8

Sampling Distributions of Moments, Statistical Tests, and Procedures

• • •

The basic function of statistical analysis is to make judgments about the real world on the basis of incomplete information. Specifically, we wish to determine the nature of some phenomenon based on a finite sampling of that phenomenon. The sampling procedure will produce a distribution of values, which can be characterized by various moments of that distribution. In the last chapter we saw that the distribution of a random variable is given by the binomial distribution function, which under certain limiting conditions can be represented by the normal probability density distribution function and the Poisson distribution function. In addition, certain physical phenomena will follow distribution functions that are non-normal in nature. We shall see that the characteristics, or statistics, of the distribution functions themselves can be characterized by sampling probability density distribution functions. Generally these distribution functions are also non-normal particularly in the small sample limit.

In section 7.4 we determined the variance of the mean which implied that the moments of any sampling could themselves be regarded as sample that would be characterized by a distribution. However, the act of forming the moment is a decidedly non-random process so that the distribution of the moments may not be represented by the normal distribution. Let us consider several distributions that commonly occur in statistical analysis.

8.1 The t, χ^2, and F Statistical Distribution Functions

In practice, the moments of any sampling distribution have values that depend on the sample size. If we were to repeat a finite sample having N values a large number of times, then the various moments of that sample will vary. Since sampling the same parent population generates them all, we might expect the sampling distribution of the moments to approach that of the parent population as the sample size increases. If the parent population is represented by a random variable, its moments will approach those of the normal curve and their distributions will also approach that of the normal curve. However, when the sample size N is small, the distribution functions for the mean, variance and other statistics that characterize the distribution will depart from the normal curve. It is these distribution functions that we wish to consider.

a. The t-Density Distribution Function

Let us begin by considering the range of values for the mean $\overline{x}$ that we can expect from a small sampling of the parent population N. Let us define the amount that the mean $\overline{x}$ of any particular sample departs from the mean of the parent population x_p as

$$t \equiv (\overline{x} - \overline{x}_p) / \sigma_{\overline{x}} . \tag{8.1.1}$$

Here we have normalized our variable t by the best un-biased estimate of the standard deviation of the mean $\sigma_{\overline{x}}$ so as to produce a dimensionless quantity whose distribution function we can discuss without worrying about its units. Clearly the distribution function of t will depend on the sample size N. The differences from the normal curve are represented in Figure 8.1. The function is symmetric with a mean, mode, and skewness equal to zero. However, the function is rather flatter than the normal curve so the kurtosis is greater than three, but will approach three as N increases. The specific form of the t-distribution is

$$f(t) = \frac{\Gamma[\frac{1}{2}(N+1)]}{\sqrt{\pi N}\,\Gamma(\frac{1}{2}N)}\left[1 + \frac{t^2}{N}\right]^{-(N+1)/2} , \tag{8.1.2}$$

which has a variance of

$$\sigma^2{}_t = N/(N-2) \ . \tag{8.1.3}$$

Generally, the differences between the t-distribution function and the normal curve are negligible for N > 30, but even this difference can be reduced by using a normal curve with a variance given by equation (8.1.3) instead of unity. At the out set we should be clear about the difference between the number of samples N and the number of degrees of freedom v contained in the sample. In Chapter 7 (section 7.4) we introduced the concept of "degrees of freedom" when determining the variance. The variance of both a single observation and the mean was expressed in terms of the mean itself. The determination of the mean reduced the number of independent information points represented by the data by one. Thus the factor of

(N-1) represented the remaining independent pieces of information, known as the degrees of freedom, available for the statistic of interest. The presence of the mean in the expression for the t-statistic [equation (8.1.1)] reduces the number of degrees of freedom available for t by one.

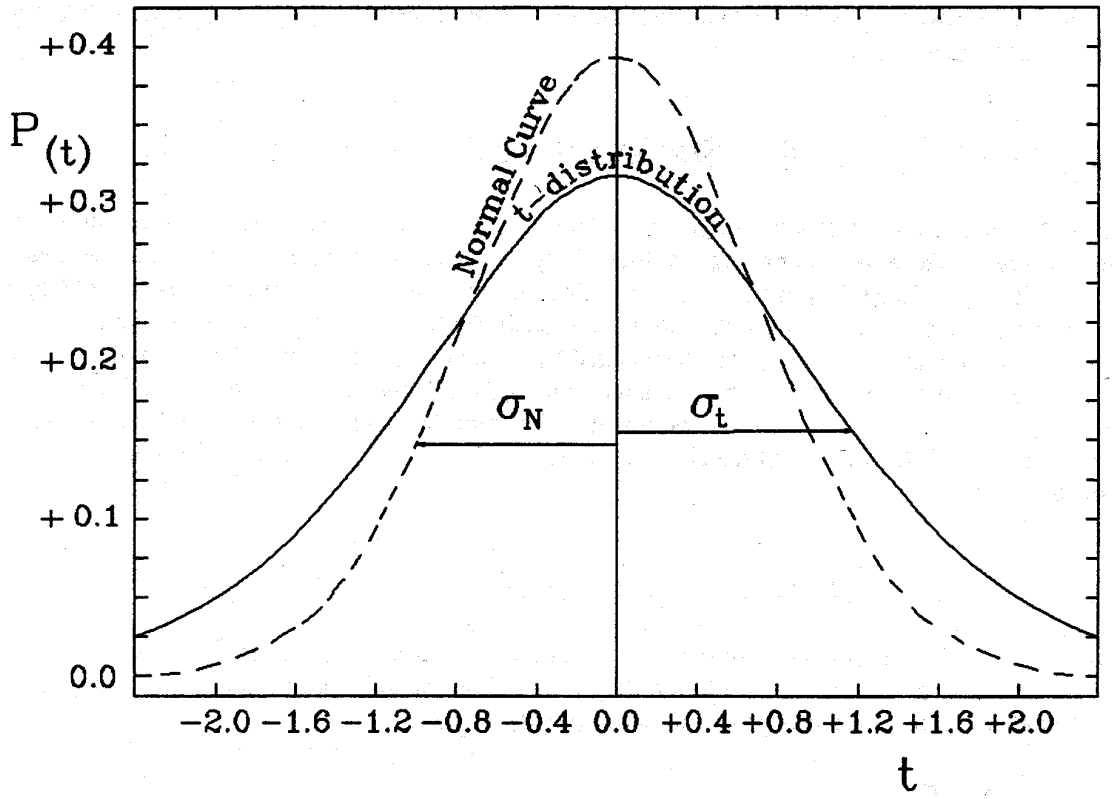

Figure 8.1 shows a comparison between the normal curve and the t*-distribution function for* N=8*. The symmetric nature of the* t*-distribution means that the mean, median, mode, and skewness will all be zero while the variance and kurtosis will be slightly larger than their normal counterparts. As* $N \to \infty$*, the* t*-distribution approaches the normal curve with unit variance.*

b. The χ^2 -Density Distribution Function

Just as we inquired into the distribution of means x that could result from various samples, so we could ask what the distribution of variances might be. In chapter 6 (section 6.4) we introduced the parameter χ^2 as a measure of the mean square error of a least square fit to some data. We chose that symbol with the current use in mind. Define

$$\chi^2 = \sum_{j=1}^{N} (x_j - \overline{x}_j)^2 / \sigma_j^2 \quad , \tag{8.1.4}$$

where σ^2_j is the variance of a single observation. The quantity χ^2 is then sort of a normalized square error. Indeed, in the case where the variance of a single observation is constant for all observations we can write

$$\chi^2 = N\overline{\varepsilon^2/\sigma^2}\,, \tag{8.1.5}$$

where ε^2 is the mean square error. However, the value of χ^2 will continue to grow with N so that some authors further normalize χ^2 so that

$$\chi^2_\nu = \chi^2/\nu\,. \tag{8.1.6}$$

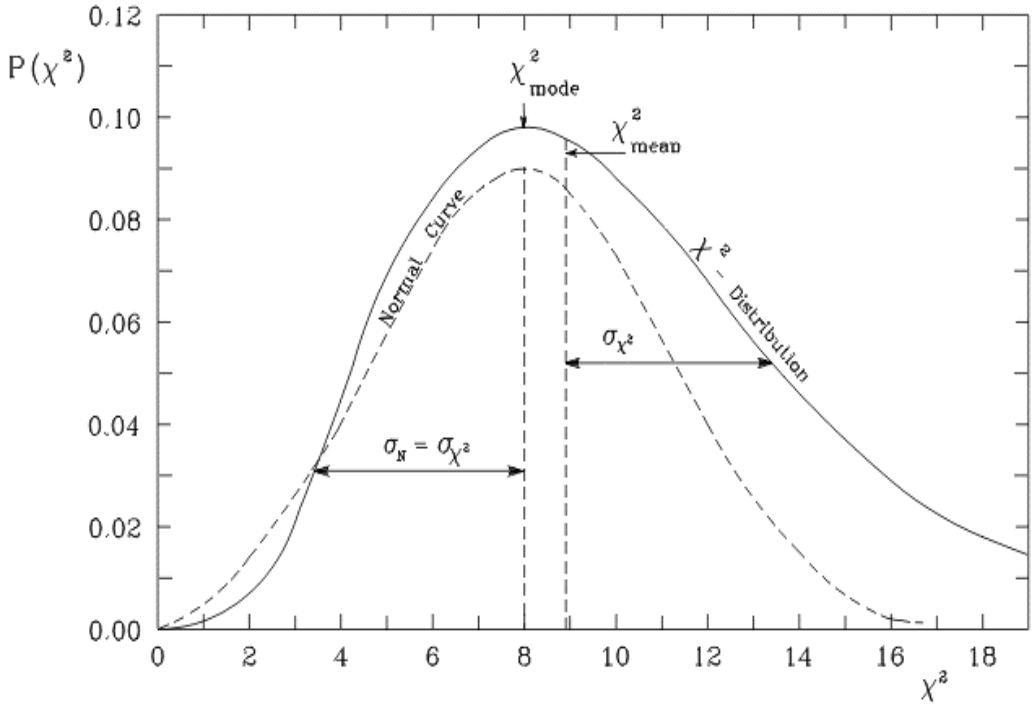

Figure 8.2 compares the χ^2- distribution with the normal curve. For N = 10 *the curve is quite skewed near the origin with the mean occurring past the mode* (χ^2 = 8). *The Normal curve has* $\mu = 8$ *and* $\sigma^2 = 20$. *For large* N, *the mode of the χ^2-distribution approaches half the variance and the distribution function approaches a normal curve with the mean equal the mode.*

Here the number of degrees of freedom (i.e. the sample size N reduced by the number of independent moments present in the expression) does not appear explicitly in the result. Since χ^2 is intrinsically positive, its distribution function cannot be expected to be symmetric. Figure 8.2 compares the probability density distribution function for χ^2, as given by

$$f(\chi^2) = [2^{N/2}\Gamma(\tfrac{1}{2}N)]^{-1}\, e^{-\chi^2/2}(\chi^2)^{\frac{1}{2}(N-2)}\,, \tag{8.1.7}$$

with the normal distribution function.

The moments of the χ^2 density distribution function yield values of the variance, mode, and skewness of

$$\left.\begin{aligned} \sigma^2_{\chi^2} &= 2N \\ \chi^2_m &= N-2 \\ s &= \sqrt{2/N} \end{aligned}\right\} . \tag{8.1.8}$$

As N increases, the mode increases approaching half the variance while the skewness approaches zero. Thus, this distribution function will also approach the normal curve as N becomes large.

c. *The* F-*Density Distribution Function*

So far we have considered cases where the moments generated by the sampling process are all generated from samples of the same size (i.e. the same value of N). We can ask how the sample size could affect the probability of obtaining a particular value of the variance. For example, the χ^2 distribution function describes how values of the variance will be distributed for a particular value of N. How could we expect this distribution function to change *relatively* if we changed N? Let us inquire into the nature of the probability density distribution of the ratio of two variances, or more specifically define F to be

$$F_{12} \equiv \left((\chi_1^2/\nu_1) \Big/ (\chi_2^2/\nu_2) \right) = \left(\chi^2_{\nu_1} \Big/ \chi^2_{\nu_2} \right) . \tag{8.1.9}$$

This can be shown to have the rather complicated density distribution function of the form

$$f(F) = \frac{\Gamma[\tfrac{1}{2}(N_1+N_2)]\, N_1^{\frac{1}{2}N_1}\, N_2^{\frac{1}{2}N_2}\, F_{12}^{\frac{1}{2}(N_1-1)}}{\Gamma(\tfrac{1}{2}N_1)\Gamma(\tfrac{1}{2}N_2)(N_1F+N_2)^{\frac{1}{2}(N_1+N_2)}} = \frac{\Gamma[\tfrac{1}{2}(\nu_1+\nu_2)]}{\Gamma(\tfrac{1}{2}\nu_1)\Gamma(\tfrac{1}{2}\nu_2)} \left[\frac{\nu_1}{\nu_2}\right]^{\nu_1/2} \frac{F_{12}^{(\nu_1-1)/2}}{(1+F_{12}\nu_1/\nu_2)^{(\nu_1+\nu_2)/2}} , \tag{8.1.10}$$

where the degrees of freedom ν_1 and ν_2 are N_1 and N_2 respectively. The shape of this density distribution function is displayed in Figure 8.3.

The mean, mode and variance of F-probability density distribution function are

$$\left.\begin{aligned} \bar{F} &= N_2/(N_2-2) \\ F_{m0} &= \frac{N_2(N_1-2)}{N_1(N_2-2)} \\ \sigma_F^2 &= \frac{2(N_2+N_1-2)N_2^2}{N_1(N_2-4)(N_2-2)^2} \end{aligned}\right\} . \tag{8.1.11}$$

As one would expect, the F-statistic behaves very much like a χ^2 except that there is an additional parameter involved. However, as N_1 and N_2 both become large, the F-distribution function becomes indistinguishable from the normal curve. While N_1 and N_2 have been presented as the sample sizes for two different samplings of the parent population, they really represent the number of independent pieces of information (i.e. the number of degrees of freedom give or take some moments) entering into the determination of the variance σ^2_n or alternately, the value of χ^2_n. As we saw in chapter 6, should the statistical analysis involve a more complicated function of the form $g(x,a_i)$, the number of degrees of freedom will depend on the number of values of a_i. Thus the F-statistic can be used to provide the distribution of variances resulting from a change in the number of values of a_i thereby changing the number of degrees of freedom as well as a change in the sample size N. We shall find this very useful in the next section.

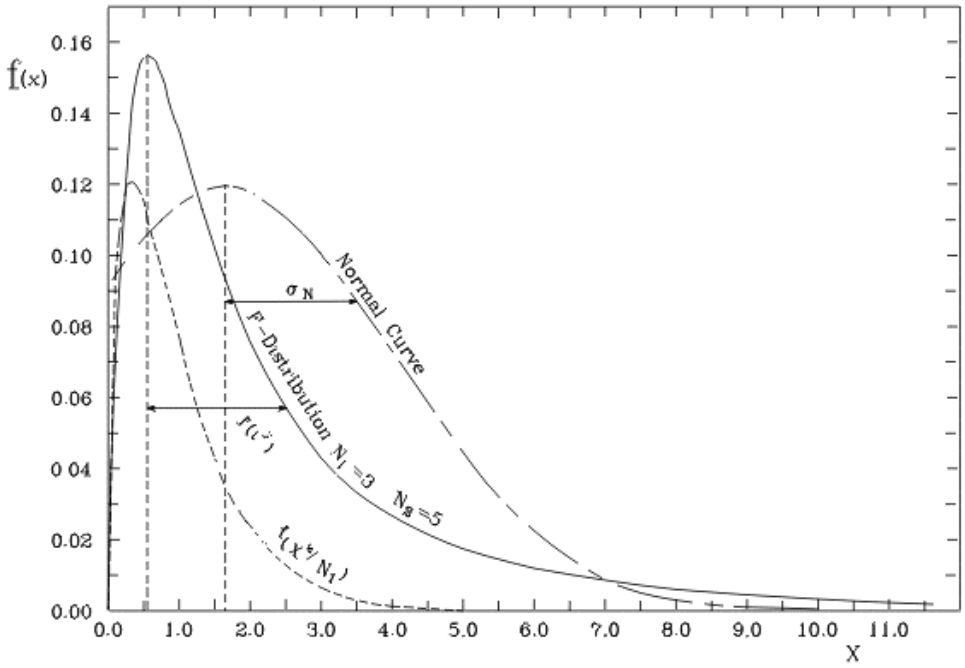

Figure 8.3 shows the probability density distribution function for the F*-statistic with values of* $N_1 = 3$ *and* $N_2 = 5$ *respectively. Also plotted are the limiting distribution functions* $f(\chi^2/N_1)$ *and* $f(t^2)$*. The first of these is obtained from* f(F) *in the limit of* $N_2 \to \infty$*. The second arises when* $N_1 \to 1$*. One can see the tail of the* $f(t^2)$ *distribution approaching that of* f(F) *as the value of the independent variable increases. Finally, the normal curve which all distributions approach for large values of* N *is shown with a mean equal to* $\bar{F}$ *and a variance equal to the variance for* f(F)*.*

Since the t, χ^2, and F density distribution functions all approach the normal distribution function as $N \to \infty$, the normal curve may be considered a special case of the three curves. What is less obvious is that the t- and χ^2 density distribution functions are special cases of the F density distribution. From the defining

equations for t [equation (8.1.1)] and χ^2 [equation(8.1.4)] we see that

$$\lim_{N \to 1} t^2 = \chi^2 \ , \tag{8.1.12}$$

From equations (8.1.5) and (8.1.6) the limiting value of the normalized or reduced χ^2 is given by

$$\lim_{v \to \infty} \chi_v^2 = 1 \ , \tag{8.1.13}$$

so that

$$\lim_{\substack{N_1 \to N \\ N_2 \to \infty}} F = \chi^2/N \ . \tag{8.1.14}$$

Finally t can be related to F in the special case where

$$\lim_{\substack{N_1 \to 1 \\ N_2 \to N}} F = t^2 \ . \tag{8.1.15}$$

Thus we see that the F probably density distribution function is the general generator for the density distribution functions for t and χ^2 and hence for the normal density distribution function itself.

8.2 The Level of Significance and Statistical Tests

Much of statistical analysis is concerned with determining the extent to which the properties of a sample reflect the properties of the parent population. This could be re-stated by obtaining the probability that the particular result differs from the corresponding property of the parent population by an amount ε. These probabilities may be obtained by integrating the appropriate probability density distribution function over the appropriate range. Problems formulated in this fashion constitute a statistical test. Such tests generally test hypotheses such as "this statistic does not differ from the value of the parent population". Such a hypothesis is often called *null hypothesis* for it postulates no difference between the sample and the value for the parent population. We test this hypothesis by ascertaining the probability that the statement is true or possibly the probability that the statement is false. Statistically, one never "proves" or "disproves" a hypothesis. One simply establishes the probability that a particular statement (usually a null hypothesis) is true or false. If a hypothesis is sustained or rejected with a certain probability p the statement is often said to be *significant* at a percent level corresponding to the probability multiplied by 100. That is, a particular statement could be said to be significant at the 5% level if the probability that the event described could occur by chance is .05.

a. The "Students" t*-Test*

Say we wish to establish the extent to which a particular mean value $\bar{x}$ obtained from a sampling of N items from some parent population actually represents the mean of the parent population. To do this we must establish some tolerances that we will accept as allowing the statement that $\bar{x}$ is indeed "the same" as $\bar{x}_p$. We can do this by first deciding how often we are willing to be wrong. That is, what is the acceptable probability that the statement is false? For the sake of the argument, lets us take that value to be 5%. We can re-write equation (8.1.1) as

$$\bar{x} = \bar{x}_p \pm \sigma_{\bar{x}}\, t \quad , \tag{8.2.1}$$

and thereby establish a range δ in $\bar{x}$ given by

$$\delta = \left|\bar{x} - \bar{x}_p\right| = \sigma_{\bar{x}} t \quad , \tag{8.2.2}$$

or for the 5% level as

$$\delta_{(5\%)} = \sigma_{\bar{x}} t_{5\%} \quad , \tag{8.2.3}$$

Now we have already established that the t-distribution depends only on the sample size N so that we may find $t_{5\%}$ by integrating that distribution function over that range of t that would allow for it to differ from the expected value with a probability of 5%. That is

$$0.05 = 2\int_{t_{5\%}}^{\infty} f(t)\,dt = 2\left(1 - \int_0^{t_{5\%}} f(t)\,dt\right) \quad . \tag{8.2.4}$$

The value of t will depend on N and the values of δ that result and are known as the *confidence limits of the 5% level*. There are numerous books that provide tables of t for different levels of confidence for various values of N (e.g. Croxton et al[1]). For example if N is 5, then the value of t corresponding to the 5% level is 2.571. Thus we could say that there is only a 5% chance that $\bar{x}$ differs from $\bar{x}_p$ by more than $2.571\sigma_{\bar{x}}$. In the case where the number of samples increases to $\bar{x}_p$, the same confidence limits drop to 1.96 σ_x. We can obtain the latter result simply by integrating the 'tails' of the normal curve until we have enclosed 5% of the total area of the curve. Thus it is important to use the proper density distribution function when dealing with small to moderate sample sizes. There integrals set the confidence limit appropriate for the small sample sizes.

We may also use this test to examine additional hypotheses about the nature of the mean. Consider the following two hypotheses:

a. *The measured mean is greater than the mean of the parent population (i.e* $\bar{x} > \bar{x}_p$*),*

and

b. *The measured mean is less than the mean of the parent population (i.e* $\bar{x} < \bar{x}_p$*)* .

While these hypotheses resemble the null hypothesis, they differ subtly. In each case the probability of meeting the hypothesis involves the frequency distribution of t on just one side of the mean. Thus the factor of two that is present in equation (8.2.4) allowing for both "tails" of the t-distribution in establishing the probability of occurrence is absent. Therefore the confidence limits at the p-percentile are set by

$$\left.\begin{aligned} p_a &= \int_{t_p}^{\infty} f(t)\,dt = 1 - \int_0^{t_p} f(t)\,dt \\ p_b &= \int_{-\infty}^{-t_p} f(t)\,dt = 1 - \int_{-t_p}^{0} f(t)\,dt \end{aligned}\right\} . \qquad (8.2.5)$$

Again one should be careful to remember that one never "proves" a hypothesis to be correct, one simply finds that it is not necessarily false. One can say that the data are consistent with the hypothesis at the p-percent level.

As the sample size becomes large and the t density distribution function approaches the normal curve, the integrals in equations (8.2.4) and (8.2.5) can be replaced with

$$\left.\begin{aligned} p &= 2\mathrm{erfc}(t_p) = 2[1 - \mathrm{erf}(t_p)] \\ p_{a,b} &= \mathrm{erfc}(\pm t_p) = 1 - \mathrm{erf}(\pm t_p) \end{aligned}\right\} , \qquad (8.2.6)$$

where erf(x) is called the error function and erfc(x) is known as the complimentary error function of x respectively. The effect of sample sizes on the confidence limits, or alternately the levels of significance, when estimating the accuracy of the mean was first pointed out by W.S. Gossett who used the pseudonym "Student" when writing about it. It has been known as "Students's t-Test" ever since. There are many other uses to which the t-test may be put and some will be discussed later in this book, but these serve to illustrate its basic properties.

b. The χ^2-test

Since χ^2 is a measure of the variance of the sample mean compared with what one might expect, we can use it as a measure of how closely the sampled data approach what one would expect from the sample of a normally distributed parent population. As with the t-test, there are a number of different ways of expressing this, but perhaps the simplest is to again calculate confidence limits on the value of χ^2 that can be expected from any particular sampling. If we sample the entire parent population we would expect a χ_v^2 of unity. For any finite sampling we can establish the probability that the actual value of χ^2 should occur by chance. Like the t-test, we must decide what probability is acceptable. For the purposes of demonstration, let us say that a 5% probability that χ^2 did occur by chance is a sufficient criteria. The value of χ^2 that represents the upper limit on the value that could occur by chance 5% of the time is

$$0.05 = 2\int_{\chi^2_{5\%}}^{\infty} f(\chi^2, N)\,d\chi^2 = N - \int_0^{\chi^2_{5\%}} f(\chi^2, N)\,d\chi^2 \quad , \qquad (8.2.7)$$

which for a general percentage is

$$p = \int_{\chi^2_p}^{\infty} f(\chi^2, N)\,d\chi^2 \quad , \qquad (8.2.8)$$

Thus an observed value of χ^2 that is greater than χ^2_p would suggest that the parent population is not represented by the normal curve or that the sampling procedure is systematically flawed.

The difficulty with the χ^2-test is that the individual values of σ^2_i must be known before the calculations implied by equation (8.1.4) can be carried out. Usually there is an independent way of estimating them. However, there is usually also a tendency to under estimate them. Experimenters tend

believe their experimental apparatus performs better than it actually does. This will result in too large a value of an observed chi-squared (i.e. χ^2_o). Both the t-test and the χ^2-test as described here test specific properties of a single sample distribution against those expected for a randomly distributed parent population. How may we compare two different samples of the parent population where the variance of a single observation may be different for each sample?

c. The F*-test*

In section 8.1 we found that the ratio of two different χ^2's would have a sampling distribution given by equation (8.1.10). Thus if we have two different experiments that sample the parent population differently and obtain two different values of χ^2, we can ask to what extent are the two experiments different. Of course the expected value of F would be unity, but we can ask "what is the probability that the actual value occurred by chance?" Again we establish the confidence limits on F_{12} by integrating the probability density distribution function so that

$$p = \int_{F_{12}^{(p)}}^{\infty} f(F)\,dF \quad . \tag{8.2.9}$$

Thus if the observed value of F_{12} exceeds $F_{12}^{(p)}$, then we may suspect that one of the two experiments did not sample the parent population in an unbiased manner. However, satisfying the condition that $F_{12} < F_{12}^{(p)}$ is not sufficient to establish that the two experiments did sample the parent population in the same way. F_{12} might be too small. Note that from equation (8.1.9) we can write

$$F_{12} = 1/F_{21} \quad . \tag{8.2.10}$$

One must then compare F_{21} to its expected value $F_{21}^{(p)}$ given by

$$p = \int_{F_{21}^{(p)}}^{\infty} f(F)\,dF \quad . \tag{8.2.11}$$

Equations (8.2.9) and (8.2.11) are not exactly symmetric so that only in the limit of large ν_1 and ν_2 can we write

$$F > F_{12} > 1/F \quad . \tag{8.2.12}$$

So far we have discussed the cases where the sampled value is a direct measure of some quantity found in the parent population. However, more often than not the observed value may be some complicated function of the random variable x. This was certainly the case with our discussion of least squares in chapter 6. Under these conditions, the parameters that relate y and x must be determined by removing degrees of freedom needed to determine other parameters of the fit from the statistical analysis. If we were to fit N data points with a function having n independent coefficients, then we could, in principle, fit n of the data points exactly leaving only (N-n) points to determine, say, ε^2. Thus there would only be (N-n) degrees of freedom left for statistical analysis. This is the origin of the (N-n) term in the denominator of equation (6.3.26) for the errors (variances) of the least square coefficients that we found in chapter 6. Should the mean be required in subsequent analysis, only (N-n-1) degrees of freedom would remain. Thus we must be careful in determining the number of degrees of freedom when dealing with a problem having multiple parameters. This includes the use of the t-test and the χ^2-test. However, such problems suggest a very powerful application of the F-test. Assume that we have fit some data with a function of n parameters. The χ^2-test and perhaps other considerations suggest that we have not achieved the best fit to the data so that we consider a function with an additional parameter so that there are now a total of (n+1) independent parameters. Now we know that including an additional parameter will remove one more degree of freedom from the analysis and that the mean square error ε^2 should decrease. The question then becomes, whether or not the decrease in ε^2

represents an amount that we would expect to happen by chance, or by including the additional parameter have we matched some systematic behavior of the parent population. Here the F-test can provide a very useful answer. Both samples of the data are "observationally" identical so that the σ^2_i's for the two χ^2's are identical. The only difference between the two χ's is the loss on one degree of freedom. Under the conditions that σ_i^2s are all equal, the F-statistic takes on the fairly simple form of

$$F = \frac{(n - n - 1)\overline{\varepsilon_n^2}}{(N - n)\overline{\varepsilon_{n=1}^2}} \quad . \tag{8.2.13}$$

However, now we wish to know if F_{12} is greater that what would be expected by chance (i.e. is $F_{12} > F_{12}^{(p)}$). Or answering the question "What is the value of p for which $F_{12} = F_{12}^{(p)}$?" is another way of addressing the problem. This is a particularly simple method of determining when the addition of a parameter in an approximating function produces an improvement which is greater than that to be expected by chance. It is equivalent to setting confidence limits for the value of F and thereby establishing the significance of the additional parameter. Values of the probability integrals that appear in equations (8.2.5), (8.2.6), (8.2.8), (8.2.9), and (8.2.11) can be found in the appendices of most elementary statistics books[1] or the CRC Handbook of tables for Probability and Statistics[2]. Therefore the F-test provides an excellent criterion for deciding when a particular approximation formula, lacking a primary theoretical justification, contains a sufficient number of terms.

d. Kolmogorov-Smirnov Tests

Virtually all aspects of the statistical tests we have discussed so far have been based on ascertaining to what extent a particular property or statistic of a sample population can be compared to the expected statistic for the parent population. One establishes the *"goodness of fit"* of the sample to the parent population on the basis of whether or not these statistics fall within the expected ranges for a random sampling. The parameters such as skewness, kurtosis, t, χ^2, or F, all represent specific properties of the distribution function and thus such tests are often called parametric tests of the sample. Such tests can be definitive when the sample size is large so that the actual value of the parameter represents the corresponding value of the parent population. When the sample size is small, even when the departure of the sampling distribution function from a normal distribution is allowed for, the persuasiveness of the statistical argument is reduced. One would prefer tests that examined the entire distribution in light of the expected parent distribution. Examples of such tests are the Kolmogorov-Smirnov tests.

Let us consider a situation similar to that which we used for the t-test and χ^2-test where the random variable is sampled directly. For these tests we shall used the observed data points, x_i, to estimate the *cumulative probability* of the probability density distribution that characterizes the parent population. Say we construct a histogram of the values of x_i that are obtained from the sampling procedure (see figure 8.4). Now we simply sum the number of points with $x < x_i$, normalized by the total number of points in the sample. This number is simply the probability of obtaining $x < x_i$ and is known as the cumulative probability distribution $S(x_i)$. It is reminiscent of the probability integrals we had to evaluate for the parametric tests [eg. equations (8.2.5),(8.2.8), and (8.2.9)] except that now we are using the sampled probability distribution itself instead of one obtained from an assumed binomial distribution. Therefore we can define $S(x_i)$ by

$$S(x_i) = \frac{1}{N}\sum_{j=1}^{i} n(x_j < x) \quad . \tag{8.2.14}$$

This is to be compared with the cumulative probability distribution of the parent population, which is

$$p(x) = \int_0^x f(z)\,dz \quad . \tag{8.2.15}$$

The statistic which is used to compare the two cumulative probability distributions is the largest departure D_0 between the two cumulative probability distributions, or

$$D_0 \equiv \text{Max} \left| S(x_i) - p(x_i) \right|, \forall x_i \quad . \tag{8.2.16}$$

If we ask what is the probability that the two probability density distribution functions are different (i.e. disproof of the null hypothesis), then

$$\left.\begin{aligned} P_{D_0} &= Q(D_0\sqrt{N}) \\ P_{D_0} &= Q[D_0\sqrt{N_1 N_2/(N_1+N_2)}\,] \end{aligned}\right\}, \tag{8.2.17}$$

where Press et al[3] give

$$Q(x) = 2\sum_{j=1}^{\infty}(-1)^{j-1} e^{-2j^2x^2} \quad . \tag{8.2.18}$$

Equations (8.2.17) simply state that if the measured value of D is D_0 then p is the probability that the null hypothesis is false. The first of equations (8.2.17) applies to the case where the probability density distribution function of the parent population is known so that the cumulative probability required to compute D_0 from equations (8.2.15) and (8.2.16) is known *a priori*. This is known as the Kolmogorov-Smirnov Type 1 test. If one has two different distributions $S_1(x_i)$ and $S_2(x_i)$ and wishes to know if they originate from the same distribution, then one uses the second of equations (8.2.17) and obtains D_0 from $\text{Max}\left| S_1(x_i)-S_2(x_i) \right|$. This is usually called the Kolmogorov-Smirnov Type 2 test.

Note that neither test assumes that the parent population is given by the binomial distribution or the normal curve. This is a major strength of the test as it is relatively independent of the nature of the actual probability density distribution function of the parent population. All of the parametric tests described earlier compared the sample distribution with a normal distribution which may be a quite limiting assumption. In addition, the cumulative probability distribution is basically an integral of the probability density distribution function which is itself a probability that x lies in the range of the integral. Integration tends to smooth out local fluctuations in the sampling function. However, by considering the entire range of the sampled variable x, the properties of the whole density distribution function go into determining the D_0-statistic. The combination of these two aspects of the statistic makes it particularly useful in dealing with small samples. This tends to be a basic property of the non-parametric statistical tests such as the Kolmogorov- Smirnov tests.

We have assumed throughout this discussion of statistical tests that a single choice of the random variable results in a specific sample point. In some cases this is not true. The data points or samples could themselves be averages or collections of data. This data may be treated as being collected in groups or bins. The treatment of such data becomes more complicated as the number of degrees of freedom is no longer calculated as simply as for the cases we have considered. Therefore we will leave the statistical analysis of grouped or binned data to a more advanced course of study in statistics.

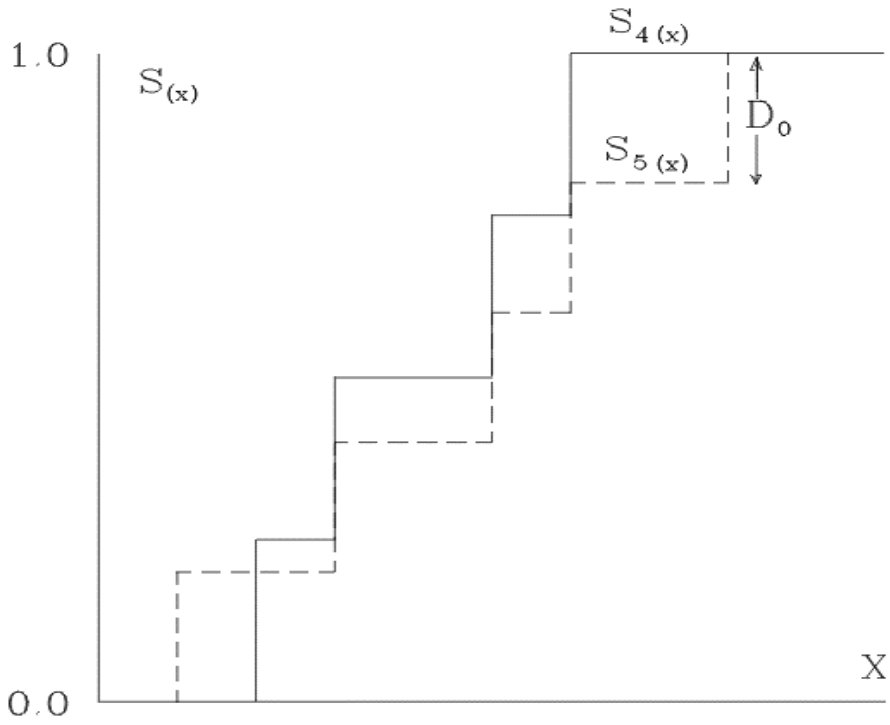

Figure 8.4 shows a histogram of the sampled points x_i *and the cumulative probability of obtaining those points. The Kolmogorov-Smirnov tests compare that probability with another known cumulative probability and ascertain the odds that the differences occurred by chance.*

8.3 Linear Regression, and Correlation Analysis

In Chapter 6 we showed how one could use the principle of least squares to fit a function of several variables and obtain a maximum likelihood or most probable fit under a specific set of assumptions. We also noted in chapter 7 that the use of similar procedures in statistics was referred to as regression analysis. However, in many statistical problems it is not clear which variable should be regarded as the dependent variable and which should be considered as the independent variable. In this section we shall describe some of the techniques for approaching problems where cause and effect cannot be determined.

Let us begin by considering a simple problem involving just two variables, which we will call X_1 and X_2. We have reason to believe that these variables are related, but have no *a priori* reason to believe that either should be regarded as causally dependent on the other. However, in writing any algebraic formalism it is necessary to decide which variables will be regarded as functions of others. For example, we could write

$$X_1 = a_{1.2} + X_2 b_{1.2} \,, \tag{8.3.1}$$

or

$$X_2 = a_{2.1} + X_1 b_{2.1} \quad . \tag{8.3.2}$$

Here we have introduced a notation commonly used in statistics to distinguish the two different sets of a's and b's. The subscript m.n indicates which variable is regarded as being dependent (i.e. the m) and which is to be regarded as being independent (i.e. the n).

a. The Separation of Variances and the Two-Variable Correlation Coefficient

In developing the principle of least squares in chapter 6, we regarded the uncertainties to be confined to the dependent variable alone. We also indicated some simple techniques to deal with the case where there was error in each variable. Here where the very nature of dependency is uncertain, we must extend these notions. To do so, let us again consider the case of just two variables X_1 and X_2. If we were to consider these variables individually, then the distribution represented by the sample of each would be characterized by moments such as $\overline{X}_1$, σ^2_1, $\overline{X}_2$, σ^2_2, etc. However, these variables are suspected to be related. Since the simplest relationship is linear, let us investigate the linear least square solutions where the roles of independence are interchanged. Such analysis will produce solutions of the form

$$\left.\begin{aligned} X_2^c &= a_{1.2} + X_1 b_{2.1} \\ X_1^c &= a_{1.2} + X_2 b_{1.2} \end{aligned}\right\} . \tag{8.3.3}$$

Here we have denoted the values of the dependent variable resulting from the solution by the superscript ^{c}. The lines described by equations (8.3.3) resulting from a least square analysis are known in statistics as *regression lines*. We will further define the departure of any data value X_i from its mean value as a *deviation* x_i. In a similar manner let x_i^c be the calculated deviation of the ith variable. This variable measures the spread in the ith variable as given by the regression equation. Again the subscript denotes the dependent variable. Thus, for a regression line of the form of the first of equations (8.3.3), $(x_2 - x_2^c)$ would be the same as the error ε that was introduced in chapter 6 (see figure 8.5). We may now consider the statistics of the deviations x_i. The mean of the deviations is zero since $a_{m.n} = X_n$, but the variances of the deviations will not be. Indeed they are just related to what we called the mean square error in chapter 6. However, the value of these variances will depend on what variable is taken to be the dependent variable. For our situation, we may write the variances of x_i as

$$\left.\begin{aligned} \sigma^2_{2.1} &= \left(\sum X_2^2 - a_{2.1}\sum X_2 - b_{2.1}\sum X_1 X_2\right)/N \\ \sigma^2_{1.2} &= \left(\sum X_1^2 - a_{1.2}\sum X_1 - b_{1.2}\sum X_1 X_2\right)/N \end{aligned}\right\} \quad . \tag{8.3.4}$$

Some authors[4] refer to these variances as *first-order variances*. While the origin of equations (8.3.4) is not immediately obvious, it can be obtained from the analysis we did in chapter 6 (section 6.3). Indeed, the right hand side of the first of equations (8.3.4) can be obtained by combining equations (6.3.24) and (6.3.25) to get the term in the large parentheses on the right hand side of equation (6.3.26). From that expression it is clear that

$$\sigma^2_{1.2} = \overline{w\varepsilon^2} \quad . \tag{8.3.5}$$

The second of equations (8.3.4) can be obtained from the first by symmetry. Again, the mean of x_i^c

is clearly zero but its variance will not be. It is simple a measure in the spread of the computed values of the dependent variable. Thus the total variance σ^2_i will be the sum of the variance resulting from the relation between X_1 and X_2 (i.e. $\sigma^2\, x^c_i$) and the variance resulting from the failure of the linear regression line to accurately represent the data. Thus

$$\left.\begin{aligned}\sigma_1^2 &= \sigma^2_{x_1^c} + \sigma^2_{1.2}\\ \sigma_2^2 &= \sigma^2_{x_2^c} + \sigma^2_{2.1}\end{aligned}\right\}\ . \tag{8.3.6}$$

The division of the total variance σ^2_i into parts resulting from the relationship between the variables X_1 and X_2 and the failure of the relationship to fit the data allow us to test the extent to which the two variables are related. Let us define

$$r_{12} = \frac{\sum X_1 X_2}{N\sigma_1\sigma_2} = \pm\left(\frac{\sigma^2_{x_1^c}}{\sigma_1^2}\right)^{1/2} = \pm\left(\frac{\sigma^2_{x_2^c}}{\sigma_2^2}\right)^{1/2} = \pm\left(1-\frac{\sigma^2_{1.2}}{\sigma_1^2}\right)^{1/2} = \pm\left(1-\frac{\sigma^2_{2.1}}{\sigma_2^2}\right)^{1/2} = r_{21}\ . \tag{8.3.7}$$

The quantity r_{ij} is known as the Pearson correlation coefficient after Karl Pearson who made wide use of it. This simple correlation coefficient r_{12} measures the way the variables X_1 and X_2 change with respect to their means and is normalized by the standard deviations of each variable. However, the meaning is perhaps more clearly seen from the form on the far right hand side of equation (8.3.7). Remember σ_2 simply measures the scatter of X_{2j} about the mean X_2, while $\sigma_{2.1}$ measures the scatter of X_{2j} about the regression line. Thus, if the variance $\sigma^2_{2.1}$ accounts for the entire variance of the dependent variable X_2, then the correlation coefficient is zero and a plot of X_2 against X_1 would simply show a random scatter diagram. It would mean that the variance $\sigma^2\, x_2^c$ would be zero meaning that none of the total variance resulted from the regression relation. Such variables are said to be uncorrelated. However, if the magnitude of the correlation coefficient is near unity then $\sigma^2_{2.1}$ must be nearly zero implying that the total variance of X_2 is a result of the regression relation. The definition of r as given by the first term in equation (8.3.7) contains a sign which is lost in the subsequent representations. If an increase in X_1 results in a decrease in X_2 then the product of the deviations will be negative yielding a negative value for r_{12}. Variables which have a correlation coefficient with a large magnitude are said to be highly correlated or anti-correlated depending on the sign of r_{12}. It is worth noting that $r_{12} = r_{21}$, which implies that it makes no difference which of the two variables is regarded as the dependent variable.

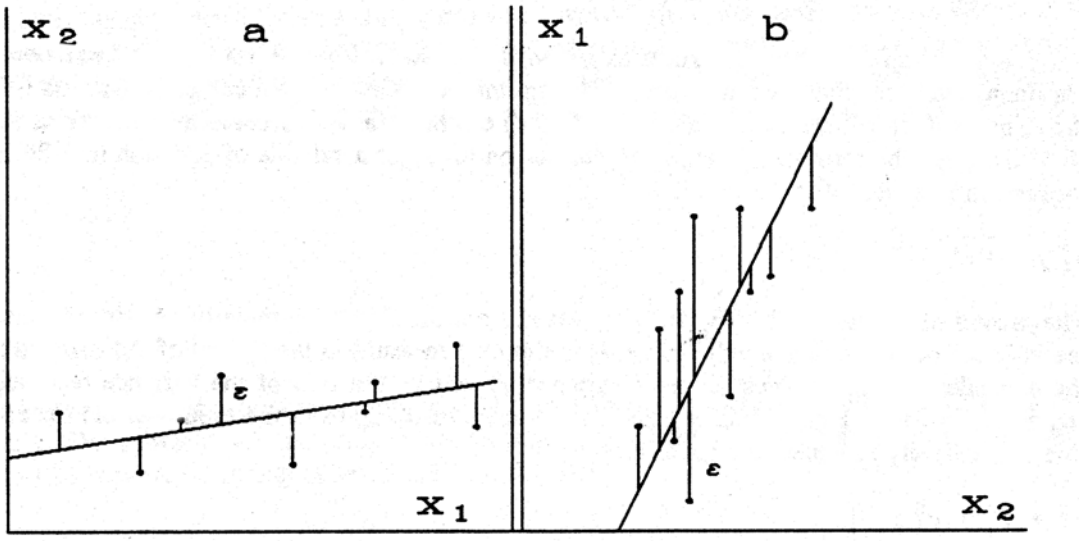

Figure 8.5 shows the regression lines for the two cases where the variable X_2 is regarded as the dependent variable (panel a) and the variable X_1 is regarded as the dependent variable (panel b).

b. The Meaning and Significance of the Correlation Coefficient

There is a nearly irresistible tendency to use the correlation coefficient to imply a causal relationship between the two variables X_1 and X_2. The symmetry of $r_{12}=r_{21}$ shows that this is completely unjustified. The correlation statistic r_{12} does not distinguish which variable is to be considered the dependent variable and which is to be considered the independent variable. But this is the very basis of causality. One says that A causes B, which is very different than B causing A. The correlation coefficient simply measures the relation between the two. That relation could be direct, or result from relations that exist between each variable and additional variables, or simply be a matter of the chance sampling of the data. Consider the following experiment. A scientist sets out to find out how people get from where they live to a popular beach. Researchers are employed to monitor all the approaches to the beach and count the total number of people that arrive on each of a number of days. Say they find the numbers given in Table 8.1.

Table 8.1

Sample Beach Statistics for Correlation Example

DAY	TOTAL # GOING TO THE BEACH	# TAKING THE FERRY	# TAKING THE BUS
1	10000	100	1000
2	20000	200	500
3	5000	50	2000
4	40000	400	250

If one carries out the calculation of the correlation coefficient between the number taking the Ferry and the number of people going to the beach one would get $r_{12}=1$. If the researcher didn't understand the meaning of the correlation coefficient he might be tempted to conclude that all the people who go to the beach take the Ferry. That, of course, is absurd since his own research shows some people taking the bus. However, a correlation between the number taking the bus and the total number of people on the beach would be negative. Should one conclude that people only take the bus when they know nobody else is going to the beach? Of course not. Perhaps most people drive to the beach so that large beach populations cause such congestion so that busses find it more difficult to get there. Perhaps there is no causal connection at all. Can we at least rule out the possibility that the correlation coefficient resulted from the chance sampling? The answer to this question is yes and it makes the correlation coefficient a powerful tool for ascertaining relationships.

We can quantify the interpretation of the correlation coefficient by forming hypotheses as we did with the mono-variant statistical tests and then testing whether the data supports or rejects the hypotheses. Let us first consider the null hypothesis that there is no correlation in the parent population. If this hypothesis is discredited, then the correlation coefficient may be considered significant. We may approach this problem by means of a t-test. Here we are testing the probability of the occurrence of a correlation coefficient r_{12} that is significantly different from zero and

$$t = r_{12}\left(\frac{(n-2)}{1-r_{12}^2}\right) \quad . \tag{8.3.8}$$

The factor of (N-2) in the numerator arises because we have lost two degrees of freedom to the constants of the linear regression line. We can then use equations (8.2.5) to determine the probability that this value of t (and hence r_{12}) would result from chance. This will of course depend on the number of degrees of freedom (in this case N-2) that are involved in the sample. Conversely, one can turn the problem around and find a value of t for a given p and v that one considers significant and that sets a lower limit to the value for r_{12} that would support the hypothesis that r_{12} occurred by chance. For example, say we had 10 pairs of data points which we believed to be related, but we would only accept the probability of a chance occurrence of .1% as being significant. Then solving equation (8.3.8) for r_{12} we get

$$r_{12} = t(v+t^2)^{-½} \quad . \tag{8.3.9}$$

Consulting tables[2] that solve equations (8.2.5) we find the boundary value for t is 4.587 which leads to a minimum value of r = 0.851. Thus, small sample sizes can produce rather large values for the correlation coefficient simply from the chance sampling. Most scientists are very circumspect about moderate values of the correlation coefficient. This probably results from the fact that causality is not guaranteed by the correlation coefficient and the failure of the null hypothesis is not generally taken as strong evidence of significance.

A second hypothesis, which is useful to test, is appraising the extent to which a given correlation coefficient represents the value present in the parent population. Here we desire to set some confidence limits as we did for the mean in section 8.2. If we make the transformation

$$z = ½\,\ell n\,[(1+r_{12})/(1-r_{12})] = \tanh^{-1}(r_{12}) \quad , \tag{8.3.10}$$

then the confidence limits on z are given by

$$\delta z = t_p \sigma_z \ , \tag{8.3.11}$$

where

$$\sigma_z \tilde{} \ [N-(8/3)]^{-1/2} \ . \tag{8.3.12}$$

If for our example of 10 pairs of points we ask what are the confidence limits on a *observed* value of r_{12}=0.851 at the 5% level, we find that t=2.228 and that δz=0.8227. Thus we can expect the value of the parent population correlation coefficient to lie between $0.411<r_{12}<0.969$. The mean of the z distribution is

$$\bar{z} = \tfrac{1}{2}\{ \ell n \ [(1+r_p)/(1-r_p)] + r_p/(N-1)\} \ . \tag{8.3.13}$$

For our example this leads to the best unbiased estimator of $r_p = 0.837$. This nicely illustrates the reason for the considerable skepticism that most scientists have for small data samples. To significantly reduce these limits, σ_z should be reduced at least a factor of three which implies an increase in the sample size of a factor of ten. In general, many scientists place little faith in a correlation analysis containing less than 100 data points for reasons demonstrated by this example. The problem is two-fold. First small sample correlation coefficients must exhibit a magnitude near unity in order for it to represent a statistically significant relationship between the variables under consideration. Secondly, the probability that the correlation coefficient lies near the correlation coefficient of the parent population is small for a small sample. For the correlation coefficient to be meaningful, it must not only represent a relationship in the sample, but also a relationship for the parent population.

c. Correlations of Many Variables and Linear Regression

Our discussion of correlation has so far been limited to two variables and the simple Pearson correlation coefficient. In order to discuss systems of many variables, we shall be interested in the relationships that may exist between any two variables. We may continue to use the definition given in equation (8.3.7) in order to define a correlation coefficient between any two variables X_i and X_j as

$$r_{ij} = \Sigma \ X_i X_j \ / \ N\sigma_i\sigma_j \ . \tag{8.3.14}$$

Certainly the correlation coefficients may be evaluated by brute force after the normal equations of the least square solution have been solved. Given the complete multi-dimensional regression line, the deviations required by equation (8.3.14) could be calculated and the standard deviations of the individual variables obtained. However, as in finding the error of the least square coefficients in chapter 6 (see section 6.3), most of the require work has been done by the time the normal equations have been solved. In equation (6.3.26) we estimated the error of the least square coefficients in terms of parameters generated during the establishment and solution of the normal equations. If we choose to weight the data by the inverse of the experimental errors ε_i, then the errors can be written in terms of the variance of a_j as

$$\sigma^2(a_j) = C_{jj} = \sigma_j^2 \ . \tag{8.3.15}$$

Here C_{jj} is the diagonal element of the inverse matrix of the normal equations. Thus it should not be surprising that the off-diagonal elements of the inverse matrix of the normal equations are the covariances

$$\sigma_{ij}^2 = C_{ij} \quad . \tag{8.3.16}$$

of the coefficients a_i and a_j as defined in section 7.4 [see equation (7.4.9)]. An inspection of the form of equation (7.4.9) will show that much of what we need for the general correlation coefficient is contained in the definition of the covariance. Thus we can write

$$r_{ij} = \sigma_{ij}^2 / \sigma_i \sigma_j \quad . \tag{8.3.17}$$

This allows us to solve the multivariant problems of statistics that arise in many fields of science and investigate the relationships between the various parameters that characterize the problem. Remember that the matrix of the normal equations is symmetric so that the inverse is also symmetric. Therefore we find that

$$r_{ij} = r_{ji} \quad . \tag{8.3.18}$$

Equation (8.3.18) generalizes the result of the simple two variable correlation coefficient that no cause and effect result is implied by the value of the coefficient. A large value of the magnitude of the coefficient simply implies a relationship may exist between the two variables in question. Thus correlation coefficients only test the relations between each set of variables. But we may go further by determining the statistical significance of those correlation coefficients using the t-test and confidence limits given earlier by equations (8.3.8)-(8.3.13).

d Analysis of Variance

We shall conclude our discussion of the correlation between variables by briefly discussing a discipline known as the *analysis of variance*. This concept was developed by R.A. Fisher in the 1920's and is widely used to search for variables that are correlated with one another and to evaluate the reliability of testing procedures. Unfortunately there are those who frequently make the leap between correlation and causality and this is beyond what the method provides. However, it does form the basis on which to search for causal relationships and for that reason alone it is of considerable importance as an analysis technique.

Since its introduction by Fisher, the technique has been expanded in many diverse directions that are well beyond the scope of our investigation so we will only treat the simplest of cases in an attempt to convey the flavor of the method. The name analysis the variance is derived from the examination of the variances of collections of different sets of observed data values. It is generally assumed from the outset that the observations are all obtained from a parent population having a normal distribution and that they are all independent of one another. In addition, we assume that the individual variances of each single observation are equal. We will use the method of least squares in describing the formalism of the analysis, but as with many other statistical methods different terminology is often used to express this venerable approach.

The simplest case involves one variable or "factor", say y_i. Let there be m experiments that each collect a set of n_j values of y. Thus we could form m average values of $\overline{y}$ for each set of values that we shall label $\overline{y}_j$. It is a fair question to ask if the various means $\overline{y}_j$ differ from one another by more than chance. The general approach is not to compare the individual means with one another, but rather to consider the means as a group and determine their variance. We can then compare the variance of the means with the estimated variances of each member within the group to see if that variance departs from the overall variance

of the group by more than we would expect from chance alone.

First we wish to find the maximum likelihood values of these estimates of $\bar{y}_j$ so we shall use the formalism of least squares to carry out the averaging. Lets us follow the notation used in chapter 6 and denote the values of $\bar{y}_j$ that we seek as a_j. We can then describe our problem by stating the equations we would like to hold using equations (6.1.10) and (6.1.11) so that

$$\phi \vec{a} = \vec{y} , \tag{8.3.19}$$

where the non-square matrix ϕ has the rather special and restricted form

$$\phi_{ik} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} n_1 \quad \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \end{pmatrix} n_1 \quad \vdots \quad \begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} n_1 \tag{8.3.20}$$

This matrix is often called the *design matrix* for analysis of variance. Now we can use equation (6.1.12) to generate the normal equations, which for this problem with one variable will have the simple solution

$$a_j = n_j^{-1} \sum_{i=1}^{n_j} y_{ij} \quad . \tag{8.3.21}$$

The over all variance of y will simply be

$$\sigma^2(y) = n^{-1} \sum_{j=i}^{m} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \quad , \tag{8.3.22}$$

by definition, and

$$n = \sum_{j=1}^{m} n_j \quad . \tag{8.3.23}$$

We know from least squares that under the assumptions made regarding the distribution of the y_j's that the a_j's are the best estimate of the value of y_j (i.e.y^0_j), but can we decide if the various values of y^0_j are all equal? This is a typical statistical hypothesis that we would like to confirm or reject. We shall do this by investigating the variances of a_j and comparing them to the over-all variance. This procedure is the source of the name of the method of analysis.

Let us begin by dividing up the over-all variance in much the same way we did in section 8.3a so that

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}\frac{(y_{ij}-\bar{y}_j^{0})^2}{\sigma^2}=\sum_{j=1}^{m}\left[\left(\sum_{i=1}^{n_j}\frac{(y_{ij}-\bar{y}_j)^2}{\sigma^2}\right)+\frac{n_j(\bar{y}_j-\bar{y}_j^{0})^2}{\sigma^2}\right] . \tag{8.3.24}$$

The term on the left is just the sum of square of n_j independent observations normalized by σ^2 and so will follow a χ^2 distribution having n degrees of freedom. This term is nothing more than the total variation of the observations of each experiment set about their true means of the parent populations (i.e. the variance if the true mean weighted by the inverse of the variance of the observed mean). The two terms of the right will also follow the χ^2 distribution function but have n-m and m degree of freedom respectively. The first of these terms is the total variation of the data about the observed sample means while the last term represents the variation of the sample means themselves about their true means. Now define the overall means for the observed data and parent populations to be

$$\left.\begin{aligned}\bar{y}&=\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}y_{ij}=\frac{1}{n}\sum_{j=1}^{m}n_j\bar{y}_j\\ \bar{y}^{0}&=\frac{1}{n}\sum_{j=1}^{m}n_j\bar{y}_j^{0}\end{aligned}\right\} . \tag{8.3.25}$$

respectively. Finally define

$$a_j^{0}\equiv\bar{y}_j^{0}-\bar{y}^{0} , \tag{8.3.26}$$

which is usually called the *effect* of the factor y^0 and is estimated by the least square procedure to be

$$a_j=\bar{y}_j-\bar{y} . \tag{8.3.27}$$

We can now write the last term on the right hand side of equation (8.3.24) as

$$\sum_{j=1}^{m}\frac{n_j(\bar{y}_j-\bar{y}_j^{0})^2}{\sigma^2}=\sum_{j=1}^{m}\frac{n_j(\bar{y}_j-\bar{y}-a_j^{0})^2}{\sigma^2}+\frac{n(\bar{y}-\bar{y}^{0})}{\sigma^2} , \tag{8.3.28}$$

and the first term on the right here is

$$\sum_{j=1}^{m}\frac{n_j(\bar{y}_j-\bar{y}-a_j^{0})^2}{\sigma^2}=\sum_{j=1}^{m}\frac{n_j(a_j-a_j^{0})^2}{\sigma^2} , \tag{8.3.29}$$

and the definition of α_j allows us to write that

$$\sum_{j=1}^{m}a_j=0 . \tag{8.3.30}$$

However, should any of the α_j^0's not be zero, then the results of equation (8.3.29) will not be zero and the assumptions of this derivation will be violated. That basically means that one of the observation sets does not sample a normal distribution or that the sampling procedure is flawed.

We may determine if this is the case by considering the distribution of the first term on the right hand side of equation (8.3.28). Equation (8.3.28) represents the further division of the variation of the first term on the right of equation (8.3.24) into two new terms. This term was the total variation of the

observations about their sample means and so would follow a χ^2-distribution having n-m degrees of freedom. As can be seen from equation (8.3.29), the first term on the right of equation (8.3.28) represents the variation of the sample effects about their true value and therefore should also follow a χ^2-distribution with m-1 degrees of freedom. Thus, if we are looking for a single statistic to test the assumptions of the analysis, we can consider the statistic

$$Q = \left[\sum_{j=1}^{m} n_j(\bar{y}_j - \bar{y})/(m-1)\right] \Big/ \left[\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2/(n-m)\right] , \tag{8.3.31}$$

which, by virtue of being the ratio of two terms having χ^2-distributions, will follow the distribution of the F-statistic and can be written as

$$Q = \frac{(n-m)\sum_{j=1}^{m}(n_j\bar{y}_j^2 - n\bar{y})/(m-1)}{\left[\sum_{j=1}^{m}\sum_{i=1}^{n_j} y_{ij}^2\right] - \sum_{j=1}^{m} n_j\bar{y}_j^2} . \tag{8.3.32}$$

Thus we can test the hypothesis that all the effects α_j^0 are zero by comparing the results of calculating Q[(n-m),(m-1)] with the value of F expected for any specified level of significance. That is, if $Q>F_c$, where F_c is the value of F determined for a particular level of significance, then one knows that the α_j^0s are not all zero and at least one of the sets of observations is flawed.

In development of the method for a single factor or variable, we have repeatedly made use of the additive nature of the variances of normal distributions [i.e. equations (8.3.24) and (8.3.28)]. This is the primary reason for the assumption of "normality" on the parent population and forms the foundation for analysis of variance. While this example of an analysis of variance is for the simplest possible case where the number of "factors" is one, we may use the technique for much more complicated problems employing many factors. The philosophy of the approach is basically the same as for one factor, but the specific formulation is lengthy and beyond the scope of this book.

This just begins the study of correlation analysis and the analysis of variance. We have not dealt with multiple correlation, partial correlation coefficients, or the analysis of covariance. All are of considerable use in exploring the relationship between variables. We have again said nothing about the analysis of grouped or binned data. The basis for analysis of variance has only been touched on and the testing of nonlinear relationships has not been dealt with at all. We will leave further study in these areas to courses specializing in statistics. While we have discussed many of the basic topics and tests of statistical analysis, there remains one area to which we should give at least a cursory look.

8.4 The Design of Experiments

In the last section we saw how one could use correlation techniques to search for relationships between variables. We dealt with situations where it was even unclear which variable should be regarded as the dependent variable and which were the independent variables. This is a situation unfamiliar to the

physical scientist, but not uncommon in the social sciences. It is the situation that prevails whenever a new phenomenology is approached where the importance of the variables and relationships between them are totally unknown. In such situations statistical analysis provides the only reasonable hope of sorting out and identifying the variables and ascertaining the relationships between them. Only after that has been done can one begin the search for the causal relationships which lead to an understanding upon which theory can be built.

Generally, physical experimentation sets out to test some theoretical prediction and while the equipment design of the experiment may be extremely sophisticated and the interpretation of the results subtle and difficult, the philosophical foundations of such experiments are generally straightforward. Where there exists little or no theory to guide one, experimental procedures become more difficult to design. Engineers often tread in this area. They may know that classical physics could predict how their experiments should behave, but the situation may be so complex or subject to chaotic behavior, that actual prediction of the outcome is impossible. At this point the engineer will find it necessary to search for relationships in much the same manner as the social scientist. Some guidance may come from the physical sciences, but the final design of the experiment will rely on the skill and wisdom of the experimenter. In the realm of medicine and biology theoretical description of phenomena may be so vague that one should even relax the term variable which implies a specific relation to the result and use the term "*factor*" implying a parameter that may, or may not, be relevant to the result. Such is the case in the experiments we will be describing.

Even the physical sciences, and frequently the social and biological sciences undertake surveys of phenomena of interest to their disciplines. A survey, by its very nature, is investigating factors with suspected but unknown relationships and so the proper layout of the survey should be subject to considerable care. Indeed, Cochran and Cox[5] have observed

> *"Participation in the initial stages of an experiment in different areas of research leads to the strong conviction that too little time and effort is put into the planning of experiments. The statistician who expects that his contribution to the planning will involve some technical matter in statistical theory finds repeatedly that he makes a much more valuable contribution simply by getting the investigator to explain clearly why he is doing the experiment, to justify experimental treatments whose effects he expects to compare and to defend his claim that the completed experiment will enable his objectives to be realized. ..."*

Therefore, it is appropriate that we spend a little time discussing the language and nature of experimental design.

At the beginning of chapter 7, we drew the distinction between data that were obtained by observation and those obtained by experimentation. Both processes are essentially *sampling* a parent population. Only in the latter case, does the scientist have the opportunity to partake in the specific outcome. However, even the observer can arrange to carry out a well designed survey or a badly designed survey by choosing the nature and range of variables or factors to be observed and the equipment with which to do the observing.

The term experiment has been defined as "a considered course of action aimed at answering one or more carefully framed questions". Therefore any experiment should meet certain criteria. It should have a

specific and well defined mission or objective. The list of relevant variables, or factors, should be complete. Often this latter condition is difficult to manage. In the absence of some theoretical description of the phenomena one can imagine that a sequence of experiments may be necessary simply to establish what are the relevant factors. As a corollary to this condition, every attempt should be made to exclude or minimize the effect of variables beyond the scope or control of the experiment. This includes the bias of the experimenters themselves. This latter consideration is the source of the famous "double-blind" experiments so common in medicine where the administers of the treatment are unaware of the specific nature of the treatment they are administrating at the time of the experiment. Which patients received which medicines is revealed at a later time. Astronomers developed the notion of the "personal equation" to attempt to allow for the bias inadvertently introduced by observers where personal judgement is required in making observations. Finally the experiment should have the internal precision necessary to measure the phenomena it is investigating. All these conditions sound like "common sense", but it is easy to fail to meet them in specific instances. For example, we have already seen that the statistical validity of any experiment is strongly dependent on the number of degrees of freedom exhibited by the sample. When many variables are involved, and the cost of sampling the parent population is high, it is easy to short cut on the sample size usually with disastrous results.

While we have emphasized the two extremes of scientific investigation where the hypothesis is fully specified to the case where the dependency of the variables is not known, the majority of experimental investigations lie somewhere in between. For example, the quality of milk in the market place could depend on such factors as the dairies that produce the milk, the types of cows selected by the farmers that supply the dairies, the time of year when the milk is produced, supplements used by the farmers, etc. Here causality is not firmly established, but the order of events is so there is no question that the quality of the milk determines the time of year, but the relevance of the factors is certainly not known. It is also likely that there are other unspecified factors that may influence the quality of the milk that are inaccessible to the investigator. Yet, assuming the concept of milk quality can be clearly defined, it is reasonable to ask if there is not some way to determine which of the known factors affect the milk quality and design an experiment to find out. It is in these middle areas that experimental design and techniques such as analysis of variance are of considerable use.

The design of an experiment basically is a program or plan for the manner in which the data will be sampled so as to meet the objectives of the experiment. There are three general techniques that are of use in producing a well designed experiment. First, data may be grouped so that unknown or inaccessible variables will be common to the group and therefore affect all the data within the group in the same manner. Consider an experiment where the one wishes to determine the factors that influence the baking of a type of bread. Let us assume that there exists an objective measure of the quality of the resultant loaf. We suspect that the oven temperature and duration of baking are relevant factors determining the quality of the loaf. It is also likely that the quality depends on the baker mixing and kneading the loaf. We could have all the loaves produced by all the bakers at the different temperatures and baking times measured for quality without keeping track of which baker produced which loaf. In our subsequent analysis the variations introduced by the different bakers would appear as variations attributed to temperature and baking time reducing the accuracy of our test. But the simple expedient of grouping the data according to each baker and separately analyzing the group would isolate the effect of variations among bakers and increase the accuracy of the experiment regarding the primary factors of interest.

Second, variables which cannot be controlled or "blocked out" by grouping the data should be

reduced in significance by randomly selecting the sampled data so that the effects of these remaining variables tend to cancel out of the final analysis. Such randomization procedures are central to the design of a well-conceived experiment. Here it is not even necessary to know what the factors may be, only that their effect can be reduced by randomization. Again, consider the example of the baking of bread. Each baker is going to be asked to bake loaves at different temperatures and for varying times. Perhaps as the baker bakes more and more bread fatigue sets in affecting the quality of the dough he produces. If each baker follows the same pattern of baking the loaves (i.e. all bake the first loaves at temperature T_1 for a time t_1 etc.) then systematic errors resulting from fatigue will appear as differences attributable to the factors of the experiment. This can be avoided by assigning random sequences of time and temperature to each baker. While fatigue may still affect the results, it will not be in a systematic fashion.

Finally, in order to establish that the experiment has the precision necessary to answer the questions it poses, it may be necessary to repeat the sampling procedure a number of times. In the parlance of statistical experiment design the notion of repeating the experiment is called *replication* and can be used to help achieve proper randomization and well as establish the experimental accuracy.

Thus the concepts of data grouping, randomization and repeatability or replication are the basic tools one has to work with in designing an experiment. As in other areas of statistics, a particular jargon has been developed associated with experiment design and we should identify these terms and discuss some of the basic assumptions associated with experiment design.

a. The Terminology of Experiment Design

Like many subjects in statistics, the terminology of experiment design has its origin in a subject where statistical analysis was developed for the specific analysis of the subject. As the term *regression analysis* arose form studies in genetics, so much of experimental design formalism was developed for agriculture. The term *experimental area* used to describe the scope or environment of the experiment was initially a area of land on which an agricultural experiment was to be carried out. The terms *block* and *plot* meant subdivisions of this area. Similarly the notion of a *treatment* is known as a *factor* in the experiment and is usually the same as what we have previously meant by a variable. A *treatment level* would then refer to the value of the variable. (However, remember the caveats mentioned above relating to the relative role of variables and factors.) Finally the term *yield* was just that for an agricultural experiment. It was the results of a treatment being applied to some plot. Notice that here there is a strong causal bias in the use of the term yield. For many experiments this need not be the case. One factor may be chosen as the yield, but its role as dependent variable can be changed during the analysis. Perhaps a somewhat less prejudicial term might be *result*.

All these terms have survived and have taken on very general meanings for experiment design. Much of the mystery of experiment design is simply relating the terms of agricultural origin to experiments set in far different contexts. For example, the term *factorial experiment* refers to any experiment design where the levels (values) of several factors (i.e. variables) are controlled at two or more levels so as to investigate their effects on one another. Such an analysis will result in the presence of terms involving each factor in combination with the remaining factors. The expression of the number of combinations of n thing taken m at a time does involve factorials [see equation (7.2.4)] but this is a slim excuse for calling such

systems "factorial designs". Nevertheless, we shall follow tradition and do so.

Before delving into the specifics of experiment designs, let us consider some of the assumptions upon which their construction rests. Underlying any experiment there is a model which describes how the factors are assumed to influence the result or yield. This is not a full blown detailed equation such as the physical scientist is used to using to frame a hypothesis. Rather, it is a statement of additivity and linearity. All the factors are assumed to have a simple proportional effect on the result and the contribution of all factors is simply additive. While this may seem, and in some cases may be, an extremely restrictive assumption, it is the simplest non-trivial behavior and in the absence of other information provides a good place to begin any investigation. In the last section we divided up the data for an analysis of variance into sets of experiments each of which contained individual data entries. For the purposes of constructing a model for experiment design we will similiarly divide the observed data so that i represents the treatment level, and j represents the block containing the factor, and we may need a third subscript to denote the order of the treatment within the block. We could then write the mathematical model for such an experiment as

$$y_{ij\,k} = \langle y \rangle + f_i + b_j + \varepsilon_{ij\,k} \ . \tag{8.4.1}$$

Here $y_{ij\,k}$ is the yield or results of the ith treatment or factor-value contained in the jth block subject to an experimental error $\varepsilon_{i\,j\,k}$. The auusmption of additivity means that the block effect b_j will be the same for all treatments within the same block so that

$$y_{1j\,k_1} - y_{2j\,k_2} = f_1 - f_2 + \varepsilon_{1j\,k_1} - \varepsilon_{2\,j\,k_2} \ . \tag{8.4.2}$$

In addition, as was the case with the analysis of variance it is further assumed that the errors $\varepsilon_{i\,j\,k}$ are normally distributed.

By postulating a linear relation between the factors of interest and the result, we can see that only two values of the factors would be necessary to establish the dependence of the result on that factor. Using the terminology of experiment design we would say that only two treatment levels are necessary to establish the effect of the factor on the yield. However, we have already established that the order in which the treatments are applied should be randomized and that the factors should be grouped or blocked in some rational way in order for the experiment to be well designed. Let us briefly consider some plans for the acqusition of data which constitute an experiment design.

b. Blocked Designs

So far we have studiously avoided discussing data that is grouped in bins or ranks etc. However, the notion is central to experiment design so we will say just enough about the concept to indicate the reasons for involving it and indicate some of the complexities that result. However, we shall continue to avoid discussing the statistical analysis that results from such groupings of the data and refer the student to more complete courses on statistics. To understand the notion of grouped or blocked data, it is useful to return to the agricultural origins of experiment design.

If we were to design an experiment to investigate the effects of various fertilizers and insecticides on the yield of a particular species of plant, we would be foolish to treat only one plant with a particular

combination of products. Instead, we would set out a block or plot of land within the experimental area and treat all the plants within that block in the same way. Presumably the average for the block is a more reliable measure of the behavior of plants to the combination of products than the results from a single plant. The data obtained from a single block would then be called grouped data or blocked data. If we can completely isolate a non-experimental factor within a block, the data can be said to be *completely blocked* with respect to that data. If the factor cannot be completely isolated by the grouping, the data is said to be *incompletely blocked*. The subsequent statistical analysis for these different types of blocking will be different and is beyond the scope of this discussion.

Now we must plan the arrangements of blocks so that we cover all combinations of the factors. In addition, we would like to arrange the blocks so that variables that we can't allow for have a minimal influence on our result. For example, soil conditions in our experimental area are liable to be similar for blocks that are close together than for blocks that are widely separated. We would like to arrange the blocks so that variations in the field conditions will affect all trials in a random manner. This is similiar to our approach with the bread where having the bakers follow a random sequence of allowed factors (i,e, T_i, and t_j) was used to average out fatgue factors. Thus randomization can take place in a time sequence as well as a spatial layout. This will tend to minimize the effects of these unknown variables.

The reason this works is that if we can group our treatments (levels or factor values) so that each factor is exposed to the same unspecified influence in a random order, then the effects of that influence should tend to cancel out over the entire run of the experiment. Unfortunately one pays a price for the grouping or blocking of the experimental data. The arrangement of the blocks may introduce an effect that appears as an interaction between the factors. Usually it is a high level interaction and it is predictable from the nature of the design. An interaction that is liable to be confused with an effect arising strictly from the arrangement of the blocks is said to be *confounded* and thus can never be considered as significant. Should that interaction be the one of interest, then one must change the design of the experiment. Standard statistical tables[2] give the arrangements of factors within blocks and the specific interactions that are confounded for a wide range of the number of blocks and factors for two treatment-level experiments.

However, there are other ways of arranging the blocks or the taking of the data so that the influence of inaccessible factors or sources of variation are reduced by randomization. By way of example consider the agricultural situation where we try to minimize the systematic effects of the location of the blocks. One possible arrangement is known as a *Latin square* since it is a square of Latin letters arranged in a specific way. The rule is that no row or column shall contain any particular letter more than once. Thus a 3×3 Latin square would have the form:

$$\begin{pmatrix} ABC \\ BCA \\ CAB \end{pmatrix} .$$

Let the Latin letters A, B, and C represent three treatments to be investigated. Each row and each column represents a complete experiment (i.e. replication). Thus the square symbolically represents a way of randomizing the order of the treatments within each replication so that variables depending on the order are averaged out. In general, the rows and columns represent two variables that one hopes to eliminate by randomization. In the case of the field, they are the x-y location within the field and the associated soil

variations etc. In the case of the baking of bread, the two variables could have been the batch of flour and time. The latter would then eliminate the fatigue factor which was a concern. Should there have been a third factor, we might have used a Greco-Latin square where a third dimension is added to the square by the use of Greek subscripts so that the arrangement becomes:

$$\begin{pmatrix} A_\alpha B_\delta C_\beta \\ B_\beta C_\alpha A_\delta \\ C_\delta A_\beta B_\alpha \end{pmatrix} .$$

Here the three treatments are grouped into replicates in three different ways with the result three sources of variation can be averaged out.

A Latin or Greco-Latin square design is restrictive in that it requires that the number of "rows" and "columns" corresponding to the two unspecified systematic parameters, be the same. In addition, the number of levels or treatments must equal the number of rows and columns. The procedure for use of such a design is to specify a trial by assigning the levels to the letters randomly and then permuting the rows and columns of the square until all trials are completed. One can find larger squares that allow for the use of more treatments or factors in books on experiment design[6] or handbooks of statistics[7]. These squares simply provide random arrangements for the application of treatments or the taking of data which will tend to minimize the effects of phenomena or sources of systematic error which cannot be measures, but of which the experimenter is aware. While their use may increase the amount of replication above the minimum required by the model, the additional effort is usually more than compensated by the improvement in the accuracy of the result.

While the Latin and Greco-Latin squares provide a fine design for randomizing the replications of the experiment, they are by no means the only method for doing so. Any reasonable modern computer will provide a mechanism for generating random numbers which can be used to design the plan for an experiment. However, one must be careful about the confounding between blocked data that can result in any experiment and be sure to identify those regions of the experiment in which it is likely to occur.

c. Factorial Designs

As with all experimental designs, the primary purpose of the factorial design is to specify how the experiment is to be run and the data sampling carried out. The main purpose of this protocol is to insure that all combinations of the factors (variables) are tested at the required treatment levels (values). Thus the basic model for the experiment is somewhat different from that suggested by equations (8.4.1) and (8.4.2). One looks for *effects* which are divided into *main effects* on the yield (assumed dependent variable) resulting from changes in the level of a specific factor, and *interaction effects* which are changes in the yield that result from the simultaneous change of two or more factors. In short, one looks for correlations between the factors and the yield and between the factors themselves. An experiment that has n factors each of which is allowed to have m levels will be required to have m^n trials or replications. Since most of the statistical analysis that is done on such experimental data will assume that the relationships are linear, m is usually taken to be two. Such an experiment would be called a 2^n *factorial experiment*. This simply means that it is an experiment with n-factors requires 2^n trials.

A particularly confusing notation is used to denote the order and values of the factors in the

experiment. While the factors themselves are denoted by capital letters with subscripts starting at *zero* to denote their level (i.e. A_0, B_1, C_0, etc.), a particular trial is given a combination of lower case letters. If the letter is present it implies that the corresponding factor has the value with the subscript 1. Thus a trial where the factors A,B, and C have the values A_0, B_1, and C_1 would be labeled simply bc. A special representation is reserved for the case A_0, B_0, C_0, where by convention nothing would appear. The symbology is that this case is represented by (1). Thus all the possible combinations of factors which give rise to the interaction effects requiring the 2^n trials for a 2^n factorial experiment are given in Table 8.2

Table 8.2

Factorial Combinations for Two-level Experiments with n = 2 $\rightarrow$ 4

NO. OF LEVELS	COMBINATIONS OF FACTORS IN STANDARD NOTATION
2 factors	(1), a, b, ab
3 factors	(1), a, b, ab, c, ac, bc, abc
4 factors	(1), a, b, ab, c, ac, bc, abc, d, ad, bd, cd, acd, bcd, abcd.

Tables[2] exist of the possible combinations of the interaction terms for any number of factors and reasonable numbers of treatment-levels.

As an example, let us consider the model for two factors each having the two treatments (i.e. values) required for the evaluation of linear effects

$$y_i = \langle y \rangle + a_i + b_i + a_i b_i + \varepsilon_i \quad . \tag{8.4.3}$$

The subscript $_i$ will take on values of 0 and 1 for the two treatments given to a and b. Here we see that the cross term ab appears as an additional unknown. Each of the factors A and B will have a main effect on y. In addition the cross term AB which is known as the interaction term, will produce an interaction effect. These represent three unknowns that will require three independent pieces of information (i.e. trials, replications, or repetitions) for their specification. If we also require the determination of the grand mean <y> then an additional independent piece of information will be needed bringing the total to 2^2. In order to determine all the cross terms arising from an increased number of factors many more independent pieces of information are needed. This is the source of the 2^n required number of trials or replications given above. In carrying out the trials or replications required by the factorial design, it may be useful to make use of the blocked data designs including the Latin and Greco-latin squares to provide the appropriate randomization which reduces the effect of inaccessible variables.

There are additional designs which further minimize the effects of suspected influences and allow more flexibility in the number of factors and levels to be used, but they are beyond the scope of this book. The statistical design of an experiment is extremely important when dealing with an array of factors or variables whose interaction is unpredictable from theoretical considerations. There are many pitfalls to be

encountered in this area of study which is why it has become the domain of specialists. However, there is no substitute for the insight and ingenuity of the researcher in identifying the variables to be investigated. Any statistical study is limited in practice by the sample size and the systematic and unknown effects that may plague the study. Only the knowledgeable researcher will be able to identify the possible areas of difficulty. Statistical analysis may be able to confirm those suspicions, but will rarely find them without the foresight of the investigator. Statistical analysis is a valuable tool of research, but it is not meant to be a substitute for wisdom and ingenuity. The user must also always be aware that it is easy to phrase statistical inference so that the resulting statement says more than is justified by the analysis. Always remember that one does not "prove" hypotheses by means of statistical analysis. At best one may reject a hypothesis or add confirmatory evidence to support it. But the sample population is not the parent population and there is always the chance that the investigator has been unlucky.

Chapter 8 Exercises

1. Show that the variance of the t-probability density distribution function given by equation (8.1.2) is indeed σ^2_t as given by equation (8.1.3).

2. Use equation (8.1.7) to find the variance, mode , and skewness of the χ^2-distribution function. Compare your results to equation (8.1.8).

3. Find the mean, mode and variance of the F-distribution function given by equation (8.1.11).

4. Show that the limiting relations given by equations (8.1.13) - (8.1.15) are indeed correct.

5. Use the numerical quadrature methods discussed in chapter 4 to evaluate the probability integral for the t-test given by equation (8.2.5) for values of p=.1, 0.1, 0.01, and N=10, 30, 100. Obtain values for t_p and compare with the results you would obtain from equation (8.2.6).

6. Use the numerical quadrature methods discussed in chapter 4 to evaluate the probability integral for the χ^2-test given by equation (8.2.8) for values of p=.1, 0.1, 0.01, and N=10, 30, 100. Obtain values for χ^2_p and compare with the results you would obtain from using the normal curve for the χ^2-probability density distribution function.

7. Use the numerical quadrature methods discussed in chapter 4 to evaluate the probability integral for the F-test given by equation (8.2.9) for values of p=.1, 0.1, 0.01, N_1=10, 30, 100, and N_2=1, 10, 30. Obtain values for F_p.

8. Show how the various forms of the correlation coefficient given by equation (8.3.7) can be obtained from the definition given by the second term on the left.

9. Find the various values of the 0.1% marginally significant correlation coefficients when n= 5, 10, 30, 100, 1000.

10. Find the correlation coefficient between X_1 and Y_1, and Y_1 and Y_2 in problem 4 of chapter 6.

11. Use the F-test to decide when you have added enough terms to represent the table given in problem 3 of chapter 6.

12. Use analysis of variance to show that the data in Table 8.1 imply that taking the bus and taking the ferry are important factors in populating the beach.

13. Use analysis of variance to determine if the examination represented by the data in Table 7.1 sampled a normal parent population and at what level of confidence on can be sure of the result.

14. Assume that you are to design an experiment to find the factors that determine the quality of bread baked at 10 different bakeries. Indicate what would be your central concerns and how you would go about addressing them. Identify four factors that are liable to be of central significance in determining the quality of bread. Indicate how you would design an experiment to find out if the factors are indeed important.

Chapter 8 References and Supplemental Reading

1 Croxton, F.E., Cowden, D.J., and Klein, S., "Applied General Statistics", (1967), Prentice-Hall, Inc., Englewood Cliffs, N.J.

2. Weast, R.C., "CRC Handbook of Tables for Probability and Statistics", (1966), (Ed. W.H.Beyer), The Chemical Rubber Co. Cleveland.

3. Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T., "Numerical Recipies the art of scientific computing" (1986), Cambridge University Press, Cambridge.

4. Smith, J.G., and Duncan, A.J., "Sampling Statistics and Applications: Fundementals of the Theory of Statistics", (1944), McGraw-Hill Book Company Inc., New York, London, pp.18.

5. Cochran , W.G., and Cox, G.M., "Experimental Designs" (1957) John Wiley and Sons, Inc., New York, pp 10.

6. Cochran , W.G., and Cox, G.M., "Experimental Designs" (1957) John Wiley and Sons, Inc., New York, pp 145-147.

7. Weast, R.C., "CRC Handbook of Tables for Probability and Statistics", (1966), (Ed. W.H.Beyer), The Chemical Rubber Co. Cleveland, pp63-65.