7

Probability Theory and Statistics

• • •

In the last chapter we made the transition from discussing information which is considered to be error free to dealing with data that contained intrinsic errors. In the case of the former, uncertainties in the results of our analysis resulted from the failure of the approximation formula to match the given data and from round-off error incurred during calculation. Uncertainties resulting from these sources will always be present, but in addition, the basic data itself may also contain errors. Since all data relating to the real world will have such errors, this is by far the more common situation. In this chapter we will consider the implications of dealing with data from the real world in more detail.

Philosophers divide data into at least two different categories, observational, historical, or empirical data and experimental data. Observational or historical data is, by its very nature, non-repeatable. Experimental data results from processes that, in principle, can be repeated. Some[1] have introduced a third type of data labeled hypothetical-observational data, which is based on a combination of observation and information supplied by theory. An example of such data might be the distance to the Andromeda galaxy since a direct measurement of that quantity has yet to be made and must be deduced from other aspects of the physical world. However, in the last analysis, this is true of all observations of the world. Even the determination of repeatable, experimental data relies on agreed conventions of measurement for its unique interpretation. In addition, one may validly ask to what extent an experiment is precisely repeatable. Is there a fundamental difference between an experiment, which can be repeated and successive observations of a phenomenon that apparently doesn't change? The only difference would appear to be that the scientist has the option in the case of the former in repeating the experiment, while in the latter case he or she is at the mercy of nature. Does this constitute a fundamental difference between the sciences? The hard sciences such as physics and chemistry have the luxury of being able to repeat experiments holding important variables constant, thereby lending a certain level of certainty to the outcome. Disciplines such as Sociology, Economics and Politics that deal with the human condition generally preclude experiment and thus must rely upon observation and "historical experiments" not generally designed to test scientific hypotheses. Between these two extremes are sciences such as Geology and Astronomy which rely largely upon observation but are founded directly upon the experimental sciences. However, all sciences have in common the gathering of data about the real world. To the analyst, there is little difference in this data. Both experimental and observational data contain intrinsic errors whose effect on the sought for description of the world must be understood.

However, there is a major difference between the physical sciences and many of the social sciences and that has to do with the notion of cause and effect. Perhaps the most important concept driving the physical sciences is the notion of causality. That is the physical biological, and to some extent the behavioral sciences, have a clear notion that event A causes event B. Thus, in testing a hypothesis, it is always clear which variables are to be regarded as the dependant variables and which are to be considered the independent variables. However, there are many problems in the social sciences where this luxury is not present. Indeed, it may often be the case that it is not clear which variables used to describe a complex phenomenon are even related. We shall see in the final chapter that even here there are some analytical techniques that can be useful in deciding which variables are possibly related. However, we shall also see that these tests do not prove cause and effect, rather they simply suggest where the investigator should look for causal relationships. In general data analysis may guide an investigator, but cannot substitute for his or her insight and understanding of the phenomena under investigation.

During the last two centuries a steadily increasing interest has developed in the treatment of large quantities of data all representing or relating to a much smaller set of parameters. How should these data be combined to yield the "best" value of the smaller set of parameters? In the twentieth century our ability to collect data has grown enormously, to the point where collating and synthesizing that data has become a scholarly discipline in itself. Many academic institutions now have an entire department or an academic unit devoted to this study known as statistics. The term statistics has become almost generic in the language as it can stand for a number of rather different concepts. Occasionally the collected data itself can be referred to as statistics. Most have heard the reference to reckless operation of a motor vehicle leading to the operator "becoming a statistic". As we shall see, some of the quantities that we will develop to represent large

amounts of data or characteristics of that data are also called statistics. Finally, the entire study of the analysis of large quantities of data is referred to as the study of statistics. The discipline of statistics has occasionally been defined as providing a basis for decision-making on the basis of incomplete or imperfect data. The definition is not a bad one for it highlights the breadth of the discipline while emphasizing it primary function. Nearly all scientific enterprises require the investigator to make some sort of decisions and as any experimenter knows, the data is always less than perfect.

The subject has its origins in the late 18th and early 19th century in astronomical problems studied by Gauss and Legendre. Now statistical analysis has spread to nearly every aspect of scholarly activity. The developing tools of statistics are used in the experimental and observational sciences to combine and analyze data to test theories of the physical world. The social and biological sciences have used statistics to collate information about the inhabitants of the physical world with an eye to understanding their future behavior in terms of their past performance. The sampling of public opinion has become a driving influence for public policy in the country. While the market economies of the world are largely self-regulating, considerable effort is employed to "guide" these economies based on economic theory and data concerning the performance of the economies. The commercial world allocates resources and develops plans for growth based on the statistical analysis of past sales and surveys of possible future demand. Modern medicine uses statistics to ascertain the efficacy of drugs and other treatment procedures. Such methods have been used, not without controversy, to indicate man made hazards in our environment. Even in the study of language, statistical analysis has been used to decide the authorship of documents based on the frequency of word use as a characteristic of different authors.

The historical development of statistics has seen the use of statistical tools in many different fields long before the basis of the subject were codified in the axiomatic foundations to which all science aspires. The result is that similar mathematical techniques and methods took on different designations. The multi-discipline development of statistics has lead to an uncommonly large amount of jargon. This jargon has actually become a major impediment to understanding. There seems to have been a predilection, certainly in the nineteenth century, to dignify shaky concepts with grandiose labels. Thus the jargon in statistics tends to have an excessively pretentious sound often stemming from the discipline where the particular form of analysis was used. For example, during the latter quarter of the nineteenth century, Sir Francis Galton analyzed the height of children in terms of the height of their parents[2]. He found that if the average height of the parents departed from the general average of the population by an amount x, then the average height of the children would depart by, say, 2x/3 from the average for the population. While the specific value of the fraction (2/3) may be disputed all now agree that it is less than one. Thus we have the observation that departures from the population average of any sub group will *regress* toward the population average in subsequent generations. Sir Francis Galton used Legendre's Principle of Least Squares to analyze his data and determine the *coefficient of regression* for his study. The use of least squares in this fashion has become popularly known as *regression analysis* and the term is extended to problems where the term regression has absolutely no applicability. However, so wide spread has the use of the term become, that failure to use it constitutes a barrier to effective communication.

Statistics and statistical analysis are ubiquitous in the modern world and no educated person should venture into that world without some knowledge of the subject, its strengths and limitations. Again we touch upon a subject that transcends even additional courses of inquiry to encompass a lifetime of study. Since we may present only a bare review of some aspects of the subject, we shall not attempt a historical development.

Rather we will begin by giving some of the concepts upon which most of statistics rest and then developing some of the tools which the analyst needs.

7.1 Basic Aspects of Probability Theory

We can find the conceptual origins of statistics in probability theory. While it is possible to place probability theory on a secure mathematical axiomatic basis, we shall rely on the commonplace notion of probability. Everyone has heard the phrase "the probability of snow for tomorrow 50%". While this sounds very quantitative, it is not immediately clear what the statement means. Generally it is interpreted to mean that on days that have conditions like those expected for tomorrow, snow will fall on half of them. Consider the case where student A attends a particular class about three quarters of the time. On any given day the professor could claim that the probability of student A attending the class is 75%. However, the student knows whether or not he is going to attend class so that he would state that the probability of his attending class on any particular day is either 0% or 100%. Clearly the probability of the event happening is dependent on the prior knowledge of the individual making the statement. There are those who define *probability as a measure of ignorance*. Thus we can define two events to be equally likely if we have no reason to expect one event over the other. In general we can say that if we have n equally likely cases and any m of them will generate an event E, then the probability of E occurring is

$$P(E) = m/n \quad . \tag{7.1.1}$$

Consider the probability of selecting a diamond card from a deck of 52 playing cards. Since there are 13 diamonds in the deck, the probability is just 13/52 = ¼. This result did not depend on there being 4 suits in the standard deck, but only on the ratio of 'correct' selections to the total number of possible selections. It is always assumed that the event will take place if all cases are selected so that the probability that an event E will *not* happen is just

$$Q(\tilde{E}) = 1 - P(E) \quad . \tag{7.1.2}$$

In order to use equation (7.1.1) to calculate the probability of event E taking place, it is necessary that we correctly enumerate all the possible cases that can give rise to the event. In the case of the deck of cards, this seems fairly simple. However, consider the tossing of two coins where we wish to know the probability of two 'heads' occurring. The different possibilities would appear to be each coin coming up 'heads', each coin coming up 'tails', and one coin coming up 'heads' while the other is 'tails'. Thus naïvely one would think that the probability of obtaining two 'heads' would be 1/3. However, since the coins are truly independent events, each coin can be either 'heads' or 'tails'. Therefore there are two separate cases where one coin can be 'head' and the other 'tails' yielding four possible cases. Thus the correct probability of obtaining two 'heads' is 1/4. The set of all possible cases is known as the *sample set*, or *sample space*, and in statistics is sometimes referred to as the *parent population*.

a. The Probability of Combinations of Events

It is possible to view our coin tossing even as two separate and independent events where each coin is tossed separately. Clearly the result of tossing each coin and obtaining a specific result is 1/2. Thus the result of tossing two coins *and* obtaining a specific result (two heads) will be 1/4, or $(1/2)\times(1/2)$. In general, the probability of obtaining event E *and* event F, [P(EF)], will be

$$P(EF) = P(E) \times P(F) \quad . \tag{7.1.3}$$

Requiring of the occurrence of event E *and* event F constitutes the use of the *logical and* which always results in a multiplicative action. We can ask what will be the total, or joint, probability of event E *or* event F occurring. Should events E and F be mutually exclusive (i.e. there are no cases in the sample set that result in both E and F), then $P(E_{or}F)$ is given by

$$P(E_{or}F) = P(E) + P(F) \quad . \tag{7.1.4}$$

This use of addition represents the *logical 'or'*. In our coin tossing exercise obtaining one 'head' and one 'tail' could be expressed as the probability of the first coin being 'heads' *and* the second coin being 'tails' *or* the first coin being 'tails' *and* the second coin being 'heads' so that

$$P(HT) = P(H)P(T) + P(T)P(H) = (1/2)\times(1/2) + (1/2)\times(1/2) = 1/2 \quad . \tag{7.1.5}$$

We could obtain this directly from consideration of the sample set itself and equation (7.1.1) since m = 2, and n = 4. However, in more complicated situations the laws of combining probabilities enable one to calculate the combined probability of events in a clear and unambiguous way.

In calculating $P(E_{or}F)$ we required that the events E and F be mutually exclusive and in the coin exercise, we guaranteed this by using separate coins. What can be done if that is not the case? Consider the situation where one rolls a die with the conventional six faces numbered 1 through 6. The probability of any particular face coming up is 1/6. However, we can ask the question what is the probability of a number less than three appearing *or* an even number appearing. The cases where the result is less than three are 1 and 2, while the cases where the result is even are 2, 4, and 6. Naïvely one might think that the correct answer 5/6. However, these are not mutually exclusive cases for the number 2 is both an even number and it is also less than three. Therefore we have counted 2 twice for the only distinct cases are 1, 2, 4, and 6 so that the correct result is 4/6. In general, this result can be expressed as

$$P(E_{or}F) = P(E) + P(F) - P(EF) \quad , \tag{7.1.6}$$

or in the case of the die

$$P(<3_{or}\text{even}) = [(1/6)+(1/6)] + [(1/6)+(1/6)+(1/6)] - [(1/3)\times(1/2)] = 2/3 \quad . \tag{7.1.7}$$

We can express these laws graphically by means of a Venn diagram as in figure 7.1. The simple sum of the dependent probabilities counts the intersection on the Venn diagram twice and therefore it must be removed from the sum.

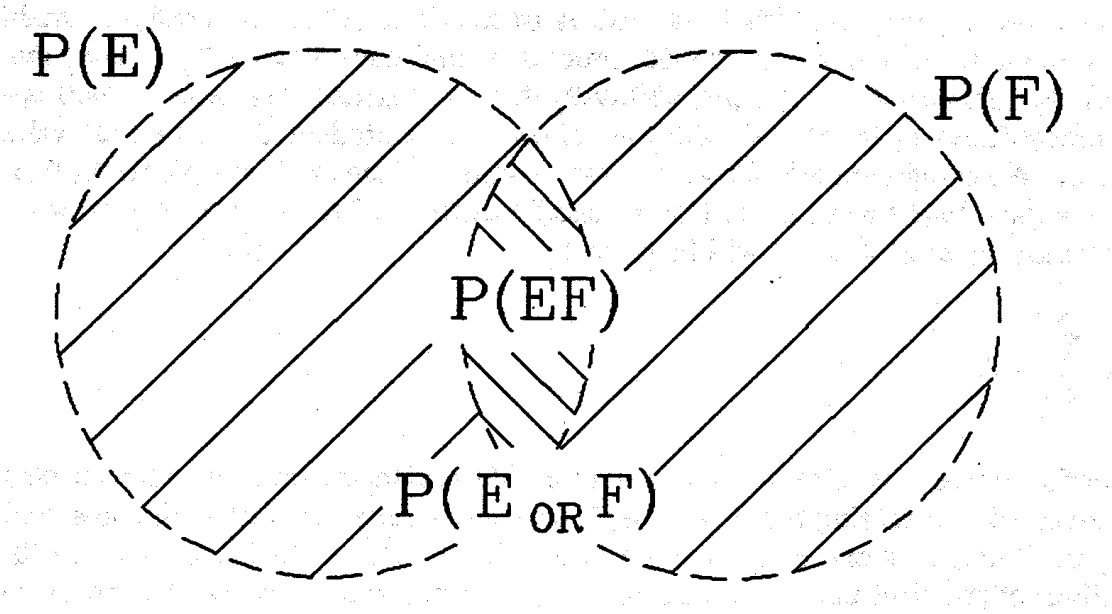

Figure 7.1 shows a sample space giving rise to events E *and* F. *In the case of the die,* E *is the probability of the result being less than three and* F *is the probability of the result being even. The intersection of circle* E *with circle* F *represents the probability of* E *and* F *[i.e.* P(EF)*]. The union of circles* E *and* F *represents the probability of* E *or* F. *If we were to simply sum the area of circle* E *and that of* F *we would double count the intersection.*

b. Probabilities and Random Variables

We can define a random process as one where the result of the process cannot be predicted. For example, the toss of a coin will produce either 'heads' or 'tails', but which will occur as the result of flipping the coins cannot be predicted with complete certainty. If we assign a 1 to 'heads' and a 0 to 'tails', then a succession of coin flips will generate a series of 1's and 0's having no predictable order. If we regard a finite sequence of flips as a binary number, then we can call it a random number since its value will not be predictable. Any secession of finite sequences of the same length will produce a secession of random binary numbers where no number can be predicted from the earlier numbers. We could carry out the same experiment with the die where the results would range from 1 to 6 and the sequences would form base six random numbers.

Now the sequence that produces our random number could be of arbitrary length even though the sample set is finite, but it will always have some numerical value. We can define a *random variable* as any numerically valued function that is defined on the sample set. In the case we have picked, it could be, say, all numbers with five digits or less. Let us define the elements of the sample set to have numerical values x_i. In the case of the coin these would be the 1's and 0's we assigned to 'heads' and 'tails'. For the die, they are simply the values of the faces. Then any random variable, which would appear through its definition as a

random process, would have a result $X_j(x_i) = X_j$ that depends on the values of the sample set x_i. The probability P_j that any particular value of X_j will appear will depend on the probabilities p_i associated with the values x_i that produce the numerical value of the random variable X_j. We could then ask "If we generate n values of the random variable X_j from the sample set, what is the most likely value of X_j that we should expect?". We will call that value of X_j the *expected* or *expectation value* of X and it will be given by

$$E(X) = \sum_{j=1}^{N} P_j X_j \quad . \tag{7.1.8}$$

Consider the simple case of tossing coins and ask "What is the expectation value for obtaining one 'head' in any given trial of tossing the two coins?". The possibilities are that both coins could turn up 'tails' yielding no 'heads', or one coin could be 'heads' and the other 'tails', or both could be 'heads'. The probabilities of the first and last occurring is ¼, but since either coin can be 'heads' while the other is 'tails' the middle possibility represents two separate cases. Thus the expected value for the number of 'heads' is just

$$E(\mathrm{H}) = 0\times(¼) + 1\times(¼) + 1\times(¼) + 2\times(¼) = 1 \quad . \tag{7.1.9}$$

The first term is made up of the number of heads that result for each trial times the probability of that trial while the other representation of that sum show the distinctly different values of X_j multiplied by the combined probability of those values occurring. The result is that we may expect one 'head' with the toss of two coins. The expectation value of a random variable is sort of an average value or more properly the most likely value of that variable.

c. Distributions of Random Variables

It is clear from our analysis of the coin tossing experiment that not all values of the random variable (eg. the number of 'heads') are equally likely to occur. Experiments that yield one 'head' are twice as likely to occur as either no 'heads' or two 'heads'. The frequency of occurrence will simply be determined by the total probability of the random variable. The dependence of the probability of occurrence on the value of the random variable is called a *probability distribution*. In this instance there is a rather limited number of possibilities for the value of the random variable. Such cases are called discrete probability distributions. If we were to define our random variable to be the value expected from the roll of two dice, then the values could range from 2-12, and we would have a more extensive discrete probability distribution. In general, measured values contain a finite set of digits for the random variables and their probability distributions are always discrete.

However, it is useful to consider continuous random variables as they are easier to use analytically. We must be careful in our definition of probability. We can follow the standard practice of limits used in the differential calculus and define the differential probability of the continuous random variable x occurring within the interval between x and x+Δx to be

$$dP(x) = \lim_{\Delta x \to 0} [f(x+\Delta x)-f(x)]/\Delta x \quad . \tag{7.1.10}$$

Thus the probability that the value of the random variable will lie between a and b will be

$$P(a,b) = \int_a^b f(x)\,dx \quad . \tag{7.1.11}$$

The function f(x) is known as the *probability density distribution function* while P(a,b) is called the *probability distribution function*. The use of probability density functions and their associated probability distribution functions constitute a central tool of analysis in science.

7.2 Common Distribution Functions

From our discussion of random variables, let us consider how certain widely used distribution functions arise. Most distribution functions are determined for the discrete case before generalizing to their continuous counterparts and we will follow this practice. Consider a sample space where each event has a constant probability p of occurring. We will let the random variable be represented by a sequence of sampling events. We then wish to know what the probability of obtaining a particular sequence might be. If we assign each sequence a numerical value, then the probability values of the sequences form a probability distribution function. Let us sample the set of equally probable events n times with m occurrences of an event that has probability p so that we obtain the sequence with total probability

$$P(S) = ppqqq \cdots pqqqppqp = p^m q^{n-m}\,, \tag{7.2.1}$$

where

$$q = 1 - p\,, \tag{7.2.2}$$

is the probability that the sampling did not result in the event. One can think of an event as getting a head from a coin toss.

Since the sampling events are considered independent, one is rarely interested in the probability of the occurrence of a particular sequence. That is, a sequence ppq will have the same probability as the sequence pqp, but one generally wishes to know the probability that one *or* the other *or* some equivalent (i.e. one having the same number of p's and q's) sequence will occur. One could add all the individual probabilities to obtain the probability of all equivalent sequences occurring, or, since each sequence has the same probability, we may simply find the number of such sequences and multiply by the probability associated with the sequence.

a. Permutations and Combinations

The term *permutation* is a special way of describing an arrangement of items. The letters in the word *cat* represent a sequence or permutation, but so do *act*, *tac*, *tca*, *atc*, and *cta*. All of these represent permutations of the same letters. By enumeration we see that there are 6 such permutations in the case of the word *cat*. However, if there are N elements in the sequence, then there will be N! different permutations that can be formed. A simple way to see this is to go about constructing the most general permutation possible. We can begin by selecting the first element of the sequence from any of the n-elements. That means that we would have at least n permutations that begin with one of the n first elements. However, having selected a first element, there are only (n-1) elements left. Thus we will have only (n-1) new permutations for each of our initial n permutations. Having chosen twice only (n-2) elements will remain. each of the n(n-1) permutations generated by the first two choices will yield (n-2) new permutations. This process can be

continued until there are no more elements to select at which point we will have constructed n! distinct permutations.

Now let us generalize this argument where we will pick a sequence of m elements from the original set of n. How many different permutations of m-elements can we build out of n-elements? Again, there are n-ways to select the first element in the permutation leaving (n-1) remaining elements. However, now we do not pick all n-elements, we repeat this process only m-times. Therefore the number of permutations, P^n_m, of n-elements taken m at a time is

$$P^n_m = n(n-1)(n-2) \cdots (n-m+1) = n!/(n-m)! \quad . \tag{7.2.3}$$

A combination is a very different thing than a permutation. When one selects a combination of things, the order of selection is unimportant. If we select a combination of four elements out of twenty, we don't care what order they are selected in only that we ended up with four elements. However, we can ask a question similar to that which we asked for permutations. How many combinations with m-elements can we make from n-elements? Now it is clear why we introduced the notion of a permutation. We may use the result of equation (7.2.3) to answer the question about combinations. Each permutation that is generated in accordance with equation (7.2.3) is a combination. However, since the order in which elements of the combination are selected is unimportant, all permutations with those elements can be considered the same combination. But having picked the m elements, we have already established that there will be m! such permutations. Thus the number of combinations C^n_m of n-elements taken m at a time can be written in terms of the number of permutations as

$$C^n_m = P^n_m/m! = n!/[(n-m)!m!] \quad \binom{n}{m} \quad . \tag{7.2.4}$$

These are often known as the binomial coefficients since they are the coefficients of the binomial series

$$(x+y)^n = C^n_0 x^n + C^n_1 x^{n-1}y + \cdots + C^n_{n-2} x^2 y^{n-1} + C^n_n y^n \quad . \tag{7.2.5}$$

As implied by the last term in equation (7.2.4), the binomial coefficients are often denoted by the symbol $\binom{n}{m}$.

b. The Binomial Probability Distribution

Let us return to the problem of finding the probability of equivalent sequences. Each sequence represents a permutation of the samplings producing events m-times. However, since we are not interested in the order of the sampling, the distinctly different number of sequences is the number of combinations of n-samplings producing m-events. Thus the probability of achieving m-events in n-samplings is

$$P_B(m) = C^n_m p^m q^{n-m} = C^n_m p^m (1-p)^{n-m} \quad , \tag{7.2.6}$$

and is known as the *binomial frequency function*. The probability of having *at least* m-events in n-tries is

$$F(m) = \sum_{i=1}^{m} P(i) = C^n_0 (1-p)^n + C^n_1 p(1-p)^{n-1} + \cdots + C^n_m p^m (1-p)^{n-m} \quad . \tag{7.2.7}$$

and is known as the *binomial distribution*.

Equations (7.2.6) and (7.2.7) are discrete probability functions. Since a great deal of statistical analysis is related to sampling populations where the samples are assumed to be independent of one another, a great deal of emphasis is placed on the binomial distribution. Unfortunately, it is clear from equation (7.2.4) that there will be some difficulties encountered when n is large. Again since many problems involve sampling very large populations, we should pay some attention to this case. In reality, the case when n is large should be considered as two cases; one where the total sample, n, and the product of the sample size and the probability of a single event, np, are both large, and one where n is large but np is not. Let us consider the latter.

c. The Poisson Distribution

By assuming that n is large but np is not we are considering the case where the probability of obtaining a successful event from any particular sampling is very small (i.e. p<<1). A good example of this is the decay of radioactive isotopes. If one focuses on a particular atom in any sample, the probability of decay is nearly zero for any reasonable time. While p is considered small, we will assume both n and m to be large. If m is large, then the interval between m and m+1 (i.e. 1) will be small compared to m and we can replace m with a continuous variable x. Now

$$\frac{n!}{(n-x)!} = n(n-1)(n-2)\ \cdots\ (n-x+1) \cong n^x\ ,\ x >> 1,\ n >> x\ . \tag{7.2.8}$$

With this approximation we can write equation (7.2.6) as

$$P_B(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x} \cong \frac{n^x}{x!} p^x (1-p)^n\ . \tag{7.2.9}$$

The last term can be written as

$$(1\text{-}p)^n = (1\text{-}p)^{\mu/p} = [(1\text{-}p)^{1/p}]^{\mu}\ , \tag{.10}$$

where

$$\mu = np\ . \tag{7.2.11}$$

The meaning of the parameter μ will become apparent later. For the moment it is sufficient to note that it results from the product of a very large number and a very small number. If expand the quantity on the right in brackets by means of the binomial theorem and take the limit as $p \to 0$, we get

$$\lim_{p\to 0}[(1-p)^{1/p} = \lim_{p\to 0}\left(1 - \left[\frac{1}{p}\right]p + \left[\frac{1}{p}\right]\left[\frac{1}{p}-1\right]\left[\frac{p^2}{2!}\right] - \left[\frac{1}{p}\right]\left[\frac{1}{p}-1\right]\left[\frac{1}{p}-2\right]\left[\frac{p^3}{3!}\right] + \cdots + \right) = e^{-1}\ . \tag{7.2.12}$$

Therefore in the limit of vanishing p equation (7.2.9) becomes

$$\lim_{p\to 0} P_B(x) \equiv P_P(x,\mu) = \mu^x e^{-\mu}/x!\ . \tag{7.2.13}$$

$P_P(x,\mu)$ is known as the *Poisson* probability density distribution function. From equation (7.1.8) and equation (7.2.13) one can show that μ is the expected value of x. However, one can see that intuitively from the

definition in equation (7.2.11). Surely if one has a large number of samples n and the probability p that any one of them will produce an event, then the expected number of events will simply be np = μ. The

Poisson distribution function is extremely useful in describing the behavior of unlikely events in large populations. However, in the case where the event is much more likely so that np is large, the situation is somewhat more complicated.

d. The Normal Curve

By assuming that both n and np are large, we move into the realm where all the elements of the binomial coefficients are large. So although the variables are still technically discrete, the unit interval between them remains small compared to their value and we can again replace m by the continuous variable x and np by the continuous variable μ. We can summarize the situation by

$$n >> x \approx np \; \mu >> 1 \quad . \tag{7.1.14}$$

Now we may use Sterling's approximation formula,

$$k! \approx e^{-k} k^{k} \sqrt{2\pi k} \quad , \tag{7.2.15}$$

for large factorials to simplify the binomial coefficients in equation (7.2.9) to get

$$P_B(x) = \frac{n^n p^x q^{(n-x)}}{x^x (n-x)^{(n-x)}} \left(\frac{n}{2\pi k (n-x)} \right)^{1/2} = \left(\frac{np}{x} \right)^x \left(\frac{nq}{n-x} \right)^{n-x} \left(\frac{n}{2\pi k (n-x)} \right)^{1/2} \quad , \tag{7.2.16}$$

Now we add the further restriction that

$$0 < p < 1 \quad . \tag{7.2.17}$$

As in the case of the Poisson distribution, np will be the expectation value of x and it is near that value that we will be most interested in the probability distribution. Thus let us describe x in the vicinity of np by defining a small quantity δ so that

$$\delta = x - np \quad , \tag{7.2.18}$$

and

$$n\text{-}x = n(1\text{-}p) - \delta = nq\text{-}\delta \quad . \tag{7.2.19}$$

Expressing the binomial distribution function given by equation (7.2.16) in terms of δ, we get

$$P_b(x) = \left(1 + \frac{\delta}{np} \right)^{-(\delta + np)} \left(1 - \frac{\delta}{np} \right)^{+(\delta - np)} \left(\frac{n}{2\pi (nq - \delta)(np + \delta)} \right)^{1/2} \quad , \tag{7.2.20}$$

which in terms of logarithms can be written as

$$\ln [P_B(x) Q] \approx - (\delta + np) \ln(1 + \delta/np) - (nq - \delta) \ln(1 - \delta/nq) \, , \tag{7.2.21}$$

where

$$Q = \sqrt{2\pi pq (1 - \delta/nq)(1 + \delta/np)} \quad . \tag{7.2.22}$$

Now we choose to investigate the region in the immediate vicinity of the expected value of x, namely near

np. Therefore δ will remain small so that

$$|\delta| < npq \ . \tag{7.2.23}$$

This implies that

$$\left.\begin{aligned} \left|\frac{\delta}{np}\right| &< 1 \\ \left|\frac{\delta}{nq}\right| &< 1 \end{aligned}\right\} , \tag{7.2.24}$$

and the terms in equations (7.2.21) and (7.2.22) can be approximated by

$$\left.\begin{aligned} Q &\cong \sqrt{2\pi npq} \\ \ln\left[1+\frac{\delta}{np}\right] &\cong \frac{\delta}{np} - \frac{\delta^2}{2n^2p^2} + \cdots + \\ \ln\left[1+\frac{\delta}{nq}\right] &\cong \frac{\delta}{nq} - \frac{\delta^2}{2n^2q^2} + \cdots + \end{aligned}\right\} . \tag{7.2.25}$$

Keeping all terms through second order in δ for the logarithmic expansions, equation (7.2.21) becomes

$$\ln[P_B(x)Q] \approx -(\delta+np)(\delta/np)(1-\delta/2np)+(nq-\delta)(\delta/nq)(1-\delta/2nq) \approx -\delta^2/2npq \ , \tag{7.2.26}$$

so that the binomial distribution function becomes

$$f_B(x) \approx \frac{e^{-\delta^2/2npq}}{\sqrt{2\pi npq}} \ . \tag{7.2.27}$$

Replacing np by μ as we did with the Poisson distribution and defining a new quantity σ by

$$\left.\begin{aligned} \sigma^2 &\equiv 2npq = 2np(1-p) \\ \delta &= x-\mu \end{aligned}\right\} , \tag{7.2.28}$$

we can write equation (7.2.27) as

$$f_N(x) \approx \frac{e^{-(x-\mu)^2/\sigma^2}}{\sqrt{2\pi\sigma}} . \tag{7.2.29}$$

This distribution function is known as the *normal distribution function* or just the *normal curve*. Some texts refer to it as the "Bell-shaped" curve. In reality it is a probability density distribution function since, in considering large n, we have passed to the limit of the continuous random variable. While the normal curve is a function of the continuous random variable x, the curve also depends on the expectation value of x (that is μ) and the probability p of a single sampling yielding an event. The sample set n is assumed to be very much larger than the random variable x which itself is assumed to be very much greater than 1. The meaning of the parameters μ and σ can be seen from Figure 7.2.

Although the normal curve is usually attributed to Laplace, it is its use by Gauss for describing the distribution of experimental or observational error that brought the curve to prominence. It is simply the

large number limit of the discrete binomial probability function. If one makes a series of independent measurements where the error of measurement is randomly distributed about the "true" value, one will obtain an expected value of x equal to μ and the errors will produce a range of values of x having a characteristic width of σ. Used in this context the normal curve is often called the *Gaussian error curve*.

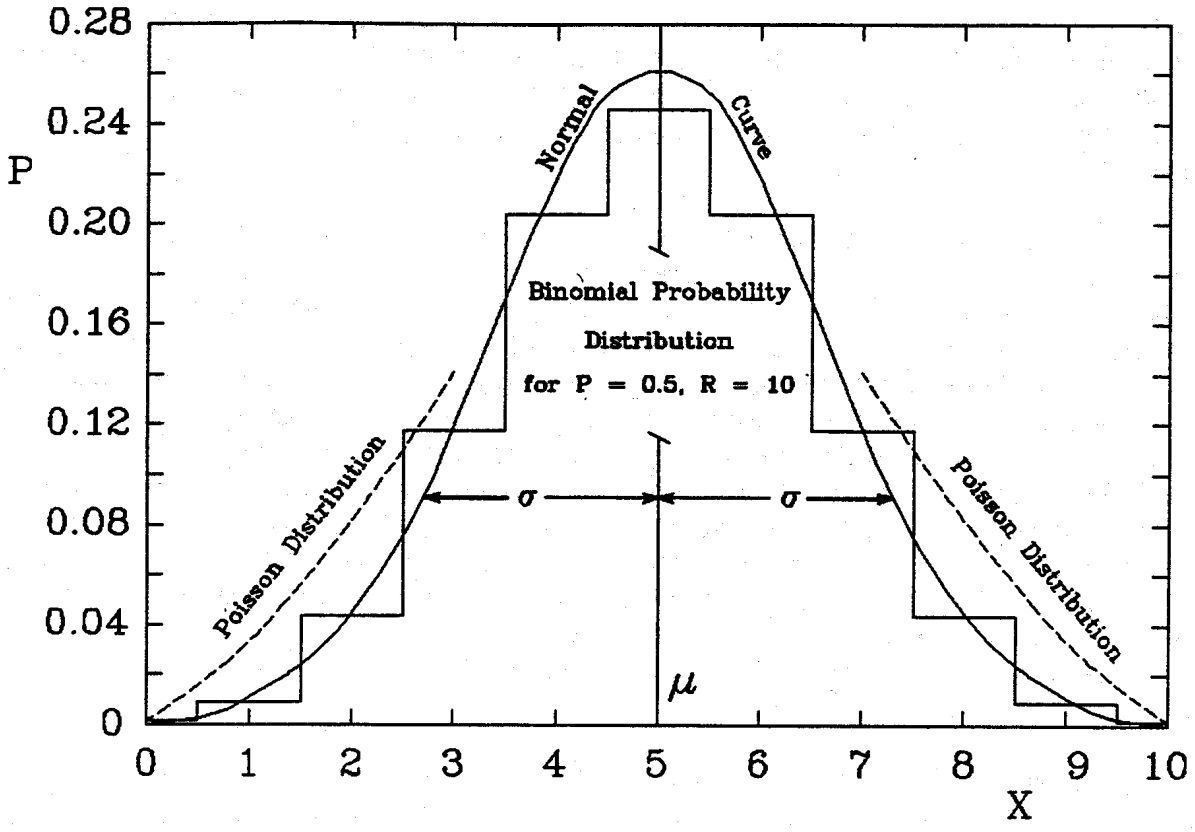

Figure 7.2 shows the normal curve approximation to the binomial probability distribution function. We have chosen the coin tosses so that p = 0.5. *Here* μ *and* σ *can be seen as the most likely value of the random variable* x *and the 'width' of the curve respectively. The tail end of the curve represents the region approximated by the Poisson distribution.*

Because of the basic nature of the sampling assumptions on which it is based, the normal curve plays a major role in testing. This is the curve that students hope will be used when they ask "Will the course be curved?". Of course there are many reasons why a test sample will depart from the normal curve and we will explore some of them in the next chapter. One of the most obvious is that the sample size is small. It should always be remembered that the continuous distribution functions such as the normal curve and the Poisson distribution are approximations which only approach validity when the sample set is very large.

Also, these are not the only distribution functions that arise from probability theory. To demonstrate this point, let us consider some important ones that occur in the physical world.

e. Some Distribution Functions of the Physical World

The foundations of statistical mechanics devote considerable effort to describing the distribution functions for particles that make up our physical world. The random variable that is used turns out to be the total energy of the particles. Most of the details of the derivations are related to the manner by which experiment effectively samples the set of available particles. In the realm of the quantum, the nature of the particles also plays a major role in determining the resulting probability distribution functions. Since the physical world can be viewed as being made up of atomic, or if necessary nuclear, particles, the number of particles in the sample set is usually huge. Therefore the derived distribution functions are usually expressed in terms of functions of the continuous random variable.

Consult a book on statistical mechanics, and you will immediately encounter the terms *microstate*, and *macrostate*. A macrostate is basically a physical distribution of particles with respect to the random variable. A microstate is an artificial concept developed to aid in enumerating the various possible macrostates in the same spirit that permutations aided in the calculation of combinations. The concept of a microstate specifically assumes that the particles are distinguishable. The detailed arrangement of which particles have which values of the random variable determines the microstate. Based on the sampling assumptions, one attempts to find the most probable macrostate which corresponds to the expectation value of the system of particles. In addition, one searches for the number of microstates within a particular macrostate. Since the relative probability of a particular macrostate occurring will be proportional to the number of microstates yielding that macrostate, finding that number is equivalent to finding the probability distribution of macrostates. The most probable macrostate is the one most likely to occur in nature. The basic differences of the distribution functions (i.e. most probable macrostates) that occur can be traced to properties attributed to the particles themselves and to the nature of the space in which they occur.

Consider the total number of particles (N) to be arranged sequentially among m volumes of some space. The total number of sequences or permutations is simply N!. However, within each volume (say the ith volume), there will be N_i particles which yield $N_i!$ indistinguishable sequences which must be removed. If we take the 'volumes' in which we are arranging the particles to be energy w_i then we get the distribution function to be

$$N_i = a_1 e^{-w_i/kT} . \tag{7.2.30}$$

Here T is the temperature of the gas, w_i is the energy of the particles, the constant a_1 depends on the detailed physical makeup of the gas, and k is the Boltzmann constant.

The statistical distribution of particles within the m 'spatial' volumes given by equation (7.2.30) is known as Maxwell-Boltzmann statistics and gives excellent results for a classical gas where the particles can be regarded as distinguishable. In the world of classical physics, the position and momentum of a particle are sufficient to make it distinguishable from all other particles. However, the quantum-mechanical picture of the physical world is quite different and results in different distribution functions. In the world of the

quantum, as a consequence of the Heisenberg uncertainty principle, there is a small volume of 'space' within which particles are indistinguishable. Thus, one may loose any number of particles into one of these 'volumes' and they would all be considered the same kind of particle. Earlier, the sampling order produced permutations that were different from combinations where the sampling order didn't matter. This affected

the probability distributions through the difference between P_m^n and C_m^n. In a similar manner we would expect the distinguishability of particles to affect the nature of the most probable macrostate. In this case the resultant distribution function has the form

$$N_i = a_2 \left(e^{w_i/kT} - 1\right) , \tag{7.2.31}$$

where the parameter a_2 can be determined in terms of the energy of the particles N_i. This is the distribution function that is suitable for the particles of light called photons and any particles that behave like photons. The distribution function is known as the Bose-Einstein distribution function.

Finally if one invokes the Pauli Exclusion Principle that says you can put no more than two of certain kinds of nuclear particles in the minimum volume designated by the Heisenberg uncertainty principle, then the particle distribution function has the form

$$N_i = a_3 \left(e^{w_i/kT} + 1\right) , \tag{7.2.32}$$

This is known as the Fermi-Dirac distribution function and again a_3 is determined by the detailed nature of the particles.

Equations (7.2.30 - 32) are just examples of the kinds of probability distribution functions that occur in nature. There are many more. Clearly the knowledge of the entire distribution function provides all the available information about the sample set. However, much of the important information can be obtained from simpler properties of the distribution function.

7.3 Moments of Distribution Functions

Let us begin by defining what is meant by the moment of a function. The moment of a function is the integral of some property of interest, weighted by its probability density distribution function, over the space for which the distribution function is defined. Common examples of such moments can be found in statistics. The mean, or average of a distribution function is simply the first moment of the distribution function and what is called the variance can be simply related to the second moment. In general, if the distribution function is analytic, all the information contained in the function is also contained in the moments of that function.

One of the most difficult problems in any type of analysis is to know what information is unnecessary for the understanding of the basic nature of a particular phenomenon. In other words, what information can be safely thrown away? The complete probability density distribution function representing some phenomenon contains much more information about the phenomenon than we usually wish to know. The process of integrating the function over its defined space in order to obtain a specific moment removes or averages out much of the information about the function. However, it results in parameters which are much easier to interpret. Thus one trades off information for the ability to utilize the result and obtain some explicit properties of the phenomenon. This is a standard 'trick' of mathematical analysis.

We shall define the kth moment of a function f(x) as

$$M_k = \int x^k f(x)dx \Big/ \int f(x)dx \quad , k \geq 1 \; . \tag{7.3.1}$$

The kth moment then is the kth power of the independent variable averaged over all allowed values of the that variable and weighted by the probability density distribution function. Clearly M_0 is unity as we have chosen to normalize the moment by $\int f(x)dx$. This has the practical advantage of making the units of M_k the same as the units and magnitude of an average of x^k in the occasional situation where f(x) is not a normalized probability density function. If the function f(x) is defined for a range of the independent variable $a \leq x \leq b$, then the moments can be written as

$$\left.\begin{aligned} < x > \equiv M_1 &= \int_a^b xf(x)\,dx \Big/ \int_a^b f(x)\,dx \\ < x^2 > \equiv M_2 &= \int_a^b x^2 f(x)\,dx \Big/ \int_a^b f(x)\,dx \\ &\vdots \qquad \vdots \qquad\qquad \vdots \\ < x^k > \equiv M_k &= \int_a^b x^k f(x)\,dx \Big/ \int_a^b f(x)\,dx \end{aligned}\right\} \; . \tag{7.3.2}$$

In equations (7.3.1) and (7.3.2) we have chosen to define moments of the continuous random variable x which is represented by a probability density distribution function f(x). However, we could just as easily define a set of discrete moments where the integral is replaced by a sum and the probability density distribution function is replaced by the probability of obtaining the particular value of the random variable itself. Such moments would then be written as

$$\overline{x^k} \equiv \sum_{i=1}^{N} x_i^k P(x_i) \Big/ \sum_{i=1}^{N} P(x_i) \; . \tag{7.3.3}$$

If the case where the probability of obtaining the random variable is uniform (which it should be if x is really a random variable), equation (7.3.3) becomes

$$\overline{x^k} \equiv \sum_{i=1}^{N} x_i^k P(x_i) \Big/ N \; . \tag{7.3.4}$$

As we shall see, much of statistical analysis is concerned with deciding when the finite or discrete moment can be taken to represent the continuous moment (i.e. when $\overline{x^k} = <x^k>$).

While a complete knowledge of the moments of a analytic function will enable one to specify the function and hence all the information it contains, it is usually sufficient to specify only a few of the

moments in order to obtain most of that information. Indeed, this is the strength and utility of the concept of moments. Four parameters which characterize a probability density distribution function, and are

commonly used in statistics are the *mean*, *variance*, *skewness*, and *kurtosis*. Figure 7.3 shows a graphical representation of these parameters for an arbitrary distribution function.

These four parameters provide a great deal of information about the probability density distribution function f(x) and they are related to the first four moments of the distribution function. Indeed, the *mean of a function* is simply defined as the first moment and is often denoted by the symbol μ. We have already used the symbol σ to denote the 'width' of the normal curve and it is called *the standard deviation* [see equation (7.2.29) and figure 7.2]. In that instance, the 'width' was a measure of the root-mean-square of the departure of the random variable from the mean. The quantity σ^2 is formally called the *variance of the function* and is defined as

$$\sigma^2 \equiv \int (x-\mu)^2 f(x)\,dx = \int x^2 f(x)\,dx - 2\mu\int x f(x)\,dx + \mu^2 \int f(x)\,dx = < x^2 > -\mu^2 \quad . \qquad (7.3.5)$$

Thus the variance clearly contains the information supplied by the second moment of f(x) and is just the *mean-square minus the square of the mean*. We can define a dimensionless parameter, the *skewness of a function*, as a measure of the cube of the departure of f(x) from its mean value so that

$$s^3 \equiv \int (x-\mu)^3 f(x)\,dx \Big/ \sigma^3 \quad = \quad [< x^3 > -\mu(3 < x^2 > -2\mu^2)]/\sigma^3 = [< x^3 > -\mu(< x^2 > -2\sigma^2)]/\sigma^3 . \qquad (7.3.6)$$

The name skewness given s^3 describes what it measures about the function f(x). If the distribution function is symmetric about μ, then the integrand of the integral in equation (7.3.6) is anti-symmetric and $s = 0$. If the skewness is positive then on average f(x) > f(-x), and the distribution function is 'skewed' to the right. The situation is reversed for $s^3 < 0$. Since this parameter describes an aspect of the relative shape of the distribution function, it should be normalized so that it carries no units. This is the reason for the presence of σ^3 in the denominator of equation (7.3.6).

As one would expect, the kurtosis involves information from the fourth moment of the probability density distribution function. Like the skewness, the kurtosis is dimensionless as it is normalized by the square of the variance. Therefore the *kurtosis of a function* is defined as

$$\beta = \int (x-\mu)^4 f(x)dx \Big/ (\sigma^2)^2 \int f(x)dx = [<x^4>-4\mu<x^3>+6\mu^2<x^2>-3\mu^4] . \qquad (7.3.7)$$

For the normal curve given by equation (7.2.29), β = 3. Thus if β < 3 the distribution function f(x) is 'flatter' in the vicinity of the maximum than the normal curve while β > 3 implies a distribution function that is more sharply peaked. Since a great deal of statistical analysis deals with ascertaining to what extent a sample of events represents a normal probability distribution function, these last two parameters are very helpful tools.

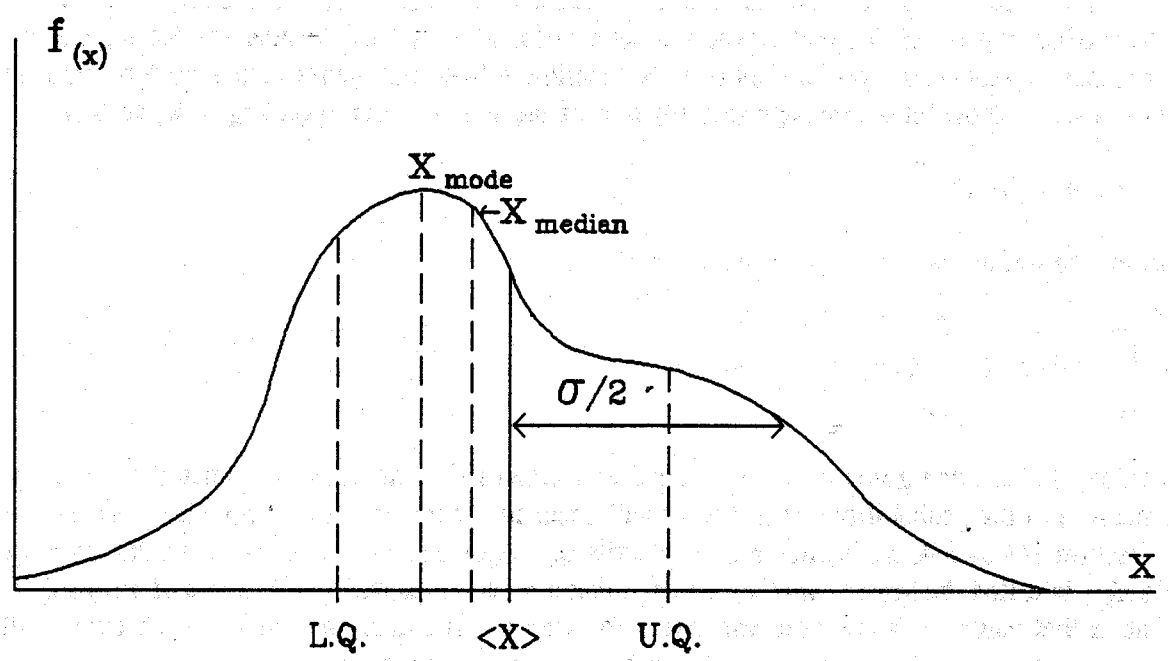

Figure 7.3 shows the mean of a function f(x) *as* <x>. *Note this is not the same as the most likely value of* x *as was the case in Figure 7.2. However, in some real sense* σ *is still a measure of the width of the function. The skewness is a measure of the asymmetry of* f(x) *while the kurtosis represents the degree to which the* f(x) *is 'flattened' with respect to a normal curve. We have also marked the location of the values for the upper and lower quartiles, median and mode.*

There are two other quantities that are often used to characterize a distribution function. These are the median and mode. To understand the notion of median, let us consider the more general concept of a percentile. Consider a probability density function defined for values of the random variable in the interval a ≤ x ≤ b. Now let α represent that fraction of the interval corresponding to x_α so that

$$\alpha = (x_\alpha - a)/(b-a) \quad . \tag{7.3.8}$$

Now we can define the αth *percentile* by

$$\alpha = \int_a^{x_a} f(x)\,dx \Big/ \int_a^b f(x)\,dx \quad . \tag{7.3.9}$$

The value of α is often given in terms of the percentage of the interval $a \rightarrow b$, hence the name for x_α. x_α is a measure of the probability that the event will occur in α-percent of the sample tries. When α is given as a fraction 1/4 or 3/4, x_α is known as a *quartile* Q_α. Specifically $x_{1/4}$ is called the *lower quartile*, while $x_{3/4}$ is called the *upper quartile*. The parameter $x_{1/2}$ acquires the special name of *median*. Thus the median is that value of the random variable x for which it is equally probable that an event will occur with x greater or less than $x_{1/2}$. Thus the median is defined by

$$\tfrac{1}{2} = \int_a^{x_{1/2}} f(x)\,dx \Big/ \int_a^b f(x)\,dx \quad . \tag{7.3.10}$$

Finally, the term *mode* is reserved for the most frequently occurring value of x. This parameter is similar to the expectation value of x discussed in section 7.1 [see equation (7.1.8)]. For continuous distribution functions, this will clearly occur where the curve has a maximum. Thus we may define the *mode of a function* as

$$\left.\frac{df(x)}{dx}\right|_{x=x_m} = 0 \quad . \tag{7.3.11}$$

In this section we have made all of the definitions in terms of the continuous probability density distribution function f(x). The reason for generating these specific parameters is to provide ways of characterizing that function without enumerating it for all values of x. These parameters allow us to compare f(x) to other distribution functions within certain limits and thereby to ascertain the extent to which the conditions that give rise to f(x) correspond to the conditions that yield known probability density distribution functions. Usually one does not have a complete continuous probability density distribution function available for analysis. Instead, one deals with finite samples and attempts to ascertain the nature of the distribution function that governs the results of the sampling. All the parameters defined in this section can be defined for finite samples. Usually the transformation is obvious for those parameters based on moments. Equations (7.3.3) and (7.3.4) give suitable definitions of their discrete definitions. However, in the case of the mode, no simple mathematical formula can be given. It will simply be the most frequently occurring value of the sampled events.

When dealing with finite samples, it is common to define skewness in terms of other more easily calculated parameters of the sample distribution. Some of these definitions are

$$\left.\begin{aligned} s_1 &\equiv (\mu - x_m)/\sigma \\ s_2 &\equiv 3(\mu - x_{1/2})/\sigma \\ s_3 &\equiv 2(x_{3/4} + x_{1/4} - 2x_{1/2})/(x_{3/4} + x_{1/4}) \end{aligned}\right\} \quad . \tag{7.3.12}$$

There are practical reasons for picking any particular one of these definitions, but they are not equivalent so that the user should be careful and consistent when using them.

Let us close this section by considering a hypothetical case of a set of grades given in a course. Suppose that there is a class of ten students who take a twenty-question test with the results given in Table 7.1. Here we encounter a common problem with the use of statistics on small samples. The values for the percentiles do not come out to be integer values so that it is necessary to simply assign them to the nearest integer value. At first look, we find that the median and mode are the same which is required if the scores are to follow the normal curve. However, we might suspect that the curve departs somewhat from the statistically desired result as there are a number of grades that equal the maximum allowed. Therefore let us

consider the moments of the grade distribution as give in Table 7.2

Table 7.1

Grade Distribution for Sample Test Results

STUDENT NO.	PERCENTAGE GRADE	PERCENTILE SCORES
1	100	
2	100	
3	95	Upper Quartile
4	90	
5	85	Median
6	85	Mode
7	85	
8	70	Lower Quartile
9	60	
10	40	

Table 7.2

Examination Statistics for the Sample Test

Statistic	Value
Mode	85
$\overline{x}$	81
$\overline{x^2}$	6890
$\overline{x^3}$	605175
$\overline{x^4}$	54319250
Standard Deviation σ	18.138
Skewness s	-1.041
s_1	-0.221
s_2	0.000
s_3	-0.061
Kurtosis β	3.087

Here we see that the mean is somewhat below the median and mode indicating that there are more extreme negative scores than there are positive ones. Or conversely that a larger fraction of the class has scores above

the mean than below then mean. This is supported by the value for the skewness. However, here we have four different choices to choose from. The values s_i are often used to allow for the small number statistics. While they would tend to imply that the curve is skewed somewhat toward negative numbers in the sense suggested by the relative values of the median and mean, the magnitude is not serious. The value of the Kurtosis is obtained from equation (7.3.7) and suggests that the curve is very similar to a normal curve in its flatness.

Thus the instructor responsible for this test could feel confident that the test grades represent a sample of the parent population. In the next chapter we will investigate quantitatively how secure he or she may be in that regard. However, this begs the issue as to whether or not this is a good test. With the mean at 81, one finds 70% of the class with grades between the mean and the top possible grade of 100. Thus 20% of the grading range has been used to evaluate 70% of the class. Excellent discrimination has been obtained for the lower 30% of the class as their grades are spread over 80% of the possible test range. If the goal of the test is to evaluate the relative performance of the class, the spread in scores indicates that this was not done in a very efficient way. Indeed, for the two students who scored 100, no upper limit on their ability has been established. The examiner when establishing the degree of difficulty of the examination so that uniform discrimination is obtained for all segments of the class should consider such factors.

7.4 The Foundations of Statistical Analysis

In making the transition to finite sample sizes we also make the transition from the theoretical realm of probability theory to the more practical world of statistical analysis. Thus we should spend some time understanding the basic tenets of statistics before we use the results.

In science we never prove a theory or hypothesis correct, we simple add confirmatory evidence to an existing body of evidence that supports the theory or hypothesis. However, we may prove a theory or hypothesis to be incorrect or at least invalid for a particular set of circumstances. We investigate the validity of a hypothesis by carrying out experiments or observations. In its purest form, the act of experimentation can be viewed as the measurement of the values of two supposedly related quantities. The relationship is said to be a functional relationship when the quantities are theoretically related [for example $y=f(x)$] where the relationship involves parameters that are to be determined by the experiment. The entire point of the dual measurement of y and x is to determine those parameters and thereby test the validity of the statement $y=f(x)$. In the physical world no measurement can be carried out with arbitrary precision and therefore there will be errors inherent in both y and x. One of the important roles of statistics is to objectively establish the extent to which the errors affect the determination of the parameters in $f(x)$ and thereby place limits on the extent to which the experiment confirms or rejects the hypothesis. Most statistical analysis is focused on answering the question "To what extent is this experimental result a matter of chance?".

In general, we assume that experiments sample some aspect of the real world producing values of y_i and x_i. We further assume that this sampling could in principle be carried out forever yielding an arbitrarily large set of values of y_i and x_i. In other words there exists an infinite sample space or set which is often called the *parent population*. As a result of sampling error, our sample values will deviate from those of the parent population by an amount, say ε. Each measured value of x_i departs from its 'true' value by some unknown value ε_i. However, we have already seen that if the errors ε_i are not correlated with each other, then

ε_i will be distributed in accordance with the binomial distribution. The notion that we are unbiasedly sampling the parent population basically assumes that our error sample will follow the binomial distribution and this is a central assumption of most statistical analysis. To be sure there are ways we may check the validity of this assumption, but most of the tests comprising statistical inference rely on the assumption being true. It is essentially what we mean when we address the question "To what extent is this experimental result a matter of chance?".

Many students find the terminology of statistics to be a major barrier to understanding the subject. As with any discipline, the specific jargon of the discipline must be understood before any real comprehension can take place. This is particularly true with statistics where the terminology has arisen from many diverse scientific disciplines. We have already noted how a study in population genetics gave rise to the term "regression analysis" to describe the use of Legendre's principle of least squares. Often properly phrased statistical statements will appear awkward in their effort to be precise. This is important for there are multitudinous ways to deceive using statistics badly. This often results from a lack of precision in making a statistical statement or failure to properly address the question "To what extent is this experimental result a matter of chance?".

a. Moments of the Binomial Distribution

Since the binomial distribution, and its associated large sample limit, the normal curve, play such a central role in statistical analysis, we should consider the meaning of the moments of this distribution. As is clear from figure 7.2, the binomial distribution is a symmetric function about its peak value. Thus the mean of the distribution [as given by the first of equations (7.3.2)] will be the peak value of the distribution. From the symmetric nature of the curve, the median will also be the peak value which, in turn, is the mode by definition. Therefore, for the normal curve the median, mean and mode are all equal or

$$\mu_N \equiv \langle x \rangle_N = (x_{½})_N = (x_m)_N \ . \tag{7.4.1}$$

Similarly the various percentiles will be symmetrically placed about the mean. We have already seen that the fourth moment about the mean called the kurtosis takes on the particular value of 3 for the normal curve and it is clear from the symmetry of the normal curve that the skewness will be zero.

The variance σ^2, is simply the square of a characteristic half-width of the curve called the standard deviation σ. Since any area under a normalized probability density distribution function represents the probability that an observation will have a value of x defined by the limits of the area, σ corresponds to the probability that x will lie within σ of μ_N. We may obtain that probability by integrating equation 7.2.29 so that

$$P_N(\sigma) = \frac{1}{\sqrt{2\pi\sigma}} \int_{\mu-\sigma}^{\mu+\sigma} e^{-(x-\mu_N)^2 / 2\sigma^2} dx - \frac{1}{\sqrt{\pi}} \int_{-1}^{+1} e^{-y^2} dy = \mathrm{erf}(1) = 0.68269 \ . \tag{7.4.2}$$

Thus the probability that a particular randomly sampled value of x will fall within σ of the mean value μ, is about 68%. Since this argument applies to the error distribution ε, σ is sometime called the *standard error of estimate*. One could ask "what is the range in x corresponding to a 50% probability of x being within that

value of the mean"? This will clearly be a smaller number than σ since we wish

$$P_N(x_p) = \frac{1}{\sqrt{2\pi\sigma}} \int_{\mu - x_p}^{\mu + x_p} e^{-(x-\mu_N)^2 / 2\sigma^2} dx = \frac{1}{2} \quad . \tag{7.4.3}$$

The quantity x_p is usually called the *probable error*.

$$x_p = 0.6745\sigma \quad . \tag{7.4.4}$$

The use of the probable error is discouraged since it has become associated with statistical arguments here the author chooses the smaller probable error over the more common standard error simply for its psychological effect.

b. Multiple Variables, Variance, and Covariance

We have discussed the behavior of events that can be characterized by a single random variable distributed according to f(x). What are we to do when the event is the result of two or more variables each characterized by their own probability density distribution functions? Say the event y is related to two variables v and w by

$$y = g(v,w) \quad . \tag{7.4.5}$$

If only two variables are involved y is said to have a *bivariant* distribution. Should the event depend on more than two variables, it has a *multivariant* distribution. Such a situation can result from an experiment where more than one variable must be measured simultaneously in order to characterize the result. Consider the Hall effect in physics where a current flowing perpendicular to a magnetic field will generate a voltage in the direction of the field. In order to investigate this effect one must simultaneously measure the strength of the field and the current as well as the resulting voltage. Each of the independent variables v and w will be characterized by probability density distribution functions that reflect the errors of measurement. Each distribution function will be characterized by the moments we developed for the single random variable. Measurement error will affect the values of both the current and magnetic field and it is a fair question to ask how those errors of measurement affect the expected value of the voltage through the function g(v,w).

Let any variation from the means of y, v, and w be denoted by δ. Then the chain rule of calculus guarantees that

$$(\delta y)^2 = \left[\delta v \frac{\partial y}{\partial v} + \delta w \frac{\partial y}{\partial w} \right]^2 = (\delta v)^2 \left[\frac{\partial g}{\partial v} \right]^2 + 2\delta v \delta w \left[\frac{\partial g}{\partial v} \right] \left[\frac{\partial g}{\partial w} \right] + (\delta w)^2 \left[\frac{\partial g}{\partial w} \right]^2 \quad . \tag{7.4.6}$$

Therefore

$$\sigma_y^2 = \sigma_v^2 \left[\frac{\partial g}{\partial v} \right]^2 + 2\sigma_{vw}^2 \left[\frac{\partial g}{\partial v} \right] \left[\frac{\partial g}{\partial w} \right] + \sigma_w^2 \left[\frac{\partial g}{\partial w} \right]^2 \quad . \tag{7.4.7}$$

Here we have introduced the parameter σ_{vw}^2 which is called the *coefficient of covariance*, or just the *covariance*, as it measures the combined variations from the mean of the variables v and w. For continuous random variables v and w, the coefficient of covariance is defined by

$$\sigma_{vw}^2 \equiv \int\int (v-\mu_v)(w-\mu_w) f(v) h(w)\, dv dw \quad . \tag{7.4.8}$$

Here f(v) and h(v) are the normalized probability density distribution functions of v and w respectively. The coefficient of covariance can be defined over a finite data set as

$$\sigma_{vw}^2 = \sum_{i=1}^{N} (v_i - \mu_v)(w_i - \mu_w) \Big/ N \quad . \tag{7.4.9}$$

Unlike the variance, which in some sense measures the variation of a single y variable against itself, the terms that make up the covariance can be either positive or negative. Indeed, if the probability density distribution functions that govern v and w are symmetric about the mean, then $\sigma_{vw}^2 = 0$. If this is true for a multivariant distribution function, then all the covariances will be zero and

$$\sigma_y^2 = \sum_{i=1}^{N} \sigma_{x_k}^2 \left(\frac{\partial g}{\partial x_k} \right)^2 \quad . \tag{7.4.10}$$

This is a result similar to that obtained in section 6.3 [see equations (6.3.9) - (6.3.11)] for the errors of the least square coefficients and rests on the same assumption of error symmetry. Indeed, we shall see in the next chapter that there is a very close relation between linear least squares, and the statistical methods of regression analysis and analysis of variance.

When one is discussing the moments and properties of the normal curve, there is no question as to the their value. This is a result of the infinite sample size and therefore is not realized for actual cases where the sample is finite. Thus there will be an uncertainty resulting from the error of the sampled items in the mean as well as other moments and it is a fair question to ask how that uncertainty can be estimated. Let us regard the determination of the mean from a finite sample to be the result of a multivariant analysis where

$$\mu = g(x_i) = \sum_{i=1}^{N} x_i \Big/ N \quad . \tag{7.4.11}$$

The partial derivative required by equation (7.4.10) will then yield

$$\frac{\partial g}{\partial x_k} = \frac{1}{N} \quad , \tag{7.4.12}$$

and taking $y = \mu$ we get the variance of the mean to be

$$\sigma_\mu^2 = \sum_{i=1}^{N} \frac{\sigma_{x_k}^2}{N^2} = \frac{\sigma^2}{N} \quad ; \tag{7.4.13}$$

the different observations are all of the same parameter x, and the values of $\sigma^2_{x_k}$ will all be equal.

In order to evaluate the variance of the mean σ^2_μ directly, we require an expression for the variance of a single observation for a finite sample of data. Equation (7.3.5) assumes that the value of the mean is known with absolute precision and so its generalization to a finite data set will underestimate the actual spread in the finite distribution function. Say we were to use one of our observations to specify the value of the mean. That observation would no longer be available to determine other statistical parameters as it could no longer be regarded as independent. So the total number of independent observations would now be N-1 and we could write the variance of a single observation as

$$\sigma_x^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{(N-1)} \quad . \tag{7.4.14}$$

Therefore, the variance of the mean becomes

$$\sigma_\mu^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N(N-1)} \quad . \tag{7.4.15}$$

The factor of (N-1) in the denominator results from the uncertainty of the mean itself. The number of independent observations that go into a statistical analysis are often referred to as the number of *degrees of freedom* of the analysis. Since the equivalent of one observation is required to specify the mean, one degree of freedom is removed from further analysis. It is that degree of freedom required to specify the value of the mean. At any point in a statistical analysis one should always be concerned with the number of degrees of freedom available to specify the solution to the problem. In some real sense, the number of degrees of freedom represents the extent to which the problem is over-determined in the absence of error. Thus in a least square problem with n coefficients to be determined from N data points, there are only (N-n) degrees of freedom. This is the statistical origin of the factor of (N-n) in equation (6.3.26) that specifies the error in the least square coefficients.

c. Maximum Likelihood

Most of statistics is devoted to determining the extent to which a sample population represents the parent population. A corollary to this task is the problem of determining the extent to which the parent population is represented by a normal distribution. We have already seen that the mean, mode, and median are all equal for a normal distribution. This means that the most probable value (i.e. the expectation value) of x is obtained from the mean, median, or mode. For a finite population, these three will not, in general be equal. Is there some way to decide if the differences result simply from chance and a finite random sample, or whether the parent population is not represented by the normal curve? One approach is to reverse the question and ask, "What is the likelihood that the finite sample will result in a particular value for the mean, median, mode or any other statistic?". To answer this question assumes that the probability density distribution for the parent population is known. If this is the case, then one can calculate the probability that a sample of known size (and characteristics) will result from sampling that distribution. Indeed the logarithm of that probability is known as the *likelihood* of the statistic. The value of the likelihood will depend on the particular value of the statistic, which should not be regarded as a variable, as well as the nature of the probability distribution of the parent population. Maximum likelihood algorithms are those that adjust the sampling procedure within the constraints imposed by the definition of the statistic so as to maximize the likelihood of obtaining a particular statistic when sampling the parent population.

Assume that we are interested in determining the most probable value of an event from a sample of a parent population, which does not follow the normal curve. If the distribution function is not symmetric about the mean, then the arithmetic mean will not, in general, be the most probable result (see figure 7.3). However, if we knew the nature of the distribution function of the parent population (i.e. its shape, not its exact values) we could devise a sampling procedure that yielded an accurate value for the mode, which then would be the most probable value for the sampled event. If the probability density function of the parent population is the normal curve, then the mean is that value. In the case of multivariant analysis, least-squares

yields the maximum likelihood values for the coefficients when the parent populations of the various variables are represented by the normal curve.

In the next chapter we will consider some specific ways of determining the nature of the parent population and the extent to which we can believe that the values of the moments accurately sample the parent population. In addition, we will also deal with the problem of multivariant analysis, small sample size and other practical problems of statistical analysis.

Chapter 7 Exercises

1. Find the probability that, from a deck of 52 playing cards, a person can draw exactly:

    a. a pair,
    b. three of a kind,
    c. four of a kind.

2. Calculate the probability that a person sitting third from the dealer in a four person game will be dealt five cards containing:

    a. a pair,
    b. three of a kind,
    c. four of a kind.

    What is the effect of having additional players in the game? Does it matter where the player is located with respect to the other players? If so, why?

3. What is the probability that a single person can draw a five-card straight *or* a flush from a single deck of cards?

4. Calculate the binomial probability distribution function of obtaining "heads" for ten throws of an unbiased coin.

5. Show explicitly how the skewness and the kurtosis are related to the third and fourth moments of the distribution function. Express them in terms of these moments and the mean and variance. Re-express the kurtosis in terms of the fourth moment, the mean variance and skewness.

6. Show that the value for the kurtosis of the normal curve is 3.

7. Obtain expressions for:

    a. the variance of the skewness of a finite sample,
    b. the variance of the kurtosis of a finite sample.

Chapter 7 References and Supplemental Reading

1. Eddington, Sir A.S. "The Philosophy of Physical Science" (1939)

2. Smith, J.G., and Duncan, A.J. "Elementary Statistics and Applications: Fundamentals of the Theory of Statistics", (1944), Mc Graw-Hill Book Company Inc., New York, London, pp. 323.

The basics of probability theory and statistics can be found is a very large number of books. The student should try to find one that is slanted to his/her particular area of interest. Below are a few that he/she may find useful.

1. DeGroot, M.H., "Probability and Statistics" (1975), Addison-Wesley Pub. Co. Inc., Reading, Mass.

2. Miller, I.R., Freund, J.E., and Johnson,R., "Probability and Statistics for Engineers", 4th ed., (1990), Prentice-Hall, Inc. Englewood Cliffs, N.J.

3. Rice, J.A. "Mathematical Statistics and Data Analysis", (1988), Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove Cal.

4. Devore, J.L., "Probability and Statistics for Engineering and the Sciences", 2nd ed., (1987), Brooks/Cole Publishing Co. Inc. Monterey Cal.

5. Larsen, R.J., and Marx, M.L., An Introduction to Mathematical Statistics and Its Applications", 2nd ed., (1986) Prentice-Hall, Englewood Cliffs, N.J.