CHAPTER SEVEN

an introduction to statistics

The *statistician* suggests probabilistic models of reality and investigates their validity. He does so in an attempt to gain insight into the behavior of physical systems and to facilitate better predictions and decisions regarding these systems. A primary concern of statistics is *statistical inference*, the drawing of inferences from data.

The discussions in this chapter are brief, based on simple examples, somewhat incomplete, and always at an introductory level. Our major objectives are (1) to introduce some of the fundamental issues and methods of statistics and (2) to indicate the nature of the transition required as one moves from probability theory to its applications for *statistical reasoning*.

We begin with a few comments on the relation between statistics and probability theory. After identifying some prime issues of concern in statistical investigations, we consider common methods for the study of these issues. These methods generally represent the viewpoint of *classical* statistics. Our concluding sections serve as a brief introduction to the developing field of *Bayesian* (or *modern*) statistics.

7-1 Statistics Is Different

Probability theory is axiomatic. Fully defined probability problems have unique and precise solutions. So far we have dealt with problems which are wholly abstract, although they have often been based on probabilistic *models* of reality.

The field of statistics is different. Statistics is concerned with the relation of such models to actual physical systems. The methods employed by the statistician are arbitrary ways of *being reasonable* in the application of probability theory to physical situations. His primary tools are probability theory, a mathematical sophistication, and common sense.

To use an extreme example, there simply is no unique *best* or *correct* way to extrapolate the gross national product five years hence from three days of rainfall data. In fact, there is no best way to predict the rainfall for the fourth day. But there are many ways to try.

7-2 Statistical Models and Some Related Issues

In contrast to our work in previous chapters, we are now concerned both with models of reality and reality itself. It is important that we keep in mind the differences between the statistician's model (and its implications) and the actual physical situation that is being modeled.

In the real world, we may design and perform experiments. We may observe certain *characteristics of interest* of the experimental outcomes. If we are studying the behavior of a coin of suspicious origin, a characteristic of interest might be the number of heads observed in a certain number of tosses. If we are testing a vaccine, one characteristic of interest could be the observed immunity rates in a control group and in a vaccinated group.

What is the nature of the statistician's model? From whatever knowledge he has of the physical mechanisms involved and from his past experience, the statistician postulates a probabilistic model for the system of interest. He anticipates that this model will exhibit a probabilistic behavior *in the characteristics of interest* similar to that of the physical system. The details of the model might or might not be closely related to the actual nature of the physical system.

If the statistician is concerned with the coin of suspicious origin, he might suggest a model which is a Bernoulli process with probability P for a head on any toss. For the study of the vaccine, he might suggest a model which assigns a probability of immunity P_1 to each member of the control group and assigns a probability of immunity P_2 to each member of the vaccinated group.

We shall consider some of the questions which the statistician asks about his models and learn how he employs experimental data to explore these questions.

1 Based on some experimental data, does a certain model seem reasonable or at least not particularly unreasonable? This is the domain of *significance testing*. In a significance test, the statistician speculates on the likelihood that data similar to that actually observed would be generated by *hypothetical* experiments with the model.

2 Based on some experimental data, how do we express a preference among several postulated models? (These models might be similar models differing only in the values of their parameters.) When one deals with a selection among several hypothesized models, he is involved in a matter of *hypothesis testing*. We shall learn that hypothesis testing and significance testing are very closely related.

3 Given the form of a postulated model of the physical system and some experimental data, how may the data be employed to establish the most desirable values of the parameters of the model? This question would arise, for example, if we considered the Bernoulli model for flips of the suspicious coin and wished to adjust parameter P to make the model as compatible as possible with the experimental data. This is the domain of *estimation*.

4 We may be uncertain of the appropriate parameters for our model. However, from previous experience with the physical system and from other information, we may have convictions about a reasonable PDF for these parameters (which are, to us, random variables). The field of *Bayesian analysis* develops an efficient framework for combining such "prior knowledge" with experimental data. Bayesian analysis is particularly suitable for investigations which must result in *decisions* among several possible future courses of action.

The remainder of this book is concerned with the four issues introduced above. The results we shall obtain are based on *subjective* applications of concepts of probability theory.

7-3 Statistics: Sample Values and Experimental Values

In previous chapters, the phrase "experimental value" always applied to what we might now consider to be *the outcome of a hypothetical experiment with a model of a physical system*. Since it is important that we be able to distinguish between consequences of a model and consequences of reality, we establish two definitions.

EXPERIMENTAL VALUE: Refers to actual data which must, of course, be obtained by the performance of (real) experiments with a physical system

SAMPLE VALUE: Refers to the outcome resulting from the performance of (hypothetical) experiments with a model of a physical system

These particular definitions are not universal in the literature, but they will provide us with an explicit language.

Suppose that we perform a hypothetical experiment with our model n times. Let random variable x be the characteristic of interest defined on the possible experimental outcomes. We use the notation x_i to denote the random variable defined on the ith performance of this hypothetical experiment. The set of random variables $(x_1,x_2, \ldots ,x_n)$ is defined to be a *sample of size n* of random variable x. A sample of size n is a collection of random variables whose probabilistic behavior is specified by our model. Hypothesizing a model is equivalent to specifying a compound PDF for the members of the sample.

We shall use the word *statistic* to describe any function of some random variables, $q(u,v,w, \ldots)$. We may use for the argument of a statistic either the members of a sample or actual experimental values of the random variables. The former case results in what is known as a *sample value* of the statistic. When experimental values are used for $u,v,w, \ldots$, we obtain an *experimental value* of the statistic. Given a specific model for consideration, we may, in principle, derive the PDF for the sample value of any statistic from the compound PDF for the members of the sample. If our model happens to be correct, this PDF would also describe the experimental value of the statistic.

Much of the field of statistics hinges on the following three steps:

1 Postulate a model for the physical system of interest.
2 Based on this model, select a desirable statistic for which:
   The PDF for the sample value of the statistic may be calculated in a useful form.
   Experimental values of the statistic may be obtained from reality.
3 Obtain an experimental value of the statistic, and comment on the likelihood that a similar value would result from the use of the proposed model instead of reality.

The operation of deriving the PDF's and their means and variances for useful statistics is often very complicated, but there are a few cases of frequent interest for which some of these calculations are not too involved. *Assuming that the x_i's in our sample are always independent and identically distributed,* we present some examples.

One fundamental statistic of the sample $(x_1,x_2, \ldots ,x_n)$ is the *sample mean* M_n, whose definition, expected value, and variance were introduced in Sec. 6-3

$$M_n = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad E(M_n) = E(x) \qquad \sigma_{M_n}^2 = \frac{\sigma_x^2}{n}$$

and our proof, for the case $\sigma_x^2 < \infty$, showed that M_n obeyed (at least) the weak law of large numbers. If characteristic x is in fact described by *any* PDF with a finite variance, we can with high probability use M_n as a good estimate of $E(x)$ by using a large value of n, since we know that M_n converges stochastically to $E(x)$.

It is often difficult to determine the exact expression for $f_{M_n}(M)$, the PDF for the sample mean. Quite often we turn to the central limit theorem for an approximation to this PDF. Our interests in the PDF's for particular statistics will become clear in later sections.

Another important statistic is S_n^2, the *sample variance*. The definition of this particular random variable is given by

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - M_n)^2$$

where M_n is the sample mean as defined earlier. We may expand the above expression,

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2}{n}M_n\sum_{i=1}^{n} x_i + M_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - M_n^2$$

This is a more useful form of S_n^2 for the calculation of its expectation

$$E(S_n^2) = \frac{1}{n}E\left(\sum_{i=1}^{n} x_i^2\right) - E(M_n^2)$$

The expectation in the first term is the expected value of a sum and may be simplified by

$$E\left(\sum_{i=1}^{n} x_i^2\right) = E(x_1^2 + x_2^2 + \cdots + x_n^2) = nE(x^2)$$

The calculation of $E(M_n^2)$ requires a few intermediate steps,

$$E(M_n^2) = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] = E\left(\frac{1}{n^2}\sum_{i=1}^{n} x_i^2\right) + E\left(\frac{1}{n^2}\sum_{l=1}^{n}\sum_{\substack{j=1\\ l\neq j}}^{n} x_l x_j\right)$$

$$= \left(\frac{1}{n}\right)^2 \{nE(x^2) + n(n-1)[E(x)]^2\}$$

In the last term of the above expression, we have used the fact that, for $l \neq j$, x_l and x_j are independent random variables. Returning to our expression for $E(S_n^2)$, the expected value of the sample variance, we have

$$E(S_n^2) = \frac{1}{n}nE(x^2) - \frac{1}{n}E(x^2) - \left(1 - \frac{1}{n}\right)[E(x)]^2$$

$$= \left(1 - \frac{1}{n}\right)\{E(x^2) - [E(x)]^2\} = \frac{n-1}{n}\sigma_x^2$$

Thus, we see that for samples of a large size, the *expected value* of the sample variance is very close to the variance of random variable x. The poor agreement between $E(S_n^2)$ and σ_x^2 for small n is most reasonable when one considers the definition of the sample variance for a sample of size 1.

We shall not investigate the variance of the sample variance. However, the reader should realize that a result obtainable from the previous equation, namely,

$$\lim_{n\to\infty} E(S_n^2) = \sigma_x^2$$

does not necessarily mean, in itself, that an experimental value of S_n^2 for large n is with high probability a good estimate of σ_x^2. We would need to establish that S_n^2 at least obeys a weak law of large numbers before we could have confidence in an experimental value of S_n^2 (for large n) as a *good* estimator of σ_x^2. For instance, $E(S_n^2) \approx \sigma_x^2$ for large n does not even require that the variance of S_n^2 be finite.

7-4 Significance Testing

Assume that, as a result of preliminary modeling efforts, we have proposed a model for a physical system and we are able to determine the PDF for the sample value of q, the statistic we have selected. In *significance testing,* we work in the event space for statistic q, using this PDF, which would also hold for experimental values of q *if* our model were correct. We wish to evaluate the hypothesis that our model is correct.

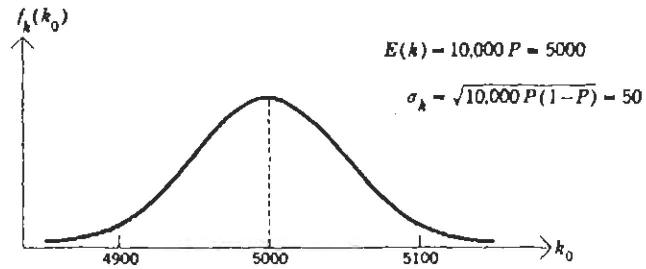
In the event space for q we define an event W, known as the *improbable* event. We may select for our improbable event any particular event of probability α, where α is known as the *level of significance* of the test. *After* event W has been selected, we obtain an experimental value of statistic q. Depending on whether or not the experimental value of q falls within the improbable event W, we reach one of two conclusions as a result of the significance test. These conclusions are

1 *Rejection of the hypothesis.* The experimental value q fell within the improbable event W. If our hypothesized model were correct, our observed experimental value of the statistic would be an improbable result. Since we did in fact obtain such an experimental value, we believe it to be unlikely that our hypothesis is correct.

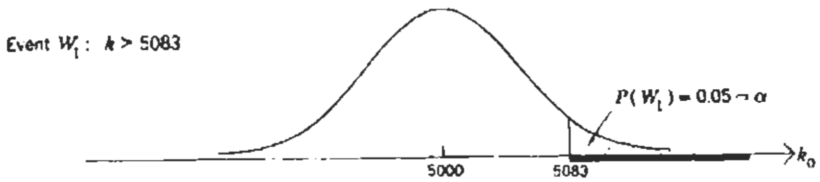
2 *Acceptance of the hypothesis.* The experimental value of q fell in W'. If our hypothesis were true, the observed experimental value of the statistic would not be an improbable event. Since we did in fact obtain such an experimental value, the significance test has not provided us with any particular reason to doubt the hypothesis.

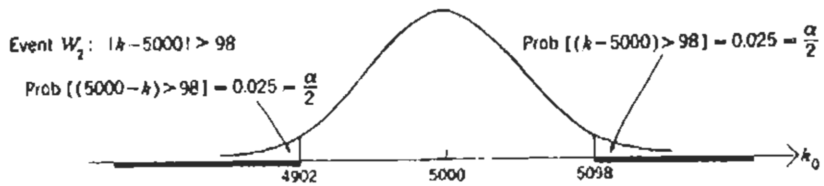
We discuss some examples and further details, deferring general comments until we are more familiar with significance testing.

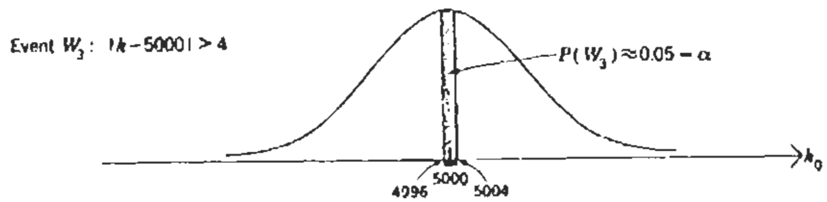
Suppose that we are studying a coin-flipping process to test the hypothesis that the process is a Bernoulli process composed of fair ($P = \frac{1}{2}$) trials. Eventually, we shall observe 10,000 flips, and we have selected as our statistic k the number of heads in 10,000 flips. Using the central limit theorem, we may, for our purposes, approximate the sample value of k as a continuous random variable with a Gaussian PDF as shown below:



Thus we have the conditional PDF for statistic k, given our hypothesis is correct. If we set α, the probability of the "improbable" event at 0.05, many events could serve as the improbable event W. Several such choices for W are shown below in an event space for k, with $P(W)$ indicated by the area under the PDF $f_k(k_0)$.







The heavy line appearing on the k_0 axis in each of the above sketches represents one possible selection of an improbable event at the 0.05 level of significance.

That part of the event space for a statistic which is included in the improbable event W is called the *critical range* of the statistic. If the experimental value of the statistic falls into the critical region, the hypothesis is "rejected"; otherwise it is "accepted." Note that the level of significance of the test is actually equal to the conditional probability that a hypothesis will be rejected, given that it is correct.

A reasonable choice of the improbable event must depend on the actual problem at hand. *In a significance test, one is, in effect, testing his hypothesis against all other hypotheses, with no particular alternatives in mind.* If the hypothesis being tested is not correct, some other hypothesis (stated or unstated) is correct. The critical region is placed where we believe other hypotheses are more likely to place the experimental value of the statistic than is the particular hypothesis under test. This may be viewed as "setting a trap" for outcomes due to other hypotheses, and it often results in a decision to make the acceptance region W' as small as possible. There can be no escape from the fact that this type of statistical reasoning is necessarily an arbitrary

and subjective procedure, but it is a procedure that most people would consider superior to guessing.

Let's return to the example of the coin-tossing process and assume that we have agreed to set α, the level of significance (or the conditional probability of the improbable event, given that the model is correct) equal to 0.05. If we have no general feelings about possible alternative hypotheses, we would expect to trap most other hypotheses most often by making our acceptance region, the complement of the critical region, as small as possible. For this purpose, we would select W_3 in the above sketch as our choice for the improbable event W.

If we had suspicions that the most likely alternative hypotheses were of the form "P greater than 0.5," we would want the critical region to cover values of our statistic most favored by the alternative hypotheses. We would therefore select W_1 of the three improbable events shown above. Most often, however, significance testing refers to testing one hypothesis with no others in mind, and the acceptance region is generally made as small as possible for a given level of significance.

The choice of the level of significance is rather arbitrary. There are a few popular conventional values of α, and these include 0.05, 0.02, and 0.01. The smaller the level of significance, the less likely we are to reject our hypothesis if it is true and the more likely we are to accept our hypothesis if it is false. In most cases, one would expect the choice of the level of significance to depend on the relative costs of the two possible types of errors which may result from the test, *false acceptance* of the hypothesis and *false rejection* of the hypothesis.

Consider one additional example of the specification of a significance test. Suppose that our model for characteristic x of a certain process is that x is a random variable described by a Gaussian PDF with $\sigma_x = 1$. We have

$$f_x(x_0) = \frac{1}{\sqrt{2\pi}} e^{-(x_0-r)^2/2} \qquad -\infty \le x_0 \le \infty$$

and we do not know the value of r, the expected value of x. Ten independent experimental values of characteristic x have been obtained, and we wish to test the hypothesis that parameter r is equal to zero.

Using x_i as the notation for the ith experimental value of x, we arbitrarily elect to use the statistic

$$y = x_1 + x_2 + \cdots + x_{10}$$

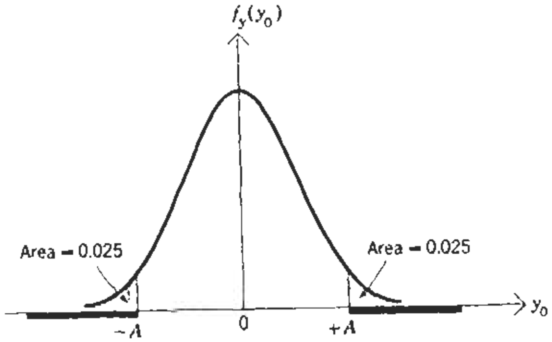
since we know (from the properties of sums of independent Gaussian random variables) that the PDF for the sample value of y is

$$f_y(y_0) = \frac{1}{\sqrt{2\pi}\sqrt{10}} e^{-(y_0-10r)^2/(2\cdot 10)} \qquad -\infty \le y_0 \le \infty$$

In a significance test we work with the conditional PDF for our statistic, given that our hypothesis is true. For this example, we have

$$f_y(y_0) = \frac{1}{\sqrt{2\pi}\,\sqrt{10}} e^{-y_0^2/20} \qquad -\infty \le y_0 \le \infty$$

Assume that we have decided to test at the 0.05 level of significance and that, with no particular properties of the possible alternative hypotheses in mind, we choose to make the acceptance region for the hypotheses in mind, we choose to make the acceptance region for the significance test as small as possible. This leads to a rejection region of the form $|y| > A$. The following sketch applies,



and A is determined by

$$\text{Prob}(y \ge A) = 0.025 = 1 - \Phi\left(\frac{A-0}{\sqrt{10}}\right)$$

$$\Phi\left(\frac{A}{\sqrt{10}}\right) = 0.975 \qquad A \approx 6.2 \qquad \text{from table in Sec. 6-4}$$

Thus, at the 0.05 level, we shall reject our hypothesis that $E(x) = 0$ if it happens that the magnitude of the sum of the 10 experimental values of x is greater than 6.2.

We conclude this section with several brief comments on significance testing:

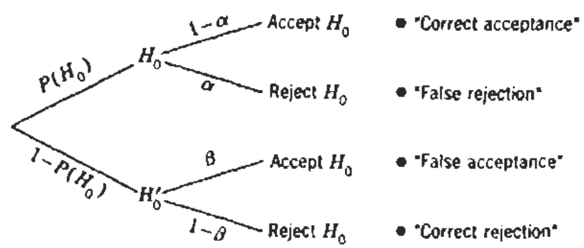
1 The use of different statistics, based on samples of the same size and the same experimental values, may result in different conclusions from the significance test, even if the acceptance regions for both statistics are made as small as possible (see Prob. 7.06).

2 In our examples, it happened that the only parameter in the PDF's for the statistics was the one whose value was specified by the hypothesis. In the above example, if σ_x^2 were not specified and we wished to make no assumptions about it, we would have had to *try* to find a statistic whose PDF depended on $E(x)$ but not on σ_x^2.

3 Even if the outcome of a significance test results in acceptance of the hypothesis, there are probably many other more accurate (and less accurate) hypotheses which would also be accepted as the result of similar significance tests upon them.

4 Because of the imprecise statement of the alternative hypotheses for a significance test, there is little we can say in general about the relative desirability of several possible statistics based on samples of the same size. One desires a statistic which, in its event space, *discriminates* as sharply as possible between his hypothesis and other hypotheses. In almost all situations, increasing the size of the sample will contribute to this discrimination.

5 The formulation of a significance test does not allow us to determine the a priori probability that a significance test will result in an incorrect conclusion. Even if we can agree to accept an a priori probability $P(H_0)$ that the hypothesis H_0 is true (before we undertake the test), we are still unable to evaluate the probability of an incorrect outcome of the significance test. Consider the following sequential event space picture for any significance test:



The lack of specific alternatives to H_0 prevents us from calculating a priori probabilities for the bottom two event points, even if we accept a value (or range of values) for $P(H_0)$. We have no way to estimate β, the conditional probability of acceptance of H_0 given H_0 is incorrect.

6 One value of significance testing is that it often leads one to discard particularly poor hypotheses. In most cases, statistics based on large enough samples are excellent for this purpose, and this is achieved with a rather small number of assumptions about the situation under study.

7-5 Parametric and Nonparametric Hypotheses

Two examples of significance tests were considered in the previous section. In both cases, the PDF for the statistic resulting from the model contained a parameter. In the first example, the parameter was P, the probability of success for a Bernoulli process. In the second example, the parameter of the PDF for the statistic was r, the expected value of a Gaussian random variable. The significance tests were performed on hypotheses which specified values for these parameters.

If, in effect, we assume the given form of a model and test hypotheses which specify values for parameters of the model, we say that we are testing *parametric* hypotheses. The hypotheses in both examples were parametric hypotheses.

Nonparametric hypotheses are of a broader nature, often with regard to the general form of a model or the form of the resulting PDF for the characteristic of interest. The following are some typical nonparametric hypotheses:

1 Characteristic x is normally distributed.
2 Random variables x and y have identical marginal PDF's, that is, $f_x(u) = f_y(u)$ for all values of u.
3 Random variables x and y have unequal expected values.
4 The variance of random variable x is greater than the variance of random variable y.

In principle, significance testing for parametric and nonparametric hypotheses follows exactly the same procedure. In practice, the determination of useful statistics for nonparametric tests is often a very difficult task. To be useful, the PDF's for such statistics must not depend on unknown quantities. Furthermore, one strives to make as few additional assumptions as possible before testing nonparametric hypotheses. Several nonparametric methods of great practical value, however, may be found in most elementary statistics texts.

7-6 Hypothesis Testing

The term *significance test* normally refers to the evaluation of a hypothesis H_0 in the absence of any useful information about alternative hypotheses. An evaluation of H_0 in a situation where the alternative hypotheses $H_1, H_2, \ldots$ are specified is known as a *hypothesis test.*

In this section we discuss the situation where it is known that there are only two possible parametric hypotheses $H_0(Q = Q_0)$ and $H_1(Q = Q_1)$. We are using Q to denote the parameter of interest.

To perform a hypothesis test, we select one of the hypotheses, H_0 (called the *null* hypothesis), and subject it to a significance test based on some statistic q. If the experimental value of statistic q falls into the critical (or rejection) region W, defined (as in Sec. 7-4) by

$$\text{Prob}(q \text{ in } W \mid H_0) = \text{Prob}(q \text{ in } W \mid Q = Q_0) = \alpha$$

we shall "reject" H_0 and "accept" H_1. Otherwise we shall accept H_0 and reject H_1. In order to discuss the choice of the "best" possible critical region W for a given statistic in the presence of a specific alternative hypothesis H_1, consider the two possible errors which may result from the outcome of a hypothesis test.

Suppose that H_0 were true. If this were so, the only possible error would be to reject H_0 in favor of H_1. The conditional probability of this type of error (called an *error of type* I, or *false rejection*) given H_0 is true is

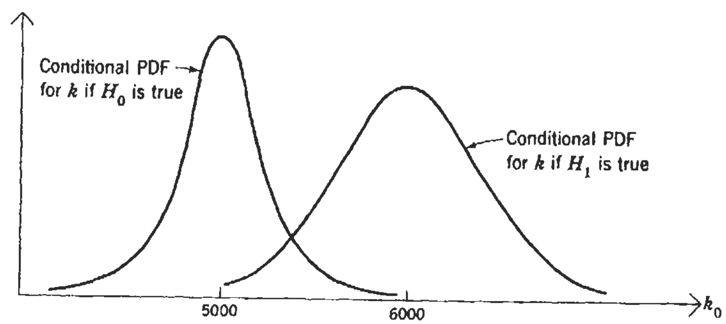
$$\text{Prob}(\text{reject } H_0 \mid Q = Q_0) = \text{Prob}(q \text{ in } W \mid Q = Q_0) = \alpha$$

Suppose that H_0 is false and H_1 is true. Then the only type of error we could make would be to accept H_0 and reject H_1. The conditional probability of this type of error (called an *error of type* II, or *false acceptance*) given H_1 is true is

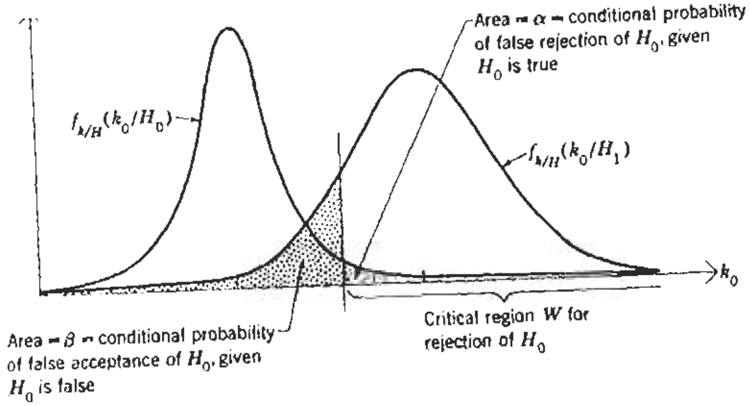
$$\text{Prob}(\text{accept } H_0 \mid Q = Q_1) = \text{Prob}(q \text{ not in } W \mid Q = Q_1) = \beta$$

It is important to realize that α and β are conditional probabilities which apply in different conditional event spaces. Furthermore, for *significance* testing (in Sec. 7-4) we did not know enough about the alternative hypotheses to be able to evaluate β. When we are concerned with a hypothesis test, this is no longer the case.

Let's return to the example of 10,000 coin tosses and a Bernoulli model of the process. Assume that we consider only the two alternative hypotheses $H_0(P = 0.5)$ and $H_1(P = 0.6)$. These hypotheses lead to two alternative conditional PDF's for k, the number of heads. We have

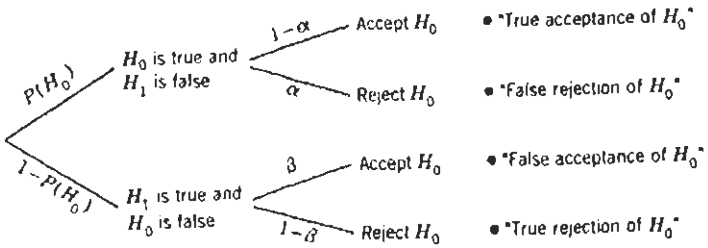


In this case, for any given α (the conditional probability of false rejection) we desire to select a critical region which will minimize β (the conditional probability of false acceptance). It should be clear that, for this example, the most desirable critical region W for a given α will be a continuous range of k on the right. For a given value of α, we may now identify α and β as areas under the conditional PDF's for k, as shown below:



In practice, the selection of a pair of values α and β would usually depend on the relative costs of the two possible types of errors and some a priori estimate of the probability that H_0 is true (see Prob. 7.10).

Consider a sequential event space for the performance of a hypothesis test upon H_0 with one specific alternative hypothesis H_1:



If we are willing to assign an a priori probability $P(H_0)$ to the validity of H_0, we may then state that the probability (to us) that this hypothesis test will result in an incorrect conclusion is equal to

$$\alpha P(H_0) + \beta[1 - P(H_0)]$$

Even if we are uncomfortable with any step which involves the assumption of $P(H_0)$, we may still use the fact that

$$0 \leq P(H_0) \leq 1$$

and the previous expression to obtain the bounds

$$\min(\alpha,\beta) \leq \text{Prob(incorrect conclusion)} \leq \max(\alpha,\beta)$$

We now comment on the selection of the statistic q. For any

hypothesis test, a desirable statistic would be one which provides good discrimination between H_0 and H_1. For one thing, we would like the ratio

$$\frac{f_{q|H_0}(q_0 \mid H_0)}{f_{q|H_1}(q_0 \mid H_1)}$$

to be as large as possible in the acceptance region W' and to be as small as possible in the rejection region W. This would mean that, for any experimental value of statistic q, we would be relatively unlikely to accept the wrong hypothesis.

We might decide that the best statistic, q, is one which (for a given sample size of a given observable characteristic) provides the minimum β for any given α. Even when such a best statistic does exist, however, the derivation of the form of this best statistic and its conditional PDF's may be very difficult.

7-7 Estimation

Assume that we have developed the form of a model for a physical process and that we wish to determine the most desirable values for some parameters of this model. The general theory of using experimental data to estimate such parameters is known as the *theory of estimation*.

When we perform a hypothesis test with a rich set of alternatives, the validity of several suggested forms of a model may be under question. For our discussion of estimation, we shall take the viewpoint that the general form of our model is not to be questioned. We wish here only to estimate certain parameters of the process, given that the form of the model is correct. Since stating the form of the model is equivalent to stating the form of the PDF for characteristic x of the process, determining the parameters of the model is similar to adjusting the parameters of the PDF to best accommodate the experimental data.

Let $Q_n(x_1,x_2, \ldots ,x_n)$ be a statistic whose sample values are a function of a sample of size n and whose experimental values are a function of n independent experimental values of random variable x. Let Q be a parameter of our model or of its resulting PDF for random variable x. We shall be interested in those statistics Q_n whose experimental values happen to be good estimates of parameter Q. Such statistics are known as *estimators*.

Some examples of useful estimators follow. We might use the average value of n experimental values of x, given by

$$Q_n = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = M_n$$

as an estimate of the parameter $E(x)$. We have already encountered this statistic several times. [Although it is, alas, known as the *sample mean* (M_n), we must realize that, like any other statistic, it has both sample and experimental values. A similar comment applies to our next example of an estimator.] Another example of a statistic which may serve as an estimator is that of the use of the sample variance, given by

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - M_n)^2 = S_n^2$$

to estimate the variance of the PDF for random variable x. For a final example, we might use the maximum of n experimental values of x, given by

$$Q_n = \max (x_1,x_2, \ldots ,x_n)$$

to estimate the largest possible experimental value of random variable x.

Often we are able to suggest many reasonable estimators for a particular parameter Q. Suppose, for instance, that it is known that $f_x(x_0)$ is symmetric about $E(x)$, that is,

$$f_x[E(x) + a] = f_x[E(x) - a] \qquad \text{for all } a$$

and we wish to estimate $E(x)$ using some estimator $Q_n(x_1,x_2, \ldots ,x_n)$. We might use the estimator

$$Q_{n1} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

or the estimator

$$Q_{n2} = \frac{\max (x_1,x_2, \ldots ,x_n) - \min (x_1,x_2, \ldots ,x_n)}{2}$$

or we could list the x_i in increasing order by defining

$$y_i = i\text{th smallest member of } (x_1,x_2, \ldots ,x_n)$$

and use for our estimator of $E(x)$ the statistic

$$Q_{n3} = \begin{cases} y_{(n+1)/2} & n \text{ odd} \\ \frac{1}{2}(y_{n/2} + y_{(n+2)/2}) & n \text{ even} \end{cases}$$

Any of these three estimators might turn out to be the most desirable, depending on what else is known about the form of $f_x(x_0)$ and also depending, of course, on our criterion for desirability.

In the following section, we introduce some of the properties relevant to the selection and evaluation of useful estimators.

7-8 Some Properties of Desirable Estimators

A sequence of estimates $Q_1, Q_2, \ldots$ of parameter Q is called *consistent* if it converges stochastically to Q as $n \to \infty$. That is, Q_n is a consistent estimator of Q if

$$\lim_{n\to\infty} \text{Prob}(|Q_n - Q| > \epsilon) = 0 \qquad \text{for any } \epsilon > 0$$

In Chap. 6, we proved that, given that σ_x^2 is finite, the sample mean M_n is stochastically convergent to $E(x)$. Thus, the sample mean is a consistent estimator of $E(x)$. If an estimator is known to be consistent, we would become confident of the accuracy of estimates based on very large samples. However, consistency is a limit property and may not be relevant for small samples.

A sequence of estimates $Q_1, Q_2, \ldots$ of parameter Q is called *unbiased* if the expected value of Q_n is equal to Q for all values

$$n = 1, 2, \ldots$$

That is, Q_n is an unbiased estimate for Q if

$$E(Q_n) = Q \qquad \text{for } n = 1, 2, \ldots$$

We noted (Sec. 7-3) that the sample mean M_n is an unbiased estimator for $E(x)$. We also noted that, for the expected value of the sample variance, we have

$$E(S_n^2) = \frac{n-1}{n} \sigma_x^2$$

and thus the sample variance is not an unbiased estimator of σ_x^2. However, it is true that

$$\lim_{n\to\infty} E(S_n^2) = \sigma_x^2$$

Any such estimator Q_n, which obeys

$$\lim_{n\to\infty} E(Q_n) = Q$$

is said to be an *asymptotically unbiased* estimator of Q. If Q_n is an unbiased (or asymptotically unbiased) estimator of Q, this property alone does not assure us of a good estimate when n is very large. We should also need some evidence that, as n grows, the PDF for Q_n becomes adequately concentrated near parameter Q.

The *relative efficiency* of two unbiased estimators is simply the ratio of their variances. We would expect that, the smaller the variance of an unbiased estimator Q_n, the more likely it is that an experimental value of Q_n will give an accurate estimate of parameter Q. We would say the *most efficient* unbiased estimator for Q is the unbiased estimator with the minimum variance.

We now discuss the concept of a *sufficient* estimator. Consider the n-dimensional sample space for the values $x_1, x_2, \ldots, x_n$. In general, when we go from a point in this space to the corresponding value of the estimator $Q_n(x_1,x_2, \ldots ,x_n)$, one of two things must happen. Given that our model is correct, either Q_n contains all the information in the experimental outcome $(x_1,x_2, \ldots ,x_n)$ relevant to the estimation of parameter Q, or it does not. For example, it is true for some estimation problems (and not for some others) that

$$Q_n = \sum_{i=1}^{n} x_i$$

contains all the information relevant to the estimation of Q which may be found in $(x_1,x_2, \ldots ,x_n)$. The reason we are interested in this matter is that we would expect to make the best use of experimental data by using estimators which take advantage of *all* relevant information in the data. Such estimators are known as sufficient estimators. The formal definition of sufficiency does not follow in a simple form from this intuitive discussion.

To state the mathematical definition of a sufficient estimator, we shall use the notation

$\underline{x} = \underline{x_1 \quad x_2 \quad \cdots \quad x_n}$ representing an n-dimensional random variable

$\underline{x_0} = \underline{x_{10} \quad x_{20} \quad \cdots \quad x_{n0}}$ representing any particular value of $\underline{x}$

Our model provides us with a PDF for $\underline{x}$ in terms of a parameter Q which we wish to estimate. This PDF for $\underline{x}$ may be written as

$f_{\underline{x}}(\underline{x_0}) = g(\underline{x_0},Q)$ where g is a function only of $\underline{x_0}$ and Q

If we are given the experimental value of estimator Q_n, this is at least partial information about $\underline{x}$ and we could hope to use it to calculate the resulting conditional PDF for $\underline{x}$,

$f_{\underline{x}|Q_n}(\underline{x_0} \mid Q_n) = h(\underline{x_0},Q,Q_n)$

where h is a function only of $\underline{x_0}$, Q, and Q_n. If and only if the PDF h does not depend on parameter Q after the value of Q_n is given, we *define* Q_n to be a sufficient estimator for parameter Q.

A few comments may help to explain the apparent distance between our simple intuitive notion of a sufficient statistic and the

formal definition in the above paragraph. We are estimating Q because we do not know its value. Let us accept for a moment the notion that Q is (to us) a random variable and that our knowledge about it is given by some a priori PDF. When we say that a sufficient estimator Q_n will contain all the information about Q which is to be found in $(x_1,x_2, \ldots ,x_n)$, the implication is that the conditional PDF for Q, given Q_n, will be identical to the conditional PDF for Q, given the values $(x_1,x_2, \ldots ,x_n)$. Because classical statistics does not provide a framework for viewing our uncertainties about unknown constants in terms of such PDF's, the above definition has to be worked around to be in terms of other PDF's. Instead of stating that Q_n tells us everything about Q which might be found in $(x_1,x_2, \ldots ,x_n)$, our formal definition states that Q_n tells us everything about $(x_1,x_2, \ldots ,x_n)$ that we could find out by knowing Q.

In this section we have discussed the concepts of consistency, bias, relative efficiency, and sufficiency of estimators. We should also note that actual estimates are normally accompanied by *confidence limits.* The statistician specifies a quantity δ for which, given that his model is correct, the probability that the "random interval" $Q_n \pm \delta$ will fall such that it happens to include the true value of parameter Q is equal to some value such as 0.95 or 0.98. Note that it is the location of the interval centered about the experimental value of the estimator, and not the true value of parameter Q, which is considered to be the random phenomenon when one states confidence limits. We shall not explore the actual calculation of confidence limits in this text. Although there are a few special (simple) cases, the general problem is of an advanced nature.

7-9 Maximum-likelihood Estimation

There are several ways to obtain a desirable estimate for Q, an unknown parameter of a proposed statistical model. One method of estimation will be introduced in this section. A rather different approach will be indicated in our discussion of Bayesian analysis.

To use the method of *maximum-likelihood estimation,* we first obtain an experimental value for some sample $(x_1,x_2, \ldots ,x_n)$. We then determine which of all possible values of parameter Q maximizes the *a priori* probability of the observed experimental value of the sample (or of some statistic of the sample). Quantity Q^*, that possible value of Q which maximizes this a priori probability, is known as the maximum-likelihood estimator for parameter Q.

The a priori probability of the observed experimental outcome is calculated under the assumption that the model is correct. Before expanding on the above definition (which is somewhat incomplete) and commenting upon the method, we consider a simple example.

Suppose that we are considering a Bernoulli process as the model for a series of coin flips and that we wish to estimate parameter P, the probability of heads (or success), by the method of maximum-likelihood. Our experiment will be the performance of n flips of the coin and our sample $(x_1,x_2, \ldots ,x_n)$ represents the exact sequence of resulting Bernoulli random variables.

The a priori probability of any particular sequence of experimental outcomes which contains exactly k heads out of a total of n flips is given by

$$P^k(1 - P)^{n-k} \qquad k = 0, 1, \ldots , n$$

To find P^*, the maximum-likelihood estimator for P, we use elementary calculus to determine which value of P, in the range $0 \leq P \leq 1$, maximizes the above a priori probability for any experimental value of k. Differentiating with respect to P, setting the derivative equal to zero, and checking that we are in fact maximizing the above expression, we finally obtain

$$P^* = \frac{k}{n}$$

which is the maximum-likelihood estimator for parameter P if we observe exactly k heads during the n trials.

In our earlier discussion of the Bernoulli law of large numbers (Sec. 6-3) we established that this particular maximum-likelihood estimator satisfies the definition of a consistent estimator. By performing the calculation

$$E(P^*) = E\left(\frac{k}{n}\right) = \frac{1}{n} E(k) = \frac{nP}{n} = P$$

we find that this estimator is also unbiased.

Note also that, for this example, maximum-likelihood estimation based on either of two different statistics will result in the same expression for P^*. We may use an n-dimensional statistic (the sample itself) which is a finest-grain description of the experimental outcome or we may use the alternative statistic k, the number of heads observed. (It happens that k/n is a sufficient estimator for parameter P of a Bernoulli process.)

We now make a necessary expansion of our original definition of maximum-likelihood estimation. If the model under consideration results in a continuous PDF for the statistic of interest, the probability associated with any particular experimental value of the statistic is

zero. For this case, let us, for an n-dimensional statistic, view the problem in an n-dimensional event space whose coordinates represent the n components of the statistic. Our procedure will be to determine that possible value of Q which maximizes the a priori probability of the event represented by an n-dimensional incremental cube, centered about the point in the event space which represents the observed experimental value of the statistic.

The procedure in the preceding paragraph is entirely similar to the procedure used earlier for maximum-likelihood estimation when the statistic is described by a PMF. For the continuous case, we work with incremental events centered about the event point representing the observed experimental outcome. The result can be restated in a simple manner. If the statistic employed is described by a continuous PDF, we maximize the appropriate *PDF evaluated at*, rather than the *probability of*, the observed experimental outcome.

As an example, suppose that our model for an interarrival process is that the process is Poisson. This assumption models the first-order interarrival times as independent random variables, each with PDF

$$f_x(x_0) = \lambda e^{-\lambda x_0} \qquad x_0 > 0$$

In order to estimate λ, we shall consider a sample (r,s,t,u,v) composed of five independent values of random variable x. Our statistic is the sample itself. The compound PDF for this statistic is given by

$$f_{r,s,t,u,v}(r_0,s_0,t_0,u_0,v_0) = \begin{cases} \lambda e^{-\lambda r_0}\lambda e^{-\lambda s_0}\lambda e^{-\lambda t_0}\lambda e^{-\lambda u_0}\lambda e^{-\lambda v_0} & \text{if } r_0,\ s_0,\ t_0,\ u_0,\ v_0 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \lambda^5 e^{-\lambda(r_0+s_0+t_0+u_0+v_0)} & \text{if } r_0,\ s_0,\ t_0,\ u_0,\ v_0 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Maximization of this PDF with respect to λ for any particular experimental outcome (r_0,s_0,t_0,u_0,v_0) leads to the maximum-likelihood estimator

$$\lambda^* = \frac{5}{r_0 + s_0 + t_0 + u_0 + v_0}$$

which seems reasonable, since this result states that the maximum-likelihood estimator of the average arrival rate happens to be equal to the experimental value of the average arrival rate. [We used the (r,s,t,u,v) notation instead of $(x_1,x_2, \ldots ,x_n)$ to enable us to write out the compound PDF for the sample in our more usual notation.]

Problem 7.15 assists the reader to show that λ^* is a consistent estimator which is biased but asymptotically unbiased. It also happens that λ^* (the number of interarrival times divided by their sum) is a sufficient estimator for parameter λ.

In general, maximum-likelihood estimators can be shown to have a surprising number of useful properties, both with regard to theoretical matters and with regard to the simplicity of practical application of the method. For situations involving very large samples, there are few people who disagree with the reasoning which gives rise to this arbitrary but most useful estimation technique.

However, serious problems do arise if one attempts to use this estimation technique for decision problems involving small samples or if one attempts to establish that maximum likelihood is a truly *fundamental* technique involving fewer assumptions than other methods of estimation.

Suppose that we have to make a large wager based on the true value of P in the above coin example. There is time to flip the coin only five times, and we observe four heads. Very few people would be willing to use the maximum-likelihood estimate for P, $\frac{4}{5}$, as their estimator for parameter P if there were large stakes involved in the accuracy of their estimate. Since maximum likelihood depends on a simple maximization of an *unweighted* PDF, there seems to be an uncomfortable implication that *all* possible values of parameter P were equally likely *before* the experiment was performed. We shall return to this matter in our discussion of Bayesian analysis.

7-10 Bayesian Analysis

A *Bayesian* believes that any quantity whose value he does not know is (to him) a random variable. He believes that it is possible, at any time, to express his state of knowledge about such a random variable in the form of a PDF. As additional experimental evidence becomes available, Bayes' theorem is used to combine this evidence with the previous PDF in order to obtain a new a posteriori PDF representing his updated state of knowledge. The PDF expressing the analyst's state of knowledge serves as the quantitative basis for any *decisions* he is required to make.

Consider the Bayesian analysis of Q, an unknown parameter of a postulated probabilistic model of a physical system. We assume that the outcomes of experiments with the system may be described by the resulting experimental values of continuous random variable x, the characteristic of interest.

Based on past experience and all other available information, the Bayesian approach begins with the specification of a PDF $f_Q(Q_0)$, the analyst's a priori PDF for the value of parameter Q. As before, the model specifies the PDF for the sample value of characteristic x, given the value of parameter Q. Since we are now regarding Q as another random variable, the PDF for the sample value of x with parameter Q is to be written as the conditional PDF,

$f_{x|Q}(x_0 \mid Q_0) =$ conditional PDF for the sample value of characteristic x, given that the value of parameter Q is equal to Q_0

Each time an experimental value of characteristic x is obtained, the continuous form of Bayes' theorem

$$f_{Q|x}(Q_0 \mid x_0) = \frac{f_{x,Q}(x_0,Q_0)}{f_x(x_0)} = \frac{f_{x|Q}(x_0 \mid Q_0)f_Q(Q_0)}{\int_{Q_0} f_{x|Q}(x_0 \mid Q_0)f_Q(Q_0)\,dQ_0} = f'_Q(Q_0)$$

is used to obtain the a posteriori PDF $f'_Q(Q_0)$, describing the analyst's new state of knowledge about the value of parameter Q. This PDF $f'_Q(Q_0)$ serves as the basis for any present decisions and also as the *a priori* PDF for any future experimentation with the physical system.

The Bayesian analyst utilizes his state-of-knowledge PDF to resolve issues such as:

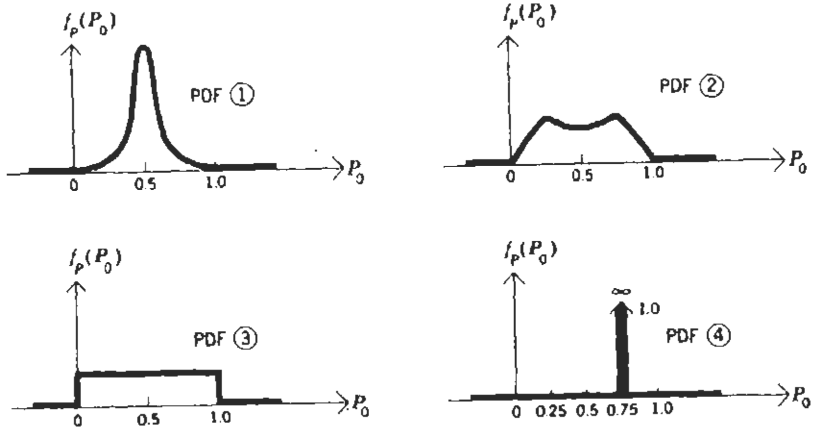
1 Given a function $C(Q^t,Q^*)$, which represents the penalty associated with estimating Q^t, the true value of parameter Q, by an estimate Q^*, determine that estimator Q^* which minimizes the expected value of $C(Q^t,Q^*)$. (For example, see Prob. 2.05.)

2 Given the function $C(Q^t,Q^*)$, which represents the cost of imperfect estimation, and given another function which represents, as a function of n, the cost of n repeated experiments on the physical system, specify the experimental test program which will minimize the expected value of the total cost of experimentation and estimation.

As one example of Bayesian analysis, assume that a Bernoulli model has been accepted for a coin-flipping process and that we wish to investigate parameter P, the probability of success (heads) for this model. We shall discuss only a few aspects of this problem. One should keep in mind that there is probably a cost associated with each flip of the coin and that our general objective is to combine our prior convictions about the value of P with some experimental evidence to obtain a suitably accurate and economical estimate P^*.

The Bayesian analyst begins by stating his entire assumptive structure in the form of his a priori PDF $f_P(P_0)$. Although this is necessarily an inexact and somewhat arbitrary specification, no estimation procedure, classical or Bayesian, can avoid this (or an equivalent) step. We continue with the example, deferring a more general discussion to Sec. 7-12.

Four of many possible choices for $f_P(P_0)$ are shown below:



A priori PDF ① could represent the prior convictions of one who believes, "Almost all coins are fair or very nearly fair, and I don't see anything special about this coin." If it is believed that the coin is probably biased, but the direction of the bias is unknown, PDF ② might serve as $f_P(P_0)$. There might be a person who claims, "I don't know anything about parameter P, and the least biased approach is represented by PDF ③." Finally, PDF ④ is the a priori state of knowledge for a person who is certain that the value of P is equal to 0.75. In fact, since PDF ④ allocates all its probability to this single possible value of P, there is nothing to be learned from experimentation. For PDF ④, the a posteriori PDF will be identical to the a priori PDF, no matter what experimental outcomes may be obtained.

Because the design of complete test programs is too involved for our introductory discussion, assume that some external considerations have dictated that the coin is to be flipped exactly N_0 times. We wish to see how the experimental results (exactly K_0 heads in N_0 tosses) are used to update the original a priori PDF $f_P(P_0)$.

The Bernoulli model of the process leads us to the relation

$$p_{K|P}(K_0 \mid P_0) = \binom{N_0}{K_0} P_0^{K_0}(1 - P_0)^{N_0-K_0} \qquad K_0 = 0, 1, 2, \ldots, N_0$$

where we are using a PMF because of the discrete nature of K, the characteristic of interest. The equation for using the experimental outcome to update the a priori PDF $f_P(P_0)$ to obtain the a posteriori PDF $f'_P(P_0)$ is thus, from substitution into the continuous form of Bayes' theorem, found to be,

$$f'_P(P_0) = f_{P|K}(P_0 \mid K_0) = \frac{\binom{N_0}{K_0} P_0^{K_0}(1 - P_0)^{N_0-K_0} f_P(P_0)}{\int_{P_0=0}^{1} \binom{N_0}{K_0} P_0^{K_0}(1 - P_0)^{N_0-K_0} f_P(P_0)\, dP_0}$$

and we shall continue our consideration of this relation in the following section.

In general, we would expect that, the narrower the a priori PDF $f_P(P_0)$, the more the experimental evidence required to obtain an a posteriori PDF which is appreciably different from the a priori PDF. For very large amounts of experimental data, we would expect the effect of this evidence to dominate all but the most unreasonable a priori PDF's, with the a posteriori PDF $f'_P(P_0)$ becoming heavily concentrated near the true value of parameter P.

7-11 Complementary PDF's for Bayesian Analysis

For the Bayesian analysis of certain parameters of common probabilistic processes, such as the situation in the example of Sec. 7-10, some convenient and efficient procedures have been developed.

The general calculation for an a posteriori PDF is unpleasant, and although it may be performed for any a priori PDF, it is unlikely to yield results in a useful form. To simplify his computational burden, the Bayesian often takes advantage of the obviously imprecise specification of his a priori state of knowledge. In particular, he elects that, whenever it is possible, he will select the a priori PDF from a family of PDF's which has the following three properties:

1 The family should be rich enough to allow him to come reasonably close to a statement of his subjective state of knowledge.
2 Individual members of the family should be determined by specifying the value of a few parameters. It would not be realistic to pretend that the a priori PDF represents very precise information.
3 The family should make the above updating calculation as simple as possible. In particular, if one member of the family is used as the a priori PDF, then, for any possible experimental outcome, the resulting a posteriori PDF should simply be another member of the family. One should be able to carry out the updating calculation by merely using the experimental results to modify the parameters of the a priori PDF to obtain the a posteriori PDF.

The third item in the above list is clearly a big order. We shall not investigate the existence and derivation of such families here. However, when families of PDF's with this property do exist for the estimation of parameters of probabilistic processes, such PDF's are said to be *complementary* (or *conjugate*) PDF's for the process being studied. A demonstration will be presented for the example of the previous section.

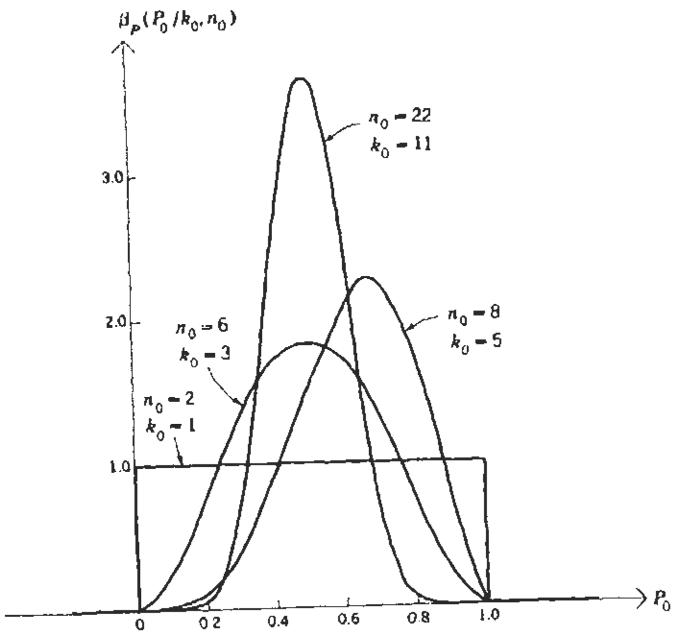
Consider the *beta* PDF for random variable P with parameters k_0 and n_0. It is convenient to write this PDF as $\mathcal{B}_P(P_0 \mid k_0, n_0)$, defined by

$$\mathcal{B}_P(P_0 \mid k_0, n_0) = C(k_0,n_0)P_0^{k_0-1}(1 - P_0)^{n_0-k_0-1} \qquad \begin{cases} 0 \le P_0 \le 1 \\ k_0 \ge 0 \\ n_0 \ge k_0 \end{cases}$$

where $C(k_0,n_0)$ is simply the normalization constant

$$C(k_0,n_0) = \left[\int_{P_0=0}^{1} P_0^{k_0-1}(1 - P_0)^{n_0-k_0-1}\,dP_0\right]^{-1}$$

and several members of this family of PDF's are shown below:



An individual member of this family may be specified by selecting values for its mean and variance rather than by selecting constants k_0 and n_0 directly. Although techniques have been developed to allow far more structured PDF's, the Bayesian often finds that these two parameters $E(P)$ and σ_P^2 allow for an adequate expression of his prior beliefs about the unknown parameter P of a Bernoulli model.

Direct substitution into the relation for $f'_P(P_0)$, the a posteriori

PDF for our example, establishes that if the Bayesian starts out with the a priori PDF

$$f_P(P_0) = \mathcal{B}_P(P_0 \mid k_0, n_0)$$

and then observes exactly K_0 successes in N_0 Bernoulli trials, the resulting a posteriori PDF is

$$f'_P(P_0) = f_{P|K}(P_0 \mid K_0) = \mathcal{B}_P(P_0 \mid k_0 + K_0, n_0 + N_0)$$

Thus, for the estimation of parameter P for a Bernoulli model, use of a beta PDF for $f_P(P_0)$ allows the a posteriori PDF to be determined by merely using the experimental values K_0 and N_0 to modify the parameters of the a priori PDF. Using the above sketch, we see, for instance, that, if $f_P(P_0)$ were the beta PDF with $k_0 = 3$ and $n_0 = 6$, an experimental outcome of two successes in two trials would lead to the a posteriori beta PDF with $k_0 = 5$ and $n_0 = 8$.

It is often the case, as it is for our example, that the determination of parameters of the a priori PDF can be interpreted as assuming a certain "equivalent past experience." For instance, if the cost structure is such that we shall choose our best estimate of parameter P to be the expectation of the a posteriori PDF, the resulting estimate of parameter P, which we call P^*, turns out to be

$$P^* = \frac{K_0 + k_0}{N_0 + n_0}$$

This same result could have been obtained by the method of maximum-likelihood estimation, had we agreed to combine a bias of k_0 successes in n_0 hypothetical trials with the actual experimental data.

Finally, we remark that the use of the beta family for estimating parameter P of a Bernoulli process has several other advantages. It renders quite simple the otherwise most awkward calculations for what is known as *preposterior analysis*. This term refers to an exploration of the nature of the a posteriori PDF and its consequences *before* the tests are performed. It is this feature which allows one to optimize a test program and design effective experiments without becoming bogged down in hopelessly involved detailed calculations.

7-12 Some Comments on Bayesian Analysis and Classical Statistics

There is a large literature, both mathematical and philosophical, dealing with the relationship between classical statistics and Bayesian analysis. In order to indicate some of the considerations in a relatively brief manner, some imprecise generalizations necessarily appear in the following discussion.

The Bayesian approach represents a significant departure from the more conservative classical techniques of statistical analysis. Classical techniques are often particularly appropriate for purely scientific investigations and for matters involving large samples. Classical procedures attempt to require the least severe possible assumptive structure on the part of the analyst. Bayesian analysis involves a more specific assumptive structure and is often described as being *decision-oriented*. Some of the most productive applications of the Bayesian approach are found in situations where prior convictions and a relatively small amount of experimentation must be combined in a rational manner to make decisions among alternative future courses of action.

There is appreciable controversy about the degree of the difference between classical and Bayesian statistics. The Bayesian states his entire assumptive structure in his a priori PDF; his methods require no further arbitrary steps once this PDF is specified. It is true that he is often willing to state a rather sharp a priori PDF which heavily weights his prior convictions. But the Bayesian also points out that all statistical procedures of any type involve similar (although possibly weaker) statements of prior convictions. The assumptive structures of classical statistics are less visible, being somewhat submerged in established statistical tests and the choice of statistics.

Any two Bayesians would begin their analyses of the same problem with somewhat different a priori PDF's. If their work led to conflicting terminal decisions, their different assumptions are apparent in their a priori PDF's and they have a clear common ground for further discussions. The common ground between two different classical procedures which result in conflicting advice tends to be less apparent.

Objection is frequently made to the arbitrary nature of the a priori PDF used by the Bayesian. One frequently hears that this provides an arbitrary bias to what might otherwise be a scientific investigation. The Bayesian replies that all tests involve a form of bias and that he prefers that *his* bias be rational. For instance, in considering the method of maximum likelihood for the estimation of parameter P of a Bernoulli process, we noted the implication that all possible values of P were equally likely before the experiments. Otherwise, the method of maximum likelihood would maximize a *weighted* form of that function of P which represents the a priori probability of the observed experimental outcome.

Continuing this line of thought, the Bayesian contends that, for anybody who has ever seen a coin, how could any bias be less rational than that of a priori PDF ③ in the example of Sec. 7-10? Finally, he would note that there is nothing fundamental in starting out with a uniform PDF over the possible values of P as a manifestation of "minimum bias." Parameter P is but one arbitrary way to characterize the process; other parameters might be, for example,

$$U = \frac{1}{P} \qquad V = P^2 + P \ln (P + 1)$$

and professing that all possible values of one of these parameters be equally likely would lead to different results from those obtained by assuming the uniform PDF over all possible values of parameter P. The Bayesian believes that, since it is impossible to avoid bias, one can do no better than to assume a rational rather than naïve form of bias.

We should remark in closing that, because we considered a particularly simple estimation problem, we had at our disposal highly developed Bayesian procedures. For multivariate problems or for tests of nonparametric hypotheses, useful Bayesian formulations do not necessarily exist.

PROBLEMS

7.01 Random variable M_n, the sample mean, is defined to be the average value of n independent experimental values of random variable x. Determine the *exact* PDF (or PMF) for M_n and its expected value and variance if:

a $f_x(x_0) = \dfrac{\lambda^k x_0^{k-1} e^{-\lambda x_0}}{(k-1)!} \qquad k = 1, 2, 3, \ldots ; \quad x_0 \geq 0$

b $f_x(x_0) = \dfrac{1}{4\sqrt{2\pi}} e^{-(x_0-8)^2/32} \qquad -\infty \leq x_0 \leq \infty$

c $p_x(x_0) = \dfrac{\mu^{x_0} e^{-\mu}}{x_0!} \qquad x_0 = 0, 1, 2, \ldots$

d $p_x(x_0) = P(1-P)^{x_0-1} \qquad x_0 = 1, 2, 3, \ldots$

7.02 Our model for a process states that x is a random variable described by the PDF

$$f_x(x_0) = \begin{cases} 1 & \text{if } r < x_0 \leq r + 1 \\ 0 & \text{otherwise} \end{cases}$$

and we do not know the value of r. For the following questions, assume that the form of our model is correct.

a We may use the average value of 48 independent experimental values of random variable x to estimate the value of r from the relation

$$M_n \approx E(x) = \int_r^{r+1} x_0 f_x(x_0)\,dx_0 = r + \tfrac{1}{2}$$

What is the probability that our estimate of r obtained in this way will be within ± 0.01 of the true value? Within ± 0.05 of the true value?

b We may use the largest of our 48 experimental values as our estimate of the quantity $r + 1$, thus obtaining another estimate of the value of parameter r. What is the probability that our estimate of r obtained this way is within $(+0, -0.02)$ of the true value? Within $(+0, -0.10)$ of the true value?

7.03 **a** Use methods similar to those of Sec. 7-3 to derive a reasonably simple expression for the variance of the sample variance.

b Does the sequence of sample variances $(S_1^2, S_2^2, \ldots)$ for a Gaussian random variable obey the weak law of large numbers? Explain.

7.04 There are 240 students in a literature class ("Proust, Joyce, Kafka, and Mickey Spillane"). Our model states that x, the numerical grade for any individual student, is an independent Gaussian random variable with a standard deviation equal to $10\sqrt{2}$. Assuming that our model is correct, we wish to perform a significance test on the hypothesis that $E(x)$ is equal to 60.

Determine the highest and lowest class averages which will result in the acceptance of this hypothesis:

a At the 0.02 level of significance

b At the 0.50 level of significance

7.05 We have accepted a Bernoulli model for a certain physical process involving a series of discrete trials. We wish to perform a significance test on the hypothesis that P, the probability of success on any trial, is equal to 0.50. Determine the rejection region for tests at the 0.05 level of significance if we select as our statistic

a Random variable r, the number of trials up to and including the 900th success

b Random variable s, the number of successes achieved in a total of 1,800 trials

The expected number of coin flips for each of these significance tests is equal. Discuss the relative merits of these tests. Consider the two ratios $\sigma_r/E(r)$ and $\sigma_s/E(s)$. Is the statistic with the smaller standard-deviation to expected-value ratio necessarily the better statistic?

7.06 Random variable x is known to be described by the PDF

$$f_x(x_0) = \begin{cases} \dfrac{1}{A} & \text{if } 0 < x_0 \le A \\ 0 & \text{otherwise} \end{cases}$$

but we do not know the value of parameter A. Consider the following statistics, each of which is based on a set of five independent experimental values $(x_1, x_2, \ldots, x_5)$ of random variable x:

$$r = 0.2(x_1 + x_2 + \cdots + x_5)$$
$$s = \max(x_1, x_2, \ldots, x_5)$$
$$t = 0.5(x_1 + x_2)$$

We wish to test the hypothesis $A = 2.0$ at the 0.5 level of significance. (A significance test using statistic r, for example, is referred to as T_r.)

Without doing too much work, can you suggest possible values of the data $(x_1, x_2, \ldots, x_5)$ which would result in:

a Acceptance only on T_r (and rejection on T_s and T_t)? Acceptance only on T_s? Acceptance only on T_t?

b Rejection only on T_r (and acceptance on T_s and T_t)? Rejection only on T_s? Rejection only on T_t?

c Acceptance on all three tests? Rejection on all three tests?

If the hypothesis is accepted on all three tests, does that mean it has passed an equivalent single significance test at the $1 - (0.5)^3$ level of significance?

7.07 Al, the bookie, plans to place a bet on the number of the round in which Bo might knock out Ci in their coming (second) fight. Al assumes only the following details for his model of the fight:

1 Ci can survive exactly 50 solid hits. The 51st solid hit (if there is one) finishes Ci.

2 The times between solid hits by Bo are independent random variables with the PDF

$$f_t(t_0) = \lambda e^{-\lambda t_0} \qquad t_0 \ge 0$$

3 Each round is three minutes.

Al hypothesizes that $\lambda = \frac{1}{30}$ (hits per second). Given the result of the previous fight (Ci won), at what significance level can Al accept his hypothesis $H_0(\lambda = \frac{1}{30})$? In the first fight Bo failed to come out for round 7—Ci lasted at least six rounds. Discuss any additional assumptions you make.

7.08 We are sure that the individual grades in a class are normally distributed about a mean of 60.0 and have standard deviation σ equal to either 5.0 or 8.0. Consider a hypothesis test of the null hypothesis $H_0(\sigma = 5.0)$ with a statistic which is the experimental value of a single grade.

a Determine the acceptance region for H_0 if we wish to set the conditional probability of false rejection (the level of significance) at 0.10.

b For the above level of significance and critical region, determine the conditional probability of acceptance of H_0, given $\sigma = 8.0$.

c How does increasing the number of experimental values averaged in the statistic contribute to your confidence in the outcome of this hypothesis test?

d Suggest some appropriate statistics for a hypothesis test which is intended to discriminate between $H_0(\sigma = 5.0)$ and $H_1(\sigma = 8.0)$.

e If we use H_1 as a model for the grades, what probability does it allot to grades less than 0 or greater than 100?

7.09 A random variable x is known to be characterized by either a Gaussian PDF with $E(x) = 20$ and $\sigma_x = 4$ or by a Gaussian PDF with $E(x) = 25$ and $\sigma_x = 5$. Consider the null hypothesis $H_0[E(x) = 20, \sigma_x = 4]$. We wish to test H_0 at the 0.05 level of significance. Our statistic is to be the sum of three experimental values of random variable x.

a Determine the conditional probability of false acceptance of H_0.

b Determine the conditional probability of false rejection of H_0.

c Determine an upper bound on the probability that we shall arrive at an incorrect conclusion from this hypothesis test.

d If we agree that one may assign an a priori probability of 0.6 to the event that H_0 is true, determine the probabilities that this hypothesis test will result in:

i False acceptance of H_0

ii False rejection of H_0

iii An incorrect conclusion

7.10 A random variable x is known to be the sum of k independent identically distributed exponential random variables, each with an expected value equal to $(k\lambda)^{-1}$. We have only two hypotheses for the value of parameter k; these are $H_0(k = 64)$ and $H_1(k = 400)$. Before we obtain any experimental data, our a priori guess is that these two hypotheses are equally likely.

The statistic for our hypothesis test is to be the sum of four independent experimental values of x. We estimate that false acceptance of H_0 will cost us \$100, false rejection of H_0 will cost us \$200, and any correct outcome of the test is worth \$500 to us.

Determine approximately the rejection region for H_0 which maximizes the expected value of the outcome of this hypothesis test.

7.11 A Bernoulli process satisfies either $H_0(P = 0.5)$ or $H_1(P = 0.6)$. Using the number of successes observed in n trials as our statistic, we wish to perform a hypothesis test in which α, the conditional probability of false rejection of H_0, is equal to 0.05. What is the smallest value of n for which this is the case if β, the conditional probability of false acceptance of H_0, must also be no greater than 0.05?

7.12 A hypothesis test based on the statistic

$$M_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

is to be used to choose between two hypotheses

$$H_0[E(x) = 0, \sigma_x = 2] \qquad H_1[E(x) = 1, \sigma_x = 4]$$

for the PDF of random variable x which is known to be Gaussian.

a Make a sketch of the possible points (α,β) in an α,β plane for the cases $n = 1$ and $n = 4$. (α and β are, respectively, the conditional probabilities of false rejection and false acceptance.)

b Sketch the ratio of the two conditional PDF's for random variable M_n (given H_0, given H_1) as a function of M_n for the cases $n = 1$ and $n = 4$. Discuss the properties of a desirable statistic that might be exhibited on such a plot.

7.13 Expanding on the statement of Prob. 7.06, consider the statistic

$$s_n = \max (x_1, x_2, \ldots, x_n)$$

as an estimator of parameter A.

a Is this estimator biased? Is it asymptotically biased?

b Is this estimator consistent?

c Carefully determine the maximum-likelihood estimator for A, based only on the experimental value of the statistic s_n.

7.14 Suppose that we flip a coin until we observe the lth head. Let n be the number of trials up to and including the lth head. Determine the maximum-likelihood estimator for P, the probability of heads. Another experiment would involve flipping the coin n (a predetermined number) times and letting the random variable be l, the number of heads in the n trials. Determine the maximum-likelihood estimator for P for the latter experiment. Discuss your results.

7.15 We wish to estimate λ for a Poisson process. If we let $(x_1, x_2, \ldots, x_n)$ be independent experimental values of n first-order interarrival times, we find (Sec. 7-9) that λ_n^*, the maximum-likelihood estimator for λ, is given by

$$\lambda_n^* = n\left(\sum_{i=1}^{n} x_i\right)^{-1}$$

a Show that $E(\lambda_n^*) = n\lambda/(n-1)$.

b Determine the exact value of the variance of random variable λ_n^* as a function of n and λ.

c Is λ_n^* a biased estimator for λ? Is it asymptotically biased?

d Is λ_n^* a consistent estimator for λ?

e Based on what we know about λ_n^*, can you suggest a desirable unbiased consistent estimator for λ?

Another type of maximum-likelihood estimation for the parameter λ of a Poisson process appears in the following problem.

7.16 Assume that it is known that occurrences of a particular event constitute a Poisson process in time. We wish to investigate the parameter λ, the average number of arrivals per minute.

a In a predetermined period of T minutes, exactly n arrivals are observed. Derive the maximum-likelihood estimator λ^* for λ based on this data.

b In 10,000 minutes 40,400 arrivals are observed. At what significance level would the hypothesis $\lambda = 4$ be accepted?

c Prove that the maximum-likelihood estimator derived in (a) is an unbiased estimator for λ.

d Determine the variance of λ^*.

e Is λ^* a consistent estimator for λ?

7.17 The volumes of gasoline sold in a month at each of nine gasoline stations may be considered independent random variables with the PDF

$$f_v(v_0) = \frac{1}{\sqrt{2\pi}\,\sigma_v} e^{-[v_0 - E(v)]^2/2\sigma_v^2} \qquad -\infty \le v_0 \le +\infty$$

a Assuming that $\sigma_v = 1$, find E^*, the maximum-likelihood estimator for $E(v)$ when we are given only V, the total gasoline sales for all nine stations, for a particular month.

b Without making any assumptions about σ_v, determine σ_v^* and E^*, the maximum-likelihood estimators for σ_v and $E(v)$.

c Is the value of E^* in (b) an unbiased estimator for $E(v)$?

7.18 Consider the problem of estimating the parameter P (the probability of heads) for a particular coin. To begin, we agree to assume the following a priori probability mass function for P:

$$p_P(P_0) = \begin{cases} 0.1 & P_0 = 0.4 \\ 0.8 & P_0 = 0.5 \\ 0.1 & P_0 = 0.6 \end{cases}$$

We are now told that the coin was flipped n times. The first flip resulted in heads, and the remaining $n-1$ flips resulted in tails.

Determine the a posteriori PMF for P as a function of n for $n \ge 2$. Prepare neat sketches of this function for $n = 2$ and for $n = 5$.

7.19 Given a coin from a particular source, we decide that parameter P (the probability of heads) for a toss of this coin is (to us) a random variable with probability density function

$$f_P(P_0) = \begin{cases} K(1-P_0)^8 P_0^8 & \text{if } 0 \le P_0 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

We proceed to flip the coin 10 times and note an experimental outcome of six heads and four tails. Determine, within a normalizing constant, the resulting a posteriori PDF for random variable P.

7.20 Consider a Bayesian estimation of λ, the unknown average arrival rate for a Poisson process. Our state of knowledge about λ leads us to describe it as a random variable with the PDF

$$f_\lambda(\lambda_0) = \frac{a^k \lambda_0^{k-1} e^{-a\lambda_0}}{(k-1)!} \qquad \lambda_0 \ge 0$$

where k is a positive integer.

a If we observe the process for a predetermined interval of T units of time and observe exactly N arrivals, determine the a posteriori PDF for random variable λ. Speculate on the general behavior of this PDF for very large values of T.

b Determine the expected value of the a priori and a posteriori PDF's for λ. Comment on your results.

c Before the experiment is performed, we are required to give an estimate λ_G for the true value of λ. We shall be paid $100 - 500(\lambda_G - \lambda)^2$ dollars as a result of our guess. Determine the value of λ_G which maximizes the expected value of the guess.