CR-171872
C.1

management and technical services company

TECHNICAL INFORMATION RELEASE

TIR 2114-MED-4005

FROM	TO
R. Srinivasan, Ph.D.	Dr. Nitza Cintron-Trevino/SD4

DATE	CONTRACT NO:	T.O. OR A.D. REF:	MIS OR OTHER NASA REF:
9/7/84	NAS9-17151		

SUBJECT

Statistical Analysis Techniques for Small Sample Sizes

A problem which is encountered when dealing with analysis of space-flight data is that of small sample sizes. Resource and cost considerations limit the number of experimental subjects available on each flight, thus greatly limiting the amount of data obtained and the power of the results derived. In the light of such a small amount of data available, careful analyses are essential in order to extract the maximum amount of information with acceptable accuracy. This report is concerned with statistical analysis of small samples. It begins with the background material necessary for understanding statistical hypothesis testing and then explains with examples the various tests which can be done on small samples. Emphasis is on the underlying assumptions of each test and on considerations needed to choose the most appropriate test for a given type of analysis.

R. Srinivasan

R. Srinivasan, Ph.D.

/db

Attachment

(NASA-CR-171872) STATISTICAL ANALYSIS TECHNIQUES FOR SMALL SAMPLE SIZES (Management and Technical Services Co.) 94 p HC A05/MF A01 CSCI 12A

N85-26279

Unclas G3/65 21219

Unit Manager	Approving Manager	NASA Concurrence
	F. A. Kutyna, Ph.D. 21 Sept 84	

DISTRIBUTION

NASA/JSC:
Dr. M. Bungo/SD3
Dr. J. Charles/SD3
Billy J. Jefferson/BE
Dr. P. Johnson/SD3
Dr. J. Logan/SD2
Dr. W. Shumate/SD
Dr. V. Schneider/SD3

MATSCO:
R. F. Meyer
Biomed. Group

TIR 2114-MED-4005

STATISTICAL ANALYSIS TECHNIQUES FOR SMALL SAMPLE SIZES

Prepared for
National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, Texas

Prepared by
Sharon E. Navard
Management and Technical Services Company
Houston, Texas

September 1984

This work was supported by NASA contract NAS9-17151

ABSTRACT

A problem which is encountered when dealing with analysis of spaceflight data is that of small sample sizes. Resource and cost considerations limit the number of experimental subjects available on each flight, thus greatly limiting the amount of data obtained and the power of the results derived. In the light of such a small amount of data available, careful analyses are essential in order to extract the maximum amount of information with acceptable accuracy. This report is concerned with statistical analysis of small samples. It begins with the background material necessary for understanding statistical hypothesis testing and then explains with examples the various tests which can be done on small samples. Emphasis is on the underlying assumptions of each test and on considerations needed to choose the most appropriate test for a given type of analysis.

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLES

STATISTICAL ANALYSIS TECHNIQUES FOR SMALL SAMPLE SIZES

1.0 INTRODUCTION

When working with estimating population parameters based on only a sample of the population, it is logical to expect that better estimates will result from larger sample sizes; in fact, values of sample parameters approach those of the population as the sample size increases. For large samples (usually $N \geq 30$ is sufficient), a powerful statistical tool called the Central Limit Theorem provides the basis for obtaining acceptable results. In many situations, however, it is not possible to obtain samples of such large size. Space flight is a prime example where the limited available resources render large samples infeasible. This problem has been dealt with in the past by combining data from several flights. For example, in analysing some of the Skylab data, the data from three separate manned flights with three crewmembers each were pooled, thus producing a combined sample size of nine (1, 2). Care should be taken when combining data from different sources, however. Experimental conditions will never be identical from one flight to the next, and these differences might undermine the underlying assumptions of the analyses and thereby falsify the results. When planning an experiment, careful consideration should be given to the effect of the sample size on the outcome of the experiment, and the type of analysis chosen should permit extraction of the maximum amount of information with desired accuracy from the available data. If data are combined from different sources, this fact should be incorporated into the analysis. Techniques that are particularly useful for analysing data from small samples taken from different types of situations are presented in this report. Emphasis is placed on the assumptions inherent in each test and on considerations needed in choosing the type of analysis.

2.0 BACKGROUND

This section gives the background information necessary to understand the statistical tests described in the next section. It is very basic, beginning with the purpose of statistics, and then developing the basics of hypothesis testing. Characteristics of tests such as the level of significance, the power, and the relative efficiency are defined, and the distinction is made between parametric and nonparametric tests and the various scales of measurement. Finally, there is a discussion of the central limit theorem. Anyone familiar with these topics may skip this section and go on to Section 3. Computation of simple sample parameters such as sample mean and standard deviation can be found in an earlier report (3).

2.1 PURPOSE OF STATISTICS

The purpose of statistics is to ascertain, within a specified degree of accuracy, the characteristics, or parameters, of a population, using observations made only on a sample of the population. The values of these parameters could be determined exactly, of course, if observations are available on every individual member of the population in question, say, astronauts. No statistics would be necessary in such a case. Unfortunately, observations on the entire population are seldom (if ever) possible, and so we must resort to the next best thing: take a sample of the population, make the observations on those few individuals, and from the data thus obtained, try to infer the characteristics of the entire population.

Unfortunately, sources of error inherent in any experiment will prevent the sample parameter values from being identical to the values of the population parameters. One of these sources is observation error; the accuracy of the results will be influenced by the precision of the instruments

and methods used to obtain the observations. One should be able to obtain a good estimate of this error before the start of the experiment, so that its effect on the results can be explained.

The main source of statistical error, i.e., the deviation of sample parameter values from population parameter values, however, will be due to the subjects themselves. This error is caused both by between-subjects variation and by within-subjects or day-to-day variation. The between-subjects variation arises from the fact that no two individuals are exactly the same and therefore observations on them will necessarily differ. Within-subjects variation arises from the fact that the characteristics of the same individual will change over a period of time and hence the observations taken on one day probably won't be the same as those taken on another day.

Because of these various sources of error in the data, it is not possible to determine exactly the "true" values of the population parameters. This is where statistics can be of help. Statistical techniques have been developed to estimate the values of the parameters in question (both single point estimates and interval estimates) and to determine the probability that these estimates are correct. Building on this, it is also possible to test whether parameter values between populations are the same, or whether different factors (e.g., weightlessness) have any effect on parameter values.

2.2 BASICS OF HYPOTHESIS TESTING

The roots of these statistical inference techniques lie in the theory of probability and probability distributions. For example, take the simple experiment of flipping a coin. For a fair coin, there are two possible outcomes, heads and tails, each with its associated probability, 1/2. Now flip the coin 10 times and count the number of heads. How many will there be?

One can't say for sure, because there is a lot of 'within-subject variation' in the coin; roughly half the time it will be one value, half the time the other. So one would expect and guess that there would be five heads, which is half of ten. But what is the probability that there will, in fact, be exactly five?

2.2.1 The Binomial Distribution

Since we know that the probability of getting a head on a single toss is 1/2, we can easily figure out the probability of getting any number of heads that we want; all that we need to do is figure out the probability of getting x heads and (10 - x) tails and then multiply it by the number of possible combinations of x heads and (10 - x) tails. This in fact follows what is called the binomial distribution with parameters n(sample size) = 10 and p(probability of a head) = 1/2. The probability density function (pdf) of the binomial distribution is given by

$$p(x)=\binom{n}{x}p^x(1 - p)^{n-x}.$$

Using the above formula, the distribution of heads in ten tosses of a coin is given by:

x	p(x)
0	.0010
1	.0097
2	.0440
3	.1172
4	.2051
5	.2460
6	.2051
7	.1172
8	.0440
9	.0097
10	.0010

Looking at this table it can be seen that, while five is the number that is most probable, it actually occurs less than one quarter of the time.

Now let's turn things around. Instead of predicting in advance how many heads will be obtained, let's count the number of heads in ten tosses and try to determine whether or not the probability of getting a head actually is equal to 1/2. To put this in statistical terms, we want to test the null hypothesis H_0 that the probability of heads (denoted by p(heads)) = 1/2.

2.2.2 Level of Significance

Common sense dictates that if the probability of heads is 1/2, then the number of heads will be close to five; but the question is, how close is "close." To determine this, the experimenter must first decide how certain he wants to be that his results are correct, that is, his "level of significance." He can never be absolutely sure that the null hypothesis is not true; even if he flips the coin ten times and doesn't get any heads, it doesn't necessarily mean that the coin is not fair. However, such an outcome is unlikely enough that its occurrance would lead one to infer that the coin was not fair. Note that since $p(x) = 0.001$ for $x = 10$, approximately one out of every thousand trials with a fair coin will result in no heads. This is the level of significance, denoted usually by α. It is the probability of obtaining by random chance a value which the investigator is willing to accept as disproving the null hypothesis. In other words, α is the probability of rejecting the null hypothesis when it is true. The value of α should always be determined before the start of the experiment; as the value of α decreases, the significance increases.

If the experimenter would be satisfied with a result that would occur by chance only one in twenty times, then he would set the significance level at 1/20 or 0.05. Suppose he does this, then flips the coin and gets two heads. Should he accept or reject the null hypothesis that p(heads) = 1/2?

Since he is only interested in whether the probability equals 1/2, and not in whether it is larger or smaller than 1/2, outcomes with both large numbers and small numbers of heads will lead to the rejection of the null hypothesis. Therefore, our observed value of two should be matched with the corresponding value on the other end, i.e., 10-2 = 8. Furthermore, the numbers even farther from our proposed value [equal to 5 for p(heads) = 1/2] than two and eight should also be considered; i.e., zero, one, nine, and ten. In other words, we are interested in the probability of getting a number as far or fartner from five than two and eight. Adding the probabilities of zero, one, two, eight, nine, and ten, one sees that the probability of this occurring by chance is 0.1094, more than twice the level of significance. The number 0.1094, denoted by $\hat{\alpha}$, is the actual level of significance of the experiment. It is the probability of obtaining by random chance a number at least as extreme as the observed value if one assumes that the null hypothesis is true. One will reject the null hypothesis only if $\hat{\alpha}$ is less than or equal to α, the predetermined level of significance. In this case, since $\hat{\alpha}$ is over twice the value of α, the experimenter must accept the hypothesis that the probability of heads is 1/2.

Suppose the experimenter was interested in knowing if the coin was biased in favor of tails. If this were so, then the number of heads would be small. The null hypothesis in a case like this is "H_0: the number of heads is greater than or equal to five" versus the alternate hypothesis "H_a: the number of heads is less than five." Suppose he flips the coin and again obtains two heads. This time only small numbers will lead to the rejection of the null hypothesis. To determine if this is significant, one need only add up the probabilities of getting two heads or less; i.e., the probability of

obtaining a zero, one, or two. Doing this, we see that the probability is 0.0547. This is still larger than the pre-determined significance level of 0.05, so the experimenter must still accept the null hypothesis; he does not have sufficient evidence to reject the hypothesis that the expected number of heads is greater than or equal to five.

2.2.3 One and Two Tailed Tests

It would, of course, be possible for the experimenter to determine in advance what kind of numbers he would have to get in order to reject the null hypothesis. For example, take the test of "H_0: the mean number of heads is five." This is what is known as a "two-tailed" test because both large and small values will lead to the rejection of the hypothesis. In a two-tailed test, the level of significance is divided as evenly as possible between both ends. In a symmetric distribution, i.e., one in which the probabilities are distributed evenly about the mean, this can be done exactly. If α = 0.05, then we want $\alpha/2$ = 0.025 to be at each end of the distribution. Thus, to determine which values will lead to the rejection of the null hypothesis we need only add up the probabilities, starting with zero, and keep going as long as the sum is less than or equal to $\alpha/2$ = 0.025. In the coin tossing experiment the probability of zero or one is 0.0107; but if two is added, it is greater than 0.05, much larger than 0.025. Therefore, zero and one, and their corresponding values of nine and ten at the other end, will constitute the rejection values or critical values for this experiment; i.e., if one flips a coin ten times and obtains zero, one, nine, or ten heads, he will reject the null hypothesis and conclude, at α = 0.05, that the coin is not fair.

A null hypothesis which specifies the mean to be greater than or equal to five (or less than or equal to five) is tested by a "one-tailed test"

because all of the critical values lie at one end of the distribution. In this case we determine whether large or small values will lead to the rejection of the null hypothesis, then go to that end of the distribution and add up the probabilities, keeping the sum less than or equal to α. In the coin tossing experiment, the critical values for "H_0: the mean number of heads is greater than or equal to five" will consist of the numbers, starting with zero, such that the sum of their probabilities is less than or equal to 0.05. Adding these probabilities, we see that this region consists of zero and one, because the addition of the probability associated with two heads makes the sum larger than 0.05. Therefore, we would reject the hypothesis that the mean is greater than or equal to five only if we obtain a value of zero or one.

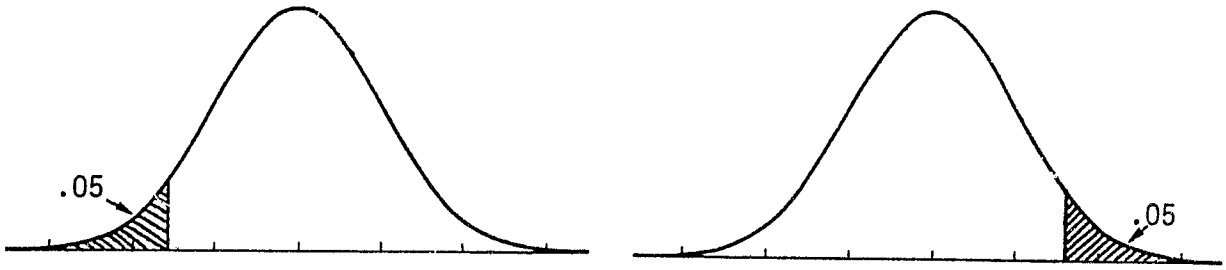
2.2.4 Discrete and Continuous Distributions

The coin-tossing experiment described above is somewhat unusual in that the one-tailed critical values are exactly the same as the two-tailed values at each end of the distribution. Generally the one-tailed critical values will be closer to the hypothesized mean than the two-tailed values. This was not the case because the underlying distribution (binomial) was discrete, meaning it had a finite number of sample points, and the increase in probabilities from one to two was relatively large. Another inconvenience that arises when working with discrete distributions is that it is usually impossible to find critical values with probabilities that sum exactly to α. Generally the sum will be less, as in the above example; in the two-tailed test, α was actually 0.0214 and in the one-tailed test, it was 0.0107.

These problems don't arise when working with continuous (having an infinite number of sample points) distributions. Since there are an infinite

number of points, the probability associated with any one point is zero; therefore, it is necessary to work with intervals. Everything is exactly the same as in the discrete case except that instead of having specific critical values, there are critical regions corresponding to the areas under the curve of the distribution function. For example, in a two-tailed test with α = 0.05 there will be two critical or rejection regions, one in each tail of the distribution (thus, the term "two-tailed" test) and each having an area of 0.025. $\hat{\alpha}$ will be twice the area under the tail of the curve starting at the observed value.

In a one-tailed test, the critical region will be under only one tail of the distribution and will have an area of 0.05.



The null hypothesis will be rejected anytime the observed value lies in the critical region. $\hat{\alpha}$ is simply the area under the tail of the curve beginning at the observed value.

2.2.5 Summary of Hypothesis Testing

To summarize, the experimenter must decide on two things before the start of the experiment: the hypothesis that he wishes to test and the level of significance, α . The hypothesis, either one-tailed or two-tailed, will be stated as a null hypothesis vs. an alternate hypothesis. Generally the experimenter states what he is trying to disprove as the null hypothesis, i.e., he assumes that it is true and tries to find sufficient evidence to say that it is not true. For example, if one is trying to show that a certain parameter M is greater than 50, then he will set up the one-tailed hypotheses as:

H_0: $M \leq 50$

H_a: $M > 50$

Failing to reject the null hypothesis does not give any statistical evidence to say that it is true; it only means there is not sufficient evidence to conclude that it is false. Rejecting H_0, on the other hand, does give statistical significance to the falsity of the null hypothesis and therefore the truth of the alternative. The level of significance, α , gives the probability of rejecting H_0 when it is in fact true.

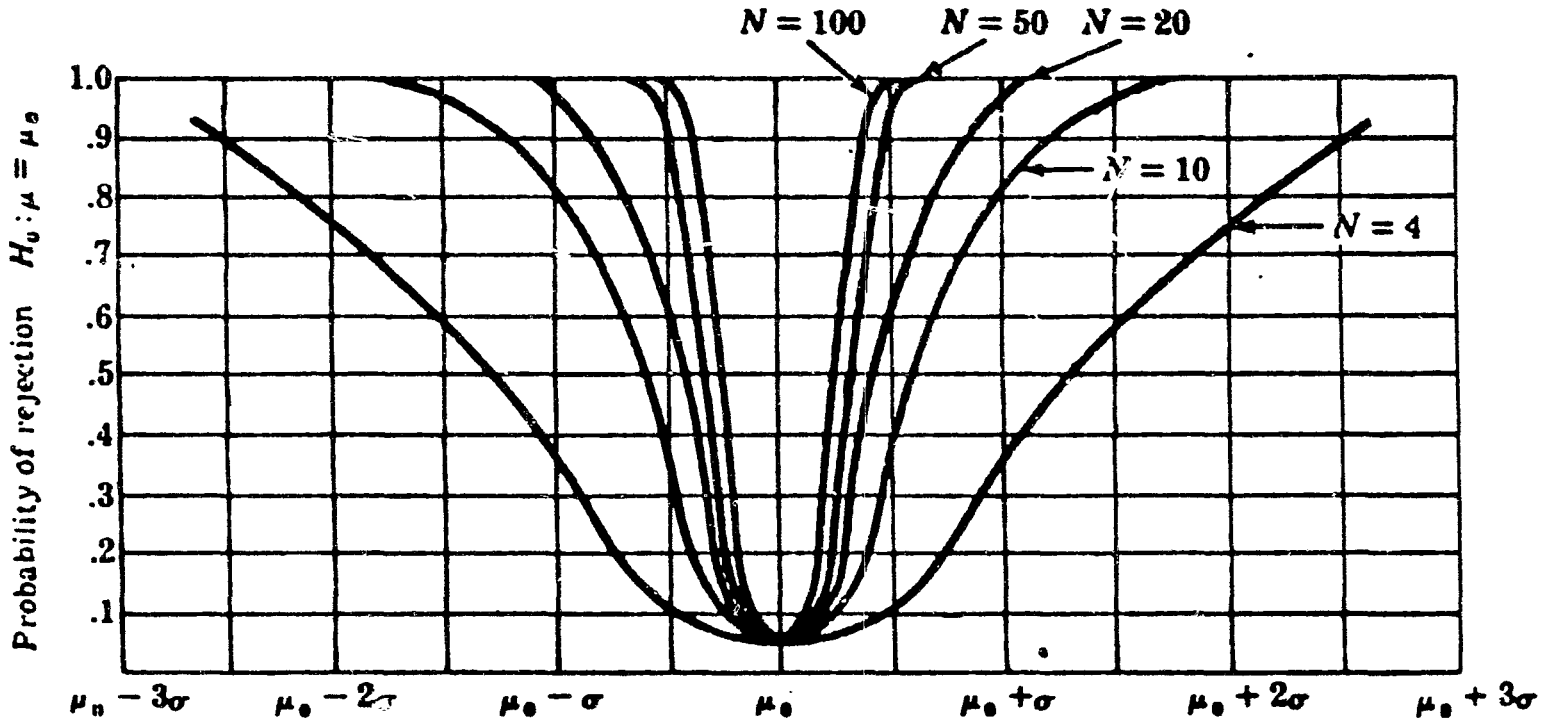
After the experiment is done, the observed value, or test statistic, is computed. This value is then compared to the distribution of all possible samples of that type. If the observed value is one that would occur less than or equal to α of the time by random chance, then the null hypothesis is rejected in favor of the alternative. The actual level of significance of the experiment, $\hat{\alpha}$, can be computed. This is simply the probability of obtaining a value at least as extreme as the observed value if H_0 is true. If $\hat{\alpha} \leq \alpha$, then H_0 will be rejected.

2.3 POWER

There is another measure of the validity of a test besides the the level of significance, α . The measure α itself is the probability of making an error, i.e., rejecting the null hypothesis when it is true. This is known as a Type I error, and is controlled in the experiment. There is also another type of error, specifically, accepting the null hypothesis when it is false. This is known as a Type II error and is denoted by β , and is not easily controlled because the true population parameter is not known prior to the experiment. The measure α is easily controlled, because it assumes that H_0 is true; therefore the value of the parameter is assumed to be a specific value and probabilities are easily computed thereafter. The measure β , however, assumes that H_0 is false, thus implying that the value of the parameter is something other than the specified value, but it is unknown.

Instead of working with β , the probability of a Type II error, statisticians generally work with the quantity $1 - \beta$, which is the probability of rejecting the null hypothesis. This quantity $1 - \beta$ is known as the power of the test. Power is a function of the sample size, the level of significance α , and the number of standard deviations of the true mean from the hypothesized mean. Generally power curves are shown as functions of α and the distance between the true and hypothesized means, with a different curve for each sample size. At zero distance, when the true and hypothesized means are the same, every curve will have a power of α , since it is the probability of rejecting the null hypothesis when true. As the distance from the mean increases, the curves change according to sample size. The smaller the sample size, the flatter the curve and thus the less the power; the larger the sample size, the greater the power. (Figure 1)

FIGURE 1: Power curves for two-tailed tests from a normal distribution, $\alpha = .05$. (Reference: Roscoe, 1969)



Power curves can be used to determine the sample sizes needed in an experiment, but to do so one would need an estimate of the variance σ^2 of the observations in the experiment. If such an estimate is available, one can determine the size difference desired to be detected, and for various values of α and $1 - \beta$, the sample size needed can be determined from the curves. For example, from Figure 1 it can be seen that to detect a difference of one standard deviation when $\alpha = 0.05$ and $\beta = 0.20$, a sample of 10 is required. To get a very powerful, highly significant test for a small difference, a very large sample size will be required. If, by the nature of the experiment, only a small sample size is possible, some compromise is needed. Either the difference to be detected must be increased, or the power and/or significance must be lowered. In planning an experiment it is often desirable to check these things in advance. It may be that with the available sample size, to detect the desired difference at a reasonable level of significance, the power would be so low that it might not justify the cost of the experiment. It will at least give the experimenter an estimate of his chances of detecting a difference.

2.4 EFFICIENCY AND ASYMPTOTIC RELATIVE EFFICIENCY

Another concept which is related to the two types of error and sample sizes, and one which can be used to compare different tests, is efficiency. The efficiency of one test relative to another is simply the ratio of the sample sizes required to test the same H_0 against the same H_a with the same values of α and β. For example, suppose we are testing a hypothesis, and we want $\alpha = 0.05$ and $\beta = 0.10$. Suppose there are two different tests that we can use; Test 1 would require a sample size of 30 to get the required accuracy, whereas Test 2 would require a sample size of only 20. Then the

efficiency of Test 2 relative to Test 1 is $n_1/n_2 = 30/20 = 1.5$. Anytime there is a choice between two possible tests, the one with the highest relative efficiency will be the better one to use because it will require a smaller sample size to obtain the same results.

The relative efficiency is not a very practical comparison to use, however, because it depends on the hypotheses and α and β, and thus would have to be computed for every situation. A measure which is independent of α and β is the asymptotic relative efficiency (A.R.E.) of one test to another, which is computed by holding α and β constant and letting n_1 approach infinity, then taking its ratio with the corresponding value of n_2. If this ratio n_2/n_1 approaches a constant for all sequences of tests with different α and β, which it frequently will, it is the A.R.E. of the first test relative to the second. Although the A.R.E. is computed for very large sample sizes, studies have shown that it is often a good approximation to the relative efficiency of small sample sizes in many practical situations, and is thus a good measure of the relative efficiency of two tests.

2.5 PARAMETRIC AND NONPARAMETRIC TESTS

One may wonder why it is even of interest to compare two tests when, as in the case of the coin tossing experiment, we know the exact distribution of the possible outcomes, i.e., the sampling distribution. The answer is simple; if the exact distribution is known, then it should be used. Most experiments, however, are more complex than tossing a coin, and in many cases, it is impossible to know exactly how the sample is distributed. When this is the case, one must use some type of test that does not depend on the distribution of the sample. Tests of this type are known as nonparametric

tests, and in many cases there are many different tests that could be used on a given set of data. In a situation such as this, the A.R.E. can be used as a guide to help determine which test should be used for maximum efficiency.

Tests which do assume that the exact form of the sampling distribution is known are called parametric tests. The coin-tossing experiment was an example of a parametric test, with the underlying distribution being the binomial. Anytime the exact distribution is known, the parametric tests will be more sensitive than any comparable nonparametric tests. However, if any of the assumptions for the parametric tests are not met, then it is possible that the nonparametric tests will be more powerful. Although parametric tests are more sensitive, they are very limited in the situations in which they can be used. All parametric tests presented in this paper will assume the normal distribution. Nonparametric tests are applicable to a much wider range of situations because they have fewer or less restrictive assumptions.

2.6 RANDOMIZATION

One very important assumption that is made by all tests, both parametric and nonparametric, is that the sample that is taken be random; that is, all elements in the population should have an equal chance of being included in the sample. If the sample is random, the sampling distribution can be estimated mathematically. If it is not random, the sampling distribution will be unknown, or at least the accuracy with which it is estimated will be unknown. A good approximation of the sampling distribution must be obtained in order to determine the precision of the inferences about the population which are made from the sample.

2.7 SCALES OF MEASUREMENT

Another consideration important to any particular test is the scale of measurement used in obtaining the data. There are four possible scales of measurement: the nominal, ordinal, interval, and ratio scales. The nominal scale uses numbers merely as a name; for example, in flipping a coin, one could assign "heads" a '0' and "tails" a '1'. These numbers are arbitrarily assigned and have no numerical meaning. In the ordinal scale, numbers can be ordered as "less than," "greater than," or "equal to." For example, in a race, the winners are assigned first, second, and third place. No measure of the amount of difference between these numbers is given. In an interval scale, the size of the difference between numbers (thus, "interval") is meaningful. An interval scale must be based on a unit distance as compared to a zero point; the zero, however, is arbitrarily assigned. Temperature is something which is usually measured on an interval scale. The last scale of measurement, the ratio scale, has all of the characteristics of the interval scale except the zero point is meaningful, thereby giving meaning to ratios between two measurements. Height and weight are measured on a ratio scale.

2.8 THE CENTRAL LIMIT THEOREM

When determining the distribution of the sample, the size of the sample plays an important role. If the sample size is large, the analysis can often be simplified by using the central limit theorem. If Y_n is some statistic based on a sample of size n from any distribution, and μ_n is its mean and σ_n^2 its variance, then the central limit theorem says that the distribution of $(Y_n - \mu_n)/\sigma_n$ approaches the standard normal distribution (normal with mean zero and variance one) as n approaches infinity. In other words, if one takes any statistic from a large enough sample (usually $n \geq 30$

is adequate for a good approximation), and subtracts its mean and divides by its standard deviation, the result will have a normal distribution with mean zero and variance one, irrespective of the form of the original distribution. The number obtained by doing this transformation is simply the number of standard deviations that the value is away from the mean of the standard normal. Probabilities for the standard normal have been extensively tabulated; one need only look up the required number in a normal table to determine the area under the curve up to that point; i.e., the probability of obtaining a value that extreme by random chance.

Unfortunately, in many situations the sample size is not adequately large to justify invoking the central limit theorem. In these cases one must either use the exact distribution of the sample or, if the distribution of the sample is unknown or if the measurement scale is insufficiently powerful, use the appropriate nonparametric tests. The most widely used tests for small sample sizes are given in the following sections.

3.0 ONE-SAMPLE TESTS FOR LOCATION

Perhaps the simplest type of test that one would wish to perform is to determine whether the mean or median of the population is equal to a specified value. Depending on the assumptions that can be made about the underlying distribution, several different types of tests can be used to test for location.

3.1 PARAMETRIC: ONE-SAMPLE T-TEST FOR A DIFFERENCE IN MEANS

Assumptions

(i) The observations $X_1,\ldots,X_n$ constitute an independent random sample from the population.

(ii) The sample is taken from a normally-distributed population.

(iii) The measurement scale is at least interval.

(iv) The measurements are continuous.

The test statistic used to test the hypothesis "H_0: $\mu = \mu_0$" is $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, where x is the mean of the sample, s is the standard deviation of the sample, μ_0 is the constant we are assuming is the value of the population mean (according to the null hypothesis) and n is the sample size. Notice that this statistic looks very much like the statistic in the central limit theorem; we are taking a number, subtracting the mean from it, and dividing the result by the standard deviation. This is proper even though we have taken a small sample because the means of samples from normal distributions are normally distributed. However, this statistic does not follow a standard normal distribution because of the necessity to estimate the standard deviation. This is accounted for by comparing the computed t to the proper quantile of a t distribution with n - 1 degrees of freedom (DF). The larger the degrees of freedom, the closer the t-distribution comes to the

standard normal because the estimate of the variance gets better. Tabled values of the t-distribution are available in most books on applied statistics.

As an example, suppose an experimenter wants to test the hypothesis that a certain population has a mean of six against the alternate that the mean is not equal to six, with $\alpha = 0.05$. He takes a random sample of size eight and obtains the following numbers: 4.6, 6.3, 5.2, 3.7, 4.8, 6.0, 4.7, 5.3. For this sample $\bar{x} = 5.075$ and $s = 0.8242$. So the t-statistic is

$$t = (5.075-6)/(.8242/\sqrt{8}) = -3.1743.$$

Since this is a two-tailed test, we compare this number to the $\pm$ $(1 - \alpha/2 = (0.025, 0.975)$ quantile of a t distribution with $(n - 1) = 7$ degrees of freedom. This value is $\pm$ 2.365. Since $-3.1743 < -2.365$, we reject the hypothesis that $\mu = 6$. To determine $\hat{\alpha}$, we would need to interpolate between the 0.975 and 0.995 values of t_7, 2.365 and 3.499, respectively. From this we obtain that $\hat{\alpha}/2 = 0.01073$, so $\hat{\alpha} = 0.02146$. In other words, in repeated trials from a population with a mean of 6, observations this extreme would occur by chance only about 2 percent of the time.

A confidence interval can also be obtained for our estimate of the true mean of the population. A $100(1 - \alpha)$ percent confidence interval gives limits between which we are $100(1 - \alpha)$ percent certain that the true mean of the population lies. For this test,

$$\bar{x} - t_{[\alpha/2;\ n-1]}\, s/\sqrt{n} \leq \mu \leq t_{[1-\alpha/2\ ;\ n-1]}\, s/\sqrt{n} + \bar{x}$$

For our example, the 95 percent confidence interval is given by $5.075 \pm 2.365(0.8242/\sqrt{8}) = (4.386, 5.764)$. Thus, we are 95 percent certain that the

true mean of the population lies somewhere between 4.386 and 5.764. Notice that this confidence interval does not contain the hypothesized value; this will be true if and only if H_0 was rejected.

There is one more aspect of this test which should be noted. If the original hypothesis had been one-tailed, i.e., "H_0: $\mu \leq 6$" or "H_0: $\mu \geq 6$," then the test statistic would have been compared to the $1 - \alpha = 0.95$ quantile of the t_7 distribution, which is 1.895. We would reject the hypothesis "H_0: $\mu \geq 6$" if $t < -1.895$, and the hypothesis "H_0: $\mu \leq 6$" if $t > 1.895$. However, the confidence interval would be exactly the same for all three tests as long as a two-sided confidence interval is desired, as is usually the case.

3.2 NONPARAMETRIC TESTS

3.2.1 One-Sample Sign Test

Assumptions

(i) The sample is a random sample from a population with unknown median.

(ii) The measurement scale is at least ordinal.

(iii) The variable of interest is continuous.

The sign test is used for testing the hypothesis that the median M of the population is equal to a certain value; i.e., "H_0: $M = M_0$". The procedure is very simple. All that one needs to do is subtract M_0 from each of the sample values and record the sign; in other words, count how many points are above and below M_0. If any point is exactly equal to M_0, it is discarded. The test statistic is simply the smaller of these numbers for "H_0: $M = M_0$"; it is the number of minuses, or the number of values lower than the median for the hypothesis "H_0: $M \geq M_0$"; and it is the number of pluses, or

the number of values greater than the median for the hypothesis "H_0: $M \leq M_0$". The test statistic is then compared with the probability values from the binomial distribution with $p = 1/2$ and n^* = the number of points left after the zero differences are discarded. This is done exactly as in the coin-tossing experiment. Binomial tables can be found in most applied, especially nonparametric, statistics books.

Using this test on the data in the previous example, we see that there are six minuses, one plus, and one zero. Therefore $n^* = 7$ and our test statistic $T = 1$ for the hypothesis "H_0: $M = 6$". Looking in the binomial tables for $n = 7$, $p = 1/2$, and $\alpha = 0.05$, we determine that the critical region, of actual size 0.0156, contains the points (0, 7). Since $T = 1$, we have insufficient evidence to reject the hypothesis that the median of this population is six. $\hat{\alpha}/2 = P(x \leq 1) = 0.0625$, so $\hat{\alpha} = 0.1250$.

Confidence intervals for the median based on the sign test can also be obtained from the binomial tables. Let K be the largest value of x for the binomial with parameters n^* and $p = 1/2$ such that $P(x \leq K) \leq \alpha/2$. The $(K+1)^{th}$ smallest observation is the lower limit and the $(K+1)^{th}$ largest observation is the upper limit.

In this example $P(x \leq 0) = 0.0078$ and $P(x \leq 1) = 0.0625$, so K=0 and K + 1 = 1; thus the smallest and largest values are themselves the endpoints for the confidence interval. Therefore we are 98.44 percent certain that the true median of the population from which the sample was drawn is between 3.7 and 6.3. Notice that this confidence interval contains the hypothesized value of six, and that the hypothesis was accepted, while the confidence interval formed using the t-test did not contain six and the hypothesis was rejected. In general, anytime the confidence interval does not contain the hypothesized value, the hypothesis will be rejected.

It can be seen from this example that for a sample of this size, the power of the sign test is not as great as that of the t-test for normal samples. For very small samples, the relative efficiency of the sign test compared to the t-test is approximately 0.95, but the efficiency decreases as the sample size increases. For a sample size of 13, the relative efficiency is only 0.75 and the A.R.E. is only 0.637. However, if the distribution begins to depart from normality, the power of the t-test becomes less and less, depending on how non-normal the distribution is. If it is too far removed, the sign test will be more powerful. Also, the sign test can be used on ordinal data while the t-test cannot.

3.2.2 Wilcoxon Signed-Ranks Test

The sign test uses only the sign of the differences between the points and the assumed value for the median. Thus, a considerable amount of information is not utilized. The Wilcoxon Signed-Ranks Test also makes use of the magnitude of the differences. This makes it a more powerful test but it also requires more limiting assumptions.

Assumptions

(i) The data constitute an independent random sample with unknown median M.

(ii) The variable of interest is continuous.

(iii) The measurement scale is at least interval.

(iv) The sampled population is symmetric.

The procedure is as follows: first subtract the assumed value for the median M_0 from each of the data points. Then, disregarding the signs, rank these differences from smallest to largest, throwing out zero differences. If any of the absolute differences are the same, assign the average of the ranks that would have been assigned to all of them. For

example, if the two smallest values are identical, assign each the rank 1.5. Then assign to each of these ranks the sign of the original difference. Take the sum of all of the positive ranks and call it T^+; likewise, sum the negative ranks and call this sum T^-.

For testing "H_0: $M = M_0$", the test statistic is $T = \min(T^+, T^-)$; for "H_0: $M \geq M_0$", T^+ is the test statistic; and for "H_0: $M \leq M_0$", T^- is the test statistic. Each of these statistics should be compared to the table values in a table of d-factors for the Wilcoxon Signed-Ranks Test for the appropriate n and d, where $d \geq T$. If the corresponding table value of α" (the probability of obtaining that particular n and d when $M = M_0$) is less than or equal to α, the null hypothesis should be rejected. The table of values for this test can be found in many nonparametric statistics books.

Using the same example as before, we obtain the following results:

X_i	$X_i - M_0$	R_i	
4.6	-1.4	-6	
6.3	.3	1	$T^+ = 1$
5.2	-.8	-3	$T^- = 27$
3.7	-2.3	-7	
4.8	-1.2	-4	$T = 1$
6.0	0	--	
4.7	-1.3	-5	
5.3	-.7	-2	

The value of the test statistic T is 1 for "H_0: $M = 6$". The table value of α" for $n = 7$ and $d = 1$ is 0.016; since this is less than 0.05 we reject H_0 and conclude that the median of the population is <u>not</u> six. Because of the way the table is set up, $\alpha" = \hat{\alpha} = 0.016$.

A confidence interval for the median can be constructed by finding the d value for the appropriate size n which is closest to the desired confidence coefficient, then taking the d^{th} smallest and d^{th} largest averages U_{ij}, where $U_{ij} = (x_i + x_j)/2$, $i \neq j$. For our example, α" for d = 3 is 0.046; α" for d = 4 is 0.078 which is larger than 0.05; so we will form a 100(1 - 0.046) = 95.4 percent confidence interval by taking the third smallest and third largest averages. The third smallest U_{ij} is given by (3.7 + 4.8)/2 = 4.25, and the third largest is given by (6.3 + 4.8)/2 = 5.55. Thus, we are 95.4 percent certain that the true median (also, since the distribution was assumed to be symmetric, the mean) lies between 4.25 and 5.55.

The A.R.E. of the Wilcoxon Signed-Ranks Test relative to the t-test is 0.955 if the differences are normally distributed. In other words, not much is lost in using this test over the t-test if the assumptions for the t are met. Furthermore, since this test is good for any symmetric distribution, it will apply to more situations than will the t-test. If the distribution is badly skewed, the sign test will be the appropriate test. The A.R.E. of the sign test to the Wilcoxon Signed-Ranks Test is 2/3 for normally distributed differences, 1/3 for uniformly distributed differences, and exceeds one as the distribution of the differences becomes skewed.

4.0 DIFFERENCES IN LOCATION FOR TWO SAMPLES

A situation which is encountered more often than merely testing to see if the mean or median of a population is a specified value is the need to compare the means of two populations to determine if they are the same. This can come about in two different ways: either the two samples which are being compared are correlated in some way or they are completely independent.

4.1 TWO RELATED SAMPLES

Anytime there is reason to believe the measurements in one sample are in some way correlated with those in the other, some kind of test for related samples should be used. Such a situation exists whenever both sets of measurements are taken on the same group of individuals before and after a treatment is applied; i.e., whenever individuals are used as their own controls. This is referred to as a repeated measures experiment. There are also instances where two individuals are paired on the basis of the variable in question before the beginning of the experiment; one in each pair receives the treatment and the other serves as the control. In either one of these situations, a test for related samples should be utilized to account for the correlation.

In a test for related samples, the two samples need not be independent (although observations within each sample should be independent) and, in the case of the parametric test, the variances of the two samples need not be the same. However, in a paired test, pairing reduces the degrees of freedom, thereby reducing the power of the test if the samples actually are independent. Given two tests, one for paired data and one for independent samples, the paired test will require almost twice as many subjects to have the same power if there are no extraneous factors; i.e., if pairing criteria is

independent of the variable of interest. These tests are described in the following sections and are merely extensions of the one-sample tests discussed previously.

4.1.1 Parametric: Paired t-test

Assumptions

(i) The subjects for repeated measures or pairs for matched pairs constitute a random sample.

(ii) The distribution of the differences is normal in the populations specified by the null nypothesis H_0.

(iii) There is no carry-over effect from treatment to treatment or from measure to measure.

(iv) The measurement scale is at least interval.

In order to test for any difference between the means of two samples, the null hypothesis is written as "H_0: $\mu_1 - \mu_2 = 0$". Note that the difference can be a specified value d_0, in which case the null hypothesis becomes, "H_0: $\mu_1 - \mu_2 = d_0$". Also, the one-sided alternatives can be used to determine if the mean of one population is larger than the other, by any desired amount. For example, if one wants to see if the mean of one population is more than five units greater than that of the other, the null hypothesis can be stated as "H_0: $\mu_1 - \mu_2 \leq 5$".

This test is very simple to perform; all that one needs to do is take the difference $D_i = X_{i1} - X_{i2}$ for each pair, then perform the one-sample t-test on the differences, D_i, as if they were the actual observations. Thus, the test statistic is $t = (D - d_0)/(S_d/\sqrt{n})$, where D is the average of the differences, S_d is the standard deviation of the differences, n is the number of pairs, and d_0 is the hypothesized difference. In most cases, when the investigator is interested only in determining if there is a difference, d_0 will be equal to zero and the test statistic reduces to $t = D/(S_d/\sqrt{n})$. The

actual testing of the hypothesis and formation of confidence intervals are then accomplished in exactly the same manner as with the one-sample test, so those procedures will not be repeated here. The only thing to keep in mind is that the inferences made and confidence intervals formed are on the differences in the means, and not on the means themselves.

4.1.2 Nonparametric Tests

Anytime the measurement scale is only ordinal or if the normality assumption is not met, one will have to resort to the nonparametric tests.

4.1.2.1 The Sign Test

Assumptions

(i) The data consist of pairs of measurements from a random sample.

(ii) The pairs of measurements are mutually independent.

(iii) The measurement scale is at least ordinal within each pair, i.e., each pair may be designated a plus, a minus, or a tie.

(iv) The pairs are internally consistent, e.g., if P(+) > P(-) for one pair, the same is true for all pairs.

The sign test is used to test for differences in the medians of the two samples. The relationship between the sign test and the one-sample sign test is the same as that between the one-sample and paired t-tests. The differences between the members of the pairs are determined, and the test statistic is the number of pluses or minuses, depending on the hypothesis. Differences of zero are once again disregarded. The hypothesis is tested and confidence intervals are formed in exactly the same manner as in the one-sample case, the only difference being that the procedures in this case pertain to differences between medians rather than to the medians themselves. Hence these procedures will not be repeated here. The efficiency of the sign test in relation to the paired t-test is also the same as in the one-sample case.

4.1.2.2 Wilcoxon Matched-Pairs Signed Ranks Test

As was the case with the previous two tests, the Wilcoxon Matched-Pairs Signed Ranks Test is merely an extension of the one-sample case.

Assumptions

(i) The sample of pairs (X_i, Y_i) is random.

(ii) The distribution of the D_i's is symmetric.

(iii) The differences are mutually independent and have the same median.

(iv) The measurement scale of the differences is at least interval.

The procedure is basically the same as in the one-sample case, except that all inferences are made about the differences rather than about the means themselves. The differences between the members of each pair are obtained, their absolute values are ranked, then the signs are returned. The test statistic, test of hypothesis, and formation of confidence intervals are the same as for the one-sample test, as is the discussion of power and relative efficiency.

4.2 TWO INDEPENDENT SAMPLES

Anytime there is no correlation between the two samples, a test for independent samples should be used. As indicated earlier, the use of a test for related samples on independent measurements will reduce the power of the test by lowering the degrees of freedom. Likewise, the use of a test for independent samples on correlated data will cause a loss of sensitivity. Therefore, it is essential to determine whether or not the samples are independent before deciding which design and analysis to use.

Unlike the tests for related samples, the tests in this section are not mere extensions of the one-sample case because of added restrictions on

the samples. For the more powerful tests, it is required that the samples have the same variance. In such cases the tests on independent samples are generally more difficult to perform.

4.2.1 Parametric: The t-test for independent samples

Assumptions

(i) The data represent a random sample.

(ii) There is independence both within the two samples and between the two samples.

(iii) The dependent variable is normally distributed in both populations.

(iv) The two populations have equal variances.

(v) The measurement scale is at least interval.

The t-test for independent samples differs from other versions of the t-test in that it requires an estimate of the combined variance of the two samples. One of the assumptions of this test is that the variances of both populations be the same; however, we have two separate estimates for it, one from each sample. These two estimates can be combined to obtain the common estimate of variance of the population. This common estimate is given by:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where n_1 and n_2 and s_1^2 and s_2^2 are the sample sizes and variances of the two samples. The sample sizes do not have to be the same as long as all of the assumptions are satisfied. The standard error, or standard deviation of the mean, becomes

$$SE = \left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{\frac{1}{2}}$$

Note that if $s_1^2 = s_2^2 = s^2$, the formula reduces to: $SE = \left[s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{\frac{1}{2}}$

This is the same form as the standard error for the one sample t-test, SE = $(S^2/n)^{1/2}$. Thus the test statistic becomes

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = \frac{\bar{x}_1 - \bar{x}_2}{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{\frac{1}{2}}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means.

This test statistic is then used exactly as before; it should be compared with the proper quantile of a t distribution with $(n_1 + n_2 - 2)$ DF.

An example is now in order. Suppose an investigator makes measurements on two independent random samples and obtains the following results:

sample 1: 8.3 7.9 6.2 9.4 5.2 9.7 7.2 8.5

sample 2: 5.2 3.9 6.7 4.6 5.3 3.5 5.2 6.1

He wants to determine whether or not the two samples have the same mean which, since they are assumed to be normal with equal variances, implies they come from the same distribution. Therefore the hypotheses would be set up as:

H_0: $\mu_1 - \mu_2 = 0$; H_a: $\mu_1 - \mu_2 \neq 0$.

Computing the means and variances of the samples, one obtains

$\bar{x}_1 = 7.8$ $\bar{x}_2 = 5.0625$

$S_1^2 = 2.3714$ $S_2^2 = 1.1227$

$n_1 = 8$ $n_2 = 8$

The common variance is

$$7(2.3714 + 1.1227)/14 = 1.7471,$$

and the standard error is

$$[1.7471(1/8 + 1/8)]^{1/2} = 0.66088.$$

Therefore, the test statistic is

$$t = (7.8 - 5.0625)/.66088 = 4.1422.$$

Since this is a two-sided test, this should be compared with the 0.025 and 0.975 quantiles of the t distribution with 8 + 8 - 2 = 14 DF. These values turn out to be ± 2.145. Since 4.1422 > 2.145, we reject the null hypothesis and conclude that the two samples are different. To determine $\hat{\alpha}$, we look in the t table and see that $t_{[0.0005,\ 14]} = 4.140$, which is very close the value of our test statistic. Therefore $\hat{\alpha}/2 = 0.0005$, so $\hat{\alpha} = 0.001$.

Confidence intervals for the difference can also be obtained in exactly the same manner as before; a $100(1 - \alpha)$ percent confidence interval for the difference between the two means is given by

$$(\overline{x_1} - \overline{x_2}) \pm t_{[1 - \alpha/2;\ n1+n2-2]}(SE).$$

A 95 percent confidence interval for the difference between the means of the two populations in our example is given by

$$(7.8 - 5.0625) \pm 2.145(.66088) = (1.3199,\ 4.1551).$$

It is of interest to examine the robustness of this test; that is, how well it holds up under the breakdown of the assumptions. Departures from normality will not have too adverse an affect as long as the variable of interest has the same distribution in both populations. Lack of homoscedasticity (equal variances) also is relatively unimportant as long as the sample sizes are the same. Violation of both assumptions will tend to increase the probability of rejecting a true hypothesis to as much as twice the level of significance. As the sample size increases, however, both departures from normality and heterogeneity of variances become less important. For sample sizes of twenty-five or more, the test is basically insensitive even to drastic violations.

There are methods of testing the validity of the assumptions, but they are not very good for small samples. Tests for normality, such as the Chi-Square Goodness of Fit test, require a fairly large number of sample points to maintain accuracy, and the usual test for homogeneous variances is very sensitive to departures from normality when the sample sizes are unequal. This test, the F-test, is very easy to perform, however; it is simply the ratio of the variances, with the larger over the smaller. This statistic is then compared to the appropriate quantile of the F distribution with the appropriate number of DF associated with the two variances. The F-tables can be found in most applied statistics books. In the example just presented, for instance,

$$F = S_1^2/S_2^2 = 2.3714/1.1227 = 2.112.$$

This is compared with $F_{[1-\alpha;\ n1-1,\ n2-1]}$ (the F-test is always one-sided). This value turns out to be 3.79, so we accept the hypothesis of equal variances.

If the variances are not the same, the t-test can be modified to take this into account and give fairly good results. When the variances are not equal, use the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^{\frac{1}{2}}}$$

and compare it to

$$t_a = \frac{\frac{t_1 s_1^2}{n_1} + \frac{t_2 s_2^2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t_i is the α quantile of a t distribution with n_i-1 DF.

4.2.2 Nonparametric Tests

4.2.2.1 The Median Test

The median test is the only nonparametric test available which will compare independent samples coming from dissimilar distributions. It is also the first test we will describe which will compare more than two samples at a time. However, it is not very good for small samples, so the discussion of it will be brief.

Assumptions

(i) Each sample is a random sample.

(ii) The variable of interest is continuous.

(iii) If all populations have the same median, then each population has the same probability p of an observation exceeding the grand median.

To perform the median test, first obtain the grand median; that is, find the number which is exceeded by exactly half of the observations from all of the combined samples. Then count the number of observations in each sample that exceed the grand median, forming a table as follows:

Sample	1	2	-------	C	Totals
≥Median	O_{11}	O_{12}	-------	O_{1c}	a
<Median	O_{21}	O_{22}	-------	O_{2c}	b
Total	n_1	n_2	-------	n_c	N

The null hypothesis is "H_0: All c populations have the same median" versus the alternate "H_a: At least two populations have different medians." The test statistic is

$$T = \frac{N^2}{ab}\sum_{i=1}^{c}\frac{O_{1i}^2}{n_i} - \frac{Na}{b}$$

This statistic should be compared with the 1 - α quantile of a Chi-Square distribution with (c - 1) DF. The Chi-Square tables can be found in most applied statistics books.

As we mentioned previously, this test is not good for small sample sizes. In general, it is not good if more than 20 percent of the n_i's are less than 10 or if any of the n_i's are less than two. This disqualifies the set of data used in the previous example, and probably a lot of space flight data as well, so no example will be presented. The A.R.E. of the median test to the t-test for normal data is only 0.64, so by the time the sample sizes are big enough to use this test, the t-test would probably be more powerful unless the assumptions for the t-test are very drastically violated.

4.2.2.2 Mann-Whitney U Test

The Mann-Whitney Test involves a rank procedure, which makes it a more powerful test than the median test. It is also good for smaller sample sizes.

Assumptions

(i) Both samples are random samples from their respective populations.

(ii) There is independence both within each sample and between the two samples.

(iii) The measurement scale is at least ordinal.

(iv) If the two distribution functions differ, they differ in location only.

To perform the Mann-Whitney Test, the data from the combined samples are first ranked from 1 to $n_1 + n_2$. As before, in the case of ties all tied points are assigned the average of the ranks that would have been assigned had there been no ties. This test can be used to test the hypothesis "H_o: $\mu_1 = \mu_2$" vs. "H_a: $\mu_1 \neq \mu_2$", or any one-tailed variation.

If there are no or few ties, the test statistic for the Mann-Whitney Test is simply the sum of the ranks from population 1, i.e., $T = \sum R(X_{1i})$. This value is then compared to the proper quantile of the Mann-Whitney test statistic, the tables of which can be found in many nonparametric statistics books. If there are many ties this statistic can be normalized, thus obtaining

$$T_1 = \frac{T - \frac{n_1(N+1)}{2}}{\frac{n_1 n_2}{N(N-1)} \sum_{i=1}^{n} R_i^2 - \frac{n_1 n_2 (N+1)^2}{4(N-1)}}$$

and then comparing this to the proper quantile of the standard normal distribution.

As an example, let us use the previous data set.

X_{1i}	$R(X_{1i})$	X_{2i}	$R(X_{2i})$
8.3	13	5.2	5
7.9	12	3.9	2
6.2	9	6.7	10
9.4	15	4.6	3
5.2	5	5.3	7
9.7	16	3.5	1
7.2	11	5.2	5
8.5	14	6.1	8
	95		41

Because of the three-way tie with 5.2, T_1 would give the more precise distribution, but one three-way tie will not effect T significantly, so we will use T as our test statistic; i.e., $T = \Sigma R(X_{1i}) = 95$. From the table we find that the 0.025 and 0.975 quantiles of the Mann-Whitney test statistic are 50 and 86. Since 95 is not in this interval, we reject H_0 and conclude that

the means are different. Looking at the other α values, we see that for α = 0.001, the value is 95; therefore $\hat{\alpha}/2 = 0.001$ and $\hat{\alpha} = 0.002$. Although this is twice the $\hat{\alpha}$ for the t-test, it is still very highly significant and there is relatively little difference between them.

To determine a $100(1 - \alpha)$ percent confidence interval for the difference, determine the number $K = w_{[\alpha/2]} - (n_1)(n_1+1)/2$, where $w_{[\alpha/2]}$ is the $\alpha/2$ quantile of the Mann-Whitney Test Statistic. Then the $100(1 - \alpha)$ percent confidence interval will be bounded by the K^{th} largest and K^{th} smallest of the $n_1 n_2$ possible differences between the sample points.

In our example, $K = 50 - 8(9)/2 = 14$. Thus the 14^{th} smallest and 14^{th} largest differences will be the lower and upper limits of the confidence interval. If the differences are computed, it can be seen that the 14^{th} smallest is 1.2 and the 14^{th} largest is 4.4. Therefore we are 95 percent certain that the true difference between the means of the two samples lies between 1.2 and 4.4. This interval is a little wider than that of the t-test, but not much.

The Mann-Whitney stands up to the t-test very well in terms of efficiency. For any case where the two distributions differ only in location, the A.R.E. is never lower than 0.864 and may be as high as infinity. For normal data it is 0.955; for uniform, it is 1.0. The A.R.E. of the Mann-Whitney test relative to the Median test is 1.5 for normal data and 3.0 for uniform data. It can be seen from this that the Mann-Whitney test is a highly powerful nonparametric test.

4.3 HOLLANDER TEST OF EXTREME REACTIONS

This test is different from others in that, rather than testing for a difference in the means of two groups, it tests to see if there are opposite

extreme reactions in the experimental group. In some situations, it is possible that not every subject in the experimental group will react the same way to a treatment; some may demonstrate increases while others may show decreases. In a case such as this, the distributions will be drastically different, but any of the tests discussed so far will show the means to be the same and thus one might conclude that the distributions are the same. This test will determine if the experimental group has extreme reactions in opposite directions.

Assumptions

(i) The data consist of two independent random samples (X_1, $X_2,\ldots,X_{n1}$ from the control group and Y_1, $Y_2,\ldots Y_{n2}$ from the experimental group)

(ii) The measurement scale is at least ordinal.

The hypotheses tested are "H_o: The two distributions are the same" vs. "H_a: One distribution has extreme reactions in both directions." To perform the test, first rank the combined samples from one to $n_1 + n_2$. The test statistic is

$$G = \sum_{i=1}^{n_1} [R(X_i) - \overline{R(X)}]^2$$

where $R(X_i)$ is the rank of the i^{th} X value from the control group and $\overline{R(X)}$ is the average of the ranks of the X's. If the reactions of the experimental group go to opposite extremes, then it should have the small and large rank and the control group will have the middle ones around the mean. Therefore G should be small if there were extreme reactions. The value of G should be compared with the table value of G for the Hollander test, which can be found in some nonparametric statistics books. If the observed value is less than or equal to the table value, H_o should be rejected at the specified α level.

There is no need to consider one and two-tailed tests with this statistic; the nature of it makes it always one-sided.

As an example, suppose an experimenter performs an experiment and obtains the following results:

X_i	$R(X_i)$	Y_i	$R(Y_i)$
10.3	10	8.3	5
9.9	8	7.1	2
10.6	12	13.2	14
8.2	4	10.4	11
9.3	6	6.2	1
11.4	13	14.7	16
9.7	7	13.9	15
10.0	9	7.5	3
	69		67

Here, $\overline{R(X)} = 69/8 = 8.625$; $G = \Sigma\,(R(Xi) - 8.625)^2 = 63.875$. Looking up in the table for $n_1 = 8$, $N = 16$, we see that the value for $\alpha = 0.01$ is 67.88; thus, this is significant at the $\alpha = 0.01$ level, and we conclude that there were extreme reactions in the experimental group; i.e., the subjects responded to the treatments in different ways. It can easily be seen by examining the two sums of ranks (69 and 67) that no test for location would have shown the difference to be significant. However, neither the t-test nor Mann-Whitney Test would have been applicable in this case because the assumption that the distributions differ only in location has been drastically violated. This is a good example for showing how an investigator can get into trouble by not checking on the validity of his assumptions. If he were not careful, he would have concluded that the treatment in this experiment had no effect.

5.0 PROCEDURES FOR COMPARING MORE THAN TWO SAMPLES

The procedures which have been examined thus far (with the exception of the median test) are useful with only two samples to compare and when there is only one treatment done on the samples. In many experimental situations, this is not the case. Often there are three or more different populations which need to be compared, with more than one treatment or levels of treatments to be examined for each one. It is possible to do a t-test or a corresponding nonparametric test between every possible pair of combinations, but this is not a good practice because the tests are not independent. Also it increases α above the predetermined level. If twenty such comparisons are done at the $\alpha = 0.05$ level, the odds are that one of them will show significance just by chance, which implies that the α for the twenty comparisons is much larger than the level at which each comparison is done.

Therefore, some techniques should be used which will allow the simultaneous comparison of all of the means at the desired level of significance. There are several techniques which will allow for this, in many types of situations. The parametric tests employ a technique known as Analysis of Variance (ANOVA).

5.1 PARAMETRIC: ANALYSIS OF VARIANCE

The Analysis of Variance is exactly what it says it is: it compares the distributions of the various samples by analyzing the total variance broken down into its components. Suppose one has several experimental groups, drawn randomly from the same population, to which different treatments are applied. If the treatments had no effect, then all the groups would be identical. The total variance of the experiment can be computed in two ways: the squared deviation of each observation from the grand mean can be computed,

or the squared deviation of each observation from its group mean can be calculated, and added to the squared deviation of each group mean from the mean of the groups.

The key to this procedure is that both of these estimates of the variance, that within the groups and that between the groups, are estimates of the population variance. If all of the groups are from the same population, these estimates should be nearly identical. The variance within the groups is the standard; if the variance between the groups is no bigger than that within the groups, then there is no reason to believe that the groups are different. If, however, the between-groups variation is larger, it means that the group means are spread more around the grand mean than the individual scores are distributed about their group means, thus indicating that the groups differ by more than random variation and are therefore different. Since this is a test of comparing variances, the F-test, which was presented in connection with the t-test for independent samples, is used.

5.1.1 Assumptions

Before going on to the procedure for the analysis of variance, let us first examine the assumptions inherent in it. These are very similar to what we have seen before.

Assumptions

(i) The samples are independent random samples.

(ii) The populations from which they are drawn are normally distributed.

(iii) The variances of the populations are equal.

(iv) The variable of interest is continuous.

For designs which have two or more factors (treatments) being compared simultaneously, another assumption must be included:

(v) The variances are additive; i.e., no interaction is present if one wishes to test the main effects.

5.1.2 Violations of Assumptions

It is generally accepted that the F-test is fairly robust with respect to these assumptions. Correlated data can be incorporated into the model by a technique known as blocking. Violations of normality do not seriously affect the results unless the data are badly skewed. If the data are skewed, the F (and t) test will produce too many significant results. As the sample size gets larger, the importance of the normality assumption grows less because of the central limit theorem. For small samples, non-normal data can often be transformed in such a way that the normality assumption is satisfied. As with the t-test, the assumption of homogeneous variances is generally considered to be robust as long as the sample size for each group is the same and the difference is not too great, such as one variance being ten times the magnitude of another. Drastic violations of these assumptions affect the test in that it will tend to give too many significant results. As in the case of non-normality, heterogeneity of variances can often be reduced by performing a transformation of the data.

5.1.3 Transformations

A transformation of scale of the data can be performed in cases where expressing the data in terms of another measurement scale will give more validity to the assumptions. Some of the more common transformations are the square root and logarithmic transformations. Both these are monotonic transformations, and thus will leave ordinal relationships the same. The square root transformation is good for count data from a Poisson process in which the mean is equal to the variance. If the mean is positively correlated

with the variance, then the logarithmic transformation will probably be good. This transformation is good for normalizing skewed distributions.

5.1.4 Fixed vs. Random Effects

There are two types of effects which can be studied by analysis of variance techniques: fixed and random effects. One of the assumptions underlying a fixed-effects design is that all levels of the factors about which any inferences are to be made are included in the experiment. In a random effects model, the factor levels (treatments) which are included in the experiment are a random sample from a larger population. In the case of replicating a fixed-effects experiment, the treatments would be exactly the same. In the case of replicating a random-effects experiment, a different set of treatments would be chosen at random every time. Only in a random-effects model can inferences be drawn about the entire population. It is possible that any one experiment can have both fixed and random effects. Such a model is known as a mixed model. One should always be careful in determining which factors in an experiment are fixed and which are random. The calculations are the same in all types of models, but the test of significance which is done at the end will vary with the nature of the model. This will be explained later in the discussions of the various designs.

5.2 TYPES OF DESIGNS

5.2.1 One-Factor ANOVA Design

This is the simplest type of design, and is merely an extension of the t-test for independent samples for testing three or more samples simultaneously. The experiment is performed by randomly assigning the

subjects to groups, then giving a different treatment to each group. The data could be arranged as follows:

		Observations			
	1	Y_{11}	Y_{12}	. . .	Y_{1n}
Treatments	2	Y_{21}	Y_{22}	. . .	Y_{2n}
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	k	$Y_k{}^1$	Y_{k2}	. . .	Y_{kn}

The calculations, which are arranged into an analysis of variance, or ANOVA, table, are shown in Table 1. Notationally, the appearance of a dot as a subscript means that the subscript in whose position it appears has been summed over; thus "$Y_{i.}$" means the i^{th} row summed over j, or merely the sum of all observations in the i^{th} row. Likewise, "$y_{..}$" means both columns and rows have been summed, making $y_{..}$ the grand total of all the observations.

It is the Expected Mean Square (EMS) column which must be examined in order to determine which mean-squares should be compared for the F-test. The two that are divided should be the same except for the treatment effects ($n\Sigma\tau$ /DF for fixed effects, σ^2 for random effects). If the treatments have no effect, then the ratio should be one. In this particular design, the F-ratio is the same for both fixed and random effects; both are compared to error. This will not be the case in any designs comparing more than one factor. Designs which have all fixed effects always compare everything to the error term, but random and mixed models will not. In these cases, the EMS column becomes important because it is the one which will determine the F-ratio to test for different effects.

An example is now in order. Suppose an experimenter wants to determine if there is any difference between four types of food for rats.

TABLE 1

ANOVA for One Factor

Source of Variation	df	Sums of Squares	Mean Squares	Expected Mean Squares		F-ratio
				Fixed	Random	
Between Treatments	$k - 1$	$SS_{Tr} = \sum_{i=1}^{k} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{kn}$	$MS_{Tr} = \frac{SS_{Tr}}{k - 1}$	$\sigma^2 + \frac{n\sum\tau^2}{k - 1}$	$\sigma^2 + n\sigma_\tau^2$	$\frac{MS_{Tr}}{MS_E}$
Error (Within Treatments)	$kn - k$	$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{k} \frac{y_{i.}^2}{n}$	$MS_E = \frac{SS_E}{nk - 1}$	σ^2	σ^2	
Total	$kn - 1$	$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{kn}$				

He takes 32 rats as subjects and randomly divides them into four groups, then assigns a food to each of the different groups. After a designated time, the weight gain of the rats is measured (in grams). The data obtained and the calculations performed are as follows:

		Weight Gain								$Y_{i.}$
	A	10	8	12	4	7	9	14	11	75
Food type	B	2	-3	0	1	0	-2	-2	4	0
	C	7	4	5	2	8	9	6	5	46
	D	18	15	22	21	15	7	17	20	135
										256

$Y_{..} = 256$

$\Sigma\Sigma y_{ij}^2 = 3546$

$$SS_{Tr} = 1/8\,[75^2 + 46^2 + 135^2] - \frac{256^2}{32} = 1197.75$$

$$SS_T = 3546 - \frac{256^2}{32} = 1498$$

$$SS_E = 1498 - 1197.75 = 300.25$$

Source of Variation	DF	Sums of Squares	Mean Square	EMS	F-ratio
Between Foods	3	1197.75	399.25	$\sigma^2 + \frac{8\Sigma\tau_i^2}{3}$	37.232
Error	28	300.25	10.72	σ^2	
Total	31	1498			

This is a fixed-effects experiment because every food that the investigator was interested in was included in the experiment. The obtained F value should be compared to the table value of F with 3 and 28 DF. This value is 7.19 for $\alpha = 0.001$, so this result is highly significant. Thus, we conclude that the means are not all identical. The F-test tells us that at least one of the means is different, but it does not tell us which ones differ from the others. To do this, some type of multiple comparison test must be applied. There are

many such tests available; several of the more common tests will be presented here.

5.2.1.1 Fisher's Least Significant Difference (LSD) Method

This test is to be applied only if the F-test shows significance, and it consists basically of applying the ordinary Student's t-test to many pairs of means. If any two means differ by more than the LSD, where

$$LSD = t_{[1-\alpha/2;\ \text{error DF}]}\sqrt{MSE(1/n_1 + 1/n_2)}$$

then those two means will be different. For our example, at $\alpha = 0.01$,

$$LSD = t_{[0.995,\ 28]}\sqrt{(10.72)(1/8 + 1/8)} = 2.7163(1.637) = 4.524.$$

In this example, the means are

B	C	A	D
0	5.75	9.375	16.875

The difference between C and A is only 3.625, so we conclude that there is no difference between C and A. All of the others differ by more than 4.524, so they are all different. This can be represented graphically as

B <u>C A</u> D.

5.2.1.2 Tukey's Honestly Significant Difference (HSD) Method

The HSD method is identical to the LSD method, except that it requires equal sample sizes. The HSD is given by

$$HSD = q_{[\alpha;\ k;\ \text{error DF}]}\sqrt{MSE/n}$$

where q is the table value of the studentised range and k is the number of means being compared. This table can be found in many design books.

In our example, for α = 0.01,

HSD = $4.83\sqrt{10.72/8}$ = 5.592.

Once again C and A are the only ones which differ by more than 5.592, so the same conclusion is reached with this method as with the Fisher's LSD Method.

5.2.1.3 Duncan's Multiple Range Test

The Duncan's Multiple Range Test differs from the Fisher and Tukey tests in that it gives a different range for different means. Instead of giving one number against which all differences in means are tested, this test gives larger intervals for means that have other means in between them.

To perform this test, determine numbers $r_{[\alpha,\ p,\ \text{error DF}]}$ for p = 2,3...,k from a table of Duncan's significant ranges, and multiply each of these numbers by $\sqrt{MSE/n}$. These will be the least significant ranges. Then rank the means. In comparing them, if they are next to each other, use the p = 2 range; if there is one other mean in between them, use p = 3, and so on.

For our example, we need values for p = 2, 3, and 4. At α = 0.01, these values are:

p =		$p\sqrt{MSE/n}$:	
2:	3.93	2:	4.550
3:	4.19	3:	4.851
4:	4.29	4:	4.967

Means:	B	C	A	D
(Ranked)	0	5.75	9.375	16.875

Applying this method, the means BC, CA, and AD would have to differ by 4.550 to be significantly different; DA and CD would have to differ by 4.851, and BD would have to differ by 4.967. As with the previous methods, C and A are the same; all others are different. Thus, as the final results of the experiment, we conclude that Food B causes the least weight gain, Food D causes the most, and types A anc C, while different from both B and D, are indistinguishable from each other.

5.2.2 Two-Factor ANOVA Design

This design is the simplest type of factorial design, that is, one in which two or more factors are being compared, and all combinations of the levels of these factors are run during the experiment. The principles behind this design are the same as those of the One-Factor Design, except now the variability is broken down into more pieces: that for Factor A, Factor B, the AB interaction, and the Error.

The data layout for this design with more than one observation per cell can be presented as:

		Factor B			
		1	2	. . .	b
Factor A	1	$Y_{111}, Y_{112}, \ldots, Y_{11n}$	$Y_{121}, Y_{122}, \ldots, Y_{12n}$		$Y_{1b1}, Y_{1b2}, \ldots, Y_{1bn}$
	2	$Y_{211}, Y_{212}, \ldots, Y_{21n}$	$Y_{221}, Y_{222}, \ldots, Y_{22n}$		$Y_{2b1}, Y_{2b2}, \ldots, Y_{2bn}$
	.				
	.				
	.				
	a	$Y_{a11}, Y_{a12}, \ldots, Y_{a1n}$	$Y_{a21}, Y_{a22}, \ldots, Y_{a2n}$		$Y_{ab1}, Y_{ab2}, \ldots, Y_{abn}$

The calculations for the analysis of this type design are presented in Table 2.

TABLE 2

ANOVA for Two Factors

Source of Variation	df	Sums of Squares	Mean Squares	Expected Mean Squares		F-ratio	
				Fixed	Random	Fixed	Random
Factor A Treatments	a - 1	$SS_A = \sum_{i=1}^{a} \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{abn}$	$\frac{SS_A}{a-1}$	$\sigma^2 + \frac{bn\sum \tau_i^2}{a-1}$	$\sigma^2 + n\sigma_{\tau\beta}^2 + bn\sigma_\tau^2$	$\frac{MS_A}{MS_E}$	$\frac{MS_A}{MS_{AB}}$
Factor B Treatments	b - 1	$SS_B = \sum_{j=1}^{b} \frac{y_{.j.}^2}{an} - \frac{y_{...}^2}{abn}$	$\frac{SS_B}{b-1}$	$\sigma^2 + \frac{an\sum \beta_j^2}{b-1}$	$\sigma^2 + n\sigma_{\tau\beta}^2 + an\sigma_\beta^2$	$\frac{MS_B}{MS_E}$	$\frac{MS_B}{MS_{AB}}$
Interaction	(a - 1)(b - 1)	$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij.}^2}{n} - \frac{y_{...}^2}{abn} - SS_A - SS_B$	$\frac{SS_{AB}}{(a-1)(b-1)}$	$\sigma^2 + \frac{n\sum\sum(\tau\beta)_{ij}^2}{(a-1)(b-1)}$	$\sigma^2 + n\sigma_{\tau\beta}^2$	$\frac{MS_{AB}}{MS_E}$	$\frac{MS_{AB}}{MS_E}$
Error	ab(n - 1)	$SS_E = SS_T - SS_A - SS_B - SS_{AB}$	$\frac{SS_E}{ab(n-1)}$	σ^2	σ^2		
Total	abn - 1	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{abn}$					

As an example suppose a chemical process is being studied, where the factors are temperature and pressure, with three levels each. The experiments are performed in random order. The data obtained and calculations performed are as follows:

		Pressure			
		Low	Medium	High	$Y_{i..}$
Temperature	Low	90	86	79	515
		89	88	83	
	Medium	85	82	86	506
		81	87	85	
	High	53	77	101	468
		60	84	93	
	$Y_{.j.}$	458	504	527	1489

$$SS_T = 125235 - (1489)^2/18 = 2061.6111$$
$$SS_A = [515^2 + 506^2 + 468^2]/6 - (1489)^2/18 = 207.44$$
$$SS_B = [458^2 + 504^2 + 527^2]/6 - (1489)^2/18 = 414.44$$
$$SS_{AB} = (250245)/2 - (1489)^2/18 - 207.44 - 414.44 = 1330.2222$$
$$SS_E = SS_T - SS_A - SS_{AB} - SS_B = 112.5$$

Source of Variation	DF	Sums of Squares	Mean Squares	F
A	2	207.444	103.722	
B	2	414.444	205.722	
AB	4	1330.222	332.556	26.604
Error	9	112.5	12.5	
Total	17	2061.611		

$F_{[0.01;\ 4,\ 9]} = 14.7$

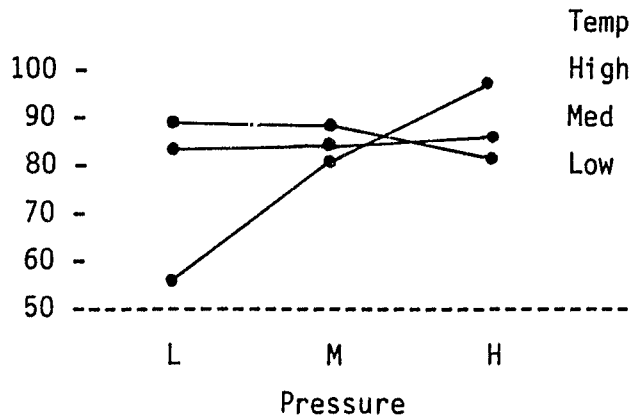
Since 26.604 > 14.7, we conclude that the interaction between temperature and pressure in this experiment is highly significant. This means

that there is a synergistic effect between temperature and pressure; i.e., they do not function independently. This can be demonstrated graphically if the cell means for pressure are plotted as a function of temperature, as indicated in Figure 2.

Figure 2

Cell Means

		Pressure		
		L	M	H
Temp.	L	89.5	87	81
	M	83	84.5	85.5
	H	56.5	80.5	97



If there were no interaction between temperature and pressure, the three figures would mirror each other, with only a difference in location. As it is, they are drastically different, so the main effects, temperature and pressure, should not be considered separately. It would not be accurate to say that high temperature produces the highest yield, because it also produces the lowest, depending on the pressure.

Because of the importance of the interaction term, it should always be tested first; if it is significant, it is often the last test to be done, because it is difficult to interpret the main effects when there is an interaction. In a K-Factor Design, the K^{th} order interaction should be tested first, then the lower order interactions, in the decreasing order of their complexity. The main effects should always be tested last and interpreted carefully if the interactions are significant.

In this example, the significance of the interaction makes testing the main effects lose its meaning. Obviously, both tests would be highly significant, implying there is a difference in yields between high, medium and low levels of temperature and of pressure. The exact nature of these differences, however, is uninterpretable without considering one variable in relation to the other. Depending on the purpose of the experiment, this may or may not be satisfactory. In this experiment, it probably does not matter to the investigator that there is an interaction because he is only interested in determining how he can get the greatest yield. He can easily determine this by doing multiple comparison tests; the only effect of the interaction is that each combination must be considered separately, rather than comparing the means of temperatures and then the means of pressure.

For performing the Fisher's LSD test at $\alpha = 0.01$,

$$LSD = t_{[0.995,\ 9]}\sqrt{12.5(1/2+1/2)} = (3.250)(3.535) = 11.490.$$

Thus any means differing by more than 11.490 are significantly different. The results are

HL	HM	LH	ML	MM	MH	LM	LL	HH
56.5	80.5	81	83	84.5	85.5	87	89.5	97

Therefore, to maximize the yield, one should use high temperature and high pressure, low temperature and low pressure, or low temperature and medium pressure. Minimum yield is obtained with high temperature and low pressure.

While this experiment worked out nicely, interactions can sometimes cause problems. For example, in working with space flight, an investigator

might be interested in determining whether or not weightlessness has an effect on some physiological parameter in man. He might have data from three flights (Skylab, for example), so he could analyze it as a two-way design, with the physiological parameter as one factor and the flight as another. A significant interaction in such an experiment can be annoying, but does not always preclude testing and interpreting the main effects. If this occurs, it would be informative to graph the means similar to those in Figure 2. If all lines increase or decrease with one or more being steepter than the others, the interaction between the two factors, i.e., the physiological parameter and the flight, may be significant. Clearly, however, the main effect may also be significant, leading to the interpretation such as: "the physiologic parameter, blood volume, decreased with exposure to weightlessness and this effect was significantly greater on the last flight."

This space-flight example leads us to the next type of design to be discussed. The differences in response on the various flights may have been caused by some extraneous factor not considered in the experiment, such as dietary changes. If the crew of the second flight, for instance, had different diets from the others, then the responses of those individuals will be correlated with each other, but not with the other crews. This not only introduces extra variability, but also defies the assumptions of independence and randomization. This can be taken care of by employing a technique known as blocking.

5.2.3 Randomized Complete Block Design

Blocking designs are the ones which correspond to the two-sample tests for related measures, and are thus the methods used for handling repeated measures. Any time there is reason to believe that particular groups

of measurements will be correlated, these groups should be separated into blocks. When this is done, an additional assumption is made: the correlations within blocks are equal. The blocking techniques originated in agriculture, where different plots of land would be blocked, because the experimenters knew that different soil conditions would lead to different yields, and the yields from the same conditions would be correlated. In repeated measures experiments, where each subject serves as his own control, each individual subject is considered to be a block. The effect of this is that the variability due to differences in the average responses of the subjects will be removed from the experimental error, thus making the test more sensitive.

In performing a randomized block experiment, the order of the treatments within the blocks should be randomized, once the blocks are determined. When the blocks are subjects, care should be taken that there are no carry-over effects between the treatments. Each experiment on the individual should be independent of the others, and if they are not, then the results will be invalid. To be a complete block, every treatment should be performed in every block.

The analysis of a blocked experiment is very similar to that of a multi-factor independent design. In the calculations, the blocks are treated as an additional factor except that no interactions are computed for blocks. (Some books do compute the block X factor interaction terms, but generally such interactions are assumed to be part of the error.)

The data layout for a one-factor randomized complete block design looks exactly like that of the two-factor randomized complete design except that instead of Factor B, we have blocks. The calculations for this design are presented in Table 3. Notice that the calculations are exactly the same as those of the completely randomized design except for the lack of an

TABLE 3

ANOVA for One Factor: Randomized Complete Block Design

Source of Variation	df	Sums of Squares	Mean Squares	Expected Mean Squares		F-ratio	
				Fixed	Random	Fixed	Random
Treatments	$a - 1$	$SS_{Tr} = \sum_{i=1}^{a} \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{abn}$	$\frac{SS_{Tr}}{a - 1}$	$\sigma^2 + \frac{bn\sum \tau_i^2}{a - 1}$	$\sigma^2 + bn\sigma_\tau^2$	$\frac{MS_{Tr}}{MS_E}$	$\frac{MS_{Tr}}{MS_E}$
Blocks	$b - 1$	$SS_B = \sum_{j=1}^{b} \frac{y_{.j.}^2}{an} - \frac{y_{...}^2}{abn}$	$\frac{SS_B}{b - 1}$	$\sigma^2 + \frac{an\sum \beta_j^2}{b - 1}$	$\sigma^2 + an\sigma_\beta^2$		
Error	$abn - a - b + 1$	$SS_E = SS_T - SS_{Tr} - SS_B$	$\frac{SS_E}{abn - a - b + 1}$	σ^2	σ^2		
Total	$abn - 1$	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{abn}$					

interaction term, and the error DF is adjusted accordingly. Also, there is only one F-test to be performed because we are only interested in whether or not the treatments had an effect. The blocks could, of course, be tested for significance. If they are not significant, it will mean that there was no need for the blocking in the first place. Because this is a one-factor design, the F-ratio is the same for both fixed and random models.

As an example, suppose an experiment is being done to test the effects of five different drugs. Four individuals are used as subjects, so each individual will be treated as a block. The order that the treatments are given to each individual is randomized, and sufficient time is given between treatments to ensure that there are no carry-over effects. There will be only one observation per cell so that the corresponding nonparametric test can be run on the same data. The data obtained and calculations performed are as follows:

		Person				
		1	2	3	4	Total
Drug	A	12	14	12	13	51
	B	9	13	8	10	40
	C	27	32	22	29	110
	D	8	22	9	11	50
	E	14	29	11	16	70
		70	110	62	79	321

$$\sum\sum Y_{ij}^2 = 6309$$

$$SS_{Tr} = [51^2 + 40^2 + 110^2 + 50^2 + 70^2]/4 - (321)^2/20 = 773.2$$

$$SS_B = [70^2 + 110^2 + 62^2 + 79^2]/5 - (321)^2/20 = 264.95$$

$$SS_T = 6309 - (321)^2/20 = 1156.95$$

$$SS_E = 1156.95 - 264.95 - 773.2 = 118.8$$

Source of Variation	DF	Sums of Squares	Mean Square	F-ratio	F-table $\alpha = .01$
Treatments	4	773.20	193.3	19.52	5.41
Blocks	3	264.95	88.316		
Error	12	118.80	9.9		
Total	19	1156.95			

Since the F-ratio is significant, we conclude that there is a difference between the drugs. To determine which ones are different, we will use Fisher's LSD. If any two means differ by more than

$$LSD = t_{[0.995,\ 12]}\sqrt{9.9(1/4 + 1/4)} = 3.055(2.225) = 6.797$$

then they will be significantly different at the $\alpha = 0.01$ level of significance. This gives the results

B	D	A	E	C
10	12.5	12.75	17.5	27.5

Therefore we conclude that drugs B, D, and A are indistinguishable, as are drugs D, A, and E; the former set gives the lowest response. Drug C shows a significantly higher response than any of the other drugs.

5.2.4 Latin Square Design

The randomized block design is applicable to any number of factors, but only for one set of blocks. Sometimes situations arise in which it is necessary to block in two directions at the same time. A simple example of this is the case of comparing different brands of tires. Suppose there are four brands to test, and it is decided to use four tires of each. Rather than using sixteen cars, the cars can be blocked and only four cars will be needed, with one of each brand of tire on each car. If the tires are randomly assigned to positions on each car, this will be a randomized complete block design. However, it is also known that tires wear differently, depending on their positions on the car, and that like positions will be correlated. Therefore, position can be blocked as well by putting four brands of tires in four different positions. This type of design, where two things are being blocked at the same time, is called a Latin Square design.

An experiment for comparing p treatments, being blocked in two directions, can be arranged into a p x p Latin Square with the rows being one set of blocks and the columns being the other. The key is that each treatment must appear once in each row and once in each column so that every combination of levels of blocks is performed. Because of the restrictions on the placements of the treatments, the randomization is lost in this type of design. However, there are different possible patterns for each size Latin Square, so one of these should be chosen at random.

The following is an example of a 5 x 5 Latin Square, with treatments denoted by A, B, C, D, and E:

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

The calculations for the Latin Square are given in Table 4.

As an example, suppose an experiment is conducted comparing the reaction times of five different catalysts on a chemical process, where only five experiments can be run per day and each batch of materials will permit only five runs. An arrangement different from the one above was utilized. The results obtained and calculations performed are as follows:

		Batch						
		1	2	3	4	5	$Y_{i..}$	
	1	A=10	B=9	D=3	C=9	E=5	36	A = 52
	2	C=13	E=4	A=9	D=5	B=10	41	B = 38
Day	3	B=6	A=11	C=12	E=3	D=7	39	C = 54
	4	D=8	C=10	E=8	B=8	A=12	46	D = 27
	5	E=6	D=4	B=5	A=10	C=10	35	E = 26
$Y_{.j.}$		43	38	37	35	44	197	

$$\Sigma\Sigma\Sigma Y_{ijk}^2 = 1759$$

$SS_{Tr} = [52^2 + 38^2 + 54^2 + 27^2 + 26^2]/5 - (197)^2/25 = 141.44$
$SS_{Rows} = [36^2 + 41^2 + 39^2 + 46^2 + 35^2]/5 - (197)^2/25 = 15.44$
$SS_{Columns} = [43^2 + 38^2 + 37^2 + 35^2 + 44^2\dagger/5 - (197)2/25 = 12.24$
$SS_T = 1759 - (197)^2/25 = 206.64$
$SS_E = 206.64 - 141.44 - 15.44 - 12.24 = 37.52$

Source of Variation	DF	SS	MS	F
Treatments	4	141.44	35.36	11.3092
Rows	4	15.44	3.86	
Columns	4	12.24	3.06	
Error	12	30.52	3.1267	
Total	24	206.64		

TABLE 4

ANOVA for p X p Latin Square Design

Source of Variation	df	Sums of Squares	Mean Square	Expected Mean Square		F-ratio	
				Fixed	Random	Fixed	Random
Treatments	$p - 1$	$SS_{Tr} = \sum_{k=1}^{p} \frac{y_{..k}^2}{p} - \frac{y_{...}^2}{p^2}$	$\frac{SS_{Tr}}{p - 1}$	$\sigma^2 + \frac{p^2 \sum \tau_i^2}{p - 1}$	$\sigma^2 + p^2\sigma_\tau^2$	$\frac{MS_{Tr}}{MS_E}$	$\frac{MS_{Tr}}{MS_E}$
Rows	$p - 1$	$SS_R = \sum_{i=1}^{p} \frac{y_{i..}^2}{p} - \frac{y_{...}^2}{p^2}$	$\frac{SS_R}{p - 1}$	$\sigma^2 + \frac{p^2 \sum \alpha_i^2}{p - 1}$	$\sigma^2 + p^2\sigma_\alpha^2$		
Columns	$p - 1$	$SS_C = \sum_{j=1}^{p} \frac{y_{.j.}^2}{p} - \frac{y_{...}^2}{p^2}$	$\frac{SS_C}{p - 1}$	$\sigma^2 + \frac{p^2 \sum \beta_j^2}{p - 1}$	$\sigma^2 + p^2\sigma_\beta^2$		
Error	$(p - 2)(p - 1)$	$SS_E = SS_T - SS_{Tr} - SS_R - SS_C$	$\frac{SS_E}{(p - 2)(p - 1)}$	σ^2	σ^2		
Total	$p^2 - 1$	$SS_T = \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} y_{ijk}^2 - \frac{y_{...}^2}{p^2}$	i = rows j = columns k = treatment (A, B, C, etc.)				

Comparing this F-value to the table value $F_{[0.01;\ 4,12]} = 5.41$, we conclude that there is a difference between the means. Applying Fisher's LSD test to determine which means are different, we see that at the $\alpha = 0.01$ level,

$$LSD = t_{[0.995,12]}\sqrt{3.1267\ (1/5 + 1/5)} = 3.055(1.118) = 3.4165.$$

Therefore, any two means differing by more than 3.4165 are significantly different. Calculating and ordering the means, we obtain the following results:

E	D	B	A	C
5.2	5.4	7.6	10.4	10.8

Thus, catalysts E, D, and B are indistinguishable, and catalysts B, A, and C are indistinguishable, with the latter set yielding the higher results.

The analysis of the Latin Square design, with two sets of blocking, is an extension of the analysis of the one-factor randomized block design. This can be extended even further for blocking in more than two directions. A three-way blocking design, for example, is called a Graeco-Latin Square and is set up and analysed in exactly the same manner except that now each treatment appears once in each row, once in each column, and once paired with each Greek letter representing the third block. The calculations or the sums of squares for the third block follow the same pattern as that of the others. Notice that, since Latin Square designs are one-factor designs, the F-ratio is the same for both fixed and random models.

5.2.5 Nested or Heirarchical Designs

Another situation which can occur in experimentation is the case where the levels of one factor are similar but not identical for levels of

another factor. For example, suppose it is desired to measure the quality of a chemical made by two different suppliers. The samples from each supplier are from different batches made by different chemists, so these need to be factors in the experiment. However, the chemicals made by the second chemist for the first supplier cannot be grouped with those of the second chemist for the second supplier, for obvious reasons; they are not on the same level. Chemists are nested within suppliers. Furthermore, suppose that each chemist uses different sources for materials in each batch. Then the batches cannot be considered to be identical, and will be nested within chemists. This is an example of a three-stage nested design. The data layout for this type design can be represented by the diagram below. The calculations for the analysis are presented in Table 5.

	Factor A_1			Factor A_2			Factor A_a		
*F B	1	2	...b	1	2	...b	1	2	...b
*F C_1	Y_{111}	Y_{121}	Y_{1b1}	Y_{211}	Y_{221}	Y_{2b1}	Y_{a11}	Y_{a21}	Y_{ab1}
	Y_{11n}	Y_{12n}	Y_{1bn}	Y_{21n}	Y_{22n}	Y_{2bn}	Y_{a1n}	Y_{a2n}	Y_{abn}
*F C_2	Y_{112}	Y_{122}	Y_{1b2}	Y_{212}	Y_{222}	Y_{2b2}	Y_{a12}	Y_{a22}	Y_{ab2}
	Y_{112n}	Y_{122n}	Y_{1b2n}	Y_{212n}	Y_{222n}	Y_{2b2n}	Y_{a12n}	Y_{a22n}	Y_{ab2n}
*F C_c	Y_{11c1}	Y_{12c1}	Y_{1bc1}	Y_{21c1}	Y_{22c1}	Y_{2bc1}	Y_{a1c1}	Y_{a2c1}	Y_{abc1}
	Y_{11cn}	Y_{12cn}	Y_{1bcn}	Y_{21cn}	Y_{22cn}	Y_{2bcn}	Y_{a1cn}	Y_{a2cn}	Y_{abcn}

*F = Factor

TABLE 5

ANOVA for Three-Stage Nested Design

Source of Variation	df	Sums of Squares	Mean Squares	Expected Mean Squares		F-ratio	
				Fixed	Random	Fixed	Random
A	$a - 1$	$SS_A = \sum_{i=1}^{a} \frac{y_{i\ldots}^2}{bcn} - \frac{y_{\ldots}^2}{abcn}$	$\frac{SS_A}{a - 1}$	$\sigma^2 + \frac{bcn\sum\tau_i^2}{a - 1}$	$\sigma^2 + n\sigma_\gamma^2 + cn\sigma_\beta^2 + bcn\sigma_\tau^2$	$\frac{MS_A}{MS_E}$	$\frac{MS_A}{MS_B}$
B(within A)	$a(b - 1)$	$SS_B = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij\cdot}^2}{cn} - \sum_{i=1}^{a} \frac{y_{i\ldots}^2}{bcn}$	$\frac{SS_B}{a(b - 1)}$	$\sigma^2 + \frac{cn\sum\sum\beta_{j(i)}^2}{a(b - 1)}$	$\sigma^2 + n\sigma_\gamma^2 + cn\sigma_\beta^2$	$\frac{MS_B}{MS_E}$	$\frac{MS_B}{MS_C}$
C(within B)	$ab(c - 1)$	$SS_C = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} \frac{y_{ijk\cdot}^2}{n} - \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij\cdot\cdot}^2}{cn}$	$\frac{SS_C}{ab(c - 1)}$	$\sigma^2 + \frac{n\sum\sum\sum\gamma_{k(ij)}^2}{ab(c - 1)}$	$\sigma^2 + n\sigma_\gamma^2$	$\frac{MS_C}{MS_E}$	$\frac{MS_C}{MS_E}$
Error	$abc(n - 1)$	$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n} y_{ijkl}^2 - \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} \frac{y_{ijk\cdot}^2}{n}$	$\frac{SS_E}{abc(n - 1)}$	σ^2	σ^2		
Total	$abcn - 1$	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n} y_{ijkl}^2 - \frac{y_{\ldots\cdot}^2}{abcn}$					

As an example, suppose the experiment comparing chemical suppliers described above was performed. The data collected and calculations done are as follows:

	Supplier 1				Supplier 2				$Y_{ijk.}$: Supplier 1		Supplier 2	
Chemist	1	2	3	4	1	2	3	4				
Batch 1	17	18	23	20	21	22	23	19	36	37	47	38
	19	19	24	18	23	20	20	18	44	42	43	37
Batch 2	25	24	25	22	24	25	19	23	47	47	52	42
	22	23	27	20	23	22	18	25	47	47	37	48
Batch 3	20	19	22	21	22	21	19	23	36	36	42	39
	16	17	20	18	20	19	22	22	42	40	41	45
Chemist Totals $Y_{ij..}$	119	120	141	119	133	129	121	130	1012			
Supplier Totals $Y_{i...}$	499				513							

$$\sum Y_{i....}^2/bcn = (499^2 + 513^2]/24 = 21340.417$$

$$\sum\sum Y_{ij..}/cn = (119^2 + 120^2 + \ldots + 130^2]/6 = 21412.333$$

$$\sum\sum\sum Y_{ijk.}/n = 1/2(36^2 + 37^2 + \ldots + 41^2 + 45^2) = 21580$$

$$\sum\sum\sum\sum Y_{ijkl}^2 = 21636$$

$$Y....^2/abcn = (1012)^2/48 = 21336.333$$

$SS_A = 21340.417 - 21336.333 = 4.083334$

$SS_{B(A)} = 21412.333 - 21340.417 = 71.916666$

$SS_{C(B)} = 21580 - 21412.333 = 167.66667$

$SS_E = 21636 - 21580 = 56$

Source of Variation	DF	SS	MS	F-ratio (fixed)	F-table α=.05	F-table α=.01	F-ratio (random)	F-table α=.05	F-table α=.01
A	1	4.0833	4.0833	1.7500	4.26	7.82	.3406	5.99	13.75
B(A)	6	71.9167	11.9861	5.1369	2.51	3.67	1.1438	2.74	4.20
C(B)	16	167.6667	10.4792	4.4911	2.31	2.86	4.4911	2.13	2.86
Error	24	56.0	2.3333						
Total	47	299.6667							

This example does a good job of showing the difference between fixed and random models. For a fixed model, there is a significant difference between both chemists and batches at the $\alpha = 0.01$ level. For the random model, however, only the batches are significantly different; chemists are not significantly different even at the $\alpha = 0.05$ level. The suppliers are indistinguishable in either case. To determine whether the effects are fixed or random, one must determine how they were chosen. It seems reasonable to assume that the two suppliers are the only ones of interest. Thus, Factor A is fixed. If the chemists used are the only ones whose work we are interested in, then Factor B is also fixed. If they were chosen as a random sample of many chemists, then Factor B is random. The same is true of the batches. Since it is more likely that batches were chosen at random, Factor C is probably a random factor. Thus if A and B are fixed and C is random, this is a mixed model, and the EMS for it is not included in Table 5. The EMS and corresponding F-ratios for a mixed model of this type are:

Factor	EMS	F-ratio	
A	$\sigma^2 + n\sigma_\gamma^2 + \frac{bn\sum \tau_i^2}{a-1}$	MS_A/MS_C	0.3897
B	$\sigma^2 + n\sigma_\gamma^2 + \frac{cn\sum \beta_{j(i)}^2}{a(b-1)}$	MS_B/MS_C	1.1438
C	$\sigma^2 + n\sigma_\gamma^2$	MS_C/MS_E	4.4911
Error	σ^2		

In this particular case, the results of the mixed and random models do not differ in significance, but it is possible that they could in some instances. It is models like these, with many factors and combinations of fixed and random effects, which make evident the importance of the Expected Mean Squares. Without them, it would be impossible to know which mean squares to divide to test the effects of a particular factor.

5.2.6 Summary of Analysis of Variance

This presentation of five different types of experimental designs is by no means complete. Each of these designs can be extended to include more factors. In addition, there are variations which have not been discussed. For example, all of the designs presented have been assumed to have equal sample sizes in each cell; that is, they are balanced designs. This is not necessary as long as the assumptions of normality and homoscedasticity are met. In fact, the computational formulas are generally the same for both balanced and unbalanced designs. However, having unequal cell numbers increases the complexity of the calculations immensely as the designs become more complex because the sample size cannot be factored out, as has been done in all of the calculations presented here. Many statistical packages are not set up to handle unequal samples sizes in complex designs.

Another possibility that has not been discussed is that of incomplete designs. These are designs in which not all of the treatment combinations are performed. This is most likely to occur in blocking designs when the blocks are not large enough to hold all of the treatments. Needless to say, such an occurrence adds complexity to the calculations and since it is an uncommon situation, the analysis will not be presented here. Techniques for analyzing incomplete designs can be found in most intermediate level design books (see bibliography).

Another concept which should be mentioned but will not be discussed is confounding and fractional replication. In factorial experiments where there are many factors, it is often desirable to run fewer experiments per block than there are treatment combinations. These experiments can be designed such that the effects of certain combinations are indistinguishable from others--that is, they are confounded--and therefore only one of these combinations needs to be performed to know about all of them. Fractional replication of a factorial design means running only a fraction of the total number of runs. Since one can determine in advance which combinations are to be confounded, the experimenter has a lot of control over such a situation and can obtain meaningful results with considerably fewer experiments. These techniques are not difficult, but they can be very involved and would require more explanation than can be given here. Once again, the reader is referred to the bibliography for further information.

5.3 NONPARAMETRIC ALTERNATIVES

As with the case of the two-sample tests, there are nonparametric tests available which will handle the one-factor experimental designs. These tests will be better than the parametric tests if the data is badly non-normal. They also can be used for ordinal data.

5.3.1 One-Factor Design: The Kruskal-Wallis Test

Assumptions

(i) All samples are random samples.

(ii) There is independence both within each sample and between the various samples.

(iii) The measurement scale is at least ordinal.

(iv) If the populations differ, they differ only in location. (Note: This is equivalent to the normal assumption of homogeneous variances.)

The Kruskal-Wallis test is another ranking method. The first step in the procedure is to rank the totality of the observations from all K samples from one to N, where $N = \sum n_i$. In the case of ties, the average of the ranks that would have been assigned to those values is assigned to all of them. Then the sum of the ranks for each sample, that is

$$R_i = \sum_{j=1}^{n_i} R(X_{ij}),\ i = 1,\ldots,K, \text{ is computed.}$$

The test statistic is:

$$T = \frac{1}{S^2}\sum_{i=1}^{K}\frac{R_i^2}{n_i} - \frac{N(N+1)}{4} \quad \text{where} \quad S^2 = \frac{1}{N-1}\sum_{\substack{\text{all}\\ \text{ranks}}} R(X_{ij})^2 - \frac{N(N+1)^2}{4}$$

If there are no ties, this simplifies to:

$$T = \frac{12}{N(N+1)}\sum_{i=1}^{K}\frac{R_i^2}{n_i} - 3(N+1)$$

This should be compared to the appropriate quantile of a Chi-Square distribution with (K - 1) DF. If the null hypothesis of no difference is rejected, then a multiple comparison test can be made. The most common, which is simply Fisher's LSD method applied to ranks, tells that two populations i and j are significantly different if the following is true:

$$\left|\frac{Ri}{n_i} - \frac{Rj}{n_j}\right| > t_{[1-\alpha/2;\ N-K]}\left(S^2\,\frac{N-1-T}{N-K}\right)^{1/2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}$$

The Kruskal-Wallis test will now be applied to the data analysed by the one-factor analysis of variance design.

A	R(A)	B	R(B)	C	R(C)	D	R(D)	
10	22	2	7.5	7	16	18	29	
8	18.5	-3	1	4	10	15	26.5	
12	24	0	4.5	5	12.5	22	32	
4	10	1	6	2	7.5	21	31	
7	16	0	4.5	8	18.5	15	26.5	
9	20.5	-2	2.5	9	20.5	7	15	
14	25	-2	2.5	6	14	17	28	$\Sigma\Sigma R(X_{ij})^2 = 11432.5$
11	23	4	10	5	12.5	20	30	
	159		38.5		111.5		219	

$$S^2 = (11432.5 - \frac{32(33)^2}{4})/31 = 87.758065$$

$$T = (1/87.758)\left[(159^2 + 38.5^2 + 111.5^2 + 219^2)/8 - \frac{32(33)^2}{4}\right] = 24.870221$$

Since $\chi^2_{[0.001;\ 3]} = 16.27$, it is highly significant and we conclude that the populations yield different values.

To determine which means are different at $\alpha = 0.001$, we determine the value

LSD = 3.674(2.1916) = 8.0519.

Thus any two means of ranks that differ by more than 8.0519 will be different. The results thus are

B	C	A	D
4.8125	13.9375	19.875	27.375

This test cannot distinguish between C and A nor between A and D. This is not as sensitive as the analysis of variance, which was able to detect a difference between food types A and D.

The A.R.E. of the Kruskal-Wallis test relative to the F-test in the analysis of variance is never less than 0.864, but it may be as high as infinity for extremely non-normal data. For normal populations, the A.R.E. is 0.955, and for uniform data it is 1. Relative to the median test, the A.R.E. of the Kruskal-Wallis test is 1.5 for normal data and 3.0 for uniform data.

5.3.2 Randomized Complete Block Design: The Quade Test

Assumptions

(i) The results within each block are independent of the results of other blocks.

(ii) Observations may be ranked within blocks.

(iii) The sample range may be determined within each block so that the blocks can be ranked.

This test, which is an extension of the Wilcoxon Signed-Ranks test, requires that an equal sample size k be taken in all b blocks. To perform the test, first rank the observations within each block from 1 to k, using average ranks in case of ties. Then go back to the original observations and obtain the sample ranges within each block, that is, the difference between the smallest and largest values, and then rank the blocks from 1 to b by their ranges. Let Q_i denote the rank of the the i^{th} block. For each X_{ij}, form its corresponding value S_{ij}, where

$$S_{ij} = Q_j \, [R(X_{ij}) - (K + 1)/2]$$

Finally, calculate the sum for each treatment, that is,

$$S_i = \sum_j S_{ij}.$$

The test statistic is

$$T_1 = \frac{B_1(b-1)}{A_1 - B_1}$$

where

$$A_1 = \sum_{i=1}^{K} \sum_{j=1}^{b} S_{ij} \quad \text{and} \quad B_1 = \frac{1}{b} \sum_{i=1}^{K} S_i^2$$

This statistic should be compared to the proper quantile of an F distribution with (K - 1) and (b - 1)(k - 1) DF.

If this test shows significance, multiple comparisons can be made. Two populations i and j will be considered significantly different if

$$S_i - S_j > t_{[1-\alpha/2;\ (b-1)(K-1)]} \left[\frac{2b(A_1-B_1)}{(b-1)(k-1)} \right]^{\frac{1}{2}}$$

As an example, we will run the Quade Test on the same set of data that was used for the randomized complete block design. The calculations are as follows:

		Persons (Blocks)							
		X_{i1}	$R(X_{i1})$	X_{i2}	$R(X_{i2})$	X_{i3}	$R(X_{i3})$	X_{i4}	$R(X_{i4})$
	A	12	3	14	2	12	4	13	3
	B	9	2	13	1	8	1	10	1
Drug	C	27	5	32	5	22	5	29	5
	D	8	1	22	3	9	2	11	2
	E	14	4	29	4	11	3	16	4
		Range = 19		Range = 18		Range = 14		Range = 19	
		$Q_1 = 3.5$		$Q_2 = 2$		$Q_3 = 1$		$Q_4 = 3.5$	

Now compute the S_{ij}'s:

$S_{11} = (3.5)(3 - (5 + 1)/2) = 0$

$S_{21} = (3.5)(2 - 3) = -3.5$, etc.

		Persons (Blocks)				
		1	2	3	4	S_i
	A	0	-2	1	0	-1
	B	-3.5	-4	-2	-7	-16.5
Drug	C	7	4	2	7	20
	D	-7	0	-1	-3.5	-11.5
	E	3.5	2	0	3.5	9

$A_1 = 0^2 + (-2)^2 + 1^2 + \ldots + 2^2 + 0^2 + 3.5^2 = 295$

$B_1 = (-1^2 + (-16.5)^2 + 20^2 + (-11.5)^2 + 9^2)/4 = 221.625$

$T_1 = \frac{3(221.625)}{295-221.625} = 9.061$

$F_{[0.01;\ 4,12]} = 5.41$

Since $T_1 = 9.061$ exceeds 5.41, we conclude that the means are different. To determine which ones are different, we need to determine which S_i's differ by more than

$$3.055 \frac{2(4)(295-221.625)}{3(4)} = 21.367$$

The results obtained are

B	D	A	E	C
-16.5	-11.5	-1	9	20

It can be seen from these results that the Quade test is not as sensitive as the ANOVA, because at the same level of significance, $\alpha = 0.01$, this test cannot distinguish between Drugs A, E, and C. The A.R.E. of the Quade test to the t-test for the case of $k = 2$ is the same as that of the Wilcoxon signed-ranks test, i.e., 0.955 for normal data. For $k > 2$, the A.R.E. of the Quade test to the F-test has never been found.

5.3.3 Randomized Complete Block Design: The Freidman Test

The Freidman test, which is easier to perform than the Quade test, appears to be more powerful than the Quade test if there are five or more treatments. It is an extension of the sign test.

Assumptions

(i) The results within each block are independent of the results of other blocks.

(ii) Observations may be ranked within blocks.

To perform this test, first find the ranks within blocks as was done in the Quade test, then find the sum of the ranks for each treatment:

$$R_i = \sum_{j=1}^{b} R(X_{ij})$$

Then calculate the terms A_2 and B_2, where

$$A_2 = \sum_{i=1}^{k}\sum_{j=1}^{b} [R(X_{ij})]^2 \text{ and } B_2 = \sum_{i=1}^{k} R_i^2$$

The test statistic is

$$T_2 = \frac{(b-1)\ B_2 - \frac{bk(k+1)^2}{4}}{A_2 - B_2}$$

This should be compared to the proper quantile of an F-distribution with (k - 1) and (b - 1)(k - 1) DF. If this results in the rejection of the null hypothesis, then multiple comparisons can be performed. Two treatments will be significantly different if their sum of ranks R_i differ by more than

$$t_{[1-\alpha/2;\ (b-1)(k-1)]}\left[\frac{2b[A_2 - B_2]}{(b-1)(k-1)}\right]^{\frac{1}{2}}$$

Running this test on the data used previously, we obtain the following results:

		Persons (Blocks)								
		X_{i1}	$R(X_{i1})$	X_{i2}	$R(X_{i2})$	X_{i3}	$R(X_{i3})$	X_{i4}	$R(X_{i4})$	R_i
	A	12	3	14	2	12	4	13	3	12
	B	9	2	13	1	8	1	10	1	5
Drug	C	27	5	32	5	22	5	29	5	20
	D	8	1	22	3	9	2	11	2	8
	E	14	4	29	4	11	3	16	4	15

$A_2 = 3^2 + 2^2 + \ldots + 3^2 + 4^2 = 220$

$B_2 = [12^2 + 5^2 + 20^2 + 8^2 + 15^2]/4 = 214.5$

$T_2 = [3[214.5 - (4)(5)(36)/4]]/(220-214.5) = 18.818$

Since $18.818 > F_{[0.01;\ 4,12]} = 5.41$, we conclude that the treatment means are different. If any two R_i differ by more than

$$3.055\ \frac{2(4)\ [220 - 214.5]}{3(4)} = 5.850$$

then they will be significantly different at the $\alpha = 0.01$ level. The results of this test are

B	D	A	E	C
5	8	12	15	20

These results are different from those of both the F-test and the Quade test; it cannot distinguish between B and D, D and A, A and E, and E and C.

The A.R.E. of the Freidman test with $k = 2$ relative to the t-test is the same as that of the sign test, that is, 0.637. For $k > 2$, the A.R.E. of the Freidman test relative to the F-test depends on k, the number of samples. It is $(0.955)k/(k + 1)$ for normal data and $k/(k + 1)$ for uniform data. It never falls below $(0.864)k/(k + 1)$. For this example, with $k = 5$, the A.R.E. of this test relative to the F-test (assuming normality) is 0.796.

6.0 REGRESSION ANALYSIS

Everything that has been discussed so far has been concerned with different types of experimental designs, that is, methods of detecting differences in population parameters. Another type of analysis that can be done on data is to develop a mathematical model which describes the relationship existing between variables. Such a model can be used to predict values of the dependent variable Y by knowing the values of the independent variables X_i. The technique used to determine the model is known as linear regression, and will be presented using matrix notation.

Assumptions

(i) The relationship between the independent variables and the response is linear; i.e., it can be expressed as $Y = X\beta + \epsilon$ (it is linear in the β's).

(ii) The ϵ_i's are uncorrelated random variables with mean zero and a common variance.

To test hypotheses, a further assumption of normality must be made:

(iii) The ϵ_i's are normally distributed.

The regression model is determined by the method of least squares; that is, it is the figure (a line, if there is only one independent variable) which minimizes the sum of the squares of the errors. The errors, or residuals, are simply the differences between the observed values of Y and the predicted values from the model. Least squares estimators are nice in that they are unbiased, i.e., their expected value is exactly equal to the value of the parameter that they are estimating. They also have the smallest standard error of any linear estimators. This makes them the "best" linear estimators.

The relationship $Y = X\beta + \epsilon$, where there are p independent variables X_i, can be written out as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & & X_{p2} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & X_{1n} & & X_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}$$

Multiplied out, this will yield the equations

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \epsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \epsilon_2$$
$$\cdot$$
$$\cdot$$
$$\cdot$$
$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \epsilon_n$$

To solve for the parameters β_i, one need only solve the equation

$$\beta = (X'X)^{-1}X'Y$$

The calculation involves inverting a $(p + 1)$ X $(p + 1)$ matrix. As a simple example, we will consider only one independent variable X and determine the regression equation

$$Y = \beta_0 + \beta_1 X.$$

Suppose a scientist is studying the relationship between the yield in a chemical reaction and the temperature at which it was run. He runs an

experiment and makes the following observations:

Temp. (X)	80	90	100	110	120	130	140
Yield (Y)	3.2	4.5	4.9	5.7	6.1	6.8	7.0

$$X'X = \begin{bmatrix} n & x_i \\ x_i & x_i^2 \end{bmatrix} = \begin{bmatrix} 7 & 770 \\ 770 & 7500 \end{bmatrix} \qquad (X'X)^{-1} = \frac{1}{19600}\begin{bmatrix} 87500 & -770 \\ -770 & 7 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} y_i \\ x_i y_i \end{bmatrix} = \begin{bmatrix} 38.2 \\ 4374.0 \end{bmatrix}$$

$$(X'X)^{-1}X'Y = \frac{1}{19600}\begin{bmatrix} 87500 & -770 \\ -770 & 7 \end{bmatrix}\begin{bmatrix} 38.2 \\ 4374.0 \end{bmatrix} = \frac{1}{19600}\begin{bmatrix} -25480 \\ 1204 \end{bmatrix} = \begin{bmatrix} -1.3 \\ 0.0614 \end{bmatrix}$$

Thus the relationship between temperature and yield can be expressed as

Yield = (0.0614)(temperature) - 1.3

and predictions of yield can be made for different temperatures. Various tests of hypotheses can be made about these estimated parameters, but these will not be covered here. The interested reader is referred to the bibliography for further information.

Several points should be made about regression before leaving the topic. First and foremost, a good relationship between variables does <u>not</u> imply a causal relationship. In this example, higher temperatures might very well cause high yields, but this is not necessarily the case in all situations. For example, it has been shown that there is a highly significant relationship between ministers' salaries and the sale of liquor in Havanan. This relationship is probably the result of an extraneous factor, namely, the economy.

Another big mistake made in regression, and one which is all too common, is the extrapolation of the model to predict Y values from X values beyond those used in the determination of the model. It is possible that the same relationship will hold, but it is also possible that extrapolation will lead to erroneous, or even meaningless, results. For example, suppose that one makes observations on children from ages 0 - 15 and forms the regression model for predicting height from age. It should be a fairly good relationship. Then, if one substitutes the age of 70 in this relationship, the predicted height would be thirty feet! In this case, it is easy to see that no predictions should be made for any ages other than zero to fifteen.

There is one more point to be made. A regression model is a <u>linear</u> model, in that it is linear in the coefficients β_i. Polynomial models such as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1 X_2 + \beta_4 X_3^2 + \beta_5 \sin(X_4) + \beta_6 \ln X_4 + \epsilon$$

can also be fitted using regression analysis. As long as it is linear in the β_i's, the regression techniques will be valid.

7.0 ANALYSIS OF COVARIANCE

The analysis of covariance, which is a combination of analysis of variance and regression analysis, is a method which can be used to remove the effects of a nuisance variable X which is linearly related to the observed variable Y. The effect of this nuisance variable, or covariate, is removed so that it will not inflate the error mean square. Examples of such situations occur when there is a possibility of a pre and post score. For example, in measuring blood volume after certain treatments are applied, it might be informative to know the original blood volume so that the variance attributable to any linear relationship between blood volume after treatment and the original volume can be removed. Or, suppose one is measuring the strength of a certain fiber. The thickness of the fiber might affect its strength, so the thickness can be treated as a covariate and the variance associated with the linear relationship between strength and thickness can be removed. An additional assumption for the analysis of covariance (in addition to those for the analysis of variance and regression analysis) is that the treatments can have no effect on the covariate, X. The computations for the analysis of covariance are given in Table 6.

As an example, suppose that an experiment is being done to determine the amount of time it takes to analyze a certain type of chemical using three different types of instruments. It is also known that the time it takes to analyze the chemical (Y) is related to the amount of chemical being analyzed (X). Observations are made, and the results and calculations are as follows:

TABLE 6

Analysis of Covariance for One Factor and One Covariate

Source of Variation	df	Sums of Squares and Products			Adjusted For Regression			
		x	xy	yy	y	df	Mean Square	F-ratio
Treatments	a - 1	T_{xx}	T_{xy}	T_{yy} *				
Error	a(n - 1)	E_{xx}	E_{xy}	E_{yy} **	$SS'_E = E_{yy} - (E_{xy})^2/E_{xx}$	a(n - 1) - 1	$MS_E = \frac{SS_E}{a(n-1)-1}$	
Total	an - 1	S_{xx}	S_{xy}	S_{yy} †	$SS'_E = S_{yy} - (S_{xy})^2/S_{xx}$	an - 2		
Adjusted Treatments					$SS_{Tr} = SS'_E - SS_E$	a - 1	$MS_{Tr} = \frac{SS_{Tr}}{a-1}$	$\frac{MS_{Tr}}{MS_E}$

* $T_{xx} = \sum_{j=1}^{a} \frac{x_{ij}^2}{n} - \frac{x_{..}^2}{an}$

$T_{xy} = \sum_{j=1}^{a} \frac{(x_{.j})(Y_{.j})}{an} - \frac{(x_{..})(y_{..})}{an}$

$T_{yy} = \sum_{j=1}^{a} \frac{y_{.j}^2}{n} - \frac{y_{..}^2}{an}$

** $E_{xx} = S_{xx} - T_{xx}$

$E_{xy} = S_{xy} - T_{xy}$

$E_{yy} = S_{yy} - T_{yy}$

† $S_{xx} = \sum_{i=1}^{n}\sum_{j=1}^{a} x_{ij}^2 - \frac{x_{..}^2}{an}$

$S_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{a} x_{ij}y_{ij} - \frac{(x_{..})(y_{..})}{an}$

$S_{yy} = \sum_{i=1}^{n}\sum_{j=1}^{a} y_{ij}^2 - \frac{y_{..}^2}{an}$

Instrument Type						
1		2		3		
Y	X	Y	X	Y	X	
30	27	28	29	43	41	$X_{..}$ = 434
47	43	38	35	25	29	$Y_{..}$ = 446
36	38	49	45	56	53	$\Sigma\Sigma Y_{ij}$ = 17862
44	43	29	28	21	23	$\Sigma\Sigma X_{ij}$ = 16606
157	151	144	137	145	146	$\Sigma\Sigma X_{ij}Y_{ij}$ = 17189

$S_{yy} = 17862 - (446)^2/12 = 1285.667$
$S_{xx} = 16606 - (434)^2/12 = 909.667$
$S_{xy} = 17189 - (434)(446)/12 = 1058.667$
$T_{yy} = [157^2 + 144^2 + 145^2]/4 - (446)^2/12 = 26.167$
$T_{xx} = [151^2 + 137^2 + 146^2]/4 - (434)^2/12 = 25.167$
$T_{xy} = [(157)(151) + (144)(137) + (145)(146)]/4 = 25.167$
$E_{xy} = 1285.667 - 26.167 = 1259.5$
$E_{xx} = 909.667 - 25.167 = 884.5$
$E_{xy} = 1058.667 - 20.9167 = 1037.75$

		Sums of Squares And Cross Products			Adjusted for Regression			
Source of Variation	DF	xx	xy	yy	y	df	MS	F-ratio
Treatments	2	25.167	20.9167	26.167				
Error	9	1037.75	1037.75	1259.5	41.947	8	5.243	
Total	11	1058.667	1058.667	1285.667	53.594	10		
Adjusted Treatments					11.647	2	5.823	1.111

Since $1.111 < F_{(0.05;\ 2,8)} = 4.46$, we conclude that there is no difference in the time required to analyse the chemical using the different instruments. If there had been a difference, multiple comparisons could have been run to determine which means were different, but the tests already presented would have to be modified. Further information can be found in the references.

As with the analysis of variance and regression techniques, this design can be extended to include more than one factor and more than one covariate for the regression. It is not difficult to see that more complex designs will require tremendous amounts of calculation, necessitating the use of a computer.

8.0 SUMMARY

A statistical problem which is encountered in analyzing space-flight data is the limited number of samples that can be obtained. Because of the small sample size available, the analysis of the data should be done in a manner which will glean the maximum amount of information from the experiments as accurately as possible. In order to determine the type of analysis to be used, one should carefully analyze the situation and determine what can be assumed about the nature of the samples.

The procedures which have been presented here should give the basic background required to determine the type of design which is needed or is being used in an experiment. Also, the factors have been specified, which need to be checked in order to insure that the requirements for using a particular test have been met.

In designing an experiment one of the major factors to remember is that the observations must be randomized; i.e., every member of the population about which inferences are to be made should have an equal chance of being observed. Randomization is the foundation of all of the statistical analyses presented here. It is an underlying assumption for every single test, and it is one which, when violated, leads to results of unknown significance when extending the characteristics of the sample to the population. If there are any restrictions on randomization, they should be considered in the analysis of the experiment (by the use of blocking, for example).

In determining the kind of analysis to be used on the data, one should try to ensure the validity of the assumptions of the test employed, be it parametric or nonparametric. For instance, the scale of measurement must be adequate for the test. Also, the statistical test must be appropriate for

the underlying distribution -- a t-test should not be used on obviously non-normal data, and so on. If the assumptions for the parametric test (normality, interval scale, homoscedasticity, etc.) are valid, then the parametric tests will be the most powerful to analyse the data. As these assumptions break down, however, the nonparametric tests become more powerful.

The number of samples being compared is another major consideration in the choice of analyses. If there are more than two samples being compared, two-sample tests should not be applied to the different combinations. This drastically raises above the predetermined level. Some type of test, such as an Analysis of Variance procedure, should be applied for simultaneous comparisons. Likewise, if there is more than one factor to be tested for an effect, a design to test all factors simultaneously should be used so that interactions can be detected. As the designs become more complex, there are no nonparametric alternatives to the analysis of variance and covariance procedures, so these will be the ones to employ.

Another major point which must be stressed is the determination of whether the samples are independent or correlated. Tests for independent samples should never be run on correlated data, because independence is one of the major assumptions in such tests; otherwise sensitivity will be lost. Likewise, running correlated tests on independent samples leads to a reduction in power. Any time that the experiment contains repeated measures, that is, when subjects are used as their own controls, the data will be correlated. Two-factor tests should be paired, and multi-factor tests should be blocked. Much space flight data is obtained by repeated measures, so this must be taken into consideration in the analysis.

Finally, when dealing with the analysis of variance, care should be taken in determining whether each factor is a fixed or random effect. Not

only are the actual tests of significance different for different types of effects, but the conclusions that are drawn from the experiment will differ as well. Any time the conclusions about a factor are to be extended to the entire population from which the factor came, one must choose and analyze the levels of that factor as a random effect.

These are the basic considerations when planning the execution and analysis of experiments involving small sample sizes. The procedures presented here are by no means all-inclusive. In many situations, an entirely new design may have to be created in order to handle the data best. However, the material presented here should give the investigator a good idea of the types of problems that must be considered in planning the experiment, and a direction in which to go to carry out the analysis.

REFERENCES

1. Johnston, R. S., and L. F. Dietlein (Ed.) Biomedical Results From Skylab. National Aeronautics and Space Administration, Washington, D.C., 1977.

2. Leonard, J. I. Energy Balance and the Composition of Weight Loss During Prolonged Space Flight. National Aeronautics and Space Administration, NASA Contractor Report CR-171745, Management & Technical Services Co. TIR 2114-MED-2012, NAS9-16328, 1982.

3. Robinson, J. Statistical Considerations in Design of Spacelab Experiments. National Aeronautics and Space Administration, NASA Contractor Report CR-160197, General Electric Company TIR 741-LSP-8012, NAS9-15487, 1978.

BIBLIOGRAPHY

Bruning, James L., and B. L. Klintz, "Computational Handbook of Statistics," Scott, Foresman and Company, Glenview, Illinois, 1968.

Conover, W. J., "Practical Nonparametric Statistics," 2nd Ed. John Wiley and Sons, New York, 1980.

Daniel, Wayne W., "Applied Nonparametric Statistics," Houghton Mifflin Company, Boston, Mass., 1978.

Dixon, Wilfrid J., and Frank J. Massey, Jr., "Introduction to Statistical Analysis," 2nd Ed. McGraw-Hill Book Company, Inc., New York, 1957.

Graybill, Franklin A., "Theory and Application of the Linear Model," Duxbury Press, North Scituate, Mass., 1976.

Klugh, Henry E., "Statistics: The Essentials for Research," John Wiley and Sons, Inc., New York, 1970.

Montgomery, Douglas C., "Design and Analysis of Experiments," John Wiley and Sons, Inc., New York, 1976.

Neter, John, and William Wasserman, "Applied Linear Statistical Models," Richard D. Irwin, Inc., Homewood, Ill., 1974.

Roscoe, John T., "Fundamental Research Statistics For the Behavioral Sciences," Holt, Rinehart and Winston, Inc., New York, 1969.

Snedecor, G. W., "Statistical Methods," 5th Ed. Iowa State, College Press, Ames, Iowa, 1956.

Winer, B. J., "Statistical Principles in Experimental Design," McGraw-Hill Book Company, New York, 1962.