CHAPTER X

PROBLEM OF INFERENCE

A. Testing Hypotheses (Sections 1-4)

1. The Basis of Statistical Inference

1.1. *The problem.* The analysis of statistical data as outlined in the foregoing chapter proceeds as follows; from the observations x_1, x_2, ..., x_n, the value of a function $x(x_1, x_2, ..., x_n)$ is derived (e.g., Pearson's X^2 or Student's t or the Lexis quotient); then, with some probabilistic assumption regarding the origin of the observations, the expectation, the variance, and if possible the distribution of $x(x_1, x_2, ..., x_n)$ are computed; finally the observed x-value and the result of the computation are "compared." Although no completely rational prescription for this "confrontation" could be given, it is in most practical cases a very workable procedure. If, for instance, the first 2000 figures in the development of π are studied on the assumption that each of the 10 digits has the same probability of appearance, one finds that X^2 can have any value between 0 and 18,000, but that it has approximately 90 % probability of lying in the limits 3.325 to 16.919; thus an observed value of $X_0^2 = 4.34$ (the value actually computed for $n = 2000$) will seem satisfactory (i.e., a confirmation of the assumptions) while a value $X^2 = 1000$ will prompt a rejection of the hypothesis.

The principal weakness of this argument is the following. The result of the computation: $\Pr\{3.325 \leqslant X^2 \leqslant 16.919\} = 0.90$ means that if a large number of investigations on sets of 2000 figures is carried out *and if each time, the assumption* (of 10 equal probabilities) *is fulfilled*, then, in the long run, an X^2 between 3.325 and 16.919 will be found in 90 % of all cases. This means we assume that a certain assumption holds and we draw conclusions from this assumption. We are solving a so-called "direct" problem (see Chapter IX, Section 5.1). What if the assumption is not fulfilled? If we want to pass judgment on the validity of an assumption, we cannot restrict ourselves to the cases for which it *is* fulfilled. To make this clear, let us assume that in the above problem

only one single parameter θ is hypothetical, e.g., that the probabilities of 3 and 7 are $0.1 + \theta$ and $0.1 - \theta$ (all others being 0.1). Then, if x is the chosen function of the variables $x_1, x_2, \ldots, x_n$, the first thing to discover would be the probability $p_n(x \mid \theta)$, that is, the probability of x under the condition that the parameter value is θ. The subscript n reminds us that x is a function of n observations.

The real *inference problem* starts indeed when two things are given: an observed value of x and a function of two variables $p_n(x \mid \theta)$, viz., the distribution of x depending on the parameter θ. Such problems have been dealt with in Chapter VII. We restate "Bayes' problem" here with a slightly changed notation: An alternative with the event probability $q = \theta$ has been tried n times; we take x as the number of events. The function $p_n(x \mid \theta)$ is given by the Bernoulli law

$$p_n(x \mid \theta) = \binom{n}{x}(1 - \theta)^{n-x}\theta^x. \tag{1}$$

The question is: *What can we say about θ if formula* (1) *and an observed value of x are given*? In general, the distribution may involve unknown parameters $\theta_1, \theta_2 \ldots \theta_l$ and we wish to draw conclusions about the values of these parameters. In the following we write θ either for one single parameter or as an abbreviation for all parameters, $\theta_1, \ldots, \theta_l$.

It must be anticipated from the beginning that no *determinate* answer can be expected, no answer of the form "θ *equals* 0.4" or "θ *lies in the limits* 0.3 *to* 0.5." It is in the nature of statistical inquiry that whatever statement we make can have a *certain chance* only of being correct. By chance we mean, of course, the frequency (relative number) of cases in which the statement is correct (if the whole procedure of testing a set of n trials is endlessly repeated in well-defined circumstances). A complete solution of the inference problem would enable us to give, for each interval H of θ, the chance $Q_n(H)$ (derived from the observed x and the function $p_n(x \mid \theta)$) that θ falls in H:

$$Q_n(H) = \text{chance}\{\theta \in H\}. \tag{2}$$

We use the term chance rather than probability to signify that *we do not assume that the sequence of repeated investigations must fulfill the randomness condition of a collective* (see Chapter I, Section 4).

Already in Chapter VII the decisive role of the magnitude of n in $Q_n(H)$ has been discussed. If, in $n = 4$ tossings of a coin, "heads" appeared three times, that is $x/n = 0.75$, no reasonable person will expect that much can be stated about the chance that $\theta > 0.5$. But, if in $n = 4000$ tossings 3000 gave the event "heads," which again results

in $x/n = 0.75$, the conclusion prompts itself that $Q_n(H)$—where H is the interval $\theta > 0.5$—practically equals 1 (that practically always when 3000 out of 4000 tossings show "heads," the event-probability θ for the coin will be greater than 0.5).

As was seen in Chapter VII, the reason lies in the fact that the value of $Q_n(H)$ depends not only on x and $p_n(x \mid \theta)$, but also on another function $p(\theta)$, whose influence on the numerical value of $Q_n(H)$ becomes less and less pronounced as n increases. This $p(\theta)$ is the *over-all chance* or *a priori chance* of a θ-value: Prior to, or independently of, the n trials to which the coin is subjected, something must be known about the occurrence of various θ-values in the universe of coins that are admitted to the statistical experiment. The statement, sometimes advanced as an objection, that θ is not a variable but an "unknown constant" having a unique value for the coin under consideration, is beside the point. Any statement concerning the chance of θ falling in an interval H (or concerning the chance of committing an error, etc.) is necessarily a statement about a universe of coins, measuring rods, etc., with various θ-values.

1.2. *Setting up inequalities for posterior distribution.* The formula giving $Q_n(H)$ if x, $p_n(x \mid \theta)$, and $p(\theta)$ are known, follows as was seen in Chapter VII, Section 1, from the rules of combining (multiplication) and partitioning (division):

$$Q_n(H) = \frac{\int_{(H)} p_n(x \mid \theta) p(\theta)\, d\theta}{\int p_n(x \mid \theta) p(\theta)\, d\theta}, \tag{3}$$

where the integral in the denominator is to be extended over all θ-values for which the density $p(\theta)$ is different from zero. It is thus seen that the value $Q_n(H)$ depends on $p(\theta)$ and is undetermined insofar as $p(\theta)$ is undetermined. The consequence is clear: on the one hand, we must use all information which we can get about $p(\theta)$; on the other hand, our answer should be no more precise than the information on which it was based.

One first approach to this consists in setting up inequalities for $Q_n(H)$ based on certain inequalities for the *a priori* probability. If the density $p(\theta)$ is assumed to be constant, it drops out and we have in this case

$$Q_n'(H) = \frac{\int_{(H)} p_n(x \mid \theta)\, d\theta}{\int p_n(x \mid \theta)\, d\theta}. \tag{4}$$

This quantity, the chance inferred under the assumption of constant over-all chance, is entirely determined by x and $p_n(x \mid \theta)$. Let B be an

interval that includes H: $B \supset H$, and denote by $Q_n(B), Q_n'(B)$ the values analogous to (3) and (4). We assume that the initial density $p(\theta)$ has the limits

$$\begin{aligned} m \leqslant p \leqslant M \quad &\text{in} \quad B \\ \bar{m} \leqslant p \leqslant \bar{M} \quad &\text{in} \quad \bar{B} \end{aligned} \tag{5}$$

where $\bar{B}$ means[1] the set of all θ-values not belonging to B. If we denote by Ω the total range of θ then $\bar{B} = \Omega - B$. Then the following two inequalities are evident from (3):

$$Q_n(H) \leqslant \frac{M\int_{(H)} p_n(x \mid \theta)\, d\theta}{m\int_{(B)} p_n(x \mid \theta)\, d\theta + \bar{m}\int_{(\bar{B})} p_n(x \mid \theta)\, d\theta} = \frac{MQ_n'(H)}{mQ_n'(B) + \bar{m}Q_n'(\bar{B})} \tag{6}$$

and, in the same way,

$$Q_n(H) \geqslant \frac{mQ_n'(H)}{MQ_n'(B) + \bar{M}Q_n'(\bar{B})}. \tag{6'}$$

Combining both relations, we obtain, since $Q_n'(\bar{B}) + Q_n'(B) = 1$,

$$\frac{m}{M}Q_n'(B) + \frac{\bar{m}}{M}[1 - Q_n'(B)] \leqslant \frac{Q_n'(H)}{Q_n(H)} \leqslant \frac{M}{m}Q_n'(B) + \frac{\bar{M}}{m}[1 - Q_n'(B)]. \tag{7}$$

This is the desired inequality. As an illustration, we consider the following application of (7). Assume that we are concerned with an inference where a great number, n, of trials is involved; then, the characteristic property of $p_n(x \mid \theta)$ is that, as n increases the distribution "concentrates" more and more around some point, say, θ_x, dependent on x. Now consider a sequence of nested intervals B which all enclose θ_x and converge toward the immediate neighborhood of θ_x and let n increase so that $Q_n'(B) \to 1$. If the density $p(\theta)$ is assumed continuous in the neighborhood of θ_x the difference between M and m tends to zero and if $m > 0$, both the first and the last expression in (7) tend to unity, that is, $Q_n'(H)/Q_n(H)$ tends toward one, or $Q_n(H) \to Q_n'(H)$. If, as n increases, H remains such a part of B that $Q_n'(H) = 50\,\%$, for example, then $Q_n(H) \to 0.50$. The result can be stated as follows:

If $p_n(x \mid \theta)$ concentrates for a given x at θ_x and if $p(\theta)$ has an upper bound and is continuous and different from zero at $\theta = \theta_x$ (or different from zero in some neighborhood of θ_x), the inferred chance $Q_n(H)$ approaches with increasing n the value $Q_n'(H)$ which holds for p = constant.

[1] Instead of $\bar{B}$ we have sometimes written B' (in particular in Chapter II). Here we write $\bar{B}$ since we use B' in a different sense. The present notation is as in Chapter VII. The prior probability $p(\theta)$ has been denoted by $p_0(\theta)$ in Ch. VII.

This is a generalization of the theorem of Chapter VII, p. 340. The content of the present theorem may be described, in a crude way, as follows. For large n, the right-hand side of (3) is practically zero, except in the immediate neighborhood of θ_x. Hence, it only matters how $p(\theta)$ behaves around $\theta = \theta_x$; in this small neighborhood, then, we may set $p(\theta) =$ constant, which amounts to replacing Q_n by Q_n'.

Our limit result $Q_n(H)/Q_n'(H) \to 1$, as $n \to \infty$, may also be interpreted from a slightly different point of view. Let H be the interval $(\theta_x - \eta, \theta_x + \eta)$ and choose η so that for a given n, $Q_n(H)$ has a certain value, for example, 50 %. Write briefly $Q_n(H) = R(\eta)$, and, similarly, $Q_n'(H) = R'(\eta)$. There is then an η' such that, for the same n and for $H' = (\theta_x - \eta', \ \theta_x + \eta')$ now $Q_n'(H') = P'(\eta') = 50\ \%$. Therefore, $Q_n(H) = R'(\eta')$, $Q_n'(H) = R'(\eta)$, and the quotient

$$\frac{R'(\eta)}{R'(\eta')}$$

takes the place of the middle term in (7). [Before, the middle term in (7) was the quotient of two different functions for the same interval H; now numerator and denominator contain the same function but the intervals η and η' are different]. We now again let n increase, keeping the value of $50\ \% = R'(\eta')$ fixed; then both η and η' will decrease and our limit result states that $R'(\eta)/R'(\eta') \to 1$. But $R'(\eta)$ is monotonic and it follows, therefore, that $\eta/\eta' \to 1$; we may say: *under the conditions of the preceding theorem the* 50 % *limits of the parameter* θ *are asymptotically the same for a non-constant a priori probability as for a constant one*, and these limits contract toward zero as $n \to \infty$.

1.3. *Using past experience.* The inequalities (7) may be of use in many cases. But, to be sure, in general, they are not the basis upon which practical estimation judgment rests. The following is an example of systematic *use of past experience.*

The fitness of a water supply is usually tested in the following way.[2] A sample of 10 cc is tested for bacteria, the reaction being positive if one or more bacteria are present. Five such samples are taken each time: $n = 5$, and $x\ (= 0, 1, \ldots, 5)$ is the number of positive tests. Denote by θ the probability of a positive test and by λ the average number of bacteria in 10 cc. Then by Poisson's law $(\lambda^z/z!)e^{-\lambda}$ is the probability of z bacteria in the sample; $e^{-\lambda}$ is the probability of no bacteria and $\theta = 1 - e^{-\lambda}$ that of at least one in 10 cc. The agreement is that $\lambda = 1$

[2] R. v. MISES, "On the correct use of Bayes' formula." *Ann. Math. Statist.* 13 (1942), pp. 156-163.

is still admissible. Hence $\theta_1 = 1 - e^{-1} = 0.63$ is the largest acceptable θ-value for this type of water. To $\theta < \theta_I$ corresponds $\lambda < \lambda_I = 1$, that means "cleaner" water.

Here

$$p(x \mid \theta) = \binom{5}{x}\theta^x(1 - \theta)^{5-x} \tag{1}$$

is the probability of x positive tests out of 5 and, with $P(\theta)$ the prior d.f.

$$Q_{n,x}(\theta_I) = \frac{\int_0^{\theta_I} p(x \mid \theta)\, dP(\theta)}{\int_0^1 p(x \mid \theta)\, dP(\theta)} \tag{8}$$

is the probability of $\theta < \theta_I$ if x tests are positive. We are mainly interested in the case $x = 0$, i.e., all tests negative. $Q_{n,0}(\theta_I)$ is the chance which we should like to compute or to estimate.

The experimenter feels fairly certain that $Q_{n,0}(\theta_I)$ is nearly one, i.e., that *the inference from* $x = 0$ (out of $n = 5$ trials) *that* $\theta < \theta_I$ *has a very large probability*. What is the source of this confidence? $n = 5$ is not a large number; the inequalities obtainable from (7) are not strong enough. Actually, the source of his confidence lies in his knowledge of *past experience*. In fact, tests, each consisting of five trials, have been performed very often in the past. They showed that the water subject to these tests was rather clean water and we expect that a new sample will be of the same type.

We know N past results $X_1, X_2, \ldots, X_N$, $X_i = 0, 1, \ldots, 5$; to each X_i belongs a θ_i, $\mid i = 1, 2, \ldots, N$. If we knew these past θ_i we could, of course, construct an empirical $P(\theta)$, but we do not know the θ_i.

From the $X_1, \ldots, X_N$ we deduce their frequency distribution N_x/N, the frequency of those past tests which gave the results x ($x = 0, 1, \ldots, 5$). Clearly N_x/N is for large N an approximation to the chance (probability)

$$r(x) = \lim_{n\to\infty} \frac{N_x}{N} \tag{9}$$

of x positive tests out of n, where

$$r(x) = \int_0^1 p(x \mid \theta)\, dP(\theta) = \binom{n}{x} \int_0^1 \theta^x(1 - \theta)^{n-x}\, dP(\theta). \tag{10}$$

For a Bernoulli distribution we have

$$\sum_{x=0}^{n} xp(x \mid \theta) = n\theta, \quad \sum_{x=0}^{n} x(x - 1)p(x \mid \theta) = n(n - 1)\theta^2;$$

therefore

$$\frac{1}{n}\sum_{x=0}^{n} x\binom{n}{x}\theta^x(1-\theta)^{n-x} = \theta, \qquad \frac{1}{n(n-1)}\sum_{x=0}^{n} x(x-1)\binom{n}{x}\theta^x(1-\theta)^{n-x} = \theta^2 \quad (11)$$

and using (10) we obtain

$$M_1 = \int_0^1 \theta\, dP(\theta) = \frac{1}{n}\sum_{x=0}^{n} xr(x)$$
$$M_2 = \int_0^1 \theta^2\, dP(\theta) = \frac{1}{n(n-1)}\sum_{x=0}^{n} x(x-1)r(x), \quad (12)$$

where we may introduce for the $r(x)$ the observed ratios N_x/N, since N is large. We have thus found the first and the second moment of the unknown distribution $P(\theta)$. In the same way we could find the 3rd, 4th, and 5th moments, if desired.

In an earlier paper[3] (which follows up a problem of A. Wald) v. Mises has shown how lower and upper bounds of a distribution function $P(\theta)$ can be found if the expected values with respect to $P(\theta)$ of two functions $f(\theta)$ and $g(\theta)$ are known, where the only condition imposed is that the curve $x = f(\theta)$, $y = g(\theta)$ be convex. In the present case we put $f(0) = g(0) = 0$ for $\theta < 0$; $f(\theta) = \theta$, $g(\theta) = \theta^2$ for $0 \leqslant \theta \leqslant 1$; $f(\theta) = g(\theta) = 1$ for $\theta \geqslant 1$. Then the results of this paper can be applied to our $P(\theta)$ and we find inequalities for the *a priori* $P(\theta)$ and from them (see the paper quoted on p. 498 for details) inequalities for the posterior probability $Q_{nx}(\theta_I)$, of Eq. (8), the probability of $\theta \leqslant \theta_I$ if x out of $n = 5$ tests are positive.

Numerical results are as follows: the "past experience" communicated to v. Mises consisted of $N = 3420$ tests (each composed of $n = 5$ trials). The results were $N_0 = 3086$, $N_1 = 279$, $N_2 = 32$, $N_3 = 15$, $N_4 = 5$, $N_5 = 3$. The overwhelming majority of tests with the result $x = 0$ is obvious. With $r(x) \sim N_x/N$ we obtain from (12)

$$M_1 = 0.02474, \qquad M_2 = 0.00401,$$

and applying the above-indicated method we find

$$Q_{n0}(\theta_I) \geqslant 0.99915,$$

i.e., the following result. *If we assume that in continuing the experiments, the distribution N_x/N of test results will be about the same as it has been in*

[3] R. v. Mises, "The limits of a distribution function if two expected values are given." *Ann. Math. Statist.* **10** (1939), pp. 99-104.

the past 3420 *cases, we have a chance of more than* 99.9 % *of being right when we state after each test which gives* $x = 0$ *that the density of the bacteria is less than* 1 per 10 cc.

In the same way we find

$$Q_{n1}(\theta_1) \geqslant 0.92.$$

1.4. *Using past experience. Another approach.* A different approach has been proposed by Robbins.[4] He considers θ as a random variable and approaches the problem as one of point estimation (Sections 6 and 7 of the present chapter deal with estimation). Denoting by $\phi(x)$ any "estimator" of the unknown θ, he proposes to choose $\phi(x)$ so that the mean square error $E[(\phi(x) - \theta)^2]$ becomes a minimum. This leads to the "*Bayes estimator*"[5]

$$\hat{\theta} = \phi(x) = \frac{\int p(x \mid \theta)\theta \, dP(\theta)}{\int p(x \mid \theta) \, dP(\theta)}, \tag{13}$$

the expected value of θ with respect to the posterior distribution.

We shall apply Robbins' approach to the problem of Section 3.1. However, before doing this, we consider another case where his approach works particularly well. Let, as an example,

$$p(x \mid \theta) = \frac{\theta^x}{x!} e^{-\theta} \qquad x = 0, 1, \ldots, \theta > 0.$$

We have then, exactly as in (10),

$$r(x) = \frac{1}{x!} \int_0^\infty e^{-\theta}\theta^x \, dP(\theta)$$

and obtain the estimate

$$\hat{\theta} = \frac{\int_0^\infty e^{-\theta}\theta^{x+1} \, dP(\theta)}{\int_0^\infty e^{-\theta}\theta^x \, dP(\theta)} = (x + 1)\frac{r(x + 1)}{r(x)}.$$

Replacement of $r(x)$ by N_x/N leads to an empirical estimate $\bar{\theta} = (x + 1)N_{x+1}/N_x$, which tends to $\hat{\theta}$ as $N \to \infty$.

Now turn to our example where $p(x \mid \theta)$ is a binomial distribution. We obtain from (13)

$$\hat{\theta}_n = \hat{\theta} = \frac{\int_0^1 \theta^{x+1}(1 - \theta)^{n-x} \, dP(\theta)}{\int_0^1 \theta^x(1 - \theta)^{n-x} \, dP(\theta)}, \tag{14}$$

which by means of (10), and writing now $P_{n,x}$ for $r(x)$ leads to

$$\hat{\theta} = \frac{x + 1}{n + 1}\frac{P_{n+1,x+1}}{P_{n,x}}, \qquad x = 0, 1, \ldots, n. \tag{15}$$

[4] H. Robbins, "An empirical Bayes approach to statistics." *Proc. 3rd Berkeley Symp.* 1955, pp. 157-163.

[5] It need hardly be said that the notation $\hat{\theta}$, $\hat{\lambda}$, etc., appearing here does not imply any relation to the notations used in Chapter IX, Section 5; there just are not enough signs!

Since a test consists of n single trials we do not know $P_{n+1,x+1}$ which would be based on $n + 1$ single trials. However, Robbins shows that the above $\hat{\theta}$ may be replaced by

$$\hat{\theta} = \frac{x+1}{n}\frac{P_{n,x+1}}{P_{n-1,x}}. \tag{15'}$$

To evaluate this we need probabilities $P_{n-1,x} = P_x'$ which relate to groups of $(n - 1)$ trials and it is proposed by Robbins to use the first $(n - 1)$ trials of each test. If (as in the present example) the data do not provide us with results relating to the first $n - 1$ trials of each test we may simply proceed as follows[6]:

(a) We have obviously the right to assume that any group of $n - 1$ out of n trials plays the same role as the group consisting of the first $n - 1$ trials (if this were not so the first $n - 1$ trials could not be considered as representative). It is then easy to derive combinatorially frequencies N_x', $x = 0, 1, \ldots, 4$ from the N_x, $x = 0, 1, \ldots, 5$, which may be used in (15'). For example, among the $N_1 = 279$ tests which have each given $x = 1$ positive trial out of five, there are on the average $\frac{1}{5} \cdot 279 = 55.8$ which have zeros in the first four places and 1 at the fifth place, so that with $n - 1 = \nu = 4$, these tests would be counted among the N_0'. In this way we obtain for $\nu = 4$:

$$\begin{aligned}
x = 0 \quad & N_0' = 3086 + 55.8 = 3141.8 \quad (= 3142)\\
x = 1 \quad & N_1' = 279 - 55.8 + 12.8 = 236\\
x = 2 \quad & N_2' = 32 - 12.8 + 9 = 28.2 \quad (= 28)\\
x = 3 \quad & N_3' = 15 - 9 + 4 = 10\\
x = 4 \quad & N_4' = 5 - 4 + 3 = 4.
\end{aligned}$$

With these values we obtain[7] from (15') if the P are replaced by the respective N

$$\text{for } x = 0: \quad \tilde{\theta} = \frac{1}{5}\frac{279}{3142} = 0.018,$$

$$\text{for } x = 1: \quad \tilde{\theta} = \frac{2}{5}\frac{32}{236} = 0.054,$$

and $\tilde{\theta} = 0.32$ for $x = 2$. We shall presently comment on these results. First consider the following:

(b) Instead of computing as in (13) the expected value of θ with regard to the posterior distribution we compute the expected value of some function $f(\theta)$, replacing thus the numerator of the right-hand side of (13) by $\int f(\theta)p(x \mid \theta)\, dP(\theta)$. In our case an appropriate $f(\theta)$ is clearly $\vartheta = \theta/(1 - \theta)$ [from which $\theta = \vartheta/(1 + \vartheta)$]. We obtain thus

$$\vartheta = \frac{\int_0^1 [\theta/(1-\theta)]\theta^x(1-\theta)^{n-x}\, dP(\theta)}{\int_0^1 \theta^x(1-\theta)^{n-x}\, dP(\theta)} = \frac{x+1}{n-x}\frac{P_{n,x+1}}{P_{n,x}}, \tag{16}$$

approximated by

$$\tilde{\vartheta} = \frac{x+1}{n-x}\frac{N_{x+1}}{N_x}, \qquad \tilde{\theta} = \frac{(x+1)N_{x+1}}{(n-x)N_x + (x+1)N_{x+1}}. \tag{16'}$$

[6] The following two pages which were needed to obtain numerical results are not in Robbins' paper.

[7] These formulas show that in Robbin's estimate of θ for a given x, only neighboring N-values are used; the total N does not enter into the formulas and for $x \geqslant 1$ the characteristically large value N_0 is not used.

The two procedures (a) *and* (b) *lead to the same result.* The proof is left to the reader. Procedure (b) may prove useful for other $p(x \mid \theta)$ as well.

It seems to us that in using this or other methods of point estimation, caution is necessary. The result $\theta = 0.02$ for $x = 0$ may be the "best" available point estimate (and there are so many "best" estimates!) but the question still remains whether there is a large probability for θ to be between 0.01 and 0.03, say. (One could try to estimate this by Mises' formulas.)

The modest precise statement of p. 500 that if $x = 0$, then, with probability very close to one, $\theta < 0.63$ (hence $\lambda < 1$, acceptable water) is safe and not over-optimistic.

Finally, we mention a more conventional procedure of practical value applied to this problem by H. A. Thomas, Jr.[8] He picks a parametric family of prior distributions and estimates the parameters by the method of moments. He takes for P an incomplete β function

$$dP(\theta) = K\theta^{p-1}(1 - \theta)^{q-1}\, d\theta.$$

The first two moments M_1, M_2 are computed by (12) and one finds then

$$p = \frac{M_1(M_1 - M_2)}{M_2 - M_1^2}, \qquad q = \frac{(1 - M_1)(M_1 - M_2)}{M_2 - M_1^2}.$$

Use of the values of p. 500 for M_1 and M_2 gives $p = 0.151$ and $q = 5.91$. Thomas' estimator for θ is the same as in (13)

$$\hat{\theta} = \int_0^1 \theta p(x \mid \theta)\, dP(\theta) \Big/ \int_0^1 p(x \mid \theta)\, dP(\theta).$$

This gives with his chosen prior probability:

$$\hat{\theta} = \frac{x + p}{n + p + q}, \qquad \lambda = \log[1/(1 - \theta)].$$

With the above values of p and q, values of θ and λ are obtained[9], which are in good accordance with direct counts of bacteria found by Thomas.

Problem 1. The data x_1, x_2, ..., x_n are the outcome of n trials on one and the same collective. It is known that the probability of obtaining any x_ν-value has a uniform distribution over the interval $(\theta - 1)$ to $(\theta + 1)$. One wants to infer from the arithmetic mean of the observations $\bar{x} = (x_1 + x_2 + \ldots + x_n)/n$ the chance of the value of θ. Show that for sufficiently large n the chance of θ falling in an interval $(\bar{x} - X)$ to $(\bar{x} + X)$ approaches $2G(\sqrt{3n}X)$ [see Eq. (53), Chapter III].

Problem 2. The probability of obtaining an average x derived from individual observations, is known to have the density

$$p_n(x \mid \theta) = \frac{h_n}{\sqrt{\pi}}\, e^{-h_n^2(x-\theta)^2}, \qquad \text{where} \quad h_n^2 = \frac{1}{2\sigma_n^2}.$$

[8] He suggested the water-supply problem to v. Mises.

[9] He obtains for $x = 0, 1, 2, 3$: $\theta = 0.014$; 0.104, 0.194; 0.284. $\lambda = 0.014$; 0.110; 0.215; 0.334, values fairly different from those following from Robbins' method.

Find, for large n, the chance inferred from x that θ falls in the interval θ_1 to θ_2.

Problem 3. If out of five trials with probability θ for success a single one was positive (i.e., $x = 1$ in Eq. (1) of p. 499) what can be concluded about the chance that $\theta < 0.63$, supposing that the maximum of the prior probability for $\theta > 0.63$ is not more than 3 times the minimum for $\theta < 0.63$?

Problem 4. A variate is distributed according to a normal law with variance $1/2$ and the unknown mean θ. Five observations have given an average $x = 0.12$. What is the chance of the inference that θ lies in the limits -0.20 to 0.20 if the over-all chance of θ is (a) assumed to be uniformly distributed over a large interval, (b) assumed to follow a normal law with mean value 0 and variance 1, (c) if it is only known that the minimum of $p(\theta)$ within the interval -0.20, 0.20 is at least one-fifth of its maximum outside this interval?

Problem 5. The variable x is known to be distributed according to a normal law with mean value zero and unknown variance:

$$p(x \mid \theta) = \frac{\theta}{\sqrt{\pi}} e^{-\theta^2 x^2}.$$

If $x = 5$ has been observed, what is the chance that θ falls in the limits 0.1 to 0.2? Make one of the following assumptions about $p(\theta)$: (a) the distribution is uniform over some large part of the positive half axis, (b) $p(\theta)$ is proportional to e^{-100}, (c) the minimum of the *a priori* probability within the interval 0.1, 0.2 is at least $1/6$ the maximum outside.

2. *Testing Hypotheses. Introduction of Neyman-Pearson Method*

2.1. *The problem.* The problem of inference discussed in the foregoing section can be restated in the following terms. The probability of the chance variable x, depending on θ, is given by a function $p(x \mid \theta)$. We omit the subscript n although x is, in general, the outcome of a group of n experiments (as in the case of the water supply example where $n = 5$), viz., $x = x(x_1, x_2, \ldots, x_n)$. Also, $p(x \mid \theta)$ may stand for $p(x_1, x_2, \ldots, x_n \mid \theta)$, where $x_1, \ldots, x_n$ are the results of $n \geqslant 1$ observations of the same random variable. There may be more than one unknown parameter, but now we assume just one. We consider an infinite sequence of trials (or an infinite sequence of groups of n trials) where each time an x-value is observed and θ unknown. The total range (continuous or discrete) of values of θ is called Ω. Let H be some interval on the θ-axis

and ξ a specified value of x. Each time ξ has been observed we make the assertion that θ lies in H. The problem is to find the chance of being right, that is, the limiting frequency of cases in which θ actually falls in H among all those cases in which the observed x coincides with ξ. This problem of inference was answered by formula (3) (and others) for $Q_n(H)$.

A certain modification of the inference problem is known as *the problem of testing hypotheses*. Again a function $p(x \mid \theta)$ or a c.d.f. $P(x \mid \theta)$ and some region H of θ-values are given and, again, an infinite sequence of trials in which x is observed will be considered. An assertion about θ falling in H is made after each trial (or after each sample of n trials, as the case may be) to wit: if the observed x falls in a specified region A of x-values, we assert that θ lies in H, and if x falls in the complementary $\bar{A}$ (= non-A), we assert that θ lies outside H (i.e., in $\bar{H}$ or non-H). The problem is again to find the chance of being right, that is, the limiting frequency of cases where either $x \subset A$ and $\theta \subset H$ or $x \subset \bar{A}$ and $\theta \subset \bar{H}$ holds true; in particular, we shall ask how the division of the x-range into A, $\bar{A}$ has to be made in order to have as high a success chance as possible for given H, $\bar{H}$ where $H + \bar{H} = \Omega$.

The region A in the n-space R_n, $n \geqslant 1$ of x-values is called the *region of acceptance* and $\bar{A}$ the *region of rejection* or *critical region*. The choice of A, $\bar{A}$, with $A + \bar{A} =$ total range of x, defines a test. We refer to H as the *hypothesis to be tested*, briefly the *null hypothesis*, and $\bar{H}$ is the *alternate hypothesis*, H can lie anywhere in the range Ω of θ-values. We do not exclude the possibility that we test H against an H_1 such that $H + H_1$ form only a part of Ω.

The over-all or *a priori* chance of θ will be described by its c.d.f. $P(\theta)$. The fact that $P(\theta)$ is, in general, unknown causes the main difficulty of the problem. As in Chapter VII, we do not require randomness for the variable θ. It is sufficient to assume that for any region which we consider, the frequency ratio of those cases for which θ falls into the region has a definite limit. On the other hand, if we want at all to make an inference on the value of θ, i.e., an assertion about the chance, given x, of θ having a certain value or falling into a certain interval, we *have* to assume that, in the long run, different θ-values may occur with certain limit frequencies, which amounts to introducing a $P(\theta)$.

As $p(x \mid \theta)$ is given, we suppose the probabilities of x falling in A or $\bar{A}$ to be known for any A:

$$P(A \mid \theta) = \int_{(A)} p(x \mid \theta)\, dx, \qquad P(\bar{A} \mid \theta) = 1 - P(A \mid \theta) = \int_{(\bar{A})} p(x \mid \theta)\, dx. \tag{17}$$

(As has been said before A is, in general, a subset of the n-space R_n for $n \geqslant 1$.)

2.2. *Errors of first and second kind.* Using the notation (17), we can easily express the chances for making a right or wrong decision. (See Fig. 36.) The chance for the coincidence of the two facts that x lies in

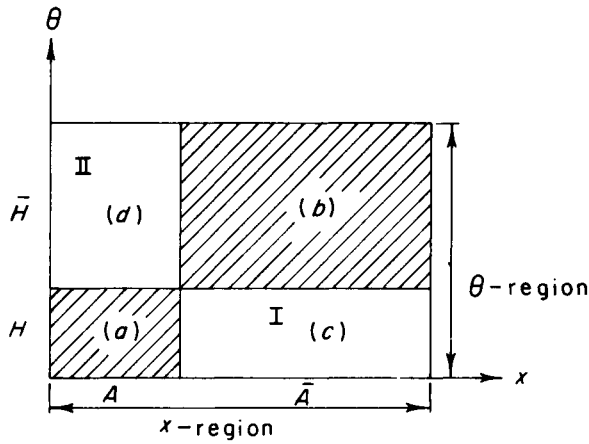

FIG. 36. Errors of first and second kind.

A and θ in H is

$$\int_{(H)} P(A \mid \theta)\, dP(\theta), \qquad x \in A, \quad \theta \in H \tag{a}$$

designated by (a) in Fig. 36. The other three combinations are

$$\int_{(\bar{H})} P(\bar{A} \mid \theta)\, dP(\theta), \qquad x \in \bar{A}, \quad \theta \in \bar{H} \tag{b}$$

$$\int_{(H)} P(\bar{A} \mid \theta)\, dP(\theta), \qquad x \in \bar{A}, \quad \theta \in H \tag{c}$$

$$\int_{(\bar{H})} P(A \mid \theta)\, dP(\theta), \qquad x \in A, \quad \theta \in \bar{H}. \tag{d}$$

The sum of (a) and (b) is the *success chance* P_S, the sum of (c) and (d) the *error chance* P_E. Obviously, $P_S + P_E = 1$. In frequency terms: in the infinite sequence of trials there will be among the first N trials N_0 trials for which the pronounced assertion "θ lies in H" or "θ lies in $\bar{H}$" proves to be correct. The limit of N_0/N is the success chance P_S and that of $1 - N_0/N$ the error chance P_E. It is usual to say that an *error of first kind is committed when the hypothesis H is rejected while it is correct,*

and an *error of second kind when the hypothesis is accepted although it is incorrect.* The expressions (c) and (d), respectively, are the chances of committing an error of first and second kind.[1] One may write

$$P_I = \int_{(H)} P(\bar{A} \mid \theta)\, dP(\theta), \qquad P_{II} = \int_{(\bar{H})} P(A \mid \theta)\, dP(\theta),$$
$$P_E = P_I + P_{II}, \qquad P_S = 1 - P_E. \tag{18}$$

2.3. *Simple hypothesis. Most powerful test.* We shall return to these general formulas in Section 3 and take up, in the present section, a special case: A hypothesis H is called a *simple hypothesis* if *the region H consists of one single point θ_0* only (see Fig. 37).

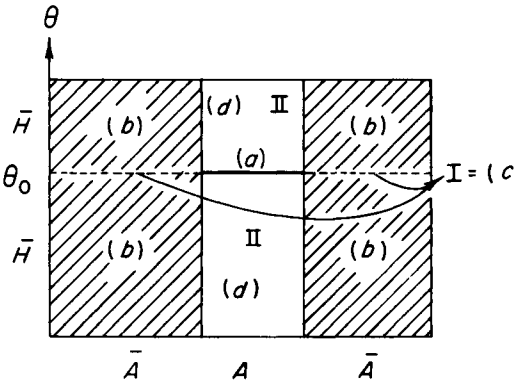

FIG. 37. Errors of first and second kind for simple hypothesis.

If $x_1, x_2, \ldots, x_n$ are successive measurements of a length and we make the hypothesis that they are independently distributed in the same normal distribution with zero mean, $\alpha = 0$, and unit variance, then H

[1] These ideas and the main publications on the subject are by J. Neyman and by E. S. Pearson. Of the numerous papers on the subject we quote:

J. NEYMAN and E. S. PEARSON, "On the problem of the most efficient tests of statistical hypotheses." *Phil. Trans. Roy. Soc. London* **A231** (1932), pp. 289-337.

J. NEYMAN, "Sur la vérification des hypothèses statistiques composées." *Bull. Soc. Math. France* **63** (1935), 246-266.

J. NEYMAN, "Outline of a theory of statistical estimation based on the classical theory of probability." *Phil. Trans. Roy. Soc. London* **A236** (1937), 333-380.

A. WALD, "Contributions to the theory of statistical estimation and testing hypotheses." *Ann. Math. Statist.* **10** (1939), 299-366.

A. WALD, *On the Principles of Statistical Inference. Notre Dame Lectures 1.* 1942.

R. v. MISES, "On the problem of testing hypotheses." *Ann. Math. Statist.* **14** (1943), 238-252.

is a simple hypothesis. A hypothesis which is not simple is called *composite*. Thus, a simple hypothesis uniquely determines the values of all parameters involved, while a hypothesis consistent with more than one value for some parameter is composite. For example the hypothesis $\theta = 2$ is simple. The hypothesis $3 \leqslant \theta \leqslant 5$ is composite or if there are two parameters, θ_1 and θ_2, the hypothesis $\theta_1 = \theta_2$ is composite. A particularly simple instance of a simple hypothesis appears if there is only one unknown parameter θ and Ω consists only of two values θ_0 and θ_1 such that $H = H_0$, or $\theta = \theta_0$, and $\bar{H}$, or $\theta = \theta_1$ are both simple. Otherwise, in the case of a simple hypothesis, $\bar{H}$ comprises the total range of θ-values except the value $\theta = \theta_0$. We have, in this case, to assume that the equality $\theta = \theta_0$ has a finite *a priori* chance, say, π_0. Should we know that $\pi_0 = 0$, the hypothesis $\theta = \theta_0$ would have no chance at all to hold true, and we could achieve the success chance 1 by consistently rejecting the hypothesis.

The region $\bar{A}$ is determined, in general, in such a way that it has a small chance α if $H = H_0$ is correct. We put

$$\alpha = P(\bar{A} \mid \theta_0) = 1 - P(A \mid \theta_0) \tag{19}$$

($\alpha = 5\,\%$ for example, or $\alpha = 1\,\%$; we shall see, however, that there are other grounds for making a choice of α.) The chance of committing a first-type error is then, with an unknown *a priori* chance π_0

$$P_I = \pi_0 P(\bar{A} \mid \theta_0) = \alpha\pi_0 . \tag{19'}$$

With $\bar{H}$ the set of all θ-values except θ_0, the probability of an error of the second kind is

$$P_{II} = \int_{(\bar{H})} P(A \mid \theta)\, dP(\theta) \tag{19''}$$

and the total error chance

$$\begin{aligned} P_E &= \alpha\pi_0 + \int_{(\bar{H})} P(A \mid \theta)\, dP(\theta) \\ &= \alpha\pi_0 + \int_{(\bar{H})} [1 - P(\bar{A} \mid \theta)]\, dP(\theta), \\ P_S &= 1 - P_E . \end{aligned} \tag{20}$$

From (19') and $\pi_0 \leqslant 1$, it follows that α *is the greatest possible chance of committing an error of first kind.* This α is usually called the *level of significance* of the test, often also the *size* of the critical region $\bar{A}$. The

choice of α is somewhat arbitrary, dependent on what probability of an error of first kind will be tolerable. [Often, the rejection of a correct hypothesis (error of first kind) seems more dangerous since it involves some finality, while to "accept" a hypothesis means only that the hypothesis does not contradict the observed material, viz., the results of the available observations. However, sometimes the opposite course is chosen; if the null hypothesis is accepted, the investigation is not pursued further.] In the following some suggestions for the choice of α will be given.

In all the work following Neyman's and Pearson's line of thought, the first and second type error chances are discussed, but the dependence on the over-all chance $P(\theta)$ is rarely considered. It is incorrect to say that α is the chance of a first-type error: $\alpha\pi_0$ is this chance; and it is inaccurate to say that the chance of a second-type error depends on θ: it depends on the distribution of θ. Also the quantity β which we shall introduce presently is not the probability of an error of second kind but the least upper bound of $P(A \mid \theta)$ for all θ in $\bar{H}$. Many authors (following R. A. Fisher) wish to avoid the use of the prior distribution. The fact that"it is not known" does not change the situation and omitting it from the formulas does not eliminate it. One may try to arrive at conclusions which do not or do not essentially depend on $P(\theta)$; otherwise our statements have to reflect our (possibly incomplete) information regarding $P(\theta)$.

If we have chosen α, (the greatest possible chance of committing an error of the first kind, of rejecting H if it is right) we then obviously desire to find among all tests of level α one that renders the probability of rejecting H if it is false as large as possible. This leads to the following definition.

The quantity $P(\bar{A} \mid \theta)$ as a function of θ is called the *power function* of the test with $\bar{A}$ as region of rejection. The value of the power function for $\theta = \theta_0$ equals α. If now, for the same hypothesis, another pair of complementary regions A' and $\bar{A}'$ is chosen with the same value of α, that is, with

$$P(\bar{A}' \mid \theta_0) = \int_{(\bar{A}')} p(x \mid \theta_0)\, dx = \alpha = P(\bar{A} \mid \theta_0) \tag{21}$$

the difference of the two error chances follows from (20) and (21) as

$$P_E' - P_E = \int_{(\bar{H})} [P(A' \mid \theta) - P(A \mid \theta)]\, dP(\theta) = \int_{(\bar{H})} [P(\bar{A} \mid \theta) - P(\bar{A}' \mid \theta)]\, dP(\theta). \tag{22}$$

Thus, it is seen that $P_E' \leqslant P_E$ and, consequently, $P_S' \geqslant P_S$ if, *for all* θ *in* $\bar{H}$, *the power* $P(\bar{A}' \mid \theta)$ *is greater than or equal to* $P(\bar{A} \mid \theta)$. Our first result for simple hypotheses can be stated:

When, for the same simple hypothesis, two different tests with A *and* A' *as regions of acceptance but with the same* α*-value are compared, the success chance of the second test will be greater than or equal to the success chance of the first, if the power of the second test, for all* θ*-values, is greater than or equal to the power of the first test.*

This statement is independent of the unknown *a priori* chance $P(\theta)$. It leads to the notion of *most powerful tests* as introduced by J. Neyman and E. S. Pearson. Let C be a class of regions A, all with the same power value α for $\theta = \theta_0$, that is, with the same maximum chance of first-type error. If within this class a region A^* exists such that

$$P(\bar{A}^* \mid \theta) \geqslant P(\bar{A} \mid \theta) \qquad \text{for all } \theta \text{ and all } A \text{ in } C, \tag{23}$$

then the test using A^* as region of acceptance is called *most powerful with respect to the class* C, also uniformly most powerful. A most powerful test has at least as high a success chance as any other test of the class C. The problem of the greatest success chance is, therefore, solved with respect to the class C if it is possible to find a most powerful test for the given α.

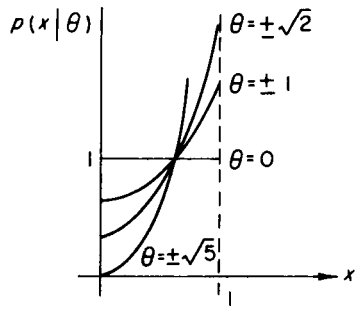

FIG. 38. A particular $p(x \mid \theta)$ for various θ-values.

Let us consider (Fig. 38) the density

$$p(x \mid \theta) = 1 + \theta^2(x^2 - \tfrac{1}{3}); \qquad 0 \leqslant x \leqslant 1, \quad -\sqrt{3} \leqslant \theta \leqslant \sqrt{3}. \tag{24}$$

(It is seen that this p is non-negative and has the integral 1 for any θ.)[2]

[2] Instead of θ^2 we may take in the density (24) any positive function of θ which is zero for $\theta = 0$.

The hypothesis to be tested might be $\theta = 0$. Let the class of all admitted regions of acceptance consist of all intervals a_1, a_2 for which (21) is fulfilled, that is, for the $p(x \mid \theta)$ of Eq. (24):

$$1 - \int_{a_1}^{a_2} p(x \mid \theta)\, dx = \alpha \qquad \text{or} \qquad a_2 - a_1 = 1 - \alpha.$$

Then the power function for the interval reaching from a_1 to $a_2 = a_1 + 1 - \alpha$, is

$$P(\bar{A} \mid \theta) = 1 - \int_{a_1}^{a_1+1-\alpha} p(x \mid \theta)\, dx = 1 - (1 - \alpha) - \theta^2 \int_{a_1}^{a_1+1-\alpha} (x^2 - \tfrac{1}{3})\, dx$$

$$= \alpha + \theta^2(1 - \alpha)\left[\frac{\alpha(2 - \alpha)}{3} - a_1^2 - a_1(1 - \alpha)\right].$$

If A^* is the interval beginning at $a_1 = 0$, the corresponding power function is

$$P(\bar{A}^* \mid \theta) = \alpha + \theta^2(1 - \alpha)\frac{\alpha(2 - \alpha)}{3}.$$

It is seen that here (23) is fulfilled. Thus, taking the interval $0 \leqslant x \leqslant 1 - \alpha$ as region of acceptance, we have a most powerful test. Figure 39

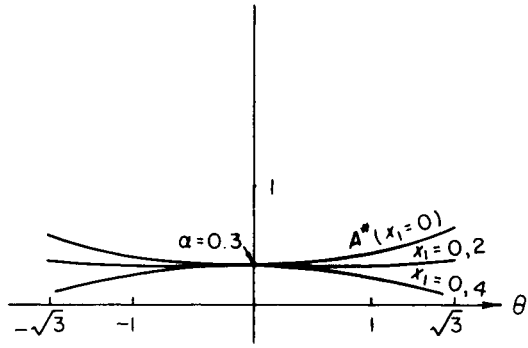

FIG. 39. Power curves for the $p(x \mid \theta)$ of Fig. 38.

shows the power functions for A^* with $\alpha = 0.3$ and the regions beginning at $a_1 = 0.1, 0.2, 0.3$. The corresponding error chance is

$$P_E^* = \alpha\pi_0 + \int_{(\bar{H})}\left[1 - \alpha - \theta^2(1 - \alpha)\frac{\alpha(2 - \alpha)}{3}\right] dP(\theta) \tag{24'}$$

$$= \alpha\pi_0 + (1 - \alpha)(1 - \pi_0) - (1 - \alpha)\frac{\alpha(2 - \alpha)}{3}\int_{(\bar{H})} \theta^2\, dP(\theta).$$

Consider the last integral; it is non-negative and can be arbitrarily near to zero for a discrete P with π_0 at θ_0 and $1 - \pi_0$ at some point $\theta \neq 0$, where $\theta^2 < \epsilon$. Thus, for $P_E{}^*$ the least upper bound (l.u.b.) for given α and π_0 is $\alpha\pi_0 + (1 - \alpha)(1 - \pi_0)$. On the other hand, if π_0 is considered variable, this is a linear function of π_0 which takes its extreme values at the ends of its interval, $\pi_0 = 0$ and $\pi_0 = 1$. Thus the larger of the two values α and $1 - \alpha$ is the l.u.b. of $P_E{}^*$. Accordingly, the g.l.b. of the success chance of the test under consideration is the smaller of the two quantities α and $1 - \alpha$. We shall return to this question in Section 2.5.

Let us assume, next:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}.$$

The hypothesis to be tested is $\theta = \theta_0$. From the results of n trials we compute $\bar{x}$ and use $u = \sqrt{n}(\bar{x} - \theta)$ as a new variable, which is normally distributed with mean value 0 and unit variance. A critical region $\bar{A}$ in R_n is defined by

$$u < u_1 \qquad \text{or} \qquad u > u_2\,, \tag{25}$$

where u_1 and u_2 will be characterized presently. The region of acceptance, A, is then

$$u_1 \leqslant u \leqslant u_2\,. \tag{25'}$$

In line with Eq. (19) the size α of $\bar{A}$ (or $1 - \alpha$ of A) is then introduced by

$$\Phi(u_2) - \Phi(u_1) = 1 - \alpha. \tag{26}$$

Let us determine the power function $P(\bar{A} \mid \theta)$ of this test:

$$P(\bar{A} \mid \theta) = 1 - P(A \mid \theta) = 1 - \Pr\{u_1 \leqslant \sqrt{n}(\bar{x} - \theta_0) \leqslant u_2\},$$

or, subtracting $\sqrt{n}(\theta - \theta_0)$ from each term in the brackets, we obtain

$$\begin{aligned} P(\bar{A} \mid \theta) &= 1 - \Pr\{u_1 - \sqrt{n}(\theta - \theta_0) \leqslant u \leqslant u_2 - \sqrt{n}(\theta - \theta_0)\} \\ &= 1 - [\Phi(u_2 - \sqrt{n}(\theta - \theta_0)) - \Phi(u_1 - \sqrt{n}(\theta - \theta_0))]. \end{aligned}$$

Each choice of u_1, u_2 which satisfies (26) defines a test of the given size α.

Since, by definition, $P(\bar{A} \mid \theta_0) = \alpha$ always, each power function passes through the point (θ_0, α), and all power curves intersect at

this point. Now consider (1) the limit case with $u_1 = -\infty$, $\Phi(u_2) = 1 - \alpha$ and (2) the limit case with $u_2 = +\infty$, $\Phi(u_1) = \alpha$. It is easily seen that all our power curves lie between these two curves. Figure 40 makes it clear

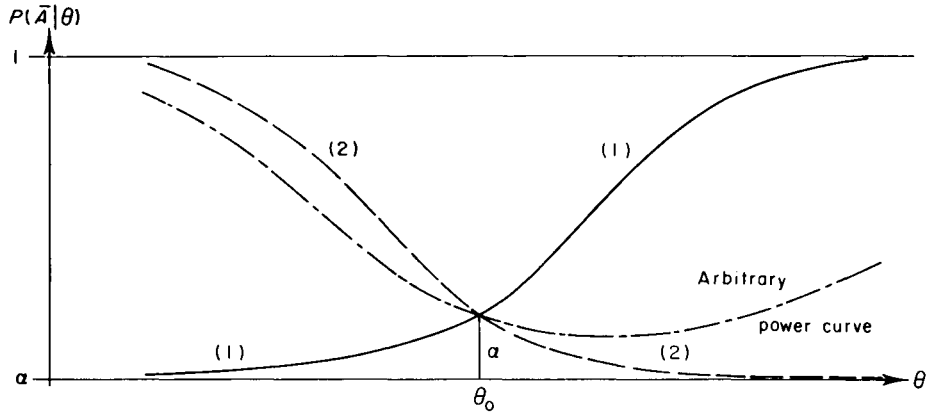

FIG. 40. Power curves for testing mean value for sample from normal population.

that (1) is most powerful for all $\theta > \theta_0$ while it is the least powerful of all tests for $\theta < \theta_0$; (2) is most powerful for $\theta < \theta_0$ and least powerful for $\theta > \theta_0$. Hence, no most powerful test exists for all θ.

2.4. *Testing a simple hypothesis H against a simple hypothesis* $\bar{H}$.[3] In this subsection we consider the simplest possible situation. Assume that only one unknown parameter θ is involved and that θ can take only two values θ_0 and θ_1. Consider x as a continuous variate: there are given two probability distributions $P(x \mid \theta_0) = P_0(x)$ and $P(x \mid \theta_1) = P_1(x)$ with continuous densities $p_0(x) = p(x \mid \theta_0)$, $p_1(x) = p(x \mid \theta_1)$; denote by $H = H_0$ the *null hypothesis* and by $\bar{H} = H_1$ the *alternate hypothesis*. We wish to test H_0 on the basis of a sample of n observations $x_1, x_2, \ldots, x_n$. The $P_0(x)$, $P_1(x)$ are n-dimensional distributions.

We choose α as the size of the critical region $\bar{A}$ and have therefore, from (19),

$$P(\bar{A} \mid \theta_0) = P_0(\bar{A}) = \int_{(\bar{A})} p_0(x)\, dx = \alpha. \tag{27}$$

Similarly, if we put, as an abbreviation

$$P_1(A) = \int_{(A)} p_1(x)\, dx = \beta, \tag{27'}$$

[3] J. NEYMAN and E. S. PEARSON, paper quoted in footnote 1, p. 507.

we have $\pi_0\alpha$ and $\pi_1\beta = (1 - \pi_0)\beta$ as the probabilities of errors of first and second kinds. Hence

$$P_1(\bar{A}) = \int_{(\bar{A})} p_1(x)\,dx = 1 - \int_{(A)} p_1(x)\,dx = 1 - \beta. \tag{27''}$$

The problem of determining a most powerful critical region $\bar{A}$ (in n-space) *is to choose* $\bar{A}$ *so that* $P_1(\bar{A}) = \int_{(\bar{A})} p_1(x)\,dx$ *is as large as possible under the condition that* $P_0(\bar{A}) = \int_{(\bar{A})} p_0(x)\,dx = \alpha$.

We shall show that a critical region which is most powerful can be chosen in a simple way.

If $p_0(x)$ is zero in some part of the sample space we add this part to the critical region without violating (27); the power (27'') can only be increased by this increase of the critical region. At all other points of the sample space $p_0(x) > 0$ and

$$r(x) = \frac{p_1(x)}{p_0(x)} \tag{28}$$

is a continuous function of x. Hence for any positive k the event $r(x) \leqslant k$ has a probability in the P_0-field, namely,

$$F_0(k) = \Pr\{r \leqslant k\} = \Pr\{p_1(x) \leqslant kp_0(x)\}. \tag{29}$$

We assume now first that the d.f. F_0 actually assumes the value $1 - \alpha$, i.e., that for some positive k

$$F_0(k) = \Pr\{r \leqslant k\} = 1 - \alpha, \tag{30}$$

and consequently

$$1 - F_0(k) = \Pr\{p_1(x) > kp_0(x)\} = \alpha. \tag{30'}$$

We state that *the region* $\bar{A}$ *defined by* $p_1/p_0 > k$ *has the desired maximum-power property and its size equals* α. Certainly, $P_0(\bar{A}) = \alpha$, by (30').

To prove the statement we compare $\bar{A}$ with another critical region $\bar{B}$ which satisfies the condition $P_0(\bar{B}) = \alpha$. We denote the intersection of $\bar{A}$ and $\bar{B}$ by E such that $\bar{A} = E + C$, $\bar{B} = E + D$; C is the part of $\bar{A}$ not contained in $\bar{B}$ and D is the part of $\bar{B}$ not contained in $\bar{A}$. In C we have everywhere $p_1/p_0 > k$ and in D we have $p_1/p_0 \leqslant k$. Also

$$P_0(\bar{B}) = P_0(\bar{A}), \qquad \text{or} \qquad P_0(D) = P_0(C).$$

Using these we have

$$\int_D p_1(x)\,dx \leqslant k \int_D p_0(x)\,dx = k \int_C p_0(x)\,dx \leqslant \int_C p_1(x)\,dx.$$

Hence

$$\int_{D+E} p_1(x)\,dx \leqslant \int_{C+E} p_1(x)\,dx,$$

or

$$\int_{\bar{A}} p_1(x)\,dx \geqslant \int_{\bar{B}} p_1(x)\,dx. \qquad \text{q.e.d.}$$

The test based on this choice of $\bar{A}$ is called the *likelihood ratio test*: *H_0 is rejected for samples $x = (x_1, x_2, \ldots, x_n)$, such that $p_1(x)/p_0(x) \geqslant k$, and accepted if $p_1(x)/p_0(x) \leqslant k$, where k is defined by the condition that $Pr\{p_1(x) > kp_0(x)\} = \alpha$.* If F_0 does not assume the value $1 - \alpha$ but has a jump there the proof is to be slightly modified.

If the sample space consists of denumerably many discrete points the proof is similar. First, if some points of the sample space have probability zero, given H_0, we add them to the critical region. At all remaining points, $p_0 > 0$. We form again $r(x) = p_1(x)/p_0(x)$. If the d.f. $F_0(k) = \Pr\{r \leqslant k\}$ takes on the value $1 - \alpha$, i.e., if $\Pr\{p_1 > kp_0\} = \alpha$, for some k, we proceed exactly as before. If no k can be found for which $F_0(k) = 1 - \alpha$ exactly, we take k such that $F_0(k) = 1 - \alpha + \epsilon$ and hence $\Pr\{r > k\} = \alpha - \epsilon$, with ϵ as small as possible for the given F_0. This $\bar{A}$ is a little smaller than the proposed one and the test is most powerful with respect to the $\bar{A}$ of "size" $\alpha - \epsilon$. If one wants the level α rather than $\alpha - \epsilon$ and if by addition of one more point the $\bar{A}$ becomes a little too large one would have to split the probability of this point into two parts, etc. We do not consider this in detail.[4]

Example. Consider the sample space R_n. Assume that H_0 states that each variate x_i, $i = 1, 2, \ldots, n$ is normally distributed $N(0, 1)$, and independence holds:

$$p_0(x) = (2\pi)^{-n/2} \exp[-\tfrac{1}{2}(x_1^2 + x_2^2 + \cdots + x_n^2)].$$

H_1 states that the x_i are normal and independent $N(\theta, 1)$ where $\theta > 0$. Then

$$p_1(x) = (2\pi)^{-n/2} \exp\{-\tfrac{1}{2}[(x_1 - \theta)^2 + \cdots + (x_n - \theta)^2]\}, \qquad \theta > 0,$$

and

$$r(x) = \frac{p_1(x)}{p_0(x)} = \exp[\theta(x_1 + \cdots + x_n) - \tfrac{1}{2}n\theta^2].$$

[4] An extension by A. Wald (1939, see footnote 1, p. 507) to the case of a general simple hypothesis (see Section 2.3) uses in a skillful way an *a priori* distribution. In this case, then our theorem of p. 510 applies. Our following example relates also to this more general case.

This $r(x)$ increases monotonically with $\Sigma x = n\bar{x}$, hence with $\bar{x}$. *The hypothesis H_0 is rejected if $\bar{x}$ is greater than a certain critical value c.* The probability that $\bar{x} > c$, given H_0, is to equal α. If H_0 holds, $\bar{x}$ is normally distributed $N(0, 1/\sqrt{n})$. Hence

$$c = \frac{1}{\sqrt{n}}\psi(1 - \alpha),$$

where ψ is the inverse function of Φ. If, for example, $\alpha = 0.05$, $n = 9$ then $c = \frac{1}{3}\psi(0.95) = \frac{1}{3}(1.64) = 0.547$. We note that this test: "reject H_0 if $\bar{x} > 0.547$" does not depend on the value of "θ." It is "uniformly most powerful" with respect to any H_1 with $\theta > 0$. Or, in other words, we are testing H_0: $\theta = \theta_0 = 0$ against $\bar{H}$: $\theta > 0$; this is a general simple hypothesis.

2.5. *Success rate.* We return to the general consideration (Section 2.3) of a simple hypothesis and denote by $1 - \beta$ *the greatest lower bound,* g.l.b. (the smallest value) *of the power function $P(\bar{A} \mid \theta)$ if all θ-values of $\bar{H}$ are considered.* Equivalently, β *is the least upper bound* l.u.b. (the greatest value) *on $\bar{H}$ of $P(A \mid \theta)$.* Then, since the integral of $dP(\theta)$ over the region H is $1 - \pi_0$, it follows from (18) that

$$P_E \leqslant \alpha\pi_0 + \beta(1 - \pi_0), \tag{31}$$

where

$$\alpha = P(\bar{A} \mid \theta_0), \qquad \beta = \underset{\theta \in \bar{H}}{\text{l.u.b.}}\, P(\bar{A} \mid \theta). \tag{32}$$

The right-hand side of (31) is not only an upper bound for P_E, but the least upper bound. In fact, the distribution $P(\theta)$ could be such that the total amount $1 - \pi_0$ is concentrated at the point of maximum $P(A \mid \theta)$ so as to make the integral in (20) *equal* to $\beta(1 - \pi_0)$.

The right-hand side of (31) depends on the unknown π_0. But as π_0 ranges from 0 to 1, the maximum of the linear expression $\alpha\pi_0 + \beta(1 - \pi_0)$ must be reached either at $\pi_0 = 0$ or at $\pi_0 = 1$, and is therefore equal either to β or α. *The least upper bound of the error chance P_E is the greater of the two quantities α and β.* From (31) it follows also that

$$P_S \geqslant \pi_0(1 - \alpha) + (1 - \pi_0)(1 - \beta), \tag{31'}$$

which for $\pi_0 = 0$ gives $1 - \beta$, for $\pi_0 = 1$ gives $1 - \alpha$. The *greatest lower bound of P_S is called the success rate S of the test. If we do not know anything about the prior probability, S is the smaller of the two quantities "one minus level of significance" and "minimum power of the test."* In other

terms: *Whatever test we use, we are sure to have a success chance P_S at least equal to the smaller of the two quantities* $1 - \alpha$ *and* $1 - \beta$. *No greater lower bound for P_S can be given* if no restriction is known to hold for $P(\theta)$.[5]

Since each power function takes the value α at $\theta = \theta_0$, the minimum power $1 - \beta$ cannot surpass α, if the power $P(\bar{A} \mid \theta)$ is a continuous function of the continuous variable θ:

$$1 - \beta \leqslant \alpha, \qquad \alpha + \beta \geqslant 1, \qquad (1 - \alpha) + (1 - \beta) \leqslant 1. \tag{33}$$

Thus the quantities $1 - \alpha$ and $1 - \beta$ cannot both be greater than $\frac{1}{2}$. It follows that the greatest lower bound of the success chance, the *success rate S, cannot be greater than* 0.50 *in the case of a continuous power function.* The value $S = \frac{1}{2}$ is only reached if both α and β equal $\frac{1}{2}$. *The situation changes decisively if we know something about* $P(\theta)$. If we know, e.g., that π_0 is great (or subjectively speaking: if one "firmly believes" that the simple hypothesis H_0 is true) $\pi_0 = 1 - \epsilon$ (ϵ small) we have from (31'), $P_S \geqslant 1 - \alpha - \epsilon(\beta - \alpha)$. Then, if α is very small S may be very close to one. If we know that π_0 is very small, all we have to do is to keep β low in order to have a high success rate. Also if θ is known to assume only discrete values, a higher success rate than $\frac{1}{2}$ is obtainable (see Section 3, p. 523).

In example (24) we have seen that the minimum power $1 - \beta$ for the most powerful test A^* coincides with α. The success rate was the smaller of the quantities α and $1 - \alpha$; in the case of Fig. 39 it is 30 %. If $P(\bar{A} \mid \theta)$ has a minimum at $\theta = \theta_0$, where $P(\bar{A} \mid \theta_0) = \alpha$, and if $P(\bar{A} \mid \theta)$ is continuous in θ we obtain $\alpha = 1 - \beta$ or $\beta = 1 - \alpha$, as in the example. Such tests for which $P(\bar{A} \mid \theta)$ has a minimum at $\theta = \theta_0$ have been denoted *unbiased* tests by J. Neyman.[6]

In order to obtain a test with the greatest possible success rate, which is $S = 0.5$ in the case of continuous θ, one has to use regions A, $\bar{A}$ such that the power function takes its minimum value at $\theta = \theta_0$ and that this minimum (which then equals α) has the magnitude 0.5. An example which justifies a century-old practice of statisticians is the following.

Assume that the distribution of x is completely known *except for its mean value*, or, which is the same, except for its location on the x-axis, and that we want to test the hypothesis that the mean (or any

[5] We might think of S as a "guaranteed" success rate, since the actual success rate might be much better.

[6] This means that $P(\bar{A} \mid \theta) \leqslant \alpha$ for $\theta \in H$ and $P(\bar{A} \mid \theta) > \alpha$ for $\theta \in \bar{H}$. In case of constant *a priori* distribution the type one error will then not exceed α while the probability to reject H if it is false will be greater than α.

other significant point of the distribution curve) has a certain location. In our formulas this means that $p(x \mid \theta)$ is a given function of the difference $x - \theta$ and that a hypothesis $\theta = \theta_0$ has to be tested. Without loss of generality, we can assume $\theta_0 = 0$. From

$$p(x \mid \theta) = f(x - \theta) \tag{34}$$

the power function for the interval of acceptance a_1 , a_2 is found:

$$P(\bar{A} \mid \theta) = 1 - \int_{a_1}^{a_2} p(x \mid \theta)\, dx = 1 - \int_{a_1}^{a_2} f(x - \theta)\, dx = 1 - \int_{a_1-\theta}^{a_2-\theta} f(z)\, dz \tag{35}$$

and its derivative is

$$\frac{dP(\bar{A} \mid \theta)}{d\theta} = f(a_2 - \theta) - f(a_1 - \theta). \tag{35'}$$

In order to place the power minimum at $\theta = \theta_0 = 0$, one must make (35') vanish at $\theta = 0$, that is,

$$f(a_1) = f(a_2) \tag{36}$$

and, in order to have $P(\bar{A} \mid \theta_0) = \alpha = \frac{1}{2}$, one must choose a_1 , a_2 such that

$$\int_{a_1}^{a_2} p(x \mid \theta_0)\, dx = \int_{a_1}^{a_2} f(x)\, dx = \tfrac{1}{2}\,. \tag{36'}$$

Conditions (36) and (36') determine the region of acceptance a_1 , a_2 in a unique way if $f(x)$ has one maximum and decreases monotonically on both sides (Fig. 41): The interval A is limited *by two equal ordinates*

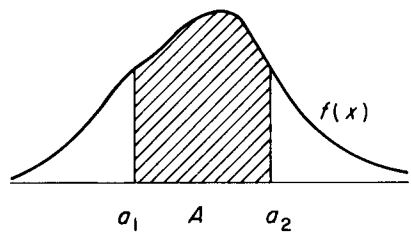

FIG. 41. Probable-limit test.

of the density curve $f(x)$ and includes one half of the total area. This region A is known as the "*probable limits*" of x. The result can be stated as follows: *If the location of an otherwise known distribution is to be*

tested, that is the hypothesis $\theta = \theta_0$ *in the case* $p(x \mid \theta) = f(x - \theta)$, *plot the curve* $f(x - \theta_0)$ *and determine the probable limits* a_1 *and* a_2 *of* x; *if the interval* a_1, a_2 *is taken as region of acceptance, the (guaranteed) success rate has its maximum value* $\frac{1}{2}$. In other terms, if the hypothesis is accepted each time x falls in a_1, a_2 and rejected in the opposite case, one will be sure to have at least 50 % correct decisions in the long run.

One might ask whether the test based on probable limits can be considered as a most powerful one. Since the derivative of the power curve for our A vanishes at $\theta = \theta_0$, all other power curves admitted should have the same property (at least). But vanishing derivative leads to $f(a_1) = f(a_2)$ according to (36). Therefore, the test cannot be most powerful with respect to the class of *all* intervals. Suppose that only such tests are admitted to the class C whose regions of rejection satisfy the condition

$$\frac{dP(\bar{A} \mid \theta)}{d\theta} = 0 \qquad \text{at} \qquad \theta = \theta_0, \tag{37}$$

i.e., "unbiased" test only. We may ask whether the probable limits test is most powerful with respect to all *these* tests or a subclass of them; this depends on the function $f(x)$. Thus, in using the probable limits test, we are not, in general, sure to have more success than with other tests, but we are sure to have at least 50 % right decisions.

In the light of our considerations we might ask whether to seek for a most powerful test is not in general too much and too little. Too much since such a test does not exist in most cases[7]; too little because another test, with a different α, may exist for the same problem whose success rate is considerably higher.

2.6. *Weighted error chance.* In many instances [assuming that we know nothing about $P(\theta)$] it might seem useful to use a test with a smaller than maximum possible success rate, for the following reason. The situation can be such that committing an error of first kind (being wrong in rejecting the hypothesis) is much less acceptable than an error of second kind (wrongly accepting the hypothesis) or vice versa. In this case, instead of minimizing the error chance $P_E = P_I + P_{II}$, one will rather try to minimize an expression like

$$P_W = \lambda P_I + \mu P_{II},$$

where λ, μ are positive factors, the ratio of which is determined by the relative weight attributed to first and second kind errors. According to (31) the least upper bound of the *weighted error chance* P_W, for given α and π_0, is

$$\lambda\alpha\pi_0 + \mu\beta(1 - \pi_0)$$

[7] See also statements to this effect by Cramér [4], p. 530; van der Waarden [28], p. 262; A. Wald, *On the Principles of Statistical Inference*, 1942, *Notre Dame Lectures*, I, p. 17; etc.

and with π_0 going over the whole range 0 to 1, we have the l.u.b. of P_W equal to the greater of the quantities $\lambda\alpha$ and $\mu\beta$. In general, it will be possible to find for each α a test the power of which takes its minimum value at $\theta = \theta_0$. We have seen that for these "unbiased" tests we have $\alpha + \beta = 1$ and the l.u.b. of P_W becomes the greater of the quantities $\lambda\alpha$ and $\mu(1 - \alpha)$. To minimize this l.u.b. for given λ, μ one has to have $\alpha : (1 - \alpha) = \mu : \lambda$ or

$$\alpha = \frac{\mu}{\lambda + \mu}, \qquad 1 - \alpha = \beta = \frac{\lambda}{\lambda + \mu}. \tag{38}$$

This may lead to choosing a level of significance α different from $\frac{1}{2}$. If, for one reason or another, we think that an error of first type is 9 times as important as a second-type error, we would choose $\alpha = 0.1$ in order to minimize the least upper bound of the weighted error chance. But the earlier result (established for continuous θ) stands: In a test on the level of significance $\alpha = 0.10$ (or $\alpha = 0.90$), whether or not it is a most powerful test with respect to any class of tests, we have to face the possibility of making 90% false decisions or even more in the long run.

Problem 6. Let θ_0 and θ_1 be the only possible values of θ.

(a) By what choice of A and $\bar{A}$ do we obtain $P_I = 0$, $P_{II} = 1$?
(b) In which case do we obtain $P_I = 1$, $P_{II} = 0$?

Problem 7. The variate x ranging from 0 to 1 has a probability density depending on θ:

$$p(x \mid \theta) = 1 - \theta(x - \tfrac{1}{2}),$$

where θ can take all values between 0 and 2. One wants to test the hypothesis $\theta = 0$ on the level of significance α. Find the power function $P(\bar{A} \mid \theta)$ and its graph if the region of acceptance is formed by intervals starting at a_1. Discuss the power functions for all possible values of a_1, and find the particular value of a_1 for which the test is most powerful with respect to the class of all intervals with the same α.

Problem 8. A variate is known to have a normal distribution with given variance and unknown mean value θ. If the hypothesis $\theta = 0$ is to be tested for a given α and any interval is admitted as the region of acceptance, prove that no most powerful test exists, but that the symmetric interval $(-a_1, a_1)$ gives the smallest possible value of the maximum second-type error chance, namely, $\beta = 1 - \alpha$.

Problem 9. The variate x is subject to a normal distribution with mean value zero and unknown variance:

$$p(x \mid \theta) = \frac{\theta}{\sqrt{\pi}} e^{-\theta^2 x^2}, \qquad \theta > 0.$$

To test the hypothesis $\theta = 1$ one will choose as the region of acceptance a pair of symmetric intervals (a_1 to a_2 and $-a_2$ to $-a_1$). Prove that, for given α, no most powerful test exists, but there is one definite test region for which the maximum possible second-type error chance β has its smallest value $1 - \alpha$. Find this region for $\alpha = 0.50$ and $\alpha = 0.10$.

3. Neyman-Pearson Method. Composite Hypothesis. Discontinuous and Multivariate Cases

3.1. *The problem.* We now take up the more general problem outlined at the beginning of Section 2. Given a function $p(x \mid \theta)$ and a region H of the θ-axis, the hypothesis $\theta \in H$ is to be tested. The region of acceptance will again be called A, the region of rejection $\bar{A}$, while $\bar{H}$ is the complementary set to H on the θ-axis (Fig. 36).

The chance of making a false decision is, as we saw, the sum of the expressions (c) and (d), of p. 506 referring to errors of first and second kind:

$$P_E = \int_{(H)} P(\bar{A} \mid \theta)\, dP(\theta) + \int_{(\bar{H})} P(A \mid \theta)\, dP(\theta). \tag{39}$$

We keep the definition of β as given above, but change those of π_0 and of α:

$$\pi_0 = \int_{(H)} dP(\theta), \qquad 1 - \pi_0 = \int_{(\bar{H})} dP(\theta) \tag{40}$$

$$\alpha = \underset{\theta \in H}{\text{l.u.b.}}\, P(\bar{A} \mid \theta), \qquad \beta = \underset{\theta \in H}{\text{l.u.b.}}\, P(A \mid \theta). \tag{41}$$

Thus, α is the maximum chance of committing a first-type error and β the same for a second-type error. From (39) follows

$$P_E \leqslant \pi_0\alpha + (1 - \pi_0)\beta$$
$$P_S \geqslant \pi_0(1 - \alpha) + (1 - \pi_0)(1 - \beta),$$

and since π_0 ranges from 0 to 1 and $P_S = 1 - P_E$, we have

$$P_E \leqslant \text{Max}\{\alpha, \beta\}, \qquad P_S \geqslant \text{Min}\{1 - \alpha,\ 1 - \beta\}; \tag{42}$$

i.e., the larger of the two quantities α and β is the l.u.b. of P_E, and the success rate S is the smaller of the two quantities $1 - \alpha$ and $1 - \beta$.

If θ is a continuous variable and $P(A \mid \theta)$ a continuous function of θ, which are assumptions not necessary for the validity of (40) and (41), then $1 - \beta$ cannot be greater than α. In fact, if θ_1 is a point on the boundary between H and $\bar{H}$, one must have $P(\bar{A} \mid \theta_1) \leqslant \alpha$ and $P(\bar{A} \mid \theta_1) \geqslant 1 - \beta$

and this implies $\alpha \geqq 1 - \beta$ or $\alpha + \beta \geqq 1$, $(1 - \alpha) + (1 - \beta) \leqq 1$ as in (33). Therefore, the success rate S, that is, the greatest lower bound of P_S, cannot be greater than $\frac{1}{2}$ in the continuous case for an unknown prior probability, since according to (42) it equals the smaller of the quantities $(1 - \alpha)$ and $(1 - \beta)$.

If, in certain cases, errors of second type are rated higher than those of first type or vice versa, one may use weights such that the weighted error chance can be reduced at the expense of the total success rate.

3.2. *Examples.* 1. Assume again as on p. 517 that the distribution of p is known except for its location, that is, $p(x \mid \theta) = f(x - \theta)$, where $f(z)$ is given. We restrict ourselves to the case where $f(z)$ is symmetrical with respect to $z = 0$ and decreases monotonically on both sides (Fig. 42).

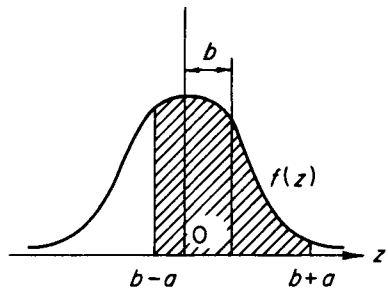

FIG. 42. Problem with maximum success rate.

The hypothesis to be tested may be $\mid \theta \mid \leqq b$. If the interval a_1, a_2 (with $a_2 > a_1$) is taken as the region of acceptance, the power function is

$$P(\bar{A} \mid \theta) = 1 - \int_{a_1}^{a_2} f(x - \theta)\, dx = 1 - \int_{a_1-\theta}^{a_2-\theta} f(z)\, dz. \tag{43}$$

Its derivative with respect to θ is $f(a_2 - \theta) - f(a_1 - \theta)$. Thus according to the assumptions about $f(z)$

$$\frac{dP(\bar{A} \mid \theta)}{d\theta} \gtreqless 0 \qquad \text{for} \qquad \theta \gtreqless \frac{a_1 + a_2}{2}.$$

If we choose $a_1 = -a$, $a_2 = a$, the power function has a minimum at $\theta = 0$ and increases on both sides monotonically (Fig. 43). The maximum power in the interval $\mid \theta \mid \leqq b$ (which is our H) and the minimum outside

H, that is, α and $1 - \beta$, both equal the power value at $\theta = \pm b$. To have both quantities $= 0.5$, we must choose a such that

$$1 - \int_{-a}^{a} f(x - b)\, dx = 1 - \int_{-a}^{a} f(x + b)\, dx = \tfrac{1}{2}.$$

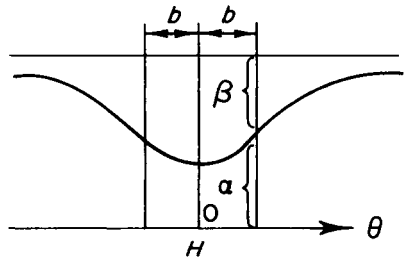

FIG. 43. Power curve to Fig. 42.

This is fulfilled if

$$\int_{-a-b}^{a-b} f(z)\, dz = \int_{b-a}^{b+a} f(z)\, dz = \tfrac{1}{2}. \tag{44}$$

For this interval $(b - a, b + a)$ the success rate has its highest possible value. In Fig. 42 the curve $f(z)$ is plotted and the meaning of (44) indicated: *one has to determine, to the right and to the left of $z = b$, two symmetrical points $b - a$ and $b + a$ such that the shaded area is $\frac{1}{2}$.* For instance, in the case of a normal distribution

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-z^2/2\sigma^2}$$

we must make

$$\Phi\left(\frac{b + a}{\sigma}\right) - \Phi\left(\frac{b - a}{\sigma}\right) = \frac{1}{2}. \tag{44'}$$

A table of the Gaussian shows, e.g., $\Phi(-0.25) = 0.4$ and $\Phi(1.28) = 0.9$. Thus, Eq. (44′) is fulfilled with $b + a = 1.28\sigma$ and $b - a = -0.25\sigma$, that is, for $b = 0.515\sigma$, the region of acceptance $|x| \leqslant a = 0.765\sigma$ has the success rate $\frac{1}{2}$.

2. We now give an example of a problem with *discontinuous* θ. In this case, even if nothing is known regarding $P(\theta)$ a higher success rate than $\frac{1}{2}$ can be achieved. Suppose we have three balls in a bag, black and white ones, but their distribution is unknown. The ratio of white in the bag can have one of the four values $0, \frac{1}{3}, \frac{2}{3}, 1$. We propose to test the

hypothesis that this ratio equals $\theta_0 = \frac{1}{3}$, by drawing a ball n times and counting the number x of white balls in each such group of trials. The probability of x is

$$p(x \mid \theta) = \binom{n}{x}\theta^x(1 - \theta)^{n-x}.$$

To fix our ideas, let us take $n = 12$ and the region A of acceptance consisting of the points $x = 1, 2, 3, 4, 5$, and 6. Then the power function is

$$P(\bar{A} \mid \theta) = 1 - \sum_{x=1}^{6} \binom{12}{x}\theta^x(1 - \theta)^{12-x}. \tag{45}$$

The four values of $P(\bar{A} \mid \theta)$ for the four possible θ can easily be computed:

$\theta =$	0	$\frac{1}{3}$	$\frac{2}{3}$	1
$P(\bar{A} \mid \theta) =$	1	0.074	0.822	1.

Our α is the power value for $\theta = \frac{1}{3}$, to wit 0.074, and $1 - \beta$ is the smallest power value among the other ones, to wit 0.822. The success rate is the smaller of the two quantities $1 - \alpha$ and $1 - \beta$ and that is 0.822. Result: If we accept the hypothesis $\theta = \frac{1}{3}$ when one to six white balls appear in 12 drawings and reject it if none appears or more than six, we are right in at least 82.2 % of all cases, whatever the (discrete) *a priori* probability of θ may be. If the region of acceptance is reduced to $A = 2, 3, 4, 5, 6$, hence, $\bar{A} = 0, 1, 7, 8, \ldots, 12$ we obtain $P(\bar{A} \mid \frac{1}{3}) = \alpha = 0.12$, $P(\bar{A} \mid \frac{2}{3}) = 0.822$ as before and $P_S \geqslant 0.822$ as before.

3.3. *Generalization to several dimensions.* All statements of this theory hold also in cases where instead of x we have a k-dimensional vector[1] or several variables $x^1, x^2, \ldots, x^k$ and instead of θ, several parameters $\theta_1, \theta_2, \ldots, \theta_l$.

Then A is a region in the k-dimensional x-space (or in the $k \cdot n$-dimensional sample space if judgment is passed after n k-dimensional observations), and H a region in the l-dimensional θ-space. The (composite) hypothesis to be tested is that the point $\theta_1, \ldots, \theta_l$ falls into a region H of the parameter space. We may again use the term "simple hypothesis" for the case where H is a single point in θ-space. The *power function* will be defined as an integral in the x-space:

$$P(\bar{A} \mid \theta_1, \ldots, \theta_l) = \iint_{(\bar{A})} \cdots \int p(x^1, \ldots, x^k \mid \theta_1, \ldots, \theta_l)\, dx^1 \cdots dx^k.$$

[1] We use superscripts for x^κ ($\kappa = 1, 2, \ldots, k$) in order to avoid confusion with the n results $x_1, x_2, \ldots, x_n$ obtained in n observations of *one* variable x.

Again, for given A and H, the maximum power for $\theta \in H$ will be called α, the minimum power for $\theta \in \bar{H}$ will be $1 - \beta$, and the success rate the smaller of the quantities $1 - \alpha$ and $1 - \beta$. If the prior distribution is unknown, then, in the continuous case the success rate cannot be larger than $\frac{1}{2}$. While the principles are exactly the same in the multivariate problem, the mathematics involved is much more complicated since the question of the *shape* of the regions A and H arises. Much work has been done to find appropriate forms for the regions of acceptance and rejection in the case of given $p(x^1, \ldots, x^k \mid \theta_1, \ldots, \theta_l)$. We cannot go into this matter here and content ourselves with discussing one very simple example with *discrete* x- and θ-values, where no mathematical difficulty intervenes.

Referring to the former example of a bag with three balls, let us now assume that two bags each containing white and black balls in the ratios θ_1 and θ_2 are given and that $n = 5$ drawings are made from each bag. The numbers of white balls appearing in the drawings will be denoted by $x^1 = x$ and $x^2 = y$. The probability function follows from the combination of two Bernoulli laws:

$$p(x, y \mid \theta_1, \theta_2) = \binom{n}{x}\binom{n}{y}\theta_1^x\theta_2^y(1 - \theta_1)^{n-x}(1 - \theta_2)^{n-y}. \tag{46}$$

The hypothesis H to be tested may be $\theta_2 > \theta_1$, that is, the assumption that the second bag holds more white balls than the first. As a suitable region A of acceptance, we choose $y > x$. Since x and y, for $n = 5$, can take the six different values 0, 1, ..., 5, the x-space includes 36 points, 15 of which belong to the region A: $x = 0$, $y = 1$; $x = 0$, $y = 2$; ...; $x = 4$, $y = 5$. The power function

$$P(\bar{A} \mid \theta_1, \theta_2) = 1 - \sum_{x<y}^{0\cdots5} \binom{5}{x}\binom{5}{y}\theta_1^x\theta_2^y(1 - \theta_1)^{5-x}(1 - \theta_2)^{5-y} \tag{47}$$

is given by a matrix of 16 quantities, the border values of which are obvious: $P = 1$ for any of the possibilities $\theta_1 = 1$ and 0, and $\theta_2 = 1$ and 0. Furthermore, it is seen that the power is decreasing when θ_2 increases at constant θ_1 or θ_1 decreases at constant θ_2. It follows that the minimum power for points in $\bar{H}$ can only appear at the two points $\theta_1 = \theta_2 = \frac{1}{3}$ and $\theta_1 = \theta_2 = \frac{2}{3}$, which have the same power, according to (47), and that the maximum power for points in H is at $\theta_1 = \frac{1}{3}$, $\theta_2 = \frac{2}{3}$. Thus,

$$1 - \beta = P(\bar{A} \mid \tfrac{1}{3}, \tfrac{1}{3}) = 1 - 3^{-10}[5 \cdot 514 + 10 \cdot 260 + 60 \cdot 136 + 55 \cdot 80 + 126 \cdot 32] = 0.631,$$

$$\alpha = P(\bar{A} \mid \tfrac{1}{3}, \tfrac{2}{3}) = 1 - 2^5 \cdot 3^{-10}[2 \cdot 210 + 4 \cdot 120 + 8 \cdot 45 + 16 \cdot 10 + 32 \cdot 1] = 0.213.$$

The smaller of the two quantities $1 - \alpha$ and $1 - \beta$ is the latter one, equal to 0.631. We therefore have in the long run a chance of at least 63 % of being right if we accept the hypothesis $\theta_2 > \theta_1$, each time y is observed to be greater than x and vice versa.[2] One can also see that the region $y > x$ is most powerful with respect to other groups of x, y-values.

3.4. *Remark on the success rate for large* n. A final more general remark may be added concerning the often mentioned limitation of the success rate to 50 % in the case of a continuous parameter θ. It seems strange that not more should be obtainable when, e.g., $x = a$ is the average of a very large number n of observations and θ the mean value of the corresponding probability distribution. One would think that with increasing sample size n, the inference upon the hypothesis $\theta \in H$ being true or not must become more and more safe. The explanation lies in the fact that in such a case we do not act in complete ignorance of the *a priori* chance of θ. In fact, a definite assumption about $P(\theta)$ is silently introduced. If we consider a sequence of like problems, with increasing n, we take as obvious that the over-all distribution of θ has nothing to do with the number of observations and, therefore, remains independent of n. It can be seen[3]—the proof will not be given here— that with an appropriate restriction on $P(\theta)$, higher success rates, asymptotically reaching 1, can be achieved. (The bound 0.50 holds if *nothing* is known or assumed about P.) The following argument might help to make this comprehensible. If, for increasing n, the function $p(x \mid \theta)$ concentrates more and more around $\theta = x$ and the interval A on the x-axis is chosen approximately coincident with the interval H in θ, then the power $P(\bar{A} \mid \theta)$ will be very small as long as θ falls in H and close to one if θ lies outside H. In Fig. 44a, the maximum of $P(\bar{A} \mid \theta)$ within H and the minimum outside H, that is, α and $1 - \beta$, coincide. But the power curve approaches more and more the discontinuous shape of Fig. 44b where evidently α is small and $1 - \beta$ large; thus the smaller of the two quantities $1 - \alpha$ and $1 - \beta$ approaches 1.

It would be worthwhile to continue the study of the Neyman-Pearson theory from the Bayesian point of view adopted here. We will, however, not go farther.[4]

---

[2] This chance could be raised by splitting up points where $x = y$.

[3] R. v. MISES, "On the problem of testing hypotheses." *Ann. Math. Statist.* **14** (1943), pp. 238-252, see p. 250.

[4] It is particularly regrettable that limitation of space prevents us from presenting the elements of "decision theory" which is in full agreement with Bayes' concepts.

Problem 10. The variate x is subject to a normal distribution with given variance σ^2 and unknown mean value θ. One wants to test the hypothesis $|\theta| \leqslant 0.2\sigma$ taking the region of acceptance to be the interval $|x| \leqslant \lambda\sigma$. Determine the maximum errors of first and second type as functions of λ. Show that if α is given, no most powerful test with respect to all intervals exists.

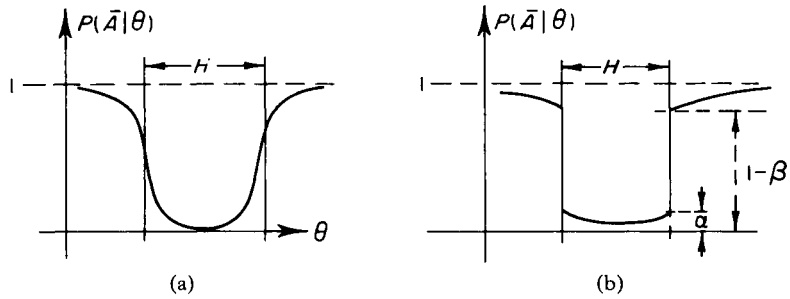

FIG. 44. Behavior of power curve as n tends to infinity.

Problem 11. In the example of two bags holding three balls each, the hypothesis $\theta_1 + \theta_2 > 1$ is to be tested. Choose a suitable region of acceptance and compute the success rate.

4. *On Sequential Sampling*

4.1. *The idea.* It might appear somehow artificial to fix in advance the size n of a sample. The more flexible approach—to let the number of observations depend on the outcome—has been advanced by H. F. Dodge and H. G. Romig,[1] by W. Bartky,[2] H. Hotelling, and others, and was developed into a theory by A. Wald[3] and collaborators.

The main idea is as follows: x_1, x_2, ... is an infinite sequence of observations. The given distribution $p(x \mid \theta)$ depends—as always—on an unknown parameter θ. Denote by $A^{(1)}$, $A^{(2)}$, $A^{(3)}$, ... the sequence of

[1] H. F. DODGE and H. G. ROMIG, "A method of sampling inspection," *Bell System Tech. J.* **8** (1929), pp. 613-631.

[2] W. BARTKY, "Multiple sampling with constant probability." *Ann. Math. Statist.* **14** (1943), pp. 363-377.

[3] A. Wald, (a) *Sequential Analysis of Statistical Data. Theory.* Statistical Research Group, Columbia Univ., NDRC Report, 1943, [29a].

(b) "On cumulative sums of random variables." *Ann. Math. Statist.* **15** (1944), pp. 283-296.

(c) "Sequential tests of statistical hypotheses." *ibid.* **16** (1945), pp. 117-186.

(d) *Sequential Analysis*, New York, 1947 [29b], and many other papers.

sample spaces for x_1, for x_1, x_2, for x_1, x_2, x_3, ..., etc. Each $A^{(m)}$, $m = 1, 2, \ldots$ is divided into three mutually exclusive parts $A_1^{(m)}$, $A_2^{(m)}$, $A_3^{(m)}$. If x_1 belongs to $A_1^{(1)}$, the null hypothesis H_0 is accepted, if it belongs to $A_2^{(1)}$, H_0 is rejected, and if it belongs to $A_3^{(1)}$ a further observation is made. We then consider x_1, x_2; if they belong to $A_1^{(2)}$, H_0 is accepted, if to $A_2^{(2)}$, it is rejected and if to $A_3^{(2)}$, one goes on, etc. The process is terminated if either the first or the second decision has been reached. Let N be the number of observations at which the process is terminated. *N is a random variable*, its value depending on the outcome of the observations.

Adopting the essential idea of sequential sampling, we have still a further choice of approach. In particular, the Neyman-Pearson procedure can be considered a special case of sequential sampling. We observe repeatedly groups of $n \geqslant 1$ trials. Accordingly we put $A_3^{(i)} = A^{(i)}$ for $i = 1, 2, \ldots, n-1$ (that means: after the first, second, ..., $(n-1)$th trial, at least one more trial follows); $A_2^{(n)} = \bar{A}$, $A_1^{(n)} = A^{(n)} - \bar{A} = A$, where A and $\bar{A}$ are the regions of acceptance and rejection.

4.2. *Simplest problem.* The first task of a sequential test is to define the regions $A_k^{(m)}$, $k = 1, 2, 3$, $m = 1, 2, \ldots$. Following Wald, we consider now the simplest case (see Section 2.4) *where θ can assume two values only, θ_0 and θ_1*, and H_0 is $\theta = \theta_0$, while H_1 is $\theta = \theta_1$. The *a priori* probability (or chance) is then discrete with $p(\theta_0) = \pi_0$, $p(\theta_1) = \pi_1$, $\pi_0 + \pi_1 = 1$, and $P_m(x_1, x_2, \ldots, x_m \mid \theta_0)$, $p_m(x_1, x_2, \ldots, x_m \mid \theta_1)$ are the probability densities (or probabilities) of m observations, under the hypotheses θ_0 or θ_1, respectively. If the trials are independent with identical distributions we have

$$p_m(x_1, \ldots, x_m \mid \theta_0) = f_0(x_1) \cdots f_0(x_m), \qquad p_m(x_1, \ldots, x_m \mid \theta_1) = f_1(x_1) \cdots f_1(x_m).$$

We still say that an error of first kind is committed if H_0 is true and is rejected and an error of second kind if H_1 is true and rejected and the respective probabilities are $\alpha\pi_0$ and $\beta\pi_1$. However, the precise definition will have to be restated since now the number of trials before a decision is a random variable, not known in advance. Before dealing with these questions we present the following sequential procedure. We abbreviate

$$p_m(x_1, \ldots, x_m \mid \theta_0) = p_{m0}, \qquad p_m(x_1, \ldots, x_m \mid \theta_1) = p_{m1}.$$

The posterior probability, formerly $q_n(\theta \mid x)$ is now denoted by $q_m(\theta_0 \mid x)$, $q_m(\theta_1 \mid x)$ or briefly $q_m(\theta_0)$, $q_m(\theta_1)$ with sum one. By Bayes' formula we have

$$q_m(\theta_0) = \frac{\pi_0 p_{m0}}{\pi_0 p_{m0} + \pi_1 p_{m1}}, \qquad q_m(\theta_1) = \frac{\pi_1 p_{m1}}{\pi_0 p_{m0} + \pi_1 p_{m1}}. \tag{48}$$

For the same reasons as before (p. 508) π_0 and π_1 will be neither zero nor one. Now let d_0 and d_1 be two positive numbers, $\frac{1}{2} < d_i < 1$, $i = 0, 1$. We consider the following sequential test: for each m, compute $q_m(\theta_0)$ and $q_m(\theta_1)$. If $q_m(\theta_0) \geqq d_0$ for $m = 1, 2, \ldots$, then θ_0 is accepted; if $q_m(\theta_1) \geqq d_1$, θ_1 is accepted; if neither holds, make an additional observation.[4] We have to show that the two inequalities

$$q_m(\theta_0) = \frac{\pi_0 p_{m0}}{\pi_0 p_{m0} + \pi_1 p_{m1}} \geqq d_0 \tag{49}$$

$$q_m(\theta_1) = \frac{\pi_1 p_{m1}}{\pi_0 p_{m0} + \pi_1 p_{m1}} \geqq d_1 \tag{50}$$

cannot hold simultaneously. This is clear since addition would give

$$1 \geqq d_0 + d_1$$

while $d_0 + d_1 > 1$. Therefore, three mutually exclusive and exhaustive regions are indeed defined for any m. Let us denote them in the present special case by $A_0^{(m)}$, $A_1^{(m)}$, $A_s^{(m)}$ instead of $A_1^{(m)}$, $A_2^{(m)}$, $A_3^{(m)}$. Then, $A_0^{(m)}$ is "acceptance of θ_0," $A_1^{(m)}$ "acceptance of θ_1," and $A_s^{(m)}$ "sequitur = go on"; $A_0^{(m)}$ is defined by $q_m(\theta_0) \geqq d_0$, $A_1^{(m)}$ by $q_m(\theta_1) \geqq d_1$, and $A_s^{(m)}$ by $q_m(\theta_0) < d_0$ and $q_m(\theta_1) < d_1$.

The inequalities (49) and (50) are equivalent to

$$\frac{p_{m1}}{p_{m0}} \leqq \frac{\pi_0}{\pi_1} \frac{1 - d_0}{d_0} \quad \text{and} \quad \frac{p_{m1}}{p_{m0}} \geqq \frac{\pi_0}{\pi_1} \frac{d_1}{1 - d_1}. \tag{51}$$

From what we said before the quotient π_0/π_1 is positive and finite.

4.3. *Sequential ratio test.* We are now going to modify the construction of the three regions, in a way suggested by (51) and by the ratio test of Section 2.4. Choose two constants[5] A and B, such that $0 < B < A$. At each stage compute p_{m0} and p_{m1}. (If $p_{m0} = p_{m1} = 0$ we set $p_{m1}/p_{m0} = 1$.) We state: accept H_0 if

$$\frac{p_{m1}}{p_{m0}} \leqq B, \tag{52}$$

accept H_1 if

$$\frac{p_{m1}}{p_{m0}} \geqq A, \tag{53}$$

[4] Note that if d_0, for example, were allowed to be less than $\frac{1}{2}$, the hypothesis H_0 might be accepted when the conditional probability of H_1 is higher than that of H_0.

[5] This A has nothing to do with "region of acceptance."

take one more observation if

$$B < \frac{p_{m1}}{p_{m0}} < A. \tag{54}$$

Thus, the number of observations required by the test is the smallest integer $m = N$ for which either (52) or (53) holds. The test procedure defined by (52), (53) and (54) is called a *sequential probability ratio test.*

Obviously, there must be relations among α, β, A, B, and π_0 where α and β have the meaning as in (27)((27′). We wish now to derive these relations. We proceed first in a not quite rigorous way which will be justified presently.

The inequality (53) is equivalent to

$$\frac{\pi_1 p_{m1}}{\pi_0 p_{m0} + \pi_1 p_{m1}} \geqslant \frac{\pi_1 A}{\pi_0 + \pi_1 A} \quad \text{or} \quad q_m(\theta_1) \geqslant \frac{\pi_1 A}{\pi_0 + \pi_1 A}, \tag{55}$$

as seen immediately. Let us temporarily denote the right-hand side of (55) by r. The left-hand side of (55) is the posterior probability $q_m(\theta_1)$ that H_1 is true, given some observed sample. For all samples for which H_1 is accepted [by reason of (53)] this probability is $\geqslant r$. It follows that for any sample for which H_1 is accepted, the conditional probability Q_1 that H_1 is true when it is accepted is $\geqslant r$, i.e.,

$$Q_1 \geqslant \frac{\pi_1 A}{\pi_0 + \pi_1 A}. \tag{56}$$

Let us compute Q_1. It is equal to the event that H_1 is true and will be accepted, divided by the probability that H_1 will be accepted; the first is $\pi_1(1 - \beta)$, the second $\pi_0\alpha + \pi_1(1 - \beta)$; hence,

$$Q_1 = \frac{\pi_1(1 - \beta)}{\pi_0\alpha + \pi_1(1 - \beta)}, \tag{57}$$

and from (56) and (57)

$$\frac{\pi_1(1 - \beta)}{\pi_0\alpha + \pi_1(1 - \beta)} \geqslant \frac{\pi_1 A}{\pi_0 + \pi_1 A}. \tag{58}$$

From (58) we have

$$A \leqslant \frac{1 - \beta}{\alpha}. \tag{59}$$

In a similar way an inequality for B can be derived, namely,

$$B \geqslant \frac{\beta}{1-\alpha}. \tag{60}$$

Since we have assumed $B < A$, *we must have* $\beta/(1-\alpha) < (1-\beta)/\alpha$, *which gives* $\alpha + \beta < 1$; *this must, therefore, hold.* From (59) and (60) we also obtain, taking reciprocals,

$$\alpha \leqslant \frac{1}{A}, \tag{61}$$

$$\beta \leqslant B. \tag{62}$$

These inequalities (59)–(62) do not involve π_0. Hence, they are valid for any positive $\pi_0 < 1$ (and one concludes from continuity considerations that they also hold for $\pi_0 = 0$ or $\pi_0 = 1$). Hence, they are true independently of the prior distribution.

4.4. *Space of sequences. Termination of the sequential process.* We wish to analyze more precisely the concepts we have been using. Let $\{x_m\} = x_1, x_2, \ldots$ be an infinite sequence of observations. The set of all such possible sequences form the *infinite sample space*, or *space of sequences*; we call it S. Each particular infinite sequence is a "point" of S.[6] P_0 and P_1, where $P_i(S) = 1$, are two probability functions (set functions) which will be defined presently. The set of all sequences of S whose first m terms are $a_1, a_2, \ldots, a_m$ is what we called in Chapter II a *basic set of order m.* Wald calls it a *cylindric point* of order m since it amounts to a point in m-space. A subset of S will be called a basic set (or a cylindric point) if there exists a positive integer N for which it is a basic set of order N.

A sample—or basic set—$(a_1, a_2, \ldots, a_N)$ *will be said to be of type* 1 *if it leads to acceptance of* H_1, *i.e., if for this sample*

$$\frac{p_{N1}}{p_{N0}} \geqslant A \quad \text{and} \quad B < \frac{p_{m1}}{p_{m0}} < A, \qquad m = 1, 2, \ldots, N-1. \tag{63}$$

The sample $(a_1, a_2, \ldots, a_N)$ *is of the type* 0 *if it leads to acceptance of* H_0, *namely, if for this sample*

$$\frac{p_{N1}}{p_{N0}} \leqslant B, \qquad B < \frac{p_{m1}}{p_{m0}} < A, \qquad m = 1, 2, \ldots, N-1. \tag{64}$$

[6] In Chapter II the present S was denoted by B; but B is used here in the ratio test.

Denote now by S_0 the sum of all basic sets of type 0. They are clearly mutually exclusive, i.e., the procedure cannot end (with the acceptance of H_0) after two different numbers N of trials. Denote by S_1 the sum of all basic sets of type 1. These sets S_0 and S_1, (subsets of the set of sequences S) generalize the previous concept of region of acceptance (denoted by A) and region of rejection (denoted by $\bar{A}$), both previously described for n-dimensional sample space. There, however, $A + \bar{A}$ added up to the whole sample space, whereas here S may contain in addition to S_0 and S_1 still (even infinitely many) "points" which belong to no basic set of order N of either type 0 or type 1, hence $S_0 + S_1 \subset S$.

Denote now by $P_0(S_0) = P(S_0 \mid \theta_0)$ the probability of S_0 if θ_0 is true and with the same meaning $P(S_1 \mid \theta_0) = P_0(S_1)$, $P(S_0 \mid \theta_1) = P_1(S_0)$, $P(S_1 \mid \theta_1) = P_1(S_1)$. *Although* $S_0 + S_1 \subset S$ *we shall prove that*

$$\begin{aligned} P_0(S_0) + P_0(S_1) &= 1 \\ P_1(S_0) + P_1(S_1) &= 1. \end{aligned} \tag{65}$$

These equations state that with probability one the sequential process will terminate,[7] since S_0 and S_1 are the subsets of S that lead to a decision.

Before proving Eqs. (65) we illustrate the concepts by two simple examples. Suppose a die is tossed repeatedly. P_0 is specified by $p_0(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6$; P_1 is such that $p_1(1) = p_1(3) = p_1(5) = p_1(6) = \frac{1}{4}$, $p_1(2) = p_1(4) = 0$. We take $A = 3$, $B = \epsilon > 0$, with the positive ϵ as small as we wish. (These values correspond to an undesirably large value of α; we choose them in order to shorten the computations.) For $m = 1$, the possible results are 1, 2, 3, 4, 5, 6 and $p_1(x)/p_0(x) = \frac{1}{4}/\frac{1}{6} = \frac{3}{2}$ for $x = 1, 3, 5, 6$, while $p_1(x)/p_0(x) = 0$ for $x = 2, 4$. Hence the basic sets (2) and (4) are seen to be of type 0 while for (1), (3), (5), (6) we have not yet a decision. A second throw leads to 1, 1 or 1, 2 or 1, 3, ..., or 6, 6. We have $p_1(1, 1)/p_0(1, 1) = (\frac{3}{2})^2 = \frac{9}{4}$. Since $\epsilon < \frac{9}{4} < 3$ there is still no decision. But $p_1(1, 2)/p_0(1, 2) = 0$. The basic set (12) is of type 0; the same holds for (14), (32), (34), (52), (54), (62), (64). The other basic sets of order two are "undecided." Now $m = 3 : p_1(1, 1, 1)/p_0(1, 1, 1) = (\frac{3}{2})^3 = 27/8 > 3$, hence (111) is of type 1 and the same holds for (113), (115), (116) while (112), (114) are of type zero. In the same way we see that (131), (133), (135), (136) are of type one but (132). (134) of type zero. We have

$$\begin{aligned} S_0 = {} & [(2) + (4)] + [(12) + (14) + (32) + (34) + (52) + (54) + (62) + (64)] \\ & + [(112) + (114) + (132) + (134) + \cdots + (164)] + [\,] + [\,] + [\,]. \\ S_1 = {} & \{[(111) + (113) + (115) + (116)] + [(131) + (133) + (135) + (136)] \\ & + [\,] + [\,]\} + \{\,\} + \{\,\} + \{\,\}. \end{aligned}$$

[7] Equations (65) also state that although there may be infinitely many "undecidable" points forming the set $S - S_0 - S_1$, its measure is zero. In (65) the P_1 and P_0 must define "regular valuations" (Chapter II, p. 83). We shall also see that P_0 and P_1 in (65) are probabilities—not only measures.

Here we have $S_0 + S_1 = S$: *all* decisions are made after at most $N = 3$ trials. We find:

$$P_0(S_0) = \frac{1}{6} + \frac{1}{6} + 8 \cdot \frac{1}{36} + 8 \cdot 4 \cdot \frac{1}{216} = \frac{152}{216} = \frac{19}{27};$$

$$P_0(S_1) = \frac{16 \cdot 4}{216} = \frac{64}{216} = \frac{8}{27}; \qquad P_1(S_0) = 0; \qquad P_1(S_1) = \frac{64}{4^3} = 1.$$

Now, consider a simple example where there *are* undecidable points. We assume independence and take $p_1(0) = p$, $p_1(1) = q$, $p + q = 1$; $p_0(0) = q$, $p_0(1) = p$, where, for example: $p = \frac{1}{5}$, $q = \frac{4}{5}$. Then, $p_1(0)/p_0(0) = p/q = \frac{1}{4}$, and $p_1(1)/p_0(1) = 4$. We take $A = 16$, $B = 1/16$. [The values are chosen so as to shorten the computations; as will be seen presently (p. 536), α and β will be less than $1/A$ and B, respectively, as stated in (61) and (62).]

The basic sets (1) and (0) are undecided for $m = 1$ since $B < 4 < A$, and $B < \frac{1}{4} < A$. For $m = 2$ consider the samples (00), (01), (10), (11). Since $p_1(0)^2/p_0(0)^2 = 1/16 = B$, and $p_1(1)^2/p_0(1)^2 = 16 = A$, while $p_1(1)p_1(0)/p_0(1)p_0(0) = 4 \cdot (\frac{1}{4}) = 1$ it is seen that the basic set (00) is of type 0, (11) is of type 1, while (01) as well as (10) are undecided for $m = 2$. Let us express (01) and (10) as basic sets of higher order. For (010) the characteristic ratio is $(\frac{1}{4}) \cdot 4 \cdot (\frac{1}{4}) = \frac{1}{4}$, hence undecided for $m = 3$, and the same holds for (011), (100), (101). For $m = 4$ we reach a decision for (0100) and (1000) which are of type 0 while (0111) and (1011) are of type 1 and the four remaining sets (0110), (1010), (0101), (1001) are undecided for $m = 4$. If we then investigate the basic sets which constitute each of them, we see that such a basic set is of type 0 (of type 1) if it contains at least two more zeros (two more ones) than ones (than zeros). Thus, eight sets of order 5 are undecided for $m = 5$; for $m = 6$ there are four basic sets of type 0, namely (010100), (011000), (100100), (101000), and similarly four basic sets of type 1, while eight basic sets are still undecided. It is seen that for any N, no matter how large, there are still "undecided" points.

Let us compute the probabilities of S_0 and S_1. We have $S_0 = (00) + (0100) + \cdots$, $P_0(S_0) = q^2 + 2pq^3 + 4p^2q^4 + \cdots = q^2(1 + 2pq + 4p^2q^2 + \cdots) = q^2/(1 - 2pq) = q^2/(p^2 + q^2)$, $P_0(S_1) = p^2/(p^2 + q^2) = P_1(S_0)$, $P_1(S_1) = q^2/(p^2 + q^2)$. For the chosen values of p and q we obtain $P_0(S_1) = P_1(S_0) = 1/17$. We see from (61), (62) that with a choice of larger A and smaller B, α and β are smaller but the number of trials needed to reach a decision is greater. Anticipating Eqs. (71) we find $\alpha = \beta = 1/17$, which are indeed less than $1/A$ and B, respectively [see Eqs. (61)].

We now prove (65). We have

$$p_{m0} = p(x_1, x_2, \ldots, x_m \mid \theta_0) = f_0(x_1)f_0(x_2) \cdots f_0(x_m) \tag{66}$$

where $p_0(x_i) \equiv p(x_i \mid \theta_0)$, and likewise

$$p_{m1} = p(x_1, x_2, \ldots, x_m \mid \theta_1) = f_1(x_1)f_1(x_2) \cdots f_1(x_m) \tag{67}$$

and put, with $i = 1, 2, \ldots, m$:

$$\log \frac{p_1(x_i)}{p_0(x_i)} = z_i, \qquad z_1 + z_2 + \cdots + z_m = Z_m, \qquad m = 1, 2, \cdots. \tag{68}$$

Denote by N the smallest integer for which either $Z_N \geqslant \log A$ or $Z_N \leqslant \log B$, and if no such integer exists we say that $N = \infty$. Obviously N is the number of observations required by the sequential test and (65) *states that* (under H_0 as well as under H_1) *the measure of all those "points" of S for which* $N = \infty$, *equals zero.* This is not hard to prove.[8]

Let $c = |\log A| + |\log B|$. If $N = \infty$ (in the sense just explained), then for any integer r the following inequalities must hold for $k = 1, 2, \ldots$:

$$|Z_{kr} - Z_{(k-1)r}| < c\,, \tag{69}$$

or equivalently, with

$$\zeta_k^2 = (Z_{kr} - Z_{(k-1)r})^2 = (z_{(k-1)r+1} + z_{(k-1)r+2} + \cdots + z_{kr})^2 \tag{69'}$$

$$\zeta_k^2 < c^2, \qquad k = 1, 2, \ldots . \tag{69''}$$

We make the restrictive assumption that $E[z_i^2]$ is positive. Then the expected value of $(\sum_{i=1}^{j} z_i)^2$ tends to ∞ as $j \to \infty$. Hence, r the number of elements in (69) can be chosen so large that the expected value of ζ_1^2 is $> c^2$. If this is so, then the probability $P_1 = \Pr\{\zeta_1^2 < c^2\}$ must be < 1. The same holds for any k, hence

$$P_k = \Pr\{\zeta_k^2 < c^2\} < 1. \tag{70}$$

Since the $z_1, z_2, \ldots$ are independently distributed, each with the same distribution, the P_k of (70) is the same for all k; hence, $P_k = P$. Since, likewise the $\zeta_1, \zeta_2, \ldots$ are independent of one another it follows that the probability of the joint event that (69) holds for $k = 1, 2, \ldots, j$ equals P^j. Therefore, the probability is zero that (69) holds for all values of k. Hence $\Pr\{N = \infty\} = 0$. We have thus proved that *with probability* 1 *the sequential ratio test will eventually terminate*, or, in other words, we have proved Eqs. (65).[9]

We have seen here that the study of the cumulative sums of random variables like $Z_m = z_1 + z_2 + \cdots + z_m$ is of great interest in sequential analysis: the "game" ends if either $Z_m \geqslant a$ or $Z_m \leqslant b$ where a and b are some real magnitudes. This problem arises in a random walk with absorbing barriers (see Chapter IV, Section 14.1.) since the particle stops whenever it arrives at a wall, i.e., whenever the cumulative sum of the displacements reaches a certain value. Our theorem states that with probability 1 this will happen after a finite time. The same problem may be translated into the "gambler's ruin" problem: here too under

[8] Wald (footnote 3b, p. 527), p. 283, and Wald [29b], Appendix A.1.

[9] We have assumed in this proof mutual independence of the z_i. The result holds, however, for certain types of dependence.

reasonable conditions the probability of an unending game is zero.[10]

We have said that with *probability* one the sequential ratio test will terminate. We now add: *if and only if the measure of the set U of undecidable sequences of S equals zero, then the measures of $S_0 + S_1 = D$ and of $U = S - D$ are genuine probabilities.* In fact, $S_0 + S_1 = D$ is a sum of basic sets, an open set D. Its complement $S - D = U$ is closed and contains no basic set since any basic set must also contain sequences (points) of type 0 and of type 1 (p. 531). Hence the boundary $\mathfrak{B}(U) = U$ and $| U | = 0$ is necessary and sufficient for both U and $D = S - U$ to have content.

One could possibly define "a game" where $| U | > 0$; then both U and D would have measure but not probability.

We return to Eqs. (65); they state that whether H_0 or H_1 is true, the respective probability of the set of all unending sequences of S is zero. Therefore, we have the right to use $P(S_1 \mid \theta_0)$ instead of the $P(A \mid \theta_0)$ of Eq. (19). Hence, we put

$$P_0(S_1) = P(S_1 \mid \theta_0) = \alpha, \qquad P_0(S_0) = P(S_0 \mid \theta_0) = 1 - \alpha$$

and (71)

$$P_1(S_0) = P(S_0 \mid \theta_1) = \beta, \qquad p_1(S_1) = P(S_1 \mid \theta_1) = 1 - \beta\,.$$

Then $\pi_0\alpha$ *is the probability of an error of first kind,* $\pi_1\beta$ *that of an error of second kind.* Equations (57), etc., are thus seen to be correct although we need Eqs. (71) in order to know exactly what is meant by α, $1 - \alpha$ and β, $1 - \beta$.

It is now very easy to recover the inequalities (59)–(62). Consider a sample—a basic set—belonging to S_1. For this sample $P_{N_1}/P_{N_0} \geqslant A$ holds by (63). Since this is true for any basic set belonging to S_1 it is true for S_1 and

$$P_1(S_1) \geqslant A \cdot P_0(S_1) \tag{72}$$

In exactly the same way we obtain

$$P_1(S_0) \leqslant B \cdot P_0(S_0), \tag{72'}$$

and using (71) we see that

$$A \leqslant \frac{1 - \beta}{\alpha}, \qquad B \geqslant \frac{\beta}{1 - \alpha},$$

which are Eqs. (59) and (60), from which (61) and (62) follow. The set of all points (α, β) which for given A, B satisfy the inequalities (59)

[10] See more in Feller [7b], p. 330.

and (60) is the interior and the boundary (shaded) of the quadrilateral bounded by the lines L_1, L_2 and the axes, where L_1, L_2 have the equations $A\alpha = 1 - \beta$ and $B(1 - \alpha) = \beta$ (Fig. 45).

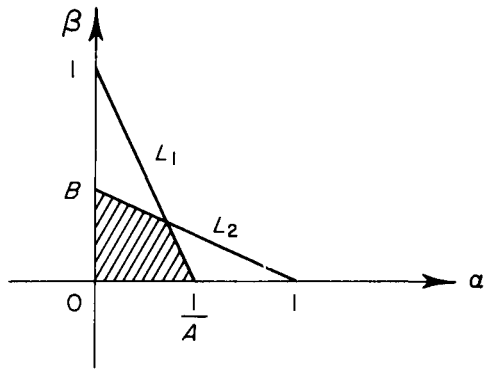

FIG. 45. Sequential ratio test: relation between A, B, α, β.

4.5. *Approximation for the bounds A and B.* Suppose that we wish to find $A = A(\alpha, \beta)$ and $B = B(\alpha, \beta)$ in such a way that α and β take certain prescribed values. The exact determination of $A(\alpha, \beta)$, $B(\alpha, \beta)$ is not easy. We know however that

$$A(\alpha, \beta) \leqslant \frac{1 - \beta}{\alpha}, \qquad B(\alpha, \beta) \geqslant \frac{\beta}{1 - \alpha}. \tag{73}$$

If we put $(1 - \beta)/\alpha = A'(\alpha, \beta)$, $\beta/(1 - \alpha) = B'(\alpha, \beta)$, then $A' \geqslant A(\alpha, \beta)$, $B' \leqslant B(\alpha, \beta)$. To these A' and B' belong values α', β' different from α, β. Among the A', B', α', β' the following inequalities hold [according to (59) and (60)]:

$$\frac{\alpha'}{1 - \beta'} \leqslant \frac{1}{A'} = \frac{\alpha}{1 - \beta} \quad \text{and} \quad \frac{\beta'}{1 - \alpha'} \leqslant B' = \frac{\beta}{1 - \alpha}, \tag{74}$$

hence

$$\alpha' \leqslant \frac{\alpha}{1 - \beta}, \qquad \beta' \leqslant \frac{\beta}{1 - \alpha}.$$

Multiplying the first Eq. (74) by $(1 - \beta)(1 - \beta')$ and the second one by $(1 - \alpha)(1 - \alpha')$ and adding we obtain

$$\alpha' + \beta' \leqslant \alpha + \beta.$$

Hence, at least one of the inequalities $\alpha' \leqslant \alpha$, and $\beta' \leqslant \beta$, must hold. In other words, by using A', B' instead of A, B at most one of the probabilities α, β may be increased.

Let us remember our examples: in the first example we had $A = 3$, $B = \epsilon > 0$ and we found $\alpha = P_0(S_1) = 8/27$, $\beta = P_1(S_0) = 0$. From these we find by (74), $A' = 27/8$, $B' = 0$. If we had taken these values instead of our $A = 3$, $B = \epsilon$ the computation would not have changed and we would have obtained the same values α and β. In the second example we took $A = 16$, $B = 1/16$. [These are the smallest values $1/A$ and B, for which, with our chosen $p = \frac{1}{5}$, $q = \frac{4}{5}$, we

can get a decision for basic sets which have two more zeros (ones) than ones (zeros).] With these values we found $P_0(S_1) = \alpha = P_1(S_0) = \beta = 1/17$ and from (74), $A' = 16$, $B' = 1/16$, equal to our A and B. We would have found the same α, β, A', B' if we had started with $A = 10$, $B = \frac{1}{8}$, say.

Wald recommends, in general, the use of A' and B' in the case of small α and β; then one will in general obtain values α', β' which are very close to α, β.

Our presentation of the idea of sequential analysis (although relating mainly to the special case where θ takes on the values θ_0 and θ_1 only) may have shown the very attractive and interesting features of sequential tests.

B. Global Statements on Parameters (Section 5)

5. *Confidence Limits*

5.1. *Confidence region.* A new way to derive statements about the values of a parameter θ from observed values x of a variate whose distribution depends on θ was devised in about 1930 by J. Neyman.[1] It is now known as the method of *confidence limits* or *confidence intervals*. Again, it is assumed that the distribution $p(x \mid \theta)$ depending on a parameter θ is known, and the *a priori* or over-all distribution $p(\theta)$ is unknown.[2] Again, we shall formulate statements about θ and try to find the chance of these statements being correct.

If both $p(x \mid \theta)$ and $p(\theta)$ are densities, the chance density for the occurrence of a definite x and a definite θ-value is the product

$$p(x \mid \theta)p(\theta). \tag{75}$$

In the x, θ-plane (Fig. 46) the total range of possible x- and θ-values may be indicated by the rectangle $ABCD$. We consider some region R inside this rectangle. In the infinite sequence of experiments, each observed x-value is connected with a definite θ-value; to each trial corresponds a certain point in the x, θ-plane. The chance (limiting frequency) of this point falling in the region R is, according to (75),

$$P(R) = \iint_{(R)} p(x \mid \theta)p(\theta)\, dx\, d\theta \tag{76}$$

[1] J. NEYMAN [23]; see also "On the problem of confidence intervals." *Ann. Math. Statist.* **6** (1935), p. 111; "Fiducial argument and the theory of confidence intervals." *Biometrika* **32** (1947), p. 128; etc.

[2] The value of x may be as before a statistic computed from a sample.

and this quantity can, in general, be computed only if $p(\theta)$ is known. We shall, however, see that there exist *special regions* R such that $P(R)$ can be found *independent of any knowledge* or *assumption about* $p(\theta)$.

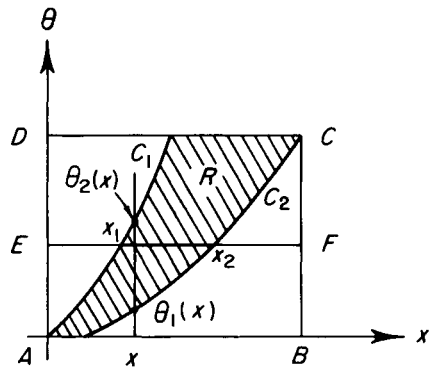

FIG. 46. A confidence belt.

On a straight line EF parallel to the x-axis, the integral of $p(x \mid \theta)\, dx$ taken from E to F has by definition the value 1. If, therefore, some quantity $\alpha < 1$, is given, one can find some smaller interval from $x_1 = x_1(\theta)$ to $x_2 = x_2(\theta)$ for which

$$\int_{x_1(\theta)}^{x_2(\theta)} p(x \mid \theta)\, dx = \alpha. \tag{77}$$

One may even choose x_1 arbitrarily (with certain restrictions) and determine x_2 so as to fulfill condition (77). Suppose that the region R in Fig. 46 is limited by two curves C_1 and C_2 such that on each horizontal line, (i.e., for each value of θ) condition (77) is satisfied with x_1 , x_2 being the abscissas of the intersections of C_1 , C_2 with the horizontal. We may also specify that the curve C_1 starts at A and the curve C_2 ends at C—still a variety of regions R fulfilling all these conditions will be available. For any such region R, the integral (76) takes the form

$$P(R) = \int_A^D p(\theta)\, d\theta \int_{x_1(\theta)}^{x_2(\theta)} p(x \mid \theta)\, dx = \int_A^D p(\theta)\alpha\, d\theta = \alpha \int_A^D p(\theta)\, d\theta = \alpha, \tag{78}$$

where the integration with respect to θ extends over all possible values of θ and, therefore $\int p(\theta)\, d\theta$ equals 1. Such a region R is a *confidence region*, which can be chosen in different ways, for given α.

Thus, for any prescribed value $\alpha < 1$, a region or "belt" R can be

found for which the chance $P(R)$ has the value α. How can the belt be used to formulate statements about θ? Assume a value x has been observed; draw, in Fig. 46, the vertical line with the abscissa x. It will intersect the border of the belt at two points, the ordinates of which may be called $\theta_1(x)$ and $\theta_2(x)$. The points can lie, as the case may be, on the curves C_1, C_2 or on the upper or lower edge of the rectangle. The statement that, for the observed x-value, θ lies between $\theta_1(x)$ and $\theta_2(x)$ is equivalent to the statement that the x, θ-point falls in the belt R. We therefore arrive at the result:

If, following each observation x, we contend that θ lies in the interval from $\theta_1(x)$ to $\theta_2(x)$, where $\theta_1(x)$ is the smallest and $\theta_2(x)$ the largest θ-value with abscissa x in the belt, we have the chance α of being right whatever the prior chance may be. If, for example, $\alpha = 0.90$, we are sure that in the long run 90 % of all our statements are correct.

Consider a different derivation. Denote by $I(\theta)$ any interval defined by Eq. (77) such that $\int_{I(\theta)} p(x \mid \theta)\, dx = \alpha$. All these intervals together cover a region R of the x, θ-plane. The same region R can be described by all intervals $J(x)$ each one for a given x [there may be no $J(x)$ for some x].

The following three statements are equivalent:

(1) A point (x, θ) belongs to R;
(2) A point with (given) θ belongs to $I(\theta)$ (between x_1 and x_2);
(3) A point with (given) x belongs to $J(x)$ (between $\theta_1(x)$ and $\theta_2(x)$).

Suppose there are many different "urns" each characterized by some (unknown) θ; for each of these urns the probability of "success" equals α. If we draw in any succession from these urns, then in the infinite sequence of drawings, α is the probability of success. Therefore, if we draw in any arbitrary succession from these urns there is a probability α that (2) holds and hence, that (3) holds. In other words, if, each time an x has been obtained in a drawing, we make the statement (3), the probability of our being right equals α.

The difference between this derivation and the previous one is that the existence of the limits $N_\theta/N \to p(\theta)$ has not been assumed here.

It is seen that a success chance as high as we wish can be reached. On the other hand, the statements for which this success chance holds are closely prescribed. In the problem of testing a hypothesis, we wanted to make, after each observation of an x-value, a contention about θ lying in a *previously determined interval*, the *same* interval whatever x has been observed. Now, by the method of confidence intervals, the contention refers each time to a different interval, and we are not free to choose these intervals. In our original inference problem, we wanted still more: we considered only those cases in which a certain $x = x_1$ had been observed and made a contention about θ falling in a given interval, given this specified x_1-value. The high success chance in the

method of confidence intervals is reached at the expense of freedom in formulating the contentions. Some examples will make this clearer.

5.2. *Examples.* 1. If the original distribution is a fairly regular slowly varying function $p(x \mid \theta)$, our method cannot give very substantial results. Let x be uniformly distributed over the interval from 0 to θ:

$$p(x \mid \theta) = \frac{1}{\theta}, \qquad 0 \leqslant x \leqslant \theta. \tag{79}$$

The range of x and θ in the x, θ-plane (Fig. 47) is limited by the positive

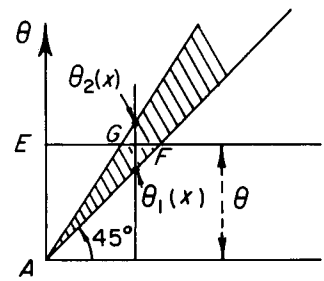

FIG. 47. A confidence belt for $p(x \mid \theta) = 1/\theta$.

θ-axis and the 45°-line $x = \theta$. On the horizontal line EF, we determine the point G so that $\overline{GF} = \alpha\overline{EF} = \alpha\theta$. Then

$$\int_{(G)}^{(F)} p(x \mid \theta)\, dx = \frac{1}{\theta}\,\alpha\theta = \alpha. \tag{80}$$

The belt R consists, therefore, of the sector between the straight lines AG and AF. The equations of these lines are $x = (1 - \alpha)\theta$ and $x = \theta$. For given x, the ordinates of the two intersections are

$$\theta_1(x) = x \qquad \text{and} \qquad \theta_2(x) = \frac{x}{1 - \alpha}. \tag{81}$$

Take $\alpha = 0.90$. If each time an x-value has been observed we state that "θ lies between x and $10x$," we have a 90 % chance of being right. Such contentions are:

$0.1 \leqslant \theta \leqslant 1$ if $x = 0.1$ has been observed

$0.2 \leqslant \theta \leqslant 2$ if $x = 0.2$ has been observed

$0.3 \leqslant \theta \leqslant 3$ if $x = 0.3$ has been observed, etc.

The infinite set of these contentions has the success chance 90 %. But it would be erroneous to think that one individual statement, e.g., the first—"if $x = 0.1$, then θ lies between 0.1 and 1"—is correct in 90 % of all cases in which $x = 0.1$ occurs. It can well happen that not in a single case where $x = 0.1$ has been observed θ falls in this interval, and that the success chance of 90 % results entirely from the correctness of the statements made in cases where $x \neq 0.1$; at the same time it may well be that the case $x = 0.1$ is just the one we are interested in.

In most cases which arise in practical statistics the function $p(x \mid \theta)$ has a quite different character than it has in example 1.

2. In the daily production of 0.5-in. steel balls, the diameter might be distributed according to a normal law with given constant variance σ^2 and unknown variable mean value θ. Each day a sample of n balls is taken from the lot and the average x of the n diameters measured. Its distribution is given by

$$p(x \mid \theta) = \sqrt{\frac{n}{2\pi}} \frac{1}{\sigma} e^{-n(x-\theta)^2/2\sigma^2} . \tag{82}$$

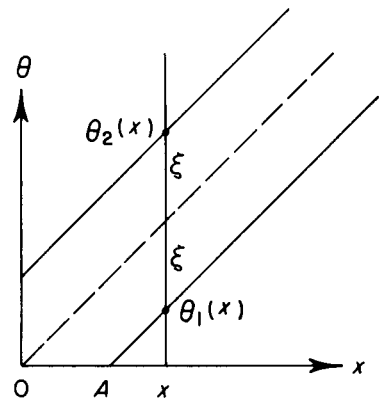

FIG. 48. A confidence belt for Gaussian $p(x \mid \theta)$ where θ is the mean value.

Here, the range of possible x, θ covers the entire plane (Fig. 48). We use as belt R the strip limited by two parallels to the bisectrix of the axes whose half width $\overline{OA} = \xi$ is determined [with $\Phi(x) = \frac{1}{2} + G(x)$] by

$$2G\left(\frac{\sqrt{n}\xi}{\sigma}\right) = \alpha. \tag{83}$$

It follows from (82) and (83) that condition (77) is fulfilled. On the other hand, it is seen from the figure that, for a given x, the limits in θ are

$\theta_1(x) = x - \xi$ and $\theta_2(x) = x + \xi$. Result: If on each consecutive day the sample average x is measured and the contention made that the unknown, varying mean value θ of the lot lies between $x - \xi$ and $x + \xi$, where ξ is the constant computed from (83), then, in the long run, the contention will be correct on approximately α N out of N days.

3. If an alternative with the unknown event probability θ is repeated very often (large n), the probability for the event frequency x is given by the Laplace-Bernoulli formula

$$p(x \mid \theta) = \sqrt{\frac{n}{2\pi\theta(1-\theta)}} \exp\Big(-\frac{n(x-\theta)^2}{2\theta(1-\theta)}\Big). \tag{84}$$

Here, the variables x and θ both range from 0 to 1. We use a belt of variable thickness 2ξ (the reader may sketch it) symmetrical with respect to the bisectrix OC, where ξ is determined for each θ by

$$\int_{\theta-\xi}^{\theta+\xi} p(x \mid \theta)\, dx = 2G\Big(\xi\sqrt{\frac{n}{\theta(1-\theta)}}\Big) = \alpha. \tag{85}$$

Let $\alpha = 0.99$, $\sqrt{n/\theta(1-\theta)}\,\xi = u$, $2G(u) = 0.99$, then $u = 2.6$. As $\xi = u\sqrt{\theta(1-\theta)/n}$ becomes small for large n, we get a narrow belt. If ξ is small, the ordinates $\theta_1(x)$ and $\theta_2(x)$ for given abscissa x equal approximately $x - \xi$ and $x + \xi$ with the ξ computed for $\theta = x$, that is from

$$\xi\sqrt{\frac{n}{x(1-x)}} = 2.6. \tag{86}$$

Our result is again a global statement, namely, that (approximately) $x - \xi \leqslant \theta \leqslant x + \xi$ where ξ is given by (86).

We have observed before that the confidence region corresponding to a given α is not unique. Obviously, we wish, in general, to obtain a *narrow* confidence belt, as in this example.

4. A slight modification takes place if x is a discrete variable and $p(x \mid \theta)$ not a density, but a probability. Assume that an alternative with the unknown event probability θ is tried $n = 5$ times. The probability for the occurrence of x events is

$$p(x \mid \theta) = \binom{5}{x}\theta^x(1-\theta)^{5-x}, \tag{87}$$

where x can take the six values 0, 1, ..., 5 and θ all values between 0 and 1.

The integral in (77) has now to be replaced by a sum, and it is no longer possible to satisfy the condition (77) in its original form of an equality since the sum takes only discrete values. We can, however, find an interval A for each θ such that if x_1 , x_2 ... are the x-values belonging to A,

$$p(x_1 \mid \theta) + p(x_2 \mid \theta) + \cdots \geqslant \alpha. \tag{88}$$

If the belt is composed of such intervals, the chance of an x, θ-point falling in the belt will be $\geqslant \alpha$.

To construct a suitable belt, we proceed as follows. Let us assume $\alpha = 0.99$. As long as θ is very small the value $p(0 \mid \theta) = (1 - \theta)^5$ will be large enough to supply the required 99 % probability, so that the belt will consist of the line $x = 0$ only, up to the θ-value for which $(1 - \theta)^5 =$

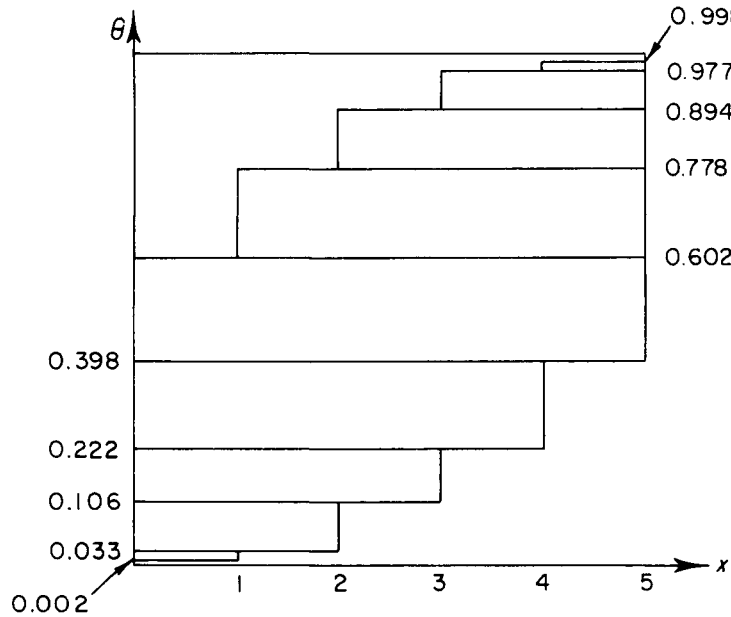

FIG. 49. A confidence belt for discrete probabilities.

0.99, that is, up to $\theta_1 = 0.002$ (Fig. 49). For $\theta > \theta_1$ we must include in the belt at least two x-values, so as to have

$$p(0 \mid \theta) + p(1 \mid \theta) = (1 - \theta)^5 + 5\theta(1 - \theta)^4 \geqslant 0.99, \tag{89}$$

which gives as the upper limit of θ the value $\theta_2 = 0.033$, and so on. The upper half of the belt is formed symmetrically. The figure shows what

contentions can be made with at least a 99 % chance of being right. One has to read off, for each of the six x-values, the smallest and the largest θ-values in the belt. This is the result: If the contention is made,

$$\begin{aligned}
&\text{each time } x = 0 \text{ is observed, that} && 0 \leqslant \theta \leqslant 0.602\\
&\text{each time } x = 1 \text{ is observed, that} && 0.002 \leqslant \theta \leqslant 0.778\\
&\text{each time } x = 2 \text{ is observed, that} && 0.033 \leqslant \theta \leqslant 0.894\\
&\text{each time } x = 3 \text{ is observed, that} && 0.106 \leqslant \theta \leqslant 0.977\\
&\text{each time } x = 4 \text{ is observed, that} && 0.222 \leqslant \theta \leqslant 0.998\\
&\text{each time } x = 5 \text{ is observed, that} && 0.398 \leqslant \theta \leqslant 1
\end{aligned}$$

the statements will be true, in the long run, in *at least* 99 % *of all cases.* Again, it is seen that this chance cannot be ascribed to any one of the individual statements. *The chance of at least* 99 % *holds only for the six joint statements.* To find the success chance of an individual statement, we have to return to the methods discussed in this chapter which dealt with the inference problem.

5. As a last and curious example, consider an alternative and assume that just one observation has been made and that the event has occurred. What can be concluded regarding the unknown probability θ of the event?[3] The obvious and correct answer is: "nothing." However, a statistician (when answering the questionnaire) dealt with this question by means of confidence intervals. Here θ is between 0 and 1. If $n = 1$ then $x = 0$ or $x = 1$; $p(0 \mid \theta) = 1 - \theta$, $p(1 \mid \theta) = \theta$. Let $\alpha = 0.9$ and construct the belt:

$$\begin{aligned}
\theta = 0, &\qquad p(0 \mid \theta) > \alpha,\\
\theta \leqslant 0.1, &\qquad p(0 \mid \theta) \geqslant \alpha,\\
0.1 < \theta < 0.9, &\qquad p(0 \mid \theta) + p(1 \mid \theta) \geqslant \alpha,\\
0.9 \leqslant \theta \leqslant 1, &\qquad p(1 \mid \theta) \geqslant \alpha.
\end{aligned}$$

Therefore, what can be concluded with 90 % chance are the following *two* answers:

$$\begin{aligned}
&\text{whenever } x = 0 \text{ is observed we say: } \theta < 0.9\\
&\text{whenever } x = 1 \text{ is observed we say: } \theta > 0.1.
\end{aligned}$$

This *pair* of statements has a 90 % chance.

[3] M. FRÉCHET, "Rapport sur une enquête internationale relative à l'estimation statistique des paramètres." *Bull. Inst. Internat. Statist.* (1947), pp. 363-422.

Suppose the "event" is the first prize in a lottery. If one concludes from $x = 1$ that $\theta > 0.1$ this is obviously wrong in most cases, that is, for most existing lotteries; but if we conclude from $x = 0$, that $\theta < 0.9$ this is obviously correct in most cases and *that* correct result, corresponding to $x = 0$ is much more frequent than the other which corresponds to $x = 1$. Both together have the limiting frequency 0.9.

5.3. *Generalization to several dimensions.* The method of confidence intervals can be extended to the case of *k chance variables and l parameters.* Let us assume that the joint probability density for two variables x and y depends on two parameters θ and σ and is given by $p(x, y \mid \theta, \sigma)$. We define a region R in four-dimensional space by means of a function $F(x, y, \theta, \sigma)$. Let R consist of all points for which $F \leqslant 0$. We specify F in such a manner that, for each pair of constants θ and σ

$$\iint_{(F<0)} p(x, y \mid \theta, \sigma)\, dx\, dy = \alpha\,, \tag{90}$$

that is, the intersection of R with a plane $\theta =$ constant, $\sigma =$ constant forms a region (in x and y) for which the probability is α. The chance of a point x, y, θ, σ falling in R is

$$P(R) = \iiiint_{(R)} p(x, y \mid \theta, \sigma) p(\theta, \sigma)\, dx\, dy\, d\theta\, d\sigma. \tag{91}$$

If (90) is fulfilled this expression equals

$$P(R) = \iint p(\theta, \sigma)\, d\theta\, d\sigma \cdot \iint_{(F<0)} p(x, y \mid \theta, \sigma)\, dx\, dy = \alpha \iint p(\theta, \sigma)\, d\theta\, d\sigma = \alpha,$$

In the same way as above, in Eq. (78). The following result can be stated: *If a function $F(x, y, \theta, \sigma)$ is chosen in such a way that Eq.* (90) *is fulfilled for any pair of constant θ and σ, one has the chance α of being right if, each time x, y have been observed, one contends that $F(x, y, \theta, \sigma) \leqslant 0$.*

Example. (Student's test, Chapter IX, Section 3.1). We take up the same problem as in Example 2 (p. 541), only the assumption that the variance σ^2 is constant and known will be dropped as being unrealistic. We thus have two unknown parameters, the daily mean value θ and the daily standard deviation σ. In addition to the sample average x, we observe the sample deviation s, that is, the square root of the dispersion

of the sample. For the joint distribution of the quantities x and s, calling the density $p(x, s \mid \theta, \sigma)$, we have

$$p(x, s \mid \theta, \sigma) = \text{const. } s^{n-2}\, e^{-n[s^2+(x-\theta)^2]/2\sigma^2}. \tag{92}$$

As the function F, we choose

$$F(x, s, \theta, \sigma) = (n-1)\left(\frac{x-\theta}{s}\right)^2 - t_\alpha^2 = t^2 - t_\alpha^2, \tag{93}$$

where t is Student's ratio and t_α a constant depending on n and α which we shall determine presently.

We know that for any t', t'' the relation $t' < \sqrt{n-1}\,(x-\theta)/s < t''$ has the probability $\int_{t'}^{t''} p_{n-1}(t)\,dt$ with p_{n-1} given by Eq. (22), Chapter IX. We take $t' = -t_\alpha$, $t'' = t_\alpha$; then denoting by $P_n(t)$ the c.d.f. of $p_n(t)$, we have to determine t_α from the condition

$$P_{n-1}(t_\alpha) - P_{n-1}(-t_\alpha) = \alpha. \tag{94}$$

This is done by using our table of Student's distribution. To find t_α for, say, $\alpha = 0.90$, one has to take the table value in the column marked 10 ($= 10\,\%$) and in the row corresponding to the $n-1$ (number of degrees of freedom) in question. For example, for $n = 10$, $n-1 = 9$, and $\alpha = 0.90$ one has $t_\alpha = 1.833$ and for $n-1 = 2$, $\alpha = 0.95$ one finds $t_\alpha = 4.303$.

The contention that can be made with the success chance α is $F \leqq 0$, that is,

$$(n-1)\left(\frac{x-\theta}{s}\right)^2 \leqq t_\alpha^2 \quad \text{or} \quad x - \frac{t_\alpha}{\sqrt{n-1}}\,s \leqq \theta \leqq x + \frac{t_\alpha}{\sqrt{n-1}}\,s. \tag{95}$$

In the case $n = 3$, $\alpha = 0.95$, this gives

$$x - 3.04s \leqq \theta \leqq x + 3.04s. \tag{96}$$

The meaning of the result (95) must be understood in the light of the preceding argument. It would be erroneous to think that if, on one day, the average in a sample of three is 0.505 and the deviation of the three values 0.001, there would be a chance of 95 % that θ, at this day, lies between 0.502 and 0.508. This high success chance refers to the entire sequence of daily varying statements of the form (95) or, more explicitly: If on many days we draw a daily sample of n from a normal population with unknown mean value θ and standard deviation σ and if for each sample we compute x and s and the limits in (95), then the frequency

of those cases for which θ is within these limits will in the long run equal α.

Sometimes the expression is used that the inequalities (95)— and similarly the contentions in the other examples—hold "*on the level of confidence of* 95 %." There is no objection to this expression as long as one is aware of what it stands for.

Problem 12. Given

$$p(x \mid \theta) = \theta e^{-\theta x}, \qquad x \geqslant 0, \quad \theta > 0,$$

find the confidence belt with left boundary $x_1 = 0$ and give the corresponding inequalities for θ.

Problem 13. Answer the same question for

$$p(x \mid \theta) = 2\theta^2 x e^{-\theta^2 x^2}, \qquad x \geqslant 0, \quad \theta > 0.$$

Problem 14. The daily output of a product is tested by a sample of 10. The variate x is supposed to follow a normal distribution law whose mean value θ and variance σ^2 are unknown. What inequalities for θ can be pronounced with a chance of 50 %, 80 %, 90 % to be correct?

C. Estimation (Sections 6 and 7)

6. Maximum Likelihood Method

6.1. *Maximum likelihood estimate.* The most restricted form of inference from an observed x upon an unknown parameter θ is the *estimation* of θ. Again, a distribution $p(x \mid \theta)$ depending on a parameter θ is given and it is supposed that a definite value of x has been observed. One asks for an "estimate" of θ, that is, for a value that is "presumably" correct or, at least, approximately correct. If θ can take discrete values only so that $p(\theta_1)$, $p(\theta_2)$, $p(\theta_3)$, ... are the *a priori* chances of the occurrences $\theta = \theta_1$, $\theta = \theta_2$, $\theta = \theta_3$, ..., respectively, we know that the product

$$p(x \mid \theta)p(\theta) \tag{97}$$

is proportional to the *a posteriori* chance $q(\theta \mid x)$ of θ. The x in (97) may stand for a statistic x, computed from n results x_1, x_2, ..., x_n. In general, however, $p(x \mid \theta)$ stands for $p(x_1, x_2, \ldots, x_n \mid \theta)$ and if the n

trials are independent and refer to the same distribution, then $p(x_1, x_2, \ldots, x_n \mid \theta) = p(x_1 \mid \theta)p(x_2 \mid \theta) \ldots p(x_n \mid \theta)$. Similarly, we may take the θ as representing several parameters, for instance $p(x \mid \theta)$ in (97) standing for $p(x_1, x_2, \ldots, x_n \mid \theta, \sigma) = p(x_1 \mid \theta, \sigma) \cdots p(x_n \mid \theta, \sigma)$, where the last equality is valid if the trials are independent and have the same distribution.

Returning to (97) we see that the value of θ, which *makes this product a maximum* for the given x, will in the long run be correct in more cases where x has been observed than any other θ-value: It is the value with the greatest posterior chance.

The trouble is again that $p(\theta)$ is, in general, not known so that this computation cannot be carried out. A simple expedient in this situation is to assume that p has the same magnitude for all possible θ-values, i.e., that the *a priori* chance is uniform (where in case of continuous θ its range must be finite). Then, one has only to maximize (with respect to θ) the given function $p(x \mid \theta)$, which is called the *likelihood function.* or *likelihood.* The likelihood of θ is the probability of the result x, given θ. This procedure really goes back to Gauss (1823). His idea was that the best value of the unknown parameter is the one which gives the greatest probability to what had been observed. Fisher[1] took this idea as the basis of a method for estimating one or several unknown parameters. Following Fisher, the value of θ for which $p(x \mid \theta)$ is maximum is called *the maximum likelihood estimate of* θ, considered also in Chapter IX, Section 5). It is in general silently assumed that one and only one maximum exists. Also, if θ is a continuously varying parameter so that $p(\theta)$ means a density, the same idea is applied: considering p as constant, we compute the θ that maximizes $p(x \mid \theta)$ for the given x and call it the maximum likelihood estimate $\hat{\theta}$ of θ.[2]

A considerable advantage of the maximum likelihood (m.l.) method is its transitivity: if t is found to be a m.l. estimate of some θ then t^2 is the m.l. estimate of θ^2, etc. This is untrue for some other methods of estimation (see Problem 18).

From what has been said, it follows that: (a) *The maximum likelihood estimate is the value of* θ *which for an observed* x, *has the greatest chance of being correct* (or the greatest chance density), *under the assumption that the a priori chance is equal for all* θ. Note that $p(x \mid \theta)$ is not a probability (or a chance) in θ; it is a probability or a chance in x.

[1] R. A. FISHER, *Messenger of Mathematics* **41** (1912), pp. 155-160.

[2] In line with our general agreement about notation we should denote any estimate of parameters θ or σ by t or s and use θ, σ for the unknown true values of the parameters. We do not attempt to do this throughout (see also Chapter IX, Section 5.1) since it is unusual. We shall often use θ_0 for the true value, $\tilde{\theta}$ for the maximum likelihood estimate.

We can give a slightly more general interpretation of the maximum likelihood estimate. Calling $q(\theta \mid x)$ the posterior chance of θ (which depends on the observed x), we have

$$Cp(x \mid \theta) = \frac{q(\theta \mid x)}{p(\theta)}, \tag{98}$$

where C does not depend on θ. The sum or integral over all θ-values of $q(\theta \mid x)$ as well as of $p(\theta)$ is necessarily 1. It follows that it is impossible that one of these two curves is situated above the other for *all* θ: they must intersect. Therefore, the quotient (98) cannot be smaller than 1 for all θ, i.e., the maximum of this quotient is greater than 1, except in the trivial case where it equals 1 for all θ, that is, where $p(x \mid \theta)$ does not really depend on θ. For some θ-values the chance $q(\theta \mid x)$ deduced from the observed x is greater than the *a priori* chance $p(\theta)$; for other θ-values it is smaller than $p(\theta)$. The maximum likelihood estimate $\tilde{\theta}$ is among the first group and we can state: (b) *The maximum likelihood estimate is that value of* θ *for which the ratio of posterior to prior chance is the greatest, for a given set of observations*, a correct, but rather weak statement.

Statement (a) above was valid under the assumption of a uniform $p(\theta)$. Now remember Chapter VII where we considered a sample $x_1, x_2, \ldots, x_n$ drawn from a population with probability $p(x \mid \theta)$, where θ denotes the unknown true value of some parameter (Cf. p. 495). We asked for the (posterior) probability $q_n(\theta)$ [denoted here by $q(\theta \mid x)$] of this parameter and found that if $p(\theta)$ satisfies certain rather general regularity conditions, the influence of the prior $p(\theta)$ on $q_n(\theta)$ becomes negligible with increasing n. Thus, in this case statement (a) holds approximately, with *any* $p(\theta)$, i.e., the m.l. estimate gives then approximately the value of θ which has the greatest chance of being correct: (c) *If* n *is large, then under certain regularity assumptions statement* (a) *holds without the restriction to a uniform prior chance.*

Fisher who introduced the terms likelihood and maximum likelihood method does not refer to the relationship between prior and posterior chances. He tries to make it plausible that $p(x \mid \theta)$, the probability of x for given θ is a measure of the "likelihood" of θ for given x and that, therefore, maximizing $p(x \mid \theta)$ for given x supplies a suitable estimate of θ. Actually, the m.l. method very often does lead to good results.

6.2. *Examples.* 1. Assume that an event whose unknown probability is q happened x times in n independent trials. This probability is $\binom{n}{x}q^x(1 - q)^{n-x}$, or omitting $\binom{n}{x}$ and writing $q = \theta$ we have the likelihood

$$p(x \mid \theta) = \theta^x(1 - \theta)^{n-x}.$$

Differentiating $\log p$ with respect to θ we find $(\log p)' = \frac{x}{\theta} - \frac{n-x}{1-\theta}$, which vanishes for $n\theta = x$. The maximum of $\log p$ and of p is obtained at $\theta = x/n$. Thus, $\tilde{\theta} = x/n$ *is the maximum likelihood estimate of* q.

We notice that in this case the expectation of x/n equals q; this fact is expressed by saying that $\tilde{\theta}$ is an *unbiased estimate* of q (see p. 559).

2. In his theory of observational errors, Gauss came to the conclusion that if some physical quantity whose "true value" is α is repeatedly measured by a certain measuring device, the probability of observing the value x_1 is given by

$$\frac{h}{\sqrt{\pi}} e^{-h^2(x_1-\alpha)^2} \tag{99}$$

with unknown constants α and h. See Chapter VII, Section 5. (This is the historical origin of the term Gaussian function.) If n observations are made, the probability of getting the set of values $x_1, x_2, \ldots, x_n$ is accordingly

$$\begin{aligned} L &= p(x_1, x_2, \ldots, x_n \mid \alpha, h) \\ &= \left(\frac{h}{\sqrt{\pi}}\right)^n \exp\{-h^2[(x_1-\alpha)^2 + (x_2-\alpha)^2 + \cdots + (x_n-\alpha)^2]\}. \end{aligned} \tag{100}$$

Using the shift-of-origin rule, calling a the average and s^2 the dispersion of the sample $x_1, x_2, \ldots, x_n$, we can write instead of (100)

$$L = \left(\frac{h}{\sqrt{\pi}}\right)^n e^{-h^2 n[s^2+(a-\alpha)^2]} \tag{100'}$$

Here the n variables $x_1, x_2, \ldots, x_n$ appear explicitly and there are two parameters α, h. The rule is: The maximum likelihood estimate for α and h is found by maximizing the right-hand side of (100) or (100′) with respect to α and h, keeping $x_1, x_2, \ldots, x_n$ (and therefore a and s^2) constant. Omitting a constant and passing to the logarithm, we have to maximize the function

$$M = n \log h - nh^2[s^2 + (a-\alpha)^2]$$

and find

$$\frac{\partial M}{\partial \alpha} = 2nh^2(a-\alpha) = 0, \qquad \frac{\partial M}{\partial h} = \frac{n}{h} - 2nh[s^2 + (a-\alpha)^2] = 0.$$

The first condition supplies $\alpha = a$, the second $n = 2nh^2s^2$. Thus, the estimates are a (or $\bar{x}$) for α and $1/2s^2$ for h^2. Considering that the variance σ^2 of the distribution (100′) is $\sigma^2 = 1/2h^2$, we can state: *If a sample is drawn from a normal distribution with unknown mean value* α *and unknown variance* σ^2 *the maximum likelihood estimate for* α *is the observation average* a *and for* σ^2 *the observation dispersion* s^2.

3. Let x be in $[0, 1]$, θ in $[\,-2 + 2\,]$ and

$$p(x \mid \theta) = 1 + \theta x - \frac{\theta}{2}, \qquad \text{where} \qquad \int_0^1 p(x \mid \theta)\, dx = 1. \tag{101}$$

A sample of $n = 2$ gave the results x_1 and x_2. Then

$$L = \left(1 + \theta x_1 - \frac{\theta}{2}\right)\left(1 + \theta x_2 - \frac{\theta}{2}\right)$$

$$= \theta^2\left(x_1 - \frac{1}{2}\right)\left(x_2 - \frac{1}{2}\right) + \theta(x_1 + x_2 - 1) + 1$$

$$L' = \frac{\partial L}{\partial \theta} = 2\theta(x_1 - \tfrac{1}{2})(x_2 - \tfrac{1}{2}) + (x_1 + x_2 - 1), \qquad \theta = \frac{1}{2}\,\frac{1 - x_1 - x_2}{(x_1 - \frac{1}{2})(x_2 - \frac{1}{2})},$$

$$L'' = \frac{\partial^2 L_1}{\partial \theta^2} = 2(x_1 - \tfrac{1}{2})(x_2 - \tfrac{1}{2}).$$

If x_1 and x_2 are both on the same side of $x = \frac{1}{2}$, then L'' is positive and there is a minimum and not a maximum. Note also that the range of θ is restricted to $-2 \leqslant \theta \leqslant 2$ and that the root of $L' = 0$ may fall outside this range and hence not be a maximum likelihood estimate even when $L'' < 0$. This and many other examples (see Problem 16) show that the method may become meaningless for small n.

6.3. *Consistency of maximum likelihood estimates.* An estimate $t(x) = t(x_1, x_2, \ldots, x_n)$ of a theoretical parameter θ is called *consistent* if for all ϵ the probability that $| t - \theta | < \epsilon$ tends toward one as $n \to \infty$. A sufficient condition is

$$\lim_{n\to\infty} E[t(x)] = \theta, \qquad \lim_{n\to\infty} \operatorname{Var}[t(x)] = 0. \tag{102}$$

Probably the first who proved the consistency of the likelihood estimate was H. Hotelling.[3] A very general proof has been given by A. Wald.[4]

[3] H. HOTELLING, *Trans. Amer. Math. Soc.* **32** (1930), p. 847.

[4] A. WALD, "On the consistency of the maximum likelihood estimate." *Ann. Math.*

Let $p(x \mid \theta)$ be continuous with respect to θ in an interval I which is supposed to contain the unknown true value θ_0 of θ as an inner point. We assume that p and θ are univalent and continuous functions of each other, i.e., to neighboring values of θ belong neighboring values of p and to different values of θ, different values of p and vice versa. We shall assume that repeated differentiation with respect to θ is allowed.

Let $x_1, x_2, \ldots, x_n$ be the results of n observations. Let us now consider *the arithmetical case where there are only k possible results,* $a_1, a_2, \ldots, a_k$. We form the likelihood function

$$L(x \mid \theta) = p(x_1 \mid \theta)p(x_2 \mid \theta) \cdots p(x_n \mid \theta). \tag{103}$$

We assume that the result a_i, $i = 1, 2, \ldots, k$ has appeared $nr_i = n_i$ times, $\Sigma_{i=1}^{k} r_i = 1$. Then, writing $L(\theta)$ for brevity,[5] we have

$$L(\theta) = [p(a_1 \mid \theta)^{r_1} p(a_2 \mid \theta)^{r_2} \cdots p(a_k \mid \theta)^{r_k}]^n \tag{103'}$$

$$\log L(\theta) = n \sum_{i=1}^{k} r_i \log p(a_i \mid \theta). \tag{103''}$$

We want to prove

(*d*) *For large n, it is almost certain that* $L(\theta)$ *has an (absolute) maximum at a point* $\tilde{\theta}$ *for which* $\mid \tilde{\theta} - \theta_0 \mid < \epsilon$ (ϵ *arbitrarily small*) *and that there is arbitrarily strong concentration of* $L(\theta)$ *around* $\theta = \tilde{\theta}$, *as* $n \to \infty$.

The idea of the proof is as follows: From the (first) law of large numbers it is almost certain that the observed frequencies $r_1, \ldots, r_k$ of the $a_1, \ldots, a_k$ are as close as we please to the true probabilities $p(a_1 \mid \theta_0), \ldots p(a_k \mid \theta_0)$, or, more generally, that the repartition $S_n(x)$ of the $x_1, x_2, \ldots, x_n$ is arbitrarily close to the true probability distribution $P(x \mid \theta_0)$. If then the $p(a_i \mid \theta_0)$ are introduced into L instead of the r_i, the resulting function $L_0(\theta)$ has, as we shall show, a maximum at $\theta = \theta_0$ and the property of concentration. Therefore, if, for large n, we determine the abscissa $\theta = \tilde{\theta}$ of the maximum of L, this $\tilde{\theta}$ must lie with very great probability arbitrarily close to the correct θ_0, toward which it converges as $n \to \infty$.

Statist. **20** (1949), pp. 595-601, and quotations of other proofs. See also J. WOLFOWITZ, "On Wald's proof of the consistency of the maximum likelihood estimate." *Ibid.* pp. 601-602. See also Cramér [4], pp. 500-504.

[5] In Chapter IX, Section 5 we used $q_i(\theta)$ (in direct generalization of the notation of Section 4) with the meaning of the present $p(a_i \mid \theta)$; the present notation is in line with our use of $p(x \mid \theta)$ all through the present chapter, a notation which emphasizes the observed (continuous or discrete) x opposed to the parameter θ.

The computation is as follows:

$$\log L_0 = n \sum_{i=1}^{k} p(a_i \mid \theta_0) \log p(a_i \mid \theta), \tag{104}$$

$$\frac{d}{d\theta} \log L_0 = n \sum_{i=1}^{k} p(a_i \mid \theta_0) \frac{p'(a_i \mid \theta)}{p(a_i \mid \theta)}, \tag{104'}$$

$$\frac{d^2}{d\theta^2} \log L_0 = n \sum_{i=1}^{k} p(a_i \mid \theta_0) \Big[\frac{p''(a_i \mid \theta)}{p(a_i \mid \theta)} - \frac{p'^2(a_i \mid \theta)}{p^2(a_i \mid \theta)} \Big]. \tag{104''}$$

Now from

$$\sum_{i=1}^{k} p(a_i \mid \theta) = 1, \tag{105}$$

valid for any θ, follows

$$\sum_{i=1}^{k} p'(a_i \mid \theta) = 0, \qquad \sum_{i=1}^{k} p''(a_i \mid \theta) = 0. \tag{105'}$$

Hence, for $\theta = \theta_0$:

$$\frac{d}{d\theta} \log L_0 = 0, \qquad \frac{d^2}{d\theta^2} \log L_0 = 0 - n \sum_{i=1}^{k} \frac{p'^2(a_i \mid \theta_0)}{p(a_i \mid \theta_0)}. \tag{105''}$$

Hence, $L_0''(\theta_0)/L_0(\theta_0) \to -\infty$, as $n \to \infty$. Therefore $L_0(\theta)$ has at $\theta = \theta_0$ a maximum and the property of concentration; that means that the c.d.f. approaches a unit step function. It may be shown that θ_0 is the only maximum of L_0.

Let us consider[6] the order of the difference between $\bar{\theta}$ and the true value θ_0. We set $p(a_i \mid \theta) = p_i$ and

$$u_i = n_i - np_i,$$

then $\sum_{i=1}^{k} u_i = 0$. We express in $\log L$ the p_i in terms of the u_i. Then from (103''):

$$N \equiv \log L = \sum n_i \log p_i = \sum n_i \log \frac{n_i - u_i}{n}$$

$$= \sum n_i \Big[\log\Big(1 - \frac{u_i}{n_i}\Big) + \log r_i \Big].$$

[6] See van der Waerden [28], §47.

Here the term $\sum n_i \log r_i$ does not contain θ and we may consider the following expression equivalent to $\log L = N$:

$$M = \sum n_i \log\left(1 - \frac{u_i}{n_i}\right) = \sum n_i \left(-\frac{u_i}{n_i} - \frac{1}{2}\frac{u_i^2}{n_i^2} - \cdots\right)$$

$$= -\sum \left(\frac{1}{2}\frac{u_i^2}{n_i} + \frac{1}{3}\frac{u_i^3}{n_i^2} + \cdots\right).$$

This is so far an identity in θ. If we introduce in p_i the true θ_0 then u_i, the difference between the observed n_i and its expected value, is with arbitrarily great probability of the order $\sqrt{n}$. Now if u_i is of the order $\sqrt{n}$ at most, then u_i/n_i is of the order $1/\sqrt{n}$ at most, and u_i^2/n_i at most of the order one, and this holds also for M. Therefore for some appropriate positive constant d we have with arbitrarily great probability $-M \leqslant d$ or

$$M \geqslant -d. \tag{106}$$

If, on the other hand, some u_i is of a higher order of magnitude than $\sqrt{n}$, then u_i/n is great compared to $1/\sqrt{n}$ and $-M$ is great compared to 1, hence the inequality opposite to (106) holds, namely, $M < -d$.

Now since for the true θ_0 and the corresponding p_i^0 and M_0 the inequality (106) holds, it follows that for the $\tilde{\theta}$ which maximizes $\log L$, and therefore M, (106) holds *a fortiori*.

Therefore not only the u_i^0 but also the $\tilde{u}_i = n_i - n\tilde{p}_i$ cannot be of a higher order than $\sqrt{n}$. Hence $p_i^0 - r_i$ as well as $\tilde{p}_i - r_i$ are each in probability of the order $1/\sqrt{n}$. And therefore, the $\tilde{p}_i - p_i^0$ *are with very great probability of the order* $1/\sqrt{n}$. On account of the assumed continuity of θ with respect to p it follows that the difference $\tilde{\theta} - \theta_0$ is also arbitrarily small (in probability) and because of the assumed differentiability it even follows that this difference is with arbitrarily great probability of the order $1/\sqrt{n}$. We add to our statement (*d*) of p. 552:

(d') *The differences* $\tilde{p}_i - p_i^0$ *as well as* $\tilde{\theta} - \theta_0$ *are with probability arbitrarily close to one of the order* $1/\sqrt{n}$. We assumed in our proof that *there are only* k *attributes* $a_1, \ldots, a_k$.

In the case where $p(x \mid \theta)$ is a density and $\int p(x \mid \theta)\, dx = 1$ the first part of the considerations is similar to the corresponding ones. Compare, as in Eqs. (105),

$$\log L = n \int p(x \mid \theta)\, dS_n(x)$$

with

$$\log L_0 = n \int p(x \mid \theta)\, dP(x \mid \theta_0)$$

and obtain the maximum of L_0 at $\theta = \theta_0$:

$$\frac{\partial}{\partial \theta} \log L_0 = n \int \frac{p'(x \mid \theta)}{p(x \mid \theta)}\, p(x \mid \theta_0)\, dx = 0 \qquad \text{for} \qquad \theta = \theta_0$$

$$\frac{\partial^2}{\partial \theta^2} \log L_0 = n \int \left[\frac{p''(x \mid \theta)}{p(x \mid \theta)} - \frac{p'^2(x \mid \theta)}{p^2(x \mid \theta)}\right] p(x \mid \theta_0)\, dx$$

$$= -n \int \frac{p'^2(x \mid \theta_0)}{p(x \mid \theta_0)}\, dx \to -\infty \qquad \text{for} \qquad \theta = \theta_0 .$$

For a proof of consistency in the case of a density refer to Wald, footnote 4, p. 551.

Reviewing some of the results obtained so far for the m.l. method we state:

1. If the number of observations, n, is small and the prior probability $p(\theta)$ is known, Bayes' formula gives all the information available about θ. The maximum likelihood method supplies nothing unless $p(\theta)$ is known to be uniform.

2. If n is small and it is known that the prior probability is uniform, the maximum likelihood method, cautiously applied, gives the θ value that has the greatest probability density (i.e., if the point is really a maximum, if it is the absolute maximum when there are more than one, and if no border value has a greater probability density).

3. If n is very large, then, under weak restriction, the value $\tilde{\theta}$ supplied by the maximum likelihood method is the most probable value and arbitrarily close, in probability, to the unknown true value θ_0.

We add the following remark. The consistency proof stated here for one parameter θ holds in the same way in the case of m parameters. However, *if the number of unknown parameters increases together with n, the number of observations, the consistency property of the estimate may get lost.* Consider the following example (see van der Waerden, [28] p. 151) where the likelihood method does not lead to a consistent estimate: n magnitudes with unknown true values $\alpha_1, \alpha_2, \ldots, \alpha_n$ have been measured, each twice. All $2n$ measurements are supposed to be normally and independently distributed with the same unknown variance σ^2. Denote by x_i, y_i, $i = 1, 2, \ldots, n$ the $2n$ results; we have

$$L = \prod_{i=1}^{n} p(x_i, y_i \mid \alpha_i, \sigma^2) = \sigma^{-2n}(2\pi)^{-n} \exp\left\{-\frac{\sum_{i=1}^{n} (x_i - \alpha_i)^2 + (y_i - \alpha_i)^2}{2\sigma^2}\right\} .$$

The maximum likelihood estimate of the α_i is the average

$$\tilde{\alpha}_i = \tfrac{1}{2}(x_i + y_i).$$

Substituting this into L we obtain

$$(2\pi)^n L = \sigma^{-2n} \exp\left\{-\sum \frac{(x_i - y_i)^2}{4\sigma^2}\right\}.$$

We obtain the maximum with respect to σ^2 by logarithmic differentiation and find

$$\tilde{\sigma}^2 = \frac{1}{4n}\sum (x_i - y_i)^2.$$

The expectation of $s^2 = (1/2n)\,\Sigma(x_i - y_i)^2$ equals σ^2, while

$$E[\tilde{\sigma}^2] = \tfrac{1}{2}\sigma^2.$$

Hence, the maximum likelihood estimate of σ^2 is highly "biased" and even as $n \to \infty$, the first Eq. (102) does not hold.

6.4. *Asymptotic distribution of maximum likelihood estimates.* Let us turn now to the asymptotic distribution of the m.l. estimate $\tilde{\theta}$. We have seen in Chapter IX, Section 5.2. (p. 465 ff) that $\tilde{\theta}$ is asymptotically normal.[7] The method used there was to establish the distribution of θ', the solution of the "linearized" Eqs. (56) of Chapter IX, and to show that, *asymptotically, $\tilde{\theta}$ has the same distribution as its linear approximation θ'*. We give here a direct computation of *the asymptotic mean value and variance of* $\tilde{\theta}$ [the latter stated in Chapter IX, Eq. (64′) without complete proof] based on two useful general formulas.

Assume an *arithmetical distribution* with k attributes $a_1, \ldots, a_k$. In n observations the result a_i has appeared $n_i = r_i n$ times, $i = 1, 2, \ldots, k$, $\Sigma_{i=1}^k r_i = 1$. Denote by $f(r_1, r_2, \ldots, r_k)$ a function of the relative frequencies r_i (this f may or may not be an estimate). Let $\bar{\theta} = \lim_{n\to\infty} E[f]$, and $\sigma^2 = \lim_{n\to\infty} \mathrm{Var}(f)$, and, let π_i denote the theoretical probabilities which correspond to the r_i; then $\Sigma_{i=1}^k \pi_i = 1$. In our present problem $\pi_i = p(a_i \mid \theta_0) = p_i(\theta_0)$. Denote by f_i the partial derivative $\partial f/\partial r_i$, where the r_i are replaced after differentiation by the π_i. Then the following easily understood formulas hold:

$$\bar{\theta} = f(\pi_1, \pi_2, \ldots, \pi_n), \qquad \sigma^2 \sim \frac{1}{n}\Big[\sum_{i=1}^k f_i^2 \pi_i - \Big(\sum_{i=1}^k f_i \pi_i\Big)^2\Big]. \tag{107}$$

[7] The likelihood defined here in Eq. (103) reduces, in the arithmetical case to Eq. (103′) and to Eq. (57) of Chapter IX.

These formulas will be proved in Chapter XI, last section [Eqs. (41) and (42)].

Now consider our specific problem. We set as abbreviations:

$$\begin{aligned} &p(a_i \mid \theta) = p_i(\theta) = p_i\,, \qquad \frac{d}{d\theta}(\log p_i) = \frac{1}{p_i}\frac{dp_i}{d\theta} = q_i(\theta) = q_i\,, \\ &p(a_i \mid \theta_0) = p_i(\theta_0) = \pi_i\,, \qquad i = 1, 2, \ldots, k. \end{aligned} \tag{108}$$

Then, as in (103″)

$$N = \log L = n \sum r_i \log p_i\,, \qquad \frac{d \log L}{d\theta} = n \sum r_i q_i\,.$$

We have for all θ:

$$\begin{aligned} &\sum_{i=1}^{k} p_i = 1, \qquad \sum_{i=1}^{k} p_i' = \sum_{i=1}^{k} q_i p_i = 0 \\ &\sum_{i=1}^{k} q_i' p_i = -\sum_{i=1}^{k} q_i p_i' = \sum_{i=1}^{k} p_i q_i^2. \end{aligned} \tag{109}$$

The likelihood estimate $\bar{\theta}$ is defined as a function $f(r_1, r_2, \ldots, r_k)$ of the relative frequencies $r_1, r_2, \ldots, r_k$ by

$$\sum_{i=1}^{k} r_i q_i(\bar{\theta}) = 0. \tag{110}$$

The first of Eqs. (107) shows that if in (110) the r_i are replaced by the true probabilities $p_i(\theta_0) = \pi_i$, then $\bar{\theta}$ is to be replaced by $\bar{\theta}$ and we obtain

$$\sum_{i=1}^{k} \pi_i q_i(\bar{\theta}) = 0; \tag{111}$$

this, in turn is satisfied for $\bar{\theta} = \theta_0$ on account of the second of Eqs. (109). Hence $\bar{\theta} = \theta_0$; the asymptotic mean value equals the true value, in agreement with the previously established property of consistency.

We have now to evaluate the second formula (107).

From Eq. (110) we compute $\partial\bar{\theta}/\partial r_\lambda$. We find using subscript for differentiation $q_\lambda(\bar{\theta})\, dr_\lambda + \Sigma_{i=1}^{k} r_i q_i'\, d\bar{\theta} = 0$, or

$$\frac{\partial \bar{\theta}}{\partial r_\lambda} = -\frac{q_\lambda(\bar{\theta})}{\Sigma_i r_i q_i'(\bar{\theta})}, \tag{112}$$

and when $\bar{\theta}$ is replaced by θ_0 and r_i by π_i we obtain,

$$f_\lambda = -\frac{q_\lambda(\theta_0)}{\sum_i \pi_i q_i'(\theta_0)}.$$

It follows, again from (109²), that,[8] with $q_i(\theta_0) = q_i$:

$$\sum_{\lambda=1}^{k} f_\lambda \pi_\lambda = -\frac{\sum_\lambda \pi_\lambda q_\lambda}{\sum_\lambda \pi_\lambda q_\lambda'} = 0$$

and, if (109³) is used

$$\sum_\lambda f_\lambda^2 \pi_\lambda = \frac{\sum q_\lambda^2 \pi_\lambda}{(\sum \pi_\lambda q_\lambda')^2} = \frac{\sum q_\lambda^2 \pi_\lambda}{(-\sum \pi_\lambda q_\lambda^2)^2} = \frac{1}{\sum q_\lambda^2 \pi_\lambda}.$$

Thus, from the second Eq. (107)

$$\frac{1}{c^2} = \frac{1}{n\sigma^2} = \sum_{i=1}^{k} q_i^2 \pi_i = \sum_{i=1}^{k} \frac{p_i'^2(\theta_0)}{p_i(\theta_0)}. \tag{113}$$

This result agrees with Eq. (64′), Chapter IX. The quotient $n/c^2 = 1/\sigma^2 = I$, is also called the *information* or *population. The likelihood estimate* $\bar{\theta}$ *is asymptotically normal with mean value* θ_0 *and variance* σ^2 *where* $1/n\sigma^2 = \sum_{i=1}^{k} [p_i'(\theta_0)]^2/p_i(\theta_0)$. *It follows that* $\sqrt{n}(\bar{\theta} - \theta_0)$ *is asymptotically normal with mean value zero and variance* c^2.

In the case of a density Eq. (113) is replaced by

$$\frac{1}{c^2} = \frac{1}{n} I = \frac{1}{n\sigma^2} = \int \left(\frac{p'}{p}\right)^2 p\, dx = \int \left(\frac{\partial \log p}{\partial \theta}\right)^2 p\, dx. \tag{113′}$$

We shall find this expression again (p. 561) in another connection. There we shall see that the asymptotic variance of the maximum likelihood estimate has a certain minimum property.[9]

We found, here and in Chapter IX, the asymptotic normality of the m.l. estimate $\bar{\theta}$ with mean θ_0 and reciprocal variance equal to population information. From the considerations at the end of Chapter VII it also follows that under the conditions given there *the posterior distribution*

[8] Writing $q_i(\theta_0) = q_i^0$ would make the following lines too clumsy.

[9] In Chapter IX, Eq. (64′) we computed the variance of the asymptotic distribution of a statistic F (function of the observations). Here, the second Eq. (107) gave the $\lim_{n\to\infty}$ Var(F) and the two results are seen to coincide. This is not necessarily so: the exact distribution of F may have infinite variance as $n \to \infty$ (or even for all n) and nevertheless the asymptotic distribution of F may have a finite variance.

of the (unknown) true θ_0 is asymptotically normal with $\tilde{\theta}$ as mean value and reciprocal variance equal to the sample information.

Problem 15. Find the maximum likelihood estimate in the cases

$$p(x \mid \theta) = \theta e^{-\theta x} \quad \text{and} \quad p(x \mid \theta) = 2\theta x e^{-\theta x^2}; \quad x \geqslant 0, \quad \theta > 0.$$

Problem 16. Let $p(x \mid \theta) = \dfrac{\sqrt{\theta} + (\sqrt{|x - \theta|^3})/(x - \theta)}{\sqrt{\theta} + \sqrt{1 - \theta}}$; $0 \leqslant x \leqslant 1$, $0 \leqslant \theta \leqslant 1$. Show that for any x, two points of minimum likelihood exist.

7. Further Remarks on Estimation

7.1. *Unbiasedness. Consistency.* Some other definitions connected with the estimation problem will be briefly recorded. We may look for such estimates $t(x)$ of θ whose expectation equals θ, for all n:

$$E[t(x)] = \theta. \tag{114}$$

Functions fulfilling (114) have been called *unbiased estimates* of θ. Examples are provided by our computations in Chapter VIII, Section 3. There it was learned that if a sample of n is taken from a population with mean value α and variance σ^2, the expectations of the sample average a and the sample dispersion s^2 are

$$E[a] = \alpha, \qquad E[s^2] = \frac{n-1}{n}\sigma^2 \qquad \text{or} \qquad E\left[\frac{ns^2}{n-1}\right] = \sigma^2. \tag{115}$$

Thus a is an unbiased estimate of α and $ns^2/(n-1)$ is an unbiased estimate of σ^2 for any kind of distribution and any n. If the distribution consists of an m times repeated alternative with unknown event probability q we have found (Chapter IX, Section 1.1) two unbiased estimates of $\sigma^2 = mq(1-q)$. They are [see Eq. 16), Chapter IX]

$$\frac{n}{n-1}s^2 \qquad \text{and} \qquad \frac{nm}{nm-1}a\left(1 - \frac{a}{m}\right). \tag{116}$$

In this sense, the Lexis theory can be interpreted as a comparison of two unbiased estimates of the same parameter. The property of unbiasedness is of questionable value. It lacks, in general, transitivity. We know, for example, that $s^2 = [1/(n-1)] \sum_{i=1}^{n} (x_i - \bar{x})^2$ is an unbiased estimate of σ^2 for any given d.f. Consider, on the other hand, a normal distribu-

tion, $N(\alpha, \sigma)$, and $s = +\sqrt{s^2}$; s is not an unbiased estimate of σ; computation shows (see Problem 18) that $E[s] = \sigma\sqrt{\frac{2}{n-1}}\,\Gamma\left(\frac{n}{2}\right)/\Gamma\left(\frac{n-1}{2}\right)$.

The more important properties of estimates relate to the case of *large* samples. The distribution $p(x \mid \theta)$ should concentrate, more and more around θ_0 with increasing n, as proved for the likelihood estimate. We recall the definition of *consistency* given at the beginning of Section 6.3.

A question often raised—whether s^2 or $ns^2/(n-1)$, i.e., the maximum likelihood estimate or the unbiased estimate of σ^2 is "better"—does not make too much sense. Both are consistent. In fact, let μ_4 be the fourth moment about the mean of a distribution. Then

$$E\left[\frac{n}{n-1}\,s^2\right] = \sigma^2, \qquad \mathrm{Var}\left[\frac{n}{n-1}\,s^2\right] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\,\sigma^4\right). \tag{117}$$

Since the last expression tends to zero as $n \to \infty$ and $n/(n-1) \to 1$ it is seen that for both statistics, Eqs. (102) hold. If for the actual purpose of an investigation the difference between the two estimates has any importance, this indicates that a larger sample should have been chosen.

7.2. *Minimum variance estimates.* We remember that the inference problem consists in deriving probability statements regarding theoretical assumptions from results of observations. The maximum likelihood estimate turned out to be interesting in various ways, from this point of view. On the other hand, the mere determination of a parameter value according to some "optimum principle" is not our aim.

From another point of view one may consider the task of point estimation to determine estimates $t(x)$ of θ, functions of the observations $x = (x_1, x_2, \ldots, x_n)$, which have certain desirable properties and are used on account of such properties. An example of this approach is the definition of an unbiased estimate (p. 559). In these last subsections, a few further concepts and results regarding point estimation and related questions will be briefly treated with a mere indication of proofs.

If for an estimation $t(x)$ of θ we set

$$E[t(x)] = \bar{t} = \int t p(x \mid \theta)\, dx$$

and

$$\bar{t} - \theta = b(\theta), \tag{118}$$

we call $b(\theta)$ *the bias of* t. The unbiased estimate has the property $b = 0$. Another characterization of estimates is, of course, the variance of $t(x)$;

$$\mathrm{Var}[t(x)] = \sigma_t^2 = E[(t - \bar{t})^2]. \tag{119}$$

One would like to determine an estimate whose bias and whose variance are both as small as possible. Accordingly, one seeks among all estimates with a given bias the estimate with minimum variance. For a certain important class of distributions the answer is contained in an interesting inequality found by Fréchet, Rao, and Cramér[1] which is often also designated the information inequality.

Denote, as before $\log L(x \mid \theta)$ by $N(x \mid \theta)$. If the x_i are mutually independent, $N(x \mid \theta) = \sum_{\nu=1}^{n} \log p(x_\nu \mid \theta)$. *Under certain regularity assumptions and if the information* $I(\theta) = E[N'^2] \neq 0$ *Fréchet's inequality or the information inequality holds*:

$$\sigma_t^2 \geqslant \frac{[1 + b'(\theta)]^2}{E[N'^2]} = \frac{[1 + b'(\theta)]^2}{I(\theta)}, \tag{120}$$

which, for an unbiased estimate, reduces to

$$\sigma_t^2 \geqslant \frac{1}{E[N'^2]} = I^{-1}. \tag{120'}$$

The derivation of (120) is simple: essentially, it uses the Schwarz inequality. For this and the following, see the presentation in [28], pp. 157-179 and a book by Rao.[2] The denominator in (120) is the previously [Eq. (113)] introduced information $I(\theta) = E[N'^2]$ which, with a continuous distribution, may be written in various equivalent forms:

$$\begin{aligned} I(\theta) = E[N'^2] &= n \int (\log p(z \mid \theta))' \, p' \, dz = n \int \left(\frac{\partial \log p}{\partial \theta}\right)^2 p \, dz \\ &= nE\left[\left(\frac{\partial \log p}{\partial \theta}\right)^2\right] = -nE\left[\frac{\partial^2 \log p}{\partial \theta^2}\right] \end{aligned} \tag{121}$$

as found previously in (113').[3]

We ask now: *under what circumstances does the equality sign hold in* (120) *or* (120')? We assume that $p(x \mid \theta) = p(x_1, x_2, \ldots, x_n \mid \theta)$ satisfies

[1] M. FRÉCHET, "Sur l'extension de certains évaluations statistiques au cas des petits échantillons." *Rev. Inst. Internat. Statist.* **11** (1943), pp. 185-205.

G. DARMOIS, "Sur les limites de la dispersion de certaines estimations." *Ibid.* **13** (1945), pp. 9-15.

C. R. RAO, "Information and accuracy obtainable in one estimation of a statistical parameter." *Bull. Calcutta Math. Soc.* **37** (1945), pp. 81-91.

H. CRAMÉR, *Skand. Akt.* **29** (1946), pp. 85-94.

[2] C. R. Rao, *Advanced Statistical Methods in Biometric Research.* New York, 1952.

[3] The name "information" relates to the fact that I is additive: if several independent observations are combined the corresponding information is the sum of the individual ones.

certain fairly general regularity conditions such that the inequality (120) holds. It can then be shown without difficulty that the equality sign is valid if and only if $p(x \mid \theta) = p(x_1, \ldots, x_n \mid \theta)$ is of the form

$$p(x \mid \theta) = e^{A(\theta)t+B(\theta)}h(x). \tag{122}$$

If $p(x \mid \theta)$ is such that (120) *holds, then the resolution of $p(x \mid \theta)$ into the product form* (122) *is necessary and sufficient for the estimate t to have the smallest variance among all estimates with the same bias.* In fact it is easily seen that if (122) holds, the equality sign holds in (120) for $t(x)$, while for all other estimates the $\geqslant$ sign is valid; on the other hand, from (120) with the equality sign, (122) follows.

An example of an unbiased estimate with minimum variance is the average $\bar{x}$ of n observations as an estimate of the mean value α of a *Gaussian* with known σ^2. The minimum variance of $\bar{x}$ is σ^2/n. Likewise if in a normal distribution α is known and σ^2 unknown then $s^2 = (1/n)\sum_{\nu=1}^{n}(x_\nu - \alpha)^2$ is a minimum-variance estimate of σ^2. The minimum variance is $2\sigma^4/n$.

There is a relation between the unbiased minimum variance estimate (equality sign in the information inequality) and likelihood estimates: *if such an estimate t of θ exists it can be found as a solution of the likelihood equation* (on account of the form of p in this case).

We note that it may happen that a minimum variance is still quite large. In the two given examples the respective variances go to zero as $n \to \infty$. If, however, n is small a consistent estimate may have a considerable variance.

We have seen that there is a close connection between "equality signs in (120)" and the "product-form" (122) of $p(x \mid \theta)$. If this particular product form does not hold, the equality sign *cannot* hold in (120). Nevertheless, a minimum-variance estimate may exist—with a different lower bound.

7.3. *Sufficiency.* These considerations are connected with the concept of sufficiency, suggested already by Fisher and elaborated by Neyman, Rao, Lehman, Scheffé, and others. We write the product form (122) more generally,

$$p(x \mid \theta) = f(t, \theta)h(x), \tag{123}$$

where $t = t(x)$ does not depend on θ. If (123) holds for t one calls $t = t(x)$ an exhaustive or sufficient estimate of θ. Since, however, t need not be an estimate of θ we call any $t(x)$ satisfying (123) a *sufficient* or *exhaustive statistic* for the family of distributions $p(x \mid \theta)$.

The factorization (123) is equivalent[4] to the following more suggestive property. Consider a family of distributions $p(x \mid \theta) = p(x_1, x_2, \ldots, x_n \mid \theta)$, each member being characterized by a value of θ. Consider a statistic $t(x) = t(x_1, x_2, \ldots, x_n)$ and its distribution (density) $p_1(t(x) \mid \theta)$ for any given θ. The quotient $p(x \mid \theta)/p_1(t \mid \theta)$ is the conditional density of the x, given $t(x)$. In general, this conditional density is different for each particular member of the family; it depends on θ. We call $t(x)$ *sufficient* for a family if the conditional density of the x, given $t(x)$, is the same whatever member of the family is considered. Thus: *The statistic $t = t(x)$ is called sufficient with respect to the family of distributions $p(x \mid \theta)$ of x, given θ, if the conditional density of x, given $t(x)$ and θ, does not depend on θ.* That means that all the information regarding θ that can be drawn from the sample $(x_1, x_2, \ldots, x_n)$ is contained in $t(x)$ alone; a sufficient statistic "exhausts the information" of a sample.

We consider what is probably the simplest example, and one that is particularly suggestive, since it deals with probabilities rather than with densities, the "simplest alternative" with probability θ for $x = 1$ (success) and $1 - \theta$ for $x = 0$ (failure): $p(x \mid \theta) = \theta^x(1 - \theta)^{1-x}$, $x = 0, 1$, $p(0 \mid \theta) + p(1 \mid \theta) = 1$. Then

$$\begin{aligned} p(x_1, x_2, \ldots, x_n \mid \theta) &= p(x_1 \mid \theta)p(x_2 \mid \theta) \cdots p(x_n \mid \theta) \\ &= \theta^{x_1}(1 - \theta)^{1-x_1} \cdots \theta^{x_n}(1 - \theta)^{1-x_n}. \end{aligned}$$

Let $t(x) = x_1 + x_2 + \cdots + x_n$; the probability distribution of t, the binomial distribution, is $p_1(t \mid \theta) = \binom{n}{t}\theta^t(1 - \theta)^{n-t}$ and

$$\frac{p(x \mid \theta)}{p_1(t \mid \theta)} = \frac{\theta^t(1 - \theta)^{n-t}}{\binom{n}{t}\theta^t(1 - \theta)^{n-t}} = \frac{t!(n - t)!}{n!} = h_1(x).$$

This conditional distribution h_1 of the x_i, given t, is independent of θ. Thus $t = x_1 + \cdots + x_n$ is a sufficient statistic for θ.[5]

In general, as in this example, the $f(t, \theta)$ and $h(x)$ of (123) can be taken to be the probability (density) $p_1(t(x) \mid \theta)$ of t for given θ, and the conditional distribution $h_1(x)$ of x, given $t(x)$; for sufficient $t(x)$, the conditional distribution $h_1(x)$ is independent of θ.

The previously mentioned relation (end of Section 7.2) between minimum variance and sufficiency, if the particular form (122) of (123) does not hold, is partly clarified by a theorem due to Rao and Black-

[4] J. NEYMAN, *Giorn. Ist. Ital. Attuari* **6** (1934), p. 330.

[5] It has actually minimum variance $n\theta(1 - \theta)$ among unbiased estimates and (122) holds, i.e., the special form of (123), hence equality sign in (120).

well.[6] Let $t(x)$ be a sufficient estimate of θ and $T(x)$ be some other estimate of θ with finite expectation and variance. We denote by $\psi(t) = E[T \mid t]$ the conditional expectation of T, given t. This expectation depends, in general, on θ, but not if $t(x)$ is sufficient. The theorem states that $\psi(t)$ *has the same bias as* T *but a smaller variance than* T *unless* T *is a function of* t.

$$\sigma_T^2 \geqslant \mathrm{Var}[\psi(t)]. \tag{125}$$

Here $\psi(t)$ itself is an expectation. The definition and use of the conditional expectation, $\psi(t)$, which is immediate if t is a *discontinuous* random variable, needs deeper investigation in the general case.

We add finally that recently a definition of sufficiency in the Bayesian sense has been given.[7] The statistic $t(x)$ is called sufficient if for all prior distributions $P(\theta)$ the posterior distribution of θ, given x, is equal to that of θ, given t:

$$q(\theta \mid t) = q(\theta \mid x);$$

i.e., the posterior distribution $q(\theta \mid x)$ is a function of θ and $t(x)$ alone. The authors show that this definition is equivalent to those above.

As an example of a sufficient statistic for which (123) but not the particular form (122) holds, let $x_1, x_2, \ldots, x_n$ be normally distributed with $\alpha = 0$ and unknown σ. The unbiased estimate of σ

$$z = \frac{1}{\sqrt{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} \left(\sum_{\nu=1}^{n} x_\nu^2\right)^{1/2}, \qquad E[z] = \sigma \tag{126}$$

is sufficient but does not satisfy the information equality.

At this point, we finish our discussion of this group of problems. To present them adequately would take up more space than is available in the framework of this book.

Problem 17. Consider $p(x \mid \theta) = \theta^x e^{-\theta}/x!$, $x = 0, 1, 2, \ldots$ and prove that $t = (\sum_{\nu=1}^{n} x_\nu)/n$ is a sufficient and unbiased estimate of θ.

Problem 18. We know that $s^2 = [1/(n-1)] \sum_{\nu=1}^{n} (x_\nu - \bar{x})^2$ is an un-

[6] See Rao's above quoted paper (footnote 1, p. 561) and D. BLACKWELL, *Ann. Math. Statist.* **18** (1947), pp. 105-110.

[7] Raiffa and Schlaifer, *Applied Statistical Decision Theory*. Division of Research, Harvard Business School, 1961.

biased estimate of the variance σ^2. Prove that even in the case of a normal distribution, s is not an unbiased estimate of σ. Prove that

$$E[s] = \sigma \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}. \tag{127}$$

Problem 19. Prove the statement of the text regarding the statistic (126). Why are Eqs. (126) and (127) consistent?

Problem 20. Prove that under some regularity assumptions, with the notation of the text

$$1 + b' = E[N'(t - \tilde{t})].$$

Use this to derive the inequality (120).