USGS
science for a changing world

Techniques of Water-Resources Investigations
of the United States Geological Survey

Chapter B4

REGRESSION MODELING OF GROUND-WATER FLOW

By Richard L. Cooley and Richard L. Naff

Book 3
APPLICATIONS OF HYDRAULICS

distribution. The density of the normal distribution is a bell-shaped curve, symmetric about its mean μ_ε, and with most of the mass concentrated within one standard deviation σ_ε of the mean (see figure 2.2-10). In the case of the titration experiment, we would hope that the most frequently found value of the error would be near-zero and expect that μ_ε would equal zero. The standard deviation σ_ε is a measure of the dispersion, or spread, of the errors about the mean and is equal to the distance from the mean to an inflection point on the curve $f(\epsilon)$. The mean and standard deviation will be formally defined in a later section.

A normal random variable is frequently standardized with its mean and standard deviation by the following transformation:

$$Z=(\mathcal{E}-\mu_\varepsilon)/\sigma_\varepsilon \quad . \tag{2.2-15}$$

The cumulative distribution for this standard normal random variable is tabulated (table 2.10-1) for use by the investigator, since its probability density function, $f_Z(z)$, is parameter free:

$$f_Z(z)=\frac{e^{-z^2/2}}{\sqrt{2\pi}} \quad . \tag{2.2-16}$$

Given the density function for the standard normal random variable, it is natural to inquire about the form of density, $f_\varepsilon(\epsilon)$, of the unnormalized random variable $\mathcal{E}$. Consider the cumulative frequency distribution for Z. By making the change of variables $z=(s-\mu_\varepsilon)/\sigma_\varepsilon$,

$$F_Z(a)=\int_{-\infty}^{a}\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz$$

$$=\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\int_{-\infty}^{\epsilon}\exp\left[-\left(\frac{s-\mu_\varepsilon}{\sigma_\varepsilon}\right)^2/2\right]ds \tag{2.2-17}$$

results where $\epsilon=a\sigma_\varepsilon+\mu_\varepsilon$ is a value of the unnormalized random variable. Since differentiation is the inverse operator of integration, equation 2.2-17 is differentiated with respect to ϵ to find $f_\varepsilon(\epsilon)$ (see also equation 2.2-8):

$$f_\varepsilon(\epsilon)=\frac{d}{d\epsilon}F_Z(a(\epsilon))$$

$$=\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\exp\left[-\left(\frac{\epsilon-\mu_\varepsilon}{\sigma_\varepsilon}\right)^2/2\right]. \tag{2.2-18}$$

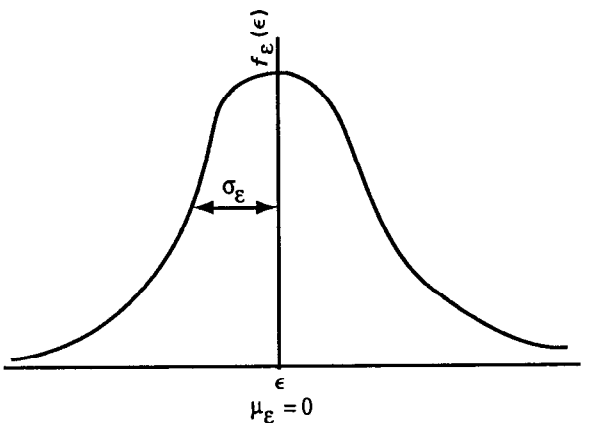

Figure 2.2-10

Note that equation 2.2-18 is not parameter free, as this density is a function of the parameters μ_ε and σ_ε.

2.3 Expectation and the Continuous Random Variable

The discussion in this section is largely presented with continuous random variables in mind. All the results, however, are applicable to discrete random variables; whenever a quantity is defined by an integration over a probability density function for the continuous case, this same quantity can almost invariably be defined by a summation over the discrete density function for the discrete case. The reader should demonstrate the veracity of this statement.

2.3.1 The Mean

The mean is a measure of central tendency of a population. As an estimator of this central tendency, consider a finite random sample consisting of n values x_i of the random variable X. If the sample frequency of occurrence f_i^* is estimated from this random sample, then a logical estimator of the central tendency is to sum the product of the central value $\bar{x}_i$ of each class interval and the frequency of occurrence for that interval:

$$\bar{x}=\sum_{i\le x_m/\Delta x} f_i^*\,\bar{x}_i=\sum_{i\le x_m/\Delta x} f_i\,\bar{x}_i\,\Delta x \tag{2.3-1}$$

where x_m is the upper limit of the largest class interval necessary to construct f_i^*. The frequencies of occurrence f_i^* in equation 2.3-1 can be looked upon as weights that sum to one, and the quantities $\bar{x}_i$ as equally spaced values of the random variable. The values of the random variable that occur more frequently, as indicated by the random sample, receive larger weights through equation 2.3-1 and will have a greater influence on $\bar{x}$.

Equation 2.3-1 should be recognized by the reader as also being the definition for the center of mass of physical weights distributed along a line. That is, if mf_i^* represents the mass of a weight located at $\bar{x}_i$, where m is the total mass of all the weights, then equation 2.3-1 would give us the center of mass of the line with respect to the origin. In the case of a histogram, the role of the weights is played by the sample frequency of occurrence for an interval, which gives us the approximate relative likelihood that any future value of the random variable will occur in that interval. For calculation purposes, this distributed weight over any interval i is replaced by a point weight having the same mass as the distributed weight, but located at the center $\bar{x}_i$ of the interval. The sum of the products of the relative masses of these point weights, f_i^*, with their relative distances from the origin, $\bar{x}_i$, gives us the center of mass, which is also a measure of the central tendency. Of course, if the sample size n were to become very large, then Δx could be made very small, refining equation 2.3-1 as an estimator of the central tendency of a random variable.

Reasoning similar to that leading to $\bar{x}$ as an estimator of the population mean can be applied directly to defining this parameter. First, given that the density function $f(x)$ is known, then the approximate frequency of occurrence of an event corresponding to an interval of size Δx that has as its central value $\bar{x}_i$ is $f(\bar{x}_i)\Delta x$. Thus, assuming that these relative frequencies are centered at each $\bar{x}_i$, an approximate measure of the population central tendency, μ_X, is

$$\mu_X \approx \sum_{(\text{all } i)} \bar{x}_i f(\bar{x}_i)\,\Delta x \qquad (2.3\text{-}2)$$

where the values $\bar{x}_i$ are equally spaced by Δx from each other. Of course, by letting Δx become smaller, a more accurate measure of μ_X is developed, until μ_X, also known as the expected value, $E[X]$, of the random variable X, is defined by the following integral expression:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f(x)\,dx \quad . \qquad (2.3\text{-}3)$$

This equation is the standard form for the expected value of a univariate random variable.

Equation 2.3-3 can also be developed directly from equation 2.3-1 by letting $n \to \infty$ and $\Delta x \to 0$:

$$\mu_X = \lim_{\substack{n\to\infty \\ \Delta x\to 0}} \bar{x} = \lim_{\substack{n\to\infty \\ \Delta x\to 0}} \sum_{i \le x_m/\Delta x} \bar{x}_i f_i\,\Delta x = \int_{-\infty}^{\infty} x f(x)\,dx. \qquad (2.3\text{-}4)$$

That is, as Δx becomes smaller and as the number of observations becomes very large, $\bar{x}_i$ becomes a unique continuous value of X, f_i becomes the continuous function $f(x)$, and the summation can be replaced by an integration. As in the case of equation 2.2-7, we can only say that the limit indicated in equation 2.3-4 is reached with a very high probability as n becomes large; however, this probability should be unity as n becomes infinite.

Note that μ_X is a population parameter that is characteristic of the random variable X, while $\bar{x}$, being derived from values of a finite random sample from the population of X, is only an estimate for μ_X. Estimators such as $\bar{x}$ will be developed in greater detail in a later section.

Problem 2.3-1

a. Find $\bar{x}$ from 100 μmho/cm histogram of problem 2.2-2.
b. Find μ_X for the random variable of problem 2.2-1.

2.3.2 Generalization and Application of the Expectation Operator

The operation of finding an expected value can be generalized by considering a function $g(X)$ of continuous random variable X. If we

wish to find the average effect of the function $g(x)$ over the outcomes of a random sampling, we again resort to the approximation

$$\overline{g(x)}=\sum_{i\leq x_m/\Delta x} f_i^* g(\bar{x}_i) \ . \qquad (2.3\text{-}5)$$

That is, we weight $g(x)$, where $g(x)$ is evaluated at the center of every class interval, by the frequency of occurrence of that interval and sum all the weighted values of $g(\bar{x}_i)$. To obtain the population equivalent of $\overline{g(x)}$, n is taken to be large, while Δx is taken to be small; this equivalent is denoted by the expectation symbol $E[g(X)]$:

$$E[g(X)]=\lim_{\substack{n\to\infty \\ \Delta x\to 0}} \overline{g(x)}=\int_{-\infty}^{\infty} g(x)f(x)dx \ . \qquad (2.3\text{-}6)$$

Equation 2.3-6 represents the general form of the expectation operator for a univariate distribution when the random variable is continuous. A similar form exists for discrete random variables, in which the integration has been replaced by summation.

A trivial but useful property of the expectation operator is that the expected value of any constant c is that constant; for the continuous case, this is easily demonstrated as

$$E[c]=\int_{-\infty}^{\infty} cf(x)dx=c\int_{-\infty}^{\infty} f(x)dx=c \qquad (2.3\text{-}7)$$

where equation 2.2-11 has been invoked. A more important property of E is that it is a linear operator; that is,

$$E[ag_1(X)+bg_2(X)]=aE[g_1(X)]+bE[g_2(X)]. \qquad (2.3\text{-}8)$$

This property results because in the continuous case, integration itself is a linear operator:

$$\int_{-\infty}^{\infty}(ag_1(x)+bg_2(x))f(x)dx=a\int_{-\infty}^{\infty} g_1(x)f(x)dx + b\int_{-\infty}^{\infty} g_2(x)f(x)dx \qquad (2.3\text{-}9)$$

As a practical example of finding the expected value of a random variable, consider the problem of finding the mean of X where X is a normal random variable with mean μ_X and standard deviation σ_X:

$$E[X]=\frac{1}{\sqrt{2\pi}\sigma_X}\int_{-\infty}^{\infty} x\exp\left[-\left(\frac{x-\mu_X}{\sigma_X}\right)^2/2\right]dx. \qquad (2.3\text{-}10)$$

By a change of variable $z=(x-\mu_X)/\sigma_X$, we see that equation 2.3-10 becomes

$$E[X]=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\mu_X+\sigma_X z)e^{-z^2/2}dz=\mu_X, \qquad (2.3\text{-}11)$$

because $z\cdot\exp(-z^2/2)$ is an odd function of z in the interval $(-\infty,\infty)$, and equation 2.3-7 holds (μ_X, being a population parameter, is constant). Equation 2.3-11 is the reason why μ_X is defined to be $E[X]$.

2.3.3 The Variance, Standard Deviation, and Coefficient of Variation

Although the mean μ_X is a measure of the central tendency of a random variable, it gives no information as to how frequently a random variable will be encountered in its vicinity. The variance σ_X^2, defined as the expected value of the function $g(X)=(X-\mu_X)^2$, is a population parameter that quantifies this concept. The variance can also be looked upon as an operator that is defined in terms of another operator (the expectation operator) as follows:

$$\sigma_X^2=\mathrm{Var}[X]=E[(X-\mu_X)^2] \ ,$$
$$=\int_{-\infty}^{\infty}(x-\mu_X)^2f(x)dx \qquad (2.3\text{-}12)$$

where Var[X] represents an operator that operates on X. The intuitive sense of σ_X^2 is that it is the sum of the frequency weighted deviations, which have been squared, from the mean. As such, σ_X^2 represents the amount of dispersion of the random variable about the mean: when σ_X^2 is relatively large, then a random variable is less likely to have values in the

immediate vicinity of the mean. The standard deviation σ_X is simply the square root of the variance: $\sigma_X=(\mathrm{Var}[X])^{1/2}$.

By exercising the linear property of the expectation operator, equation 2.3-12 can be expressed in an alternate form:

$$\sigma_X^2=\mathrm{Var}[X]=E[(X-\mu_X)^2]=E[X^2-2X\mu_X+\mu_X^2],$$
$$=E[X^2]-\mu_X^2. \qquad (2.3\text{-}13)$$

The variance operator, like the expectation operator, can be generalized to operate on any function $g(X)$:

$$\mathrm{Var}[g(X)]=E[g^2(X)]-(E[g(X)])^2. \qquad (2.3\text{-}14)$$

The variance operator, however, is not a linear operator, as demonstrated with the function $g(X)=a+bX$:

$$\mathrm{Var}[g(X)]=E[(a+bX)^2]-(E[a+bX])^2,$$
$$=b^2E[X^2]-b^2\mu_X^2=b^2\sigma_X^2, \qquad (2.3\text{-}15)$$

because $E[a+bX]=a+b\mu_X$. By letting $b=0$ in the above example, one can demonstrate that the variance of a constant, as expected, is zero.

When the standard deviation is normalized by the mean of the random variable ($\mu_X\neq 0$), it is referred to as the coefficient of variation V_X:

$$V_X=\sigma_X/\mu_X. \qquad (2.3\text{-}16)$$

Estimators for this population parameter, as well as the variance and standard deviation, are discussed in a later section of this report.

As an example of an application of the variance operator, consider an application on the standard normal random variable $Z=(X-\mu_X)/\sigma_X$:

$$\mathrm{Var}[(X-\mu_X)/\sigma_X]=\mathrm{Var}[X]/\sigma_X^2=1 \qquad (2.3\text{-}17)$$

which results by analogy with equation 2.3-15. Thus, if X is a normal random variable with mean μ_X and variance σ_X^2, then Z is a zero-mean random variable with a variance of unity, which is commonly denoted $N(0,1)$.

Problem 2.3-2

a. For $f(x)=\begin{cases}2x & 0\leq x\leq 1\\ 0 & \text{otherwise}\end{cases}$

i. Plot $f(x)$.
ii. Derive and plot $F(x)$.
iii. Calculate $E[X]$ and $\mathrm{Var}[X]$.

b. An estimator of the variance σ_X^2 of the random variable X can be developed directly from equation 2.3-12. First, $f(x)dx$ is estimated by f_i^* of equation 2.2-2. Then x is replaced by $\bar{x}_i$, the center of each class interval corresponding to f_i^*. Finally, μ_X is estimated by $\bar{x}$ from equation 2.3-1. Then

$$s_X^{*2}=\sum_{i\leq x_m/\Delta x}(\bar{x}_i-\bar{x})^2f_i^*$$

gives an estimate of σ_X^2. Apply this estimator to the log-transmissivity data of table 2.2-3.

2.4 Jointly Distributed Random Variables

The investigator frequently encounters the problem that he or she has to deal with two (or more) random variables in the same probability statement. As an example, in the case of random variables X and Y, where X and Y are possibly correlated, one might desire the probability that X is less than or equal to a, and Y is less than or equal to b. If the investigator should know the form of the joint probability density function $f(x,y)$ for these two random variables, then this probability statement is definable:

$$P(X\leq a \text{ and } Y\leq b)$$
$$=\int_{-\infty}^{a}\int_{-\infty}^{b}f(x,y)dydx=F(a,b) \qquad (2.4\text{-}1)$$

where $F(a,b)$ is the equivalent cumulative distribution function. (The statement $P(X\leq a$ and $Y\leq b)$ is also denoted frequently as $P(X\leq a, Y\leq b)$; the more explicit form will be used in this discussion.) As in the univariate case, it is required that the mass under the joint probability density function equal unity:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)dxdy=1. \qquad (2.4\text{-}2)$$

The concept of joint probability density functions applies to any number of multiple random variables; the following discussion, however, is largely restricted to the bivariate case.

As an example of an experiment yielding jointly distributed random variables, consider the results from a simple nonsteady-state pumping test of a confined aquifer: When the Theis equation is used to evaluate data from these tests, information concerning the storativity and transmissivity of the aquifer results. Indeed, we can easily imagine that these quantities are random variables, varying from location to location in response to the local distribution of materials composing the aquifer. More important, however, would be the manner in which they vary with regard to each other: Should the clay content of the aquifer increase at some point, it might be expected that the transmissivity will decrease while the storativity, reflecting the compressibility of the aquifer, would increase. Thus, quite possibly these quantities, with regard to the aquifer in question, could be treated as jointly distributed random variables which are, in some manner, interdependent.

Assume for the moment that we have determined the form of the joint density function of storativity and transmissivity. For argument's sake, let X represent the transmissivity random variable (or its logarithmic transformation) and Y represent the storativity (or a functional transformation thereof) and then denote the joint density as $f(x,y)$. Now assume that we are interested in the probability that X is less than or equal to a, regardless of the value of Y; that is, we wish to evaluate the probability that our measure of the transmissivity will take on a specific range of values, whereas the exact value of storativity is unimportant to us. For our probability statement regarding X to be meaningful, all values of Y which influence the joint density function must be taken into consideration, for different values of Y would surely influence a statement on X alone. To obtain the total contribution of Y to the joint density function, we allow that Y may take on any value in the interval $(-\infty, \infty)$ and write our probability statement as

$$P(X<a \text{ and } -\infty<Y<\infty)=\int_{-\infty}^{a}\int_{-\infty}^{\infty} f(x,y)dydx$$

$$=\int_{-\infty}^{a} f_X(x)dx \qquad (2.4\text{-}3)$$

in which the evaluation of the inner integral with respect to y results in a function $f_X(x)$ that meets all requirements to be a probability density function. Thus, in general, univariate density functions can be recovered from joint density functions by integration, and this integration has the effect of summing the total contribution of one random variable in the bivariate joint density onto the axis of the other variate, the second variate giving the relative frequency of occurrence of the event in question. These densities are referred to as marginal probability density functions and, with respect to the bivariate joint density $f(x,y)$, they are defined as

$$f_X(x)=\int_{-\infty}^{\infty} f(x,y)dy \ , \qquad (2.4\text{-}4)$$

$$f_Y(y)=\int_{-\infty}^{\infty} f(x,y)dx \ , \qquad (2.4\text{-}5)$$

where $f_X(x)$ is the marginal density for the X random variable and similarly $f_Y(y)$ for the Y random variable. The marginal-density concept is easily extended to multiple random variables when they are jointly distributed.

2.4.1 Expectation of Jointly Distributed Random Variables

The expectation operator for jointly distributed random variables is defined in the same manner as in the univariate case. Thus, if X and Y are jointly distributed, and $g(X,Y)$ is a function of these two random variables, then a general definition of the expectation operator is

$$E[g(X,Y)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dydx \ . \qquad (2.4\text{-}6)$$

If on the other hand, we desire the expected value of $h(X)$, which is a function of X only, we set $g(x,y)$ equal to $h(x)$ and proceed as in equation 2.4-6). The result,

$$E[h(X)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x)f(x,y)dydx$$

$$=\int_{-\infty}^{\infty} h(x)f_X(x)dx \qquad (2.4\text{-}7)$$

shows that, in such cases, finding the expected value reduces to finding the marginal density and integrating. By letting $h(X)$ equal X, one realizes that the mean μ_X is equal to the integral of the product of x and the marginal density $f_X(x)$, as might be expected.

Consider the case where $g(X,Y)$ equals the product $(X-\mu_X)(Y-\mu_Y)$. The expected value of this product gives an indication of how X and Y vary together. If the absolute value of the expected value of this product is exceptionally large, then one would expect that X and Y are highly correlated. This expected value of X and Y is referred to as the covariance of X and Y, and is denoted Cov$[X,Y]$ or σ_{XY}:

$$\text{Cov}[X,Y]=E[(X-\mu_X)(Y-\mu_Y)]$$
$$=E[XY]-\mu_Y\mu_X \ . \qquad (2.4\text{-}8)$$

Note that the covariance of X with itself is Cov$[X,X]$=Var$[X]$.

Returning to the example of transmissivities and storativities of the previous section, we see that the covariance provides a measure of the degree of interdependence between random variables. That is, because X and Y are both random, we would not expect observations of X and Y to show a perfect relationship; rather, the relationship will be clouded with noise. Because the expected value implies a frequency-weighted average of the function in question, and because the frequency distribution will reflect the amount of relationship between X and Y, summing the product of these frequency weights with $(X-\mu_X)(Y-\mu_Y)$ over the total variate space will give the average relationship between X and Y. It will be demonstrated in the next section that if X, the measure of transmissivity, and Y, the measure of storativity, were independent, then the covariance would theoretically be zero. However, if our intuition is correct, we would not expect this; rather we might expect, should the aquifer have a rather high clay content, that the two variables will be negatively correlated.

If the covariance is normalized with the standard deviations of the two random variables, then it is referred to as the correlation coefficient ρ_{XY}:

$$\rho_{XY}=\text{Cov}[X,Y]/(\sigma_X\sigma_Y) \ . \qquad (2.4\text{-}9)$$

The correlation coefficient, as a measure of the linear relationship between X and Y, has the property that its absolute value is less than or equal to unity:

$$|\rho_{XY}|\leq 1 \ . \qquad (2.4\text{-}10)$$

That is, when X and Y are precisely linearly related, then $|\rho_{XY}|$ will equal unity. If there is no relationship between X and Y, as shown in the next section, Cov$[X,Y]$ and therefore ρ_{XY} will be zero. This property is demonstrated in appendix 2.11.1, but this appendix requires some knowledge of the next section.

2.4.2 Independent Random Variables

Two random variables X and Y are said to be independent if, for all a and b,

$$P(X\leq a \text{ and } Y\leq b)$$
$$=P(X\leq a)P(Y\leq b)$$
$$=\int_{-\infty}^{a}\int_{-\infty}^{b} f_X(x)f_Y(y)dydx \ , \qquad (2.4\text{-}11)$$

where $f_X(x)$ and $f_Y(y)$ are the densities of X and Y, respectively. Equation 2.4-11 implies that the joint density function of two independent random variables is the product of their individual densities, that is,

$$f(x,y)=f_X(x)f_Y(y) \ . \qquad (2.4\text{-}12)$$

Of course, an event corresponding to $X\leq a$ and $Y\leq b$ would be expected to occur with equal or less frequency than an event corresponding to either $X\leq a$ or $Y\leq b$ separately. Only in the case of a complete lack of dependence between these events can we say that $P(X\leq a \text{ and } Y\leq b)=P(X\leq a)P(Y\leq b)$. This is a somewhat intuitive result that has already been used in connection with the two-dice experiment; if X is an outcome of the first die and Y the second, then $P(X=1 \text{ and } Y=2)=P(X=1)P(Y=2)=1/36$.

A random sample is, ideally, a collection of independent random variables. That is, prior to their observation, each element of a random sample is a random variable; its value is not known until after the observation process is

completed. These outcomes should not have any interdependence which might affect the sample density. This generally requires careful design of the experiment from which the observations result so that all $X_i, i=1, \ldots, n$, are independent.

The question of independence of two random variables X and Y has important implications on their covariance, for if X and Y are independent, then

$$\text{Cov}[X,Y]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f_X(x)f_Y(y)dxdy$$

$$=\int_{-\infty}^{\infty}(x-\mu_X)f_X(x)dx\int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy$$

$$=E[X-\mu_X]\,E[Y-\mu_Y]=0\ . \qquad (2.4\text{-}13)$$

However, if the covariance of two random variables is zero, it does not necessarily follow that they are independent. One may only suspect that independence is the cause of a zero covariance.

2.4.3 Conditional Probabilities

The marginal probability density function, as developed in equation 2.4-3, can be considered to be a special case of a more general concept referred to as conditioning. Generally speaking, a multivariate probability statement is subject to conditioning when a subset of the random variables pertaining to an experiment falls under some restriction, causing the remaining variables to be conditioned by this restriction. In the case of the marginal density function, we examined the probability that X is less than a, given that Y can take on any value in the interval $(-\infty, \infty)$. Thus, the restriction that Y take on a specific set of values conditions the probability that X is less than a. Formally, we state this as

$$P(X<a|-\infty<Y<\infty)\ .$$

In general, the restriction can be applied to any interval (b,c), where $b<c$, and need not be limited to the interval $(-\infty, \infty)$. However, as in the case of the marginal density, the variable or variables subject to restriction are effectively removed from the probability statement; the variable or variables being conditioned are the ones over which the frequency of occurrence of a specific event may be questioned.

When an experiment which results in a bivariate random variable is conditioned over a range other than $(-\infty, \infty)$, a reduction of the potential sample space available to the experiment results. In the previous example of jointly varying transmissivities X and storativities Y, if we were interested in the conditioned results that X is less than a, given that we are only interested in a specific range of values (b,c) for storativities, then the specific value which Y takes on does not interest us, as long as it falls between b and c. One could proceed as in equation 2.4-3 to evaluate this probability, except for an obvious pitfall: The resulting probability statement over X, where X can take on any value less than a, would not necessarily have the property of cumulative distribution functions noted in equation 2.2-11. That is, as X is the remaining active random variable in the probability statement, its probability of occurrence over the interval $(-\infty, \infty)$ should be unity:

$$\lim_{a\to\infty} P(X<a|b<y<c)=1\ .$$

However, by restricting Y to a specific interval (b,c), then as a goes to infinity, an integral of the form of equation 2.4-3 will most probably have a lesser value than unity when the inner integral over y is restricted to a range of something less than $(-\infty, \infty)$. Thus, an integration with the form of equation 2.4-3 alone will not produce a form suitable to serve as a cumulative distribution function for the conditioned variable X.

So that a probability statement resulting from conditioning has the limiting value of unity, these statements must be appropriately normalized. If, as in the bivariate case, we desire $P(X<a|b<Y<c)$, then we must normalize by $P(-\infty<X<\infty \text{ and } b<Y<c)$; that is,

$$P(X<a|b<Y<c)=\frac{P(X<a \text{ and } b<Y<c)}{P(-\infty<X<\infty \text{ and } b<Y<c)}$$

$$=\frac{\int_{-\infty}^{a}\int_{b}^{c}f(x,y)dydx}{\int_{-\infty}^{\infty}\int_{b}^{c}f(x,y)dydx}$$

$$=\frac{\int_{-\infty}^{a}\int_{b}^{c}f(x,y)dydx}{\int_{b}^{c}f_Y(y)dy}\ . \qquad (2.4\text{-}14)$$

Thus, the conditional probability density function, $f(x|b<Y<c)$, for the conditioned random variable X may be defined as

$$f(x|b<Y<c)=\int_b^c f(x,y)dy \Big/ \int_b^c f_Y(y)dy \quad , \quad (2.4\text{-}15)$$

which of course gives the limiting value of unity when integrated with respect to x over the interval $(-\infty, \infty)$. Note that when $b=-\infty$ and $c=\infty$, then $f(x|b<Y<c)=f_X(x)$, as indicated by the previous discussion of marginal densities.

Remarkably, the conditional density exists even when the restriction is that, in the example of the bivariate case, Y take on a specific value. To see this easily, consider $P(X<a|Y=c)$; then equation 2.4-15 may be written as

$$f(x|Y=c)=\lim_{\delta\to 0} \frac{\int_c^{c+\delta} f(x,y)dy}{\int_c^{c+\delta} f_Y(y)dy}$$

$$=\lim_{\delta\to 0} \frac{[F(x,c+\delta)-F(x,c)]/\delta}{[F_Y(c+\delta)-F_Y(c)]/\delta}$$

$$=\frac{dF(x,y)/dy}{dF_Y(y)/dy}\Bigg|_{y=c}$$

$$=\frac{f(x,c)}{f_Y(c)} \quad . \quad (2.4\text{-}16)$$

Thus, we may recover the density function for X for any particular slice, $Y=c$, through the joint density function $f(x,y)$. If $f(x,y)$ were defined for the example of transmissivity and storativity random variables, equation 2.4-16 would enable us to predict the probability of events concerning transmissivity X for any given value of storativity Y.

The student should also note that some remarkable simplifications result if X and Y are independent random variables. That is, if X and Y are independent, then from equations 2.4-12 and 2.4-14 we see that

$$P(X<a|b<Y<c)=P(X<a) \quad . \quad (2.4\text{-}17)$$

Indeed, this is yet another way in which we can define independence of random variables.

The following problem is intended to familiarize the student with the concept of conditioning; it is not intended to be rigorous. The key to understanding conditioning, especially for discrete random variables, is to understand how it restricts the sample space and realize that the probability of occurrence of an event which contains the entire remaining sample space must be unity.

Problem 2.4-1

a. Given two dice that are thrown sequentially, what is the probability that the first is a three and the second is a two? That is,
$P(X=3 \text{ and } Y=2)$?

b. What is the probability that the sum of the dice is five? That is,
$P(X+Y=5)$?

c. Given that the first die is three, what is the probability that the second is two? That is,
$P(Y=2|X=3)$?

d. Given that the first die is three, what is the probability that the sum of the two dice is five? That is,
$P(X+Y=5|X=3)$?

e. Given that the first die is three, what is the probability that the sum of the two dice is less than or equal to five? That is,
$P(X+Y\leq 5|X=3)$?

Parts c, d, and e are conditional probability statements; that is, the probability statement is conditioned by prior information.

2.4.4 Variance of a Column Vector

Our purpose in this section is to develop a representation for the variance of a column vector. As a vehicle to this end, consider the linear equation

$$Y=a_1X_1+a_2X_2+a_3X_3 \quad (2.4\text{-}18)$$

where Y, X_1, X_2, and X_3 are random variables and a_1, a_2 and a_3 are constants. The variance of Y is

$$\begin{aligned}\mathrm{Var}[Y]&=E\{[(a_1X_1+a_2X_2+a_3X_3)\\&\quad -E(a_1X_1+a_2X_2+a_3X_3)]^2\}\\&=a_1^2\sigma_{X_1}^2+a_2^2\sigma_{X_2}^2+a_3^2\sigma_{X_3}^2\\&\quad +2a_1a_2\sigma_{X_1X_2}+2a_1a_3\sigma_{X_1X_3}\\&\quad +2a_2a_3\sigma_{X_2X_3}\end{aligned}\qquad (2.4\text{-}19)$$

where correlations between X_1, X_2, and X_3 have been allowed for. A vector representation for equation 2.4-19 is[1]

$$\begin{aligned}\mathrm{Var}[Y]&=E[Y^2]-(E[Y])^2=E[\underline{a}\underline{X}\underline{X}^T\underline{a}^T]\\&\quad -E[\underline{a}\underline{X}]E[\underline{a}\underline{X}]\end{aligned}\qquad (2.4\text{-}20)$$

where

$$\underline{a}=[a_1,a_2,a_3] \text{ and } \underline{X}=\begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix}.$$

Because expectation is a linear operator, the right side of equation 2.4-20 can be expressed as

$$\begin{aligned}&E[\underline{a}\underline{X}\underline{X}^T\underline{a}^T]-E[\underline{a}\underline{X}]E[\underline{X}^T\underline{a}^T]\\&=\underline{a}E[(\underline{X}-E[\underline{X}])(\underline{X}-E[\underline{X}])^T]\underline{a}^T\end{aligned}\qquad (2.4\text{-}21)$$

where $\underline{a}\underline{X}=\underline{X}^T\underline{a}^T$. The expected value of a matrix is the matrix of expected values of each element. Thus,

$$E[(\underline{X}-E(\underline{X}))(\underline{X}-E(\underline{X}))^T]=$$

$$\begin{bmatrix}\sigma_{X_1}^2 & \sigma_{X_1X_2} & \sigma_{X_1X_3}\\ \sigma_{X_1X_2} & \sigma_{X_2}^2 & \sigma_{X_2X_3}\\ \sigma_{X_1X_3} & \sigma_{X_2X_3} & \sigma_{X_3}^2\end{bmatrix}. \qquad (2.4\text{-}22)$$

This matrix is defined to be the variance of a 3×1 column vector $\underline{X}$, and allows one to express equation 2.4-19 in matrix notation as

$$\mathrm{Var}[Y]=\underline{a}\mathrm{Var}[\underline{X}]\underline{a}^T. \qquad (2.4\text{-}23)$$

[1]Throughout this text singly underlined symbols represent vectors and doubly underlined symbols represent matrices.

If the variances $\sigma_{X_1}^2$, $\sigma_{X_2}^2$, and $\sigma_{X_3}^2$ are all equal, the matrix 2.4-22 becomes

$$\mathrm{Var}[\underline{X}]=\begin{bmatrix}1 & \rho_{X_1X_2} & \rho_{X_1X_3}\\ \rho_{X_1X_2} & 1 & \rho_{X_2X_3}\\ \rho_{X_1X_3} & \rho_{X_2X_3} & 1\end{bmatrix}\sigma^2 \qquad (2.4\text{-}24)$$

where $\rho_{X_iX_j}$ is the correlation coefficient for X_i and X_j, and σ^2 is the common variance. A further reduction in equation 2.4-22 occurs if X_1, X_2, and X_3 are uncorrelated, causing the correlation coefficients in equation 2.4-24 to be zero. In this case,

$$\mathrm{Var}[\underline{X}]=\underline{I}\sigma^2 \qquad (2.4\text{-}25)$$

where $\underline{I}$ is a 3×3 identity matrix. These forms have practical importance in regression.

Problem 2.4-2

a. Carry out the expectation indicated and show that equation 2.4-19 holds.

b. Demonstrate that equation 2.4-21 holds and that

$$\mathrm{Var}[\underline{a}\underline{X}]=\underline{a}\mathrm{Var}[\underline{X}]\underline{a}^T.$$

c. Let $Y_i=\underline{a}_i\underline{X}$, where $\underline{a}_i=[a_{i1},a_{i2},a_{i3}]$, and $\underline{X}$, defined as in equation 2.4-20, is a column vector of random variables. Further, let $\underline{Y}=\underline{\underline{A}}\underline{X}$, where

$$\underline{Y}=\begin{bmatrix}Y_1\\Y_2\\\cdot\\\cdot\\\cdot\\Y_p\end{bmatrix} \quad \text{and} \quad \underline{\underline{A}}=\begin{bmatrix}\underline{a}_1\\\underline{a}_2\\\cdot\\\cdot\\\cdot\\\underline{a}_p\end{bmatrix}.$$

That is, $\underline{\underline{A}}$ is a $p\times3$ matrix composed of the row vectors $\underline{a}_i$, $i=1,\ldots,p$. Show that $\mathrm{Var}[\underline{Y}]=\mathrm{Var}[\underline{\underline{A}}\underline{X}]=\underline{\underline{A}}\mathrm{Var}[\underline{X}]\underline{\underline{A}}^T$. (Hint: Equation 2.4-22 still defines the variance of a column vector;

$$\begin{aligned}\sigma_{Y_iY_j}&=E[Y_iY_j]-E[Y_i]E[Y_j]\\&=E[\underline{a}_i\underline{X}\underline{X}^T\underline{a}_j^T]-E[\underline{a}_i\underline{X}]E[\underline{X}^T\underline{a}_j^T].)\end{aligned}$$

2.5 Estimators of Population Parameters

A statistic is defined as any computation from a random sample resulting in a specific value. As such, a statistic is considered to be a random variable, since it is highly probable that the computed value would change from random sampling to random sampling. Note that this definition precludes that a statistic contain any unknown parameters. Estimators of population parameters are considered to be statistics and, therefore, random variables. Consider equation 2.3-1 as an estimator of the mean:

$$\bar{x}=\sum_{i\leq x_m/\Delta x} f_i^*\bar{x}_i \ . \qquad (2.5\text{-}1)$$

The estimated frequency f_i^* is computed from values of observations x_i originating from a random sampling of the sample space. However, prior to sampling, a random sample is merely an abstract collection of random variables X_i, $i=1,...,n$. Any function of random variables, as equation 2.5-1 would be prior to sampling, is also a random variable, perhaps having a completely different distribution than those individuals composing the collection.

Our discussion of statistics will largely be from the a priori viewpoint; that is, in the case of equation 2.5-1, $\bar{x}$ is the value of the random variable $\bar{X}$, which is an estimator for the population mean, as developed from some arbitrary random sample.

2.5.1 Mean Estimator

As an estimator for the population mean, equation 2.5-1, in addition to being cumbersome to compute, has the debility that it is dependent upon an arbitrary selection of a class interval. That is, since f_i^* is dependent upon Δx, the value of $\bar{X}$ will depend upon the choice of Δx used in the computation. As a means of pursuing this problem, assume that we have at our disposal a random sample consisting of n observations, and at some point their distribution appears as in figure 2.5-1. Because Δx is arbitrary, it can be reduced to δ, the minimum of all differences in neighboring values of the random variable. Then f_i^* would take on only two values, $1/n$ or 0, and would have the ragged sawtoothed shape shown in figure 2.5-2. The teeth in figure 2.5-2 will be concentrated in regions where f_i^* in figure 2.5-1 is larger. Of course, that a repeat value of a random variable could occur is highly unlikely, as the probability of such an occurrence is essentially zero for a continuous random variable (see equation 2.2-14). By using the class interval δ of figure 2.5-2 in our computation, we see that the problem of estimating frequency weights for $\bar{x}_i$, the central location in each class interval, from a discrete data set has essentially been removed, as these weights now take on only two specific values for this and any other smaller class interval.

The use of the smaller class interval, δ, is expected to produce a better estimator of the population mean because a value of $\bar{X}$ would contain less measurement error associated with the arbitrary selection of the class interval. It is still cumbersome, however, to calculate the central value $\bar{x}_i$ of these possibly very small class intervals, especially when one considers that many do not contribute to the estimator. We ask ourselves if it is not possible to use the observations x_i in their place. By reducing Δx even further until every observation is isolated in the center of its own infinitesimally small class interval, in which case f_i^* would remain at the $1/n$ level, observations x_i can be used in place of $\bar{x}_i$ in equation 2.5-1 without significantly altering the basis of our estimator. Using future observations X_i of the process in place of central-interval values, $\bar{x}_i$, an estimator based on an infinitesimally small interval would appear as

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i \qquad (2.5\text{-}2)$$

where n is the size of the random sample, and X_i, $i=1,...,n$, is the collection of random variables from the random sample. Equation 2.5-2 is the preferred estimator for the population mean.

A sample statistic is said to be unbiased if its expected value is equal to the population parameter that it estimates. Consider the expected value of the sample mean, derived from random variables X_i, $i=1,...,n$. Since $E[X_i]=\mu_X$,

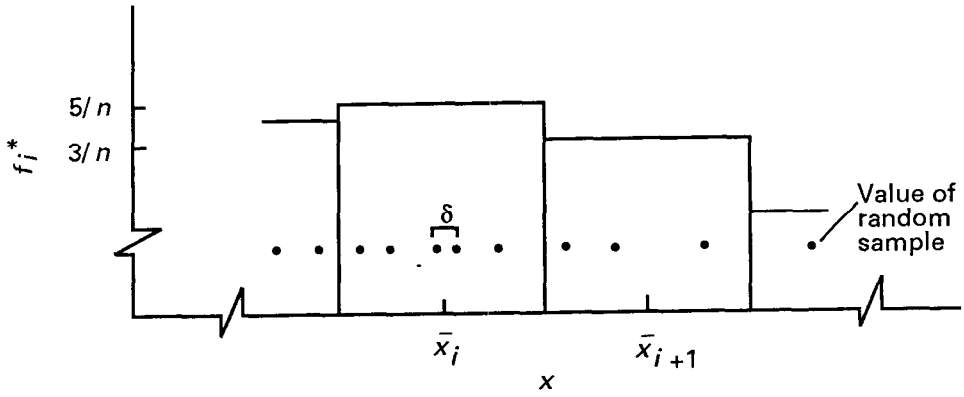

Figure 2.5-1

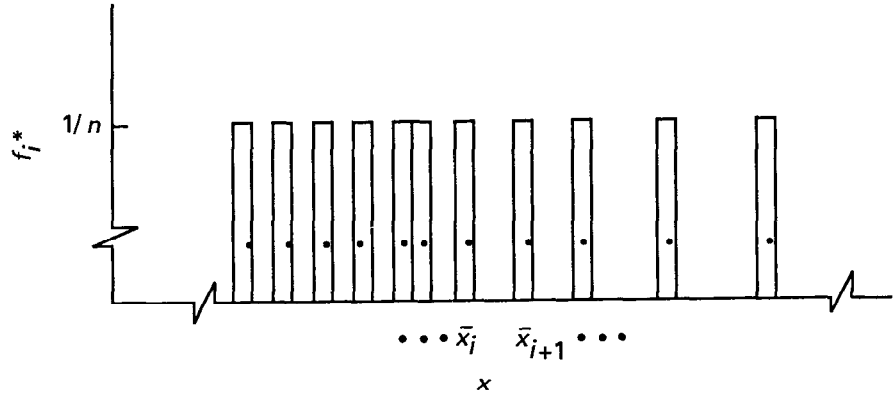

Figure 2.5-2

$$E[\bar{X}]=\frac{1}{n}\sum_{i=1}^{n}E[X_i]=\frac{1}{n}\sum_{i=1}^{n}\mu_X=\mu_X \ . \qquad (2.5\text{-}3)$$

Hence, $\bar{X}$ is an unbiased estimator of μ_X. Although examining an estimator for unbiased qualities is important, it does not necessarily insure that the estimator is the most efficient (or best) in the sense that the variance of the estimator is the smallest. It is, however, an important quality, and the variance estimator is examined for this quality in the next section.

Problem 2.5-1

a. Recompute the sample mean for the data set in problem 2.2-2 using equation 2.5-2 as the mean estimator. How does this result vary from that of problem 2.3-1? How do you, in light of equation 2.2-14, explain the repeat values in the data set (note that this data set represents a random sampling of a continuous random variable)?

b. With regard to a large regional aquifer, well data such as that in table 2.2-2 represent point estimates of transmissivities. The best estimate of the effective transmissivity (the one to use in modeling the flow field) is generally considered to be the geometric mean of these point estimates. The statistic for the geometric mean is defined as

$$\bar{T}_g=(T_1\cdot T_2\cdot T_3\cdot\cdots\cdot T_n)^{1/n}$$

where T_i, $i=1,...,n$, is a random sample from the sample space of the T random variable. Letting $X_i=\log_{10}T_i$, we see that

$$\log_{10}[\bar{T}_g]=\frac{1}{n}\log_{10}(T_1\cdot T_2\cdot T_3\cdot\cdots\cdot T_n)$$

$$=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X} \quad .$$

Therefore,

$$\bar{T}_g=10^{\bar{X}} \quad .$$

What is the geometric mean of the transmissivity data in table 2.2-2?

As a measure of the dispersion about the geometric mean, one could use the estimator

$$D_g=10^{(\bar{X}\pm S_X)} \quad .$$

What is the dispersion D_g about the geometric mean? Use results of problem 2.3-2, part b, as values for $\bar{X}$ and S_X^2. Considering the dispersion, how do you feel about $\bar{t}_g$ being the effective transmissivity of the carbonate rocks of central Pennsylvania?

2.5.2 Variance Estimator

As an estimator of the variance σ_X^2, consider using an estimator S_X^{*2} whose value s_X^{*2} is calculated from the equation

$$s_X^{*2}=\sum_{i\leq x_n/\Delta x} f_i^*(\bar{x}_i-\mu_X)^2 \quad , \qquad (2.5\text{-}4)$$

which is analogous to equation 2.3-12 for the population parameter. If the class interval is taken to be small enough so as to isolate every future observation X_i in a class interval, then equation 2.5-4 can be rewritten in terms of random observations X_i, $i=1,...,n$, as

$$S_X^{*2}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_X)^2 \quad , \qquad (2.5\text{-}5)$$

because $f_i^*=1/n$ and $X_i=x_i$ prior to sampling. When the underlying population X_i, $i=1,...,n$, is normally distributed, it can be shown that equation 2.5-5 is the most efficient, unbiased estimator of the variance σ_X^2 in the sense that its variance is the least of all possible unbiased estimators for σ_X^2.

On occasion, the population mean μ_X can be determined from other considerations, as was done in the titration experiment in section 2.2.5. However, usually μ_X is also unknown, requiring that μ_X be replaced by $\bar{X}$ in equation 2.5-5:

$$S_X^{*2}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad . \qquad (2.5\text{-}6)$$

To test whether equation 2.5-6 is an unbiased estimator of σ_X^2, the expected value of S_X^{*2} is determined. The actual mechanics of this operation are presented in appendix 2.11.2; only the result is presented here:

$$E(S_X^{*2})=\frac{n-1}{n}\sigma_X^2 \quad . \qquad (2.5\text{-}7)$$

Thus, S_X^{*2} is a biased estimator of σ_X^2. To produce an unbiased estimator of σ_X^2, S_X^{*2} is multiplied by the ratio $n/(n-1)$:

$$S_X^2=\frac{n}{n-1}S_X^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad . \qquad (2.5\text{-}8)$$

This estimator is unbiased but less efficient than equation 2.5-6. However, it is the preferred estimator for small sample sizes.

Heuristically, one can argue that this adjustment to the estimator is necessary, because the population mean μ_X is being estimated by the sample statistic $\bar{X}$. The sample mean will be located at the centroid of the random sample, regardless of whether its value is near that of the population mean. Thus, an equation that estimates the variance about this centroid will produce a smaller value than if the estimate were made about the population mean. The adjustment, then, merely compensates for the smaller deviates produced by using $\bar{X}$ in place of μ_X.

Equation 2.5-8 can be rewritten, with the aid of some algebraic manipulation, to produce a slightly more useful form for hand calculations:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] . \qquad (2.5\text{-}9)$$

The estimator for the standard deviation is taken to be the square root of the variance estimator. Values x_i, $i=1,...,n$, obtained by sampling the population of X randomly, are used in place of X_i in equation 2.5-9 to obtain a value s_X^2 for the sample statistic S_X^2.

2.5.3 Estimator of Correlation Coefficient

In a manner analogous to the variance, an estimator for the covariance, and therefore the correlation coefficient, can be derived. Let R_{XY} represent the estimator for the correlation coefficient ρ_{XY}; then, for paired data,

$$R_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]^{1/2}} \qquad (2.5\text{-}10)$$

or, provided that S_X and S_Y originate from the paired data,

$$R_{XY} = \frac{1}{(n-1)S_X S_Y}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) , \qquad (2.5\text{-}11)$$

which can be written, for purposes of hand calculation, as

$$R_{XY} = \frac{1}{(n-1)S_X S_Y}\left(\sum_{i=1}^{n} X_iY_i - n\bar{Y}\bar{X}\right) \qquad (2.5\text{-}12)$$

where S_X and S_Y are calculated by taking the square root of either equations 2.5-8 or 2.5-9. The actual value r_{XY} of R_{XY} is obtained by using values x_i, $i=1,...,n$, from a random sample in place of X_i.

2.5.4 Summary

In summary, population parameters and equivalent sample statistics can be tabulated as follows:

Population parameter	Sample statistic
μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
σ_X^2	$S_X^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$
V_X	$C_X = S_X/\bar{X}$
ρ_{XY}	$R_{XY} = \frac{1}{(n-1)S_X S_Y}\left(\sum_{i=1}^{n} X_iY_i - n\bar{Y}\bar{X}\right)$

These estimators can also be stated in matrix form. For instance, let $d_i = x_i - \bar{x}$; then a value for S_X^2 is

$$s_X^2 = \frac{1}{n-1}\underline{d}^T\underline{d} \qquad (2.5\text{-}13)$$

where $\underline{d}$ is a column vector of deviates and $\underline{d}^T$ is its transpose. If $e_i = y_i - \bar{y}$, then a value for R_{XY} is

$$r_{XY} = \frac{\underline{e}^T\underline{d}}{(n-1)s_X s_Y} . \qquad (2.5\text{-}14)$$

Forms similar to equations 2.5-13 and 2.5-14 are commonly encountered in linear regression.

Problem 2.5-2

Using the following data set, calculate the sample mean, variance, and standard deviation of both dissolved solids and specific conductance; then calculate their correlation coefficient.

Specific conductance and dissolved solids data for wells in carbonate rocks of Maryland

[From Nutter, 1973, p. 63-68]

Specific conductance (μmho/cm)	Dissolved solids (ppm)
278	257
1,120	610
533	338
723	458
462	264
1,030	562
357	231
304	175
469	268
641	388
969	638
876	532
721	405
895	610
501	304
323	171
310	201
1,230	736
504	290
319	208
704	464
1,130	688
600	342

2.6 Transformation of Random Variables

As we have noted previously, statistics are combinations of random variables and, as such, must be random variables themselves. If the population from which the random sample is selected can be identified, then one can frequently identify the probability density functions of statistics, which are estimators of the population parameters. If a density function is identified, then one should be able to develop criteria for testing the accuracy of these estimators. With these objectives in mind, we proceed to identify density functions that result from the several types of transformations that produce statistics.

Before proceeding with this identification process, we make note of two general results from expectation which are applicable to all random variables, regardless of their distribution. In general, if $X_1, X_2, \ldots, X_n$ are independent variables with identical mean μ_X and identical standard deviation σ_X, then

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i \qquad (2.6\text{-}1)$$

is also a random variable with mean

$$\mu_{\bar{X}}=E[\bar{X}]=\frac{1}{n}\sum_{i=1}^{n} E[X_i]=\mu_X \qquad (2.6\text{-}2)$$

and variance

$$\sigma_{\bar{X}}^2=\mathrm{Var}[\bar{X}]=\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i]=\sigma_X^2/n \quad . \qquad (2.6\text{-}3)$$

Equation 2.6-2 was used previously to show that $\bar{X}$ is an unbiased estimator of μ_X, and equation 2.6-3 is demonstrated more fully in appendix 2.11.2. Note that equation 2.6-3 only succeeds because $\mathrm{Cov}[X_i,X_j]=0$, $i \neq j$; that is, the X_i's are independent.

The square root of equation 2.6-3, $\sigma_{\bar{X}}$, is also known as the standard error of $\bar{X}$. The standard deviation of any statistical measure is referred to as the standard error of that statistic.

2.6.1 Sum of Independent Normal Random Variables

Let X_1 and X_2 be independent normal random variables, X_1 with mean zero and variance one ($N(0,1)$) and X_2 with mean zero and variance k ($N(0,k)$). How, then, is their sum distributed? To answer this question, consider

$$P(Y \leq y)=P(X_1+X_2 \leq y) \ ,$$

where $Y=X_1+X_2$. By noting that $P(X_1+X_2 \leq y)=P(X_1 \leq y-X_2$ and $-\infty \leq X_2 \leq \infty)$, comparison with equation 2.4-11 shows that

$$F_Y(y)=\int_{-\infty}^{\infty}\int_{-\infty}^{y-x_2} \frac{e^{-x_1^2/2}}{\sqrt{2\pi}} \frac{e^{-x_2^2/(2k)}}{\sqrt{2\pi k}}\, dx_1 dx_2 \ ,$$

$$=\int_{-\infty}^{\infty} F_{X_1}(y-x_2) \frac{e^{-x_2^2/(2k)}}{\sqrt{2\pi k}}\, dx_2 \ . \qquad (2.6\text{-}4)$$

To find the probability density function of Y, we differentiate $F_Y(y)$ with respect to y; that is

$$f_Y(y)=\frac{d}{dy}F_Y(y)=\int_{-\infty}^{\infty}\frac{d}{dy}F_{X_1}(y-x_2)\frac{e^{-x_2^2/(2k)}}{\sqrt{2\pi k}}dx_2$$

$$=\int_{-\infty}^{\infty}\frac{e^{-(y-x_2)^2/2}}{\sqrt{2\pi}}\frac{e^{-x_2^2/(2k)}}{\sqrt{2\pi k}}dx_2$$

$$=\frac{1}{2\pi\sqrt{k}}$$

$$\cdot\int_{-\infty}^{\infty}\exp[-(y^2-2yx_2+(k+1)x_2^2/k)/2]dx_2 \quad (2.6\text{-}5)$$

which, after some algebraic manipulation, yields

$$f_Y(y)=\frac{\exp[-y^2/(2k+2)]}{2\pi\sqrt{k}}$$

$$\cdot\int_{-\infty}^{\infty}\exp\left[-\left(x_2\sqrt{\frac{k+1}{2k}}-y\sqrt{\frac{k}{2k+2}}\right)^2\right]dx_2. \quad (2.6\text{-}6)$$

By letting $u=\sqrt{2}\left(x_2\sqrt{\frac{k+1}{2k}}-y\sqrt{\frac{k}{2k+2}}\right)$, then

$$f_Y(y)=\frac{\exp[-y^2/(2k+2)]}{\sqrt{2\pi(k+1)}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-u^2/2}du$$

$$=\frac{\exp[-y^2/(2k+2)]}{\sqrt{2\pi(k+1)}} \quad (2.6\text{-}7)$$

which follows from equation 2.2-11. Thus, the sum of two independent zero-mean normal random variables, one with variance unity and the other with variance k, is a normal random variable with variance $k+1$; that is, $N(0,k+1)$.

If, in the previous problem, k were to equal one, then we see that the sum of two $N(0,1)$ independent random variables is a $N(0,2)$ random variable. By adding yet another independent $N(0,1)$ random variable to the previous two $N(0,1)$ random variables, induction tells us that a $N(0,3)$ random variable results. Thus, in general, the sum of n independent $N(0,1)$ random variables results in a $N(0,n)$ random variable.

We are now in a position to determine the distribution of the statistic $\bar{X}$, as shown in equation 2.6-1, if $\bar{X}$ is determined from a random sample in which all the observations X_i, $i=1,\ldots,n$, are independent normal random variables with common mean μ_X and variance σ_X^2; that is, $N(\mu_X,\sigma_X^2)$. We note from equation 2.6-3 that $\bar{X}$ has the standard deviation $\sigma_X/\sqrt{n}$. If we standarize $\bar{X}$ by its mean and standard deviation, and multiply this result by $\sqrt{n}$, then

$$\sqrt{n}\left|\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}}\right|=\sqrt{n}\left|\frac{\sum_{i=1}^{n}X_i-n\mu_X}{\sqrt{n}\sigma_X}\right|=\sum_{i=1}^{n}\left|\frac{X_i-\mu_X}{\sigma_X}\right| \quad (2.6\text{-}8)$$

results. We see that this new statistic is the sum of n normal random variables, each with mean zero and variance one. From the previous paragraph, equation 2.6-8 must be a normal random variable with mean zero and variance n. To obtain a random variable with mean zero and variance unity, one would divide equation 2.6-8 by the square root of n. By inspection, then, the quantity $(\bar{X}-\mu_X)/(\sigma_X/\sqrt{n})$ must be a standard normal random variable, and $\bar{X}$ must be normal with mean μ_X and variance σ_X^2/n. Thus, if one knew that a random sample were composed of normal random variables with a particular mean and variance, then one could investigate the probability that a future determination of the sample mean could take on a particular range of values.

2.6.2 The Chi-Square Distribution

We are frequently concerned with the square of a random variable and may wish to know its density function. Assuming that the random variable X is normally distributed with mean zero and variance one ($N(0,1)$), we may inquire as to the nature of the distribution of its square, $Y=X^2$. Proceeding as in the previous section, we find the cumulative distribution of Y:

$$F_Y(y)=P(Y\le y)=P(X^2\le y)$$

$$=P(-\sqrt{y}\le X\le\sqrt{y})$$

$$=\int_{-\sqrt{y}}^{\sqrt{y}}\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx \quad . \quad (2.6\text{-}9)$$

By taking the derivative of $F_Y(y)$, one finds the density function of Y:

$$f_Y(y)=\frac{d}{dy}F_Y(y)=\frac{1}{\sqrt{2\pi}}\frac{e^{-y/2}}{\sqrt{y}},\ y\geq 0 \qquad (2.6\text{-}10)$$

which is the chi-square density function with 1 degree of freedom.

Chi-square random variables have a useful additive property similar to that exhibited by independent normals. Namely, if Y_1 and Y_2 are independent chi-square random variables with degrees of freedom ν_1 and ν_2, then Y_1+Y_2 is a chi-square random variable with degrees of freedom $\nu_1+\nu_2$. Consequently, if $Y_1,Y_2,\ldots,Y_n$ are independent chi-square random variables each with 1 degree of freedom, then Y_1+Y_2 is a chi square with 2 degrees of freedom, $(Y_1+Y_2)+Y_3$ is a chi square with 3 degrees of freedom and, in general, $\sum_{i=1}^{n} Y_i$ is a chi square with n degrees of freedom. Values for the cumulative distribution function of the chi-square distribution with ν degrees of freedom are to be found in table 2.10-2.

If X_i, $i=1,\ldots,n$, are independent normal random variables, each with mean μ_X and variance σ_X^2, then $\sum_{i=1}^{n}((X_i-\mu_X)/\sigma_x)^2$ must be a chi-square random variable with n degrees of freedom. This follows from the previous argument by letting $Y_i=(X_i-\mu_X)^2/\sigma_X^2$ and noting that Y_i is the square of $N(0,1)$ random variable. Furthermore, because

$$\sum_{i=1}^{n}\frac{(X_i-\mu_X)^2}{\sigma_X^2}=n\frac{(\bar{X}-\mu_X)^2}{\sigma_X^2}+\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma_X^2}, \qquad (2.6\text{-}11)$$

the statistic S_X^2 can be written in terms of this sum as

$$\frac{(n-1)S_X^2}{\sigma_X^2}=\sum_{i=1}^{n}\frac{(X_i-\mu_X)^2}{\sigma_X^2}-\frac{(\bar{X}-\mu_X)^2}{(\sigma_X/\sqrt{n})^2} \quad . \qquad (2.6\text{-}12)$$

Under the condition that the underlying population is independent and normal, it was demonstrated in section 2.6.1 that $\bar{X}$ is normal with mean μ_X and standard deviation $\sigma_X/\sqrt{n}$. Thus, $(\bar{X}-\mu_X)^2/(\sigma_X^2/n)$, under this condition, is chi square with 1 degree of freedom. One might reasonably expect, then, that

$$\frac{(n-1)S_X^2}{\sigma_X^2}\sim\chi^2(\nu) \qquad (2.6\text{-}13)$$

is a chi-square random variable with $\nu=n-1$ degrees of freedom, which is indeed the case when the underlying population of X_i's are independent normal random variables.

2.6.3 The *F* Distribution

The density function for the ratio of two independent chi-square random variables can be calculated rather easily by the method used in the previous sections. However, because we have little need of the actual form of this density function, known as the F distribution, we relieve the student of working through the actual calculation if he or she will accept the following statement: If X_1 is a chi-square random variable with ν_1 degrees of freedom, and X_2 is a chi-square random variable with ν_2 degrees of freedom, and X_1 and X_2 are independent, then

$$\frac{X_1/\nu_1}{X_2/\nu_2}\sim F(\nu_1,\nu_2) \qquad (2.6\text{-}14)$$

defines the F distribution with ν_1 and ν_2 degrees of freedom.

Table 2.10-3 is a tabulation of values of $F(\nu_1,\nu_2)$ which satisfies the probability statement

$$P(F(\nu_1,\nu_2)\leq F_\alpha(\nu_1,\nu_2))=1-\alpha\ , \qquad (2.6\text{-}15)$$

where α equals 0.05; the meaning of equation 2.6-15 is illustrated in figure 2.6-1. Note that the reciprocal of an entry $F_\alpha(\nu_1,\nu_2)$ in table 2.10-3 is equal to $F_{1-\alpha}(\nu_2,\nu_1)$. That is, if equation 2.6-15 holds, then

$$\begin{aligned}&P(F(\nu_2,\nu_1)>F_\beta(\nu_2,\nu_1))\\&=P(1/F(\nu_2,\nu_1)\leq 1/F_\beta(\nu_2,\nu_1))\\&=P(F(\nu_1,\nu_2)\leq 1/F_\beta(\nu_2,\nu_1))\\&=\beta\ ,\end{aligned} \qquad (2.6\text{-}16)$$

as, by equation 2.6-14, $1/F(\nu_2,\nu_1)$ is an $F(\nu_1,\nu_2)$ random variable. By comparing equation 2.6-15 with the third line in equation 2.6-16, we see that, when β equals $1-\alpha$,

$$F_\alpha(\nu_1,\nu_2)=\frac{1}{F_{1-\alpha}(\nu_2,\nu_1)} \quad . \qquad (2.6\text{-}17)$$

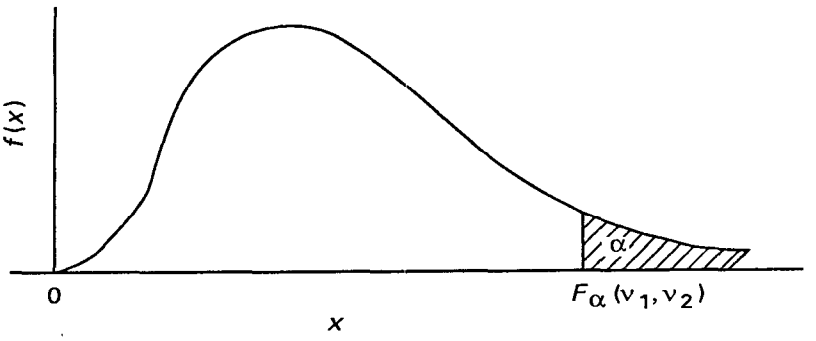

Figure 2.6-1

Thus, if we wish to evaluate $F_{1-\alpha}(n_1, n_2)$ for the statement

$$P(F(n_1,n_2)\leq F_{1-\alpha}(n_1,n_2))=\alpha \qquad (2.6\text{-}18)$$

where α is the relative mass indicated in figure 2.6-2, then we need only find $F_\alpha(\nu_1,\nu_2)$, where $\nu_1=n_2$ and $\nu_2=n_1$, in a table of values for the F distribution and calculate its reciprocal.

As an example of a practical statistic associated with the $F(\nu_1,\nu_2)$ random variable, consider two random samples of size n_1 and n_2, which have been selected from two normal populations with variances σ_1^2 and σ_2^2, respectively. Let

$$X_1=(n_1-1)S_1^2/\sigma_1^2 \qquad (2.6\text{-}19)$$

and

$$X_2=(n_2-1)S_2^2/\sigma_2^2 \qquad (2.6\text{-}20)$$

where S_1^2 and S_2^2 are sample variances that are independent, because they originate from separate random samples. From equations 2.6-13 and 2.6-14, we see that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(\nu_1,\nu_2) \qquad (2.6\text{-}21)$$

is an $F(\nu_1,\nu_2)$ random variable with $\nu_1=n_1-1$ and $\nu_2=n_2-1$ degrees of freedom.

If σ_1^2 were to equal σ_2^2, then equation 2.6-21 would undergo an obvious simplification. As a case in point, consider the ratio of $(\bar{X}-\mu_X)^2/(\sigma_X^2/n)$, which is the square of a $N(0,1)$ random variable, and $(n-1)S_X^2/\sigma_X^2$, which is a $\chi^2(n-1)$ random variable, where $\bar{X}$ and S_X^2 are statistics developed from the same population. One can show (rather arduously) that $\bar{X}$ and S_X^2 are independent, even though they originate from the same random sample. Thus,

$$\frac{(\bar{X}-\mu_X)^2/(\sigma_X^2/n)}{S_X^2/\sigma_X^2}=\frac{(\bar{X}-\mu_X)^2}{S_X^2/n} \sim F(1,n-1)\,. \qquad (2.6\text{-}22)$$

The square root of equation 2.6-22 is also known as a T random variable with $n-1$ degrees of freedom. However, as the T random variable is, in general, equal to the square root of an $F(\nu_1, \nu_2)$ random variable with $\nu_1=1$, no additional time will be devoted to it.

Problem 2.6-1

Residuals ε from a titration experiment (section 2.2.5) have the following values in moles of acid:

$-0.011,\ +0.003,\ +0.004,\ -0.01,\ +0.005,$

$+0.014,\ +0.004,\ +0.001,\ -0.01,\ +0.003.$

Calculate $\bar{\epsilon}$ and s_ε^2 from this random sample. Assume $\mu_\varepsilon=0$; from equation 2.6-22 derive the probability statement

$$P\left(-\sqrt{F_\alpha(1,n-1)}\leq\bar{\varepsilon}/(S_\varepsilon/\sqrt{n})\leq\sqrt{F_\alpha(1,n-1)}\right)=1-\alpha.$$

(Hint: $a^2\leq b$ is equivalent to $-\sqrt{b}\leq a\leq\sqrt{b}$). Find the interval corresponding to this statement when $\alpha=0.05$ (that is, go to table 2.10-3 and find $F_\alpha(1,n-1)$ and then calculate $-\sqrt{F_\alpha(1,n-1)}$ and $\sqrt{F_\alpha(1,n-1)}$). How does the value of the statistic

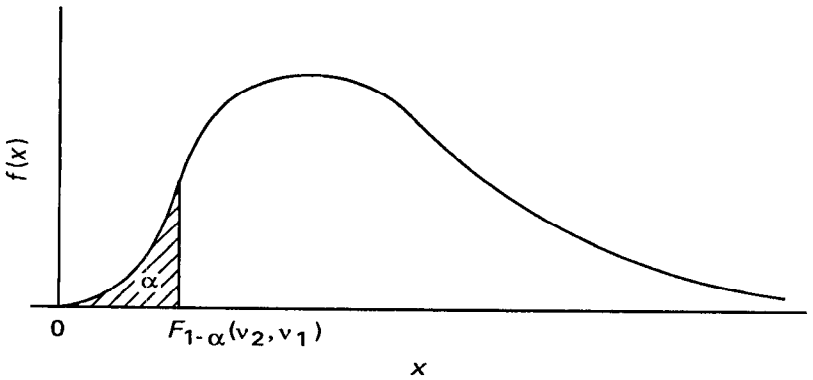

Figure 2.6-2

$\bar{\varepsilon}/(S_\varepsilon/\sqrt{n})$, calculated from the above random sample, compare with this interval? Would you expect this result? Would it bother you if the value fell outside the interval?

2.7 Central Limit Theorem

An interesting and difficult-to-prove theorem of statistics and probability, known as the Central Limit Theorem, concerns the sum of random variables:

Let $X_1,X_2,...,X_n$ be a sequence of identically distributed, independent random variables each with mean μ_X and variance σ_X^2. Then, the distribution of

$$\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}}$$

tends to a standard normal random variable as n goes to infinity. That is,

$$\lim_{n\to\infty} P\left(\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}} \le a\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-t^2/2} dt \qquad (2.7\text{-}1)$$

regardless of the distribution of X_i, $i=1,...,n$.

The key to understanding the impact of this theorem is to realize that the underlying distribution of the X_i random variables can be that of any random variable. Thus, for instance, a chi-square random variable with n degrees of freedom is the sum of other chi-square random variables, and, if n becomes large enough, the chi-square random variable approaches the normal random variable (in fact, the normal tables are used to approximate the chi-square random variable for large values of n).

Another result of this theorem concerns the robustness of some of the distributions developed in the previous section. In particular, if the underlying distribution of X_i's was not normal in equation 2.6-22, this statistic would still be approximately an $F(1,n-1)$ random variable, provided the sample size n were large. The argument for this statement proceeds as follows: Because the numerator of equation 2.6-22 is, when n is large, the square of an approximately normal random variable (by the central limit theorem), it will tend to be a chi-square distributed random variable with 1 degree of freedom. The denominator, on the other hand, will approach unity for large n, because another law of probability dictates that, as n becomes large, S_X^2 approaches σ_X^2 (this phenomenon occurs regardless of the underlying distribution). The net result is that, regardless of the distribution of the X_i random variables, equation 2.6-22 tends, for large n, to be a chi-square distributed random variable with 1 degree of freedom. However, one can show that, for large ν_2, the $F(\nu_1,\nu_2)$ random variable (equation 2.6-14) tends to a chi-square random variable whose value has been diminished by a factor of $1/\nu_1$. Thus, regardless of the underlying distribution of the X_i random variables, we say that equation 2.6-22 behaves asymptotially as an $F(1,n-1)$ random variable when n is large, as both equations 2.6-22 and 2.6-14 have the same distribution for the limiting case where the degrees of freedom in the denominator become large.

2.8 Confidence Limits

We have already noted that statistics are random variables themselves. Now we wish to use the information developed in the previous sections concerning the form of these random variables to make a statement about the reliability of these statistics as estimators. We attempt to define an interval, based upon the statistic, such that a certain percentage of all such intervals, as constructed from different random samples, contain the population parameter that the statistic is thought to estimate. For example, if 5/6 of all possible intervals constructed from repeated sampling contain the population parameter Θ, then there is a probability of 5/6 that the interval we construct from any given random sample actually contains Θ (see figure 2.8-1).

As an example of the interval-construction process, consider the statistic $\bar{X}$ and the population parameter μ_X. We know, from the central limit theorem, that this statistic is approximately normally distributed with mean μ_X and standard deviation $\sigma_X/\sqrt{n}$ and that

$$\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1) \qquad (2.8\text{-}1)$$

is approximately true. Of course, when $\bar{X}$ is based on a random sampling of a normal population, then equation 2.8-1 is exactly true. This standard normal random variable will be used to devise a $(1-\alpha)100\%$ confidence interval for μ_X. This objective is achieved by first looking at the probability statement

$$P(-N_{\alpha/2}(0,1) \leq N(0,1) \leq N_{\alpha/2}(0,1)) = 1-\alpha \qquad (2.8\text{-}2)$$

and finding the values $\pm N_{\alpha/2}(0,1)$ which correspond to $1-\alpha$. This probability statement says that, $(1-\alpha)100\%$ of the time, a value of $N(0,1)$, obtained from a repetition of the experiment, will fall between $-N_{\alpha/2}(0,1)$ and $N_{\alpha/2}(0,1)$. Assuming that σ_X is known, and with a little help from equation 2.8-1, equation 2.8-2 can be rewritten as

$$P\left(\bar{X}-N_{\alpha/2}(0,1)\frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{X}+N_{\alpha/2}(0,1)\frac{\sigma_X}{\sqrt{n}}\right) = 1-\alpha \; . \qquad (2.8\text{-}3)$$

This probability statement says that, $(1-\alpha)100\%$ of the time, the interval $(\bar{X}-N_{\alpha/2}(0,1)\sigma_X/\sqrt{n},\ \bar{X}+N_{\alpha/2}(0,1)\sigma_X/\sqrt{n})$ constructed with a value of $\bar{X}=\bar{x}$ from a particular random sample will contain μ_X. Thus,

$$\bar{x}-N_{\alpha/2}(0,1)\frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{x}+N_{\alpha/2}(0,1)\frac{\sigma_X}{\sqrt{n}} \qquad (2.8\text{-}4)$$

is a $(1-\alpha)100\%$ confidence interval for a random sample of size n, whose variance is known and whose sample mean $\bar{x}$ can be calculated. The investigator would be able to say that the probability is $1-\alpha$ that this interval contains μ_X; however, in interpreting this statement, one

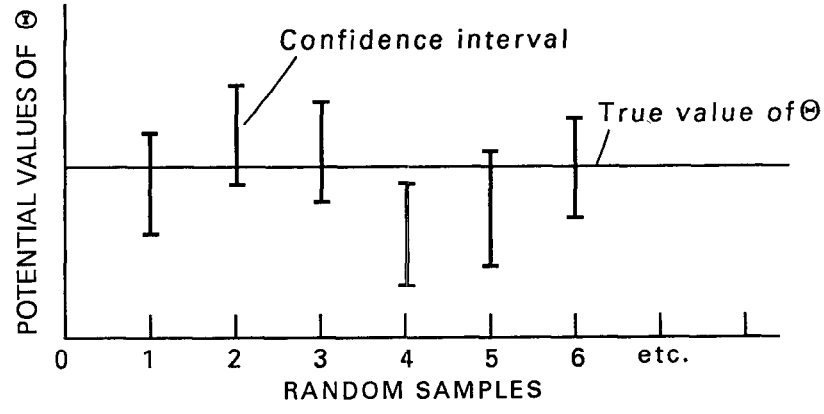

Figure 2.8-1

should realize that the interval is random, not μ_X.

As an actual numeric example of an application of equation 2.8-4, consider the following data: $\sigma_X=0.3$, $\bar{x}=2.6$, $n=36$. Find a 95% confidence interval. From table 2.10-1, for $\alpha/2=0.025$, we see that $N_{\alpha/2}(0,1)=1.96$. Hence a 95% confidence interval is

$$2.6-(1.96)(0.3/\sqrt{36})\leq\mu_X\leq 2.6+(1.96)(0.3/\sqrt{36})$$

or

$$2.50\leq\mu_X\leq 2.70 \ .$$

Thus, as 95 of 100 intervals so constructed contain the mean, there is a 95% probability that this one contains μ_X.

If σ_X is not known, equation 2.8-4 cannot be used. However, if the underlying population is nearly normal, then σ_X^2 can be estimated by S_X^2 as discussed in section 2.6.2 and one can use either the T distribution (with $n-1$ degrees of freedom) or the F distribution as given in equation 2.6-22 to make an appropriate probability statement that can be converted to an interval on μ_X.

Problem 2.8-1

a. Seven gold assays from stockpiled ore are: 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 grams per metric ton. Find a 95% confidence interval for the mean grade of the ore assuming an approximate normal distribution (hint: use equation 2.6-22; why?).
b. Write an interpretation of this interval.

2.9 Hypothesis Testing

Assume that you have determined by a method which you consider to be very good that a population parameter Θ should take on a particular range of values. On the other hand, another independent source suggests that the parameter Θ should take on a value b, which lies outside this range. This discrepancy is disconcerting, and you need some method of testing this independent estimate of Θ. You construct a hypothesis, referred to as the null hypothesis, H_0, that b is the true value of the parameter; symbolically this may be stated:

$$H_0:\Theta=b.$$

Ideally, of course, you wish to reject this hypothesis, but a procedure is needed whereby you can approach the problem objectively. When a random sample is available to the investigator, hypothesis tests can provide this procedure.

If, from a random sample $X_1, X_2,...,X_n$, a test statistic ψ can be constructed which, in some manner, is a measure of Θ, then often a statistical method can be devised to test the probable veracity of the null hypothesis. It is assumed that the distribution of the test statistic is known under the assumption of the null hypothesis or at least can be approximated. Of course, by definition a statistic cannot contain any unknown parameters. Any population parameters which it may contain must be known either by hypothesis or some other means; otherwise ψ would cease to be a statistic. The statistical test will consist of finding a critical interval with a low probability of occurrence under the null hypothesis such that, should a value of ψ as determined from a random sample fall into this interval, the null hypothesis would be rejected and the alternate hypothesis H_1, which usually consists of one of the following, would be accepted:

$$H_1:\Theta>b$$

$$H_1:\Theta<b$$

$$H_1:\Theta\neq b.$$

The alternate hypothesis chosen depends on the nature of the test. A method of intelligently selecting critical intervals must be devised before the test can be completed, for not any arbitrary interval with a small probability of occurrence will do.

2.9.1 Type I Error

In hypothesis-testing procedures, one is ultimately concerned with the possibility of rejecting the null hypothesis when it is true. The objective of hypothesis testing is to make as

small as possible the probability of committing this error, referred to as a type I error. That is, the probability statement

$$P(\text{reject } H_0|H_0 \text{ true})=\alpha \qquad (2.9\text{-}1)$$

is constructed, and α, the level of significance of the test, is chosen as small as the investigator deems reasonable. For continuous random variables, the probability α must be associated with some interval about the test statistic ψ, the statistic fulfilling the requirements of the null hypothesis. Generally speaking, the test statistic ψ will contain an estimator $\hat{\Theta}$ of the population parameter Θ. If $\hat{\Theta}$, as evaluated from some arbitrary random sampling of the experiment, were to have a value close to b, the assumed value of Θ under the null hypothesis, we would not expect to reject the null hypothesis. Rather, only when this value of $\hat{\Theta}$ was distant from b would the null hypothesis be rejected. Thus, the logical choice of an interval in ψ would be one in which all possible values of $\hat{\Theta}$ used in the calculation of ψ would be as distant as possible from b. When the distribution of ψ has infinite tails, then this procedure will cause the interval to include one or both tails, depending on the nature of the test. This interval, whose exact starting and (or) ending point(s) will be determined by the significance level α of the test, will correspond to the critical region where H_0 will be rejected should a calculated value of ψ fall into this region. In most cases, this procedure will cause the critical interval to obtain its maximum length at the chosen significance level α.

When transforming equation 2.9-1 into a probability statement over the test statistic ψ, it is often preferable to first consider the impact of the alternate hypothesis on $\hat{\Theta}$. Consider again the null hypothesis where $\Theta=b$; then, should $\Theta>b$ properly represent the alternate hypothesis, it is useful to consider that, heuristically if not exactly, the probability of committing a type I error can be stated $P(\hat{\Theta}>a|H_0)$, where a is some value of $\hat{\Theta}$ such that $b<a<\infty$, thus giving one an understanding that $\hat{\Theta}$ must take on, relatively, a large positive value in order for H_0 to be rejected. For this alternate hypothesis, a more accurate statement of equation 2.9-1 usually takes the form

$$P(\psi>c|\Theta=b)=\alpha\,, \qquad (2.9\text{-}2)$$

as the distribution of ψ is always assumed to be known.

To complete the above test, a value for c corresponding to α is obtained from a table of cumulative probabilities. Values of ψ less than c correspond to a region where the probability of committing a type I error may not be small. Therefore, if a value of ψ is less than c, we are forced to accept the null hypothesis to avoid committing a type I error. If this value is larger than c, then the probability of committing this error is considered small, and we can confidently reject H_0 at the α significance level.

Examples of hypothesis testing, which should clarify the actual mechanics of the procedure, are presented subsequently; however, before proceeding to these examples, note that we are frequently required to play the role of the devil's advocate in hypothesis testing. Often, we really desire to test the acceptability of a hypothesized value of a parameter. To accomplish this task, we first attempt to reject this value by making it the subject of the null hypothesis. If we cannot reject the null hypothesis, then we must admit that the hypothesized value is indeed a candidate for the true value of the parameter in question.

2.9.2 One-Tailed Test

As an example of developing the probability statement associated with equation 2.9-2, assume that we wish to test the hypothesis that the mean of a population is μ_0, versus the alternate hypothesis that the population mean is greater than μ_0 (assume that the standard deviation σ_X is known):

$$H_0{:}\mu_X=\mu_0$$

versus

$$H_1{:}\mu_X>\mu_0.$$

This test is referred to as a one-tailed test because the alternate hypothesis only allows for a mean greater than that indicated by H_0.

A random variable is needed whereby we may build a probability statement around the type

I error. Assume that data in the form of a random sample $X_1, X_2, \ldots, X_n$ from the population exist; a natural random variable for this purpose would be the estimator of the mean $\bar{X}$. As $\bar{X}$ is an estimator of μ_X, and as the alternate hypothesis presupposes that μ_X is large, one would reject H_0 if a value of $\bar{X}$, as determined from a random sample, were significantly larger than μ_0. From the central limit theorem, assume that $\bar{X}$ is approximately normally distributed with mean μ_X and standard deviation $\sigma_X/\sqrt{n}$. Equation 2.9-2 can be represented in terms of the statistic $\bar{X}$ as

$$P(\bar{X}>a|H_0)=\alpha\,. \tag{2.9-3}$$

Although the distribution of $\bar{X}$ is known, a statistic which will allow us to incorporate the null hypothesis that $\mu_X=\mu_0$ is needed. A statistic meeting this requirement and for which values of all the parameters can be supplied is $(\bar{X}-\mu_X)/(\sigma_X/\sqrt{n})$. With this test statistic, equation 2.9-3 can be restated as

$$P\left(\frac{\bar{X}-\mu_x}{\sigma_X/\sqrt{n}}>c|\mu_X=\mu_0\right)=P\left(\frac{\bar{X}-\mu_0}{\sigma_X/\sqrt{n}}>N_\alpha(0,1)\right)=\alpha \tag{2.9-4}$$

where μ_0 is used in place of μ_X to satisfy the null hypothesis. Note that under the null hypothesis $(\bar{X}-\mu_0)/(\sigma_X/\sqrt{n})$ is a normal random variable with mean zero and variance unity; thus, $N_\alpha(0,1)$ becomes the lower limit c of the critical region for this test.

All possible values of the test statistic greater than $N_\alpha(0,1)$, where α is the level of significance of the test, constitute the critical region where H_0 would be rejected. In other words, $N_\alpha(0,1)$ is the critical value, corresponding to the limit c in equation 2.9-2, which determines whether we accept or reject the null hypothesis. If a value of the test statistic is greater than $N_\alpha(0,1)$, we would reject H_0 at the α significance level. If the value were less, then we would be forced to accept the null hypothesis for fear of making a type I error.

As a sample application of this procedure, consider the data used to construct the confidence interval at the end of section 2.8: $\sigma_X=0.3$, $\bar{x}=2.6$ and $n=36$. We are told that the population mean is really zero, a statement that seems rather dubious to us as we believe it to be some positive real number. We set the null hypothesis that μ_X is indeed zero, $H_0{:}\mu_X=0$, and hope that we can confidently disallow it. Our alternate hypothesis consists of our own belief, $H_1{:}\mu_X>0$. As we wish to be very sure that we do not commit a type I error, we set the level of significance of our test at $\alpha=0.025$. We determine the critical value of our test statistic from table 2.10-1: $N_\alpha(0,1)=1.96$. We evaluate the test statistic under the assumption of the null hypothesis: $\bar{x}/(\sigma_X/\sqrt{n})=52$. Because this value of the test statistic is considerably larger than the critical value, we reject H_0 at the 0.025 significance level, realizing that, although we may have committed a type I error, it is highly unlikely.

2.9.3 Two-Tailed Test

Suppose that σ_X^2 is unknown, but we are given a random sample $X_1, X_2, \ldots, X_n$ from a normal population. We wish to test the hypothesis

$$H_0{:}\mu_X=\mu_0$$

versus

$$H_1{:}\mu_X\neq\mu_0$$

at a significance level α. This is referred to as a two-tailed test: We reject H_0 if a measure of μ_X is either significantly greater or less than μ_0.

To construct this test, recall the statistic from equation 2.6-22:

$$\frac{(\bar{X}-\mu_X)^2}{S_X^2/n}\sim F(1,\,n-1) \tag{2.9-5}$$

which is the $F(1,n-1)$ random variable. This statistic fulfills our requirement for a test statistic; it can be used to satisfy the null hypothesis, and the remaining statistics or parameters are either known or can be evaluated from a random sample. Now consider the probability of a type I error:

$$\begin{aligned} &P(\text{reject } H_0|H_0 \text{ true})\\ &=P(\bar{X}<a|H_0)+P(\bar{X}>b|H_0)\\ &=\alpha\,, \end{aligned} \tag{2.9-6}$$

where two critical values are necessary as it is possible to reject the null hypothesis if a value of $\bar{X}$ is either larger or smaller than μ_0. In terms of the test statistic, under the condition that the null hypothesis holds, we see that

$$P\left(\left(\frac{\bar{X}-\mu_0}{S_X/\sqrt{n}}\right)^2 > F_\alpha(1,\, n-1)\right)$$

$$= P\left(\frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} < -\sqrt{F_\alpha(1,n-1)}\right.$$

$$\left. \text{or } \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} > \sqrt{F_\alpha(1,n-1)}\right)$$

$$= P\left(\frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} < -\sqrt{F_\alpha(1,n-1)}\right)$$

$$+ P\left(\frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} > \sqrt{F_\alpha(1,n-1)}\right)$$

$$= \alpha \qquad (2.9\text{-}7)$$

which is equivalent to equation 2.9-6.

To complete the test, we need only to evaluate the test statistic with a random sample. If $(\bar{x}-\mu_0)^2/(s_X^2/n)$ be greater than $F_\alpha(1,n-1)$, we would reject the null hypothesis at the α significance level.

2.9.4 Type II Error

A test statistic is commonly selected for its ability to determine the probability of committing a type II error, as well as a type I error. A type II error is committed by accepting the null hypothesis when the alternate is true. By calculating the probability that the test statistic does not fall in the critical region, given that Θ takes on any value other than that assumed under the null hypothesis, the probability β of committing this error can be evaluated. Thus, for tests indicated previously, β is a continuous function of possible values of the population parameter Θ, other than the value b assumed under the null hypothesis. For a critical region corresponding to a given α, a good test statistic should produce small values of β for hypothetical values of Θ rather distant from b. However, β should increase sharply in value as possible values of Θ approach b and obtain a value as close to one as feasible in the immediate vicinity of b. Means are available for determining test statistics which, for certain tests, excel at the above characteristics, but a presentation of these methods is beyond the scope of this course. In most cases, a statistic which contains an estimator of the population parameter being tested and for which all other parameters are either known, or estimators of said parameters are contained in the statistic, will suffice as a test statistic; however, it may not be the best test statistic.

Note that if α, the probability of committing a type I error, were made extremely small, then the null hypothesis would almost always be accepted. At first glance, one would assume that something was amiss in the hypothesis testing procedure, as it is apparently possible to bias the test by selecting an extreme value for α. However, when the value of α is decreased, the probability of committing a type II error, $\beta(\Theta)$, is increased for all values Θ. Thus, an investigator who seeks to avoid committing a type I error by intentionally selecting a small value for α runs an increased risk of committing a type II error, which is equally as damaging. If need be, a plot of $\beta(\Theta)$ can be made for various hypothetical values of α and Θ; this can often be a rather complicated task. A rule-of-thumb value for α is 0.05, which appears to serve hypothesis test users well in most cases.

2.9.5 Summary of Method

To summarize, the steps for testing a hypothesis concerning a population parameter Θ are:

1. Define the null hypothesis H_0:$\Theta=\Theta_0$.
2. Decide upon the nature of the test; that is, H_1:$\Theta<\Theta_0$, H_1:$\Theta>\Theta_0$ or H_1:$\Theta \neq \Theta_0$.
3. Choose a level of significance α.
4. Select an appropriate test statistic and establish the critical region.
5. Compute the value of the statistic from a random sample of size n.
6. Draw conclusion of test: reject H_0 if the statistic has a value in the critical region; otherwise accept H_0.

Problem 2.9-1

a. Set up problem 2.6-1 as a hypothesis test (do not complete the test).

b. Given two random samples from independent normal populations with the following sample statistics:

Statistic	Random sample 1	Random sample 2
n	25	16
$\bar{x}$	82	78
s_x	8	7

test the following hypotheses at a significance level of $\alpha=0.05$:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1 .$$

c. An outside source informs you that the stockpiled ore of problem 2.8-1 actually only assays an average of 9.8 grams per metric ton. Can you refute this claim at a significance level of 0.05? (Construct a hypothesis test for this purpose.)