ON THE RELATION BETWEEN CALCULUS OF PROBABILITY AND STATISTICS

by

Karl Menger

In a paragraph following his elegant presentation of the Calculus of Probability, Professor Copeland touches on the problem of the application of the theory to actual observations. In the sequel we shall elaborate further on the connection between the calculus of probability and statistics in order to establish a bridge from Copeland's exposition to Wald's introduction to modern statistics contained in Number 1 of the Notre Dame Mathematical Lectures.

In applying the concept of p(x) to sequences of observations, we encounter the difficulty that all sequences of observations are finite whereas p(x) refers to the infinite sequence x of Zeros and Ones. By definition, p(x) is the limit of the relative frequencies of Zeros in the finite initial segments of the sequence x, as these segments increase in length. Clearly, the knowledge of a finite initial segment of x, no matter how long the segment may be, does not permit any apodictical conclusion concerning the relative frequency in any more extended initial segment or concerning the limit p(x) of these relative frequencies. In other words, logically, each finite initial segment of x is compatible with each hypothesis concerning p(x). Or, in still other words, on the basis of a finite sequence of observations we can neither assert nor

deny any hypothesis concerning p(x).

However, conventions have been made by which we test hypotheses concerning the value of p(x) on the basis of a finite sequence of observations. More specifically, given a finite sequence of observations, stipulations have been made as to when the hypothesis that p(x) has a certain value, should be rejected. We decide to follow these stipulations in full realization of the possibility of errors. In this respect the decision is not safer than many decisions which as practical individuals we take every day, and which we have to take at the risk of errors underlying, or dangers implied by, the decision - since without a decision we could not act at all.

In fact, we realize that our decision concerning the rejection of a hypothesis concerning p(x) (or of any hypothesis, at that) is subject to errors of two types. We may commit what is called an error of first type by rejecting a hypothesis although it is true, or an error of second type by not rejecting a hypothesis although it is false. In setting up a test for the rejection of a hypothesis our main aim is to avoid the former danger. The probability of rejecting a true hypothesis by applying a test, is called the standard of inaccuracy of the test. The smaller it is, the more desirable is the test. We shall try to develop a test of a preassigned standard of inaccuracy. In doing so we shall find that for a given percentage α there are many tests whose standard is α percent. Among these tests we shall select the one for which the probability of errors of the second type is as small as possible. However, only if the finite sequence of observations on which the test is based is very long, can we hope that also the

probability of errors of second type will be really small.

As a first example, we shall develop stipulations as to when, on the basis of n observations, one should reject the hypothesis that $p(x) = 1/2$. The rule will have a preassigned standard of α percent. It will test the hypothesis $p(x) = 1/2$ against the class of hypotheses that $p(x) = y$ for some value y between 0 and 1. That a test of a hypothesis should be against a well-defined class of hypotheses, and that the stipulation at which we arrive depends upon that class, is an important insight of modern statistics. For instance, when we shall modify our example and test the same hypothesis $p(x) = 1/2$ against the class of two hypotheses $p(x) = 1/2$ or $p(x) = 1/4$, we shall arrive at a completely different stipulation.

We form the set of all the 2^n possible ordered n-tuples of Zeros and Ones, each representing one possible outcome of a sequence of n observations. In statistics this set is called the sample space of our problem. Our task is to select a subset of the sample space with the stipulation that the hypothesis $p(x) = 1/2$ be rejected if the observed n-tuple of observations should belong to this subset. Statisticians call this subset of rejection the critical region of our problem. The probability that under the assumption of the truth of the hypothesis $p(x) = 1/2$ a point falls within the subset of rejection, is to be α percent in order to meet the preassigned standard. Now under the hypothesis that $p(x) = 1/2$, the probability that the observed n-tuple of Zeros and Ones be a given point of the sample space, i.e., a given ordered n-tuple of Zeros and Ones, is $1/2^n$ for each point of the sample space. For under the assumption $p(x) = 1/2$ it is as likely that the outcome of 8 observations will be 00000000 as that it will be

00000001 or 10101010 or 11111111. Of course, we are more likely to observe 1 One and 7 Zeros than 8 Zeros. But 1 One and 7 Zeros correspond to 8 different points of the sample space (00000001, 00000010, etc.) each of which under the hypothesis $p(x) = 1/2$ is as likely as 00000000. Under any other assumption concerning the value of $p(x)$ some n-tuples would be more likely than other ones. E.g., under the assumption $p(x) = 1/4$ the probability of observing k Zeros and n-k Ones would be $(1/4)^k \cdot (3/4)^{n-k}$. In this case, as we shall see in another example, in forming a subset of rejection of standard α percent, that is, in uniting points for which the sum of the probabilities is α percent, we have to weigh, rather than merely count, the different points. But under the assumption $p(x) = 1/2$, in order to select a subset of rejection of a standard of α percent, it is necessary and sufficient to unite in any way α percent of the 2^n points of the sample space, thus to pick out the right number of points.

Obviously, this can be done in many ways. We thus propose to select a subset of rejection for which the probability of an error of second type (i.e., of not rejecting a false hypothesis) is as small as possible, in other words, a set for which, under the assumption that $p(x)$ is $\neq 1/2$, the probability of a point belonging to the set is as large as possible. This is where the class of hypotheses against which the hypothesis $p(x) = 1/2$ is tested, comes in. E.g, in the modified example in which we test the hypothesis against the class consisting of the two hypotheses $p(x) = 1/2$ and $p(x) = 1/4$ only, we have to select α percent of the 2^n points of the sample space in

such a way that under the hypothesis $p(x) = 1/4$ a point is more likely to belong to this set of rejection than to any other set comprising α percent of the 2^n points.

Let us consider the modified example for $n = 8$ and $\alpha = 3.5$. Of the 256 points of the sample space each has the probability $1/2^8$. Thus no matter against which class of hypotheses we wish to test the hypothesis $p(x) = 1/2$, we shall have to select a set of rejection comprising 9 of the 256 points. Now under the assumption $p(x) = 1/4$, the point (i.e., the octuple of Zeros and Ones) with 8 Ones has the probability $(3/8)^8$, each of the 8 points with 7 Ones and 1 Zero has the probability $(1/4)\cdot(3/4)^7$, each of the 28 points with 6 Ones and 2 Zeros has the probability $(1/2)^2\cdot(3/4)^6$, etc. Of all the subsets comprising 9 of the 256 points, under the hypothesis that $p(x) = 1/4$, the set of 9 points with at least 7 Ones has by far the largest probability. Under the hypothesis $p(x) = 1/4$, the probability of a point belonging to this set, that is, of avoiding an error of the second type, is $(3/4)^8 + 8(1/4)\cdot(3/4)^7$ which is approximately .367. Clearly, this set of 9 points will be our set of rejection. That is to say, in testing the hypothesis $p(x) = 1/2$ against the alternative $p(x) = 1/2$ or $p(x) = 1/4$ on the basis of 8 observations and with a standard of inaccuracy of 3.5 percent, we decide to reject the hypothesis $p(x) = 1/2$ if and only if there are at least 7 Ones among our 8 observations. This decision has a 96.5 percent chance of avoiding the error of rejecting a true hypothesis, and a 36.7 percent chance of avoiding the error of not rejecting a false hypothesis.

The latter aspect of our test is not too impressive. But then the sequence of 8 observations is short. If we base our test on 20 observations, we shall considerably reduce the probability of an error of the second type though at the same time we will preassign a lower standard of inaccuracy, namely, 2 percent as compared with 3.5 percent in the preceding discussion. Among the 2^{20},(i.e., about 10^6)points of our sample space, each has the probability $1/2^{20}$ under the assumption that $p(x) = 1/2$. Under the assumption that $p(x) = 1/4$, a point representing k Ones and 20 - k Zeros has the probability $(3/4)^k \cdot (1/4)^{20-k}$. The space contains

1, 20, 190, 1140, 4845, 15504 points for k = 20, 19, 18, 17, 16, 15, respectively. These points together form about 2 percent of the approximately one million points of the space and, under the assumption that $p(x) = 1/4$, this particular set has a greater probability than any other subset containing equally many points. This probability can easily be seen to be about 63 percent. For n = 100 the probability of avoiding errors of the second type would be still much larger.

Another way of decreasing the danger of errors of second type is, of course, to relax the requirements concerning the errors of first type. By admitting a higher standard of inaccuracy, that is, a larger set of rejection, we enhance the probability that under the hypothesis $p(x) = 1/4$ a point belongs to the set of rejection, and that an error of second type is avoided. For instance, if in the case n = 20 of our modified example we are satisfied with a standard of 6 percent,

thus with a 94 percent chance of avoiding an error of first type, then we have to reject the hypothesis $p(x) = 1/2$ if there are 14 or more Ones among the 20 observations. By doing so we have an 85 percent chance of avoiding an error of second type. If we admitted a standard of 14 percent, then we should have to reject the hypothesis $p(x) = 1/2$ whenever we find 13 or more Ones among 20 observations. In this case the probability of an error of second type would be only about 5 percent, thus considerably smaller than that of an error of first type.

Returning to the test of $p(x) = 1/2$ against the class of hypotheses that $p(x)$ has some value y between 0 and 1, we notice that, unfortunately, for different values of y, under the assumption that $p(x) = y$, different sets will minimize the probability of errors of the second type. For instance, it is clear that in testing $p(x) = 1/2$ against $p(x) = 1/4$ (or against $p(x) = y$ for any value of y which is $< 1/2$) the best set of rejection consists of points with many Ones and few Zeros. (We studied the cases $n = 8$ and 20 in some detail). On the other hand, we should find in an analogous way that in testing $p(x) = 1/2$ against $p(x) = 3/4$ (or against $p(x) = y$ for any value of y which is $> 1/2$) the best set of rejection consists of points with many Zeros and few Ones. For a preassigned standard of inaccuracy of α percent we thus make the following compromise: We select $\alpha/2$ percent of the 2^n points containing mostly Zeros and $\alpha/2$ percent of the points containing mostly Ones into the set of rejection. With this selection, if n is large, the probability of a point falling into the subset will be large even under the assumption that $p(x)$ deviates only slightly from 1/2.

Next we wish to test a hypothesis other than $p(x) = 1/2$ since in testing the latter we have made use of simplifications which are not possible in other cases. We shall test on the basis of 8 observations the hypothesis $p(x) = 1/4$ against the alternative that $p(x) = 1/4$ or $p(x) = 1/2$ with a standard of inaccuracy of 3 percent. The sample space consists of 2^8 points all of which are equally likely under the hypothesis $p(x) = 1/2$. But under the assumption that $p(x) = 1/4$ each of the

	1,	8,	28,	56,	70,	56,	28,	8,	1 points
with	0,	1,	2,	3,	4,	5,	6,	7,	8 Ones

has the probability of
$1/4^8$, $3/4^8$, $3^2/4^8$, $3^3/4^8$, $3^4/4^8$, $3^5/4^8$, $3^6/4^8$, $3^7/4^8$, $3^8/4^8$, respectively.

We have to select a set of points for which the sum of the probabilities is about .03. This can be done in many ways. But if we wish that under the assumption $p(x) = 1/2$ the probability of the set of rejection be as large as possible, we obviously shall select the points with few Ones. If we unite the 93 points with at most 3 Ones, we obtain the proper set of rejection since under the assumption that $p(x) = 1/4$ its probability is $(1 + 3.8 + 9.28 + 27.56)/4^8$ which is about .03. We thus see: In testing, on the basis of 8 observations, the hypothesis $p(x) = 1/4$ against the alternative with a standard of inaccuracy of 3 percent, we shall decide to reject the hypothesis $p(x) = 1/4$ if and only if among the 8 observations we find less than 4 Ones.

The reader should realize that the following two stipulations which we have made in this paper are not contradictory. Both refer to tests against the alternative of hypotheses, that $p(x) = 1/4$ or $p(x) = 1/2$, on the basis of 8 observations with a standard of 3 percent.

1) If we test the hypothesis $p(x) = 1/4$, then we decide to reject this hypothesis, and thus to adopt the hypothesis $p(x) = 1/2$, if and only if among the 8 observations we find 3 or less than 3 Ones.

2) If we test the hypothesis $p(x) = 1/2$, then we decide to reject this hypothesis, and thus to adopt the hypothesis $p(x) = 1/4$, if and only if among the 8 observations we find 7 or 8 Ones.

We see that in case of 5 or 6 Ones among 8 observations we adopt that hypothesis which we are testing: we adopt $p(x) = 1/4$ if we are testing $p(x) = 1/4$, and we adopt $p(x) = 1/2$ if we are testing $p(x) = 1/2$. This procedure is not inconsistent because in both cases the probability of an error of first type is to be 3 percent, and this requirement has a different meaning in the two cases. In testing $p(x) = 1/4$ it means that the danger of $p(x) = 1/4$ being true and yet rejected is 3 percent. In testing $p(x) = 1/2$ it means that the danger of $p(x) = 1/2$ being true and yet rejected is 3 percent.

In concluding this brief exposition of the connection between the calculus of probability and statistics, we mention that a purely statistical theory would not transcend the domain of the observable, thus would exclusively deal with finite

sequences without referring to probability, except as an abbreviated way of discussing relative frequencies in long sequences. In such a theory we should not stipulate the rejection or adoption of a hypothesis $p(x) = y$. We would not even mention such a hypothesis because x stands for an unobservable infinite sequence. We should describe, on the basis of one n-tuple of observations, our expectation concerning other n-tuples of observations, or on the basis of a sequence of n observations, our expectation concerning an $(n+1)$-st observation or concerning a more extended finite sequence of observations. It is, of course, easy to rephrase according to this point of view the stipulations discussed in this outline. Somewhat more complicated would be the rephrasing of the claims made about these stipulations, especially concerning the "probability" of excluding errors of first and second type by forming the stipulated decision.