I INTRODUCTION

The purpose of statistics, like that of geometry or physics, is to describe certain real phenomena. The objects of the real world can never be described in such a complete and exact way that they could form the basis of an exact theory. We have to replace them by some idealized objects, defined explicitly or implicitly by a system of axioms. For instance, in geometry we define the basic notions "point," "straight line," and "plane" implicitly by a system of axioms. They take the place of empirical points, straight lines and planes which are not capable of exact definition. In order to apply the theory to real phenomena, we need some rules for establishing the correspondence between the idealized objects of the theory and those of the real world. These rules will always be somewhat vague and can never form a part of the theory itself.

The purpose of statistics is to describe certain aspects of mass phenomena and repetitive events. The fundamental notion used is that of "probability." In the theory it is defined either explicitly or implicitly by a system of axioms. For instance, Mises[1] defines the probability of an event as the limit of the relative frequency of this event in an infinite sequence of trials satisfying certain conditions. This is an explicit definition of probability. Kolmogoroff[2] defines probability as a set function which satisfies a certain system

---

1) See references 10 and 11

2) See reference 9

of axioms. These idealized mathematical definitions are related to the applications of the theory by translating the statement "the event E has the probability p" into the statement "the relative frequency of the event E in a long sequence of trials is approximately equal to p." This translation of a theoretical statement into an empirical statement is necessarily somewhat vague, for we have said nothing about the meanings of the words "long" or "approximately." But such vagueness is always associated with the application of theory to real phenomena.

It should be remarked that instead of the above translation of the word "probability" it is satisfactory to use the following somewhat simpler one: "The event E has a probability near to one" is translated into "it is practically certain that the event E will occur in a single trial." In fact, if an event E has the probability p then, according to a theorem of Bernoulli, the probability that the relative frequency of E in a sequence of trials will be in a small neighborhood of p is arbitrarily near to 1 for a sufficiently long sequence of trials. If we translate the expression "probability nearly 1" into "practical certainty," we obtain the statement "it is practically certain that the relative frequency of E in a long sequence of trials will be in a small neighborhood of p."

In statistics we always construct some probability schemes which we believe to be adequate to describe certain real phenomena. For instance, we describe the situation concerning the possible outcomes in tossing a coin by saying that the probability of obtaining a head in one toss is 1/2, for in a long se-

quence of trials we would expect to have about half as many heads as total tosses. Or, if we measure the length of a bar by some instrument, we sometimes assume that the result is a normally distributed random variable. The notions of a random variable and a distribution function are defined as follows: if $F(x)$ is a function expressing the probability that a real variable $X < x$, we say that X is a random variable and that $F(x)$ is the probability distribution of X. Then, if $F(x)$ is given by the formula

$$(1) \qquad F(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} e^{\frac{-1}{2}\frac{(y-\mu)^2}{\sigma^2}}\, dy$$

we say that X is normally distributed. The quantities σ and μ are real parameters. Thus, if in measuring the length of a bar by some instrument we assume that the outcome of the measurement is a normally distributed random variable, we may express the probability that a measurement will be less than a given value x by (1).

If $X_1, X_2, X_3, \ldots, X_n$ represent n random variables and $x_1, x_2, \ldots, x_n$ any set of real numbers, we use the symbol $F(x_1, x_2, \ldots, x_n)$ to express the probability of the composite event that $X_1 < x_1$, $X_2 < x_2, \ldots, X_n < x_n$ simultaneously. This function will be called the joint probability distribution of the n random variables. We shall say that n random variables are independently distributed if the function $F(x_1, x_2, \ldots, x_n)$ is the product of n functions such that only x_1 is involved in the first, only x_2 in the second, and so on. That is

$$F(x) = f_1(x_1)f_2(x_2)\ldots f_n(x_n).$$

For example, if n measurements $X_1, X_2, \ldots, X_n$ of a bar are independently and normally distributed with the same normal distribution, we would obtain

(2) $F(x_1, x_2, \ldots, x_n) =$

$$= \frac{1}{(2\pi)^{n/2}\sigma^2} \int_{-\infty}^{x_1} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy \int_{-\infty}^{x_2} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy \ldots \int_{-\infty}^{x_n} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy$$

If we measure the length of a bar n times by some instrument, we sometimes find it appropriate to adopt the probability scheme that the results of the n measurements have a joint probability distribution given by (2).

One of the fundamental problems of statistical inference is that of testing statistical hypotheses. The most general form of a <u>statistical hypothesis</u> we have to deal with in statistical theory may be expressed as follows. Let $X_1, \ldots, X_n$ be a finite set of random variables and let $F(x_1, \ldots, x_n)$ be its joint probability distribution function. Then the statistical hypothesis is the statement that the unknown distribution function $F(x_1, \ldots, x_n)$ is an element of a certain class ω of distribution functions. For instance, if $X_1, \ldots X_n$ are successive measurements on the length of a bar, we may consider the hypothesis that $X_1, \ldots, X_n$ are independently distributed with the same normal distribution. In this case ω is a two parameter family given by (2), σ being any positive number and μ any real number.

If we consider the hypothesis that $X_1,\ldots,X_n$ are normally, independently distributed with zero means ($\mu=0$) and unit variances ($\sigma^2=1$), then ω consists of a single element. When the class ω consists of a single element, we shall say that the hypothesis we are considering is a simple hypothesis. Otherwise, it will be called composite.

The question of testing a given hypothesis may be formulated in the following manner. We should like to know, on the basis of n observations $x_1,\ldots,x_n$ where x_α is the observed value of the random variable X_α ($\alpha=1,\ldots,n$), whether to accept or reject the hypothesis H_ω that the unknown distribution function $F(x_1,\ldots,x_n)$ belongs to the class ω. The set of n observations can be represented by a point E of n-dimensional Cartesian space, called the sample space. To test the hypothesis H_ω on the basis of n observations we must choose a subset R of the sample space and then reject the hypothesis H_ω if the sample point E falls within R. Otherwise, we maintain the hypothesis. It is evident that the fundamental problem here is the choice of the subset R, which we shall call the critical region. The solution of this problem depends, to some extent, upon any a priori knowledge we may have about the unknown distribution function $F(x_1,\ldots,x_n)$. One of the most important and most frequent a priori assumptions is that the random variables $X_1,\ldots,X_n$ are independently distributed, each having the same distribution. Thus, we have the assumption that F is of the form

$$F(x_1,\ldots,x_n) = \prod_1^n \varphi_i(x_i) \quad \text{where } \varphi_i = \varphi_j \text{ for all } i, j.$$

Such a priori knowledge about our unknown distribution function can always be expressed by saying that the function

$F(x_1,\ldots,x_n)$ is an element of a certain class Ω of distribution functions. The class ω which is being considered is then always a subclass of Ω. We shall see that the choice of the critical region R for testing the hypothesis H_ω will depend upon the a priori knowledge Ω.

It is now seen that the problem of testing hypotheses can be formulated as follows: Taking for granted that the unknown distribution function F is an element of a class Ω, we wish to test the hypothesis that F belongs to a certain subclass ω of Ω. The problem to be solved is the question of how the critical region in the sample space should be chosen.

For instance, Ω may be defined by the statement that $X_1,\ldots,X_n$ are independently and normally distributed each of them having the same distribution, and ω may be the subclass of Ω defined by the additional restriction that the mean values of $X_1,\ldots,X_n$ are zero. In this case, according to certain standards we will discuss later, the adequate critical region is given by the inequality

$$\left|\frac{\bar{x}\sqrt{n}}{s}\right| \geq c$$

where $\bar{x} = \frac{x_1+\ldots+x_n}{n}$ and $s^2 = \frac{\sum_{\alpha=1}^{n}(x_\alpha-\bar{x})^2}{n-1}$

and c is a certain constant. If, however, Ω is a much broader class defined by the statement that $X_1,\ldots,X_n$ are independently distributed each having the same distribution, the above critical region for testing H_ω is not adequate, and some other critical region has to be chosen.

Before we proceed farther it might be well for us to list a few of the mathematical terms used together with their meanings in statistics. We can do this in tabular form.

MATHEMATICAL TERMINOLOGY	STATISTICAL INTERPRETATION
n space, E_n (sample space)	Possible outcome of n observations.
Ω, class of functions on E_n	Class of possible probability distributions.
ω, subclass of Ω	The statistical hypothesis. The true distribution is a member of ω.
R, (critical region), a subset of E_n	Criterion for rejecting the hypothesis that the true distribution is a member of ω.
Association of R with Ω and ω.	Choice of the critical region for testing the hypothesis.

The problem of testing hypotheses is only one of the problems of statistical inference. Another is the problem of estimation. Given that the unknown distribution function F belongs to a certain class Ω of distribution functions, how can we choose a function $\varphi(E)$ defined for all points E of E_n such that the value of $\varphi(E)$ is always an element of Ω and can be considered a "good" estimate of the unknown distribution function F? We may say that $\varphi(E)$ is a "good statistical estimate" of F if the probability is as large as possible that $\varphi(E)$ is in a small neighborhood of F. We will formulate this principle more precisely in chapter III.

If, for instance, Ω is given by the statement that $X_1,\ldots,X_n$ are independently and normally distributed with the same means and unit variances, then Ω is a one parameter family of distribution functions and an element of Ω is completely specified by specifying the value of the unknown mean μ.

Hence, to estimate the unknown distribution function F is the same as to estimate the unknown mean μ. In this case the problem of estimation is the problem of finding a real function $\varphi(E)$ defined for all points E of the sample space such that $\varphi(E)$ can be considered as a statistical estimate of the unknown mean μ. The classical solution of this problem in this particular case is given by

$$\varphi(E) = \frac{x_1+\ldots+x_n}{n} .$$

The two types of problems of statistical inference mentioned so far do not cover all possible problems.[3] The following problem, for example, is neither a problem of testing a hypothesis nor one of estimation: Consider three subclasses ω_1, ω_2, ω_3 of the class Ω of distribution functions, and denote by H_{ω_1} the hypothesis that the unknown distribution F is an element of ω_1. The problem considered is to decide on the basis of the n observations which of the three hypotheses should be accepted (assume that the sum of the three subclasses ω_1, ω_2, ω_3 is equal to Ω). Such a situation may arise, for instance, in the case of a manufacturer who has to keep the quality of his product between two limits, and wants to test, by sampling, whether the quality is actually between these limits, below the lower limit, or above the upper limit. (Assume that the quality is measurable and can be represented by a real number.)

---

3) See in this connection 16, pp 299-300.

The reasons why such a "trilemma" is a problem different from testing a hypothesis or estimation can only be indicated here. It will be seen that there are many approaches to each problem of inference, and that the theory provides means of choosing among them by deciding that certain approaches are "better" than certain others. Now, one might suggest the reduction of the above "trilemma" to a problem of, say, estimation by estimating the unknown distribution function F and accepting that hypothesis which corresponds to the subclass in which the estimate of F is contained. This would be one answer to the trilemma, but by no means the "best" answer according to the standards developed.

The most general formulation of the problem of statistical inference is this: Let S be a system of subclasses of the class Ω of distribution functions. For each element s of S, consider the hypothesis H_s which states that the unknown distribution F is an element of s; denote by H_S the system of all such hypotheses; the problem is to decide, by means of a sample which element of H_S should be accepted.

The problems enumerated before are special cases of this general problem. If S consists of two elements only, one being a subclass ω of Ω and the other its complement in Ω, the problem is the same as that of testing the hypothesis that the true distribution function F is an element of ω. If S is the system of all elements of Ω, we have the problem of estimation. If S consists of three classes ω_1, ω_2, ω_3 with the sum Ω, we have the trilemma.