

An Overview of Experimental Design

I. Hypothesis Testing

While much research in Biology consists of data collection for descriptive purposes, there is a burgeoning trend toward collecting information with the hope of answering particular questions or to recast information collected for descriptive purposes in light of particular hypotheses. This is due to a number of causes, among them; the growing body of descriptive information on all aspects of natural history and the inappropriateness of existing data sets to answer specific questions. It may also represent an increase in the awareness of scientists of the logical structure of Scientific Method. Whatever the cause, the effect is that biologists are asking questions and designing their research efforts to answer questions. Hence, our concern with asking answerable questions and for developing procedures upon which to base probabilistic inferences regarding the answers to these questions. However, before we discuss particular procedures and their application, some discussion of the form of the questions we ask and the possible outcomes of our attempts to answer a particular question is warranted.

A. Null, tested, and alternative hypotheses

When presented as a statement, rather than a question, the question of interest in a particular research program is called the "tested hypothesis." It may be either of the general form; "Factor A is responsible for phenomenon B," or "Factor A is not responsible for phenomenon B." This latter form, which is the negation of any relationship between Factor A and phenomenon B is also called a "null hypothesis." Each of these hypotheses can serve as the tested hypothesis or as the corresponding alternative hypothesis. The alternative hypothesis is that set of hypotheses implied by the rejection of the tested hypothesis. Once again, a null hypothesis is simply an hypothesis of "no effect" or "no relationship" between some factor of interest and some observable phenomenon.

B. Intimacy, advocacy, and impartiality in hypothesis testing

Given a specified tested and alternative hypothesis, with what goal in mind should evidence be gathered to "test" the hypothesis. Should our "test" be an attempt to find evidence consistent with our tested hypothesis or to find evidence inconsistent with our tested hypothesis? Do we wish to prove or disprove our tested hypothesis? As a linguistic convenience and as a widespread practice, many scientists strive to prove particular hypotheses, or at least say they do. However, for both logical and practical reasons it may be desirable to strive to falsify the tested hypothesis.

First, the practical reasons. If the researcher attempts to prove the tested hypothesis by gathering evidence in its support, they run the risk of ignoring evidence that might

controvert the tested hypothesis. Furthermore, if the tested hypothesis is not a null hypothesis, the process of attempting to prove the tested hypothesis amounts to advocating the tested hypothesis. This may lead the researcher to develop "pet theories" and seek to support them rather than exposing these theories to critical scrutiny. Such intimacy between researcher and hypothesis can impair the researcher's ability to cast aside favorite, yet untenable hypotheses.

Logically speaking, Hume and Popper illustrate the asymmetry that exists between proof and disproof. The weight of evidence found consistent with a particular hypothesis is no match for the instance inconsistent with that hypothesis. That single instance is sufficient to falsify the hypothesis where no amount of evidence consistent with a hypothesis is sufficient to prove it. So to insure an impartial, detached evaluation of competing hypotheses and to efficiently and rigorously assess the relative merits of competing hypotheses, one should attempt to gather evidence capable of falsifying the tested hypothesis, not evidence designed to prove the tested hypothesis.

C. The statistical hypothesis, error, and power

A **statistical hypothesis** is a statement about a statistical population that, on the basis of information obtained from the observed data, one seeks to refute. A **statistical test** is a set of rules whereby a decision about the hypothesis is reached. Associated with the **decision rules** is some indication of the accuracy of the decisions reached by following the rules. The measure of accuracy is a statement about the probability of making the correct decision when certain conditions are true in the population in which the hypothesis applies. The accuracy of the decisions based upon information supplied by an experiment depends to a great extent upon the design of the experiment. The decision rules are set up by the experimenter and depend upon what the experimenter considers to be the critical bounds for arriving at the wrong decision. However, the statistical hypothesis does not become false when it exceeds the critical bounds, nor is it true when it does not exceed the bounds. The decision rules are merely guides in summarizing the results of the statistical test - following the guides enable the experimenter to attach probability statements to their decisions. The probability statements associated with the decision rules of the statistical test are predictions as to what may be expected to be the case if the experiment were repeated many times.

The logic of a statistical hypothesis test is as follows:

1. One assumes the tested hypothesis to be true
2. One examines the consequences of this assumption in terms of a sampling distribution that depends upon the truth of this hypothesis.
3. If, as determined from the sampling distribution, the observed data have relatively high probability of occurring, the decision is made that the data do not contradict the hypothesis.

4. If the probability of an observed data set is low when the tested hypothesis is true, the decision is made that the data contradict the tested hypothesis.
5. Again, the tested hypothesis is often stated in such a way that when the data contradict it, the experimenter has demonstrated the presence of some experimental effect. The experimenter has been able to nullify the tested hypothesis, in favor of the alternative hypothesis that some effect is detectable. Voila, a null tested hypothesis.

The level of significance, α , defines the probability level that is too low to warrant support of the tested hypothesis. It is one of the decision rules. If the probability of occurrence of the observed data (when the tested hypothesis is true) is smaller than the level of significance, then the data contradict the hypothesis being tested, and the decision is made to reject the tested hypothesis. This rejection is equivalent to supporting one of the possible alternative hypotheses that are not contradicted by the data. If the tested hypothesis is symbolized by H_0 , and the set of alternative hypotheses that remain tenable when H_0 is rejected is H_a : then the decision rules in a statistical test can be specified with respect to rejection or non-rejection of H_0 :

1. The rejection of H_0 may be regarded as the acceptance of H_a .
2. The non-rejection of H_0 may be regarded as a rejection of H_a .

If the decision rule rejects H_0 when it is in fact true, the rule has led to an erroneous decision. The probability of making such an error is at most equal to α , the level of significance. This kind of error is known as a Type I error; rejecting the tested hypothesis when true.

If the decision rules do not reject H_0 , when it is in fact false, it also leads to an erroneous decision. This kind of error is known as a Type II error; failing to reject the tested hypothesis when it is false. The potential magnitude of a Type II error depends in part upon the level of significance of the test, and in part upon which of the alternative hypotheses the data actually supports. Associated with each possible alternative hypothesis is a different probability of a Type II error.

Type I errors can only occur if the decision is made to reject H_0 and Type II errors may occur when the decision is made to not reject H_0 .

The probability of making a Type I error is under the direct control of the experimenter, since the experimenter sets the level of significance, α . However, Type II error is controlled indirectly, primarily through the design of the experiment. If possible the tested hypothesis is stated in such a way that the more costly error is the Type I error, since its magnitude can be directly controlled by the experimenter. This is why the tested hypothesis is often stated as a null hypothesis, since rejection of the tested null hypothesis when it is true amounts to finding an experimental effect when none exists. Such an error could have a great impact on a research program since it will most likely

lead the experimenter to consider the question answered. Better to be conservative and fail to find an experimental result. The experimenter would then be forced to repeat the experiment, possibly with modifications, or to perform other experiments to test the same hypothesis. Err on the side of innocence.

Nevertheless, it is best, to try to minimize both sources of error. However, Type I and Type II errors are not independent. The smaller the probability of a Type I error, α , the larger numerically the potential Type II error can be.

The relationship between Type I and Type II errors can be best represented graphically. The rejection region for H_0 is defined relative to the sampling distribution of the statistic of interest when H_0 is true (solid line, α). The dashed line represents the sampling distribution of the same statistic when a particular alternative hypothesis is true, H_a . β , the probability of a Type II error is that area under the dashed curve that lies within the region of non-rejection of the sampling distribution of the statistic when H_0 is true.

Part b illustrates the effect of altering α on the value of β . If α is smaller, the area of the dashed curve that falls within the region of non-rejection of the sampling distribution when H_0 is true is larger, hence β is larger when α is smaller.

The power of a test is equal to $(1 - \beta)$. The power is the area under the dashed curve that falls in the region of rejection of the sampling distribution when H_0 is true. Since β is the probability of failing to reject the tested hypothesis when false, $(1 - \beta)$, the power is the probability of rejecting the tested hypothesis when false.

$$1 = \underset{\text{Power}}{P(\text{rejecting when false})} + \underset{\beta}{P(\text{failing to reject when false})}$$

$$\text{Power} = (1 - \beta)$$

Power is the probability that the decision rule rejects H_0 when a specified H_a is true. The closer the H_a to H_0 (the greater the overlap in the corresponding sampling distributions) the lower the power of the test with respect to that particular alternative. A well-designed experiment will both be conservative, have low α , and have high power, β , with respect to all alternative hypotheses which are in a practical sense different from H_0 . For an H_0 of $\mu_1 = \mu_2$, and an H_a of $\mu_1 = \mu_2 + 0.001$, this H_a may not be, for all practical purposes, a different hypothesis than H_0 . Hence power with respect to this alternative is of no practical consequence.

The most common means to minimize the probability of a Type II error and to increase the power of a test relative to all possible reasonable alternative hypotheses, for fixed α , is to increase the sample size or replication in the experiment. This is because the dispersion of the sampling distribution of a statistic decreases by a factor of $(1/\sqrt{N})$. Hence, the overlap in the sampling distributions of the tested and alternative hypotheses decreases as N increases.

While an inordinate emphasis has been placed on the level of significance of a statistical test, the power of a test is usually ignored. This is partially due to the reluctance of experimenters to present results in which a tested null hypothesis could not be rejected, and to a lack of information as to what constitutes a reasonable alternative hypothesis. While it is easy for the experimenter to control Type I error, and the experimenter may wish to perform a conservative test (minimizing the probability of a Type I error), these tests may suffer from a complete lack of power to discriminate between the tested hypothesis and reasonable alternative hypotheses. If this is true then the best solution may be to allow higher Type I error in order to increase the power of the test versus fixed alternatives. In such instances α values of 0.1, 0.2, or even 0.3, may be reasonable.

However, given a properly designed experiment in which the power of the test has been investigated for specified alternatives for specified α values, any desired power can be obtained simply by providing an adequate sample size, even though this may be very expensive. Given an initial preliminary survey the necessary sample size to discriminate a particular alternative hypothesis can be estimated. Green (1979, p.43) outlines such a procedure for a 2 x 2 factorial experiment in which a test of the interaction between time and treatment was the statistical hypothesis of interest.

D. The statistical hypothesis test and its relationship to ruling hypotheses or theories

In practice our hypothesis test usually involve a specific set of observations regarding the effects of a particular factor (say, soil moisture) on a particular response variable (say, crop yield). If our tested hypothesis is in the form of a null hypothesis, H_0 : Soil moisture content has no effect on crop yield, a logical alternative hypothesis might be that; H_a : Soil moisture content effects crop yield. Two outcomes are possible from our test, we can either reject or fail to reject the tested hypothesis. If we reject our null hypothesis, and our observations were derived from controlled experiments so that only soil moisture was allowed to vary among replicate fields, then we might safely conclude that soil moisture does affect crop yield, or at least that our measures of soil moisture and crop yield would suggest so. However, if we fail to reject our null hypothesis we cannot conclude that soil moisture has no effect on crop yield, rather only that from the data at hand one is not compelled to posit that it does. Possibly a better-controlled or designed experiment would have succeeded in falsifying the tested null hypothesis. How do these experimental outcomes relate to the general hypothesis that soil moisture effects plant growth and yield? In the instance where we rejected the null hypothesis we have demonstrated that this appears to be true for one crop, given our experimental design. In the instance of failing to reject the null hypothesis we are back to the drawing board to design a more critical experiment. In either event, a single experiment is insufficient to lead us to believe that the truth content of the alternative or tested hypothesis is high. Repeated attempts to reject or failure to reject such an hypothesis with more and more critical experiments are necessary to establish its verisimilitude. It is seldom that a single experiment will have a major impact on how scientists in a discipline view a ruling theory.

II. A Typology of Evidence

A. Non-experimental research

1. Data-dredging

Data dredging, as described by Selvin and Stuart (1966), occurs in the process of examining data sets that are often collected for other purposes. If an hypothesis and an hypothesis test are stated prior to the examination of the data, and the results of the test, regardless of the outcome are reported, then data dredging can be useful and result in a considerable savings in effort. Why collect more data to test an hypothesis if an adequate set already exists?

However, if a specific quantitative hypothesis is not stated in advance, but rather emerges during the data analysis, perhaps along with a novel "test" variable, then the strength of the test is compromised, since a mechanistic explanation for the result must also be developed *a posteriori*. In the initial stages of study on a new topic this may be a useful process to help develop new hypotheses and to formalize critical hypothesis tests. But, when the topic has been the subject of much study, a specific *a priori* hypothesis should be available for testing.

Care must also be taken when engaging in three other types of data dredging; "snooping", "fishing", and "hunting".

Snooping is testing a large set of hypotheses. The problem arises since some tests are expected to be significant by chance alone, and because the hypotheses may not be independent. These problems also occur with experimenter generated data sets.

Fishing is choosing test variables based on an examination of the data rather than because of their importance to an *a priori* hypothesis. Also, by relegating variables to two classes, those chosen and those discarded, the interpretation of the results are clouded. In the absence of a specific *a priori* hypothesis, why not report the tests for all variables?

Hunting is the process of searching through many data sets to find some relationships worth testing. We never know how many data sets were found not to display the desired relationship since negative results are seldom reported.

2. Uncontrolled Observations

By uncontrolled observations I mean observations on a test variable under experimental conditions which cannot be compared to a set of observations obtained on the same test variable in the absence of the experimental conditions. Uncontrolled observations are often, but not always, experimenter generated. They arise either because of poor experimental design, or because of the nature of the problem under study. An example of this latter problem can be seen in studies of trends in global atmospheric chemistry.

These studies postulate recent anthropogenic changes in atmospheric chemistry based on observations on present conditions and knowledge of the increase in anthropogenic inputs. The only available control observations are theoretical or budgetary predictions removing the anthropogenic inputs, or a reconstructed fragmentary historical and prehistoric record. While these might be the only sorts of "control" observations possible in these circumstances, the absence of good estimates of experimental and control parameters and their variance makes rigorous hypothesis testing difficult. Similar problems plague explanations of the impact of introduced plants and animals on native populations. Usually no information is available on population trends in the native biota prior to the introduction of the exotic species.

The only cure for uncontrolled observations is to generate controlled observations. This can be accomplished by repeating an experiment with proper controls, if possible, generating theoretical expectations of the control observations by either deterministic models or by Monte Carlo procedures, or by seeking observations that may be considered "natural controls." While the first of these three alternatives is most desirable, it is often not possible. Using theoretical expectations is only as reasonable as the existing theory. The use of natural controls is plagued with the problem of determining if the so called "natural control" observations were recorded under conditions that truly differ from the experimental conditions in only that the experimental conditions are absent. Ecologists have tried for the last two decades to use natural controls in so called "natural experiments". However, in most cases no effort was made to demonstrate the similarity, in all but the experimental conditions, of the control and experimental observations.

B. Experimental Evidence

Experimental evidence is data collected by the experimenter for the express purpose of answering a particular question or to test a particular hypothesis. I do not mean to suggest that all experimental evidence is created equal as a basis for causal inferences. In fact, I can discern several kinds of experiments that in the order I will present them represent an increasing degree of intervention on the part of the experimenter into the workings of nature. And, I believe an increasing ability to intimately connect cause and effect.

1. Controlled observations

Controlled observations are collected by design to test a particular hypothesis. The design includes samples under the experimental conditions of interest and under putative control conditions (lacking the experimental treatment). However, the observations are derived from a sampling program that involves nature only passively. The only activity of the experimenter is to make the observations, analyze the data, and interpret the result. For example, I conducted a sampling program to determine if the amount of folivory on White Oak trees is related to the timing of leafing and leaf development in the spring. Much of the folivory that occurs on woody plants, in general, occurs during spring when leaves are young and supple. If the amount of damage

received by Oak trees is determined by the age of the foliage relative to the emergence time of leaf feeding insects, then trees that either foliate sufficiently prior to insect emergence and feeding, or after insect emergence may receive less damage. To test this hypothesis, I sampled groups of trees similar in size, but differing in the timing of foliation. Three classes of foliation were established, early, mid, and late, which by making inter-comparisons serve as corresponding controls as well (mid and late serve as controls for early, etc.). In this "experiment", I have intervened only to record my observations on folivory and foliation. I have neither altered the leafing time of the trees to observe the subsequent damage received, nor have I manipulated the herbivores to create or destroy a pattern of synchronous emergence and foliation. The key to distinguish controlled observations from a more elaborate experiment is the passive role of both nature and the experimenter.

In as much as the experimental observations have good control observations, such sampling programs can be a reasonable basis for causal inferences. However, the danger exists that the control observations are not true controls, since the experimenter may not be able to insure that subjects used for experimental observations differ only in the experimental treatment from the subjects used in the control observations. In the above example it is conceivable that some aspect of leaf chemistry, either nutritional quality, the concentration of volatile compounds (which may be used in host location and stimulate feeding) or the concentration of chemicals that deter feeding may co-vary with time of foliation. These chemical characteristics may be the proximal causal agents responsible for any observed relationship between foliation and folivory. Foliation time may either affect these characteristics or simply, as I said, co-vary with them. In either instance, further experiments would be necessary to determine the actual causal mechanism.

Therefore, although more convincing than inferences based on uncontrolled observations it is still difficult to base firm causal inferences purely on controlled observations.

2. Mensurative Experiments

The next step in experimental interventions I call mensurative experiments (*sensu* Hurlburt 1984). They involve the experimenter and a part of nature a bit more actively in the hypothesis test, but only to passively measure another part of nature. A common example of a mensurative experimental technique in ecology is the use of litter bags to examine the rate of litter decomposition in aquatic environments or on the forest floor. The experimenter packages a bit of nature in these litter bags (which have mesh that allow colonization by bacteria and invertebrates who along with chemical weathering are responsible for the litter decomposition) and exposes the bags under different experimental conditions or times to determine if significant variation in decomposition rates is detectable between the experimental conditions of interest. In this case, the experimental intervention into the workings of nature is solely to create a replicable sampling device with which to passively measure a natural process.

As with controlled observations, mensurative experiments suffer because we cannot uniquely associate a causal mechanism with the variety of experimental situations into which we have placed our mensurative device. Were we to find different rates of litter decomposition in temperate and tropical forests, to what would we attribute these differences? The list of reasonable causes is lengthy. Once again further experiments are necessary to establish the specific causal mechanisms involved.

3. Manipulative Experiments

In a manipulative experiment, the experimenter may exercise total control over a portion of nature to create all the desired experimental and control conditions. To repeat the folivory experiment mentioned above as a manipulative experiment would involve direct modification of the foliation time and exposure of all trees to identical herbivore populations, possibly via massive rearing and release of leaf-eating insects. The litter decomposition experiment would involve a series of experiments in which several factors such as temperature, humidity, bacterial populations, fungal populations, and invertebrate populations are controlled or allowed to vary singly or in combination to test for simple effects and interactions of factors in determining decomposition rates. Obviously, it is often easiest to perform manipulative experiments in laboratories or experimental enclosures where there is some hope of success in actually controlling the multitude of environmental and biological variables.

If it is in any way possible to perform a manipulative experiment, it is preferred over the previously mentioned kinds of experimental exercises. However, it is extremely difficult to perform these kinds of experiments under field conditions. Manipulative experiments do, however, have a greater ability to associate cause and effect since if properly designed and executed they remove the possibility of co-varying potential causal factors.

Manipulative experiments, even where logistically feasible, are not without their problems. The most important of which is the danger of introducing some experimental artifact via some aspect of the manipulation. This problem also besets mensurative experiments, but to a lesser degree since the experimental intervention in nature is less drastic.

III. Allocation of sampling effort

A. What is a sampling program to do?

A statistical population is the collection of all elements about which one seeks information, or about which one desires to make some inference. It is incumbent upon the experimenter to state *a priori* the population of elements about which they wish to make some statement. It is crucial that this population is defined in advance of designing a sampling program or experiment. The reason for this will be apparent, shortly.

Usually only a small portion, or a sample, of this population can actually be observed. It is from data on the elements of the population that are members of the sample that conclusions or inferences are drawn about the characteristics of the entire statistical population. Quantities computed from sample data are commonly termed statistics while those characterizing populations are known as parameters. Sample statistics then serve two roles:

- 1) to describe the data obtained in the sample
- 2) to estimate or test hypotheses about characteristics of the population

Had we enumerated the values of a particular characteristic for all elements in a population, and then tabulated the frequencies with which the elements of the population take on different values, the resulting tabulation would be the population distribution of the character of interest. The population distribution can be described by this sort of enumeration or by a series of parameters. The number of parameters necessary to describe a particular distribution depends on its form, but is in general a more parsimonious approach than enumeration. For example, the Poisson distribution can be specified by one parameter and the normal distribution by two parameters.

If we are interested in estimating a population mean, μ , the sample mean, \bar{x} , generally provides a good estimate. Similarly the sample standard deviation, s , provides a good estimate of the population standard deviation, σ . The precision of these estimates depends on four factors:

- 1) the size of the sample
- 2) the manner of sampling
- 3) the characteristics of the underlying population
- 4) the principle used in estimating the parameter

If a sample is drawn such that:

- 1) all elements of the population have an equal chance of being drawn at all times, and
- 2) all possible samples of size n have an equal (or fixed and determinable) chance of being drawn,

then, the sample is a random sample of size n from the underlying population. Of course to meet the conditions the population must be defined in advance or a sampling program that insures that all elements have an equal chance of being drawn cannot be designed.

Random samples insure that all elements in a population are at equal risk of being sampled and that the probability of sampling any individual element from the population is independent of which other elements may or may not be sampled.

Suppose 10,000 samples each of n elements are drawn from a population. Sample means, \bar{x} , and variances, s^2 , could be computed from each sample. The tabulation of the frequencies with which our sample statistics take on different values is the sampling distribution of the statistic. In this instance, we have determined the distribution empirically. The form of these distributions depends in part upon the sampling method. As with population distributions sampling distributions can be described more economically with parameters than by enumeration. Frequently the parameters of the sampling distribution of a statistic are related to the parameters of the underlying population. The mean or average value of the sampling distribution of a statistic and its standard deviation is the standard error of the statistic. The form of the sampling distribution as well as the magnitude of its parameters depend on:

- 1) the form of the population distribution,
- 2) the manner of sampling, and
- 3) the size of the sample.

If, for example, the underlying population is normally distributed and the samples are random samples, then if one draws a large number of samples, the sampling distribution of the sample mean, \bar{x} , will be approximately normal with mean, μ and standard error s/\sqrt{n} . This same consequence is derivable mathematically based on the properties of random samples. This is the importance of random samples - they permit the estimation of sampling distributions from purely mathematical considerations without necessitating the laborious kinds of enumerations I have mentioned. The key aspect of random sampling which allows this is that random samples ensure all elements of the population are at equal risk of being sampled and the probability that any single element is sampled is independent of which other elements are sampled. Statistics obtained from samples drawn by other sampling plans which are not random have sampling distributions which are either unknown or which can only be approximated with unknown precision. Good approximations to sampling distributions required if one is to evaluate the precision of the inferences made from sample data.

Of course, the Central Limit Theorem generalizes this result for populations that are non-normal. The sampling distribution of a statistic derived from a non-normal population is also approximately normal and the approximation improves as the sample size increases. For the sample mean, \bar{x} , its expected value is still, μ and its standard error, is s/\sqrt{n} .

In the context of hypothesis testing, the role of sampling is to enable the experimenter to discover something about the sampling distribution of the statistic of interest, under the

experimental conditions of interest, based on the underlying population of interest. This is because the statistical hypothesis test is based upon the sample estimates of the parameters of the sampling distribution of the statistic, not upon the sample estimates of the underlying population parameters. The close relationship between the parameters of the sampling distribution of a statistic and the population parameters estimated by the sample statistic tends to obscure this fact. The statistical hypothesis test involves a comparison of sampling distributions. Upon this comparison inferences about population distributions and their parameters can be made.

B. Bias, precision and random sampling

So we sample to estimate population parameters and to learn through knowledge of the sampling distribution of our statistics just how good our estimates are. Two criteria are commonly used to judge the accuracy of an estimate: bias and variance.

A statistic is an unbiased estimate of a parameter if the expected value of the sampling distribution of the statistic is equal to the parameter of which it is an estimate. Bias therefore, is a property of the sampling distribution not of a single statistic. This implies that in the long run the mean of a statistic computed from a large number of samples of equal size will be equal to the parameter, if it is unbiased. In addition to insuring that the elements included in a sample are independent, random sampling also helps to prevent biasing estimates of population parameters. If all elements in a population were not at an equal risk of being sampled it is easy to see that values systematically above or below the true population value may be represented disproportionately in the sample.

The precision of an estimator is measured by the standard error of its sampling distribution. The smaller the standard error the greater the precision. The standard error is only a good measure of precision if the sampling distribution is asymptotically normal. If this is true, then the best-unbiased estimator is the one with the smallest standard error. This is called a minimum variance unbiased estimator. Increasing sample size will also increase the precision of an estimator.

C. The preliminary survey

Once the statistical population of interest has been defined, the attributes to be examined are selected, and the experimental conditions decided upon, the experimenter is left with the task of deciding where to invest sampling effort. First and foremost, this is dictated by the questions of interest. Assuming limited resources, there is no reason to expend extra effort to test ancillary hypotheses that are not of pressing interest. It is easy to compromise all the hypothesis tests you wish to perform by attempting to design an all encompassing sampling plan which allows you to test several hypotheses, but none with any power. You just cannot answer all the important questions in biology in one MS. or Ph.D. thesis. Believe me, I tried. State the questions you wish to answer and rank them in importance. If the cost of sample collection or processing in time or money, or the inherent variation in the attributes of the populations you wish to study are high then pare down the number of questions so that at least

some can be answered with adequate confidence and power.

Second try out your sampling gear to assess its accuracy and to estimate the cost per sample. In nature you can rest assured that appearances will be deceiving and that field work always costs more in time and money than anticipated. If you are using some sort of sampling gear that you cannot normally observe during its operation, try to observe its behavior at least once. If there is any subjective component to sample selection or any other aspect of the collection, sorting, or enumeration of samples have more than one observer repeat the same procedure to see if any systematic bias is being introduced.

Third, carry out a preliminary survey so that you can estimate the amount of variation to be expected under each set of experimental conditions. If you know where the variation lies in your subject populations you can increase your replication to improve the precision of your estimates and thereby (by decreasing the standard error of the sampling distributions) increase the power of your tests against fixed alternatives for fixed values of α .

D. Optimal allocation of sampling effort

As I mentioned before, the precision of our estimates of population parameters depends upon the form of the population distribution, the manner of sampling, and the size of the sample. The experimenter has control only over these last two aspects. So the allocation of sampling effort must involve variations in the manner of sampling and the size of the sample.

Sample size

Increasing sample size will increase the precision of our estimates by decreasing the standard error of our sample statistic. Increasing sample size should not decrease our estimate of the standard deviation of the underlying population. For fixed α , this increases the power of an hypothesis test against all alternatives. For example, the sample size required to be 95% confident that our estimate of the sample mean lies within an allowable error, L , of the true population mean is:

$$n = \frac{4s^2}{L^2}$$

where n is sample size and s is estimated by the sample standard deviation. Increasing precision is synonymous with decreasing the allowable error, L , and for fixed confidence we must increase n to achieve increased precision.

Sampling manner

The results concerning sampling distributions that I mentioned earlier hold for other types of sampling than just simple random sampling. It is sufficient for the sampling

method to sample all elements independently and with known probabilities. These probabilities need not be equal for all elements of the population (as in simple random sampling), as long as we take account of these probabilities when constructing our estimates. Sampling plans that follow these criteria are known as probability sampling. Simple random sampling being the most common of these. Two other commonly used methods of probability sampling are stratified sampling and 2-stage sampling.

Stratified sampling involves dividing a population into a number of parts, called strata, drawing simple random samples from each strata, and computing the parameter of interest as a weighted mean of the parameter estimates from each strata. For the sample mean we have

$$\bar{x}_{st} = \frac{\sum_{h=1}^k N_h \bar{x}_h}{N},$$

where n is the total number of elements in the h th stratum, \bar{x}_h is the sample mean for the h th stratum and $\sum_{h=1}^k N_h$ is the size of the population. Stratified sampling is useful

because differences between strata means do not contribute to the standard error of the mean, \bar{x} . That is, the sampling error arises solely because of variation among elements within strata. If we can stratify an otherwise heterogeneous population into strata which are fairly homogeneous, we can increase the precision of our estimate over that achievable by simple random sampling. The size of the sample we choose in any stratum is determined by the experiments. This freedom of choice allows the experimenter to allocate sampling effort efficiently. This control over the allocation of sampling effort is often the principal reason for the gain in precision derived from stratification. If equal fractions of the elements in each stratum are sampled the weighting factors are equal for all strata and we need not modify our sample statistics to account for the unequal probabilities of sampling elements in different strata. This is known as stratified sampling with proportional allocation of sampling effort. The optimum allocation of sampling effort in a stratified design is not necessarily a proportional allocation program where n_h/N_h is equal for all strata. Rather the optimal solution is to take n_h elements proportional to $N_h s_h / C_h$, where s_h is the within stratum standard deviation, and C_h is the cost per sample in the h th stratum. This method gives the smallest standard error of the estimated sample statistic for a given total cost of sampling. In other words, take a larger sample in a stratum that is unusually variable (s_h large), and a smaller sample where sampling is unusually expensive (C_h large). If the within strata standard deviations are all approximately equal and the cost of sampling in each strata is also equal then this method reduces to the method of proportional allocation. Of course, in order to allocate effort optimally rough estimates of standard deviations and costs must be made.

Two stage sampling

In a two stage sampling program the sample is derived by first collecting a sample of primary sampling units, and then by sub-sampling within each of these units. The oak tree experiment I described earlier is an example of a two-stage sampling program. The primary units are the trees selected from the forests and the sub units are the leaves or leaf clusters sampled within a tree. Sometimes two-stage sampling is the only practical sampling method. On a live oak tree 5 meters in height I once counted 4,000 leaves on just one branch. Obviously enumerating all the leaves on the tree would be very tedious. In general, it is easy to sample the primary sampling units but difficult to sample the sub-unit. The observation on each sub-unit is considered to be the sum of two independent terms. One term, associated with the primary unit, has the same value for all second-stage units in the primary unit, and varies from one primary unit to the next with variance s_1^2 . The second term, which serves to measure differences between second stage units varies independently from one sub-unit to the next with variance s_2^2 . If a sample consists of n_1 primary units from each of which n_2 sub-units are drawn, then the sample as a whole contains n_1 independent values of the first term and $n_1 n_2$ values of the second term. The variance of the sample mean, \bar{x} , per sub-unit is:

$$s_{\bar{x}}^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2}.$$

These two components of variation can be estimated from an analysis of variance.

$$s_1^2 = s_1^2 = \left(\frac{\text{MS Primary Units} - \text{MS Sub - units}}{n_1} \right),$$

$$s_2^2 = s_2^2 = (\text{Ms Sub - units}),$$

$$s_{\bar{x}}^2 = s_{\bar{x}}^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2}.$$

Therefore, we can juggle the number of primary and secondary units to minimize $s_{\bar{x}}^2$.

But what choice of values is best? Naturally the answer to this question depends on the relative costs of primary and secondary sampling units. If the costs associated solely with sampling primary units is C_1 and the cost associated with sampling secondary units is C_2 then the total cost (C_T) of a 2-stage program is

$$C_T = C_1 n_1 + C_2 n_1 n_2$$

If advance estimates of these individual costs and of the variation due to each sampling stage are known then one can allocate sampling effort to minimize the standard error of

the statistic of interest for fixed cost, or to achieve a desired precision of our estimate by minimizing the product

$$VC_T = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2} \right) (C_1 n_1 + C_2 n_1 n_2)$$

where V is the variance of the sample mean in this case and C_T is total cost.

Since n_1 drops out of this expression we can solve for the value of n_2 that minimizes this expression:

$$\begin{aligned} \frac{\partial(C_T, V)}{\partial n_2} &= \frac{-s_2^2}{n_2} (C_1 + C_2 n_2) + \left(s_1^2 + \frac{s_2^2}{n_2} \right) C_2 = 0 \\ &= -s_2^2 C_1 - s_2^2 C_2 n_2 + s_1^2 n_2^2 C_2 + s_2^2 C_2 n_2 = 0 \\ s_1^2 C_2 n_2^2 &= s_2^2 C_1 \\ n_2 &= \frac{\sqrt{s_2^2 C_1}}{s_1^2 C_2} \end{aligned}$$

Then for known total cost C_T ,

$$n_1 = C_T / \left\{ C_1 + \left[\frac{\sqrt{s_2^2 C_1 C_2}}{s_1^2} \right] \right\},$$

and for known total variance V ,

$$n_1 = \left\{ s_1^2 + \left[s_2^2 / (s_2^2 C_1 / s_1^2 C_2) \right] \right\} / V.$$

Therefore, the value of n_2 required for an optimal allocation of sampling effort can be obtained, and a similar value for n_1 can also be obtained contingent on being able to specify the total cost or variance desired.

IV. Experimental Design

I have tried to illustrate that the goal of a sampling program is both to produce unbiased estimates of population parameters and to learn something about the sampling distributions appropriate for the underlying population distributions. Also, the reason for choosing a particular sampling program is to improve the power or sensitivity of the statistical hypothesis test motivating the sampling.

In testing a statistical hypothesis one uses sampling distributions which are largely chosen for mathematical convenience (i.e., whose forms can be specified if certain preconditions are met by the sampling program). One proposes a model, imposes specific conditions upon the model, and derives the model's consequences in terms of sampling distributions which are valid given the properties assumed for these sampling distributions. To the extent that the model and conditions imposed upon it approximate the actual experiment, the model can be used as a guide in drawing inferences from the data.

To use models that allow the properties of the sampling distributions to be specified in advance, the experiment must be designed to meet the preconditions associated with the particular model. If an experiment does not meet the specifications of existing models, the experimenter may be able to develop a model tailored to the specific experiment. However, the resulting data must still be analyzed. If sampling distributions with known and manageable characteristics appropriate for an experiment can be derived, the specific model can lead to inferences with known precision. Without knowledge of the properties of the appropriate sampling distributions, inferences drawn from an experiment have unknown precision.

The analysis of experimental data is dependent upon the experimental design and the sampling distributions appropriate for the underlying population distributions. The design, in part, determines what the sampling program will be. For standard designs the sampling distribution necessary to test the hypotheses of interest have known and manageable properties (i.e., asymptotic normality), which lead to the widespread use of these designs. Alternative designs are often available for an experiment having specified objectives. Depending upon the specific situation, one design may be more efficient - that is have power in the associated tests and narrower confidence intervals - for a given amount of experimental effort. The goal in planning experiments is to find the design that is most efficient per unit effort relative to the primary objectives of the experiment.

Increasing sample size, improving measurement techniques, and introducing various kinds of controls all may decrease experimental error and therefore improve power. Which method results in the greater increase in power for a given unit of effort will depend upon conditions unique to each experimental situation.

An examination of purely statistical aspects of experimental designs will help the experimenter find the model best suited for their experiment. The model chosen should allow the experimenter to reach decisions regarding all the objectives of the experiment. Whether or not a particular model actually corresponds to a specific experimental situation requires an in-depth knowledge of the subject matter addressed by the experiment. A careful assessment of the adequacy of alternative models may lead the experimenter to more fully understand the sources of variation inherent in the experiment. This may ultimately lead to a better design and therefore a more clear-cut interpretation of the experimental result.

Five criteria for evaluating experimental designs can be stated.

1. The model chosen and its underlying assumptions should be appropriate for the experimental material.
2. The design should provide as much information as possible with regard to the major objectives of the experiment for a given amount of experimental effort
3. The design should provide some information with regard to all the experimental objectives.
4. The design must be feasible within the working conditions that exist for the experimenter.
5. The analyses based upon the design should provide unambiguous information on the primary objectives of the experiment.

In the following discussion several broad categories of experimental designs will be presented. The benefits and costs of choosing one particular category of design over another will be examined.

A. Factorial Designs

Factorial experimental designs involve the comparison of the effects of two or more factors acting simultaneously on a common response or criterion variable. A factor can be considered a set of related treatments or related classifications. Each member of the set of related treatments belonging to Factor A is considered a level of Factor A. The principal advantage of using a factorial design versus a series of single factor experiments is that it allows one to examine the effects of the interaction of each factor combination on the criterion variable. The presence of an interaction effect attributable to the combination of factors above and beyond the effects of the factors singly can be determined. However, the additional effort necessary to test an hypothesis of interaction can be considerable. For example, if five replications are made at each level to test for the effects of 2 - four level factors singly then such a design requires a total of 40 replications. To test for the effects of 2 - four level factors and their interaction requires 80 replications. If prior information indicates that no interaction exists, a factorial design will not be as economical as several single-factor designs.

Figures 1-4 illustrate the data layout and analysis of variance for a single factor and a two-factor factorial experiment. In each instance an equal number of independent and randomly sampled elements are sampled at each factor level or combination of factor levels. Fully factorial designs, those with no confounding between factors and independent observations at all factor levels, are the most common and widely used designs. Other kinds of factorial experiments are sometimes useful.

Figure 1. Single Factor Factorial Experiment - data layout

Treatment 1	Treatment 2	...	Treatment k
X_{11}	X_{12}	...	X_{1k}
X_{21}	X_{22}	...	X_{2k}
X_{31}	X_{32}	...	X_{3k}
...
X_{n1}	X_{n2}	...	X_{nk}

Figure 2. Single Factor Experiment - ANOVA Table

Source of Variation	SS	df	MS	F
Treatments	SS_{treat}	$k-1$	$SS_{treat}/(k-1)$	MS_{treat}/MS_{error}
Error	SS_{error}	$kn-k$	$SS_{error}/kn-k$	
Total	SS_{total}	$kn-1$	$SS_{total}/kn-1$	

Figure 3. Two Factor Factorial Experiment - data layout

		Factor A			
		Level 1	Level 2	...	Level p
Factor B	Level 1	X_{111} X_{112} X_{113} ... X_{11n}	X_{121} X_{122} X_{123} ... X_{12n}	...	X_{1p1} X_{1p2} X_{1p3} ... X_{1pn}
	Level 2	X_{211} X_{212} X_{213} ... X_{21n}
	Level 3

	Level r	X_{r11} X_{r12} X_{r13} ... X_{r1n}	X_{rp1} X_{rp2} X_{rp3} ... X_{rpn}

Figure 4. Two Factor Factorial Experiment - ANOVA Table

Source of Variation	SS	df	MS	F-Ratios		
				Model I	Model II	Model III (A fixed, B random)
Factor A	SS_A	$p-1$	$SS_A/(p-1)$	MS_A/MS_e	MS_A/MS_{AB}	MS_A/MS_{AB}
Factor B	SS_B	$r-1$	$SS_B/(r-1)$	MS_B/MS_e	MS_B/MS_{AB}	MS_B/MS_e
AB Interaction	SS_{AB}	$(p-1)*(r-1)$	$SS_{AB}/(p-1)*(r-1)$	MS_{AB}/MS_e	MS_{AB}/MS_e	MS_{AB}/MS_e
Within cell (error)	SS_e	$pr*(n-1)$	$SS_e/pr*(n-1)$			
Total	SS_{total}					

Occasionally in executing a single-factor experiment a limited number or amount of primary sampling units are available to receive the experimental treatments. Those that are available may not be considered strict "replicates" because uncontrolled variation exists between primary units prior to the experiment. In order to incorporate enough replicates for each experimental treatment it is often necessary or maybe even desirable to use more than one primary unit. The best design in this situation is a randomized complete block design. Each primary unit is considered a block and each treatment is randomly assigned to sub-blocks within each block. A 2-factor analysis of variance is performed with blocks as one factor, in order to remove variation due to blocks from the experimental error. If hypothesis tests are only performed on the treatment effect then the blocking factor can be considered a fixed factor. If analyzed in this manner the treatment block interaction is implicitly considered to be zero.

B. Nested Designs

Three kinds of nested designs are used in agricultural and psychological research and have many applications in biology. These designs are *hierarchical*, *split-plot*, and *repeated measures*. The primary purpose of these designs is to eliminate uncontrolled variation due to *a priori* differences in primary sampling units from the estimate of experimental error. In this sense we can see that these designs are a way to remove confounding variation by adding classificatory controls or strata. Another reason for the use of these kinds of nested designs in biological research is that we often wish to make inferences concerning hierarchically arranged environments, habitats, and species.

1. Hierarchical Factors

Consider the example depicted in Figure 5, ignoring for the moment the high and low marsh categories. We wish to test the hypothesis that some characteristic, say above ground biomass, does not differ between estuaries. We have estimates of biomass per marsh in each estuary. Our Factor B, marshes, is not completely crossed with Factor A,

estuaries, since no marsh is found in both estuaries. The marsh factor is nested within each level of the factor estuaries. Since all levels of the factor marsh do not occur in combination with all levels of the factor estuaries, we cannot examine the effects of a marsh by estuary interaction. The degrees of freedom and SS for estuaries can be computed as in a normal 2-Way ANOVA. The SS for marshes is computed as the sum of the SS marshes within level 1 of factor A and the SS marshes within level 2 of Factor A. The degrees of freedom for each of these components is $(q - 1)$ where q is the number of marshes in each estuary. If marshes are considered a random factor the F ratio to test the hypothesis that $\sigma_a^2 = 0$ is $F = MS_A / MS_B$. If marshes is a fixed factor the F is MS_A / MS_{WC} . Designs with more levels of nesting are possible.

If we include data on biomass at locations high and low in each marsh the resulting design is a partially hierarchical design. The high-low factor is not nested within marshes or estuaries, but rather completely crossed with them. If we consider the estuary and high-low factors fixed and marshes random then these hypotheses may be tested: an estuary effect, a high-low effect, and a high-low/estuary interaction. An outline of the degrees of freedom, mean squares, and F ratios are given in Figure 6. Note that the within cell variation has been partitioned into two orthogonal components which are used as error terms to evaluate different hypotheses.

Figure 5. Nested Analysis of Variance - Data Layout

		Estuary 1			Estuary 2			Estuary 3		
		Marsh 1	Marsh 2	Marsh 3	Marsh 4	Marsh 5	Marsh 6	Marsh 7	Marsh 8	Marsh 9
	High	n	n	n	n	n	n	n	n	n
	Low	n	n	n	n	n	n	n	n	n

Figure 6. Nested Analysis of Variance - ANOVA Table Data Layout

Source of Variation	SS	df	MS	F
Estuaries	$SS_{\text{Estuaries}}$	$(p-1)$	$\frac{SS_{\text{Estuaries}}}{(p-1)}$	$\frac{MS_{\text{Estuaries}}}{MS_{\text{Marshes w Estuaries}}}$
Marshes within Estuaries	$SS_{\text{Marshes w Estuaries}}$	$p^*(q-1)$	$\frac{SS_{\text{Marshes w Estuaries}}}{p^*(q-1)}$	
High-Low	$SS_{\text{High-Low}}$	$r-1$	$\frac{SS_{\text{High-Low}}}{(r-1)}$	$\frac{MS_{\text{High-Low}}}{MS_{\text{Marshes w (Estuary by High-Low)}}}$
Estuary by High-low interaction	$SS_{\text{Estuary by High-Low}}$	$(p-1)*(r-1)$	$\frac{SS_{\text{Estuary by High-Low}}}{(p-1)*(r-1)}$	$\frac{MS_{\text{Estuary by High-Low}}}{MS_{\text{Marshes w (Estuary by High-Low)}}}$
Marshes within (Estuaries by High - Low)	$SS_{\text{Marshes w (Estuary by High-Low)}}$	$p^*(q-1)*(r-1)$	$\frac{SS_{\text{Marshes w (Estuary by High-Low)}}}{p^*(q-1)*(r-1)}$	

2. Split Plot Designs

Split-plot designs are equivalent to repeated measures designs and are used widely in agriculture. They are useful when one of the treatments is more difficult to apply than the other, or at least that one-treatment is easier to apply at a larger scale. Figure 7 depicts the layout of a split-plot design. This design is similar in form to the randomized complete block design except that in this instance a treatment level is applied to each whole block or plot. Within each whole plot, each level of a second treatment is randomly assigned to sub-plots. The effects of factor A are confounded with differences between whole plots while the effects of factor B are part of the within plot variation. The estimates of the effect of Factor B are free from variation due to whole plots. The interaction between Factor A and B is also free from whole-plot effects. The analysis of this design is outlined in Figure 8.

3. Repeated Measures Designs

In repeated measures experiments observations are made on the same subject at all levels of at least one factor. For example, the paired t - test can be considered the simplest instance of a repeated measure design. Each subject is observed before and after the application of some treatment. The advantage of such a design is that the subject acts as a self-control. Variation between subjects that occurs for reasons unrelated to the experiment can then be removed from one estimate of experimental error. This may lead to a more sensitive test of the hypothesis of interest. For example in a t - test with un-correlated observations (no repeated measures) the estimate of

experimental error is

$$s_{\bar{x}_a - \bar{x}_b}^2 = \frac{s_a^2}{n_a} + \frac{s_b^2}{n_b},$$

while for a *t* - test on paired observations it is

$$s_d^2 = s_{\bar{x}_a - \bar{x}_b}^2 = \frac{s_a^2}{n} + \frac{s_b^2}{n} - \frac{2r_{ab}s_a s_b}{n}$$

Where $r_{ab}s_a s_b$ is the covariance of *a* and *b*. If the covariance is positive then the estimated experimental error from a design involving correlated observations will be smaller than that from un-correlated observations by a factor of $2r_{ab}s_a s_b$. However, the degrees of freedom for the estimate with correlated observations are only (*n* - 1), while they are (*n_a* - 1) + (*n_b* - 1) for un-correlated observations.

Figure 7. Split -Plot or Repeated Measures Analysis of Variance - Data layout

		Factor A			
		A1	A2	A3	A4
Factor B	Plot 1-n	Plot 1-n	Plot 1-n	Plot 1-n	Plot 1-n
	B1	B2	B3	B2	
	B2	B1	B1	B3	
	B3	B3	B2	B1	

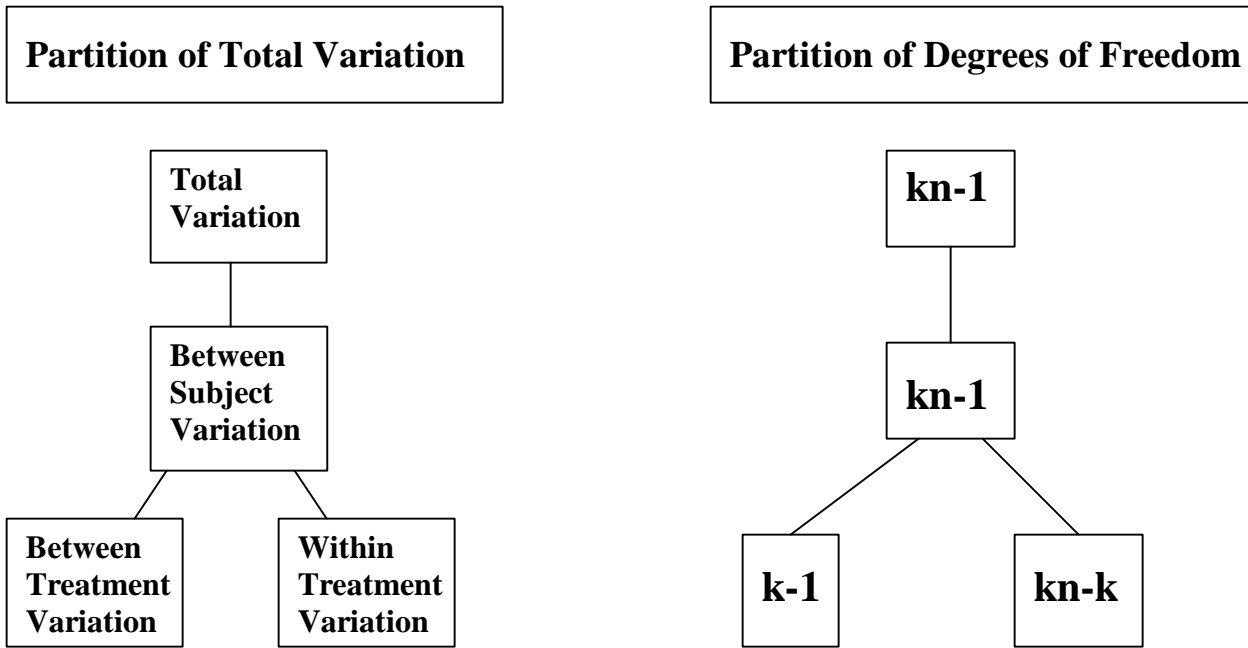
Figure 8. Split -Plot or Repeated Measures Analysis of Variance - ANOVA Table

Source of Variation	SS	df	MS	F
A	SS _A	(<i>p</i> -1)	SS _A / <i>p</i> -1)	MS _A /MS _{plots w A}
Plots w A	SS _{plots w A}	<i>p</i> *(<i>n</i> -1)	SS _{plots w A} / <i>p</i> *(<i>n</i> -1)	
B	SS _B	(<i>q</i> -1)	SS _B / <i>q</i> -1)	MS _B /MS _{B x plots w A}
AB	SS _{AB}	(<i>p</i> -1)*(<i>q</i> -1)	SS _{AB} / <i>p</i> -1)*(<i>q</i> -1)	MS _{AB} / MS _{B x plots w A}
B by Plots w A	SS _{B x Plots w A}	<i>p</i> *(<i>q</i> -1)*(<i>n</i> -1)	SS _{B x plots w A} / <i>p</i> *(<i>q</i> -1)*(<i>n</i> -1)	

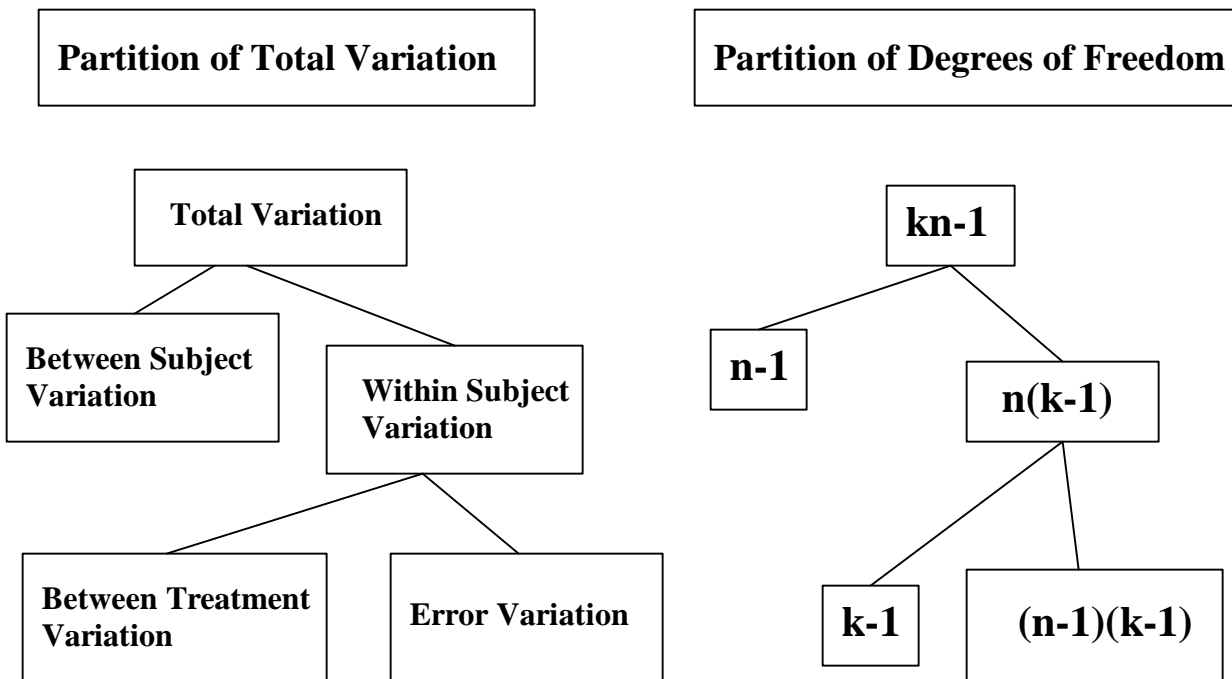
For a repeated measures experiment to be more efficient than a design with un-correlated observations the reduction in experimental error associated with controlling for extraneous between subject variation must offset the reduction in degrees of freedom. In field biological research where our subjects might be trees, lakes, marshes, streams, or grids, even given an effort to chose them to be as similar in physical characteristics as possible, a repeated measures design may help to remove

uninteresting between subject variation. One, two, three, and more complicated multi-factor repeated measures designs are possible with from one to all factors with repeated measures. The main effects and interaction sum of squares are computed as in a factorial experiment. The error variation is partitioned into a series of orthogonal terms that are used to evaluate the main effects and interactions. Figure 9 shows a comparison of the partition of variation in a repeated measures and a non-repeated measures single-factor experiment.

Figure 9. Single Factor Experiment Uncorrelated Observations



Single Factor Experiment (Repeated Measures)



C. Square Designs

Square designs like Latin and Greco-Latin squares are useful in controlling for individual differences between experimental units. They are also useful in instances where the main effects of three factors are of interest, but the number of subjects available is low or the cost of making observations is high. The loss of information in square designs involves the interaction terms. Square designs can involve repeated measures and in that context are used for controlling sequence effects associated with the order of applying treatments to subjects.

For Further Reference

- Chamberlin, T.C. 1965. The method of multiple working hypotheses. *Science* 148: 754-759.
- Cochran, W.G. and G.M. Cox. 1957. *Experimental Designs*. John Wiley and Sons.
- Hurlburt, S.H. 1984. Pseudoreplication in the design of ecological field experiments. *Ecological Monographs* 54: 187-211.
- Keppel, G. 1982. *Design and Analysis: A Researcher's Handbook*. Prentice-Hall: Englewood Cliffs, NJ.
- McCall, R.B, and M.L, Applebaum. 1973. Bias in the analysis of repeated-measures designs: Some alternative approaches. *Child Development* 44:401-415.
- Nowell, A.R.M.et al. 1982. High energy benthic boundary layer experiment: Hubble. *EOS*, August 1982. P. 594-595.
- Snedicor, G.W. and and W.G. Cochran. 1967. *Statistical Methods*. Iowa State University Press.
- Selvin, H.C. and A. Stuart. 1966. Data dredging procedures in survey analysis. *American Statistician* 20: 20-23.
- Winer, B.J. 1971. *Statistical Principles in Experimental Design*. McGraw-Hall: New York.