CHAPTER 12. MODELS AND ESTIMATION

As we have seen, probability is the study of certain models (for experimental situations) and of how to make predictions based on those models. Statistics is the study of how to choose useful models and of how to make predictions and decisions in an experimental situation when we are not sure what model for the situation is correct. Recall that a model for an experimental situation is a probability space, that is to say, a sample space (set of possible outcomes) together with a probability function (assignment of probability values) on that sample space. An observation is a result obtained in one, or several, actual trials of the experiment.

The crucial concept connecting models and observations is the descriptive level of significance (DLS). The DLS, as we have seen, depends on a model, an observation, and a chosen metric. (If μ is a model, ω is an observation, and s is a metric (for the model μ), then we shall sometimes use the notation $\underline{\text{DLS}}_s(\omega;\mu)$ for the value of the DLS of the observation ω given the model μ and the metric s.) Once the metric is chosen, the DLS gives us a measure of how "reasonable" the given observation is, if we assume that the given model is correct. We shall use the DLS in two major systematic ways: first, to decide, for a given fixed model, what observations are reasonable under that model, and second, to decide, for a given fixed observation, what models make that observation a reasonable one. We consider these two ways further in Chapters 14 and 15.

In the present chapter, we present some further concepts having to do with observations, models, and the DLS. Many of the statistical techniques that we shall later study depend upon these concepts. The presentation in this chapter will appear somewhat abstract, however, and, in order to gain full understanding, the reader may find it helpful to return again to this chapter after studying some of the specific techniques described later.

Universe of models. In a statistical problem, we consider various possible models. Hence, we usually ask the following question at the outset: what is the range of possible models to be considered? Sometimes this range is very broad. At other times it may be clear from the experimental situation that the correct model must come from a limited class of models (such as binomial distributions or Poisson distributions). When, at the beginning of a problem, we decide on the appropriate set of possible models to be considered, we call this set the universe of models for the problem.

Example A. A polyhedral die is a flat-sided solid that can come to rest, when thrown, on any one of its sides. (Such a die could have the shape of a pyramid or prism, for instance.) Assume that an irregular 11 sided polyhedral die has its sides marked 0, 1, ..., 10. This die is thrown once and we observe the number upon which the die comes to rest. What is an appropriate universe of models for this experiment? Unless we have further physical knowledge, including knowledge of the distribution of mass in the interior of the die,

we must take the set of all probability spaces with 11 points as our universe of models.

Example B. A thumbtack is tossed 10 times, and we observe the number of times that the tack comes to rest on its side. What is an appropriate universe of models? As in the previous example, each model must be a probability space with 11 points. However, we also know from the nature of the experiment (provided the tack is tossed from a sufficient height and in a sufficiently irregular way) that the experiment is binomial and that the correct model must be a binomial distribution. Hence we can say that the universe of models is the set of all binomial distributions with 10 trials. To choose a particular model in the universe, we need only fix the value of p, the probability of success in a single trial.

Example C. We have two groups of 10 rats each. All 20 rats are of similar age, weight, and genetic type. The first group (the "control" group) is fed a standard diet for a certain period. The second group (the "experimental" group) is fed a special supplemented diet for the same period. At the end of the period, all 20 rats are exposed to a certain bacterium. We then observe the number of rats in each group that develop signs of infection. What is an appropriate universe of models for this experiment? Assume that we know, from past experience, that observing the control group can be viewed as a binomial experiment of 10 trials with the number of rats showing infection as the number of successes. We let p_1 be the probability of success for a

single trial. Similarly, assume that we know, from past experience, that observing the experimental group can be viewed as a second binomial experiment of 10 trials (independent of the first binomial experiment). Let p_2 be the probability of success for a single trial in this second binomial experiment. The entire experiment can then be called a double binomial experiment of 10 trials and 10 trials with single success probabilities p_1 and p_2. The universe of models will be the set of all pairs of binomial distributions of 10 trials each. We choose a particular model in the universe by fixing values for p_1 and p_2. Then the probability of getting x_1 successes in the control group and x_2 successes in the experimental group is given by the formula

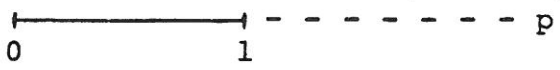
$$D(x_1, x_2) = \binom{10}{x_1} p_1^{x_1} (1-p_1)^{10-x_1} \binom{10}{x_2} p_2^{x_2} (1-p_2)^{10-x_2}$$

Example D. We observe the number of fire alarms in a given town each week. What is the appropriate universe of models? Here there are three alternatives. First, (D_1) if we believe that the conditions of randomness and independence for a Poisson experiment hold, we can take the set of all Poisson distributions as our universe of models (for what happens each week). Or, second, (D_2) if we believe that results are independent from week to week, but that within a given week fires may not occur entirely independently of one another (for example, there might be an arsonist who, when he occasionally sets fires, always makes sure to set several fires during the same day), then we may take the set of all

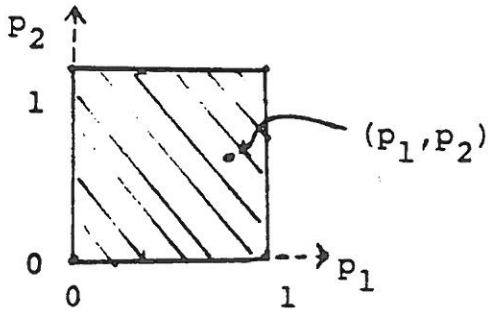
probability spaces with the infinite sample space {0,1,2,...} as our universe of models (for what happens each week). Or, finally, (D_3) we may believe that results from week to week are not independent (for example, the likelihood of fires might vary with the season of the year or with a level of public vigilance that is, in turn, influenced by the recent frequency of fires.) In this case, we may decide to look at models that are stochastic processes, with sample points that represent the experience of an entire year. Our universe of models in this case would be an appropriate class of stochastic processes.

Parametric and non-parametric universes. In Example B above, we can select a model from the universe by giving a value of p. In Example C , we can select a model by giving values of p_1 and p_2. In Example D_1 , we can select a model by giving a value for m, the parameter occurring in the Poisson formula. In each of these cases we say that the universe is parametric because we can select a model from the universe by giving a value for a single parameter (p in B , m in D_1) or by simultaneously giving values for several parameters (p_1 and p_2 in C). In D_2 and D_3, we say that the universe is non-parametric because there is no natural parameter (or simple set of parameters) that can be used in this way. In A, we could use the 11 values of the probability function itself as parameters, and say that the universe is parametric, but this is somewhat clumsy, and we might prefer to speak of the universe as non-parametric.

It is often possible, in the case of parametric universes, to give a simple diagram for the entire universe of models by giving a geometric picture for the corresponding possible parameter values. For example, in B , we can picture the universe of models itself as the interval [0,1] of possible values for p.



Each model in the universe corresponds to a single point in this interval. In C , we can picture the universe of models as the square region of points whose coordinates are possible values for p_1 and p_2.



Each model in the universe corresponds to a single point in this square. In D_1 we can picture the universe of models by giving the half-line (positive axis) of possible values for the parameter m.

Each model in the universe corresponds to a single point on this positive axis.

A variety of special techniques have been developed for statistical problems where the universe is parametric. These are called parametric methods. Much basic work in statistics (work now viewed as classical) has been done in the area of parametric methods. Most, but not all, of our later work in this text is in this area. When we work with parametric methods, geometric universe-diagrams like those described above are often useful.

A number of special techniques have also been developed for problems where the universe is non-parametric. These are called non-parametric (or sometimes distribution-free) methods. In Chapter 16, we shall look at several non-parametric methods.

Universes and metrics. If we have an experiment and model, and wish to calculate the DLS of an observation, we need to choose a metric. How we choose a metric can be affected by the choice that we make of a universe of possible models for the experiment. We see this in the following example.

A multiple choice test has 100 questions. Each question has five possible answers. A student gets the correct answer for 28 of the questions. If we assume that the student has chosen each answer by making a random choice among the five possible answers, what is the DLS of the observed result (of 28 correct answers)?

We can approach this problem in two different ways.

(a) In the first way, we take the universe of possible models to be the set of all binomial distributions with $0 \leq p \leq 1$ and $n = 100$. We take X, the number of successes, to be the number of correct answers. We seek the DLS of the observation $X = 28$ under the particular model $p = 0.2$. Using the usual metric for a binomial experiment and model, we get $|X - np| = |28-20| = 8$. Then DLS $= P(|X-20| \geq 8) = P(X \geq 28$ or $X \leq 12) \approx 1 - 2A(7.5) = 0.06$ by normal approximation.

(b) In the second way, we take the universe to be the set of all binomial distributions with $0.2 \leq p \leq 1$ and $n = 100$. X is again the number of correct answers, and we again seek the DLS of the observation $X = 28$ under the chosen model $p = 0.2$. In this case we have, by assumption, ruled out as impossible any binomial model with $p < 0.2$. It follows that any observation ≤ 20 must be viewed as strongly agreeing with (or confirming) the model $p = 0.2$, because it agrees with that model more strongly than with any other possible model. We therefore adopt the metric

$$s(X) = \frac{1}{2}(|X-20|+(X-20)) = \begin{cases} |X-20| & \text{for } X \geq 20, \\ 0 & \text{for } X \leq 20. \end{cases}$$

If we now calculate the DLS, we get DLS $= P(s(X) \geq s(28)) = P(s(X) \geq 8) = P(X \geq 28) \approx \frac{1}{2} - A(7.5) = 0.03$.

We thus see that the DLS value with the second and more restricted universe (b) is only half as large as the DLS obtained in (a).

What would lead us to choose universe (a) as opposed to universe (b) (or vice-versa)? We would choose (a) if we wanted to allow for the possibility that a student making a deliberate effort (and not doing the test in a random way) might, as a result, do worse rather than better. This could occur, for example, if the questions on the test were composed so that they invited certain wrong answers from students with limited knowledge. We would also choose (a) if we thought that there were students who might purposely seek to get a low score. On the other hand, we would choose (b) if we believed that every student who worked in a non-random way would, on the average, do at least as well as a student who made purely random choices.

The above example is somewhat artificial. In a real situation, we might believe that universe (b) was likely to apply, but that there was some slight possibility of one of the additional models (with $p < 0.2$) from (a). We might then seek to find a metric with effects (in terms of DLS values) part way between the effects of (a) and the effects of (b). We will discuss these matters further in Chapters 14 and 20. Our example above illustrates how our choice of universe can affect our choice of metric and hence can affect the DLS values we get. Two further examples of this kind were given in Exercises 10-15 and 10-16 above.

Estimation. Sometimes in a statistical problem, after we have settled on a universe of models, we make an observation and seek to find the single model in that universe which "best explains" the observation that we got. The problem of finding such a "best" model is called the estimation problem in statistics, and the model that we find is called our estimate of the correct model. In the case of a parametric universe, we need only find values of the parameter (or parameters) in order to find a model. When we do this in a particular case, we say that we have estimated the parameter (or parameters).

Example. A thumbtack is tossed 10 times and we observe that it lands on its side 7 times. Here (Example B above) it is natural to take our universe to be the one-parameter family of all binomial distributions of 10 trials. Furthermore, there is an obvious solution to the estimation problem: we take $p = 7/10$. The binomial distribution with $n = 10$ and $p = 0.7$ is our estimate.

In more complex examples, there may be no such single obvious choice of an estimate. How can we go about finding an estimate in such examples? There are several distinct and systematic ways of approaching the estimation problem. One method (the DLS method) can be applied if we have a metric for every model in the universe. We then take, as our estimate, the model which gives the highest DLS value to the given observation. A second method (the maximum-likelihood method) is to find that model which gives the highest probability to the given observation. This model is called the maximum-likelihood estimate (or maximum-likelihood model) for the given observation. These two methods often give the same estimate. There may, of course, be statistical problems where we obtain under the maximum-likelihood method (or the DLS method), instead of a single best model, a set of models all of which maximize the likelihood (or the DLS). We shall usually assume that we are working with problems where a single best model exists for each observation, except that in the case of universes of two or more parameters, we shall sometimes use an observation to determine a value for only one of those parameters. In that case, our estimate is the set of models with that parameter value.

In parametric universes, maximum-likelihood models can often be found by the maximum-minimum techniques of calculus.

Example. In a binomial experiment with 10 trials, 7 successes are observed. What is the maximum-likelihood model for this observation? The probability of an observation x is $\binom{10}{x}p^x(1-p)^{10-x}$. We abbreviate this probability as $L(x;p)$ the likelihood of x under model p. We seek that value of p which makes $L(7,p)$ a maximum. Hence we find

$$\frac{d}{dp}L(7;p) = \binom{10}{7}\left[7p^6(1-p)^3 - p^7\,3(1-p)^2\right]$$

and set this equal to 0. Solving for p, we get

$$7(1-p) - 3p = 0.$$

$$p = 0.7.$$

Example. We show that our previous method of fitting a Poisson distribution to an observation amounts to choosing the maximum-likelihood model in the universe of all Poisson distributions. Let $x_1, x_2, \ldots, x_n$ be specific independent observations (for a Poisson experiment) in the order obtained. The probability of getting these particular values in this order, given a Poisson model with parameter m, must be

$$L(x_1,\ldots,x_n;m) = p(x_1;m)p(x_2;m)\ldots p(x_n;m)$$

$$= \frac{e^{-nm}\,m^{x_1 + \ldots + x_n}}{x_1!\ldots x_n!}$$

We find the value of m which maximizes L by differentiating L and setting the derivative = 0. We get:

$$\frac{d}{dm} L(x_1,\ldots,x_n;m) =$$

$$\frac{-ne^{-nm} m^{x_1+\ldots+x_n} + e^{-nm}(x_1+\ldots+x_n)m^{x_1+\ldots+x_n-1}}{x_1!\ \ldots\ x_n!} = 0$$

This gives $\quad -nm + (x_1+\ldots+x_n) = 0$

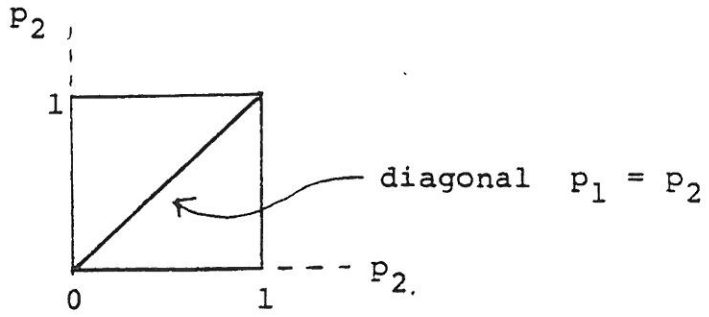
or $\quad m = \dfrac{x_1 + \ldots + x_n}{n}$, as stated.

Maximum-likelihood methods have a special role, as we shall later see, in statistical theory.

Note. A method of estimation is sometimes called an estimator. More abstractly, an estimator can be viewed as a mapping from the set of possible observations to a chosen universe of models, where the estimate given by each observation is obtained by solving a maximization problem of some given kind.

Composite models. Let us return to Example C above. An experimenter, who conducts this double binomial experiment, will be interested in whether or not, and how, the difference in diets affects the susceptibility of the rats to infection. As experimenters, we have as our first question: does the difference in diets appear to have any effect at all? We can restate this question as: is $p_1 = p_2$? To put it in geometric terms, does the point representing the

correct model fall on the diagonal marked in the figure?



As experimenters, we may wish to think of models with points on this diagonal as models of a special kind. We might even sometimes use the word "model" to indicate the set of all models of this kind. For example, as we make observations, we might ask how well our observations agree with the "model" that the difference in diet has no effect. In this case, we would not care about the particular numerical values of p_1 and p_2, but only about whether $p_1 = p_2$.

When there is a subset (of the universe of models) in which we have such a special interest (like the diagonal set $p_1 = p_2$ above), we shall refer to this subset as a composite model. We shall see that in certain circumstances it is possible to think of, and work with, a composite model almost as if it were a single model. We shall sometimes use a single symbol to stand for a composite model. In further discussion of Example C, we shall use M_o to refer to the composite model made up of all models for C with $p_1 = p_2$.

As a second example of a composite model, consider Example D above (fire alarms). If we believe that the results are independent from week to week, but we are not sure ahead of time that the experiment is Poisson, we will take all possible probability spaces with sample space $\{0,1,2,...\}$ as our universe of models (D_2). We may, of course, then wish to ask whether our observations lead us to believe that the experiment is Poisson. (We might even say "lead us to believe that the experiment fits the model of being Poisson". In Chapter 11 we saw how to use chi-square methods to answer this question). In such a case we are treating the set of all Poisson distributions as a composite model within our larger universe of all possible probability spaces on $\{0, 1, 2, ...\}$.

In a parametric universe of two or more parameters, composite models also occur when we specify the values of some but not all of the parameters. In Example C the set of all models with $p_2 = 1/2$ is such a composite model.

If we have a given composite model M and an observation ω, we can ask the maximum-likelihood question: what model in the composite model M gives the highest probability to the observation ω? Such a model is called the maximum-likelihood model in M for the given observation. (Thus we are temporarily taking the composite model to be the entire universe for the purposes of finding a maximum-likelihood estimate.) In Example C, what is the maximum-likelihood model in M_o for the observation $X_1 = 0$, $X_2 = 3$? Taking M_o itself as a parametric universe, with $p = p_1 = p_2$ as parameter, we have

$$L(x_1,x_2;p) = D(x_1,x_2) = \binom{10}{x_1} p^{x_1} (1-p)^{10-x_1} \binom{10}{x_2} p^{x_2} (1-p)^{10-x_2}.$$

Setting $\frac{d}{dp} L(0,3;p) = 0$, we obtain $p = 0.15$. More generally, the maximum-likelihood estimate for observation (x_1,x_2) with a control group of n_1 rats and an experimental group of n_2 rats will be $p = \frac{x_1+x_2}{n_1+n_2}$. See exercise 12-2.

Remark. In order to get a maximum-likelihood model in a parametric universe, we need only find certain parameter values. These values are usually called the maximum-likelihood estimates of the parameters (from the given observation). Thus, $X_1 = 0$, $X_2 = 3$ gives 0.15 as the maximum-likelihood estimate for p in the composite model M_o.

DLS for composite models. It is helpful, as we shall see, to be able to use the idea of DLS with composite models. We must proceed carefully, however. Given a universe of models (say the universe of double binomial models in Example C), given a composite model M (say the composite model M_o in C), given an observation (say $X_1 = 0$ and $X_2 = 3$), and given a metric, the DLS of the given observation may be large for some of the models in M and small for other models in M. (For example, under the natural metric $(X_1-10p_1)^2 + (X_2-10p_2)^2$, the observation $X_1 = 0$, $X_2 = 3$ would have a large DLS for the model $p_1 = p_2 = 0.1$ in M_o but a very small DLS for the model $p_1 = p_2 = 0.9$ in M_o.) In such a case, there does not seem to be a single value of the DLS that we can associate with the composite model M.

The best solution to this difficulty is to find, if we can, a reasonable new metric under which, for any given observation, each model in M gives nearly the same DLS value to that observation (that is to say, a metric such that

all models in M will give nearly the same DLS values to the same observations.) Surprisingly, such remarkable metrics can often be found and are highly useful in applications. If a metric has this special property for a composite model, we say that is well-defined for that composite model. (In Chapter 11, in the example of the chocolate bars, we found such a well-defined metric (we called it s) for the composite model of all Poisson distributions.) Thus a well-defined metric serves to measure confirmation of a composite model rather than, as in the cases of the metrics given in Chapter 10, to measure confirmation of a single given model.

Well-defined metrics. The definition of well-defined metric can be given somewhat more formally as follows. Let U be a given universe of models and M be a given composite model. We use the symbol μ to refer to individual models in U. Let ω be the particular observation obtained. Let $s(\Omega,\mu)$ give a family of metrics on M. That is to say: for each μ in M, $s(\Omega,\mu)$ defines a metric for the model μ. Then, as in Chapter 10, the DLS of an observation ω for each μ in M is defined as $\underline{DLS}^s(\omega;\mu) = P_\mu(s(\Omega,\mu) \geq s(\omega))$. We say that s is well-defined on M if, for each observation ω, it is the case that for every choice of μ_1 and μ_2 in M,

$$\underline{DLS}^s(\omega;\mu_1) \approx DLS^s(\omega;\mu_2) .$$

(Of course, for a given μ in M, if $\omega_1 \neq \omega_2$, we may have $\underline{DLS}^s(\omega_1,\mu) \neq \underline{DLS}^s(\omega_2,\mu)$.)

for a well-defined

Usually, in the cases we consider, the formula $s(\Omega,\mu)$ will not explicitly mention μ, and hence will have metric values which depend only on Ω. In this case, we shall often speak of $s(\Omega,\mu)$ as a metric rather than as a family of metrics and write it as $s(\Omega)$ rather than as $s(\Omega,\mu)$.

Example. Consider the composite model M from Example C above. Let $X_1 = 0$ and $X_2 = 3$ be our given observation. (No rats in the control group show signs of infection, but 3 rats in the experimental group do so.) Let us try the function $s(\Omega,\mu) = f_1(X_1,X_2) = |X_1-X_2|$ and see if it gives us a metric that is well-defined on M_o. In the following table we give the DLS, under f_1, of the observation (0,3) for each of the nine models p = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

p	$\underline{\text{DLS}}_{f_1}$ for $X_1 = 0$, $X_2 = 3$
0.1	0.06
0.2	0.16
0.3	0.21
0.4	0.23
0.5	0.24
0.6	0.23
0.7	0.21
0.8	0.16
0.9	0.06

We see that the DLS varies significantly from model to model within M_o and hence that the metric f_1 is not well-defined on M_o. (Note. In each case, the DLS is calculated by using the double binomial formula $D(x_1,x_2)$ of Example C and finding

$$\sum_{\substack{0\le i,j\le 10 \\ |i-j|\ge 3}} D(i,j) \qquad .)$$

We see from the table that f_1 also fails to be a reasonable metric (for M_o) for the following intuitive reason. The observation $X_1 = 0$, $X_2 = 3$ suggests that the model $p = 0.1$ is more appropriate than, say, the model $p = 0.5$, since the observation seems intuitively closer to the TER for 0.1 than for 0.5; yet the metric $|X_1 - X_2|$ gives a smaller DLS for $p = 0.1$ than for $p = 0.5$.

Can we find a metric that is well-defined for the model M_o ? We now look at the metric

$$f_2(X_1, X_2) = \frac{(X_1 - X_2)^2}{(X_1 + X_2) - \frac{1}{20}(X_1 + X_2)^2}$$

The following table gives the DLS for $X_1 = 0$ and $X_2 = 3$ calculated from this metric.

p	DLS
0.1	0.05
0.2	0.09
0.3	0.08
0.4	0.06
0.5	0.05
0.4	0.06
0.3	0.08
0.2	0.09
0.1	0.05

We see that this metric is more nearly well-defined than the previous one (and, indeed, is nearly well-defined to two decimal places.) (Note. To form this table, calculate

$$\sum_{\substack{0\le i,j\le 10 \\ f_2(i,j)\ge f_2(0,3)}} D(i,j) \quad .)$$

How was the formula for the metric f_2 obtained? It is, in fact, the metric that we get if we (i) view the observation as an observation of four categories with $\Omega = (X_1, 10-X_1, X_2, 10-X_2)$, (ii) use the observation to indicate a model in M_o by taking the maximum likelihood estimate $p = \frac{X_1+X_2}{20}$, and (iii) use the indicated model to form a CS metric for a four-category observation (with expected result given by the indicated model). Thus, from (iii), we get

$$\frac{(X_1-10p)^2}{10p} + \frac{((10-X_1)-10(1-p))^2}{10(1-p)} + \frac{(X_2-10p)^2}{10p} + \frac{((10-X_2)-10(1-p))^2}{10(1-p)}$$

and this simplifies directly to the formula given above for f_2.

In addition to being well-defined, a metric must, of course, embody a notion of distance that is appropriate for our statistical purposes. In the metric f_2 above, the numerator $(X_1-X_2)^2$ helps to provide this. Appropriateness of metrics is discussed, in a general setting, in Chapter 20.

Ease of calculation. In searching for a well-defined metric for a given composite model, it is important to get a metric for which the approximate common DLS value can be easily calculated. The metric f_2 above is highly satisfactory in this regard. In Chapter 13, we shall see that probabiliti-

for f_2 follow the chi-square curve with 1 degree of freedom. Hence the approximate common DLS value can be got by setting

$$\underline{\text{DLS}}(X_1, X_2) = C_1(f_2(X_1, X_2))$$

where C_1 gives area under the chi-square curve with 1 degree of freedom. (This approximation improves as n increases.) For our example with $n = 10$, $X_1 = 0$, and $X_2 = 3$, we have $f_2(0,3) = 3.53$ and we get $C_1(3.53) = 0.06$ as the approximate common value for the DLS. (This agrees well with our table of exact DLS values.) Thus we can say that under our well-defined metric f_2, we expect an observation as "far" as the observation $X_1 = 0$, $X_2 = 3$ only about 6 percent of the time. Thus the observation $X_1 = 0$, $X_2 = 3$ throws serious doubt on the composite model M_o, and suggests that the given difference in diet does make a difference in susceptibility to infection.

How does the reduction to 1 degree of freedom arise in the above use of chi-square approximation? In briefest summary, it occurs as follows. We begin with 4 degrees of freedom corresponding to our four categories. Reduction by one degree occurs because we first view our observation as if it were an observation from a multinomial experiment with four categories (and we used three degrees of freedom for such an

experiment). Reduction by one further degree occurs because we use the observation to indicate a model (we use a single algebraic fact: $p = \frac{X_1+X_2}{20}$). Reduction by a third degree occurs because our experiment is not in fact multinomial but is instead a double binomial experiment (it is like a multinomial experiment, but with the added constraint that we must always have, in any observation, the sum of the first two categories = 10, and hence the sum of the second two categories = 10). (In Chapter 13, we shall explain and justify, in a more general and systematic way, uses of chi-square approximation of this kind. Such uses were also briefly considered in the Appendix to Chapter 11.)

Historical note. Statistics emerged as a branch of mathematics in the first third of the twentieth century. Virtually all of the work done in this classical period was on parametric methods, and many of the deepest results had to do with finding, in convenient and useful forms, well-defined metrics for certain composite models (in various universes of models). We shall see examples of this in Chapters 17 and 18.

Note on terminology. The uses of the terms "model", "universe", and "composite model" in this text differ from other more customary choices of terminology. What we call models are sometimes (but not always) called "distributions" and are also sometimes called "states of nature." What we call universes are sometimes called "models". (In Example B above, one customary usage would be to say that we select a binomial "model" (with unspecified

parameter).) What we call composite models are sometimes called "models", sometimes called "hypotheses" or "composite hypotheses", and sometimes called "parameter values". (In Example C , if we define the parameter $q = p_1 - p_2$, then the composite model M_o corresponds to the parameter value $q = 0$.) The terminology used in this book has advantages for the purpose of a uniform and common approach to parametric, non-parametric, and decision-theoretic methods in statistics.

EXERCISES FOR CHAPTER 12

12-1. You are given a multinomial experiment with $c = 3$ and $n = 20$. You observe $(2,6,12)$. Let M be the composite model consisting of all models (p_1,p_2,p_3) with $p_2 = p_3$. Find the maximum likelihood estimate in M given by your observation.

12-2. In a double binomial experiment with n_1 and n_2 trials, you observe $X_1 = a$ and $X_2 = b$. Let M_o be the composite model with $p_1 = p_2$. Find the maximum likelihood estimate in M_o given by the observation.

12-3. In Exercise 11-16, average observed interval length (between successive sunny days) was used to indicate a Bernoulli-trial model. Show that this procedure gives us the maximum-likelihood estimate in the composite model of all Bernoulli-trial models.

Problems 4 through 8 concern certain parametric inverses of continuous probability distributions for a random variable X. In each case, given a parametric inverse M and given a finite set of independent observed values $x_1, x_2, \ldots, x_n$, you are asked to make a maximum-likelihood estimate in M. You may do so as follows. Let $f(x;\alpha)$ be the probability density function in M corresponding to parameter value α. Use as likelihood function:

$$L(x_1, x_2, \ldots, x_n;\alpha) = f(x_1;\alpha)\ f(x_2;\alpha)\ \ldots\ f(x_n;\alpha)\ .$$

For a given observation $x_1, \ldots, x_n$, find the value of α which maximizes L.

12-4. Consider the universe of all <u>exponential</u> density functions given by

$$f(x;m) = me^{-mx} \text{ (where } m > 0).$$

Find the maximum-likelihood estimate (for m) given by $(x_1, \ldots, x_n)$.

12-5. Consider the universe of all <u>translated Cauchy distributions</u> given by

$$f(x;m) = \frac{1}{\pi(1 + (x-m)^2)} .$$

Show that the maximum-likelihood estimate (for m) given by the observation (x_1, x_2) is $\frac{x_1+x_2}{2}$.

12-6. Consider the universe of all <u>normal distributions with fixed standard deviation</u> b_o. Thus $f(x;a) = N(x; a, b_o)$. Find the maximum-likelihood estimate (for a) given by $(x_1, \ldots, x_n)$.

12-7. Consider the universe of all <u>normal distributions with fixed mean</u> a_o. Then

$$f(x;b) = N(x; a_o, b) .$$

Find the maximum-likelihood estimate (for b) given by $(x_1, \ldots, x_n)$.

12-8. Consider the two-parameter universe of all <u>normal distributions</u>. Then

$$f(x; a,b) = N(x; a,b) .$$

Find the maximum-likelihood estimate (for the pair of values (a,b)) given by $(x_1, \ldots, x_n)$.