

Topic 1: INTRODUCTION TO PRINCIPLES OF EXPERIMENTAL DESIGN

[S&T Ch 6] plus review [S&T Ch 1-3]

1. 1. Purpose

"The purpose of statistical science is to provide an objective basis for the analysis of problems in which the data depart from the laws of exact causality.

D. J. Finney,
An Introduction to Statistical Science in Agriculture

1. 2. Types of reasoning

Two types of inferences are commonly used to derive a scientific conclusion, namely, deductive inference and inductive inference

Deductive reasoning

Deductive reasoning is reasoning from the general to the specific. It is what we colloquially think of as "logic". A **deductive inference** is a judgment or generalization based on axioms or assumptions. For example, if two coins are assumed perfectly balanced then one can conclude that the mean number of heads of tossing these two coins must be one. The deductive generalization is correct only if the assumptions are correct.

Inductive reasoning

Inductive reasoning, as the term is used by statisticians, means reasoning from the particular to the general. In the example of two coins, a conclusion about the mean number of heads will be based on the actual results of a number of trials. Experiments are conducted to provide specific facts from which general conclusions are established and thus involve inductive reasoning. Here is a slightly humorous inductive "argument" that all odd numbers are prime numbers: 1 is a prime, 3 is a prime, 5 is a prime, 7 is a prime, (there are some problems with 9 that seem to be due to temperature effects), 11 is a prime, and 13 is a prime, therefore all odd numbers are prime. From this example, it is clear that inductive reasoning does not always produce valid conclusions. It is, however, the source of almost all advances in science, since its intelligent use allows one to use observations to motivate the formulation of new scientific theories.

1. 3. The scientific method

The power of inductive reasoning is that it permits the scientist to formulate general theories about the world based on particular observations of the behavior of the world. The problem with inductive reasoning is that these theories are often wrong. The "scientific method" is a formal statement of procedure designed to facilitate the scientist's making the most effective use of his or her observations. The scientific method is usually defined to consist of the following four steps (Little and Hills, 1978):

Formulation of the hypothesis. Based on preliminary observations, this is the tentative explanation.

Planning the experiment. The experiment must be constructed to objectively test the hypothesis. This is what this course is all about.

Careful observation and collection of the data

Interpretation of the results. The results of the experiment may lead to confirmation, alteration, or rejection of the hypothesis.

1. 3. 1. Some important characteristics of a well-planned experiment are (Cox 1958):

1. *Degree of precision.* The probability should be high that the experiment will be able to measure differences with the degree of precision the experimenter desires. This implies an appropriate design and sufficient replication.
2. *Simplicity.* The design should be as simple as possible, consistent with the objectives of the experiment.
3. *Absence of systematic error.* Experimental units receiving one treatment should not differ in any systematic way from those receiving another treatment so that an unbiased estimate of each treatment effect can be obtained.
4. *Range of validity of conclusions.* Conclusions should have as wide a range of validity as possible. An experiment replicated in time and space would increase the range of validity of the conclusions that could be drawn from it. A factorial set of treatments is another way of increasing the range of validity of an experiment.
5. *Calculation of degree of uncertainty.* The experiment should be designed so that it is possible to calculate the possibility of obtaining the observed result by chance alone.

1. 3. 2. Steps in experimentation (Little and Hills 1978):

Define the problem

Determine the objectives

Select the treatments

Select the experimental material

Select the experimental design

Select the experimental unit and number of replications

Ensure proper randomization and layout

Ensure proper means of data collection

Outline the statistical analysis before doing the experiment

Conduct the experiment

Analyze the data and interpret the results

Prepare complete and readable reports

1. 4. Experimental design

1. 4. 1. The role of experimental design

Experimental design concerns the validity and efficiency of the experiment. The experimental design in the following diagram (Box et al., 1978), is represented by a movable window through which certain aspects of the true state of nature, more or less distorted by noise, may be observed. The position and size of the window depend on the questions being asked by and the quality of the experiment. A poorly used design may not generate any useful information for meaningful analysis. A wisely designed experiment can provide factual evidence which can easily be analyzed and understood by the researcher.

Obviously methods of experimental design are at least as important as methods of data analysis in a research program.

The diagram emphasizes that, although the conjectured state of nature may be false or at least inexact, the data themselves are generated by the true state of nature. This is the reason why the process of continually updating the hypothesis and comparing the deduced states of nature with actual data can lead to convergence on the right answers.

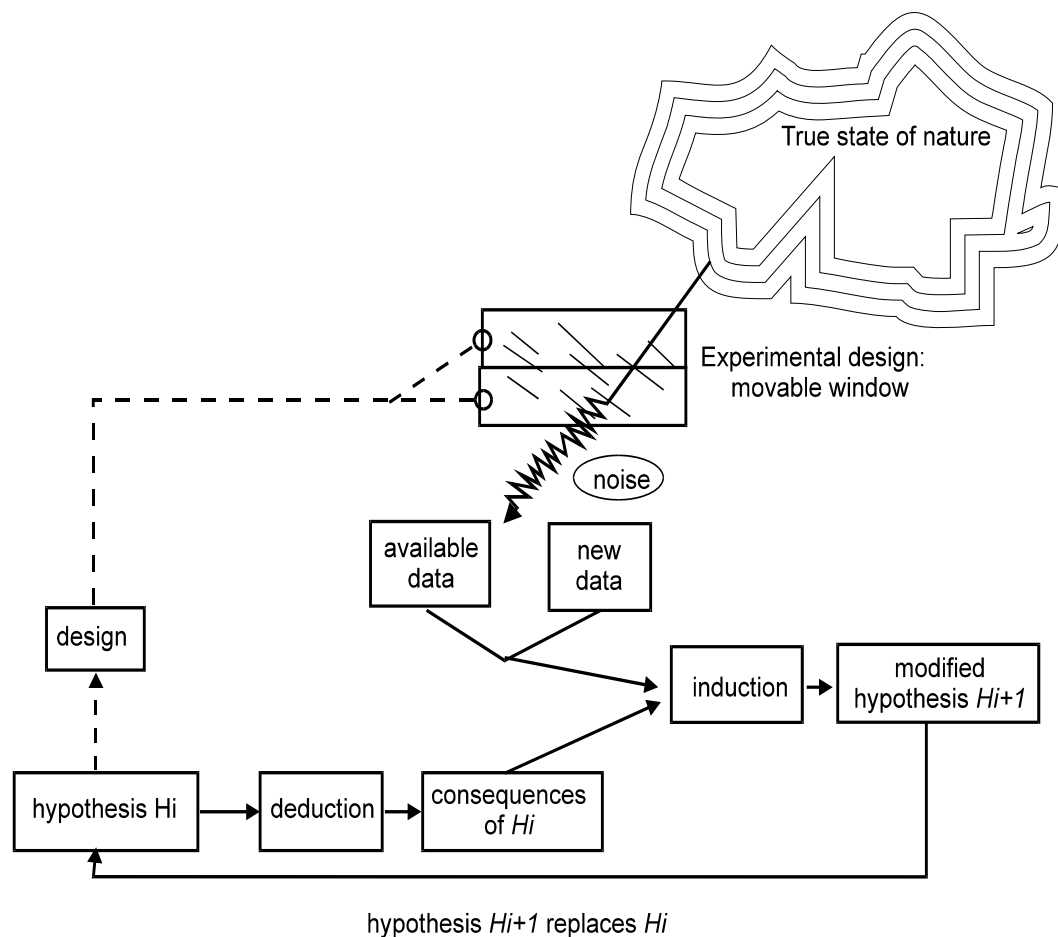


Fig. 1. Role of experimental design

A complete scientific research process therefore includes two concurrent and parallel approaches: theoretical and experimental. Statistics offers a powerful tool to help researchers to conduct experiments and analyze data. These two approaches must be integrated to gain knowledge of the phenomenon.

A simple example of a research program may be shown as follows:

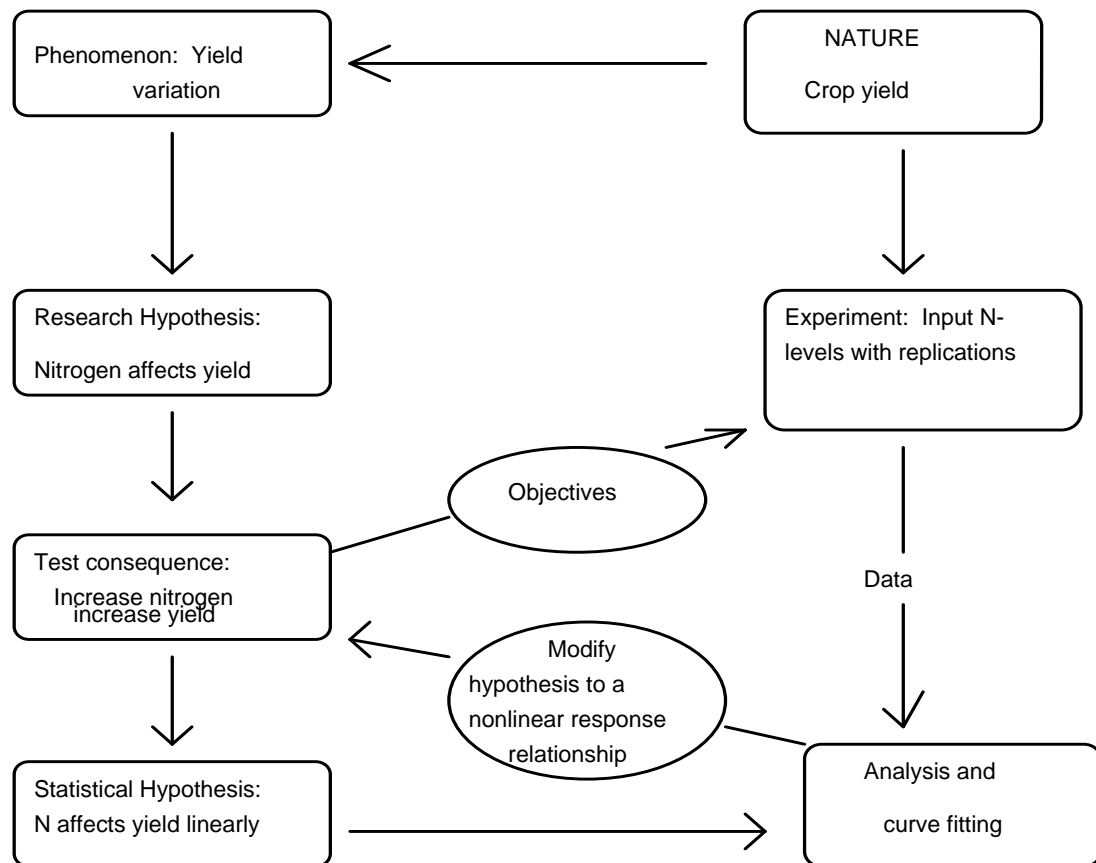


Figure 2. Example of the process of research

A designed experiment must satisfy all requirements of the objectives of a study but is also subject to the limitations of available resources. Below we will give examples of how the objective and hypothesis of a study influences the design of an experiment.

1. 4. 2. Objectives and experimental design

First, the type of an experiment depends on the objective. This point may be illustrated by the following examples:

a) In variety trials the objectives of a **screening trial** and a critical evaluation trial can be very different. In a screening trial (exploratory type of study), one

would like to compare as many entries as possible even at some expense of precision, that is, at a reduced number of replications. On the other hand, in a rigorous variety trial the entries can be few, but high precision must be maintained so that the differences among the varieties can be *compared* with a high level of confidence.

b) Environmental studies can be classified into *baseline, monitoring, or impact studies* according to their objectives. Baseline studies are used to establish the “normal environment.” Monitoring studies are used for detecting changes, and impact studies are designed to relate environmental changes to certain specific impact factors.

For example, in evaluating the environmental impact of the application of an herbicide, periodical soil samples may have to be taken from farms with herbicide application and farms without the herbicide application, and from farms before and after the application. Thus, in an impact study, two controls, spatial and temporal, are needed to have an optimal impact design. An area-by-time factorial design provides evidence for impact effects in a significant area-by-time interaction.

Second, the *scope* of the experiment is also defined by the objectives. For instance, a comparison of several varieties in several years but only in one location, does not permit the researcher to make inferences about their behavior in other soil environments. Thus, how far and to what populations the inferences can be extended depends on the scope of the study, which is determined in turn by the objectives of the study.

1. 4. 3. Hypotheses and experimental design

1. 4. 3. 1. Concepts about hypotheses

Curiosity leads to investigational questions that can be posed in the form of hypotheses. A hypothesis is the simplest possible answer to a question, stated in a way that is testable and falsifiable.

Hypothesis must be falsifiable

Hypothesis formulation is a prerequisite to the application of statistical design and analysis. A null hypothesis (H_0) can never be proved correct, but it can be rejected with known risks of being wrong, i.e. falsifiable. (Thus, a chemical can never be statistically proved as perfectly safe.)

Example: If H_0 says that late maturing cultivars have the same yields as early maturing cultivars, then you must reject H_0 if an increased yield of late cultivars were observed in the experiment. You can't then say “come to think of it, these differences may be due to different environments where they were grown, not really the cultivar differences.” If a result can be interpreted in that way, then the hypothesis was unfalsifiable. This can result only from a poor experimental design.

A hypothesis can be expressed as a model. All models are abstractions and simplifications or approximations of reality but not the reality itself. In biological or environmental studies, a model that closely approximates a complex real system would need hundreds of simultaneous partial differential equations with time lags and hundreds of parameters. Simplification is both legitimate and necessary. Thus, *all models may be considered as hypotheses* that can be further improved or refined by experimental work.

Initially, a proposed hypothesis can be considered as an *assumption* which can only be tested by comparing the predicted results of the assumption with the experimental data. An initial hypothesis is called a *research hypothesis* (RH), and the predicted result is called the *test consequence* (TC). Many alternative TCs can be deducted from an RH, and each TC can be tested by a specifically designed experiment. Therefore, it is important to select a TC so that the RH can be most efficiently and correctly tested by experiment. The method of deriving TC from RH is called *hypothetical-deductive* (HD) method. A poorly derived TC leads to inefficient experiments, or possibly nonfalsifiable RH. In general, a TC is stated in a form of a statistical hypothesis so that statistically sound methods can be applied to test the TC. Since a TC dictates the objective, type, and scope of an experiment and potential errors of conclusions of analysis, it is extremely important for experimenters to formulate well thought out and the best possible TC from an RH.

1. 4. 3. 2. Significance level of testing a hypothesis

Significance level is the probability that one rejects H_0 when it is in fact true. Of course, this is an incorrect decision that should be kept as small as possible. Conventionally, this is kept less than 0.05 (alpha).

If H_0 “no yields reduction is caused by increased ozone” is tested at $\alpha = 0.05$ or 5%, then a significant result means that on the evidence of the experiment, there is less than 1-in-20 chance that the observed yield reduction would have occurred without the damage by ozone and that one out of 20 or more times is an acceptable low risk of being wrong in the conclusion that ozone damage affects crop yield. If one finds a 3.5 pounds significant decrease in yield this can't be interpreted as the real decrease has a 5% probability of being 3.5 pounds. The probability of the difference to be statistically detected, depends on the true magnitude of the difference.

However, lowering alpha will increase Type II error, beta, which is the probability of concluding that H_0 is true when in fact it is not. There is always a tradeoff, and the only way to reduce one error without increasing the other is to improve the design. For example, one could increase the number of samples to reduce either error, or both.

Sometimes a significant biological difference between treatments is not statistically significant. This problem is due to inadequate design. On the other hand, statistical significance may not be biologically significant. Sometimes too many measurements are made without real purpose, and the chance of a non-real biological difference being declared significant statistically is greatly increased. With a 5% alpha

and 200 comparisons one can obtain about 10 false significances. Thus, it is important to define “biological significance;” then experiments can be designed to detect this amount, no more and no less.

1. 4. 4. Specific issues of experimental design

Once the objectives, interesting questions, and the hypothesis are defined, the scope, type, and requirements of an experiment are also more or less determined. Thus, *the experiment should be designed to meet those requirements*. Specifically, experimental design is concerned with the following issues:

1. 4. 4. 1. *The size of the study*: number of replications, and the size and shape of experimental units.
1. 4. 4. 2. *Type and number of measurements*: availability of a measuring device, precision and accuracy of the measurement, and the timing of making measurements.
1. 4. 4. 3. *Treatments*: the type of treatments, the levels of treatment, and the number of treatments.
1. 4. 4. 4. *Assignment of treatments to experimental units*: completely random, restricted randomization, etc.
1. 4. 4. 5. *Error control*: the error control can be accomplished by blocking techniques, the use of concomitant observations, the choice of size and shape of the experimental units, and the control of the environment using a growth chamber or greenhouse.
1. 4. 4. 6. *Relative precision of designs involving few treatments*: to compare two experimental designs one compares amounts of information.

Among the above considerations, replication and randomization are the most important basic principles in designing experiments.

1. 4. 4. 1. Replication

1. 4. 4. 1. 1. Functions of replication. [ST&D pg. 130]

- To provide an estimate of the experimental error. The *experimental error* is the variation which exists among observations on experimental units treated alike. When there is no method of estimating experimental error, there is no way to determine whether observed differences indicate real differences or are due to inherent variation.
- To improve the precision of an experiment by reducing the standard deviation of a treatment mean. Increased replication usually improves precision, decreasing the lengths of confidence intervals and increasing the power of statistical tests.

- To increase the scope of inference of the experiment by selection and appropriate use of more variable experimental units. Example: replication in time and space in yield trials.
- To effect control of the error variance. The aim is to assign the total variation among experimental units so that it is maximized among groups and, simultaneously, minimized within. Experimental error must not be inflated by differences among groups.

1. 4. 4. 1. 2. Distinguishing between replications, subsamples, and repetitions

Replication refers to the number of experimental units that are treated alike (experimental unit or experimental plot is the unit of material to which one application of a treatment is applied). Misconception of number of replications has often occurred in experiments where subsamples or repeated observations on some unit were mistaken as experimental units. Hurlbert (1984) reported that pseudoreplication occurred in 27% of the studies published in ecology journals since 1960. A 1982 survey of articles in plant pathology journals showed no true replications in 19% of papers on phytopathology and 37% of the papers on plant disease. Without proper replication, no valid scientific conclusion can be drawn from such a study.

There are two important points to remember in determining what constitutes a replication.

(1) Each replication must be independent of every other.

(2) Each replication must be part of a randomized trial; that is, any one plot must have the same chance of getting each treatment.

Example: Field trial. We want to compare 3 fertilizer treatments, denoted A, B, and C. Except as discussed in Case 4, the field is an ordinary one, and assumed uniform. The crop is an annual.

1	2	3
4	5	6
7	8	9
10	11	12




Figure 3. Replications, subsamples and repetitions.

Case 1

4 plots are selected at random for each fertilizer. This is done by moving through a table of random numbers. The first 4 between 1 and 12 selected are assigned to treatment 1, the next 4 to treatment 2, and the last 4 to treatment 3. This is a *completely randomized design*. There are 4 replications for each treatment. If the experiment is run in succeeding years then each year provides a different set of replications.

Case 2

Same as case 1 except the crop is a perennial. In this case the succeeding years are not replications. The fact that the crop is a perennial violates the independence requirement above. This is a repeated measures experiment.

Case 3

Same as case 1 except each of the 12 plots are further divided into three subplots. These subplots are *not* replications, they are subsamples. The randomization requirement above is violated.

Case 4

The numbers 1, 2, and 3 are arranged in a random order. Each treatment is applied to all plots in a north-south column containing the randomly selected number (e.g. treatment 1 is applied to plots 2, 5, 8, and 11, etc.). In this case there is no replication for the treatment. The separate plots are subsamples.

The definition of replication can have an even broader meaning. For instance, in a breeding program of yield trials, varieties are frequently compared in a number of locations and years. In this case, the varieties are treatments and the experimental units are plots in location and year. Over the years, these plots may not be the same spots in locations, or not even in the same locations. The point here is that sometimes the locations and years may also be considered as replications so as to enable us to examine the consistency of the varietal yield performance. However, it should be noted that sometimes the locations may be of special interest and these are considered as treatments rather than replications.

1. 4. 4. 2. Type and number of measurements**1. 4. 4. 2. 1. Type: practical and theoretical considerations**

Practical considerations are the availability of the measuring device and the costs. Sometimes *indirect measurement* is much cheaper than direct measurement. For example, it may be very expensive to measure herbicide residues in the soil by taking soil samples and analyzing them chemically. Instead we may select an indicator plant and use a bioassay to measure the amount of residue indirectly from the plant response. At times it is not practical to make an observation on the complete experimental unit and the unit is sampled.

From the theoretical point of view it is necessary to establish the reliability of the measurements, i.e. their precision and accuracy.

Precision has to do with the concept of random errors and the precision of an average can always be improved by increasing the sample size. Precision, sensitivity, or amount of information is measured as the reciprocal of the variance of a mean. If we let I represent the amount of information contained in a sample mean, then

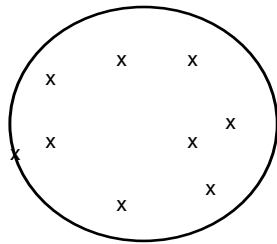
$$I = 1 / \sigma_{\bar{y}}^2 = n / \sigma^2.$$

If $s_{\bar{y}}^2$ is used to estimate $\sigma_{\bar{y}}^2$ there is a correction to the I formula

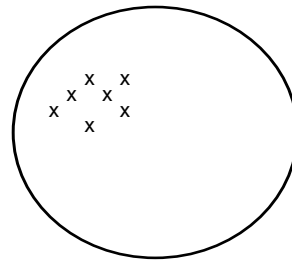
$$I = \frac{(n_1 + 1)}{(n_1 + 3)} \frac{1}{s_{\bar{y}}^2}$$

Note that when $n \rightarrow \infty$, then the **correction factor** $(n_1 + 1) / (n_1 + 3) \rightarrow 1$.

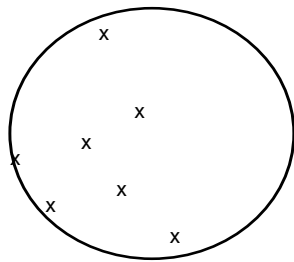
Accuracy is associated with the concept of bias or systematic errors. It is caused by the procedure of taking the measurement or the device itself. The next figure shows the differences between accuracy and precision.



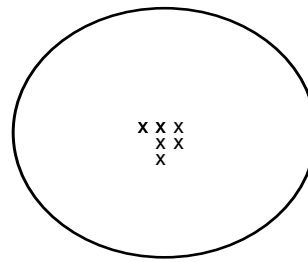
Accurate but not precise



Not Accurate but precise



Not accurate not precise



Accurate and precise

Figure 4. Accuracy and precision

1. 4. 4. 2. 2. Number of measurements

Sometimes, researchers collected too many variables and unnecessarily complicated the experiment. Example: in a clinical study, several systems, such as lymphoreticula system, pulmonary system, gastrointestinal system and cardiovascular system are examined in addition to chemical and enzymatic analyses of blood and urine samples. As a result, it is very difficult to perform a meaningful data analysis.

If subsampling is used, there is little point in making a large number of determinations in each experimental unit since experimental error must be based on variation among experimental units, not on variation among samples from within the experimental units.

1. 4. 4. 3. Type and level of treatments [ST&D p. 136] [Cox 1958]

For most of the experimental designs it is assumed that the observation obtained when a particular treatment is applied to a particular experimental unit is:

(a quantity depending only on the unit) + (a quantity depending on the treatment)

The essential points about this assumption are that

- a) the treatment term adds on to the unit term rather than for example, multiplying
- b) the treatment effects are constant

c) the observation on one unit is unaffected by the treatment applied to other units (no carry-over)

A particular type of treatment is the control, included in many situations as a standard to check the effect of other treatments.

The actual levels and number of treatment levels are important considerations, particularly when the treatments are quantitative. Choice of the number of levels and their spacing is important to determine the nature of the response (for example linear or curvilinear response).

1. 4. 4. 4. Assignment of treatments to experimental units [ST&D p. 137]

Treatments are generally randomly assigned to experimental units. The function of randomization is to ensure that we have a valid estimate of experimental error and of treatments means and differences among them. Randomization avoids possible bias on the part of the experimenter, thereby increasing the accuracy of the estimation (See 1. 3. 1. 3.) and guarantees that, on the average, errors will be independently distributed. Randomization tends to destroy the correlation among errors and make valid the usual tests of significance. Putting samples in “representative or typical” places is not random sampling.

Under special circumstances, systematic assignments may be applicable but we are not going to study them in this course.

1. 4. 4. 5. Error control [S&T p. 133]

Control of experimental error through experimental design consists of designing an experiment so that some of the natural variation among the set of experimental units is handled so as to contribute nothing to differences among treatment means.

Sometimes covariates are used to adjust environmental differences. The precision can be increased by the use of accessory observations and the analysis of covariance.

In field plot experiments, the size and shape of the experimental unit and the block are important in relation to precision. Uniformity trials have shown that the individual plot should be relatively long and narrow for greatest precision. Blocks should be approximately square to maximize variation among blocks. In fields with definite gradients the most precision is obtained when the long sides of the plots are parallel to the direction of the gradient.

Depends on the purpose, a study will require different level of environmental control, i.e., growth chambers or greenhouses.

1. 4. 4. 6. Relative precision of designs involving few treatments [ST&D p. 132-133]

As we said before (1. 4. 4. 2. 1.), the precision or amount of information in an experiment is measured by

$$I = 1 / \sigma_{\bar{Y}}^2 = n / \sigma^2$$

The number of degrees of freedom affect the estimation of σ^2 by s^2 and the estimate of the amount of information. The degrees of freedom depend on the number of replicates, number of treatments, and experimental design.

The *relative efficiency* of design 1 relative to design 2 is calculated as the ratio of the amount of information in design 1/design 2 (after the manipulation of the equation the of s_1^2 design ends in the denominator):

$$RE_{1 \text{ to } 2} = \frac{I_1}{I_2} = \frac{(n_1+1)/[(n_1+3)s_{Y_1}^2]}{(n_2+1)/[(n_2+3)s_{Y_2}^2]} = \frac{(n_1+1)(n_2+3)s_{Y_2}^2}{(n_2+1)(n_1+3)s_{Y_1}^2}$$

If this ratio is >1 , design 1 provides more information and is more efficient.

If $n_1=n_2$ this equation simplifies to

$$RE_{1 \text{ to } 2} = \frac{(n_1+1)(n_2+3)s_2^2}{(n_2+1)(n_1+3)s_1^2}$$

When comparing two experimental designs s_1^2 and s_2^2 are the mean square errors (MSE) of the first and second designs, respectively (obtained from the ANOVA table) and n_1 and n_2 are their degrees of freedom.

Review of basic material

Terminology

Treatment (also level or treatment level) [ST&D p. 128]

A dosage or method to be tested in an experiment

Plot (also sampling unit experimental unit or individual) [ST&D p. 128]

An experimental unit to which a treatment is applied, The unit where measurements are made. Examples: If the experimental unit is a field and the measurement is total yield then the sample unit is the experimental unit. If the measurement is yield per plant then the sample unit is a single plant.

Replication [ST&D p. 130]

When a treatment appears more than once in an experiment, it is said to be replicated.

Subsample [ST&D p. 157]

A measurement not including the whole experimental unit. For example, if there are four plants to a pot, and the pot is the experimental unit, and each plant is recorded individually, then the individual plants are subsamples.

Repeated measurement

The measurement of the experimental unit more than once. For example, if there are four plants to a pot, and the pot is the experimental unit, and pooled data from all four plants are recorded on each day for a week, then each individual measurement is part of a set of repeated measurements.

Variable [ST&D p.8]

A measurable characteristic of a plot

Population [ST&D, p11]

The set of all possible values of a variable.

Observation (also variate) [ST&D p. 9]

An individual measurement of a variable

Sample [ST&D p. 11]

A set of measurements that constitutes part of a population

Random sample [ST&D p. 12]

A sample in which each member of the population has the same chance of being included. Note that a *random* sample is NOT the same as a *haphazard* sample.

Parameter [ST&D p. 17]

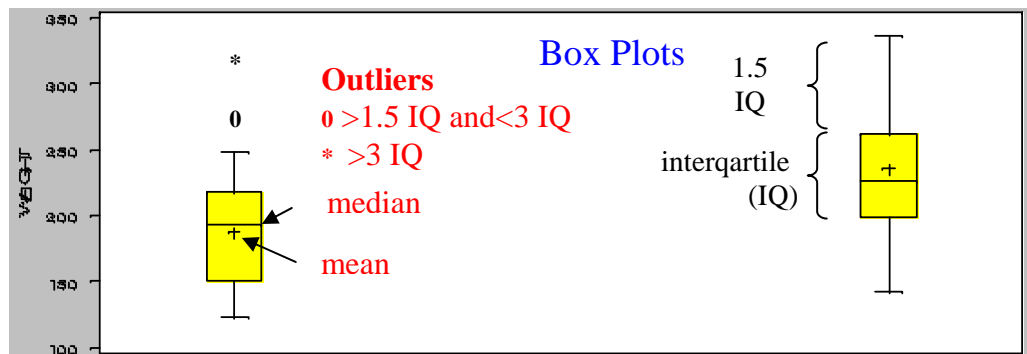
A fixed value that characterizes a population

Statistic [ST&D p. 17]

The value of a parameter as applied to a sample

Measures of Central Tendency and Dispersion [ST&D p. 16-27]

Graphic representations of the means of samples often indicate sample variation by a simple method. The method of **boxplots** has become quite common recently (ST&D p. 31). The sample is represented as a box whose top and bottom are drawn at the lower and upper quartiles (the length of the box is the interquartile range). The box is divided at the median. A vertical line ('whisker') is drawn from the top of the box to the largest observation within 1.5 interquartile ranges of the top. A comparable line is drawn from the bottom to the smallest observation within 1.5 interquartile ranges of the bottom. All observations beyond these limits (outlier) are plotted individually. The great utility of boxplots is that they furnish measures of location (median) dispersion (box and whiskers length), skewness (asymmetry of the upper and lower portions of the box, whiskers and outliers). Boxplots are also useful to visually compare two or more samples.



Degrees of freedom

The reason that the quantity '**n-1**' is used in the denominator of the sample variance rather than '**n**' is that the quantity $\sum (Y_i - \bar{Y}) / n$ is a *biased estimator* for σ^2 . To say that a statistic is an unbiased estimator of a parameter means that if sets of variates are repeatedly drawn from a population and used to compute the statistic, then the expected value of all these computed statistics is the value of the parameter. For example, the sample variance s^2 defined above is an unbiased estimator of the population variance σ^2 . That is, if a very large (infinite) set of samples of r variates are drawn from the population of N measurements and the sample variance s^2 is computed for each of these, and the average $E\{s^2\}$ of all of these sample variances is computed, then $E\{s^2\} = \sigma^2$. If you look at the mathematical proof of this, you will find that it hinges on the observation that the sum of the n individual deviations $(Y_i - \bar{Y})$, which are used to compute s^2 , must be zero. This is called a *linear constraint* on these quantities.

Any set of r quantities that do not need to satisfy any such constraints is said to have ***n* degrees of freedom**. Each constraint that a set of quantities must satisfy removes one degree of freedom. Thus, the set of r values Y_i has ***n* degrees of freedom** and the set of r deviations $(Y_i - \bar{Y})$ has ***n-1* degrees of freedom**. In ANOVA the number of degrees of freedom can be interpreted as the number of independent comparisons that can be made among means in an experiment. Any constraint that the means must satisfy removes one degree of freedom.

Probability distributions and random variables

Populations, samples, and observations come from the real world of experiments. Probability distributions and random variables come from the imaginary world of mathematics.

Binomial distribution [ST&D p. 39]

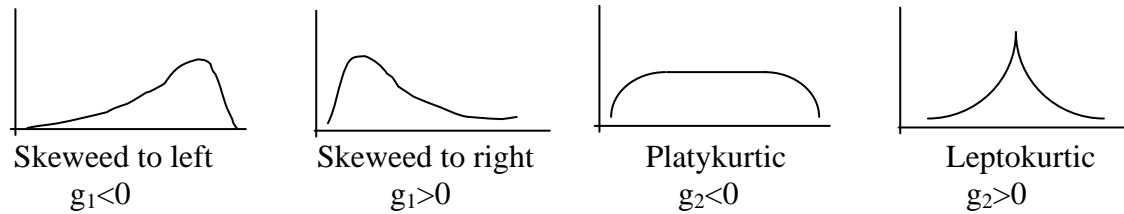
In the binomial trial there are only two possible outcomes. Trials are independent and the probability of occurrence of a certain event remain constant from trial to trial. The mean and variance of a binomial distribution are given by:

Mean $\mu = np$

Variance $\sigma^2 = np(1-p)$

The normal distribution [ST&D p. 45]

The normal curve is bell-shaped and symmetrical around the mean. Therefore the mean, median, and mode are all at the same point. Two normal departures from normality are skewness (measured by g_1) and kurtosis (measured by g_2).



The central limit theorem

Consider the coin tossing experiment. The real physical system consists of repeatedly tossing a coin and determining whether it is heads or tails. Suppose the coin has probability p (say, $p = 0.4$) of landing heads and a probability $q = 1 - p$ of landing tails. The probability model would be the selection of an infinite number of zeroes and ones, with each selection being a 1 with probability p . Out of the infinite population of values we select a sample of size r (say, 500). What is the probability that the number of heads will be within a certain range (say, 220 to 230)? The exact number can be computed by a complicated formula involving the binomial distribution.

It turns out that *because we are dealing with a sum* of heads the probability distribution of this sum is approximately normally distributed.

As sample size increases, the means of samples drawn from a population of any distribution will approach the normal distribution.

This theorem, when rigorously stated is known as the **central limit theorem**. The importance of this theorem is that if n is large enough, it permits us to use the normal distribution to make statistical inferences about means of populations even though the items are not normally distributed. The necessary size of n depends upon the distribution (skewed populations require larger samples).

Let Y_1, Y_2, Y_3, \dots be an infinite sequence of random variables drawn from the same distribution, which has mean μ and variance σ^2 . Then if a sample of size n is drawn from this sequence (or population), the expected value of the sum of the variates of this sample is obviously $n\mu$ and the variance can be shown to be $n\sigma^2$ (review ST&D p. 113-115). As the size n approaches infinity, the mean of the samples, becomes normally distributed with mean μ and variance σ^2/n . The last formula makes it clear that the standard deviation of means is a function of the standard deviation of the items, as well as of the sample size of the means

The standard error [ST&D p. 76]

One of the most important quantities in this course is the *standard error*. Returning to our coin tossing experiment, it is obvious that we will get a better estimate

of the mean μ if we toss the coin 500 times than if we toss it 10. The *standard error* makes this precise. Suppose we have a population with mean μ and variance σ^2 , and that a sample of n observations are drawn from this population. The mean \bar{Y} of this sample is itself a random variable. By the central limit theorem this mean is a member of a normally distributed population of means with mean μ and variance σ^2/n . When working with samples from a population we do not know the parametric standard deviation but can only obtain a sample estimate s .

Therefore, we usually have to estimate the standard deviation of means from a single sample by s / \sqrt{n} . This is the *standard error*, which we will denote $s_{\bar{Y}}$. This equation shows one value of a large number of replications: a reduction in the standard error. The **standard error of the mean is central to ANOVA.**

It is important to realize that the variance, population or sample, of a mean decreases inversely as n whereas the standard deviation of a mean decreases inversely as \sqrt{n} .