treatments, reduce experimental error through appropriate treatment allocation, and how to best characterize the experimental system. The section, "Data Analysis," discusses how data generated from the various designs are analyzed, while the section, "Quality Control," presents ways to improve overall quality control of data collection, processing, and evaluation. The "Conclusion" section briefly addresses issues regarding publication of research data. Standard calculations and equations presented here are not cited to any particular source, as they are found and more thoroughly discussed in most general statistics books. Among the basic reference books relied upon for this chapter are Heath (1970), Parker (1979), Steel and Torrie (1980), and Baily (1981). This chapter, however, does not present every possible analytical technique relevant to aquaculture research. Intelligent use of more complicated procedures, such as multivariate analysis, requires a solid foundation in basic research and experimental design theory. A critical examination of current aquaculture literature indicates that such a foundation is often lacking. This chapter tries to promote quality research and build that foundation by communicating with words and examples in a way fellow researchers can appreciate.

BASIC CONCEPTS

Although it is assumed that the reader already has had some exposure to statistics, a quick refresher of basic terminology may be useful. The field of statistics examines methods and procedures for collecting, classifying, processing, and analyzing data for the purpose of making inferences from those data. Statistics is an inductive process where attempts to understand the whole are based on examining representative parts through sampling and experimentation (Finney, 1988b).

The first task for the researcher, therefore, is to define what she means by the "whole." In statistical terms the "whole" is defined as the ***population***, and can be as broad or narrow as is scientifically reasonable. A population could be all the catfish raised in a particular pond, or all the catfish raised in all the earthen ponds of similar dimensions situated within tropical latitudes, or raised in concrete recirculating freshwater systems maintained within a specified temperature range. Characteristics of populations, such as mean fish weight, are called ***parameters*** (from the Greek words ***para*** and ***metron*** which mean "beyond measure"). Parameters are fixed values without variability. If there are 10,000 fish in a pond, there is only one true mean weight at harvest for those 10,000 fish.

For populations that are either too large, indefinite, or indeterminate (e.g., all tropical earthen ponds), it is generally both impractical and impossible to determine parameter measurements (hence the name parameter). Always remember that it is the population, however you have defined it, that you want to understand. In order to gain this understanding ***samples*** are taken from the population. Characteristics of the samples, such as sample mean fish weight, are ***statistics***. Unlike parameters, statistics can vary with each sample. Sample statistics estimate population parameters. For example, the sample mean estimates the true population mean, the sample variance estimates the true population variance, and so on. Since we do not know the exact values of the population parameters (if we did there would be no need to take samples), sample statistics should always be given with some indication of their variability and level of confidence of how well they represent corresponding parameter values.

The starting point for determining variability and levels of confidence is knowing the underlying population distribution. Frequencies of measurements or observations often follow a ***normal distribution***, with values distributed symmetrically on either side of the true mean (Figure 1). This normal distribution is described by two parameters, the true arithmetic mean (μ) and the true

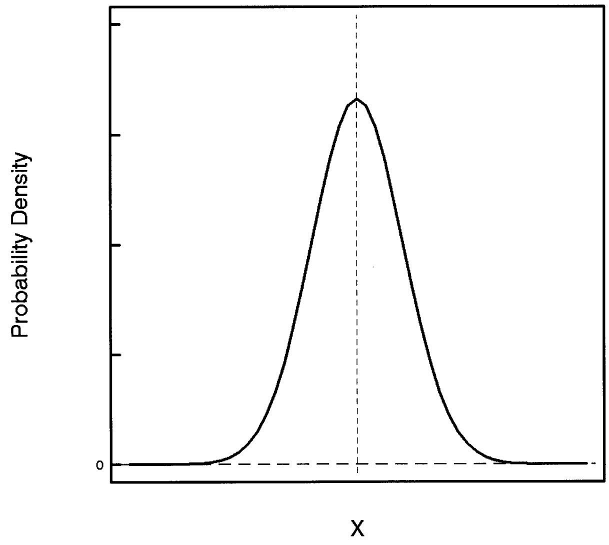


Figure 1. A typical normal distribution curve of X illustrating the characteristic "bell-shaped" curve, symmetrical on either side of the mean.

variance (σ^2). If n = the total number of observations and x_i = the i*th* observation, then μ = sum of all values divided by n, or

$$\mu = \frac{\sum_{i=1}^{n}(x_i)}{n} \tag{1}$$

The ***variance*** measures the dispersion of individual values in relation to the mean, and equals the average square of the difference between each observation and the true mean.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \tag{2}$$

As can be seen from Equation 2, the greater the dispersion (i.e., more observations farther away from μ) the greater the variance. If all the observations were the same, then the variance would be zero. The variance can be calculated using an alternative equation, which is both simpler to use and illustrates two major components of the variance, the “Sum of Squares” and the “Square of Sums.” The variance then becomes the sum of squares minus the square of sums divided by n, or

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n} \tag{3}$$

Since variances are in squared units, a more practical indicator of population variability is by taking the square root of the variance. This value, called the ***standard deviation*** (σ) of the population, now has the same units as the observations from which it was calculated.

When a sample is taken from a population, the parameter values of μ , σ, and σ^2 are estimated by the sample statistics ($\bar{x}$, s, and s^2), respectively. Calculations are similar, except $n - 1$ is used instead of n when calculating the s and s^2.

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i)}{n} \tag{4}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n - 1} \tag{5}$$

When using electronic calculators or computers to determine variances and standard deviations, you must determine whether they use n or $n - 1$ in these calculations. Most experiments are designed to represent larger (often theoretical) populations. For example, a simple experiment comparing shrimp growth using feed A vs. feed B may use three ponds for each treatment. If you are only interested in those specific ponds, then the variance is calculated using n (= 3). If, however, data are to be used for inferring possible results in other nontested ponds (which is nearly always the case), then the three ponds per treatment are really samples of the indeterminate number of shrimp ponds that may utilize either feed A or B in the future. Since the three ponds per treatment are now considered samples of a larger population of ponds, the variance of each treatment should be calculated using $n - 1$ (= 2) instead of n. When n is small, the resulting error in estimating true population variance can be great (see Hypothesis Testing in the section on Scientific Approach to Aquaculture Research). For larger sample sizes (i.e., $n > 35$), however, using $n - 1$ instead of n loses computational relevance.

Another measurement of variability is the ***standard error*** (SE). The standard error is the standard deviation of the distribution of a statistic for a random sample of size n. Let's say, for example, that you want to know the average weight of the 10,000 tilapia being raised in one of your ponds. You collect 25 fish, weigh them individually, and calculate the $\bar{x}$ and s^2. Then you collect 25 more fish and do the same thing (most likely getting different values for $\bar{x}$ and s^2), and repeat this process 20 more times. If you plot your means and variances on a graph, you will find they too have a normal distribution. The standard deviations of these distributions of means and variances represent their respective standard errors. In aquaculture we are often concerned primarily with the SE of the mean. And rather than taking a series of samples, the SE can be calculated dividing the standard deviation (s) by the square root of n.

$$\text{SE of } \bar{x} = \frac{s}{\sqrt{n}} \tag{6}$$

Note the distinction between standard deviation and standard error. Standard deviation describes the variability of ***observations*** about a sample mean, while standard error describes the variability of ***means*** about a sample mean. Whenever the scientific interest concerns comparing sample means, therefore, report mean values with standard errors (e.g., $\bar{x} \pm 1$ SE), *not* with standard deviations. On the other hand, population variability should be reported with a standard deviation.

Also note some important characteristics of the standard error. It is proportional to the standard deviation, and inversely proportional to n. That is, as the sample size increases, the variability of sample means gets smaller. Notice too that the size of the population is totally unrelated to the standard error of the mean. As long as observations are taken randomly, how well our sample mean estimates the true mean depends *only* on the sample standard deviation and the number of observations taken for that calculation. Some field manuals erroneously recommend making sample sizes equal to 10% of total population size. But as indicated by Equation 6, population size is totally irrelevant with regard to estimating the true population mean and variance.

The sample mean estimates the true mean, and the standard error describes the variability of that estimation. This variability can be conveniently expressed in terms of probabilities by calculating ***confidence intervals*** (CI). For example, assume a sampling of 25 catfish (in the pond of 10,000) gives a mean weight of 350 g/fish and a standard deviation (using $n - 1$, Equation 5) of 75 g/fish. We know that the true mean is unlikely to be exactly 350 g/fish, but what is the range of means within which we can say with a certain level of confidence that the true mean falls? First we calculate the standard error (Equation 6), which gives the variability of the theoretical distribution of sample means. Here, the SE of the mean is $75\,(\sqrt{25})^{-1}$, or 15 g/fish. If we make our confidence interval ±1 SE, we are 68% confident that the true population mean is between 350 + 15 g/fish and 350 – 15 g/fish. In other words, there is a 0.68 probability (P) that the true mean is somewhere between 335 and 365 g/fish. The upper and lower mean values are called the ***confidence limits*** (CL) of the mean.

In aquaculture, as with most fields of science, it is common to give sample means with 95% confidence ($P = 0.95$) intervals. This interval is calculated by simply multiplying the SE by a t value (Equation 7) taken from a Student's t Table located in the Appendix of virtually every statistics book on the market. On the left-hand side of the table is a column of ***degrees of freedom*** (*df*), which equals *n* – 1. The horizontal heading values are probabilities, often ranging from 0.5 to 0.001. If we make our desired level of confidence = α, then the probabilities in the Student's *t* Table equal 1 – α. Within the table matrix are *t* values. In our example we have *n* – 1 or 24 *df*, and the *t* value at *P* = 0.05 (= 1 – 0.95) is 2.064. So the 95% CI and CL equals:

$$\begin{aligned}\text{CI at } P = \alpha,\text{ or } \alpha(100)\% &= \left(t_{(1-\alpha)}\right)(\text{SE}) \\ &= (2.064)(15\text{ g/fish}) = 31.0\text{ g/fish}\end{aligned} \tag{7}$$

$$\begin{aligned}\text{CL at } P = \alpha \qquad &= \bar{x} \pm \left(t_{(1-\alpha)}\right)(\text{SE}) \\ &= 350 \pm 31.0\text{ g/fish} = 319.0 - 381.0\text{ g/fish}\end{aligned} \tag{8}$$

Therefore, we are 95% confident that the true mean of the 10,000 catfish is somewhere between 319.0 and 381.0 g/fish. In other words, there is a probability (*P*) of 0.05 that the true mean is less than 319.0 g/fish or greater than 381.0 g/fish. Since the normal distribution is symmetrical, *P* = 0.025 that μ < 319.0 g/fish, and *P* = 0.025 that μ > 381.0 g/fish.

Notice in the Student's *t* Table that *t* values increase as we demand greater and greater confidence. For instance, if we wanted to be 99.9% confident, our CI would be (3.745) (15 g/fish) or 56.2 g/fish. Notice also that *t* values decrease as the sample size increases. If we had sampled 250 fish instead of 25, assuming the same SE, the 95% CI would decrease to (1.960) (15 g/fish) or 29.4 g/fish. The mean with *n* = 250 is not necessarily more accurate (i.e., closer to the true mean) than the mean with *n* = 25, but the CI about the mean gets smaller. In actuality, increasing *n* should decrease the SE (see Equation 6) as well as the *t* value, thereby increasing the precision of estimating the true mean.

The last descriptive analysis discussed in this section is the ***coefficient of variation*** (CV). The CV is a relative, unitless measure of variation that describes as a percent the relationship between the sample standard deviation and the sample mean (Equation 9).

$$\text{CV} = \frac{s(100)}{\bar{x}} = \text{standard deviation as percent of the mean} \tag{9}$$

Researchers often report CVs, but there is very little analytically that can be done with them. Since their values are entirely relative, there is no such thing as an inherently good or bad CV. They can be used, however, to check consistency of data collection. For example, standard deviations of mean fish weights tend to increase proportionally with mean values, so CVs are often fairly consistent whether the fish mean weight is 50 g/fish or 500 g/fish. A relatively low or high CV may indicate a problem with the sampling technique.

SCIENTIFIC APPROACH TO AQUACULTURE RESEARCH

Introduction

The starting point for scientific research is the ***hypothesis*** (H), which is an educated guess consistent with observations from a "natural" system (i.e., natural history) and/or existing scientific literature. The hypothesis predicts how a system will respond when manipulated. Predictions may be anything, including nonquantitative effects or mathematical relationships.

To test hypotheses, experiments are conducted under controlled conditions. Experiments must be designed carefully so that the nature, timing, frequency, and accuracy of variable measurements account for all important sources of variation. Variation not accounted for in data analysis can be referred to interchangeably as ***noise***, ***experimental error***, or ***residual error*** (often represented by the Greek letter ε, or by R). If the noise is great enough, it may obscure actual significant relationships. After data collection and analysis, experimental results are compared to predicted hypothetical results. If the data analysis does not support the original hypothesis, further hypothesis modification may be necessary before designing more focused experiments and repeating the scientific process.

Hypothesis Testing

It is important to keep in mind that although experimental conclusions may have matched predicted results, we still cannot say that the experiment "proved" a causal relationship. It is usually more constructive to analytically test the ***null hypothesis*** (H_0) rather than the hypothesis itself. If the hypothesis predicts that a particular relationship exists, the corresponding null hypothesis predicts that this relationship does not exist. From data analysis a probability (P, where $P = 1.0$ is absolute certainty) is obtained, which equals the probability that the observed relationship or treatment differences did not just happen by chance. This probability forms the basis for accepting or rejecting null hypotheses.

For example, assume that a researcher believes (i.e., hypothesizes) that her homemade formulated pelleted feed will give greater tilapia yields than a high-priced commercial variety. The null hypothesis to be tested is that the two types of feed will not produce significantly different yields of tilapia. To test her hypothesis, the researcher conducts an outdoor tank experiment with four tanks for each type of feed. Fish growth measurements after 3 months gave mean weights (±1 SE, $n = 4$) of 157.3 ± 10.1 g/fish and 189.6 ± 9.8 g/fish for the homemade and commercial feeds, respectively. Should the null hypothesis be accepted, or is the difference between the two means so great that differences could not have happened by chance and therefore the null hypothesis should be rejected?

The first step is to understand what is meant by significance. Significance can be viewed from two perspectives, ***"true" significance*** and ***statistical significance***. True significance is to statistical significance what a parameter is to the corresponding statistic. Similar to a parameter, whether the difference between sample means is truly significant (i.e., $P = 1.0$) is not really known. Statistical significance, therefore, provides a level of confidence or probability < 1.0, by which we can infer that a truly significant difference/relationship does or does not exist. Statistical significance is based on probabilities determined from experimentation. In aquaculture, a $P < 0.05$ that the observed relationship ordifference between means, or whatever happened just by random chance is the most common level of statistical significance. That is, we are more than 95% confident ($P > 0.95$) that there is a true significant relationship. This 0.05 probability is the dividing line between statistical significance and statistical nonsignificance in many other scientific disciplines as well.

It is important to remember, however, that there is no magic probability for statistical significance, and it is the researchers themselves who determine that level. Statistical significance is actually a highly flexible concept. For example, scientists working on a cure for AIDS may consider $P < 0.40$ an adequate limit for accepting statistical significance (i.e., there is a $P > 0.60$ that a certain chemical treatment is effective against AIDS). On the other hand, scientists working on a cure for the common cold may consider a probability of 0.0001 the highest acceptable probability that the alleged cure does not cause lethal side effects. The scientist's real objective is to understand how the system works, not to find mindless and mechanistic statistical significance (see Yoccoz, 1991).

In our tilapia feed example above, the researcher has four possible outcomes based upon her decision whether to accept or reject the null hypothesis and what is actually the truth. These four outcomes are best understood by comparing the statistical vs. true significance of her experiment (Figure 2). She will be correct if she rejects the null hypothesis when the null hypothesis is false (saying the observed difference is statistically significant when it truly is), or if she accepts the null hypothesis when the null hypothesis is true (saying there is no statistically significant difference when there truly is no difference). With the two other outcomes she will be incorrect. ***Type I error*** occurs when, based on statistical significance, she rejects the null hypothesis and claims a truly significant difference when there is none. In contrast, if she says there is no truly significant difference when there actually is one (accepting the null hypothesis when it is true), this is called a ***Type II error***. The probabilities of committing these two errors are typically represented by the Greek letters α and β, respectively.

		Null Hypothesis	
		True	False
Your Decision to:	Reject	Type I Error No true significance Statistical significance	Correct True significance Statistical significance
	Accept	Correct No true significance No statistical significance	Type II Error True significance No statistical significance

Figure 2. A diagram showing a results' matrix for hypothesis testing based on the researcher's decision to reject or accept the null hypothesis when the null hypothesis is either true (i.e., no true significance) or false (i.e., true significance).

Since much of science involves determining probabilities whether or not observed hypothesized treatment effects, relationships, etc. are truly causal, we are never certain when making our conclusions that we are not committing a Type I or Type II error. If we want to reduce the possibility of committing a Type I error, we can increase our level of statistical significance from 95 to 99% confidence by reducing α from 0.05 to 0.01. In other words, we will not claim true significance for a relationship or difference between means unless there is a $P < 0.01$ that our observations happened by chance. By doing this, of course, we also increase the probability of committing a Type II error. Scientists generally prefer to decrease the possibility of committing a Type I error because it is more professionally prudent to miss a significant relationship when there was one than to claim significance when it was not there (e.g., claims of energy released from cold fusion). Nevertheless, missing a significant relationship when there was one (i.e., committing a Type II error) should also be avoided. As the probability of committing a Type II error equals β, the probability of *not* committing a Type II error equals 1 – β. This latter probability is know as the ***statistical power*** (Cohen, 1988). This is an important concept, particularly when analyzing nonsignificant results. Statistical power calculations involve treatment means, experimental variability, and the number of replicates per treatment. For example, statistical power analysis can determine if there were too few replicates to reveal any significant differences between treatment means given the closeness of the mean values, extent of experimental variability, and the desired power (Cohen, 1988). A probability of 0.80 (i.e., $\beta = 0.20$) has been proposed as the minimum acceptable statistical power, and may be higher depending on the financial consequences of the results (e.g., environmental impact assessments) (Searcy-Bernal, 1994; Cohen, 1988). Searcy-Bernal (1994) provides a clear and concise application of power analysis to aquaculture research, and thorough reading is highly recommended (see also Peterman, 1989).

Using a $P < 0.05$ (i.e., $\alpha = 0.05$) is generally an acceptable balance between committing Type I or Type II errors. But whether or not the researcher subjectively thinks a difference is significant is secondary to the actual probabilities generated by the analysis. Therefore, always report probabilities with statistical analyses. A relationship with a $P < 0.10$ may not be statistically significant, but it may be truly significant though not "seen" statistically because of too few replicates or too much noise in the experiment. With probabilities less than 0.01, it is often better to report probabilities to the nearest number other than zero (e.g., $P < 0.005$, $P < 0.001$, $P < 0.0008$, etc.). Reporting probabilities allows the reader to decide for him or herself just how much the data support or refute the tested null hypothesis and not to rely only on an author's claim of "highly significant" differences.

With the feeding experiment described above, the researcher conducted a *t*-test (see the section on Data Analysis) to determine whether the two means (157.3 ± 10.1 g/fish and 189.6 ± 9.8 g/fish for the homemade and commercial feeds, respectively) were significantly different from each other. The analysis gave a probability of $0.2 < P < 0.3$ that the commercial feed was truly no better than the homemade feed; or in other words, there is >70% chance that the commercial feed was truly better than the homemade feed. Although mean weights were not statistically significant, she may wish to review her experimental methodology (e.g., confirm all tanks were environmentally equal, all fish came from the same stock, etc.), calculate the statistical power, and perhaps rerun the experiment with more replicates before accepting the null hypothesis that the two feeds were not really different with respect to tilapia growth.

TREATMENTS

Introduction

After a general hypothesis has been formulated, the next step is to identify appropriate treatments necessary to test the resultant null hypothesis. Treatments are selected because the researcher hypothesizes they will (or will not) make a difference to a particular response variable(s). Remember, the main scientific objective is to identify and quantify sources of variability in response variables. In aquaculture, response variables are often related to aspects of growth, reproduction, productivity, and water quality. Understanding variability and how to manage it is at the heart of experimental science. Variability does not just happen, it happens for a reason, and the job of aquaculture researchers is to find out both how and why.

To understand the role of treatments, it is useful first to examine a system without treatments. For example, anyone who has worked with earthen ponds knows that if you have 16 tilapia ponds stocked and fertilized identically, you will have 16 different yields at harvest. This result can be expressed by the following model:

$$Y_i = \mu + \varepsilon_i \qquad (10)$$

where Y_i = yield in the ith pond, μ = the overall mean of all ponds, and ε_i = the residual (or "experimental error") for the ith pond.

To better understand residuals, let us assume that for these 16 ponds the overall mean for the tilapia harvest was a net fish yield (NFY) of 23.5 kg/ha/d, and ponds 3 and 7 had yields of 19.9 and 25.1 kg/ha/d, respectively. As mentioned earlier, the residual represents the sum of all unidentified sources of variation. If the overall mean is thought of as the predicted value of the experiment, then the residual is simply the observed value minus the predicted value. The NFY residuals for ponds 3 and 7 would then be –3.6 and +1.6 kg/ha/d, respectively. Although residuals will be discussed throughout this chapter, the idea that the residual equals the observed value minus the predicted value will be particularly evident with regression analysis discussed in the section on Data Analysis.

At this point, however, the researcher's primary objective is to improve the predictability of μ by decreasing ε. Decreasing ε can be achieved by focusing research on *manageable* sources of variation, and evaluating factors that influence these sources. For example, Nile tilapia will grow proportionally (within limits) to the rate of natural food production (Knud-Hansen et al., 1993). The issue then becomes identifying manageable factors that influence the variability of natural food production, which in turn affects the variability of tilapia production.

More generally, treatments are selected for their hypothesized effect (or relationship) with the response variable. A far from exhaustive list of treatment possibilities for affecting tilapia yields includes stocking density, pond sediments, natural food availability, temperature, inorganic turbidity in the water column, and rates of nutrient input. Treatment effects on Y can be shown by expanding the above model (Equation 10) as follows:

$$Y_i = \mu + \tau_i + \varepsilon_i \tag{11}$$

where τ_i = deviation due to treatment i. Now there are two sources of variation in the equation, the treatment and the residual. The hypothesis tested here is that τ is a significant source of variation of Y. The section on Data Analysis discusses how to quantitatively test this hypothesis.

Since one of the purposes of science is to improve upon existing knowledge or understanding of a particular system, the choice of treatment(s) to test a given hypothesis must be carefully thought out. The subsections below describe different aspects and guidelines for determining basic treatment selection, depending on the type of hypothesis the researcher wishes to test.

Unstructured, Structured, and Factorial Experiments

The first cross-road in choosing treatments is to decide whether the experiment will be ***unstructured*** or ***structured***. Unstructured experiments are those in which the researcher wishes to compare, in a matter of speaking, apples with oranges. One treatment cannot be expressed as a function of another. In aquaculture, examples of such experiments include comparing different culture species (or genetic varieties of the same species) under identical conditions, or comparing the efficacy of different brands of shrimp feeds. Often the researcher's primary interest is to determine which of the lot performs the best and whether it is significantly better than the rest. Such treatments are unstructured because there is no quantitative way to rank or arrange them or to demonstrate relationships between treatments. For this reason results from unstructured experiments are presented as vertical rankings of data (in either ascending or descending order) and are not based on any logical order of treatments. It is worth mentioning here and will be repeated in the section on Data Analysis that multiple range tests should be used *only* with unstructured experiments.

There are two types of unstructured experiments, both based on the nature of the researcher's hypothesis. The first type is when he tests the null hypothesis that none of the treatments (e.g., feed types, species of fish) produces a response significantly different from any other treatment. Screening trials typify this type of experiment, where frequently the objective is to determine whether the best is significantly better than the rest. In the second type of unstructured experiment, the researcher wishes to compare unrelated treatments to some benchmark treatment. This benchmark could be, for example, the local strain of tilapia or the main commercial brand of feed used at the research station. In a sense this benchmark treatment could be thought of as the "control." But in actuality, the researcher is testing the null hypothesis that none of the other treatments produces significantly different results than the benchmark treatment. The distinction between the two types of unstructured experiments determines how the data are analyzed (see the section on Data Analysis).

Structured experiments are those in which there is a logical ordering of quantitatively or qualitatively defined treatments. Each treatment can be expressed as a function of another. Simply stated, if you can draw a line graph of the results with treatments aligned along the x-axis, then the experiment was structured. The primary objective with structured experiments is to determine *relationships* and/or *treatment interactions*, and *not* just to find which treatment gave the "best" results. For this reason, among many others discussed in the section on Data Analysis, multiple range tests should *never* (repeat *never*) be used to analyze structured experiments.

The two most common types of experiments used to evaluate relationships are those examining changes over time and those examining response changes with increasing levels/concentrations of a treatment variable. The latter is also known as a dose–response experiment. The purpose of such experiments is to estimate and identify trends. The hypotheses tested generally reflect a mathematical relationship (e.g., linear, quadratic, exponential, asymptotic, etc.) that the tested relationship hypothetically demonstrates. Regression analysis (described below in the section on Data Analysis) is normally used to test the null hypothesis that there is no relationship over time or with variable treatment dosages.

Structured experiments are among the most common in aquaculture. Examples of the "dose–response" variety include fertilization experiments (e.g., relationship between increasing nutrient input levels and yield), stocking density experiments (e.g., relationship between stocking density and yield), hatchery studies (e.g., relationship between flow rate and hatching success), and feeding trials (e.g., relationship between lipid content in feed and its digestibility). The key concept is relationship, and for that reason such investigations are particularly useful for model building. Note that each treatment level is a function of another (e.g., stocking densities of 1 m^{-2}, 2 m^{-2}, 3 m^{-2}, and 4 m^{-2} can be expressed as multiples of 1, 2, 3, and 4 times 1 m^{-2}, and therefore the experiment is structured.

When the experimental objectives include testing the hypothesis that two treatment variables interact with each other to produce a nonadditive response, a ***factorial*** design is chosen. To illustrate a positive interaction, consider a nutrient-poor pond. If you add only phosphorus (P), you may get a little algal growth. If you add only nitrogen (N), you may again get some minor algal response. But if you add P and N together, you may get an algal bloom, a response much greater than the individual N and P responses added together. Our model equation would now be as follows:

$$Y_{ij} = \mu + N_i + P_j + (NP)_{ij} + \varepsilon_{ij} \tag{12}$$

where N_i = ith level of nitrogen treatment, P_j = jth level of phosphorus treatment, and NP_{ij} = deviation due to the interaction between N and P.

N and P in the above example are called ***factors***. Within each factor there may be two or more ***levels*** of that factor. What characterizes a factorial experiment is that every factor is represented in all treatments, and the treatments represent every possible factor and level combination. A simple factorially designed experiment would have two factors (e.g., N and P input), and two levels of each factor (e.g., no input and input). This experiment is called a 2 by 2 factorially designed experiment because there are 2 levels (input amounts) of 2 factors (N and P). There would be a total of 2 × 2, or four different treatments in this experiment: (1) no N input, no P input, (2) N input, no P input, (3) no N input, P input, and (4) N input, P input (Figure 3). Although there are only four treatments, three different null hypotheses are tested:

1. There is no significant response with the addition of N.
2. There is no significant response with the addition of P.
3. There is no significant interaction between and N and P with regard to the response variable.

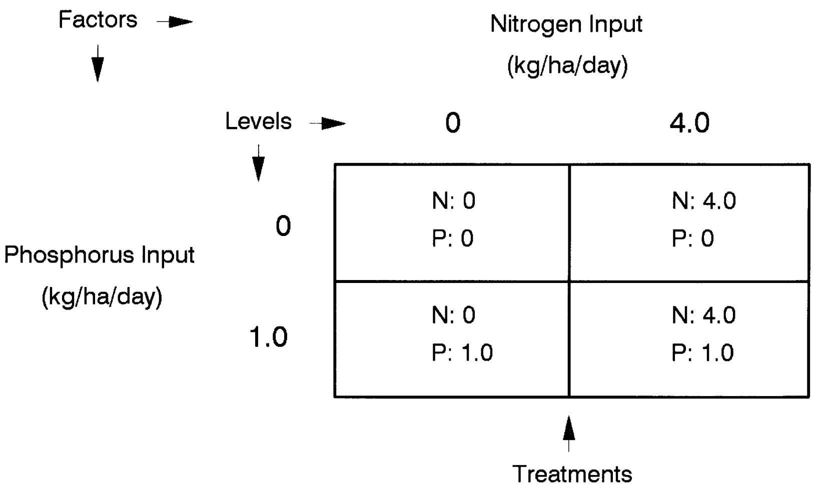


Figure 3. A diagram illustrating the four treatment combinations of a 2 by 2 factorially designed experiment with two levels (hypothetical input rates) for two factors (nitrogen and phosphorus input).

More complicated factorial experiments can reveal a tremendous amount of information because of the great number of hypotheses tested. For example, assume a researcher wants to test for any interaction between lipid and protein concentrations in shrimp feeds as reflected in the growth of two different varieties of shrimp. In the experiment she tests three levels (i.e., concentrations in the feed) of lipid and four levels of protein. As a factorial experiment, there are three factors (shrimp variety, lipid concentration in feed, and protein concentration in feed), with two levels of shrimp, three levels of lipid, and four levels of protein. The experiment is, therefore, a 2 by 3 by 4 factorially designed experiment, with a total of 2 × 3 × 4 = 24 treatments (Figure 4).

	Shrimp Variety					
	1			2		
	Lipid			Lipid		
Protein	1	2	3	1	2	3
1	Shrimp 1 Lipid 1 Protein 1	Shrimp 1 Lipid 2 Protein 1	Shrimp 1 Lipid 3 Protein 1	Shrimp 2 Lipid 1 Protein 1	Shrimp 2 Lipid 2 Protein 1	Shrimp 2 Lipid 3 Protein 1
2	Shrimp 1 Lipid 1 Protein 2	Shrimp 1 Lipid 2 Protein 2	Shrimp 1 Lipid 3 Protein 2	Shrimp 2 Lipid 1 Protein 2	Shrimp 2 Lipid 2 Protein 2	Shrimp 2 Lipid 3 Protein 2
3	Shrimp 1 Lipid 1 Protein 3	Shrimp 1 Lipid 2 Protein 3	Shrimp 1 Lipid 3 Protein 3	Shrimp 2 Lipid 1 Protein 3	Shrimp 2 Lipid 2 Protein 3	Shrimp 2 Lipid 3 Protein 3
4	Shrimp 1 Lipid 1 Protein 4	Shrimp 1 Lipid 2 Protein 4	Shrimp 1 Lipid 3 Protein 4	Shrimp 2 Lipid 1 Protein 4	Shrimp 2 Lipid 2 Protein 4	Shrimp 2 Lipid 3 Protein 4

Figure 4. A diagram illustrating the 24 treatment combinations of a 2 by 3 by 4 factorially designed, hypothetical shrimp feed experiment. There are three factors of shrimp variety, and feed concentrations of lipid and protein, with two, three, and four levels for the three factors, respectively.

This shrimp feed experiment tests the following null hypotheses for *each* response variable (e.g., shrimp growth, productivity, etc.):

1. Null hypotheses based on factors:
   a. No difference in response between the two shrimp types (results pooled over all protein-lipid combinations)
   b. No difference in response between the three lipid levels (results pooled over all shrimp-protein combinations)
   c. No difference in response between the four protein levels (results pooled over all shrimp-lipid combinations)
2. Null hypotheses based on nested designs within the factorial experiment at different levels of each factor:
   a. No difference in response between the two shrimp varieties at
      1. Lipid level 1 and protein level 1
      2. Lipid level 1 and protein level 2
      3. Lipid level 1 and protein level 3
      4. Lipid level 1 and protein level 4
      5. Lipid level 2 and protein level 1
      6. Lipid level 2 and protein level 2
      7. Lipid level 2 and protein level 3
      8. Lipid level 2 and protein level 4
      9. Lipid level 3 and protein level 1
      10. Lipid level 3 and protein level 2
      11. Lipid level 3 and protein level 3
      12. Lipid level 3 and protein level 4
   b. With each shrimp variety, there is no relationship between the response variable and
      1. Increasing lipid concentration when protein is kept at level 1
      2. Increasing lipid concentration when protein is kept at level 2
      3. Increasing lipid concentration when protein is kept at level 3
      4. Increasing lipid concentration when protein is kept at level 4
      5. Increasing protein concentration when lipid is kept at level 1
      6. Increasing protein concentration when lipid is kept at level 2
      7. Increasing protein concentration when lipid is kept at level 3
3. There is no two-way interaction as reflected in the response variable between
   a. Lipid and protein concentrations (pooled over both shrimp varieties)
   b. Shrimp variety and lipid concentration (pooled over all protein levels)
   c. Shrimp variety and protein concentration (pooled over all lipid levels)
4. There is no three-way interaction between shrimp variety, lipid concentration, and protein concentration

Schematic representations of treatments (e.g., Figure 4) greatly facilitate identifying testable hypothesis. An advantage of factorially designed experiments is that unusual treatment combinations may reveal relationships or interactions not previously considered. On the other hand, some interactions may not make any biological sense. A statistically significant three-way interaction between shrimp variety, lipid, and protein concentration would probably defy biological understanding. In reality, however, only a few of the above possible hypotheses may have real importance to the researcher. By identifying testable hypotheses *prior* to conducting the experiment, the researcher can adjust treatment levels to ensure the primary hypotheses will be adequately tested.

Note that the shrimp feed experiment has both unstructured and structured components when levels of each factor are analyzed. The factor of shrimp variety is unstructured since the two varieties are distinctly different with no definable relationship (like apples and oranges). Of the lipid and protein factors, however, each has different levels related to each other; i.e., each level is a fraction of another. Both protein and lipid treatments are of the dose–response kind; therefore, that part of the experiment (and analysis) is considered structured.

Factorially designed experiments are probably the most complicated experiments aquaculture scientists will encounter. A review of the literature shows that most researchers employing factorial experiments have grossly underutilized the analytical potential of their research. A great deal more is yet to be learned from existing data sets, and hopefully this brief discussion will induce interested scientists to review their favorite statistics book with a renewed purpose.

Scope of Experiments

Existing scientific knowledge together with the researcher's ambitions (as well as available funding) usually determine the scope of research. The two most common problems are trying to "prove" too much and not developing an overall game plan. Not surprisingly, these problems tend to go hand-in-hand.

The researcher who tries to establish a major scientific principle with a single experiment or analysis almost always ignores the myriad of other relevant factors and influences, which eventually show up as unidentified sources of variation (i.e., noise). Instead of testing an hypothesis, the researcher attempts to "prove" a particular personal belief. This is both dangerous and unethical. Ironically, it is the narrow mind that produces an overbroad experimental design and an open mind that produces a narrowly focused design. The latter is by far preferred, while the former is a waste of time, resources, and money.

The second problem of lack of game plan can be avoided by having a coherent, systematic approach to the problem. This is the heart of a good research proposal, and it is at that stage where designs of *all* experiments necessary to attain stated objectives must be made. To revive the cooking analogy, in making a cake there is a time to add the dry ingredients, a time to add the wet ingredients, a time to mix, a time to bake, etc. In order to produce the desired objective (i.e., an edible cake), the whole process must be visualized and then followed according to plan (i.e., recipe).

Good scientific research is conceptually no different from baking a cake, except that the results are published instead of eaten. The objectives must be clearly stated up front. Most objectives are to gain scientific understanding through testing specific hypotheses. Too often, however, objectives are stated as "to measure . . . , to monitor . . . ," etc. These are not objectives, but means of attaining objectives. The difference is significant. The objective is not to measure flour or heat a mixture of ingredients at 350°F, but to produce a cake.

And like good cooking, good research is rarely conducted *ad hoc*. The first ingredient is a thorough understanding of existing scientific knowledge in order to know what assumptions can be made or should first be tested. The importance of identifying potential sources of error right from the start cannot be overemphasized. These assumptions include all aspects of research from equipment (Are all pumps really operating the same? Do all ponds have the same size and characteristics?), to methodology (Is my sampling truly random or representative? Should I trust my DO meter? should I believe my primary productivity or chlorophyll measurements?), to design (Is this variable truly not important and therefore should not be measured?). It is difficult to repair the damage after discovering that assumptions were falsely made and probably not even considered. It is particularly embarrassing when a reviewer of the subsequent manuscript points out these unconsidered (and possibly fatal) assumptions. Not all assumptions should necessarily be tested, but the competent researcher will carefully consider all possible sources of error (i.e., sources of variation) and make every reasonable attempt to minimize them before proceeding.

Identifying and evaluating necessary assumptions is a critical step in any comprehensive research design. If more than one experiment is required to satisfy research objectives, then these experiments must be planned and coordinated logically and efficiently. Two common schemes are the "wagon-wheel" approach and the "funnel" approach. In the wagon-wheel plan, experiments are like the wheel's spokes, which address a central objective (the wheel's hub) from different angles. An example would be looking at the role of chicken manure fertilization in the production of Nile tilapia (Knud-Hansen et al., 1993). In that study, chicken manure was examined as source of N, P, and inorganic carbon for phytoplankton production, particulate carbon as a direct food source, impacts on pond ecology, and economic feasibility. Another example is systems modeling, where each experiment may evaluate an identified relationship in the system. The hub of the wheel is like a jigsaw puzzle, and each experiment provides missing piece(s).

The funnel approach is appropriate where the objectives are more focused and background scientific knowledge is less understood. Experiments are more linearly planned, beginning with broad ones and progressing to those more narrowly defined. Assume, for example, that the research objective was to determine the optimal input rate of the best available local plant to feed a culture of native herbivorous fish. The first experiment would be to test all possible candidates (i.e., local plants) in an unstructured experiment. The clear winner(s) (based on fish yield, plant availability, convenience, etc.) would then be examined in a dose–response experiment to determine optimal input rates based on predetermined criteria.

The wagon-wheel and funnel approaches are not mutually exclusive, and both can be easily incorporated in a comprehensive research scheme. Regardless of the approach taken, however, the researcher must be clear from the start about how results from each experiments fit into the big picture. The overall plan must be flexible where necessary, but potential contingencies should be already outlined in the research proposal. One research direction may be chosen over another depending on whether or not the preliminary experiment showed any significant relationships or differences.

EXPERIMENTAL UNITS

At this point the researcher has identified the hypothesis(es) to be tested, and has determined the most appropriate designs (e.g., unstructured, dose–response, factorial) to meet specified research objectives. Part of this preparatory process also entails defining what the experimental units for each experiment will be.

Types

Experimental units are individual entities, representing a population of similar entities, each of which receives an individual treatment. In aquaculture the most common experimental units are aquaria, tanks, hapas (net enclosures), and ponds. For example, in a fertilization study conducted in ponds, each pond is an experimental unit because treatments (i.e., different fertilization strategies) are applied on a per pond basis. Similarly, with a fry density experiment conducted in 30 hapas with 10 hapas in each of three ponds, the hapas would be the experimental units if fry density (i.e., treatment) varied on a per hapa basis.

Recall the discussion in the section on Basic Concepts regarding samples and populations. Experimental units can be thought of as a subset of a population of like units (aquaria, ponds, etc.) similarly situated; or, the population could also be the same aquaria but at different times. For example, results in your experimental ponds will be used to predict what will happen in other similarly situated ponds receiving the same treatments or in the same ponds at a latter time.

Individual organisms can also be experimental units as long as there is some way to mark, tag, or otherwise distinguish one individual from another. Types of data collected include observed changes within each organism over time. For example, assume eight common carp (tagged for identification) were treated with a test solution to kill known parasites. The null hypothesis is that the solution has no effect on parasite infestation. A parasite count is performed on each fish before and after treatment. The change in parasite number for each fish is determined, and the eight values are analyzed using a paired sample *t*-test or Wilcoxon's Signed Rank test (See the section on Data Analysis) to see if the treatment had a significant effect on parasite infestation. In this experiment each individual fish is an experimental unit.

Replications

To appreciate the importance of treatment replication in designing experiments, the researcher must first understand two points, first, why replication is necessary and, second, what needs to be replicated. First, replication of treatments is necessary because of noise. Remember that there are generically two sources of variation, known and unknown. Whatever variation we cannot account for by treatments and other measured variables is called the residual experimental error, or noise. Through identical replicated treatments, the residual error can be partitioned from total experimental variability (as indicated, for example, in Equations 10 through 12). As will be discussed more fully in the section on Data Analysis, the ratio of known to unknown variation is the foundation of analysis of variance (ANOVA). With regard to the second point, experimental units (aquaria, ponds, etc.) must be replicated to determine within-treatment variability.

Treatments can be replicated spatially or temporally. With spatial replication the same treatment is repeated in two or more experimental units in the same experiment. This is by far the more common approach and is generally preferred over temporal replication, in which a treatment is replicated in sequential experiments over time. Temporal replication may be used where experimental and environmental conditions are highly controlled, such as with some hatchery experiments. For outdoor investigations, however, variable climatic conditions add too much experimental error to make temporal replication a reasonable option.

Now comes the important question: how many replicates are enough? Simplistically, the greater the anticipated experimental error (noise) and/or desired precision of analytical estimates, the more replicates you need. Practically, however, the type of experimental units (e.g., ponds vs. aquaria) and the scope of the experiment often influence the chosen number of replicates per treatment. For example, inherent variability would be expected to be relatively high in earthen ponds as compared to tanks or aquaria. Similarly, flow-through systems with constant water quality may be less variable than static water cultures, and incubation under controlled indoor conditions would likely be less variable than outdoor culture systems. A limitation of the number of available experimental units can also affect the number of different treatments or testable hypotheses for a given experiment. Remember from Equation 6, however, that decreasing the number of replicates per treatment generally increases the standard error of the treatment mean and therefore decreases the precision of how well this sample mean estimates the true mean. So trying to squeeze in too many treatments at the expense of too few replicates only increases the possibility of erroneous conclusions. For unstructured aquaculture experiments, where the hypothesis involves comparing one treatment mean with another, three should be the minimum number of replicates per treatment; four replicates per treatment is generally better. In fact, statistical power analysis (Searcy-Bernal, 1994; Cohen, 1988; and discussed briefly in the section on Scientific Approach to Aquaculture Research) may indicate the need for even more replicates to reduce the probability of missing significant treatment differences (i.e., committing Type II errors).

Replication in dose–response type experiments can be viewed differently from unstructured experiments. There is often only one treatment, with several levels of that treatment. Stocking density may be the treatment, and stocking densities of 1 m^{-2}, 2 m^{-2}, 3 m^{-2}, and 4 m^{-2} are different levels of that treatment. In contrast to unstructured experiments, where the null hypothesis involves comparing two or more treatment means, with dose–response experiments the null hypothesis involves comparing the observed relationship with a predicted model. The residual (experimental error) is measured by summing the absolute differences between observed and predicted values based on the model. Since there is really only one treatment with several levels, there is no reason to replicate each level (i.e., stocking density) since the residual is *not* based on within-level variation.

For example, assume a researcher wants to know the relationship between the rate of urea input and the growth rate of a particular herbivorous fish. She has twelve ponds in which to conduct the experiment. Since the experimental objective is to determine a ***relationship*** (and not differences between means), it is better to have twelve different levels of urea input with no replicate levels (Figure 5) than to have, say, four levels with three replicates of each level (Figure 6a). Rather than testing a relationship, the latter design tests whether one treatment mean is significantly different from another (Figures 6b, 6c). Remember that different urea input rates just represent different levels of a single treatment, namely urea input. In a sense both designs, therefore, have twelve replications of a single treatment.

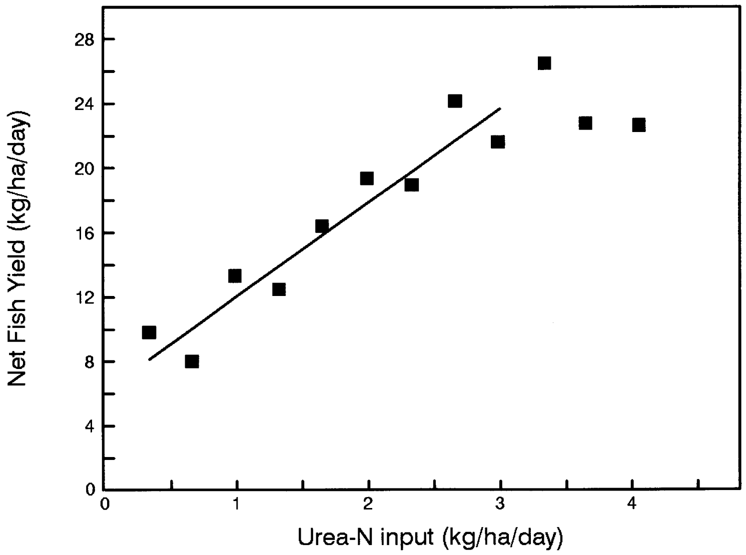


Figure 5. Results from a hypothetical dose-response experiment using 12 different input levels to examine the relationship between urea-N fertilization and net fish yield. The regression line shows a linear relationship ($r^2 = 0.91$, $P < 0.001$) up to about 3 kg urea-N/ha/d.

There are at least two reasons why it is better not to replicate levels in a single treatment, dose–response experiment. First, the greater the number of levels (i.e., urea input rates), the more likely any true relationship between urea input and growth will be detected (compare Figure 5 with Figure 6). Second, any relationship between urea input and the individual residuals (observed fish growth minus predicted fish growth) which constitute experimental error can be identified. A significant relationship here may reveal the need to transform and reanalyze the data, and any relatively unusual or extreme data can be readily identified and reexamined if necessary. These aspects of residual analyses are discussed more fully in the section on Data Analysis.

Unfortunately, nearly all dose–response experiments in aquaculture have replicated treatment levels to test differences between treatment means and then add regression lines to show relationships. It is not that these studies are wrong; it is just that researchers could have attained more valuable information with the same amount of effort. Since researchers have taken this hybrid approach, the reported relationships have limited interpretive value because either there were sizeable gaps between treatment levels or the range of levels was too narrow to infer a broader, meaningful relationship. Analysis of treatment means is only meaningful for those few, somewhat arbitrarily selected treatment levels. As an example, if the underlying objective in the urea input study mentioned above were to find which input rate gives the best fish yield, increasing the number of treatment levels clearly provides more useful results (again, compare Figure 5 with Figure 6).

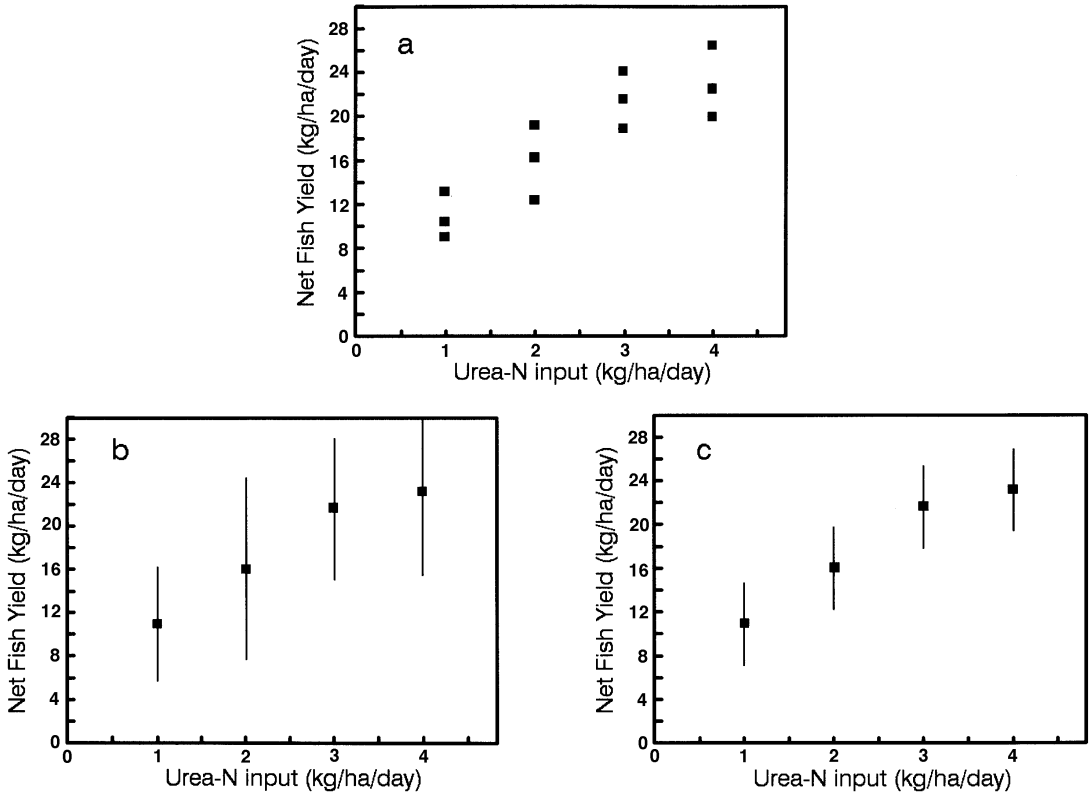


Figure 6. Results from a hypothetical treatment experiment using four different levels of urea-N input as four treatments with three replicates per treatment. Graph (a) shows individual treatments, graph (b) gives treatment means with 95% confidence limits based on individual treatment standard errors (i.e., *t* value based on 2 degrees of freedom), and graph (c) gives treatment means with 95% confidence limits based on the pooled standard error for all treatments with *t* value based on degrees of freedom for residual error in an ANOVA (i.e., 8 degrees of freedom, see Tables 1 and 2).

The threshold question in choosing one design over the other is whether the ultimate objective is to compare treatment means or to determine relationships.

However, there are situations in which replicates in dose–response experiments are justified, such as when relationships and treatment mean comparisons are both important experimental objectives. Furthermore, factorially designed experiments with a dose–response component (e.g., the lipid-protein shrimp feed experiment described above) also represent a hybrid approach in which replicates must be utilized in order to properly analyze possible interactions between factors (see the section on Data Analysis).

Allocation of Treatments to Experimental Units

Estimating experimental error is an essential element of data analysis. As was just discussed, whether residual determinations are based on within-treatment variation or on observed variation about a hypothesized model depends upon the chosen experimental design. The greater the experimental error (residual variation) as a percentage of total variation, the more difficult it becomes to find a statistically significant treatment effect, even if there truly is one. There is also a greater likelihood of committing a Type II error. A primary objective in designing experiments, therefore, is to identify and account for other possible sources of variation.

Perhaps the most common source of experimental error comes from the allocation of treatments to experimental units. By better understanding any nonuniformity of experimental units, variation due to suspected nonuniformity can be partitioned from other sources of variation, including experimental error. For example, are all the ponds identical? Are some ponds leakier than others? Are they really all the same size? In a tank experiment, are some tanks shaded more than others? Do aquaria on the top shelf have the same water temperatures as those on the bottom shelf? Such examples of potential sources of error are endless. Depending on the which experimental units are different, treatment effects could be artificially enhanced or diminished.

Completely Randomized Design (CRD)

If the researcher believes that all experimental units are identical, then treatments should be allocated to experimental units in a completely random fashion. This type of treatment allocation is called a ***completely randomized design*** (*CRD*). With this design, experimental units are either presumed to be identical to each other or any actual differences are nonsystematic. That is, there is no way to characterize any differences. The representative model is the same given in Equation 11 above:

$$Y_i = \mu + \tau_i + \varepsilon_i \tag{13}$$

where μ = overall mean, τ = deviation due to treatment, and ε = residual, or deviation due to experimental error.

The simplest way to randomize treatment allocation is by using a random numbers table. Such tables are located in the appendices of most statistics books or can be computer-generated using statistical software. First, make a vertical column listing all treatments. Then identify each experimental unit with a number from 1 to however many are needed. For example, a feeding trial conducted in tanks that tests four treatments with four replicates per treatment requires 16 tanks. Number the tanks 1 through 16. Then randomly pick a point in the random number table and read the last two digits in the column. If that number is not between 01 and 16, then continue reading down. The first suitable number, say 08, will be the tank number for the first replicate of the first treatment as listed in the column of treatments. Then, continue down the random number column until another number between 01 and 16 (but not 08) appears, and that will be the tank number for the second replicate of the first treatment listed. Continue this process until every treatment and replicate has a different tank number. You have now completely randomized treatment allocation to your experimental tanks.

Some advantages and disadvantages of the CRD design, as well as other designs to follow, are discussed in the Data Analysis section using example analysis of variance (ANOVA) tables for each treatment allocation scheme. Such analytical considerations are essential at the experimental design stage, and the reader is strongly encouraged to familiarize him/herself with the section on Data Analysis prior to finalizing any research plans.

Randomized Complete Block Design (RCBD)

Complete randomization is ideal when appropriate but will result in unnecessary experimental error if there are systematic differences between experimental units. If experimental errors are not random but instead relate to some definable differences between experimental units, then this variation can be separated out.

The simplest case occurs when variability among experimental units can be blocked out. For example, you find that water in aquaria on the top shelf is warmer than in those on the bottom shelf; or you have 16 cement tanks in a 4 by 4 square, and a dirt road passes along one side (noise/cars make catfish very nervous, and dust has a lot of phosphorus in it); or a water source canal passes by some ponds but not others; or you are doing a fry nursing experiment in hapas (i.e., experimental units) with 15 hapas in each of three ponds, but you know the ponds are not identical. In each of these examples, variability among experimental units can be blocked out by allocating treatments according to a ***randomized complete block design*** (*RCBD*).

The main idea of a RCBD design is to *maximize* experimental unit variation *between* blocks, and *minimize* variation *within* blocks. So with the above examples, each shelf of aquaria would be a block; each row of tanks parallel to the road would be a block; ponds adjacent to the canal would be one block, while those away would be another block; and each pond (with 15 hapas in each pond) would be a block. The model for each of these designs would be the following:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \tag{14}$$

where β = variation due to blocks.

The RCBD restricts randomization of treatment allocation in order to better isolate treatment variation while minimizing residual error. There are three requirements in setting up a RCBD. First, there must be the same number of blocks as there are replicates per treatment. Second, each treatment must be represented in each block. And third, treatment allocation within each block should be done randomly. In the hapa experiment, for example, there would be as many replicates per treatment as there are ponds. Each treatment would be represented once in each pond, and randomly allocated to a hapa within each pond.

An added benefit to using a RCBD is the extra hypothesis being tested with no additional effort. The RCBD tests the null hypothesis that the blocking (or the reason for blocking) has no significant effect on the response variable. Using the above examples, null hypotheses may include (1) that there is no road effect on catfish growth in tanks adjacent to a road or (2) that there is no difference in shrimp growth due to the horizontal placement of aquaria on shelves. In the former example, three blocks may be statistically the same, with only the block adjacent to the road causing a significant block effect. In the latter experiment, a significant block effect may reflect a horizontal gradient in water temperatures. For example, with four shelves and accurate temperature measurements, a dose–response experiment revealing a growth relationship with water temperature can be incorporated solely by allocating treatments to aquaria using a RCBD. The section on Data Analysis discusses how to analyze data, test hypotheses, and make all relevant comparisons.

To most effectively use the power of a RCBD, blocks must be selected carefully. Remember that the goal of blocking is to maximize differences between blocks while keeping experimental units within blocks as similar as possible. And always keep in mind that the reason for restricting randomization of treatment allocation is to identify and quantify sources of variation.

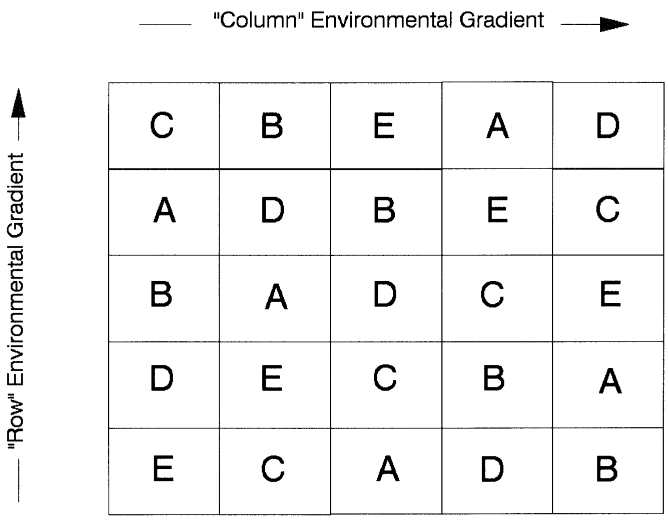
Latin Square Design

Whereas the RCBD identifies differences between experimental units across one dimension (e.g., between ponds in the hapa experiment or between shelves in the aquaria experiment), the ***Latin Square design*** attempts to quantify differences in experimental units when two sources of variation are suspected. The representative model looks like the following:

$$Y_{ijk} = \mu + \tau_i + r_j + c_k + \varepsilon_{ijk} \tag{15}$$

where r = variation due to rows, and c = variation due to columns.

The Latin Square design severely restricts randomization and therefore has fairly rigid design requirements. Typically, experimental units are laid out in the shape of a square (Figure 7), although variations are possible. The design requires that the number of replicates per treatment equals the number of rows and columns in the square. It is also essential that a treatment be represented only once in each row and column (see, for example, Figure 7). For analytical reasons (see the section



C	B	E	A	D
A	D	B	E	C
B	A	D	C	E
D	E	C	B	A
E	C	A	D	B

Figure 7. A schematic diagram of an experiment where five replicates of each of five treatments (A through E) were allocated to experimental units using a Latin Square design to account for suspected environmental gradients in two directions. Notice that each column and row is represented by exactly one replicate of each and every treatment.

on Data Analysis) a 4 by 4 Latin Square is the smallest size that should be considered. For practical reasons, a 6 by 6 (with 36 experimental units) is probably the largest Latin Square manageable.

Thankfully, the Latin Square design has limited utility in aquaculture. One possible situation may be in an experiment using aquaria as experimental units. If we assume that the aquaria are in a poorly air-conditioned room with large windows on a wall perpendicular to the aquaria, there could be a horizontal temperature gradient (i.e., row variation) and a vertical light gradient (i.e., column variation). A Latin Square design could identify variations caused by these gradients, thereby reducing experimental error. It might be easier, however, to move the aquaria to the wall opposite the windows, fix the air conditioner, and use a CRD.

The RCBD and Latin Square designs are complete block designs, where every treatment is represented in each block. There are also more complicated ways to block out suspected variation among experimental units utilizing incomplete block designs such as the ***split-plot design***, ***balanced incomplete blocks***, and ***partially balanced lattices*** (Steel and Torrie, 1980). Although discussion of these designs is beyond the scope of this chapter and best left for statistics textbooks, it is worth noting that the best designs are often the simplest, and increasing design complexity without increasing precision makes no sense.

Analysis of Covariance (ANCOVA)

The last design consideration applies when you suspect differences in your experimental units, but there is no way to systematically block out those differences. This is less of an issue when aquaria or cement tanks are used as experimental units but can be a serious concern when earthen ponds are used. The way to account for suspected nonuniform sources of variation is through an analytical technique called ***Analysis of Covariance*** (ANCOVA). The section on Data Analysis presents further aspects regarding ANCOVAs, but it is important to understand some basic concepts when designing experiments.

As an example, this technique was employed when effects of previous pond fertilizations (i.e., pond history) were shown to significantly affect tilapia production in a subsequent grow-out experiment (Knud-Hansen, 1992a). Since the ponds had received variable inputs (mainly chicken manure, urea, and TSP) during prior experiments, each pond had a unique fertilization history.

Analysis of variance of the CRD grow-out experiment revealed that treatments accounted for only 55% of the observed variation, while the residual accounted for about 45% of total variation. ANCOVA showed that over 90% of the residual variation, however, was due to variable chicken manure, urea, and TSP fertilizations from prior experiments within the previous two years. The greater the amount of previous fertilization, the greater the positive influence on algal productivity, and therefore tilapia productivity (see also the section on Data Analysis under residual analysis).

Based on the pond history example, it may be argued that it would be better to scrape the pond bottoms prior to each experiment in order to make the ponds more uniform. But this would be wrong for many reasons, not least of which would be the high cost and the practical impossibility of scraping pond sides and bottoms uniformly. In fact, anything but precision scraping would likely cause more harm by reducing the ability to identify and partition out sources of between-pond variability. Experimental units do *not* have to be uniform as long as the reasons for their lack of uniformity can be identified and tested. The message, therefore, is to identify possible differences among your experimental units *before* beginning the experiment and make the necessary measurements to be later used in ANCOVA.

SYSTEM CHARACTERIZATION AND MONITORING

Before beginning this section on system characterization and monitoring, it is useful to see where we are with respect to the experimental design process. So far we have conceptually:

1. Identified a research topic
2. Identified primary objectives of the research
3. Stated these objectives in the form of null hypotheses
4. Determined which type(s) of experiment(s) (e.g., unstructured, dose–response, etc.) are best suited to test stated hypotheses
5. For each experiment determined the scope, treatments, treatment levels, and number of replicates per treatment most appropriate for testing stated hypotheses
6. If more than one experiment is necessary, arranged in logical systematic order the appropriate sequence for conducting several experiments
7. Identified the most appropriate allocation of treatments to experimental units (e.g., CRD, RCBD, and Latin Square)
8. Identified any differences among experimental units that may affect experimental results and that can be measured before beginning the experiment

The last remaining task is to identify which variables (i.e., system characteristics) should be measured during the experiment and to determine how to go about measuring these variables.

Identification of Variables

Everyone agrees that you cannot measure everything, so there must be some sort of limitation on both what you measure and how frequently you measure it. Factors that limit the scope of variables include time, money, effort, and utility of data collected. As a practical matter, utility of data affects more the decision of which variables should be measured, while time, money, and effort affect more the determination of measurement frequency.

There are three types of variables generally measured in an experiment. The first type represents those variables that are directly part of your null hypothesis. If the null hypothesis states that brand X feed has no greater beneficial effect on fish growth than brand Y, then you must measure fish growth in order to test the hypothesis. All variables essential for testing stated null hypotheses must be measured. These will be the response variables in unstructured experiments, and the relationship variables in dose–response experiments. For example, in a fertilization experiment that also includes the null hypothesis that algal productivity has no effect on fish yield, both algal productivity and fish yield must be measured.

The second type includes those variables that are not essential for testing null hypotheses but may reflect other sources of variation affecting response variables. Water quality measurements most frequently fall into this category. Which water quality variables should be measured depends on the nature of experimental units and how the researcher intends to analyze resulting data. With outdoor tank and pond studies, for example, variables that affect both the culture species and algal growth should be included. Culture species are generally affected by temperature, dissolved oxygen (DO) concentrations, un-ionized ammonia concentrations, and turbidity (Meade, 1985). Algal productivity is affected by temperature, light availability, and availability of inorganic N, P, and C (Knud-Hansen and Batterson, 1994). Monitoring the following variables would provide information used to evaluate possible effects on response variables: water temperature, DO, total ammonia-N, nitrate-nitrite-N, soluble reactive P, pH, Secchi depth, total and ash-free suspended solids, and total alkalinity. Chlorophyll *a* (i.e., primary production) would not necessarily be included because its relationship with the *rate* of algal growth (i.e., primary productivity) is extremely tenuous. What affects the *rate* of growth for detrivorous species like Nile tilapia is primary productivity, not primary production (Knud-Hansen et al., 1993).

Variables that are used for broader purposes beyond the immediate experiment represent the third type. Such data provide baseline information useful for comparing separate experiments. For example, weather measurements may not provide any insight into variations observed in one experiment but can be extremely useful when making comparisons between seasons or climatic regions.

It is essential to understand how you intend to use and analyze the data being collected. Measuring Kjeldahl N simply because you have the equipment available is a waste of your time, money, and energy if all you intend to do with the data is create summary tables. Carefully consider all reasonable variables and eliminate those that do not help test your hypotheses and meet your experimental objectives. If the Kjeldahl digester needs to be used, then design an experiment where those data have importance.

Selection of Methods

Once you have identified which variables should be measured to test your hypotheses, the next issue is how should they be measured. There are a number of method manuals and books offering a variety of ways to get the same information. But the fact that these methods are in print does not necessarily make them the best or even necessarily adequate. Standard Methods for the Analysis of Water and Wastewater (APHA, 1985) is one of the few that present methods that have been actually comparatively tested, but they too might not be suitable for your particular needs.

It is worth the effort for an aquaculture research facility or program to select methods based on rational comparisons. There are five factors that should influence the choice of one method over another: (1) precision or sensitivity, (2) accuracy, (3) reproducibility, (4) ease of operation, and (5) cost. For field and laboratory instruments, decisions often can be made by comparing hardware specifications. An instrument's reputation regarding ease of operation and repair history can also be quite useful. Inherent in such decisions is an understanding of how the data can foreseeably be used, because accuracy and precision must often be balanced against cost. For example, is a digital pH meter accurate to 0.001 pH unit really cost effective? Is that level of sensitivity really important?

With water chemistry such comparisons are essential, not only to understand the limits of a particular method but to ensure that the most rational method is being used. As an example of how the above decision criteria can be applied, several years ago the aquaculture laboratory at the Asian Institute of Technology compared the nitrate-nitrate analysis using the cadmium reduction column

(APHA, 1985) with another method using hydrazine reduction (Kamphake et al., 1967). To determine precision, the average slope of five standard curves (linear relationship between nitrate concentration and spectrophotometric absorption; APHA, 1985) for each method was compared. Hydrazine was slightly more precise (i.e., had a higher slope), but both methods were adequate for aquaculture investigations. For accuracy, five internal standards per method were used. Internal standards gave the percent recovery of known nitrate and nitrite spikes added to pond water (Knud-Hansen, 1992b). Cadmium reduction gave slightly higher percent recoveries than hydrazine reduction. Reproducibility was based on standard deviations of five replicate samples, and there was little difference between methods. Ease of operation favored hydrazine reduction. A liquid reagent was much easier to handle than dealing with the vagaries of the cadmium reduction column. Differences in cost were not determinative. What tipped the scales in favor of the cadmium column method, however, was the current laboratory expertise in the method, the fact that the laboratory had been using the cadmium column method for years, and the lack of significant analytical improvement by switching to the hydrazine method. The important point here is that a rational decision was made using specific comparable criteria.

Aquaculture laboratories should also be on the look-out for new methods that offer significant improvement in one or more of the above decision criteria. A good example is with chlorophyll *a* measurements. A review of the aquaculture literature indicates continued widespread use of the acetone extraction method (Strickland and Parsons, 1968), whereas a review of the limnological and oceanographic literature suggests that most laboratories have rejected that method for others that are much easier and less prone to experimental error (e.g., Holm-Hansen and Riemann, 1978).

Finally, field methods should be tested before blind adoption. For example, primary productivity is commonly measured by *in situ* incubation of pond water suspended in light and dark bottles at multiple depths (Boyd, 1990). This method may be appropriate in oceans and lakes with photic zones of several meters or greater, but it also has been commonly used in aquaculture studies in ponds with Secchi depths of < 15 cm. Data are reported as if resulting numbers actually represented pond algal productivity (severe vertical variations and unmeasured floating algal mats notwithstanding). There may be no practical way to measure true algal productivity in shallow, highly productive aquaculture ponds. Nevertheless, it is absolutely essential that researchers understand the analytical limitations of all methodologies employed in their research.

Representativeness of Samples

Both sampling and field measurements are founded on the implicit assumption that they represent larger populations. Fish samples characterize the pond's population of fish in the same way water and sediment samples characterize the pond itself. In order to obtain representative samples, however, the researcher must appreciate the spatial and temporal heterogeneity of aquaculture systems.

Studies conducted in aquaria and recirculating systems present the fewest problems regarding collecting representative samples. Water in aquaria can be easily mixed, and recirculating systems by their nature are generally more homogenous.

Perhaps the greatest concern is in earthen ponds, where so many dynamic processes occur at once. There are diel temperature and oxygen changes, daily thermal stratifications, and variable weather-induced effects just to name a few. The following discussion does not pretend to be exhaustive but only alerts researchers to the more common considerations regarding biological, water, and sediment sampling.

The main problems dealing with representative biological sampling concern collecting fish, zooplankton, and phytoplankton samples. Fish sampling can be problematic, depending on the method of sampling. Seining can give a biased representation if net avoidance is related to size or species. One way to test your method is to sample each pond the day before draining and completely harvesting the ponds. A comparison of sample means and distributions with each pond's population mean and distribution should indicate the adequacy of samples (see Hopkins and Yakupitiyage, 1991). Collecting representative zooplankton samples is particularly difficult because of their diel vertical migrations, the inherent patchiness of their distributions, and their ability to avoid nets and traps (Wetzel and Likens, 1979). The primary difficulty with representative phytoplankton sampling is how to quantify surface mats of blue-green algae. Unless the experimental hypothesis focuses on zooplankton and/or phytoplankton population dynamics, sampling (and subsequent analyses) should probably be restricted to a qualitative rather than quantitative level.

Water samples collected for chemical analyses should also be checked for representativeness. A 200-m^2 pond 1-m deep contains 2×10^5 L of water. How well does a one liter sample represent that pond (see Likens, 1985)? Spatial heterogeneity occurs both areally and vertically with depth. Afternoon soluble ammonia and phosphorus concentrations can vary several mg L^{-1} in highly productive fish ponds. Thermal destratification (i.e., complete pond mixing) can occur nightly or after an afternoon thunderstorm. Based on these and other similar considerations, researchers must select water sampling methods and times of collection carefully to provide the necessary representative data.

Water samplers that collect an integrated vertical sample of the entire water column are useful for whole pond nutrient budgets but are entirely inadequate for studying dynamics of thermally stratified ponds. Collecting and mixing water samples from different parts of the pond may help reduce problems associated with areal heterogeneity, but it would be extremely useful to determine the extent of variability through at least one synoptic sample of a pond (i.e., sampling many locations of the pond "simultaneously"). Timing of sampling should be based on pond dynamics, not convenience. If anoxia is a concern, DO samples should be taken at pre-dawn and not some fixed hour regardless of sunrise. If un-ionized ammonia is a concern, temperature, pH, and total ammonia should be measured several times during the day and particularly at mid-afternoon when the pH is the highest.

Trying to collect representative sediment samples may be the most difficult of all. Spatial heterogeneity over the pond bottom must be determined before any meaningful analysis can be made. Just collecting ten or so samples and mixing them without knowing the variability of these samples can readily lead to false conclusions. Synoptic sampling may be required in three or four ponds, but the effort is necessary to confirm that mixing a given number of scattered samples does in fact represent well the pond bottom. An additional problem is determining how deep the samples should be. Decisions regarding both the number of samples per pond and the depth of the samples should be rationally based after testing necessary assumptions. Sediment analyses tend to be time-consuming and costly, so it is important that the samples truly represent what the researcher believes they represent.

One last point on representativeness regards the concept of samples vs. subsamples. Individual samples are used to infer characteristics of a population. A water sample may represent the entire pond volume or just the upper photic zone, depending on how the researcher defines the population. Collecting and analyzing many samples acquired at the same time characterizes that population, whether it be the pond's water or sediments. These are replicate samples taken from a single population. But taking one sample and splitting it into several parts is called taking subsamples. Subsamples are good for evaluating variability of analyses (see discussion above on selection of methods) and as a backup reserve if/when rerunning an analysis becomes necessary. Subsamples do not, however, provide any indication of pond water or sediment variability.

Sample Size

The question of how large a sample size must be collected to adequately represent a population comes up often in pond studies. In particular, these concerns arise mainly with respect to sampling culture organisms and pond water. Sampling culture organisms in studies conducted in aquaria or tanks usually involves collecting entire populations (of each tank, etc.) for routine monitoring.

When sampling larger populations in pond and large tank studies, however, an adequate sample size must be determined.

One fallacy often perpetuated in aquaculture is that the sample size of culture organisms should be 10% of the total population. Recall the discussion in the section on Basic Concepts regarding the standard error of the mean (Equation 6) and confidence intervals. The standard error of the mean is based on the standard deviation of sample observations and the number of observations made. Confidence intervals about the mean are based on the standard error of the mean, the sample size, and the level of confidence (e.g., 95%) desired by the researcher. Nowhere does *population* size play a role in determining these statistics. Sample size simply does *not* depend on population size.

The sample size does depend on how precise (i.e., how small a confidence interval about the sample mean) you want or, in other words, the maximum acceptable difference at a given probability between the sample mean and the true mean. To determine the appropriate sample size, first take a preliminary sample, and use the information in the following standard equation:

$$t_{\infty} = \frac{(\bar{x} - \mu)}{s\ \sqrt{n}} \tag{16}$$

where: t_{∞} = tabular t value at ∞ degrees of freedom at the desired probability (level of confidence)

$\bar{x} - \mu$ = maximum acceptable difference between a sample mean and true mean at a desired probability

s = standard deviation of preliminary sample

n = number of observations in preliminary sample

$s/\sqrt{n}$ = standard error of preliminary sample mean

As an example, assume a farmer needed to know the mean weight of about 20,000 fish raised in a pond. He also wants to be 95% confident that the sample mean weight is no more than 10 g away from the true mean. A preliminary random sample of 30 fish gave a standard deviation of 42.0 g. By restructuring Equation 16 and noting from the Student's *t* Table in his handy statistics book that t_{∞} at $P = 0.05$ equals 1.96, the farmer does the following calculation:

$$n = \left[\frac{ts}{(\bar{x} - \mu)}\right]^2 = \left[\frac{(1.96)\,(42)}{10}\right]^2 = 67.8 \approx 68 \text{ fish}$$

Therefore, the farmer must randomly sample 38 more fish to attain a sample mean that he can be 95% confident is within 10 g of the true mean of all 20,000 fish. Again note that population size, as well as the sample mean value, was unimportant in the sample size determination.

In practice, however, most researchers give little thought to how well the sample mean should represent the true mean. It is obvious that the appropriate sample size will differ depending upon population variability. But as a general rule, good random samples greater than 30 to 40 organisms per experimental unit are rarely cost effective. To be sure, use Equation 16.

Water samples collected for chemical analyses should be as large as manageable for two reasons. First, the greater the sample volume, the more likely it will reasonably represent the entire volume. Multiple subsamples combined in a mixing vessel helps reach this first objective. Second, the greater the sample size, the less relative effect any contamination (e.g., from handling, dirty glassware, etc.) may have on the sample. These two general principles help combat the twin problems of system heterogeneity and sample contamination.

Frequency of Sampling

The first criterion for determining sampling frequency is to identify the purpose of sampling in the first place. You want to strike a balance between sampling frequently enough to ensure the utility of the data, and yet not so frequently as to become a relative waste of time and money. With culture organisms, increased handling and labor costs often define the maximum reasonable frequency. With water sampling, the decision becomes more difficult.

For example, assume you are conducting a 5-month grow-out experiment and monitor pond water chemistry to help understand dynamic relationships that hypothetically affect yields. Monthly sampling may be sufficient, but changes in weather, etc. may cause short-term effects, thus increasing overall variability. Because of such variability, any real trends over time may or may not be seen. The conscientious scientist would find any conclusion based only on monthly samples unconvincing. At the other extreme are data loggers, which can amass mountains of data at the flip of a switch. They should be used only where cost effective, that is, where all the information they provide has planned utility. Data loggers that make many measurements, for example, every half hour, are best used for short-term (< 2 weeks) experiments (e.g., Green and Teichert-Coddington, 1991). Unless there are specific analytical objectives for mountains of data, they are most likely wasted when used for monitoring long-term experiments. For a 5-month grow-out experiment, sampling every 1 to 2 weeks would probably be adequate for the purposes of identifying other sources of variability. Or, as an alternative, sample daily for 1 week each month.

Again, the real issue for the researcher is knowing the reasons for monitoring chosen variables. Make sure that you sample enough to allow the data to achieve analytical objectives. Most researchers have a sense of how often they need to sample. But it is well worth the time and effort to think through each variable while designing the experimental protocol. When analyzing data from one experiment, researchers should also ask themselves whether or not sampling frequencies were adequate or sufficient. Addressing such questions will only improve future research.

DATA ANALYSIS

If by chance you jumped to this section in the hope of learning how to analyze your recently collected data set, sorry. Shortcuts do not work. To benefit from this section, you must get here by starting at the beginning of the chapter. Those who started at the Introduction section should now well appreciate that data analysis begins with the first hypothesis formulation. By the time you have completed the experimental design, you already know why, what, and conceptually how to analyze your data.

From your selection of treatments, experimental design, and treatment allocation to experimental units you already understand the basics of how you will analyze your results. This section focuses more on post-experiment considerations, although example calculations will be kept to a bare minimum. As indicated in the Introduction, this chapter minimizes recipes common to most statistics books. Those who wish to do ANOVAs and regressions by hand can choose one of those books for guidance. It is assumed here that the majority of researchers rely on computers to run their statistical calculations. A primary objective of this section is to help the researcher know what she is feeding her computer, and to better understand what the computer gives back in return.

Data Reduction

Conceptually, there are two types of data reduction, distinguished by their different purposes. The first purpose is to summarize data descriptively using means, standard errors (see the section on Basic Concepts) and sometimes ***ranges***. Ranges are just the differences between minimum and maximum observed values. Ranges have limited analytical utility since they represent extreme values, and tend to increase with increasing number of observations (Steel and Torrie, 1980).

Another way of summarizing data is through ***frequency tables*** and ***histograms***. Frequency tables are commonly used to summarize size classes of culture organisms or any other types of data which can be expressed in discrete form (i.e., discrete counts or continuous data (e.g., lengths, weights, etc.) expressed in discrete form). Histograms are bar-graphs used to illustrate frequency distributions (e.g., Figure 8). To adequately present a frequency distribution, you need enough classes to show differences in frequencies, but not too many or the histogram will appear flat. Usually about 10 to 15 classes adequately represent underlying distributions (compare Figure 8b with Figures 8a and 8c).

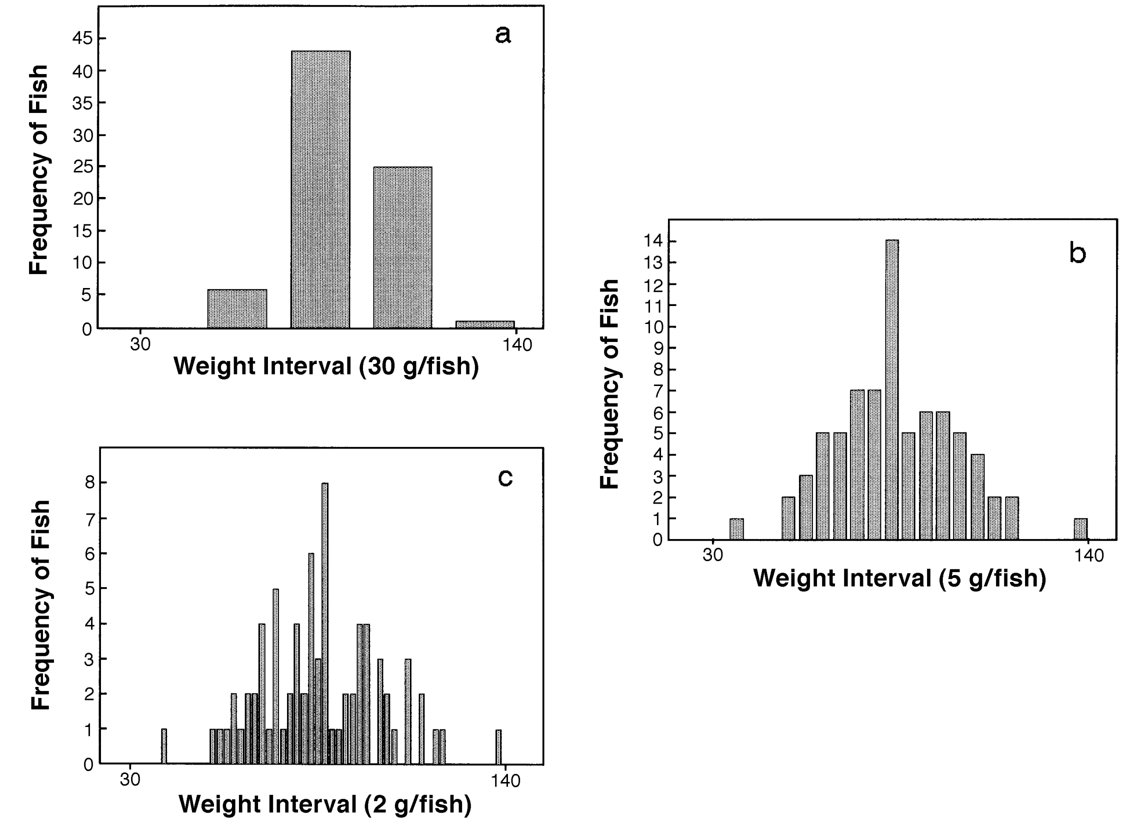


Figure 8. Three histograms of weight frequencies of 75 fish (weights ranged from 34 to 138 g/fish) using weight intervals of 30 g/fish (a), 5 g/fish (b), and 2 g/fish (c).

The second purpose of data reduction, particularly for relatively large data sets, is to facilitate data analysis. Data may require some form of data reduction or consolidation to become more analytically manageable. As a general rule, reduce data *within* an experimental unit and not between experimental units. For example, imagine that you conducted a five-treatment grow-out experiment in tanks with three replicates per treatment. The experiment lasted 10 weeks, and water chemistry was monitored weekly. Assuming alkalinity did not show any trend over time but did vary with treatments, you could then average the ten measurements to give one mean value per tank (i.e., one value per experimental unit). Treatment means would then be based on the three tank values per treatment (i.e., $n = 3$). Do not average the 30 alkalinity values per treatment in order to get a treatment mean; otherwise, you will lose all information regarding within-treatment variability (i.e., variability among experimental units with the same treatment).

Some data reduction may be planned in the experimental design, but always make sure that no important information or insightful data analysis is lost in the process. Reduce data for a purpose, and understand that purpose and consequential implications before reducing.

Data Transformations

An essential requirement for employing *t*-tests, ANOVAs, regressions, correlations, and any other parametric analysis (e.g., using the statistics to estimate the parameters of true mean and variance, see Basic Concepts section) is that observations must have distributions not significantly different from normality. There are statistical tests, such as the Chi-Square test, which test the null hypothesis that your data set is not significantly different from normality. If the null hypothesis is rejected, non-normal data sets can often be normalized through transformations.

A few basic transformations are commonly used for certain types of data that are already known or assumed not to be normal. The ***square root transformation*** ($\sqrt{x}$) normalizes counting data, which typically exhibit a Poisson distribution. If counts are low and include 0, $\sqrt{(x + 0.5)}$ can be used. If counts are all much greater than 100, then transformation is probably not necessary before parametric statistical analysis.

The square root transformation also normalizes binomial percentage data that are generally <30% or >70%. Percent mortalities and survivals often fall in this range and therefore should be transformed before running ANOVAs. Note that the entire data set would be transformed, not just the ones falling within the specified range. If percentage data range widely from near 0% up to around 100%, the ***arcsine of the square root*** (arcsine $\sqrt{x}$) is used. On the other hand, if percentages range between around 30 to 70%, data transformations are generally unnecessary.

The ***logarithmic transformation*** is used when the standard deviation (or variance) is proportional to the mean. A classic example are fish grow-out trials with widely differing results. A treatment mean of 80 g/fish may have a standard deviation of 15 g, whereas a treatment mean of 500 g/fish may have a standard deviation of 100 g. As the mean gets bigger, so does the standard deviation. The Bartlett's test can be used to test the null hypothesis that treatment variances are not significantly different from each other (Steel and Torrie, 1980). The logarithmic transformation should be used with positive numbers only. If there are numbers less than 10, data transformation should be with log(x + 1) instead of log(x). Logarithms can be at any base (e.g., base 10 or the natural log).

It is a wise practice always to check data for normality before examining treatment comparisons and relationships. Virtually all statistical software packages have this capability, and with a computer it only takes a few seconds. If a particular data set is not normally distributed, try using the square root and logarithmic transformations, and retest the data for normality.

When data transformation is necessary, you must use the transformed data for making treatment comparisons, confidence intervals, etc., because all these analyses assume that the data have normal distributions. Therefore, do not change variances and standard deviations back to the original scale. Only treatment means and ranges should be given/reported in the original scale.

Comparisons of Means and ANOVA

This relatively small subsection covers what many researchers think of when they hear the words, "data analysis." Although an oversimplification, comparisons and relationships are often at the heart of hypothesis testing. All the analyses presented here are parametric and assume normality of the data. Other important assumptions are best left to the statisticians to describe, and consultation of your favorite statistics book is highly recommended.

The basic idea of treatment comparisons is to see whether treatment means are significantly different from each other. When the experiment consists of only two treatments, the *t*-test is used to test the hypothesis that one treatment mean is not significantly greater than the other. The *t* statistic is calculated by the following equation:

$$t = \frac{\text{difference between means}}{\text{standard error of difference between means}} \tag{17}$$

The standard error of the difference between means is based on the weighted mean of the two treatment variances, also called the ***pooled variance*** (s_p^2).

$$s_p{}^2 = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{1/2} \tag{18}$$

The pooled variance allows for comparisons of treatments with different sample sizes (n_1 and n_2). Conceptually, consider two ponds (numbered 1 and 2) each containing thousands of fish. You sample both ponds and get a difference between mean fish weight between the ponds. You sample both ponds several more times and get more differences between mean fish weights. The *t*-test calculates the standard error of the difference between means using the sample variances for each pond. In practice, however, each pond is usually sampled just once. The probability that the true means for ponds 1 and 2 are significantly different from each other is based on the ratio of the observed difference between sample means from the two ponds and the standard error of the difference between means (Equation 17). In other words, the difference between mean fish weights in ponds 1 and 2 is more likely to be significant as the variability of fish weights within each pond gets smaller.

The computed *t* statistic is compared to *t* values in the *t*-Table at $n_1 + n_2 - 2$ degrees of freedom. Probabilities are given horizontally in the Table. The probability range (e.g. $0.01 < P < 0.05$) for a given calculated *t* value is found by moving horizontally in the table at the appropriate degree of freedom. These probabilities, as well as those given by statistical software, represent the probability that there was no true difference between the two means, i.e., accepting the null hypothesis.

Experiments with more than two treatment means use ***analysis of variance** (ANOVA)* to test the null hypothesis that no two means are significantly different from each other. Simply stated, this hypothesis is tested by comparing the variation *between* treatments with the variation *among* treatments. A proportionally large between-treatment variation in relation to within-treatment variation (i.e., experimental error) results in a small probability that observed treatment differences happened by chance. The following discussion will examine ANOVAs for the experimental designs described in the section, Experimental Units.

Completely Randomized Design (CRD)

Table 1 provides the basic ANOVA table for a CRD experiment. There are only two sources of error, from treatments (between-treatment variation) and from experimental error (within-treatment variation). The sum of squares represents total variation for each source, and the mean square represents the amount of variation per degree of freedom. The F statistic is the ratio between mean square treatment (MST) and mean square error (MSE, which also equals the pooled variance, s_p^2).

Table 1 Example ANOVA Table of a Completely Randomized Designed (CRD) Experiment

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments	SST	k – 1	$\frac{SST}{k-1} = MST$	$\frac{MST}{MSE}$
Residual error	SSE	k(n – 1)	$\frac{SSE}{k(n-1)} = MSE = s_p{}^2$	
Total	SS	kn – 1		

Note: k = number of treatments; n = number of replicates per treatment.

This calculated F statistic is compared to a table of F values at (k–1) degrees of freedom for the numerator and k (n–1) for the denominator. At given degrees of freedom, the greater the F value (i.e., ratio between MST and MSE), the smaller the probability that the difference between at least two of the treatment means happened by chance.

Table 2 gives an abbreviated ANOVA for a hypothetical CRD experiment with five treatments and five replicates per treatment. Note that 20 out of 24 degrees of freedom are for the residual. As indicated by Table 1, the more replicates per treatment, the more degrees of freedom for error. If there were only three replicates per treatment, 10 out of the available 14 degrees of freedom would be for the residual; with the same amount of experimental variation, MSE would be twice as large (and F half the value) than with five replicates and 20 degrees of freedom for residual. The more degrees of freedom for error, the smaller the MSE and the larger the MST/MSE ratio (i.e., F statistic). This observation highlights the importance of both reducing experimental error and maximizing the number of degrees of freedom for error. The high proportion of degrees of freedom for the residual is an advantage of the CRD, and so the CRD is better suited for smaller experiments with fewer experimental units. The main disadvantage of the CRD occurs when experimental units are not homogenous, thereby increasing experimental error.

Table 2 Example ANOVA Table of a Completely Randomized Designed (CRD) Experiment with Five Treatments and Five Replicates Per Treatment

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments	SST	5 – 1 = 4	$\frac{SST}{4} = MST$	$\frac{MST}{MSE}$
Residual error	SSE	5(5 – 1) = 20	$\frac{SSE}{20} = MSE = s_p^2$	
Total	SS	(5)(5) – 1 = 24		

Randomized Complete Block Design (RCBD)

As discussed in the section Experimental Units, when the researcher suspects systematic differences in experimental units, a RCBD design is often chosen to account for those differences. Table 3 provides the generalized ANOVA formulation, while Table 4 illustrates an example ANOVA with five treatments and five replicates per treatment, with each replication representing an individual block. Notice that the example ANOVA for the RCBD has the same total number and treatment degrees of freedom as the CRD. The only difference is that four degrees of freedom have been moved from the residual in order to test a second null hypothesis that there are no significant differences between blocks. The F value generated by the block analysis (MSB/MSE) is analyzed in the same way as that for treatments, and compared to tabular F values (and associated probabilities) at (b–1) degrees of freedom for the numerator and (k–1)(b–1) for the denominator (Table 3).

Although the RCBD experiment restricts randomization and reduces the number of degrees of freedom for residual error (e.g., from 20 to 16, compare Tables 2 and 4), the variation due to blocks is removed from residual variation (SSE). A large block variation will reduce the SSE so that even with fewer degrees of freedom for error, the MSE is smaller than it would have been without the block design. Not only has the RCBD experiment tested a second hypothesis without any additional effort, but the smaller MSE results in a larger F value for treatment making that analysis more sensitive.

Table 3 Example ANOVA Table of a Randomized Complete Block Design (RCBD) Experiment

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments	SST	k – 1	$\frac{SST}{k-1} = MST$	$\frac{MST}{MSE}$
Blocks	SSB	b – 1	$\frac{SSB}{b-1} = MSB$	$\frac{MSB}{MSE}$
Residual error	SSE	(b – 1)(k – 1)	$\frac{SSE}{(b-1)(k-1)} = MSE = s_p^2$	
Total	SS	bk – 1		

Note: k = number of treatments and b = number of blocks.

If you do not find a significant block effect, reanalyze the data as if it were a CRD experiment (in effect adding the block variation (SSB) and degrees of freedom to the residual variation (SSE) and degrees of freedom). Report, however, that the original design was a RCBD and there was no significant block effect with that particular response variable (i.e., you accepted the null hypothesis for blocks). If you had blocked against a measurable variable such as temperature or light, the lack of a block effect could nevertheless be scientifically significant. If you had blocked based on the physical location of tanks or ponds, a lack of a block effect would provide further insight into the nature of your research system. Remember also that your blocks may affect one response variable and not another, so always first analyze your data as a RCBD.

Latin Square Design

Whereas the RCBD tests one block hypothesis, the Latin Square design tests two block hypotheses. As discussed in the section Experimental Units, this design severely restricts randomization and has limited application in aquaculture. Tables 5 and 6 provide the generalized ANOVA and an example ANOVA with five treatments and five replicates, respectively. With the 5 by 5 Latin Square, only 12 of the 24 degrees of freedom are for residual error (Table 5). A 4 by 4 and 3 by 3

Table 4 Example ANOVA Table of a Randomized Complete Block Design (RCBD) Experiment with Five Treatments and Five Blocks (i.e., Replicates per Treatment)

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments	SST	5 – 1 = 4	$\frac{SST}{4} = MST$	$\frac{MST}{MSE}$
Blocks	SSB	5 – 1 = 4	$\frac{SSB}{4} = MSB$	$\frac{MSB}{MSE}$
Residual error	SSE	(5 – 1)(5 – 1) = 16	$\frac{SSE}{16} = MSE = s_p^2$	
Total	SS	(5)(5) – 1 = 24		

Table 5 Example ANOVA Table of a Latin Square Designed Experiment

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments	SST	k – 1	$\frac{SST}{k-1} = MST$	$\frac{MST}{MSE}$
Rows	SSR	r – 1	$\frac{SSR}{r-1} = MSR$	$\frac{MSR}{MSE}$
Columns	SSC	c – 1	$\frac{SSC}{c-1} = MSC$	$\frac{MSC}{MSE}$
Residual error	SSE	(k – 1)(k – 2)	$\frac{SSE}{(k-1)(k-2)} = MSE = s_p^2$	
Total	SS	$k^2 - 1$		

Note: k = number of treatments, r = number of rows, and c = number of columns.

Latin Square has only six and two degrees of freedom for residual error, respectively. Both latter totals are too few to get a meaningful MSE, and so a 5 by 5 Latin Square is probably the smallest size statistically manageable. On the other hand, a 4 by 4 Latin Square could be finessed by analyzing it as a RCBD first using rows as the block and then reanalyzing the data using columns as the block. This technique would increase degrees of freedom for experimental error; and if the block effects from rows or columns are not significant, the data can be analyzed as either a RCBD or even as a CRD (when neither rows nor columns show significant block effects).

Factorial Experiments

If we go back to the CRD ANOVA (Table 1), there are only two sources of variation, treatments and residual error. The main purpose of RCBD and Latin Square designs is to identify and partition

Table 6 Example ANOVA Table of a 5 by 5 Latin Square Designed Experiment with Five Treatments, Five Rows and Five Columns (i.e., Five Replicates Per Treatment

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments	SST	5 – 1 = 4	$\frac{SSE}{4} = MST$	$\frac{MST}{MSE}$
Rows	SSR	5 – 1 = 4	$\frac{SSR}{4} = MSR$	$\frac{MSR}{MSE}$
Columns	SSC	5 – 1 = 4	$\frac{SSC}{4} = MSC$	$\frac{MSC}{MSE}$
Residual error	SSE	(5 – 1)(5 – 2) = 12	$\frac{SSE}{12} = MSE = s_p^2$	
Total	SS	(5)(5) – 1 = 24		

out sources of variation from the residual error. The main purpose of factorial experiments, however, is to identify and partition out sources of variation (additive and interactions) from two or more factors. The former designs concern location of experimental units, while factorial experiments describe treatment combinations. So depending on the layout of experimental units, the researcher may allocate factorially determined treatments in either a CRD or a RCBD. (Although theoretically possible, a Latin Square design would be extremely impractical.)

The simplest factorial is the 2 by 2 in a CRD (Figure 3). There are a total of four treatment combinations (2 factors with 2 levels for each factor). The four treatments use three degrees of freedom, which are partitioned out one for each factor, and one for their interaction (Table 7). The example given in Figure 3 illustrates two factors (nitrogen (N) and phosphorus (P) input) and two levels of each factor. The result is a four-treatment experiment examining individual effects of N and P individually and together.

Table 7 Example ANOVA Table of a Two Factor Factorially Designed Experiment in a CRD

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments				
A	SS(A)	a – 1	$\frac{\text{SS(A)}}{\text{a}-1} = \text{MS(A)}$	$\frac{\text{MS(AC}}{\text{MSE}}$
B	SS(B)	b – 1	$\frac{\text{SS(B)}}{\text{b}-1} = \text{MS(B)}$	$\frac{\text{MS(B)}}{\text{MSE}}$
AB	SS(AB)	(a – 1)(b – 1)	$\frac{\text{SSB}}{\text{b}-1} = \text{MS(AB)}$	$\frac{\text{MS(AB)}}{\text{MSE}}$
Residual error	SSE	ab(n – 1)	$\frac{\text{SSE}}{\text{ab(n}-1)} = \text{MSE} = \text{s}_\text{p}{}^2$	
Total	SS	abn – 1		

Note: a = number of levels of treatment A, b = number of levels of treatment B, and n = number of replicates per treatment combination.

Factorial experiments become progressively more complicated as the researcher adds more factors and levels to an experimental design. For example, Figure 4 illustrates schematically the shrimp feed experiment first discussed in the section, Treatments. In that experiment there were three factors (shrimp variety, and protein and lipid concentrations in shrimp feed), with two varieties of shrimp, three concentrations of lipid, and four concentrations of protein, resulting in a 2 by 3 by 4 factorial design with a total of 24 treatments. Table 8 gives the generalized ANOVA for this experiment. Assuming three replicates per treatment with each replicate in a separate block, Table 9 gives the ANOVA for the resulting RCBD factorially designed experiment. Notice that the ANOVA analyzes interactions between all factor combinations. Remember also that this ANOVA only represents treatment analysis and does not include individual relationship analyses (e.g., the relationship of protein concentrations in feed and shrimp yield at one level of lipid concentration) as previously listed. The following subsection discusses how to proceed with relationship analyses.

Table 8 Example ANOVA Table of a Three Factor Factorially Designed Experiment in a RCBD

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments				
A	SS(A)	a – 1	$\frac{SS(A)}{a-1} = MS(A)$	$\frac{MS(AC}{MSE}$
B	SS(B)	b – 1	$\frac{SS(B)}{b-1} = MS(B)$	$\frac{MS(B)}{MSE}$
C	SS(C)	c – 1	$\frac{SS(C)}{c-1} = MS(C)$	$\frac{MS(C)}{MSE}$
AB	SS(AB)	(a – 1)(b – 1)	$\frac{SS(AB)}{(a-1)(b-1)} = MS(AB)$	$\frac{MS(AB)}{MSE}$
AC	SS(AC)	(a – 1)(c – 1)	$\frac{SS(AC)}{(a-1)(c-1)} = MS(AC)$	$\frac{MS(AC)}{MSE}$
BC	SS(BC)	(b – 1)(c – 1)	$\frac{SS(BC)}{(b-1)(c-1)} = MS(BC)$	$\frac{MS(BC)}{MSE}$
ABC	SS(ABC)	(a – 1)(b – 1)	$\frac{SSB}{b-1} = MS(AB)$	$\frac{MS(AB)}{MSE}$
Blocks	SSB	n – 1	$\frac{SSB}{n-1} = MSB$	$\frac{MSB}{MSE}$
Residual error	SSE	(abc – 1)(n – 1)	$\frac{SSE}{(abc-1)(n-1)} = MSE = s_p{}^2$	
Total	SS	abcn – 1		

Note: a = number of levels of treatment A, b = number of levels of treatment B, c = number of levels of treatment C, and n = number of blocks (replicates per treatment combination).

Relationships

Evaluating relationships between variables represents a key component of most data analyses and the primary objective of structured experiments. Similar to many other topics presented in this chapter, however, one does not have to look too closely in the aquaculture literature to find basic conceptual and analytical errors regarding such analyses. But because of the vast complexity of such analyses, the following discussion is necessarily limited in scope and aims primarily to equip the reader with a more solid foundation of elementary regression and correlation analyses. This foundation will hopefully enable the reader to explore more knowledgeably (i.e., through better understanding of purposes, assumptions, and limitations) advanced analytical techniques (e.g., path analysis, canonical analysis, etc.) found in some standard and specialty statistics books.

Table 9 Example ANOVA Table of a Three Factor Factorially Designed Experiment in a RCBD

Sources of variation	Sum of squares	Degrees of freedom	Mean squares	F
Treatments				
Shrimp	SS(A)	2 – 1 = 1	$\frac{SS(A)}{1} = MS(A)$	$\frac{MS(AC}{MSE}$
Lipid	SS(B)	3 – 1 = 2	$\frac{SS(B)}{2} = MS(B)$	$\frac{MS(B)}{MSE}$
Protein	SS(C)	4 – 1 = 3	$\frac{SS(C)}{3} = MS(C)$	$\frac{MS(C)}{MSE}$
Shrimp × lipid	SS(AB)	(2 – 1)(3 – 1) = 2	$\frac{SS(AB)}{2} = MS(AB)$	$\frac{MS(AB)}{MSE}$
Shrimp × protein	SS(AC)	(2 – 1)(4 – 1) = 3	$\frac{SS(AC)}{3} = MS(AC)$	$\frac{MS(AC)}{MSE}$
Lipid × protein	SS(BC)	(3 – 1)(4 – 1) = 6	$\frac{SS(BC)}{6} = MS(BC)$	$\frac{MS(BC)}{MSE}$
Shrimp × lipid × protein	SS(ABC)	(2 – 1)(3 – 1) (4 – 1) = 6	$\frac{SSB}{6} = MS(AB)$	$\frac{MS(AB)}{MSE}$
Blocks	SSB	3 – 1 = 2	$\frac{SSB}{2} = MSB$	$\frac{MSB}{MSE}$
Residual error	SSE	(2 · 3· 4 – 1)(3 – 1) = 46	$\frac{SSE}{46} = MSE = s_p^2$	
Total	SS	2 · 3 · 4 · 3 – 1 = 71		

Note: There are two levels of factor A (shrimp variety), three levels of factor B (lipid concentration in feed), four levels of factor C (protein concentration in feed), and three blocks (replicates per treatment combination).

Regression Analysis

Regression analysis determines the presence or lack of relationships between a dependent variable (y), and an independent variable (x). The independent variable is assumed to be constant, not subject to sampling or measuring error. Examples would include the number of eggs produced (dependent) vs. time after injection (independent), fish yield (dependent) vs. nitrogen input (independent), shrimp growth rate (dependent) vs. water temperature (independent), and standard curves in water chemistry (absorption (dependent) vs. concentration (independent)). Two other assumptions for regression analysis are that for any x value the distribution of y is normal, and the variance of the distribution of y is the same for any value of x (Draper and Smith, 1981; Steel and Torrie, 1980).

The simplest regression assumes a linear relationship between x and y. Equation 19 below represents total experimental variability assuming a linear relationship.

$$y = a + bx + \varepsilon \tag{19}$$

where a = y-intercept, b = slope of line (regression coefficient), and ε = sum of residuals (sum of observed minus expected values predicted from equation). Computers determine the linear equation that best fits the observed points (Figure 9) through a process known as analysis of Least Squares. Least Squares refers to the sum of the square of each individual residual (i.e., Σ (observed y_i – predicted $y_i)^2$). The regression line determined through Least Squares analysis gives the smallest sum of residuals squared of all possible straight lines. In another words, any other line would give a greater total of the sum of individual residuals squared.

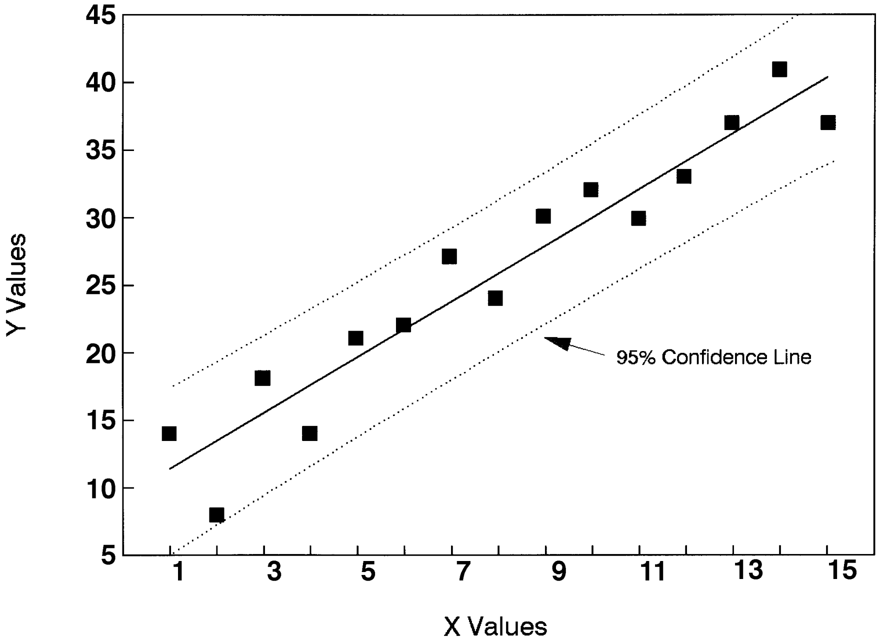


Figure 9. Linear relationship between X and Y, with both the regression line and 95% confidence limits indicated.

The linear equation provided through regression analysis is $y = a + bx$, and does not include the residual term. Implicit in this equation are several hypotheses that can be easily tested using standard information provided by most software packages when conducting a regression analysis. The ***standard error of the y estimate*** (which describes the variability of y at a given x value) can be used to test the null hypothesis that a is not significantly different from a value of 0. If $a > (t)$ (standard error of y estimate), where n = number of pairs of observations and t is the tabular value at n – 2 degrees of freedom at P = 0.05, then a is not significantly different from 0. For example, an a value significantly greater than 0 in the regression between nitrogen input and fish yield may indicate that the fish had an additional source of food not necessarily related to experimental inputs.

A second hypothesis utilizes the ***standard error of b*** (i.e., slope of regression line), and tests whether b is significantly different from 0. If it is not, then there is not a statistically significant linear relationship between x and y. To test the null hypothesis that b is not significantly different from 0, calculate a t value based on the following equation:

$$t = \frac{b}{\text{standard error of } b} \tag{20}$$

Compare the computed t value with tabular values at n – 2 degrees of freedom to determine the probability (P) that the observed hypothesized linear relationship happened by chance.

To test the null hypothesis that one linear relationship is significantly different from another (i.e., b_1 is not significantly different from b_2), calculate the 95% confidence intervals for both *b*s using the formula $b \pm (t_{0.05})$ (standard error of *b*) using the *t* value at $n - 2$ degrees of freedom. If 95% confidence limits for b_1 and b_2 do not overlap, then the two linear relationships have significantly different slopes.

An important analytical question is how well does the predicted linear equation fit the observed data points. The standard error of *b* gives some idea with respect to the actual data. If the standard error of $b = 0$, then the predicted values equal the observed values (i.e., sum of residuals squared = 0). If the standard error of *b* is relatively large in relation to *b*, then the hypothesized linear fit is not statistically significant. The ***correlation coefficient (r)*** standardizes the variability of *b* between –1 and +1, where $r = 0$ indicates no relationship at all. When $r = -1$ or +1, the standard error of $b = 0$, and *b* (slope of regression line) is either negative or positive, respectively.

Conveniently, r^2 (= ***coefficient of determination***) equals the observed variability explained by *b*, whereas $1 - r^2$ represents the remaining residual variability (i.e., ε in Equation 19). Values of r^2 vary from 0 to 1.0, with 1.0 a perfect relationship without any residual variation.

It is extremely important to understand that r^2 only gives a relative measure of the variability of *y* "explained" by the relationship with *x*, and alone does not impute statistical significance. A high r^2 does not necessarily mean that there is a significant linear relationship. As shown above, that hypothesis is based on both the standard error of *b* and the number ($n - 2$) of degrees of freedom. Similarly, statistical significance (i.e., *P*) is also based on *r* (or r^2) at $n - 2$ degrees of freedom. For example, a linear relationship is not significant ($P > 0.05$) when $r = 0.87$ ($r^2 = 0.76$) with 3 degrees of freedom (i.e., $n = 5$), but is statistically significant ($P < 0.05$) when $r = 0.28$ ($r^2 = 0.08$) with 50 degrees of freedom. In the former case 76% of the observed variation was "explained" by *b*, but there were too few observations to give sufficient confidence to the relationship. In the latter case, *b* "explained" only 8% of total variation, but the large number of observations rendered the hypothesized linear relationship statistically significant. Since the foundational experimental objective is to understand sources of variation in a particular system, the former result (although not statistically significant) gave strong support for further experimentation; whereas the latter result (although statistically significant) indicated that *x* really had very little impact on the variation of *y*.

Multiple linear regression examines the relative impacts on the variability of dependent variable *y* with several independent variables, as indicated by Equation 21.

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + \varepsilon \tag{21}$$

where b_1, b_2, and b_3 are regression coefficients for independent variables x_1, x_2, and x_3, respectively. For example, such an analysis could examine the hypothesized relationship that fish yield is a function of nitrogen input, stocking density, and survival (see Van Dam, 1990 for an example using rice-fish culture data). There are several types of multiple linear regressions available with many statistical software packages, and the reader is highly encouraged to consult proper authorities (e.g., Draper and Smith, 1981) regarding assumptions, limitations, etc. before blindly punching in numbers.

Many biological relationships are nonlinear and may be better described by ***curvilinear regression*** equations. Figure 10 illustrates representative graphs and equations for four common nonlinear relationships of quadratic, polynomial, exponential, and logarithmic. Linear regressions also include those nonlinear relationships in which (1) the model parameters are additive and multiple regression techniques are available or (2) when a transformation will make the relationship linear (Draper and Smith, 1981; Steel and Torrie, 1980). Quadratic and higher order polynomials fall into the first category, while exponential and logarithmic relationships (which can be linearized through log transformation) are examples of the second category. Nonlinear models that cannot satisfy either (1) or (2) above (e.g., logistic or Gompertz) can be analyzed by nonlinear regressions, discussion of which is beyond the scope of this chapter.

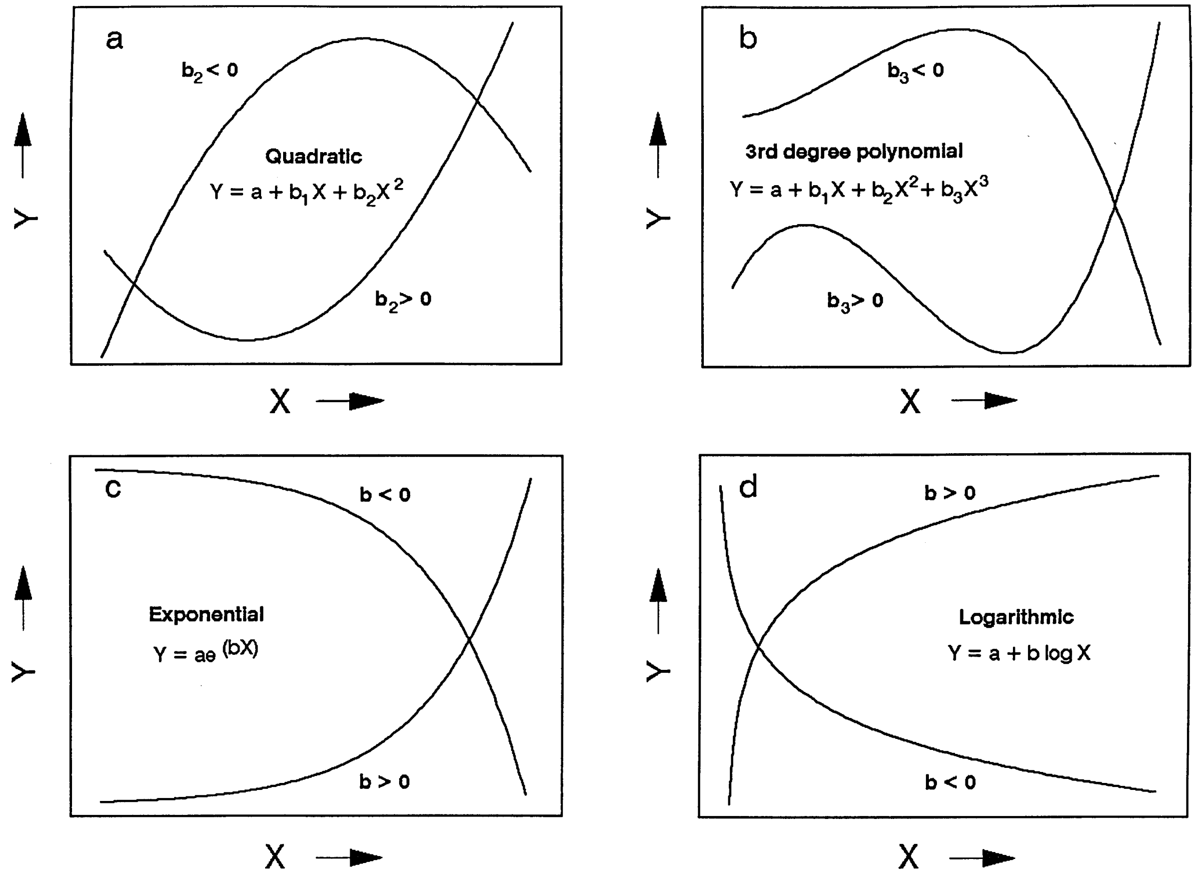


Figure 10. Nonlinear equations and examples of associated regression lines for quadratic (a), 3rd degree polynomial (b), exponential (c), and logarithmic (d) functions.

Nonlinear equations should be used *only* when sound biological reasoning supports such usage. However, because of the unfortunate and common misconception that r (or r^2) equals significance, the quadratic model (Equation 22) as well as third and fourth order polynomial equations have been reported only because the researcher found marginally higher r values by adding terms.

$$y = a + b_1x + b_2x^2 + \varepsilon \tag{22}$$

Two points require emphasizing. First, even though r may increase slightly, P could decrease slightly because of one less degree of freedom for each additional term. Second, and more importantly, P values must be determined for *each b* in the equation. For example in Equation 22, *both* b_1 and b_2 must be significantly different from 0. If b_2 is not significantly different from 0, the relationship cannot be said to be quadratic regardless of any increase in r by adding the quadratic term.

There are several other points worth mentioning before beginning any regression analysis. First, *always* plot your data points to visually see what sort of relationship you have before assuming linearity. A residual plot of independent variable x vs. residuals of y will provide a visual tool to help evaluate linearity, as well as indicate any need for data transformation (e.g., Figure 11 is a residual plot of the linear relationship in Figure 9). Second, never plot your regression line beyond your data points. The range of your data limits the predictive powers of your regression equation (e.g., Figure 9). This is one of the primary abuses of regression analysis by economic forecasters. Third, be aware that 95% confidence limits for y are not parallel to the regression line but increases (i.e., widens) toward each end of the line, giving a slightly fluted appearance (Figure 9).

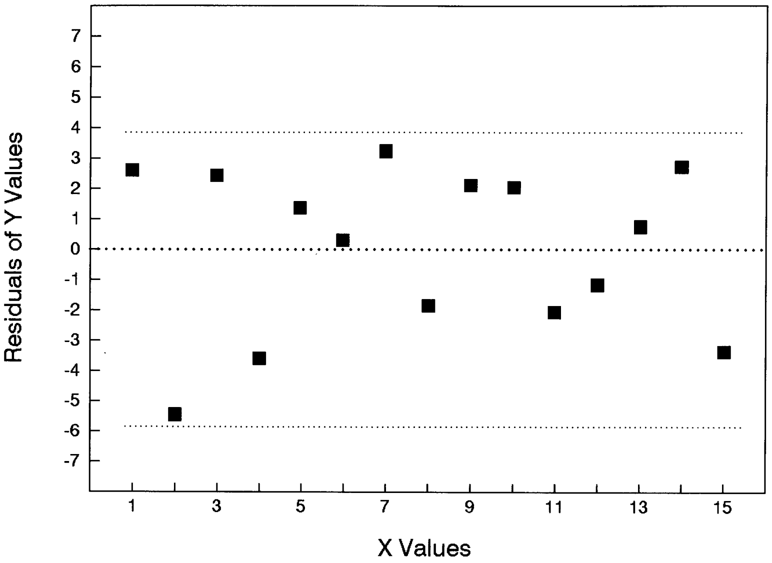


Figure 11. Plot of the residuals of Y values from Figure 9 against values of X. This residual plot illustrates no relationship or systematic pattern between X and experimental errors from predicting Y values.

Linear Correlation

Linear correlation is conceptually similar to linear regression, except there are no clear dependent and independent variables. All variables (there are more than two with multiple correlation) must be continuous, each one normally distributed, and collected from random samples (Steel and Torrie, 1980). This is different from regression analysis, where the independent variable was fixed with no variation. Like regression analysis, the correlation coefficient *r* can range from –1.0 to +1.0, with 0.0 indicating no linear relationship whatsoever.

Although correlation analysis can be quite revealing and frequently an important component of experimental data analyses, interpretation of these results should be done very carefully. Always remember that resulting probabilities and correlation coefficients only provide the degree of association; a causal relationship between variables is *never implied.* A serious danger arises when a superficially plausible explanation for an observed association is accepted without further inquiry. For example, suppose that a researcher found a nice inverse relationship between tilapia yield and ammonia concentration in ponds fertilized with TSP and urea. Without supporting evidence, it may appear that high ammonia concentrations reduced fish yield (e.g., Meade, 1985). Knud-Hansen and Batterson (1994) found such a situation, but additionally found that high ammonia concentrations were also related to low algal productivity. Ammonia accumulated in the water when algal productivity was reduced by limitations of light, phosphorus, and/or inorganic carbon. The causal relationship with ammonia concentrations was with phytoplankton productivity and not tilapia yield. Tilapia yield, on the other hand, was strongly related to algal productivity. Further analysis of ammonia concentrations showed that the toxic un-ionized component was well below toxic levels. The observed inverse association between ammonia concentrations and tilapia yield, therefore, was without any directly causal links. For another example, see Bird and Duarte (1989) regarding possible spurious correlations between bacteria and organic matter in sediments.

To guard against jumping to false conclusions of causality from observed associations, treat results from correlation analyses as a single piece to a larger puzzle. A scientific conclusion should not be based solely on a causal interpretation from a single correlation analysis. Use whatever available data to provide other pieces of the puzzle and to help understand the basis for the observed association. If no other relevant information was collected or exists, then treat the observed correlation as the foundation for a hypothesis and design further research to test it. To repeat, a significant correlation only indicates a statistically significant association and not a causal relationship between variables. Maintaining this proper perspective promotes scientific objectivity and reduces the possibility of committing Type I errors.

Multiple Comparisons of Means and Range Tests

A review of the literature would suggest that, regardless of the experimental design employed, most aquaculture scientists, primary (if not only) objective is to compare treatment means. Such an experimental objective is appropriate for unstructured experiments, but too often scientists use multiple range tests to compare treatment means for structured experiments as well, completely ignoring hypothesized relationships and interactions incorporated into the experimental design. It is clear these scientists did not appreciate the inherent power of their designs. It is also clear that the virulent nature of multiple range tests has gotten out of control via a positive feedback loop; the more scientists see multiple range tests used, the more they feel encouraged to use them (whether or not they understand implicit assumptions and limitations), and so on. Other fields of science have also been inflicted and dealt with the multiple range test plague (e.g., ecology; Day and Quinn, 1989). This subsection will hopefully be a statistical antibiotic for aquaculture scientists, providing some guidance on when and how to compare means following ANOVA.

After conducting an ANOVA, most software packages provide a table of means by treatments, blocks, etc. In some cases both the ***actual*** individual standard errors as well as standard errors ***based on the pooled variance*** (s_p^2 or MSE from ANOVA; e.g., Table 1) are given. The general formula for the latter is:

$$\text{Pooled standard error of mean} = \sqrt{\left(s_p^2\left(1/n_1 + 1/n_2\right)\right)} \qquad (23)$$

where s_p^2 = pooled variance = MSE and n_1 and n_2 = number of replicates means of treatments 1 and 2, respectively. The pooled standard error is then used to calculate confidence limits for individual treatments or blocks (see Figures 6b and 6c). If individual treatment variances are more or less similar, then using the pooled standard error to determine confidence limits for all treatment means makes sense. The difficulty arises if there is great variability among treatment variances, in which case the pooled standard error of the mean hides and possibly misrepresents observed experimental variability.

Unstructured Experiments

As discussed earlier, comparing means (with each other or with a control) is often the primary objective of unstructured experiments. The general analytical procedure is to rank the means and determine which ones are significantly different (at a given P value) from each other or the control. There are a great number of multiple comparison and range tests (e.g., the ***Least Significant Difference (lsd) test***, ***Duncan's Multiple Range test***, and ***Tukey's w Procedure***, to name a few) with which to carry out the analysis. Interested researchers are strongly encouraged to consult Day and Quinn (1989) before choosing one method over another. The wrong way to make multiple comparisons, however, is to conduct a series of individual *t*-tests.

To illustrate both the inefficiency of individual *t*-tests and general characteristics of multiple comparison tests, the ***lsd*** test will be briefly described (for a humorous explanation and defense of the lsd test, see Carmer and Walker, 1982). Assume an aquaculturist examined six unrelated types/sources of shrimp feed for their ability to increase shrimp yield. There were three replicates per treatment, and an ANOVA revealed that at least two feeds gave significantly different results. In doing individual *t*-tests, the *t* value (at a given *P* value) would be at $n_1 + n_2 - 2$ degrees of freedom, or 3 + 3 – 2 = 4 degrees of freedom (see above discussion of *t*-tests). In contrast, the lsd test uses the s_p^2 to calculate the standard error of the difference between the means (Equation 18) and the degrees of freedom for SSE. In this case, the degrees of freedom would equal k(n – 1) = 6(3 – 1) = 12. The lower *t* value with the higher degrees of freedom (2.179 vs. 2.776 at *P* = 0.05) makes the lsd much more sensitive (i.e., less likely to commit a Type II error) than a series of individual *t*-tests.

The lsd test calculates the largest difference between means that is not significant at a specified *P* value, hence the name Least Significant Difference test. For example, assuming s_p^2 = 12.1 g, the lsd value at *P* = 0.05 between any two treatments in the feeding trial would be

$$\begin{aligned} \text{lsd}_{0.05} &= t_{0.05}\left(\sqrt{\left[s_p^{\,2}\left(1/n_1 + 1/n_2\right)\right]}\right) \qquad \text{and, df} = k(n-1) \\ &= 2.179\left(\sqrt{\left[12.1(1/3 + 1/3)\right]}\right) \\ &= 6.19 \text{ g} \end{aligned} \tag{24}$$

where $\sqrt{(s_p^2\ (1/n_1 + 1/n_2)}$ = the standard error of the difference between means, and s_p^2 = pooled variance = MSE. For presentation purposes, shrimp yields (or any other response variable) would be ranked either in ascending or descending order. If descending, next to the highest yield would be placed the letter “a.” The same letter would be placed next to all lesser yields where the difference between means with the highest yield was less than 6.19 g (the lsd value). If all yields have the letter “a”, then there is no significant difference between any means at the specified *P* value. Otherwise, the first yield in descending order with a difference greater than 6.19 g when compared to the highest yield would then be designated “b.” All other yields (both ascending and descending) with means within 6.19 g will also be designated with the letter “b.” The process continues until all means are designated with a letter(s). The end result will be that all means designated with the same alphabet notation are not significantly different from each other. This provides an easy visual description of experimental results.

It is important to note that multiple range tests should be conducted only *after* an ANOVA has revealed significant differences between two or more treatments. Remember also that the lsd test, as well as other multiple range tests, provides only the most elementary form of analysis and does not provide any insight into reasons for observed differences. So even if the experiment were unstructured and a multiple range test were appropriate, more analyses may be warranted. In the feeding trial, for example, although the feeds came from independent sources they may have some identical ingredients. Was there a relationship between shrimp yield and protein or lipid concentrations or perhaps digestibility? Did any feeds have any unusual ingredients? These and many other questions may help explain your results or at least develop new research hypotheses.

Structured Experiments

The objective of structured experiments should be to evaluate system relationships. There are times in structured experiments, however, where comparing means represents an important part of data analysis. When this need arises, *never* use a multiple range test. Although the statistical community is quite clear on this point (see Chew, 1976; Peterson, 1977), the aquaculture community (among other scientific disciplines) has yet to appreciate the inappropriateness and gross inefficiency of multiple range test usage in analyzing structured experiments.

The way to compare two means in a structured experiment is with the *t*-test (Equation 17) using the pooled variance (s_p^2) to calculate the standard error of the difference between two means (Equation 18). The following is the standard equation:

$$t = \frac{\text{mean } 1 - \text{mean } 2}{\sqrt{\left[s_p^2\left(1/n_1 + 1/n_2\right)\right]}} \tag{25}$$

where n_1 = number of observations for mean 1, n_2 = number of observations for mean 2, and degrees of freedom = total for SSE in the ANOVA. This formula can be used with any design and with any pairwise comparison. Since n_1 and n_2 can be different values, the researcher can aggregate treatment or block results for more meaningful comparisons. For example, in a sixteen tank/four treatment RCBD catfish grow-out experiment with four tanks next to a dirt road as one of the four blocks (as described in the RCBD discussion in the Treatments section), a ***pairwise comparison*** of the four "road" tanks (mean 1, n_1 = 4) vs. the "non-road" tanks (mean 2, n_2 = 12) tests the null hypothesis that the road had no effect on experiment response variables (e.g., fish growth). Similar to Equation 17, Equation 25 also allows comparison of means with an unequal number of replicates, which could happen if something went wrong with an experimental unit (e.g., an aquarium accidently breaks during mid-experiment, or the wrong pond gets fertilized with a half a ton of chicken manure).

Residual Plots

Residual plots are graphs that illustrate the relationship between experimental error (i.e., unidentified variation) and some variable. The variable can be a treatment, or it can be any other measured variable such as dissolved oxygen or net fish yield. Remember that residuals are just the differences between observed and expected values. A negative residual means that the observed value for that experimental unit was less than expected. Large residuals (either positive or negative) indicate large experimental errors for those experimental units. A random residual pattern suggests that experimental errors were also random with respect to the variable that residuals were plotted against (Figure 11).

A residual plot with a discernable nonrandom pattern provides valuable insight into the nature of experimental relationships. For example, a residual plot that looks like a funnel on its side (Figure 12b) indicates increasing residual variability with increasing variable values (Figure 12a). Logarithmic transformation of the data may reduce experimental variability in such situations. Similarly, a convex- or concave-shaped residual plot (Figure 13b) may reveal a linear model incorrectly applied to a quadratic relationship (Figure 13a).

Residual analysis can be a powerful tool to identify a significant linear effect of a nontreatment variable. The relationship between experimental errors and pre-existing experimental conditions is the conceptual heart of analysis of covariance (ANCOVA). As discussed earlier, Knud-Hansen (1992a) used residual analysis to demonstrate a significant positive relationship between the amount of previous fertilization inputs to individual ponds and residual net fish yields (i.e., observed minus expected yields) for each pond in a subsequent fertilization experiment (Figure 14). Most of the experimental error could be attributed to nutrient carryover effects from previous fertilizations. Residual analysis identified previous fertilizations as a significant source of net fish yield variability and, therefore, removed this variability from experimental error.

Used in this fashion, residual analysis can (and should) become an important means for testing further hypotheses. For example, algal productivity and tilapia yields often show a strong linear relationship (e.g., Knud-Hansen et al., 1993). Tilapia yield variability not explained by a regression equation (i.e., residuals) can be plotted against hypothetical sources of variation to determine the significance of that variable. For example, tilapia yield residuals could be plotted against mean (for each experimental unit) dissolved oxygen concentrations at dawn or un-ionized ammonia concentrations to see whether any sublethal growth inhibition occurred during the experiment. Care must be taken not to test variables that also correlate with the primary independent variable. For example, where algal productivity is the independent variable, plotting the tilapia yield residuals against

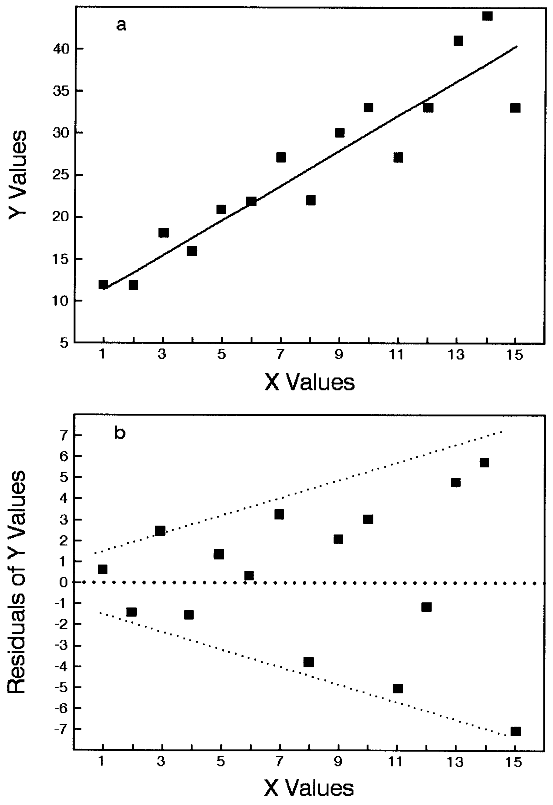


Figure 12. Example of an X vs. Y relationship where the variability of Y increases with X values. This becomes more evident when a linear regression line is made (a), and residuals of Y plotted against X show a "funnel on its side" shape (b).

suspended solid concentrations could provide confusing results, since turbidity (i.e., light availability) can affect algal productivity.

Thanks to computers, making residual plots, like testing for normality, should become a routine part of data analysis. Again, most statistical software packages will make residual plots of your results at the push of a button (or depression of a key). This is a quick way for determining if experimental errors are random or if other identifiable sources of variation are present.

Nonparametric Tests

All discussions so far have concerned parametric statistics, where sample means and variances were used to estimate corresponding parameters of normally distributed populations. In contrast, nonparametric tests make no assumptions about population distributions and have their greatest application in situations where data can be ranked rather than averaged. Surveys, which often collect qualitative data (e.g., preference rankings), frequently employ nonparametric statistics. Nonparametric statistics can also be used where the underlying distributions are not normal or when sample data have heterogenous variances, and subsequent transformations fail to adequately normalize the data. Some researchers prefer to use nonparametric tests because they are quick and easy to do, and require minimal mathematical skills.

The decision to use nonparametric statistics, however, should be made carefully and reluctantly. To begin with, they are only tests. They do not give any information regarding variances, standard errors, or confidence limits. Generally, nonparametric tests are less sensitive than corresponding parametric tests, and the researcher is more likely to commit a Type II error (i.e., accept the null hypothesis when it is false). In short, nonparametric statistics are useful only in situations when

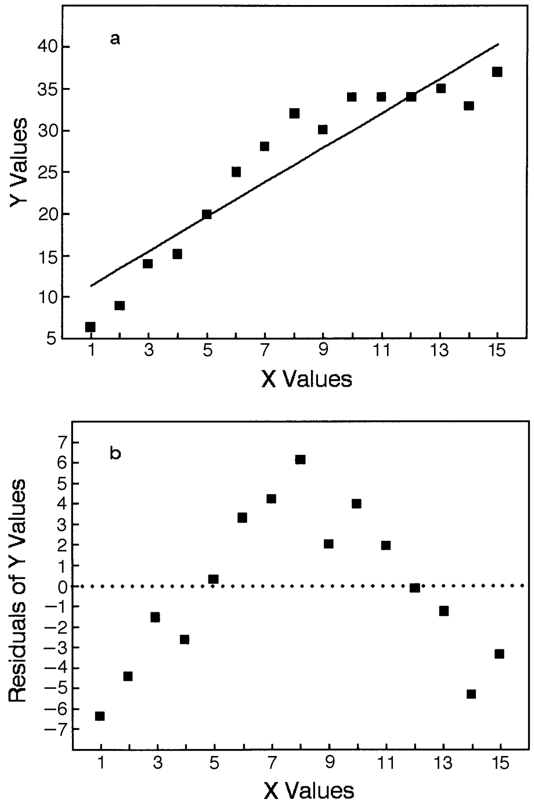


Figure 13. Example of a linear regression line plotted through a X vs. Y relationship (a). The relationship reflects more of a quadratic or asymptotic function, which is reflected in the convex shaped residual plot (b).

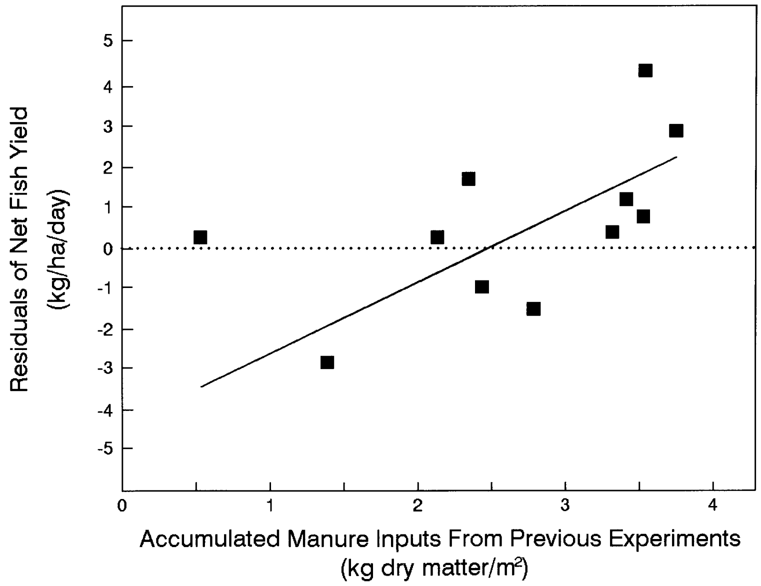


Figure 14. Relationship between the residuals of net fish yield calculated from a tilapia grow-out experiment with different fertilization inputs and total chicken manure inputs accumulated in ponds from fertilizations from previous experiments. The graph shows an overall positive effect ($P < 0.01$) on net fish yields in ponds with historically greater manure inputs. (Adapted from Knud-Hansen, C. F., 1992a.)

parametric statistics cannot be used. If parametric statistics can be used, then that should be the preferred choice.

Most statistical software packages are capable of doing a variety of nonparametric tests. The following is a representative list and description of several tests, with the analogous parametric analysis given in parentheses. The ***Kolmogorov-Smirnov One-Sample Test*** is a goodness-of-fit test that examines the null hypothesis that the distribution of your data is not significantly different from a specified parametric distribution (e.g., normal, Poisson, binomial, etc.). The ***Kolmogorov-Smirnov Two-Sample Test*** tests the null hypothesis that two independent samples have the same underlying population distribution. ***Wilcoxon's Signed Rank Test*** (paired sample *t*-test) is used for detecting differences with paired treatments (i.e., changes due to treatment within experimental units). The ***Mann-Whitney Test*** (*t*-test) compares two sample means based on ranks. The ***Kruskel-Wallis Test*** (ANOVA) compares three or more sample means based on ranks. ***Spearman's Rank Correlation*** (correlation) measures the association of two variables based on ranked observations of each variable. The calculated Spearman's rank correlation coefficient (r_s) only gives a probability of association and says nothing about the nature of the tested association.

Perhaps the most common nonparametric test used in aquaculture investigations is the ***Chi-square Test***. Chi-square analysis deals with frequency data, such as fish size distributions, and tests the null hypothesis that observed frequencies are not significantly different from expected. The greater the difference between observed and expected frequencies, the higher the Chi-square value and the less likely such differences happened by chance. Expected frequencies can be theoretical (e.g., probabilities based on genetics), or based on a given distributional frequency (e.g., normal and Poisson distributions) similar to the Kolmogorov-Smirnov One-Sample goodness-of-fit test. As with other nonparametric tests, the primary sampling requirement is that observations are made randomly and independent of each other.

Chi-square analysis can also test the null hypothesis that there is no association between two mutually exclusive classifications. For example, Chi-square analysis could be used to see if sex reversal of tilapia fry has any effect on their size distribution when they reach maturity. Frequency data are put in a contingency table with rows (r) representing classes of one variable (e.g., sex reversed or not), and columns (c) representing classes of the other variable (e.g., intervals of fish lengths). Expected frequencies are calculated from observed frequencies.

Although your actual contingency table may have classes with only a few or even no observations, it is important that the *expected* frequencies follow several general rules of thumb (Steel and Torrie, 1980). First, there should be no expected frequency < 1. Second, not more than 20% of the expected frequencies should be < 5. For example, if you make eight size classes each of sex-reversed and non-sex-reversed tilapia (i.e., a total of 16 frequencies) and find either rules of thumb violated, you have two options. You can either make more observations (i.e., measure more fish), or you can reduce the number of size classes, thus increasing the average number of fish per frequency. Since calculated Chi-square values are compared with tabular values at $(r - 1)(c - 1)$ degrees of freedom, reducing the number of classes will also reduce the test's sensitivity. Optimally, therefore, you want to make as many observations as practically feasible, and make as many size classes (e.g., class intervals every 1 cm vs. 2 cm) as possible within the above constraints.

QUALITY CONTROL

A potentially major source of experimental error comes from inadequate quality control of data collection, processing, and examination. Although most preventive measures are just plain common sense, they still merit discussion. Some useful information may be salvaged from a poorly designed experiment, but even the best designed investigation can be rendered completely worthless by carelessness and poor quality control.

Data Collection

The key here is to collect data systematically. Use the same instruments, methods, people, and times of day throughout the study. Make as few changes as possible to minimize any additional sources of variation. If a technical staff is used to collect field or laboratory data, make sure they understand the importance of both the data they are collecting and their roles in the research. Good research is frequently a team effort, and participants perform better when they feel like they are part the research team.

Essential to systematic data collection are comprehensive data sheets for all field/laboratory measurements and routine calculations. Each data sheet should provide (1) name of the project, (2) name of experiment within the project, (3) name of person(s) collecting the data, (4) date and time of day samples were collected or field measurements made, (5) date samples were analyzed/processed, (6) table headings with appropriate units, and (7) a place to add any miscellaneous notes (e.g., unusual weather, instrument malfunctions, and notable observations).

Data recording should be made with an indelible pen and should be sufficiently legible so that anyone could understand what was written (e.g., avoid confusion with the numbers 1 and 7). Unusual abbreviations should be defined on the data sheet. If a data entry needs correcting, *do not* erase or "white-out" the alleged incorrect number. Simply draw a single line through it and write the "correct" value next to it. Too often the suspected incorrect value turns out to be the correct one. And even if the correction is legitimate, the previous value could reveal a problem with the instrument or method of measurement.

Finally, file all raw data sheets systematically for future easy access. Raw data sheets represent the foundation for all subsequent data analyses. Original data sheets allow for easy detection of copying errors and discovery of systematic variations (e.g., a particular technician was later found to be routinely careless, or a reagent used during part of the experiment was later found to be incorrectly made). When appropriate, discovered errors can either be corrected or removed from further analysis. Additionally, written information that may be useful for future analyses might never enter the computer spreadsheet (e.g., miscellaneous notes, or individual fish weights and lengths when only sample means and standard deviations are analyzed).

Data Processing

Nowadays nearly all data end up on a computer spreadsheet. The *only* time data should be hand copied is when transferring data from raw data sheets to the spreadsheet. It can be quite helpful if both data sheets and spreadsheets list data column headings, treatment names, and experimental unit classifications in the same order. This foresight greatly facilitates data entry and minimizes copying errors.

No matter how careful the person, however, copying errors always occur. At a minimum, the person entering data should understand what the numbers mean. A data processor may not recognize that field data entries for ponds A1 and B1 were inadvertently reversed, or that pH values were accidently written in the DO column and vice versa, or the incongruity when on one sampling date all soluble reactive P measurements were greater than total P values for the same water samples. On the other hand, entering data is a good way for the researcher to get the feel of the numbers, sense trends, and discover errors, and thus can provide an important check on data quality control. Although time-consuming, thoughtful and careful data entry is often a cost-effective way to reduce experimental error and improve research quality.

Two further data processing considerations are the concepts of ***significant numbers*** and ***rounding numbers***. The latter can be a minor source of data bias, while abuse of the former can be a major source of misrepresentation and reader irritation. In simple terms, significant numbers refer to their perceived numerical accuracy and analytical utility. Never use more significance for a raw data value than is warranted. If the temperature meter can be read to the nearest 0.1°C, do not record a temperature to the nearest 0.01°C. It is important to understand that 25.3°C and 25.30°C may be the same temperature, but they *do not* have the same analytical meaning. The latter temperature, with four significant numbers, gives meaning (and implies confidence of measurement) to a hundredth of a degree, while the former only to a tenth of a degree. As another example, assume for illustration that three dissolved oxygen values were read and recorded as 8, 7.2, and 3.14 mg/L. Recording different levels of precision may occur between researchers or with older nondigital meters with nonlinear and/or variable scales. Because of differences in significant numbers among the recorded values, the sum may equal 18.34 mg/L but should be reported as 18 mg/L. The sum (or average) cannot be reported with greater confidence than the nearest mg/L since confidence of the first value (8 mg/L) was only to the nearest mg/L.

As a general rule three significant numbers are sufficient. Exceeding three significant numbers usually provides unnecessary numbers and may misrepresent analytical accuracy. For example, a phytoplankton count of 137,593 cells/L (i.e., six significant numbers) is better represented and more easily read as 1.38×10^5 (three significant numbers). Rarely will any scientific conclusion be based on values of greater than three significant numbers. Two common examples of analytical misrepresentation are found in water chemistry. Because some digital pH meters give measurements to the nearest 0.001 pH unit, some researchers feel obligated to report pH values to that same level of precision. Although precise, if the last two digits never stabilize, measurements are certainly not accurate to 0.001 pH units. Similarly, spectrophotometric standard curves (i.e., linear regressions) can give precise nutrient concentration calculations down to parts per trillion if so desired. However, most aquaculture water chemistry laboratories use cuvettes with 1 cm pathlengths, and minimal levels of detection are usually around 0.01 mg/L (Knud-Hansen, 1992b).

Rounding numbers should be done only *after* any and all intermediate calculations are completed. Most calculators and spreadsheet software packages have rounding functions, but they can introduce a subtle bias if they always round the number 5 either up or down. A simple way to eliminate this bias is the odd-even rule, in which rounding is always to the nearest even number (EPA, 1973). For example, 2.5 becomes 2, 3.5 becomes 4, 4.5 becomes 4, 5.5 becomes 6, and so on. Rounding should be done in one step, so if 5.451 were rounded to the nearest 0.1, the answer would 5.5 and not 5.4 (which would be the case if 5.451 were first rounded to 5.45 and then to 5.4). Rounding biases may be only a minor source of experimental error/variation, but a potential source, nonetheless.

Data Scrutiny

Data scrutiny is a quality control mechanism, a source of scientific inspiration, and an essential initial part of any data analysis (for a good review on data scrutiny, see Finney, 1988a). Data scrutiny involves the three steps of (1) a validity check on the actual data, (2) a quick examination of all possible relationships between sets of data, and (3) a check on data distributions.

The validity check requires examining the data for sources of error. Just as a gourmet chef would not cook with dirty pots and pans, the careful scientist should not analyze a "dirty" set of data. First look for extreme values (maximum-minimum spreadsheet functions make this an easy task). Misrecorded data may be discovered (e.g., a temperature of 276°C instead of 27.6°C, or a pH of 0.78 instead of 7.8), or inexplicably extreme values may appear. Do not, however, throw out extreme measurements unless you have an *independent* reason for suspecting their invalidity (e.g., the high ammonia value corresponds to the same beaker/sample in which a fly was found floating during analysis). Extreme values may reflect actual variation of your system or could also provide valuable insights into other nontested sources of variation.

The validity check also involves issues of data recording and data entry. Were there any mistakes? Were appropriate levels of precision given? Were there any changes in precision over time, which could happen if different recorders read the same instrument? For example, one person may try to read a mercury maximum-minimum thermometer to the nearest °C, while another may read it to nearest 0.5 or 0.2°C. Were there any differences due to different laboratory or field technicians, mid-experiment changes in instrumentation, or nonsystematic use of instruments (e.g., not standardizing field instruments before every use)?

The second step of data scrutiny is to examine critically all possible relationships by means of frequency distributions and scatter plots. Examine measured variables against each other and over time. This process may reveal previously unconsidered relationships, develop new hypotheses, or highlight possible experimental error. An unpublished example from a field laboratory in Thailand best illustrates the last consideration. During an experiment, diel measurements of total alkalinity in ponds consistently rose during the night and returned to lower levels during the day. This seemed peculiar since total alkalinity tends to be conservative in oxygenated freshwaters with calcium concentrations below saturation (Wetzel, 1983). At that time alkalinity was measured using a methyl orange indicator (APHA, 1985). As it turned out, this color change was harder to detect at night when the field laboratory was not as well lit; the laboratory technicians, therefore, apparently titrated with a bit more acid to make sure of reaching the endpoint. This conclusion was verified when alkalinity was subsequently measured potentiometrically (using a pH endpoint instead of a visual color change; APHA, 1985), and all diel variation (i.e., higher nighttime values) of alkalinity disappeared. Changing methods also decreased variability between replicate subsamples as well as between laboratory technicians.

The third and last data scrutiny requirement before diving into ANOVAs is to check data for distributional assumptions. This involves verifying normality of your data, or identifying other distributions and testing whether subsequent transformations adequately normalized non-normal data. At this point you are ready to conduct selected ANOVAs or perhaps resort to nonparametric analyses. It would also be wise, however, to examine residual plots from all ANOVAs to help identify possible analytical sources of variation.

CONCLUSION

Publication of Research

A logical conclusion to this chapter is a few words on writing up experimental results for publication. It seems ironic that publications represent the cornerstone of most scientific careers, yet writing papers is often the most difficult task of the overall research process. Not surprisingly, the difficulty in writing is inversely related to how well the research proposal and experimental designs were thought out prior to collecting data. Poorly focused designs, inadequate literature research, and unfounded assumptions (particularly if the researcher wants to "prove" something) generally produce papers of scientific quality comparable to the researcher's efforts.

Everyone has his or her own style and approach to writing scientific papers. Those who find writing a frustrating experience may wish to try the following approach. First, write down the two or three most significant research conclusions you wish the reader to remember a week after reading your paper (i.e., the "take home message"). Write your conclusion based on these points. As we all know, if you can remember more than two points from a research article, that paper was probably outstanding. Second, write your discussion, focusing only on your stated conclusions. How you logically present your conclusions should serve as the conceptual outline for your discussion, results, materials and methods, introduction, and abstract, so give sufficient thought to how the conclusion is written and organized. Third, write the results, presenting only the data that were discussed in the discussion and used to formulate your conclusions. If the data have no value in the discussion, they have no value in the results. Fourth, write the materials and methods, describing only those methods used to collect the data presented in the results. Fifth, write the introduction, giving necessary background information (i.e., setting the stage for your research). The introduction should include a discussion of the research problem, of prior relevant research highlighting current understanding by the scientific community, and of how the following research was designed to address one or more information gaps in this research area. A clear statement of objectives and tested null hypotheses mirroring the conclusion is a logical end to the introduction. Finally, write a clear and concise abstract, focusing on the scientific problem addressed by your research, objectives/hypotheses, and the "take home messages," along with a few notable supporting results/data. Avoid including statements detailing methods (unless methods are the focus of the paper) or statements that give no information beyond "so and so is discussed." Abstracts are read far more frequently than entire articles and should be informational. The end result of this process is a clearly focused and efficient scientific contribution.

An underlying theme of this chapter is that carefully designed research requires some serious effort up front, but this effort pays dividends when it is time to write up the results. At the time of the first measurement, the researcher already has the literature researched for the introduction and much of the discussion, has all the information for materials and methods, has the objectives and null hypotheses clearly articulated, and knows how all the data being collected will be analyzed to address these stated objectives and null hypotheses. Much of the paper could be written while the experiments are being conducted. When you have a clear vision of where you are going, it is easier to get there, which is why writing the conclusion first makes sense. It also may be useful to review one or more of the available guides to scientific writing (e.g., Gopen and Swan, 1990).

Collaborative Research

A final word should be mentioned regarding collaborative research involving two or more people. Research projects can be separated into four phases: experimental design and proposal, execution of experiment(s), data analysis, and manuscript writing. Duties and responsibilities of each participant should be clearly delineated from the outset so the partnership goes smoothly. If opportunities for routine communication are not available (e.g., researchers are based in different countries), then scheduled meetings should be written into the research protocol.

Often the most difficult decision collaborators must make are the names and order of names on subsequent publications. Preferably this decision should be made at the first organizational meeting. A general guideline requires all listed authors to have contributed substantially to at least two of the four research phases (Jackson and Prados, 1983). The lead author certainly has the discretion to add names of those who may have contributed to only one phase of the research (e.g., laboratory technicians). Adding honorary authors (often incapable of explaining or defending the paper's published results), however, is a disservice to everyone.

REFERENCES

APHA (American Public Health Association), *Standard Methods for Examination of Water and Wastewater,* 16th ed., American Public Health Association, American Water Works Association, and Water Pollution Control Federation, Washington, D.C., 1985, 1268 pp.

Baily, N. J., *Statistical Methods in Biology,* Edward Arnold, London, 1981, 216 pp.

Bird, D. F. and Duarte, C. M., Bacteria-organic matter relationship in sediments: a case of spurious correlation, *Can. J. Fish. Aquat. Sci.,* 46, 904–908, 1989.

Boyd, C. E., Water Quality in Ponds for Aquaculture, Auburn University, Auburn, AL, 1990, 482 pp.

Carmer, S. G. and Walker, W. M., Baby bear's dilemma: a statistical tale, *Agron. J.,* 74, 122–124, 1982.

Chew, V., Uses and abuses of Duncan's multiple Range test, *Proc. Fl. State Hort. Soc.,* 89, 251–253, 1976.

Cohen, J., *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Erlbaum Associates, Hillsdale, N.J., 1988, 567 pp.

Day, R. W. and Quinn, G. P., Comparisons of treatments after an analysis of variance in ecology, *Ecol. Monogr.,* 59(4), 433–463, 1989.

Draper, N. R. and Smith, H., *Applied Regression Analysis,* 2nd ed., Wiley Interscience, New York, 1981.

Edwards, P., Integrated fish farming, *INFOFISH Int.,* 5/91, 45–52, 1991.

EPA (U.S. Environmental Protection Agency), Biological Field and Laboratory Methods For Measuring the Quality of Surface Waters and Effluents; Chapter Biometrics, Weber, C. I., Ed., Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, OH, 1973, 27 pp.

Finney, J., Was this in your statistics textbook? II. Data handling, *Expl. Agric.,* 24, 343–353, 1988a.

Finney, J., Was this in your statistics textbook? III. Design and Analysis, *Expl. Agric.,* 24, 421–432, 1988b.

Gopen, G. D. and Swan, J. A., The science of scientific writing, *Am. Sci.,* 78, 550–558, 1990.

Green, B. W. and Teichert-Coddington, D. R., Comparison of two samplers used with an automated data acquisition system in whole-pond, community metabolism studies, *Progr. Fish-Cult.,* 53, 236–242, 1991.

Heath, O. V. S., *Investigation by Experiment, No. 23 in Studies in Biology,* The Camelot Press, Ltd., London, 1970, 74 pp.

Holm-Hansen, O. and Riemann, B., Chlorophyll a determination: improvements in methodology, *Oikos,* 30, 438–447, 1978.

Hopkins, K. and Yakupitiyage, A., Bias in seine sampling of tilapia, *J. World Aquaculture Soc.,* 22(4), 260–262, 1991.

Jackson, C. I. and Prados, J. W., Honor in science, *Am. Sci.,* 71, 462–464, 1983.

Kamphake, L. J., Hannah, S. A., and Cohen, J. M., Automated analysis for nitrate by hydrazine reduction, *Water Res.,* 1, 205–216, 1967.

Knud-Hansen, C. F., Pond history as a source of error in fish culture experiments: a quantitative assessment using covariate analysis, *Aquaculture,* 105, 21–36, 1992a.

Knud-Hansen, C. F., Analyzing standard curves in water chemistry, International Center for Aquatic Living Resource Management (ICLARM), *NAGA ICLARM Q.,* January, 16–19, 1992b.

Knud-Hansen, C. F. and Batterson, T. R., Effect of fertilization frequency on the production of Nile tilapia (*Oreochromis niloticus*), *Aquaculture,* 123, 271–280, 1994.

Knud-Hansen, C. F., Batterson, T. R., and McNabb, C. D., The role of chicken manure in the production of Nile tilapia (*Oreochromis niloticus*), *Aquaculture Fish. Manage.,* 24, 483–493, 1993.

Likens, G. E., Importance of perspective in limnology, in *An Ecosystem Approach to Aquatic Ecology,* Springer-Verlag, New York, 1985, 84–88.

Meade, J. W., Allowable ammonia for fish culture, *Progr. Fish-Cult.,* 47(3), 135–145, 1985.

Parker, R. E., *Introductory Statistics for Biology, No. 43 in Studies in Biology,* The Camelot Press, Ltd., London, 1979, 122 pp.

Peterman, R. M., Application of statistical power analysis to the Oregon coho salmon (*Oncorhynchus kisutch*) problem, *Can. J. Fish. Aquat. Sci.,* 46, 1183–1187, 1989.

Peterson, R. G., Use and misuse of multiple range comparison procedures, *Agron. J.,* 69, 205–208, 1977.

Searcy-Bernal, R., Statistical power and aquacultural research, *Aquaculture,* 127, 371–388, 1994.

Steel, R. G. D. and Torrie, J. H., *Principles and Procedures of Statistics,* 2nd ed., McGraw-Hill, New York, 1980, 633 pp.

Strickland, J. D. H. and Parsons, T. R., A Practical Handbook of Seawater Analysis, Fisheries Research Board of Canada, Ottawa, 1972, 310 pp.

Van Dam, A. A., Multiple regression analysis of accumulated data from aquaculture experiments: a rice-fish culture example, *Aquaculture Fish. Manage.,* 21, 1–15, 1990.

Wetzel, R. G., *Limnology,* 2nd ed., Saunders Publishing, Philadelphia, 1983, 767 pp.

Wetzel, R. G. and Likens, G. E., *Limnological Analyses,* Saunders Publishing, Philadelphia, 1979, 357 pp.

Yoccoz, N. G., Use, overuse, and misuse of significance tests in evolutionary biology and ecology, *Bull. Ecol. Soc. Am.,* 72(2), 106–111, 1991.