CHAPTER 10. DESCRIPTIVE LEVEL OF SIGNIFICANCE.

Chapters 1 through 9 have been an introduction, with examples, to the basic ideas and methods of the theory of probability. The central idea of this theory has been the concept of probability space, and the usefulness of the theory has rested, in part, on the empirical stability of relative frequencies. Probability spaces have served as mathematical pictures or models for given physical situations.

We now turn, in the remaining chapters of this book, to mathematical statistics. Mathematical statistics has to do with using observed data to choose the most suitable models we can for the purpose of describing a physical situation, of making a prediction about the situation, or of making a practical decision about the situation. In what follows, we shall, for brevity, use the word "model" to mean probability space. We shall sometimes use the symbol μ for a model. Given a model μ, if A is an event in the sample space of μ, the probability assigned to A by μ will usually be written $P(A)$ (as before); but sometimes, for clarity, it will be written $P_{\mu}(A)$.

Assume that we have an experiment, and that we have chosen a probability model for this experiment. Each time we do the experiment, we get an observed result. We call this observed result an observation. (For example, if our experiment is a binomial experiment of n trials, then we choose a model by taking a binomial distribution with (i) n trials and (ii) a particular assigned value for p, the probability of success in a single trial. When we do the experiment, we get x, the total number of successes observed. x is our observation.)

Given an experiment, an assumed model, and an observation, we can ask the basic question: how well does the observation confirm or agree with the chosen model for the experiment? Such questions lie at the heart of statistics and its applications. By asking and answering such questions, we can choose and modify models, and we can make decisions based on those models.

Example. Consider the experiment of rolling a single die 600 times and observing the number of times X that a six appears. We choose, as our model, a binomial distribution with $n = 600$ and $p = \frac{1}{6}$ If we observe $X = 102$, we would say that this observation agrees well with the chosen model (because the observed relative frequency, 102/600, is close to the probability value $p = 1/6$.) On the other hand, if we observe $X = 130$, we would say that this agrees less well, and if we observe $X = 200$, we would say that this agrees hardly at all.

How can a way to measure this confirmation or agreement be made precise? One natural way is simply to take the assumed model and calculate the probability of the observation that we have obtained. This direct approach is not satisfactory because, in some cases, the probability of every possible observation may be very small.

Example (continued). The probability of a given observation $X = x$, under our chosen model, is given by $b(x;600,1/6)$. In particular, if we get $X = 130$, this observation has probability $b(130;600,\frac{1}{6})$. Using normal approximation, we evaluate this to be approximately

$$\varphi(z)\cdot d, \text{ where } z = \frac{130 - 100}{\sqrt{600\ pq}} = 3.29 \text{ and } d = \frac{1}{\sqrt{600\ pq}} = \frac{1}{9.13}.$$

From tables for ϕ we get $\phi(3.29) = 0.0018$ and hence

$$b(130;600,\tfrac{1}{6}) \approx \frac{0.0018}{9.13} = 0.0002.$$

(Note. The correct value is 0.00026.)

If, on the other hand, we get $X = 100$, this observation has approximate probability $\phi(0)\cdot d = \frac{0.3989}{9.13} = 0.04$. The observation $X = 100$ is more probable than the observation $X = 130$, and is, in fact, the most probable observed result of all. Even so, its probability is only 0.04. Thus every observation in this experiment is, taken by itself, highly improbable. (Indeed, recall from our discussion of binomial distributions in Chapter 5 that the largest probability value becomes smaller as n increases, and is proportional to $\frac{1}{\sqrt{n}}$.) Thus the above approach to our basic question is not satisfactory because it leads to answers in the form of probability values that may be small for all observations.

We therefore take a different approach to our problem of measuring "how well an observation confirms a chosen model". Our approach is to calculate a quantity which we call the descriptive level of significance (abbreviated DLS) of the given observation under the chosen model. We calculate the DLS in two steps.

First step. For our chosen model, we find a mathematical formula s which associates, with every possible observation Ω, a number $s(\Omega)$ that represents *how far away the observation is from giving us the strongest possible confirmation of our chosen model*. In the above example, we might take this formula to be $s(X) = |X-100|$. Such a formula is called a *metric*. (It is sometimes also called a *statistic*.) In this book, we shall always choose the metric so that the *larger* the value of the metric (for a given observation and model), the *farther* we think of that observation as being from giving strongest possible confirmation of the model. For the above example and the above chosen metric, we would get $s(130) = 30$ and $s(90) = 10$. Thus the metric s places the observation $X = 130$ at a greater distance than the observation $X = 90$, and hence it ranks $X = 90$ as giving better confirmation of the model $p = 1/6$ than the observation $X = 130$ gives. How do we decide what formula to use as our *metric*? We return to this question below.

Second step. Given a metric s and a specific observation ω_o, we apply s to the observation to get $s(\omega_o)$, the corresponding *observed value* of the metric. We then ask: *if we were to do the experiment again, in identical circumstances, and make another observation, what is the probability (under our chosen model) of getting a new observation that is at least as far from agreement with the model (as measured by the metric) as the observation we actually first got?* Denoting our chosen model as μ, we can express this probability as

$$P_\mu(s(\Omega) \geq s(\omega_o))$$

where the event being described is the set of all observations Ω such that $s(\Omega) \geq s(\omega_o)$. In the expression "$P(s(\Omega) \geq s(\omega_o))$" we use the capital letter "Ω" to indicate the entity which varies. This is, in effect, the same notational convention as was adopted for numerical observations in Chapter 9 where we used capital letters to represent random variables. The probability $P_\mu(s(\Omega) \geq s(\omega_o))$ is called the descriptive level of significance (abbreviated DLS) of the given observation ω_o under the chosen μ. (It is also sometimes known as the nominal level or cumulative level of significance and sometimes as the p-value.) The value of the DLS depends, of course, upon the metric s that we have decided to use.

Example. Consider the observation of 130 sixes in 600 rolls of a die. Using the model $p = \frac{1}{6}$ and the metric $|X - 100|$, we get $s(\omega_o) = |130 - 100| = 30$. Our descriptive level of significance is then given by $P(s(\Omega) \geq s(\omega_o)) = P(|X - 100| \geq 30)$. Its value can be obtained from tables for the binomial distribution or by normal approximation. Normal approximation gives:

$$P(|X - 100| \geq 30) \approx 1 - \underline{\text{Normal Area}^{z}}_{-z} = 1 - 2A(z),$$

$$\text{where} \quad z = \frac{30 - \frac{1}{2}}{\sqrt{npq}} = \frac{29.5}{9.13} = 3.23.$$

From tables, $A(3.23) = 0.4994$; hence

$$P(|X - 100| \geq 30) \approx 1 - 0.9988 = 0.0012.$$

(Note. The correct value is 0.0013.)

Thus our descriptive level of significance is approximately 0.0012. This means that, on the basis of our assumed model, the probability is only 0.0012 of getting a result as far as this from $X = 100$. ($X = 100$ represents strongest possible confirmation.) That is to say, we can expect only about 0.12% of all observations to be as far as this from $X = 100$.

How about an observation of 90 <u>sixes</u>? This gives $s(\omega_o) = 10$, and by a similar calculation we have

$$P(s(\Omega) \geq s(\omega_o)) = P(|X - 100| \geq 10) \approx 1 - \underline{\text{Normal Area}}_{-z}^{z} = 1 - 2A(z),$$

$$\text{where} \qquad z = \frac{10 - 1/2}{9.13} = 1.04.$$

Then $A(1.04) = 0.3508$, and

$$P(|X - 100| \geq 10) \approx 1 - 0.7016 = 0.30.$$

Thus we can expect about 30% of all results to lie as far as this from $X = 100$. [<u>Footnote</u>. The normal curve is symmetric. Why can it be used, as above, to get approximate values for a binomial distribution which is centered at 100 and ranges to 0 on the left but to 600 on the right (and hence is highly non-symmetric in its extent)? (We could have asked this question about some of our previous examples of normal approximation to binomial distributions.) The answer is that in a certain region (about the value np), in which virtually all of the probability occurs, the binomial distribution <u>is</u> nearly symmetric, and the normal curve is an excellent approximation to it. Outside this region, both the normal and binomial values are so close to zero that the asymmetry of the binomial distribution in total extent does not matter. (The conditions $np/q \geq 9$ and $nq/p \geq 9$ for normal approximation guarantee the existence of the symmetric region and the closeness to zero outside it.)]

The *DLS* will be a fundamental concept in our future work. It is our basic measure of how well an observation agrees with a model. While the *metric* used can tell us, by itself, which of two observations disagrees more with a chosen model, the *DLS* is a probability value and measures the agreement in probability terms. The reader should note and remember that the *more* an observation disagrees with a model, then the *larger* the value of the metric will be; but the *smaller* the value of the *DLS* will be. Smaller values of the metric represent more agreement. Smaller values of the *DLS* represent more disagreement.

Finding a metric. The *first step*, that of finding a metric for a chosen model, can sometimes (but not always) be done in two parts, as follows.

Part one (of finding a metric). We identify an observation which we call the *theoretically expected result* (abbreviated *TER*). This corresponds to a hypothetical observation in which all observed relative frequencies agree exactly with probability values of the chosen model. For example, in the experiment of 600 rolls of a die, with the model $p = 1/6$, the *TER* will be 100. In general, in a binomial experiment, the *TER* will be np. There may be several different but reasonable ways of defining a *TER* in a given problem. These can then lead to different metrics. In the special case where we are observing the value of a random variable X (as in the case of a binomial experiment), we shall usually take the TER to be E_X. (Note that we allow non-integer values for a TER, even when the experiment itself has only integer outcomes. Thus, for n = 500 and p = 1/6, the TER is $500 \cdot \frac{1}{6} = 83.33$ to two decimal places.)

Part two (of finding a metric). We find a simple mathematical formula which will measure how far, in some sense, any given observation is from the TER. This formula is our metric. For example, in the case of 600 rolls of a die with p = 1/6, the TER is 100, and it is natural to take $s(X) = |X-100|$ as our metric.

In what follows, we will often be able to find a metric in this way: by first finding a TER; and by then getting a formula to measure distance of any observation from the TER. We shall also (in later chapters) see examples where a metric cannot be found in this way, but where other direct and simple arguments lead to a metric.

Choosing between metrics. Several different metrics may be possible, and seem natural, in a given situation. For example, in the case of rolling a die 600 times and observing the number of sixes occurring, we could have also used the metric

$$s_1(X) = \left|\frac{X}{600} - \frac{1}{6}\right|$$

which gives a direct measurement of the difference between observed and predicted relative frequencies. It is easy to see that the metric s_1 will give exactly the same DLS as the metric s for any observation x_o (because $\left|\frac{X}{600} - \frac{1}{6}\right| \geq \left|\frac{x_o}{600} - \frac{1}{6}\right|$ if and only if $|X-100| \geq |x_o-100|$). If two metrics for a chosen model always give the same DLS values (as in the example just given) the metrics are said to be equivalent metrics for the chosen model.

Unfortunately, it is possible to have two different available metrics (for a given situation and model) both of which seem natural but which are not equivalent. For example, in the case of 600 rolls of a die with the model $p = 1/6$, consider the metric

$$s_2(X) = \begin{cases} \frac{X-100}{500}, & \text{for } X > 100; \\ \\ \frac{100-X}{100}, & \text{for } X \leq 100. \end{cases}$$

(This gives distance as the difference in number of successes divided by the maximum possible difference in the same direction.) When $X = 130$, we get $s_2(130) = \frac{30}{500} = 0.06$, and when $X = 90$, we get $s_2(90) = \frac{10}{100} = 0.1$. Notice that s and s_2 are basically different (and non-equivalent) in that s makes 90 closer than 130 to 100, while s_2 makes 90 farther than 130 from 100. In particular, under s_2, the DLS of the observation $X_0 = 90$ will be $P(X \leq 90) + P(X \geq 150) \approx \text{Normal Area}_{z_1}^{z_2} = 1 - A(-z_1) - A(z_2)$, where $z_1 = \frac{-10 + 1/2}{9.13} \doteq -1.04$ and $z_2 = \frac{50 + 1/2}{9.13} = 5.42$. Thus the DLS $\approx 1 - 0.3508 - 0.5000 = 0.15$. We saw above that the DLS of the observation $X_0 = 90$ under the metric $s(X) = |X-100|$ is 0.30.

It is not always clear what the best metric to choose is. The choice will often depend on experience and on mathematical simplicity. From now on, in the case of a binomial experiment, we shall use the metric $|X-np|$ (the metric s above). In later chapters, we shall consider, in a more systematic and theoretical way, the question of how to choose a metric and the question of what metric is most useful in a given situation.

Multinomial experiments. If we view an experiment as having only two outcomes, which we call success and failure, and if we make n independent trials of the experiment and ask how many successes occur, then the resulting larger experiment is called a binomial experiment. We have studied binomial experiments in previous chapters.

Let us now look at an individual experiment which we view as having c distinct outcomes, where c is some number ≥ 2. If we make n independent trials of this experiment and ask how many times each of the outcomes occurs, we call the resulting larger experiment a multinomial experiment. c is called the number of classes or categories in the given multinomial experiment.

Example. When Boston plays Montreal at hockey, there are three possible outcomes for Boston: win, lose, or tie. If we assume that probabilities for win, lose,and tie remain fixed throughout the season, and if we assume that successive games between these opponents are independent trials, then the larger experiment of looking at 10 games between Boston and Montreal and asking how many wins, losses, and ties occur is a multinomial experiment with three categories. An observation for this experiment will list the number of wins, the number of losses, and the number of ties in the 10 games observed.

Multinomial formula. Just as the binomial formula serves to calculate probability values for results of a binomial experiment, under a chosen model which assigns a probability value p to success, there is a *multinomial formula* which gives probability values for the various possible results of a multinomial experiment under a chosen model which assigns probability values $p_1,\ldots,p_c$ to the different categories. This formula takes the following form (and can be derived in the same way that the binomial formula was derived). We give it for the case $c = 3$. Let n = the number of trials. We call the three possible outcomes $\underline{o}_1$, $\underline{o}_2$, and $\underline{o}_3$. Let p_1, p_2, and p_3 be the respective probabilities that outcomes $\underline{o}_1$, $\underline{o}_2$, $\underline{o}_3$ occur in a single trial. (Of course, we must have $p_1 + p_2 + p_3 = 1$.) Let X_1, X_2, X_3 be the observed number of respective occurrences of $\underline{o}_1$, $\underline{o}_2$, $\underline{o}_3$ when n trials are carried out. (Of course, we must have $X_1 + X_2 + X_3 = n$.) Then the probability $P(X_1 = x_1 \ \underline{\text{and}}\ X_2 = x_2 \ \underline{\text{and}}\ X_3 = x_3)$ is given by the formula

$$m(x_1,x_2,x_3;p_1,p_2,p_3) = \frac{n!}{x_1!x_2!x_3!}\, p_1^{x_1}p_2^{x_2}p_3^{x_3} = \binom{n}{x_1,x_2,x_3} p_1^{x_1}p_2^{x_2}p_3^{x_3},$$

where we use the notation from Chapter 3 for the *multinomial coefficient*. In general, where c is the number of categories, this multinomial formula takes the form:

$$m(x_1,x_2,\ldots,x_c;p_1,p_2,\ldots,p_c) =$$

$$\frac{n!}{x_1!x_2!\ldots x_c!}\, p_1^{x_1}\, p_2^{x_2} \ldots p_c^{x_c} \quad \text{(where } n = x_1 + \ldots + x_c)$$

$$= \binom{n}{x_1,x_2,\ldots,x_c} p_1^{x_1}\, p_2^{x_2} \ldots\, p_c^{x_c} \quad .$$

When $c = 2$, this reduces to the binomial formula. Indeed, a multinomial experiment with two categories is obviously equivalent to a binomial experiment, if we call one category success and the other failure. (Since the number of trials is given, knowing the number of successes is the same as knowing both the number of successes and the number of failures.) It is thus immediate that $b(x;n,p) = m(x,n-x;p,1-p)$.

The fact that the probabilities for the possible observations in a multinomial experiment total to 1 is asserted in the multinomial theorem stated in Chapter 3.

Example. When Boston plays Montreal, assume that the probabilities of win, lose and tie are 0.5, 0.3, and 0.2. What is the probability that in 10 games, Boston wins 3, loses 4, and ties 3? We get

$$m(3,4,3;0.5,0.3,0.2) = \frac{10!}{3!4!3!}(0.5)^3(0.3)^4(0.2)^3 = 0.034.$$

What is the probability that in 10 games, Boston wins 5, loses 3, and ties 2? We get

$$\frac{10!}{5!3!2!}(0.5)^5(0.3)^3(0.2)^2 = 0.085.$$

Descriptive level of significance in multinomial experiments. To find a descriptive level of significance in a multinomial experiment, we apply the two general steps described above. We illustrate with the multinomial experiment for Boston and Montreal in the example given, taking the model $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. Assume that we have, in fact, observed 5 wins, 1 loss, and 4 ties. What is a descriptive level of significance for this observation?

First step (finding a metric).

Part one. We find the theoretically expected result. In the general multinomial experiment, it will be $X_1 = E_1$, $X_2 = E_2, \ldots$ where $E_1 = E_{X_1} = np_1$, $E_2 = E_{X_2} = np_2, \ldots$. In our example, the theoretically expected result is 5 wins, 3 losses, 2 ties. (As with a binomial experiment, we allow non-integer values in the TER for a multinomial experiment even though the experiment itself has only integer outcomes. If we took the model $p_1 = 0.5$, $p_2 = 0.35$, and $p_3 = 0.15$ for the Boston-Montreal 10 game experiment, for example, the TER would be 5 wins, 3.5 losses, and 1.5 ties.)

Part two. We choose a metric. The formula most commonly used in probability theory for a metric for multinomial experiments is the chi-square metric, which we shall also call the CS metric. This metric is usually abbreviated as χ^2. The formula is

$$\chi^2(X_1,\ldots,X_c) = \frac{(X_1 - E_1)^2}{E_1} + \frac{(X_2 - E_2)^2}{E_2} + \ldots + \frac{(X_c - E_c)^2}{E_c} .$$

In our example we have,

$$\chi^2(5,1,4) = \frac{(5 - 5)^2}{5} + \frac{(1 - 3)^2}{3} + \frac{(4 - 2)^2}{2} = 0 + \frac{4}{3} + \frac{4}{2} = 3.33.$$

If the observed result had been 3 wins, 2 losses, and 5 ties, we would have got

$$\chi^2(3,2,5) = \frac{(3-5)^2}{5} + \frac{(2-3)^2}{3} + \frac{(5-2)^2}{2} = \frac{4}{5} + \frac{1}{3} + \frac{9}{2} = 5.63.$$

The *CS* metric thus places the second result (3,2,5) at a greater distance from the expected result (5,3,2) than it places the first result (5,1,4) from the expected result. Note that it is not obvious ahead of time how to decide which of two observations is farther from the expected result. The *CS* metric decides this for us. A different choice of metric might have given a different answer. We use the *CS* metric because it is simple, mathematically convenient, and, as we shall see, practically and intuitively useful. In particular, we shall find, in Chapter 10, that it is easy, in most cases, to calculate values for the *DLS* under this metric by a simple approximation method.

Second step (calculating the DLS). In our example, if the observation is 5 wins, 1 loss, and 4 ties, we want $P(\chi^2(\Omega) \geq \chi^2(\omega_o))$ where $\chi^2(\omega_o) = 3.33$ as shown above. We shall usually abbreviate the expression "$P(\chi^2(\Omega) \geq \chi^2(\omega_o))$" as "$P(\chi^2 \geq \chi_0^2)$." Here χ_o^2 stands for the value of the *CS* metric for the observation actually obtained. Thus, in the example just given, $\chi_0^2 = \chi^2(5,1,4) = 3.33$, and "$\chi^2 \geq \chi_0^2$" stands for the event which consists of all triples (x_1,x_2,x_3) such that $\chi^2(x_1,x_2,x_3) \geq 3.33$. The direct calculation of $P(\chi^2 \geq \chi_0^2)$ can be carried out on a computer (or, in simple cases, on a programmable calculator). We do it in the following way. We list all possible observations (there are 66 of them). For each possible observation we calculate the value of χ^2. We then take the observations for which $\chi^2 \geq 3.33$. For each of these, we calculate its probability by the multinomial formula. Finally, we sum these probabilities.

In the present example, this gives

$$P(\chi^2 \geq 3.33) = 0.21.$$

Thus the probability is 0.21 of getting a result which is as extreme as, or more extreme than, the observation (5,1,4). That is to say, the probability is 0.21 of getting a result whose <u>CS</u>-value is $\geq$ the <u>CS</u>-value of the observation (5,1,4).

For two more examples with the same Boston-Montreal model, consider the outcome (3,2,5) and the outcome (4,4,2). In the first case, we get $\chi_0^2 = 5.63$ (as noted above), and $P(\chi^2 \geq \chi_0^2) = 0.06$. In the second case we get $\chi_0^2 = 0.53$, and $P(\chi^2 \geq \chi_0^2) = 0.92$. Thus the first of these observations is much more unexpected than the second. This may have been clear to begin with, but the <u>CS</u> metric and the descriptive level of significance give us a precise measure of this unexpectedness. (Note, again, that the smaller the descriptive level of significance, the more unexpected or surprising or extreme the observation will be for the chosen model.)

In the examples just above, we have used lengthy exact calculations to find the value of the <u>DSL</u> under the <u>CS</u> metric. In the next chapter, we shall see that a simple approximation method exists for getting easy and accurate values for the <u>DLS</u> under the

the *CS* metric.

As noted above, if we take a multinomial experiment of 2 categories and label one category *success* and the other category *failure*, we have a binomial experiment. The usual binomial metric for this binomial experiment gives the same *DLS* values as the *CS* metric does when we view the experiment as a multinomial experiment. To show this, compare the binomial metric, which gives $|X-np|$, with the *CS* metric which gives

$$\frac{(X-np)^2}{np} + \frac{((n-X) - nq)^2}{nq} .$$

The latter reduces, by elementary algebra, to

$$\frac{(x-np)^2}{npq} ,$$

and this expression is evidently equivalent, as a metric, to $|X-np|$. See Exercise 10-18 below.

Note. In the example of rolling a die, an *observation* consisted of a single number x, the number of successes in 600 trials. In the experiment of looking at 10 hockey games, a single *observation* consisted of a triple of numbers. The *theoretically expected result* also consisted of a triple of numbers, and the *metric* was a formula which was applied to the triple of numbers of the observation to give some kind of combined measure of how far this triple might lie from the theoretically expected triple. Typically in statistics, we work with observations which take the form of pairs, triples, quadruples, or n-tuples of numbers.

Three comments. We conclude with three comments on topics that will be considered further in later chapters.

(1) More on choosing a metric. In Chapter 20 we shall see that, given a model, a best metric (in a sense to be defined) can be found, provided: (a) that we have a practical decision to be made on the basis of the observation; and (b) that we agree on

(i) the collection of all models to be considered as possible alternatives to the given model,

(ii) the degree of belief (in a sense to be defined) that we have, before any observation is made, in each of the possible alternative models, and

(iii) the eventual cost (in a sense to be defined), for each pair of possible models μ_1 and μ_2, of basing our decision on the assumption of μ_1 when, in reality, μ_2 is correct. We shall see that any two such best metrics are equivalent in that they give the same DLS values.

This theorem and the concepts of cost and degree of belief upon which it rests will not be available to us until Chapter 20. In the meantime, we will usually not have, in any given situation, an obvious metric which is uniquely determined by that situation. In each such case, the choice of metric that we make will seem, to some extent, arbitrary. As we shall see, the collection of alternative models which we allow will play a major role in our choice of metric, and we will also be influenced, in that choice, by intuitive considerations having to do with cost of making mistakes and strength of belief in various alternative models.

These intuitive considerations are later made precise in Chapter 20.

(2) Sets of models. In each example above, we have had a single fixed model, have chosen a metric, and have used the metric to calculate a DLS for a given observation. In the next few chapters we shall also see examples where we begin not with a single fixed model, but with a fixed set of models (a subset of the collection of all possible alternative models), and where we seek to find a useful metric that leads, for any given observation, to an appropriate DLS value that is the same for all models in that fixed set. Such a useful and remarkable metric will thus enable us, by this common DLS value, to measure the extent to which the observation confirms the general conclusion that there is some (unspecified) member of this fixed set with which the observation agrees. (For example, in Chapter 10, we shall see, for certain experiments, how to measure the extent to which observed data confirm the general conclusion that there is some (unspecified) member of the set of all Poisson distributions with which the observation agrees; that is to say, the extent to which the data may confirm that the experiment is a Poisson experiment.)

(3) The fundamental metric (a technical comment). In the discussion and examples above, we have introduced the concept of metric in order to describe how well a given observation confirms a chosen model. We have then used the metric (which gives distance as a real number) to calculate a DLS for the

given observation and chosen model. Clearly, however, the only information needed in order to determine a *DLS* value is information as to the *relative* strength with which different observations confirm the chosen model, because the *DLS* only depends on knowing which observations do not confirm the model more strongly than the given observation. It is therefore enough, in order to get *DLS* values, to have a certain *relation* defined among possible observations. We write the relation as $\succsim_\mu$, and we read "$\omega_1 \succsim_\mu \omega_2$" as "$\omega_1$ confirms the chosen model μ at least as strongly as ω_2".

If we have such a relation, it must satisfy the following formal laws (in order to be intuitively acceptable):

(1) For any observations ω_1 and ω_2 , either

$$\omega_1 \succsim_\mu \omega_2 \text{ or } \omega_2 \succsim_\mu \omega_1 \text{ ;}$$

(2) For any observations ω_1,ω_2 , and ω_3 , if

$$\omega_1 \succsim_\mu \omega_2 \text{ and } \omega_2 \succsim_\mu \omega_3 \text{ , then } \omega_1 \succsim_\mu \omega_3 .$$

A relation which satisfies (1) and (2) is called a *pre-ordering*.

Given a pre-ordering $\succsim_\mu$, we can then define the notation $\omega_1 \succ_\mu \omega_2$ to mean that $\omega_1 \succsim_\mu \omega_2$ holds but that $\omega_2 \succsim_\mu \omega_1$ does not hold, and we can read "$\omega_1 \succ_\mu \omega_2$" as "$\omega_1$ confirms μ more strongly than ω_2."

It is immediate that we can define an appropriate *DLS*, of a given observation ω_o , from this relation as

$$\text{DLS of. } \omega_o \text{ under } \mu = P_\mu(\omega_o \succsim_\mu \Omega).$$

We can also, if we wish, define a natural metric s^*_μ from $\succsim_\mu$ as follows. For each observation ω_o, we define

$$s^*_\mu(\omega_o) = P_\mu(\Omega \succ \omega_o).$$

This metric measures the distance of ω_o (from giving strongest confirmation of μ) as the probability of finding an observation that gives *stronger* confirmation. We call s^*_μ the *fundamental metric* for the given relation $\succsim_\mu$, and we see that since $P_\mu(\omega_o \succsim_\mu \Omega) + P_\mu(\Omega \succ_\mu \omega_o) = 1$, the *DLS* and the fundamental metric are related in a very simple way:

$$\underline{\text{DLS}} \text{ of } \omega_o \text{ under } \mu = 1 - s^*_\mu(\omega_o).$$

Of course, any given metric s for a chosen μ defines a relation $\succsim_\mu$ by:

$$\omega_1 \succsim_\mu \omega_2 \iff s(\omega_1) \leq s(\omega_2);$$

and the relation $\succsim_\mu$, in turn, defines a fundamental metric s^*_μ. It is possible to prove the following theorem relating the various concepts above.

Theorem. Two metrics (for a chosen model μ) are equivalent if and only if they determine the same relation $\succsim_\mu$, and they determine the same relation $\succsim_\mu$ if and only if they determine (and are equivalent to) the same fundamental metric.

[The proof for discrete models is immediate. For continuous models, one must assume continuity of the probability density function of the model.]

From a conceptual point of view, metrics, as introduced earlier in this chapter, may be viewed as a convenient way of introducing, describing, and working with a pre-ordering $\succsim_\mu$ of observations from which DLS values can be calculated. There are also other ways of directly defining and describing such a pre-ordering. One choice, for example, would be to define $\omega_1 \succsim_\mu \omega_2 \Longleftrightarrow P_\mu(\Omega = \omega_1) \geq P_\mu(\Omega = \omega_2)$. This is the same pre-ordering as we would get by using the metric $s(\omega) = 1 - P_\mu(\Omega = \omega)$.

EXERCISES FOR CHAPTER 10.

Note. (1) In the following exercises, normal approximation should be used wherever appropriate. (2) Finding a DLS requires that we have a chosen model μ. In some of the exercises below, this model will be indicated only indirectly by the wording used. Exercise 8-2, for example, suggests that we use a binomial model with $p = 0.4$ and $n = 100$. (3) In the case of binomial models, unless otherwise indicated, the metric $s(X) = |X-np|$ should be used.

10-1. A fair coin is tossed 1000times. 490 heads are obtained. What is the DLS of this observation?

10-2 A thumbtack has probability 0.4 of landing on its back when tossed. In 100 tosses, we observe that it lands on its back 50 times. What is the DLS of this observation?

10-3. A bridge player played 12 bridge hands and got no aces in 8 of them.

(a) Assuming good shuffling, find the DLS of his observation. (As normal approximation does not apply, a direct calculation, using a calculator, must be made.) Is it reasonable for the player to complain of poor shuffling?

(b) What number of heads in 100 tosses of a fair coin would give approximately the same DLS?

10-4. The analysis of professional hockey games given in Chapter 7 suggested a simplified binomial model in which the probability of a tie game is 0.158. Assuming this model, find the DLS of an observation of 132 ties in 720 games.

10-5. A fair die is rolled 600 times and 130 sixes are observed. What is the DLS of this observation if we use the metric

$$s(X) = \begin{cases} \frac{X-np}{n-np} & \text{for} \quad X \geq np \\ \\ \frac{np-X}{np} & \text{for} \quad X < np? \end{cases}$$

(This was the metric called s_2 in the text above.)

10-6. Let A be the event that either 0 or 1 occurs when a single digit is taken from a table of random digits. In a given table, you examine a sequence of 100 digits and find that the event A occurs 12 times.

(a) Find the DLS of this observation using the usual metric for a binomial model.

(b) Find the DLS of this observation using the metric given in Exercise 10-5.

10-7. For a series of ten games between the Boston and Montreal hockey teams, assume that Boston has probabilities 0.4, 0.5, and 0.1 of winning, losing, and tying respectively. Assume further that the result is independent from game to game. Give the value of the *CS* metric for each of the following observations, and use these values to arrange the observations in order from most confirming to least confirming:

(3 wins, 3 losses, 4 ties), (5 wins, 5 losses, 0 ties),
(4 wins, 4 losses, 2 ties), (6 wins, 1 loss, 3 ties).

10-8. A fair die is rolled 60 times. Let $(x_1, x_2, \ldots, x_6)$ indicate the observation of obtaining x_1 *ones*, x_2 *twos*,..., x_6 *sixes*, where $x_1 + x_2 + \ldots + x_6 = 60$. Give the value of the *CS* metric for each of the following observations, and use these values to arrange the observations in order from most confirming to least confirming: (9,10,11,9,12,9) (7,14,9,10,9,11), (8,9,10,12,10,11).

10-9. In 100 times at bat, a baseball player gets 25 single-base hits and 5 extra-base hits. Assume that his probability of getting a single-base hit is 0.2 and of getting an extra-base hit is 0.1. Find the value of the *CS* metric for his performance.

10-10. In a multinomial experiment of 3 categories with $n = 3$, we take the model $p_1 = 1/2$, $p_2 = 1/3$, $p_3 = 1/6$. Using the *CS* metric, find by direct calculation the *DLS* of each of the following observations: (1,0,2), (1,1,1), (0,2,1), and (0,0,3). (*Hint*. Make a table with three columns; list observations in the first column, CS-values in the second, and probabilities in the third.)

10-11. Use random digits to simulate 25 trials of the multinomial experiment in Exercise 10-10. What relative frequency do you observe for outcomes with CS-values at least as great as the CS-value for the outcome $(0,2,1)$? (Use the table made for Exercise 10-10.) (Suggestion. Use lines 37 through 39 on page 233. Let digits 1, 2, 3 represent the first category, 4, 5 the second category, and 6 the third. Ignore 7, 8, 9, 10.)

10-12. A single observation of a certain Poisson experiment yields the observation $X = 8$. Assume the model $m = 4$, use the metric $s(X) = |X-m|$, and find the DLS of the observation.

10-13. A manufacturer supplies scissors in boxes, loosely packed. He claims that the average number of pairs of scissors per box is 100. You open one box and find 75 pairs of scissors. Assuming a Poisson distribution model, what is the DLS of your observation if the manufacturer's claim is correct? (Use the metric $s(X) = |X-m|$.)

10-14. A trial is made of a binomial experiment with $n = 100$ and fixed but unknown p. The result is $X = 40$. For what values of p will the DLS of this observation be greater than 0.05? (Use normal approximation.)

10-15. A multiple choice test has 100 questions. Each question has 5 possible answers. A student gets the correct answer to 33 of the questions. Assume that the student has chosen all answers at random.

(a) Estimate the DLS of this result using the usual binomial metric $s(X) = |X-np|$.

(b) Estimate the DLS of this result using the metric $s(X) = \frac{1}{2}(|X-np| + (X-np))$.

Comment. Exercise 9-15 is an example where the choice of a metric (and hence the view that one takes of what constitutes confirming or disconfirming evidence) will depend upon what one considers as possible alternative models. If (as would normally be the case) the only alternatives are that the student has a better than 20 percent chance of getting each question correct, then the metric (b) should be used, since an observation such as $X = 7$ would not disconfirm the assumed model at all in comparison with the other possible models. On the other hand, if the multiple choice test is cunningly designed so that a student with incomplete knowledge will be tempted into wrong answers, then the possible alternative models would include models in which the student has a less than 20 percent chance of getting each question right, and metric (a) should be used. We discuss this matter further in later chapters. Exercise 10-15 should be compared with Exercise 10-16 below. (In Chapter 20, we shall see that a metric intermediate between (a) and (b) can also be used, and may be the best choice of all.)

10-16. In an experiment for extra-sensory perception (ESP), a card guessing test is used in which the probability of a correct response by pure chance (in the absence of real ESP) is 1/5.

(a) In 50 independent trials, a subject gives 14 correct responses. If you assume the pure chance model, what is the DLS of this observation?

(b) What is the DLS of 24 correct responses in 50 trials?

Comment. What metric should be used in this problem? One might argue that the only alternative models are models in which the subject does better than chance, and hence that the metric of Exercise 9-15b should be used. On the other hand, as investigators sympathetic to the existence of ESP have themselves suggested, it is conceivable that a subject could have ESP powers, but could misinterpret them in a way that would lead to a worse than pure chance binomial model. For this reason, the usual binomial metric $s(X) = |X-np|$ seems to be the best choice for this exercise.

10-17. Five repeated observations are made of a binomial experiment of 2000 trials, and the results 950, 906, 1020, 984, and 1005 are obtained. Assume the model

$p = 1/2$, and consider the five values as a single observation (for the larger experiment of conducting 5 trials of the binomial experiment). What is the DLS of this observation?

Comment. What metric should be used in Exercise 9-17? The reader will most likely choose to use the usual binomial metric $(s(X) = |X-np|)$ for a binomial model with $n = 10,000$. This is correct, since the exercise asserts, as a given fact, that the original experiment (with $n = 2000$) is binomial. If non-binomial models were possible alternatives to the chosen model, however, the usual metric might not be the best choice. Consider, for example, the five observed values 1000, 1000, 1000, 1000, 1000. These are highly confirming under the usual metric (giving DLS = 1). If it were a possible alternative that the original 2000 trial experiment was not binomial and in fact always gave the same observed value, then these observations would be seen as strongly disconfirming the binomial model with $p = 1/2$.

10-18. Show that, for a given binomial model μ, the usual binomial metric and the CS metric (for a multinomial experiment of 2 categories) are equivalent.

10-19. Prove the following statements:

(a) Given a chosen discrete model μ, two metrics for μ are equivalent if and only if they determine the same relation $\succsim_\mu$.

(b) Given a chosen discrete model μ, two metrics for μ determine the same relation $\succsim_\mu$ if and only if they determine (and are equivalent to) the same fundamental metric.