CHAPTER VII

PROBABILITY INFERENCE. BAYES' METHOD

A. Inference from a Finite Number of Observations (Sections 1-2)

1. Bayes' Problem and Solution

1.1. *The problem.* If a die is thrown six times and deuce appears twice, nobody would suppose that the probability of throwing a deuce is $\frac{2}{6}$ or near to $\frac{2}{6}$. One would not even suspect, on the basis of six trials, that the die is not unbiased or, in other words, that the probability of a deuce is different from $\frac{1}{6}$. On the other hand, if not six but 600 trials are made and deuce appears 200 times, the assumption prompts itself that the die is loaded in such a way as to favor considerably the deuce, i.e., as to make the probability of a deuce significantly different from $\frac{1}{6}$ and rather near to $\frac{2}{6}$. A sound mathematical basis for dealing with this problem was suggested by Thomas Bayes.[1] The aim is to make an inference on the unknown probability of deuce by means of the known outcome of 600 $(=n)$ trials. The theory is known under various headings —probability of hypotheses, inverse probability, probability of inference, etc.

The idea leading to Bayes' solution is this: Suppose we have in a bag a collection of various dice—correct ones, nearly correct ones, and others that are loaded to various degrees. According to our definition of probability, to each die corresponds a certain value $q = x(0 \leqslant x \leqslant 1)$, the probability of throwing the deuce with this individual die. The collection of dice in the bag is characterized by a certain distribution of the x values. What now happens is that (1) one die is drawn from the bag at random

[1] THOMAS BAYES, "An essay towards solving a problem in the doctrine of chances." *Phil. Trans. Royal Soc. London* **53** (1763), pp. 370–418. The paper was communicated to John Canton, F. R. S., by Richard Price after the death of Bayes (1761) who had left a number of manuscripts.

A facsimile of this paper with R. Price's foreword and discussion, and a commentary by E. C. Molina, was published by *E. Deming*, 1939, Washington, The Graduate School, the Department of Agriculture.

and (2) a number n of throws are executed with this die and the number n_1 of two-spots is observed. *We ask for the probability $q_n(x)$ that for the die which was used in the experiment, the unknown probability q of deuce has a certain value x.*

Opponents of Bayes' approach object that the idea of the "collection of dice with various x" is unrealistic since one deals with one single die whose x is an unknown constant, and not with a universe of dice. This objection is not to the point. By attempting an inference on the probability of the x of the die used in our experiment we imply that various values of x are possible. The "collection of dice in a bag" is a model, an abbreviating simplification; the "collection" may consist of the dice in the various stores of a city or of the United States, and we are uncertain about the x of our chosen die. (See further discussion in Ch. X.)

1.2. *Solution.* The description of the experiment indicates that we have before us the combination of two independent collectives, which leads to the multiplication of the two probabilities. The first factor in the product is a function $p_0(x)$, the probability of drawing from the bag a die for which the probability of deuce equals x. We consider $p_0(x)$ in this first example as an arithmetical distribution. The second factor will be the probability of obtaining y_1 times in n throwings of the same die the result "two" or, more generally, the probability of obtaining y_1 times in n trials the event whose probability is x. According to the formula which solves the Bernoulli problem this probability is $\binom{n}{y_1}(1-x)^{n-y_1}x^{y_1}$. This is formula (11) of Chapter IV with y_1 written for x, and x written for q and this probability can be considered a function of the chance variable y_1, depending on the parameter x. We therefore call it, for present purposes, $p_n(y_1 \mid x)$:

$$p_n(y_1 \mid x) = \binom{n}{y_1}(1-x)^{n-y_1}x^{y_1}. \tag{1}$$

The two probabilities $p_0(x)$ and $p_n(y_1 \mid x)$ fulfill the two conditions

$$\sum_x p_0(x) = 1; \qquad \sum_{x_1=0}^{n} \binom{n}{y_1}(1-x)^{n-y_1}x^{y_1} = 1. \tag{2}$$

The product of the two probabilities

$$p(x, y_1) = p_0(x)p_n(y_1 \mid x) \tag{3}$$

is, according to the multiplication rule, the probability of the combined event, namely, that of "drawing" a die with deuce-probability equal to x, and that of getting deuce y_1 times in n trials with this die. This probability depends on the two variables x and y_1.

Before making the next step in the solution of our problem, we insert two remarks. First, it was stated in Chapter I that the product formula (3) which, there, was written

$$p(x, y) = p'(x)p''(y \mid x) \tag{4}$$

also holds if $p'(x)$, now $p_0(x)$, is not a "probability" but a "chance"; that means that randomness is not required in the process of drawing the die from the bag. What we really do in the present problem, is to take a die from among a certain quantity of available dice and subject it to the n trials. We have only to assume that in the entirety of available dice, there exists a limiting frequency $p_0(x)$ for drawing a die whose q-value is x. But we need not assume that the succession of individual x-values fulfills the condition of randomness. If it does not, the product (3) must be considered as the *chance* of the combined event.

The second remark refers to the admissible values of x. It can be assumed that x is susceptible only of discrete values in the interval zero to one, and then each $p_0(x)$ is the chance of this individual x-value; or, we may assume that x may have all values between 0 and 1, and then $p_0(x)$ as well as the product (3) would be a density at the point x. An illustration of the first assumption would be the following. Instead of a die, let the object of the n trials be a ball taken out of a bag containing ten balls, white and black in an unknown proportion. The probability q of getting a white ball, under the usual circumstances, will then have one of the eleven values, $0, \frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}, 1$, and $p_0(x)$ will be supposed to be known for these eleven x-values. In the continuous case, the first equation (2) has to be replaced by

$$\int_0^1 p_0(x)\,dx = 1. \tag{2'}$$

Our problem is not solved by setting up the product (3). We do not ask for the probability (3) where y_1 may be any integer between 0 and n; i.e., we do not ask for the probability that an arbitrary die which gave any value of y_1 has a q-value equal to x; we are only interested in the probability of the q-value of a die for which $y_1 = n_1$. This last probability is obtained by applying the operation of partitioning to the collective whose probability is $p(y_1 \mid x)$. Remember the basic example of partitioning: if a die with probabilities $p_1, p_2, \ldots, p_6$ for the six faces has shown an even number, then the probability that the result was "two" is the quotient $p_2/(p_2 + p_4 + p_6)$. In the present case the two-dimensional probability $p(x, y_1)$ in (3) is given and we know that the value of y_1 is n_1. The conditional probability or inferred probability, $q_n(x)$ computed under the condition of given n_1, is the quotient $p(x, n_1)$ divided by the sum of all $p(x, n_1)$-values. Hence

$$q_n(x) = \frac{p_0(x)p_n(n_1|x)}{\sum_x p_0(x)p_n(n_1|x)} \tag{5}$$

if x is a discrete variable, or

$$q_n(x) = \frac{p_0(x)p_n(n_1|x)}{\int p_0(x)p_n(n_1|x)\,dx} \tag{5'}$$

in the case of a continuous x. *These formulas solve the Bayes problem* in its simplest form. The problem plays a basic role in theoretical statistics.[2] Equations (5) and (5′) give the probability or chance (or corresponding density) of an x value inferred from an observed number n_1 of successes in n trials. It will be seen later that, in agreement with what was said at the beginning of this section, important conclusions can be drawn from Eqs. (5) and (5′) if n is large.

Since the denominator does not depend on x, we can write formulas (5) and (5′) in the form

$$q_n(x) = \text{constant} \cdot p_0(x)p_n(n_1|x), \tag{6}$$

and simply add that the constant has to be determined from the condition that

$$\sum_x q_n(x) = 1 \qquad \text{or} \qquad \int_0^1 q_n(x)\,dx = 1. \tag{7}$$

If the value for $p_n(n_1 \mid x)$ is introduced from (1), the factor $\binom{n}{n_1}$ cancels out since it appears in the numerator as well as in the denominator. We thus obtain, say, in the case of continuous x,

$$\begin{aligned} q_n(x) &= \text{constant} \cdot p_0(x)(1-x)^{n-n_1}x^{n_1}, \\ \text{constant} &= 1 \div \int_0^1 p_0(x)(1-x)^{n-n_1}x^{n_1}\,dx. \end{aligned} \tag{8}$$

[2] The Bayes' approach as presented here in (5) or (5′) is also essential in the mathematical analysis of decision problems when x, the "state of the world" is considered as a chance variable. The "prior" knowledge regarding x is incorporated into a $p_0(x)$ and further information about x can be obtained by experimentation. See A. Wald, *Statistical Decision Functions*, New York, 1950. See H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*, Graduate School of Business Administration, Harvard University, 1961. Bayesian approach is often connected with a "subjective" or "personal" probability concept. *Our* probability concept is and remains "objective." Our Chapter X is based on Bayes' approach.

In the present example the limits 0 and 1 apply to the integral sign, since x is in this case a probability and, therefore, $p_0(x)$ must vanish outside the interval (0, 1). We see from (8) that only *that* part of $p_n(n_1 \mid x)$ which represents the probability of one particular sample will enter into $q_n(x)$; we may denote this part by $L_n(x)$. In our case we have $L_n(x) = x^{n_1}(1 - x)^{n-n_1}$.

Using the notation of the Stieltjes integral and introducing a c.d.f. $Q_n(x)$ corresponding to $q_n(x)$ we can write both cases (the discrete and the continuous) in the common form,

$$Q_n(x_2) - Q_n(x_1) = \text{constant} \int_{x_1}^{x_2} p_n(n_1|x)\, dP_0(x), \tag{9}$$

where $P_0(x)$ is a distribution function and the constant has to be determined from the condition that

$$Q_n(1) = 1. \tag{10}$$

1.3. *Discussion.* In formulas (5)–(9), two different probabilities (chances) appear for the same variable x; first, $p_0(x)$ referring to the choice of the object to be subjected to the n trials, a probability or chance that is independent of the outcome of the n trials, and second, $q_n(x)$, inferred (partly) from the outcome of the n trials. In order to distinguish them, $p_0(x)$ is often called the *a priori* and $q_n(x)$ the *a posteriori* chance of x. These terms have, of course, nothing to do with what philosophers understand by "*a priori.*" It is meant that $p_0(x)$ is the chance prior to the experiments and $q_n(x)$ the chance after n experiments have been carried out and the number n_1 is known. An alternative form of "*a priori*" chance or "prior" chance is "over-all" chance. Both $p_0(x)$ and $q_n(x)$ are frequency limits in definite sequences of observations. If all available dice or, more generally, all available objects of experimentation are scrutinized with respect to the value of some x, the limiting frequency of x is $p_0(x)$. Among all those objects which in n trials gave n_1 successes, the limiting frequency of those whose label value equals x is $q_n(x)$.

Let us illustrate the meaning of $q_n(x)$ and of $p_0(x)$ in terms of the example mentioned on p. 329. Suppose there are in this bag r different kinds of dice characterized by their different "deuce-probabilities" $q_1, q_2, \ldots, q_r$. If we ask for the *a posteriori* chance $q_n(x)$ of a particular x-value, we have in mind a sequence of trials each consisting of two steps—first picking a particular die out of the bag and then casting it n times. For the first N experiments of this kind, $N' \leqslant N$ is the number of those trials where in n throws the result "deuce" appeared n_1 times.

Among these N' trials there are N'_ρ, $\rho = 1, 2, ..., r$, for which the q-value equals q_ρ. Then, $\lim_{N\to\infty} N'_\rho/N'$, $\rho = 1, 2, ..., r$, constitutes the *a posteriori* distribution $q_n(x)$. Note that $N'_\rho/N' = (N'_\rho/N)/(N'/N) = (N'_\rho/N)/(\sum_{\rho=1}^{r} N'_\rho/N)$, corresponding to (5). On the other hand, denote by N_ρ the number of those dice (among the N) whose q equals q_ρ; the limit of N_ρ/N is the *a priori* probability; and $N'_\rho/N = (N_\rho/N) \cdot (N'_\rho N_\rho)$ corresponds to the numerator in (5).

It should be understood that, of course, formulas (5) and (5′) are not restricted to the case considered in this section. In the present discussion $p_n(y_1 \mid x)$ was a binomial distribution. But formulas (5) and (5′) can be used for drawing inference on an unknown parameter x from $p_0(x)$ and $p_n(y_1 \mid x)$ for the most varied $p_n(y_1 \mid x)$. In Chapter X we will deal with many forms of this distribution.

Problem 1. An urn holds 10 balls, partly white and partly black. The probability of drawing a white ball from an urn with $10x$ white balls is supposed to be x where x can take one of the eleven values 0, 0.1, 0.2, ..., 0.9, 1. Each of these values is supposed to be equally likely *a priori*, that is, $p_0(x) = 1/11$. Compute the probability that x has the value 0.4 if in 6 observations $n_1 = 2$ white balls appear.

Problem 2. Compute in the case of Problem 1 the probability that x is smaller (or greater) than $\frac{2}{6}$.

Problem 3. In 6 experiments with a certain object, the event whose probability equals x appeared 3 times. Assume that all values between 0 and 1 are equally likely *a priori*, $p_0(x) =$ constant. Compute the conditional probability density for x equal to $\frac{1}{3}$.

Problem 4. If it is known that an event has happened n_1 times in n observations one asks for the probability P that it will happen another m_1 times out of m. Prove the formula

$$P = \int_0^1 q_n(z)p_m(m_1|z)\,dz = \binom{m}{m_1}\frac{\int_0^1 p_0(z)z^{n_1+m_1}(1-z)^{n+m-n_1-m_1}\,dz}{\int_0^1 p_0(z)z^{n_1}(1-z)^{n-n_1}\,dz}.$$

Assuming constant *a priori* probability and $m = m_1 = 1$, prove that

$$P = \frac{n_1+1}{n+2} \qquad \text{(Laplace's "law of succession")}$$

2. Discussion of $p_0(x)$. Assumption $p_0(x) =$ constant

2.1. *Remarks on $p_0(x)$.* The main difficulty in applying the results of the foregoing section to an actual problem consists in the fact that knowledge of $p_0(x)$ is required. As long as we know nothing about $p_0(x)$, we cannot draw an inference about the probability $q_n(x)$ of an x-value after n experiments. This is in agreement with the fact that if nothing but the results of a small number n of trials is known, we would not expect to arrive at a substantial conclusion about the values of x. The situation for large n will be discussed in the following sections. At any rate, if we are to apply formulas (5) and (9) in the case of finite n, we have to have some knowledge regarding $p_0(x)$ or to make an assumption about the function $p_0(x)$.

The simplest assumption would be that all x are equally likely, that is, that the over-all probability $p_0(x)$ equals $1/m$ if m values of x are possible, or, that the over-all density $p_0(x)$ equals 1 if x can take all values between zero and one. In fact, in the original paper of Bayes only the case $p_0(x) =$ constant was considered. Sometimes one refers to an alleged "Bayes principle" "Bayes' postulate" or the "principle of insufficient reason" that would state: If nothing is known about $p_0(x)$, one *must* assume $p_0(x) =$ constant. Whether this was really Bayes' point of view is irrelevant. Indeed, in investigating the correctness of a die, the assumption that the probability x of a certain face is likely to possess any value between zero and one does not make much sense. It is, in fact, very difficult to manufacture a solid body of cubical shape which could pass for an unbiased die and would fall almost always on the same face. Rather than assume $p_0(x)$ as uniform, it would be appropriate to assume for $p_0(x)$ comparatively large values in a certain small neighborhood of $x = \frac{1}{6}$ and values close to zero outside this interval.

2.2. *The case $p_0(x)$ = constant. Mean value and variance.* Nevertheless, the case $p_0(x)$ = constant deserves attention from a mathematical point of view, particularly as a preliminary step in the study of the more general case. If $p_0(x)$ has a constant value, this constant can be combined with the constant appearing in (8), which amounts to writing (in the continuous case)

$$q_n(x) = C_n(1 - x)^{n-n_1}x^{n_1}, \qquad \frac{1}{C_n} = \int_0^1 (1 - x)^{n-n_1}x^{n_1}\,dx, \tag{11}$$

and a quite similar formula holds in the case of a discrete $p_0(x)$. To compute the integral in the denominator, we apply the following integration formula obtained by successive partial integrations:

$$\begin{aligned} I(a, b) &= \int_0^1 (1 - x)^a x^b\,dx = \frac{a}{b+1} I(a - 1, b + 1) \\ &= \frac{a(a-1)}{(b+1)(b+2)} I(a - 2, b + 2) = \cdots \\ &= \frac{a!}{(b+1)\cdots(b+a)} \int_0^1 x^{b+a}\,dx = \frac{a!b!}{(a+b+1)(a+b)!}\,. \end{aligned} \tag{12}$$

In the integral (11) we have $a = n - n_1$, $b = n_1$, and

$$I(n - n_1, n_1) = \frac{n_1!(n - n_1)!}{(n + 1)n!} = \frac{1}{(n + 1)\binom{n}{n_1}}, \qquad \text{and} \qquad C_n = (n + 1)\binom{n}{n_1}. \tag{13}$$

Thus

$$q_n(x) = (n + 1)\binom{n}{n_1}(1 - x)^{n-n_1}x^{n_1}; \tag{14}$$

this $q_n(x)$ is also called Bayes' distribution or, mathematically speaking, the Beta distribution with a slight redefinition of the parameters.[1]

To discuss this distribution, we first ask for the extreme values of $q_n(x)$. The logarithmic derivative equated to zero gives

$$\frac{n_1}{x} - \frac{n - n_1}{1 - x} = 0, \qquad x = \frac{n_1}{n}. \tag{15}$$

The function $q_n(x)$ vanishes at $x = 0$ and $x = 1$, these points being zeros of order n_1 and $n - n_1$, respectively. Since $q_n(x)$ is non-negative the point $x = n_1/n$ following from (15) is a *maximum.* The curve decreases monotonically to both sides of the maximum. Figure 19 shows $q_n(x)$, computed from Eq. (14), for $n_1/n = \frac{1}{3}$ and various values of n (the unit of x being proportional to $\sqrt{n}$).

We have the result: *If all x between 0 and 1 are a priori equally probable then the maximum of the conditional (a posteriori) probability is at n_1/n. If the prior probability is discrete and uniform the most probable (posterior) value x is n_1/n, if this is one of the a priori possible x-values or otherwise it is one of the two values which lie closest to n_1/n.*

The *mean value* of $q_n(x)$ is given by

$$a = \int_0^1 xq_n(x)\,dx = C\int_0^1 (1 - x)^{n-n_1}x^{n_1+1}\,dx = \frac{I(n - n_1, n_1 + 1)}{I(n - n_1, n_1)} = \frac{n_1 + 1}{n + 2}. \tag{16}$$

If both n and n_1 are large, this approaches n_1/n, that is, mean value and maximum value tend to coincide.

To find the *variance* we first compute the second zero-moment which is

$$\frac{I(n - n_1, n_1 + 2)}{I(n - n_1, n_1)} = \frac{(n_1 + 1)(n_1 + 2)}{(n + 2)(n + 3)}. \tag{17}$$

[1] In a usual notation $I(a, b)$ would be $B(b + 1, a + 1)$ where B is the beta function. Then, in Eq. (13), $(n + 1)\binom{n}{n_1} = 1/B(n_1 + 1, n - n_1 + 1)$.

From this we obtain the variance by subtracting a^2:

$$s^2 = \frac{(n_1 + 1)(n_1 + 2)}{(n + 2)(n + 3)} - \left(\frac{n_1 + 1}{n + 2}\right)^2 = \frac{(n_1 + 1)(n - n_1 + 1)}{(n + 2)^2(n + 3)} . \tag{18}$$

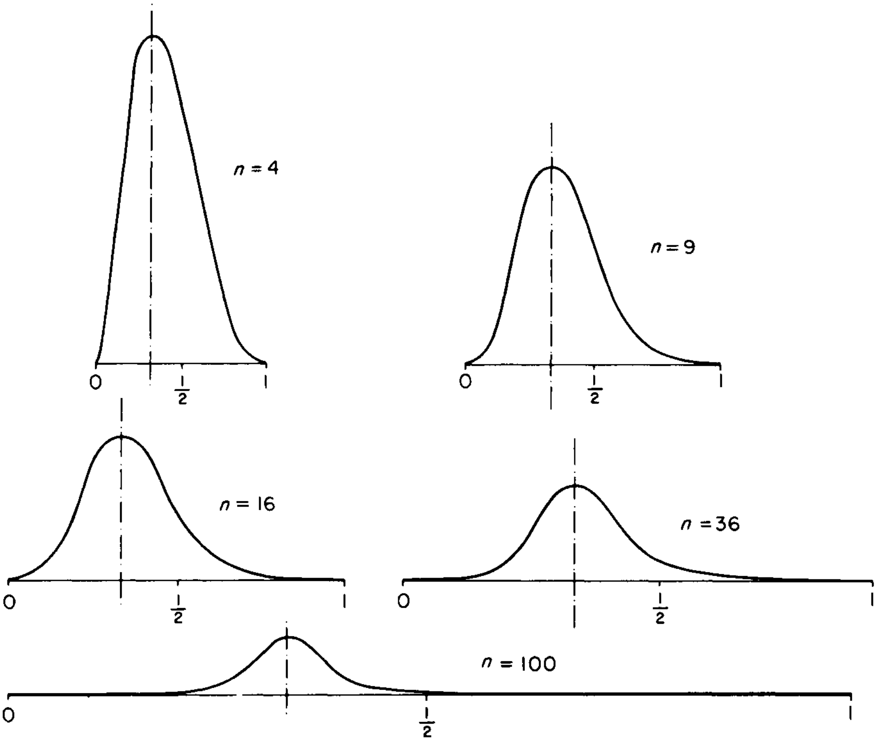


FIG. 19. Bayes' distributions. $n_1/n = \frac{1}{3}$.

It is seen that for large n and n_1 this approaches the value

$$s^2 \rightarrow \frac{1}{n}\frac{n_1}{n}\left(1 - \frac{n_1}{n}\right) . \tag{19}$$

2.3. *Analogies of Bayes' distribution to Bernoulli's distribution*. Some analogies between the distribution (14) and the Bernoulli distribution are worth mentioning. We consider the Bernoulli distribution with the relative number of successes $z = x/n$, as the independent variable [see Chapter IV, Eq. (26)]:

$$\binom{n}{nz}(1 - q)^{n-nz}q^{nz}. \tag{20}$$

Its mean value lies at $z = q$, the maximum value near to it and approaching it as n increases; the variance was found to be $(1 - q)(q/n)$. The quantity corresponding to z in (14) is $r = n_1/n$. The Bayes distribution has its maximum at $x = r$; its mean value approaches r as n (and n_1) increase while the variance tends toward $(1 - r)(r/n)$.

We also note that the c.d.f. of the binomial can be expressed in terms of the beta c.d.f. and vice versa. The following formula may be proved by integration by parts. With c a constant,

$$\sum_{x=c}^{n} \binom{n}{x} p^x (1 - p)^{n-x} = \frac{1}{B(c, n - c + 1)} \int_0^p x^{c-1}(1 - x)^{n-c}\, dx\,.$$

The most significant analogy, however, between the Bayes distribution (14) and the Bernoulli distribution (20) is the fact that the variance goes to zero as n increases indefinitely (with r or q kept constant). This means that with increasing n both distributions become more and more concentrated around the point whose abscissa is, at the same time, the mean value and the location of the maximum. The implications of this concentration in the case of the Bayes problem will be discussed in the next section for general $p_0(x)$.

In the special case of (14) the concentration also finds its expression in the value of the second derivative at the point $x = n_1/n = r$. The first derivative of (14) according to (15) is

$$\frac{dq_n}{dx} = q_n \left[\frac{n_1}{x} - \frac{n - n_1}{1 - x}\right], \tag{21}$$

and since the first derivative vanishes at $x = r$, the second derivative is there

$$\left.\frac{d^2q_n}{dx^2}\right]_{x=r} = q_n(r)\left[-\frac{n_1}{r^2} - \frac{n - n_1}{(1 - r)^2}\right] = -n(n + 1)\binom{n}{n_1}(1 - r)^{n-n_1-1} r^{n_1-1}. \tag{22}$$

If the Stirling formula is used to estimate $\binom{n}{n_1}$ and r is replaced by n_1/n it is found that the right-hand side of (22) becomes infinite as $n^{3/2}$. Thus, the curvature of the curve representing $q_n(x)$ goes to minus infinity and the radius of curvature at the top point goes to zero.

Problem 5. In 20 trials of an alternative the event has shown up 8 times. Supposing that the assumption of constant $p_0(x)$ is justified, compute the maximum value of the *a posteriori* density, the mean, and the variance of the *a posteriori* distribution. Using Tchebycheff's inequality give an estimate for the probability that x lies in an interval of width $2c$ symmetric with respect to the mean value.

Problem 6. If the assumption $p_0(x) = 1$ for $0 \leqslant x \leqslant 1$ is replaced by the assumption

$$p_0(x) = \frac{1}{2c}, \quad \text{for} \quad |x - r| \leqslant c$$
$$p_0(x) = 0, \quad \text{for} \quad |x - r| > c, \quad \text{where} \quad r > c, \quad 1 - r > c,$$

the value (14) of $q_n(x)$ is multiplied in $(r - c,\ r + c)$ by the quotient $\int_0^1 (1 - x)^{n-n_1} x^{n_1}\, dx / \int_{r-c}^{r+c} (1 - x)^{n-n_1} x^{n_1}\, dx$. Prove this statement and show that due to the concentration, this quotient tends toward 1 as n increases toward infinity for constant r.

Problem 7. How many trials have to be made to have at least 99% probability for the inference that x lies between 0.38 and 0.42, if the observed $r = n_1/n = 0.4$? Assume first that $p_0(x)$ is constant over the whole interval and then discuss the influence of a restriction of the interval, such as in the preceding problem. Use Tchebycheff's inequality.

B. Law of Large Numbers (Section 3)

3. Bayes' Theorem. Irrelevance of $p_0(x)$ for large n

3.1. *Derivation of Bayes' theorem.* In this section we shall discuss a certain aspect of the effect of n being a *large number* on the inference to the unknown value of x. It will be seen that the inference becomes more and more independent of the assumptions about $p_0(x)$ as n increases.

Let us divide the interval zero to one into two parts,[1] A and $\bar{A}$:

$$\begin{aligned} A&: \ |x - r| \leqslant \epsilon \\ \bar{A}&: \ |x - r| > \epsilon. \end{aligned} \tag{23}$$

Integrals over these two domains will be designated by $\int_A$ and $\int_{\bar{A}}$. The *a posteriori* probabilities of x falling into A and $\bar{A}$, respectively, may be denoted by Q_A and $Q_{\bar{A}}$. The same probabilities when computed under the assumption $p_0(x) = 1$ may be called Q_A' and $Q_{\bar{A}}'$. The variance of the distribution of x in the latter case, which we now call s'^2, has been computed in (18) and its value for large n given in (19).

[1] In previous chapters we used A' for the complement of A. Here, the present notation is more convenient.

From the Tchebycheff inequality we have

$$Q'_A \geqslant 1 - \frac{s'^2}{\epsilon^2}, \qquad Q'_{\bar{A}} \leqslant \frac{s'^2}{\epsilon^2}. \tag{24}$$

Hence

$$\frac{Q'_A}{Q'_{\bar{A}}} = \frac{\int_A p_n(n_1 \mid x)\, dx}{\int_{\bar{A}} p_n(n_1 \mid x)\, dx} \geqslant \frac{\epsilon^2}{s'^2} - 1. \tag{25}$$

Now we introduce the following assumptions about the *a priori* density $p_0(x)$:

(a) $p_0(x)$ is continuous and does not vanish at $x = r$,

(b) $p_0(x)$ has an upper bound M.

It follows from (a) that for a sufficiently small ϵ the density $p_0(x)$ has a minimum value m within A. Therefore, in view of (25)

$$\frac{Q_A}{Q_{\bar{A}}} = \frac{\int_A p_0(x) p_n(n_1 \mid x)\, dx}{\int_{\bar{A}} p_0(x) p_n(n_1 \mid x)\, dx} \geqslant \frac{m}{M} \cdot \frac{Q'_A}{Q'_{\bar{A}}} \geqslant \frac{m}{M}\left(\frac{\epsilon^2}{s'^2} - 1\right) = \lambda. \tag{26}$$

Due to $Q_A + Q_{\bar{A}} = 1$, we find

$$Q_A \geqslant \lambda(1 - Q_A), \qquad \text{or} \qquad Q_A \geqslant \frac{\lambda}{1+\lambda}. \tag{27}$$

Equations (18) and (19) show that $1/s'^2$ goes to ∞ with increasing n and so does λ, according to its definition in (26), for any given ϵ. Thus (27) expresses the fact

$$\lim_{n\to\infty} Q_A = 1, \qquad \text{or} \qquad \lim_{n\to\infty} \int_{r-\epsilon}^{r+\epsilon} q_n(z)\, dz = 1, \tag{28}$$

independent of the (small) ϵ used to determine the interval A. We formulate (28) in the following statement:

If the observation of an n-times repeated alternative shows a relative frequency r of "success," then, if n is sufficiently large, the chance that the probability of success lies between r — ϵ *and r* + ϵ *is arbitrarily close to one, no matter, how small the* ϵ. In other words, somewhat less precisely: For large n it is almost sure that the probability q of success is arbitrarily close to the ratio $r = n_1/n$ of the observed successes. This theorem will be called *Bayes' theorem* as a parallel to Bernoulli's theorem (Chapter IV) and may also be called a second law of large numbers. It bears out the statement on the concentration of the distribution $q_n(x)$, which was found, for the case of constant $p_0(x)$, in the foregoing section. Here, the theorem has been proved under the assumptions that $p_0(x)$

is bounded and, at $x = r$, continuous and not vanishing. This already shows that the numerical values of $p_0(x)$ or its specific shape are entirely irrelevant.

It should be understood that in all considerations of this chapter, which concern the inference problem, the observed frequency n_1/n (and its generalizations) is to be considered known. We ask for $q_n(x)$ or some property of $q_n(x)$ with n_1/n given. On the other hand, the corresponding "given" quantity in the Bernoulli problem (Chapter IV) is the basic probability q, and the situation is similar in all generalizations (Chapters IV–VI) of this "direct" problem.

3.2. *Weaker assumption.* By a more detailed analysis of the function $p_n(n_1 \mid x)$ it can be shown that the conditions for the *a priori* chance can still be considerably reduced. $p_0(x)$ may even vanish at $x = r$; it is sufficient that at the point $x = r$ the differential $dP_0(x)$ behaves like a positive power of dx, that is, for small $\eta > 0$

$$P_0(r + \eta) - P_0(r - \eta) > c\eta^{\kappa}, \qquad c, \kappa > 0. \tag{29}$$

To prove the above, we set

$$q_n(x) = \frac{x^{n_1}(1 - x)^{n-n_1}p_0(x)}{\int_0^1 x^{n_1}(1 - x)^{n-n_1}p_0(x)\,dx} = \frac{\left[\left(\frac{x}{r}\right)^r\left(\frac{1-x}{1-r}\right)^{1-r}\right]^n p_0(x)}{\int_0^1 p_0(x)\left[\left(\frac{x}{r}\right)^r\left(\frac{1-x}{1-r}\right)^{1-r}\right]^n dx} \tag{30}$$

and put

$$g(x) = \left(\frac{x}{r}\right)^r\left(\frac{1-x}{1-r}\right)^{1-r}, \qquad 0 \leqslant x \leqslant 1. \tag{31}$$

We cover the discrete and continuous cases by the use of Stieltjes integrals. Then

$$Q_A = \frac{\int_{r-\eta}^{r+\eta} g^n(x)\,dP_0(x)}{\int_0^1 g^n(x)\,dP_0(x)}, \qquad Q_A = 1 - Q_A, \tag{32}$$

If A denotes the region (23), where we write now η instead of ϵ, and g^n is the nth power of $g(x)$. We differentiate $g(x)$ and find

$$g'(x) = g(x)\frac{r - x}{x(1 - x)}, \qquad g''(x) = -g(x)\frac{r(1 - r)}{x^2(1 - x)^2}, \qquad g''(r) = \frac{-1}{r(1 - r)};$$

hence $g(x)$ takes on its maximum $g = 1$ for $x = r$ and is convex in the whole interval (0, 1).

Take $r < \frac{1}{2}$ and $\eta < r$ and compare $g(x)$ with the parabola $y = 1 - (x - r)^2$. The two curves are in contact at $x = r$, but outside a neighborhood of this point the parabola is above the $g(x)$ curve. Therefore

$$g(x) \leqslant 1 - \eta^2 \qquad \text{for} \qquad |x - r| \geqslant \eta. \tag{33}$$

Comparing on the other hand $g(x)$ with the straight line $y = 1 - (x - r)/n$ we see that for sufficiently large n, this line is below $y = g(x)$:

$$g(x) \geqslant 1 - \frac{\eta}{n} \qquad \text{for} \qquad | x - r | < \frac{\eta}{n}. \tag{34}$$

From (33)

$$\int_0^{r-\eta} + \int_{r+\eta}^1 g^n \, dP_0(x) \leqslant (1 - \eta^2)^n \left[\int_{\bar{A}} dP_0(x)\right] < (1 - \eta^2)^n . \tag{35}$$

Likewise we have for the denominator of(32) if we use (34) and the hypothesis (29):

$$\int_0^1 g^n \, dP_0 \geqslant \int_{r-\eta/n}^{r+\eta/n} \left(1 - \frac{\eta}{n}\right)^n dP_0 \geqslant c\left(\frac{\eta}{n}\right)^\kappa \left(1 - \frac{\eta}{n}\right)^n > c(1 - \eta) \left(\frac{\eta}{n}\right)^\kappa . \tag{36}$$

Therefore

$$Q_{\bar{A}} < \frac{(1 - \eta^2)^n n^\kappa}{c(1 - \eta)\eta^\kappa} . \tag{36'}$$

Since $n^\kappa(1 - \eta^2)^n$ converges toward zero with increasing n, our sharper statement has been proved. We reformulate:

Suppose an alternative with unknown probability q has been observed n times with n_1 times "success." Then the probability $Q_{\bar{A}}$ that q differs from n_1/n by more than η is below a bound which, for any η, no matter how small, tends to zero as both n and n_1 tend toward infinity and $n_1/n \rightarrow r$ if it is assumed that the a priori probability of the inequality $| q - (n_1/n) | \leqslant \eta$ is at least equal to $c\eta^\kappa$ where c and κ are positive constants.

3.3. *Discussion and comparisons.* Most criticism of Bayes' theory is directed toward the frequent lack of knowledge of the *a priori* probability. But we have proved that this lack of knowledge loses its importance if the number n of observations is large; in this case we obtain a valid inference without (practically) any assumption regarding $p_0(x)$. It is not right to state[2] that, in addition to a large number of observations knowledge of $p_0(x)$ is needed. One or the other is sufficient.

The two laws of large numbers and the first of the axioms that define probability (or chance) values as frequency limits express three intimately connected facts. The three propositions, however, are not identical. To make this clearly understood we have to translate each time the word "probability" (or chance) in terms of its statistical definition. We give here a synopsis of the propositions somewhat informally stated and specified for the alternative of getting "six" or "non-six" in casting a die ("six" is the "success").

1. *The first axiom* leading to the definition of probability states: In an indefinitely long sequence of castings with one and the same die, the rate n_1/n of successes approaches a limiting value q; this q is called the

[2] Uspensky [26], see p. 70.

probability of getting six with this particular die when the randomness condition is also fulfilled.

2. *The first law of large numbers* (Bernoulli theorem) states: If a group of n castings is repeated again and again with the same die, for which q is the success probability, almost all groups will yield a rate n_1/n of successes close to q if n is large enough. [In Chapter IV, Section 4, it has been shown that this is not true in cases where the above first axiom alone holds, but the sequence of casts is not a random sequence.]

3. *The second law of large numbers* (Bayes' theorem) says: If from a collection of different dice some die is repeatedly selected and each time the selected die is cast n times, then, if n is sufficiently large, almost every die among those for which the ratio of success had the value r will have a q-value near to r.

The first proposition is an assumption of an axiomatic character; the second and third are theorems proved mathematically on the basis of our axioms of Chapter I.

3.4. *Generalization of the law of large numbers.* We take as a starting point the following more general problem. Consider n drawings from an urn which contains $k + 1$ different labels $c_0, c_1, ..., c_k$ in unknown proportions $x_0 : x_1 : \cdots : x_k$ with $\Sigma_{\kappa=0}^{k} x_\kappa = 1$. In the last section, $k = 1$ was assumed. We observe that c_κ appeared n_κ times, $\kappa = 0, 1, ..., k$; let $n_\kappa/n = r_\kappa$, $\Sigma_{\kappa=0}^{k} r_\kappa = 1$. We denote the prior density by $p_0(x_1, x_2, ..., x_k)$, the posterior density by $q_n(x_1, x_2, ..., x_k)$. The prior density is *bounded*; $p_0(r_1, r_2, ..., r_k) \neq 0$; and p_0 is continuous in the neighborhood of $(r_1, r_2, ..., r_k)$. (See Section 3.2 for more general assumptions.)

We wish to show that, for large n, there is a very large probability that the unknown probabilities $x_0, x_1, ..., x_k$ are as close as we please to the observed $r_0, r_1, ..., r_k$.

This follows without computation from our previous results, since we may argue as follows: the label c_1 has been observed $n_1 = nr_1$ times and the label "non-c_1" has been observed $n - n_1 = n(1 - r_1)$ times. Then Bayes' theorem asserts that if n is sufficiently large it is almost certain that the unknown probability x_1 of c_1 is arbitrarily close to r_1. In the same way we may argue regarding the alternatives "c_2 or non-c_2", "c_3 or non-c_3", etc., by applying Bayes' theorem each time. The result is:

If in the n-fold observation of a collective with attributes $c_0, c_1, ..., c_k$ the label c_κ has been observed $r_\kappa = n_\kappa/n$ times, then, if n is large enough, the probability of a maximum deviation greater than ϵ between the observed $r_1, ..., r_k$ and the unknown $x_1, ..., x_k$ is arbitrarily close to zero no matter *how small an ϵ has been chosen. This holds for any bounded prior density with $p_0(r_1, ..., r_k) \neq 0$.*

We derive from this theorem a law of large numbers regarding the unknown mean value

$$x = c_0x_0 + c_1x_1 + \cdots + c_kx_k .$$

Let d and d_1 be two real numbers; we ask for the probability that x lies between d and d_1. To answer this question we need the k-dimensional posterior probability $q_n(x_1, x_2, ..., x_k)$. It is given by the following formula which generalizes Eq. (8) or Eq. (30):

$$q_n(x_1, x_2, ..., x_k) = C_n p_0(x_1, x_2, ..., x_k) \cdot [x_1^{r_1} x_2^{r_2} \cdots x_k^{r_k} x_0^{r_0}]^n. \tag{37}$$

This q_n has to be integrated between the two planes

$$\begin{aligned} c_0x_0 + c_1x_1 + \cdots + c_kx_k &= d, \\ c_0x_0 + c_1x_1 + \cdots + c_kx_k &= d_1 , \end{aligned} \tag{38}$$

and we ask for the asymptotic value of this integral. We argue as follows. If, as n tends toward infinity, the space bounded by (38) does not contain the point $x_1 = r_1, ..., x_k = r_k$ the integral in question tends toward zero as $n \to \infty$. If however, we choose d and d_1 such that they differ from each other arbitrarily little but such that the value $c_1r_1 + \cdots + c_kr_k + c_0(1 - r_1 - \cdots - r_k)$ remains between them, as $n \to \infty$, then the integral under consideration taken over *this* region tends toward one. Hence, with the same assumptions on p_0 as in the preceding result: *it is almost certain that the mean value $c_0x_0 + c_1x_1 + \cdots + c_kx_k$ of the unknown distribution lies arbitrarily close to the observed average $c_0r_0 + c_1r_1 + \cdots + c_kr_k$ if the number n of observations which yielded the values $r_\kappa = n_\kappa/n$, $\kappa = 0, 1, ..., k$ was large enough.*

This statement includes Bayes' theorem (Section 3.1) to which it reduces for $k = 1$.

Actually, it can even be seen that the above simple conclusion which followed from Eq. (37) remains valid if instead of the mean value $\sum_{i=0}^{k} c_ix_i$ we consider some other bounded and continuous function $f(x_1, ..., x_k)$ of the unknown probabilities x_κ in the arithmetical collective under consideration. We assume that f is continuous in the neighborhood of the point $(r_1, r_2, ..., r_k)$; the prior probability p_0 is assumed different from zero at this point and continuous in its neighborhood. *It is then almost certain that, for n sufficiently large, $f(x_1, x_2, ..., x_k)$ differs arbitrarily little from $f(r_1, r_2, ..., r_k)$.*[3]

[3] v. Mises [21], p. 196.

A more general *second law of large numbers* will be proved in Chapter XII, Section 4, to which we also refer for more careful formulation of our last extension.

C. Asymptotic Distributions (Sections 4-6)

4. Limit Theorems for Bayes' Problem

4.1. *Bayes' problem for large n.* Laplace was the first to consider the limit of $q_n(x)$ as $n \to \infty$, assuming that simultaneously $n_1 \to \infty$, while $n_1/n = r$ remains fixed. The result is that $q_n(x)$ converges toward a Gaussian distribution according to the following formula:

$$\lim_{n\to\infty} \sqrt{\frac{r(1-r)}{n}}\, q_n\left(r + u\sqrt{\frac{r(1-r)}{n}}\right) = \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2} = \phi(u). \tag{39}$$

Again we should stress that in the inference problem n is given. Results such as (39) should be understood as approximation formulas valid for large n. In Laplace's investigation $p_0(x)$ was a constant, an unnecessary restriction which will now be removed. Equation (39) is a "local theorem" like the theorems of Chapter VI, Sections 3.1 and 5.1.

Formula (8), which solves the Bayes' problem for any n, can be written in a slightly modified form as

$$q_n(x) = C_n p_0(x) \left(\frac{1-x}{1-r}\right)^{n-n_1} \left(\frac{x}{r}\right)^{n_1}. \tag{30'}$$

The factor C_n , to be determined from the condition $\int_0^1 q_n\, dx = 1$, takes care of the constant in (8) and the arbitrarily introduced factors $(1-r)^{n-n_1}$ and r^{n_1}. Setting $x' = x - r$ we write for the $g(x)$ of (31)

$$\left(\frac{1-x}{1-r}\right)^{1-r} \left(\frac{x}{r}\right)^{r} = \left(\frac{1-r-x'}{1-r}\right)^{1-r} \left(\frac{x'+r}{r}\right)^{r} = f(x'), \quad -r \leqslant x' \leqslant 1-r, \tag{40}$$

and thus have, since $n_1 = nr$, $n - n_1 = n(1-r)$,

$$\frac{q_n(x)}{C_n} = p_0(x)[f(x')]^n. \tag{41}$$

Our present task is to find the limit of the right-hand side of this equation for infinite n. Except for the factor $p_0(x)$ [which is assumed continuous and bounded in (0, 1) and with $p_0(r) \neq 0$] we have here a product of n equal functions of x'.

It can easily be seen that $f(x')$ satisfies the conditions of the product theorem of Chapter VI, Section 1. As seen from (31), for $x = r$, i.e., $x' = 0$, the function has the value 1. Its first derivative at this point is zero and the second is $-1/r(1 - r) = -R^2$.[1] The third derivative has certainly a finite value if r is an inner point of the interval 0, 1. It thus follows that the product theorem of Chapter VI, Section 1 applies to the present case.

We set

$$s_n^2 = nR^2 = \frac{n}{(1 - r)r}, \qquad x' = \frac{z}{s_n}, \qquad x = r + \frac{z}{s_n}$$

and have the result

$$\lim_{n \to \infty} f^n\left(\frac{z}{s_n}\right) = e^{-z^2/2}. \tag{42}$$

The limit of $q_n(x)/C_n$ from (39) is therefore (for any finite z)

$$\lim_{n \to \infty} \frac{q_n(x)}{C_n} = \lim_{n \to \infty} p_0\left(r + \frac{z}{s_n}\right) e^{-z^2/2}. \tag{43}$$

Assuming, as above, that $p_0(x)$ is continuous and not zero at the point $x = r$, this supplies [with a new constant $C_n' = C_n p_0(r)$]

$$\lim_{n \to \infty} \frac{1}{C_n'} q_n\left(r + \frac{z}{s_n}\right) = e^{-z^2/2}. \tag{44}$$

If we reintroduce x instead of z and write (44) as an approximation formula for large n, it reads

$$q_n(x) \sim C_n' \exp\left(-\frac{n(x - r)^2}{2(1 - r)r}\right), \tag{44'}$$

where C_n' has to be determined from $\int_0^1 q_n(x)\, dx = 1$. If we permit ourselves to use the approximation (44′) for the whole range $0 \leqslant x \leqslant 1$ (which implies z going to infinity) we have, with the abbreviation

$$u = \sqrt{\frac{n}{(1 - r)r}}(x - r), \qquad x = r + u\sqrt{\frac{(1 - r)r}{n}}, \tag{45}$$

the condition

$$1 = C_n' \int_0^1 e^{-u^2/2}\, dx = C_n' \sqrt{\frac{r(1 - r)}{n}} \int e^{-u^2/2}\, du = C_n' \sqrt{\frac{r(1 - r)}{n}} \sqrt{2\pi},$$

[1] We use here R^2 instead of r^2 of Chapter VI in order to avoid confusion with the present $r = n_1/n$.

which supplies

$$C_n' = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{r(1-r)}}, \tag{46'}$$

which has to be carried into (44'). A rigorous computation of C_n' (which leads to the same result) would require a study of the function f outside the neighborhood of the point $x = r$, as in the case of the second product theorem in Section 1.2 of Chapter VI.[2]

From Eqs. (44') and (46') we can pass to the integral formula

$$Q_n(x_2) - Q_n(x_1) = \int_{x_1}^{x_2} q_n(x)\,dx \sim \sqrt{\frac{n}{2\pi(1-r)r}} \int_{x_1}^{x_2} \exp\left(-\frac{n(x-r)^2}{2(1-r)r}\right) dx,$$

and this gives

$$Q_n(x_2) - Q_n(x_1) \sim \frac{1}{\sqrt{2\pi}} \int_{u_1}^{u_2} e^{-u^2/2}\,du \tag{48}$$

with

$$u_1 = \sqrt{\frac{n}{(1-r)r}}(x_1 - r), \qquad u_2 = \sqrt{\frac{n}{(1-r)r}}(x_2 - r).$$

In particular, for $x_1 = r - X$, $x_2 = r + X$, with $G = \Phi(x) - \frac{1}{2}$,

$$Q_n(r + X) - Q_n(r - X) = 2G\left(\sqrt{\frac{n}{r(1-r)}}X\right). \tag{49}$$

This formula, given by Laplace (1812) and known as the *Laplace-Bayes* formula, includes the Bayes theorem. In fact, however small X is, the argument of G on the right-hand side becomes infinite as n increases and $2G$ accordingly goes to unity.

4.2. *Discussion.*

Two things are noteworthy in these results. First, the *a priori* probability has disappeared completely from the formulas. This is in agreement with the fact mentioned repeatedly in this chapter, that *for large n* inferences can be made from the observed value of $r = n_1/n$ without any additional knowledge. On the other hand, if n is a moderate number the approximation formulas do not apply and the original solution (30') (p. 345) includes the unknown $p_0(x)$. It remains an

[2] Cf. v. Mises [21], pp. 159 and 160; "Fundamentalsätze." *Math. Z.* 4 (1919), pp. 1–96, see p. 83.

invariable fact, pervading all problems in mathematical statistics, that *no substantial inference can be drawn from a small number of observations if nothing is known prior to the experiments about the object of experimentation.*

The second remark again concerns the analogy of the present formulas with those giving the solution of Bernoulli's problem. If in Chapter VI, Eq. (35) the number x of events is replaced by zn, so that z denotes the ratio of successes, and if the density is referred to z instead of to x, so that $q_n(x)$ has to be replaced by $q_n'(z)/n$, that formula reads

$$q_n'(z) \sim \sqrt{\frac{n}{2\pi(1-q)q}} \exp\left(-\frac{n(z-q)^2}{2q(1-q)}\right) \tag{50}$$

while, in the Bayes problem (44′) and (46′)

$$q_n(x) \sim \sqrt{\frac{n}{2\pi(1-r)r}} \exp\left(-\frac{n(x-r)^2}{2(1-r)r}\right). \tag{44''}$$

Here, the formal correspondence between r, x on the one hand and q, z on the other is apparent. In (50) q is the mean value and $q(1-q)/n$ the variance of the Bernoulli distribution. For Bayes' $q_n(x)$, the mean, Q'_n and variance $s_n'^2$ depend on $p_0(x)$; but $\lim_{n\to\infty} a'_n = r$, $\lim_{n\to\infty} ns_n'^2 = r(1-r)$.

We consider a numerical example. A coin is tossed 100 times and heads appear 53 times. We call x the unknown head-probability and assume an *a priori* probability constant over the interval A from 0.45 to 0.55, and zero outside A. We can then apply Eq. (44′) and determine C_n' from the condition $Q_A = 1$ (in this way the given *a priori* probability is used). We ask, for example, for the probability P of $x > 0.50$ and find from (48) with $x_1 = 0.50$, $x_2 = 0.55$, $P = 0.633$, a result that makes sense. This is at the same time the probability that x be between 0.50 and 0.55. If we make the (rather silly) assumption that $p_0(x) = 1$ all over (0, 1) we find less than 45% probability for x between 0.50 and 0.55. If in $n = 10{,}000$ trials, heads appear 5300 times, the corresponding probabilities differ from 1 by less than the fifth decimal in both cases.

Problem 8. In 1000 tossings of a coin, heads appears 504 times. What is the probability that the coin is biased in favor of heads i.e., that the probability of heads is greater than 0.50?

Problem 9. A coin is considered unbiased if the probability of heads does not deviate by more than 0.001 from $\frac{1}{2}$. How many trials giving $r = \frac{1}{2}$ must be made to secure the unbiasedness with 99% probability?

Problem 10. For not too small n, the influence of $p_0(x)$ can be judged by using the equation

$$q_n(x) \sim C_n p_0(x)\, e^{-u^2/2}, \qquad u = \sqrt{\frac{n}{(1-r)r}}\,(x - r).$$

Compute from this formula the probability of x falling in the interval 0.499 to 0.501 in the case of $n = 10{,}000$, $r = 0.5$, under the following three assumptions:

(a) $p_0(x) = 1$ in $0 \leqslant x \leqslant 1$

(b) $p_0(x) = 10$ in $0.45 \leqslant x \leqslant 0.55$

(c) $p_0(x) = 2x$ in $0 \leqslant x \leqslant 1$.

4.3. *Bayes-Laplace formula for $(k + 1)$-valued distributions.* We return to the problem of Section 3.4. There are n observations of the same arithmetical collective with the attributes c_0, c_1, ..., c_k. The observations yield $n_\kappa = nr_\kappa$ times the attribute c_κ, $\kappa = 0, 1, \ldots k$; the unknown probability of c_κ is x_κ and $\sum_{\kappa=0}^{k} x_\kappa = 1$, $\sum_{\kappa=0}^{k} r_\kappa = 1$. The *a priori* density is $p_0(x_1, x_2, \ldots, x_k)$ and the *a posteriori* density is $q_n(x_1, x_2, \ldots, x_k)$, where

$$q_n(x_1, \ldots, x_k) = C_n p_0(x_1, x_2, \ldots, x_k) x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k} x_0{}^{n_0}. \tag{51}$$

The constant C_n is to be determined from the condition that the k-dimensional integral between 0 and 1 of $q_n(x_1, \ldots, x_k)$ equals 1.

The observed average of the observations is

$$a = r_1 c_1 + \cdots + r_k c_k + r_0 c_0 = \frac{1}{n}(n_1 c_1 + \cdots + n_k c_k + n_0 c_0). \tag{52}$$

The unknown mean value of the collective is

$$x = c_1 x_1 + \cdots + c_k x_k + c_0 x_0. \tag{53}$$

We desire a statement regarding the asymptotic value of the *a posteriori* density $q_n(x)$ of (53).

To this purpose we have to submit (51) to an integration (mixing). Denote by S the region of the k-dimensional space which is bounded for fixed x by the plane (53) and the neighboring plane

$$x + dx = c_1 x_1 + \cdots + c_k x_k + c_0 x_0\,. \tag{53'}$$

Then

$$q_n(x) = \int_{(S)} q_n(x_1, x_2, \ldots, x_k)\, dx_1\, dx_2 \cdots dx_k\,. \tag{54}$$

We wish to show that as $n \to \infty$, both the $q_n(x_1, ..., x_k)$ of (51) and the $q_n(x)$ of (54) converge, under certain restrictions, toward normal distributions.

The first part of the investigation is very similar to that in Section 4.1; essentially, we apply the product theorem of Chapter VI, Section 1 to functions of several variables. We define a function

$$f(x_1, x_2, ..., x_k) = \left(\frac{x_1}{r_1}\right)^{r_1} \left(\frac{x_2}{r_2}\right)^{r_2} \cdots \left(\frac{x_0}{r_0}\right)^{r_0}. \tag{55}$$

Then (51) takes the form

$$q_n(x_1, x_2', ..., x_k) = C_n' p_0(x_1, x_2, ..., x_k)[f(x_1, x_2, ..., x_k)]^n. \tag{56}$$

The f of (55) equals 1 at

$$x_1 = r_1, \quad x_2 = r_2, ..., x_k = r_k,$$

and it is easily seen that its first derivatives at this point vanish. Consider $\log f = F$ and differentiate F, remembering that $x_0 = 1 - x_1 - x_2 - \cdots - x_k$. Writing

$$\left(\frac{\partial F}{\partial x_i}\right)_{x_1=r_1, \ldots, x_k=r_k} = F_i, \quad \left(\frac{\partial f}{\partial x_i}\right)_{x_1=r_1, \ldots, x_k=r_k} = f_i$$

we obtain, with $i, \kappa = 1, 2, ..., k$,

$$F_i = f_i = \left(\frac{r_i}{x_i} - \frac{r_0}{x_0}\right) \cdot f = 0$$

$$F_{i\kappa} = f_{i\kappa} = -\frac{r_0}{x_0^2} \cdot f = -\frac{1}{r_0}, \qquad i \neq \kappa$$

$$F_{ii} = f_{ii} = -\left(\frac{r_i}{x_i^2} + \frac{r_0}{x_0^2}\right) \cdot f = -\left(\frac{1}{r_i} + \frac{1}{r_0}\right). \tag{57}$$

We set up Taylor's formula for $F = \log f$ and obtain, since $F(r_1, r_2, ..., r_\kappa) = 0$ and $F_i = 0$,

$$F(x_1, x_2, ..., x_k) = \tfrac{1}{2} \sum_{i,\kappa=1}^{k} F_{i\kappa}(x_i - r_i)(x_\kappa - r_\kappa) + \text{terms of 3rd order in } (x_i - r_i).$$

The coefficients of the terms of third order are third derivatives of F at some intermediate point of the interval between $x_1, ..., x_k$ and $r_1, ..., r_k$.

We introduce the new variables u_i,

$$u_i = \sqrt{n}(x_i - r_i), \qquad x_i = r_i + \frac{u_i}{\sqrt{n}}, \qquad i = 1, 2, ..., k. \tag{58}$$

Then

$$nF(x_1, x_2, ..., x_k) = \frac{1}{2}\sum_{i,\kappa}^{1...k} F_{i\kappa}u_i u_\kappa + \frac{\text{terms of 3rd order in } u_i}{\sqrt{n}}.$$

If we keep the u_i bounded, the terms of third order will vanish as $n \to \infty$ on account of $1/\sqrt{n}$, and we obtain

$$\lim_{n\to\infty} nF\left(r_1 + \frac{u_1}{\sqrt{n}}, r_2 + \frac{u_2}{\sqrt{n}}, ..., r_k + \frac{u_k}{\sqrt{n}}\right) = \frac{1}{2}\sum_{i,\kappa}^{1...k} F_{i\kappa}u_i u_\kappa.$$

If we assume $p_0(x_1, x_2, ..., x_k)$ to be continuous and non-zero at the point $r_1, r_2, ..., r_k$, then

$$\lim_{n\to\infty} p_0\left(r_1 + \frac{u_1}{\sqrt{n}}, r_2 + \frac{u_2}{\sqrt{n}}, ..., r_k + \frac{u_k}{\sqrt{n}}\right) = p_0(r_1, r_2, ..., r_k).$$

Therefore, with $C_n'' = C_n' p_0(r_1, r_2, ..., r_k)$:

$$\lim_{n\to\infty} \frac{1}{C_n''} q_n\left(r_1 + \frac{u_1}{\sqrt{n}}, ..., r_k + \frac{u_k}{\sqrt{n}}\right) = \exp\left(+ \tfrac{1}{2}\sum F_{i\kappa}u_i u_\kappa\right). \tag{59}$$

We write

$$\tfrac{1}{2}\sum_{i,\kappa} F_{i\kappa}u_i u_\kappa = -Q(u_1, u_2, ..., u_k), \tag{60}$$

where Q is a quadratic form in the $u_1, u_2, ..., u_k$ and we have to show that Q is positive definite. Since $u_0 = \sqrt{n}(x_0 - r_0)$ and $u_1 + \cdots + u_k + u_0 = 0$, we have, using (57)

$$Q = \frac{1}{2}\left[\sum_{i=1}^{k} \frac{u_i^2}{r_i} + \frac{1}{r_0}\sum_{i,\kappa=1}^{k} u_i u_\kappa\right] = \frac{1}{2}\left[\sum_{i=1}^{k} \frac{u_i^2}{r_i} + \frac{1}{r_0}\left(\sum_{i=1}^{k} u_i\right)^2\right] = \frac{1}{2}\sum_{i=0}^{k} \frac{u_i^2}{r_i}. \tag{61}$$

To find C_n'' we use, as always, the fact that the probability $q_n(x_1, ..., x_k)$ integrated over all x_i - values from 0 to 1 gives unity. Integrating over the x_i from 0 to 1 means integrating over the u_i from $-\infty$ to $+\infty$. We must show that it is legitimate to perform this integration on the limit result

(59), although this result has been derived under the assumption of bounded u_i. This can be shown as in Chapter VI, Section 1.2.

We introduce once more new variables v_i:

$$u_i = \sqrt{r_i}\,v_i\,, \qquad du_i = \sqrt{n}\,dx_i = \sqrt{r_i}\,dv_i\,, \qquad i = 0, 1, \ldots, k$$

and have from (59)

$$\frac{1}{C''_n} = \int_0^1\int_0^1\cdots\int_0^1 e^{-Q}\,dx_1\,dx_2\cdots dx_k$$

$$= \sqrt{\frac{r_1 r_2\cdots r_k}{n^k}}\iint\cdots\int \exp\left[-\tfrac{1}{2}(v_1^2 + v_2^2 + \cdots + v_0^2)\right]dv_1\,dv_2\cdots dv_k\,. \tag{62}$$

In the $(k+1)$-dimensional space with coordinates $v_1, v_2, \ldots, v_k, v_0$ we have on account of $\sum_0^k u_i = 0$ also $\sqrt{r_1}v_1 + \cdots + \sqrt{r_k}v_k + \sqrt{r_0}v_0 = 0$, which is the equation of a plane whose direction cosines are the $\sqrt{r_i}$. Since the integrand in (62) depends on the distance $v_0^2 + v_1^2 + \cdots + v_k^2$ only, we can rotate the coordinate system in such a way that this plane will have the equation $v_0' = 0$ or simply $v_0 = 0$. The differential $dv_1\,dv_2\cdots dv_k$ must then be multiplied by the Jacobian, $\sqrt{r_0}$.[3] Thus

$$\sqrt{r_0}\int\cdots\int \exp\left[-\tfrac{1}{2}(v_1^2 + \cdots + v_k^2)\right]dv_1\cdots dv_k = \sqrt{r_0}\,\sqrt{(2\pi)^k}$$

and from (62), (59), and (60) the following approximation formula results:

$$q_n(x_1, x_2, \ldots, x_k) \sim \sqrt{\left(\frac{n}{2\pi}\right)^k \frac{1}{r_0 r_1\cdots r_k}}\,\exp\left(-\frac{n}{2}\sum_{i=0}^{k}\frac{(x_i - r_i)^2}{r_i}\right). \tag{63}$$

This formula gives the generalization of the Laplace-Bayes' result of Section 4.2 to a $(k+1)$-valued arithmetical distribution.

4.4. *Posterior probability of unknown mean value.*[4] We have still to study the asymptotic form of the *a posteriori* density $q_n(x)$ defined by (54). Instead of determining the limit of the integral to the right in (54) we carry out the integration over the region S in (54) on the limit expression

[3] $\sqrt{r_0}$ equals the cosine of the angle of revolution.

[4] Cf. R. v. Mises, "Fundamentalsätze." *Math. Z.* 4 (1919), p. 88 ff; [21], p. 227 ff. The theorem (for constant prior probability) has been stated by Laplace and proved (essentially) by Bienaymé.

(63); in other words, we introduce the right-hand side of (63) into (54); this interchange of two limits is allowed since the f of (55) satisfies the conditions of the second theorem in Chapter VI, Section 1.2.

Instead of the x_i we use again the v_i. The equation of the plane (53) is then [with a given by (52)]

$$(x - a)\sqrt{n} = c_1\sqrt{r_1}v_1 + \cdots + c_k\sqrt{r_k}v_k + c_0\sqrt{r_0}v_0 \,. \tag{64}$$

and the v_i satisfy the equation

$$0 = \sqrt{r_1}v_1 + \sqrt{r_2}v_2 + \cdots + \sqrt{r_0}v_0 \,. \tag{65}$$

The two "planes" (64) and (65) intersect along a "straight line" and we need the distance h of this line of intersection from the origin O (Fig. 20).

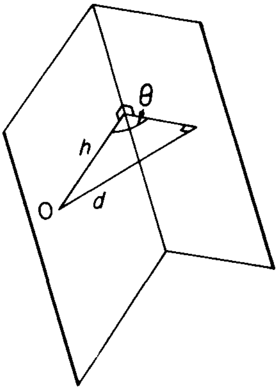


FIG. 20. Notations.

If d is the distance of the plane (64) from O and θ the angle between the two planes, then there is a right-angled triangle, such that $h = d/\sin\theta$. Let us find $\sin\theta$.

The direction cosines of (65) are $\sqrt{r_i}$, those of (64) equal $c_i\sqrt{r_i}/\sqrt{\sum c_i^2 r_i}$. Elementary geometry gives

$$\cos\theta = \sum_{i=0}^{k}\left(\sqrt{r_i}\,\frac{c_i\sqrt{r_i}}{\sqrt{\sum c_i^2 r_i}}\right) = \frac{\sum c_i r_i}{\sqrt{\sum c_i^2 r_i}} = \frac{a}{\sqrt{\sum c_i^2 r_i}}\,,$$

where (52) has been used, and

$$\sin^2\theta = 1 - \cos^2\theta = \frac{\sum c_i^2 r_i - a^2}{\sum c_i^2 r_i} = \frac{\sum (c_i - a)^2 r_i}{\sum c_i^2 r_i}\,.$$

The distance d of the plane (64) from the origin equals the left-hand side of the normal form of the equation of the plane, hence

$$d = (x - a)\sqrt{n}/\sqrt{\sum c_i^2 r_i}\,.$$

Therefore

$$h^2 = \frac{(x-a)^2 n}{\sum c_i^2 r_i \sin^2\theta} = \frac{(x-a)^2 n}{\sum (c_i - a)^2 r_i} = \frac{(x-a)^2 \cdot n}{s^2}. \tag{66}$$

We introduce now the v_i into (54) and obtain

$$q_n(x)\,dx \sim \text{const.} \int_{(S')} \exp\left[-\tfrac{1}{2}(v_0^2 + v_1^2 + \cdots + v_k^2)\right] dv_1 \cdots dv_k\,,$$

where the "const" is so far undetermined and S' corresponds to the plane (65). Again, we rotate this plane in such a way that its equation is $v_0 = 0$; this merely changes the constant in front of the integral. We obtain

$$q_n(x)\,dx \sim \text{const.} \int_{(S'')} \exp\left[-\tfrac{1}{2}(v_1^2 + v_2^2 + \cdots + v_k^2)\right] dv_1 \cdots dv_k\,. \tag{67}$$

S'' is now that part of the plane $v_0 = 0$ which lies between the two parallel lines whose distance from the origin are h and $h + dh$, respectively. Since the integrand in (67) depends on the sum of the squares of the coordinates only, we do not change the value of the integral by rotating the axes in such a way that these two lines have the equations $v_k = h$ and $v_k = h + dh$. The integration is, therefore, from $-\infty$ to $+\infty$ with respect to $v_1, v_2, ..., v_{k-1}$ and from h to $h + dh$ with respect to v_k. Thus

$$q_n(x)\,dx \sim \text{const.} \int e^{-\frac{1}{2}v_1^2}\,dv_1 \cdots \int e^{-\frac{1}{2}v_{k-1}^2}\,dv_{k-1} \int_h^{h+dh} e^{-\frac{1}{2}v_k^2}\,dv_k$$

$$= \text{const.}\sqrt{(2\pi)^{k-1}}\,e^{-\frac{1}{2}h^2}\,dh.$$

Now from (66) we see that dh equals dx to within a constant factor; hence, we "cancel" dx and dh and have asymptotically, using (66),

$$q_n(x) \sim \sqrt{\frac{n}{2\pi}}\,\frac{1}{s}\exp\left(-\frac{n(x-a)^2}{2s^2}\right), \tag{68}$$

where

$$a = \sum_{i=0}^{k} c_i r_i\,, \qquad s^2 = \sum_{i=0}^{k} (c_i - a)^2 r_i\,. \tag{68'}$$

If we set $k = 1$ and use the notation of Section 4.1, $r_1 = r, r_0 = 1 - r$, $c_1 = 1$, $c_0 = 0$ we have

$$a = r, \qquad s^2 = r(1 - r)^2 + (1 - r)r^2 = r(1 - r)$$

and (68) reduces to (44″).

In (68) and (68′), a and s^2 are known from the observations; a is the average, s^2 the dispersion of the observations. The variance of the right-hand side of (68) is s^2/n.

As a limit formula (68) is written, with $u = [(x - a)/s] \sqrt{n}$, $x = a + us/\sqrt{n}$,

$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} s q_n \left(a + u \sqrt{\frac{1}{n}} s\right) = \phi(u), \tag{69}$$

or, in integrated form, with the usual notation:

$$\lim_{n \to \infty} Q_n(x) = \Phi(u). \tag{69′}$$

We state our result[5] as follows.

If n observations on the same arithmetical $(k + 1)$-valued collective have given average a and dispersion s^2, then the posterior density $q_n(x)$ of the *unknown mean value of the distribution tends, as $n \to \infty$, toward the normal distribution $N(a, s^2/n)$ if the $r_i = n_i/n$, $i = 1, 2, \ldots, k$ are kept fixed as $n \to \infty$. The arbitrary bounded prior density $p_0(x_1, x_2, \ldots, x_k)$ is assumed continuous and different from zero at the point $(r_1, r_2, \ldots, r_k)$.*

This result may also be applied to the following situation. The attribute space may be divided into $k + 1$ parts $L_0, L_1, L_2, \ldots, L_k$ and we denote by x_κ the probability of a result falling into L_κ.

5. Application of the Two Basic Limit Theorems to the Theory of Errors

5.1. *The law of error. The hypothesis of elementary errors.* If the same object is measured repeatedly in the same way one obtains, in general, a sequence of slightly different results. *We assume that these results of repeated measurements form the elements of a collective.* It follows that there exists a distribution $P(x)$ which gives the probability of a result being less than or equal to x. The mean value of $P(x)$, viz.,

$$\alpha = \int x \, dP(x) \tag{70}$$

[5] v. Mises called this theorem the "second fundamental limit theorem" since it is a kind of converse (though much less general) to the "first" or central limit theorem.

is considered the (unknown) *true value* of the magnitude under consideration. What should we assume regarding $P(x)$, the *law of error*?

The central limit theorem of Chapter VI gives a certain answer to this question. One might imagine that each of the actual results differs from the correct true value by a deviation, the *error* of this measurement $\epsilon_\nu = a_\nu - \alpha$; this ϵ_ν is due to a variety of independent causes, like inaccuracy of the tools, of the scales, parallax, perhaps the influence of temperature, of air pressure, "subjective errors" of the investigator, etc. This leads to the hypothesis that the deviation between the result obtained and the true value *is a sum of a great number of "elementary errors,"* of independent variates, each of which is subject to a probability distribution. If we apply the central limit theorem to this situation, it follows that the distribution of the deviation and, therefore, *the law of error $P(x)$ is a normal curve with mean value α.* However, we do not know α and we do not know the variance σ^2 of this normal curve. The probability density of an error $z = x - \alpha$ is then

$$p(z) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-z^2/2\sigma^2} \tag{71}$$

with α and σ^2 unknown.

Note that the above conclusion regarding the normality of $P(x)$ is independent of the number and results of the observations.

5.2. *Inference on the true value.* If our measurements have given n results $a_1, a_2, \ldots, a_n$ it is natural to form their average

$$a = \frac{1}{n}(a_1 + a_2 + \cdots + a_n),$$

and to ask for the posterior probability—inferred from those observations—that the deviation between the observed a and the unknown true value α remains below a certain value. We thus assume that the true value α is subject to chance, i.e., that a density $p_0(x)$ exists, that the true value equals x. Then the product

$$p_0(x)p(a_1 - x) \cdots p(a_n - x)$$

is the joint probability that the measurements of an object whose true "length" equals x are $a_1, a_2, \ldots, a_n$. To find the inference density $q_n(x)$ we—as usual—apply a partition and have

$$q_n(x \mid a_1, a_2, \ldots, a_n) = Cp_0(x)p(a_1 - x) \cdots p(a_n - x) \tag{72}$$

with

$$1/C = \int p_0(x)p(a_1 - x) \cdots p(a_n - x)\, dx. \tag{72'}$$

If we now substitute for p the expression (71) we obtain, writing $q_n(x)$ for $q_n(x \mid a_1, \ldots, a_n)$ with the same σ^2 as in (71),

$$q_n(x) = C_n p_0(x) \exp\left(-\frac{1}{2\sigma^2} \cdot \sum_{\nu=1}^{n} (a_\nu - x)^2\right). \tag{73}$$

Let

$$s^2 = \frac{1}{n}\sum_{\nu=1}^{n} (a_\nu - a)^2, \qquad \text{where} \qquad a = \frac{1}{n}\sum_{\nu=1}^{n} a_\nu . \tag{74}$$

Then

$$\sum_{\nu=1}^{n} (a_\nu - x)^2 = ns^2 + n(x - a)^2$$

and (73) becomes

$$q_n(x) = C_n \exp\left(-\frac{ns^2}{2\sigma^2}\right) p_0(x) \exp\left(-\frac{n(x-a)^2}{2\sigma^2}\right)$$

$$= C_n' p_0(x) \exp\left(-\frac{n(x-a)^2}{2\sigma^2}\right). \tag{75}$$

If we assume $p_0(x)$ = constant we put $p_0 C_n' = C_n''$, and, determining C_n'' in the usual way, we have

$$q_n(x) = \frac{1}{\sigma}\sqrt{\frac{n}{2\pi}} \exp\left(-n\frac{(x-a)^2}{2\sigma^2}\right), \tag{76}$$

for the probability density inferred from the mean a, that the true value is at x. Here σ^2 is not known.

Let us now *assume that n is large*. Then a decisive statement follows from the "second basic limit theorem" of the preceding section. Since we consider the repeated measurements as a sequence of trials performed on the same collective[1] we obtain the asymptotic result: *Whatever the law $P(x)$ and the prior probability distribution $P_0(x)$, the posterior probability density that the true value is x is asymptotically normal with mean value a and variance s^2/n, where a and s^2 are computed from* (74). Except for the remark on $p_0(x)$ this result is due to Bienaymé, 1838, but was not known

[1] See the end of Section 4.

to Gauss. Thus, asymptotically, $q_n(x)$ is entirely determined from the measurements. The hypothesis of elementary errors does not enter this deduction.

Reviewing, we see that the following relations hold:

(A) *The hypothesis of elementary errors gives for* $p(z)$, $z = x - \alpha$, *the form* (71), *with true value* α *and variance* σ^2, *both undetermined.*[1] *If we assume constant prior probability* $p_0(x)$, *then for the inferred probability of the true value, the result* (76) *holds with the same* (*unknown*) σ^2 *as in* (71). *The "precision"*[2] *of the true value is* $\sqrt{n}$ *times that of each single observation. All this holds for any n, small or large.*

(B) *If n is large, both constants in* (76) *are known*; *they are asymptotically equal to a and* s^2 *of* (74) *and this result holds for any prior probability. No information about the law of error,* $p(z)$ *follows from this consideration.*

6. Inference on a Statistical Function of Unknown Probabilities

6.1. *The problem.* In this section we wish to prove an inference theorem where instead of the theoretical mean value (53) of the unknown arithmetical distribution, a general function of this distribution is considered. In Chapter XI we shall deal at greater length with such "statistical functions" and Chapter XII is exclusively devoted to them. It should, however, be emphasized that in the case of an arithmetical distribution, the statistical function reduces to an ordinary function of k variables either of the relative frequencies $r_1, r_2, ..., r_k$ or of probabilities $x_1, x_2, ..., x_k$. Here we deal with the second case. We denote now by k, rather than by $(k + 1)$, the total number of attributes and have therefore $\Sigma_{i=1}^k r_i = \Sigma_{i=1}^k x_i = 1$, instead of the corresponding equation at the beginning of Section 4.3. Equation (51), which we shall need, reads then,

$$q_n(x_1, ..., x_{k-1}) = C_n p_0(x_1, ..., x_{k-1}) x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}. \qquad (51')$$

Denote by $f(x_1, x_2, ..., x_k)$ a function of the (unknown) probabilities; for brevity we write $\mathbf{x} = \mathbf{r}$ as an abbreviation for $x_1 = r_1$, $x_2 = r_2$ ..., $x_k = r_k$. We put

$$\left(\frac{\partial f}{\partial x_\kappa}\right)_{\mathbf{x}=\mathbf{r}} = f_\kappa, \qquad \left(\frac{\partial^2 f}{\partial x_\kappa \, \partial x_\lambda}\right)_{\mathbf{x}=\mathbf{r}} = f_{\kappa\lambda}; \qquad (77)$$

[1] A famous hypothesis (returning in the idea of "maximum likelihood") regarding α has been advanced by Gauss.

[2] Remember that in a normal distribution with variance s^2, $h^2 = \frac{1}{2}s^2$ is often called the "precision."

f is supposed bounded and has bounded and continuous derivatives of first and second order. *All expectations will be computed with respect to the inference distribution* (51′).

In analogy to the a and s^2 of (68′) we define now

$$a = \sum_{i=1}^{k} f_i r_i , \qquad s^2 = \sum_{i=1}^{k} (f_i - a)^2 r_i = \sum_{1}^{k} f_i^2 r_i - a^2. \tag{78}$$

We shall show that, under slight restrictions, formula (68) or (69) remains valid if we make an inference on $f(x_1 , x_2 , ..., x_k)$, a function of the unknown probabilities.

6.2. *Preparatory considerations.* We shall need mean value and variance of x_i with respect to the distribution (51′). Just as in Section 3.4 (p. 343) we may then consider the following situation. The ith label has appeared $n_i = nr_i$ times, and the label "non-i" has appeared $n(1 - r_i)$ times. We have, therefore, to compute mean value and variance of a one-dimensional probability only [a marginal distribution of (51′)], given in (8) or in similar formulas. It is obvious that $E_n[x]$ with respect to $q_n(x)$ will depend on p_0 as long as n is small. We shall deal, however, with large n only. Now the larger n, the more the distribution $q_n(x)$ of (8) concentrates around $n_1/n = r$ and eventually only an arbitrarily small vicinity of r enters into the computation of $E_n[x]$. If, as before in this chapter, p_0 is assumed bounded, continuous in the neighborhood of $\mathbf{r}$ and different from zero at $\mathbf{r}$, we can consider p_0 as constant in this very small neighborhood of $\mathbf{r}$, which enters into the computation of $\lim_{n\to\infty} E_n[x]$ and obtain [compare with Eq. (16)] $\lim_{n\to\infty} E_n[x] = r$. [The same result follows from the last generalization of the law of large numbers (p. 344), since each x_i is a statistical function $f(x_1 , x_2 , ..., x_k)$ and r_i is the value corresponding to x_i .]

We conclude in the same way that we may determine $\lim_{n\to\infty} E_n[(x-r)^2]$ by computing it for constant p_0 . We had from (18) and (19) the value of the variance, which is asymptotically $(1/n)\, r(1 - r)$ [in agreement with our result for the variance of the limit distribution (47) of $q_n(x)$].

We obtain in this way the formulas

$$\begin{aligned} &\lim_{n\to\infty} E_n[x_i] = r_i , \qquad i = 1, 2, ..., k, \\ &\lim_{n\to\infty} nE_n[(x_i - r_i)^2] = r_i(1 - r_i). \end{aligned} \tag{79}$$

We shall need these results.

We shall use a lemma (v. Mises, 1936) which will be proved and put to full use in Chapter XII. It is, however, easily understandable without proof. In sufficient generality for present purposes it reads:

LEMMA. Consider a k-dimensional collective with distribution $P_n(x_1, x_2, ..., x_k)$ and two functions A_n and B_n of the k variables and of n. *Let $F_n(x)$ and $G_n(x)$ be the distributions of A_n and B_n*, viz.,

$$F_n(x) = \text{Prob}\{A_n \leqslant x\}, \qquad G_n(x) = \text{Prob}\{B_n \leqslant x\}. \tag{80}$$

If $\lim_{n\to\infty} G_n(x) = G(x)$ *exists and if for any* $\epsilon > 0$

$$\lim_{n\to\infty} \text{Pr}\{ \mid A_n - B_n \mid \geqslant \epsilon\} = 0, \tag{81}$$

then, at all points where $G(x)$ is continuous,

$$\lim_{n\to\infty} F_n(x) = G(x) \tag{82}$$

holds. Note that (81) certainly holds if

$$\lim_{n\to\infty} E[\mid A_n - B_n \mid] = 0, \tag{83}$$

where $E[\;]$ is with respect to P_n.

6.3. *Proof of the inference theorem.* We use $\mathbf{x}$ and $\mathbf{r}$ for $(x_1, x_2, ..., x_k)$ and for $(r_1, r_2, ..., r_k)$ and apply Taylor's formula of first order to the given statistical function $f(x_1, ..., x_k) = f(\mathbf{x})$

$$f(\mathbf{x}) - f(\mathbf{r}) = \sum_{\kappa=1}^{k} (x_\kappa - r_\kappa) f_\kappa + R \tag{84}$$

$$R = \tfrac{1}{2} \sum_{\kappa,\lambda}^{1...k} (x_\kappa - r_\kappa)(x_\lambda - r_\lambda) f_{\kappa\lambda}(x'), \tag{85}$$

where x' is a point on the segment from $\mathbf{r}$ to $\mathbf{x}$; $f(x_1, ..., x_k)$ is assumed bounded in the domain

$$D: \quad x_1 \geqslant 0, ..., x_k \geqslant 0, \qquad x_1 + x_2 + \cdots + x_k = 1 \tag{86}$$

and has continuous and bounded derivatives of first and second order at least in a subdomain D_1 of D which includes the point $\mathbf{r}$.

Consider a and s^2 defined by Eqs. (78). In order to be sure that

$s^2 > 0$, we assume that there are at least two subscripts α, β, such that, for $\eta > 0$, n large, $r_\alpha > \eta$, $r_\beta > \eta$, $f_\alpha \neq 0$, $f_\beta \neq 0$, $|f_\alpha - f_\beta| > \eta$. Then

$$\begin{aligned} s^2 &\geqslant r_\alpha(f_\alpha - a)^2 + r_\beta(f_\beta - a)^2 \\ &\geqslant 2\eta\left[\left(\frac{f_\alpha - f_\beta}{2}\right)^2 + \left(\frac{f_\alpha + f_\beta}{2} - a\right)^2\right] > \frac{\eta^3}{2}. \end{aligned} \tag{87}$$

We now define

$$A_n = \frac{\sqrt{n}}{s}[f(\mathbf{x}) - f(\mathbf{r})], \qquad B_n = \frac{\sqrt{n}}{s}\sum_{\kappa=1}^{k}(x_\kappa - r_\kappa)f_\kappa . \tag{88}$$

Then, from (84),

$$A_n - B_n = \frac{\sqrt{n}}{s}R. \tag{89}$$

The a and s^2 of (78) are mean value and variance of an arithmetical distribution with jumps of magnitude r_κ at f_κ. Actually, this is an observed or empirical arithmetical distribution and a, s^2 would be better called average and dispersion. The theoretical mean value of the arithmetical distribution which at the points f_κ has the probabilities x_κ, $\kappa = 1, 2, \ldots, k$, is

$$x = \sum_{\kappa=1}^{k} f_\kappa x_\kappa . \tag{90}$$

We know from the "second basic theorem" of Section 4.4 that x is asymptotically normal with mean value a and variance s^2/n.

Now consider our B_n:

$$B_n = \frac{\sqrt{n}}{s}\left(\sum x_\kappa f_\kappa - \sum r_\kappa f_\kappa\right) = \frac{\sqrt{n}}{s}(x - a). \tag{91}$$

Therefore, B_n, like the u of (69) is normally distributed, $N(0, 1)$. Hence, the $G(x)$ of our lemma equals $\Phi(x)$, which is everywhere continuous.

We wish now to estimate the expectation of $|A_n - B_n| = (\sqrt{n}/s)|R|$, with R given by (85). Since all second derivatives of f are bounded, $|f_{k\lambda}| < M_2$, we obtain

$$E_n[|R|] \leqslant \tfrac{1}{2}M_2 E_n\left[\sum_{\kappa,\lambda}^{1\ldots k} |(x_\kappa - r_\kappa)| \cdot |(x_\lambda - r_\lambda)|\right].$$

For the expectation of $(x_\kappa - r_\kappa)^2$ we use (79) and the double products are estimated by applying Schwarz's inequality. Then

$$E_n[|\, R \,|] \leqslant \frac{k}{2} M_2 E_n \left[\sum_1^k (x_i - r_i)^2\right] = \frac{k}{2n} M_2 E_n \left[n \sum_{i=1}^k (x_i - r_i)^2\right] .$$

Therefore

$$\lim_{n\to\infty} E_n[|\, A_n - B_n \,|] \leqslant \lim_{n\to\infty} \frac{\sqrt{n}}{s} \frac{k}{2} M_2 \cdot \frac{1}{n} E_n \left[\sum_{i=1}^k (x_i - r_i)^2 \cdot n\right]$$

$$\leqslant \lim_{n\to\infty} \frac{k}{2s} M_2 \frac{1}{\sqrt{n}} \sum_{i=1}^k r_i(1 - r_i) = 0. \tag{92}$$

Thus (82) has been proved and we have the result:

Let n observations be performed on the same k-valued arithmetical collective with (unknown) probabilities $x_1, x_2, \ldots, x_k$. Let $x = f(x_1, x_2, \ldots, x_k)$ be an arbitrary function of these x_κ, which is bounded, and has continuous and bounded first and second derivatives in a certain region. With the notations (77) we assume that there are at least two subscripts α and β such that

$$r_\alpha > \eta, \qquad r_\beta > \eta, \qquad |\, f_\alpha - f_\beta \,| > \eta. \tag{93}$$

For the prior density the assumptions of the theorem of Section 4.4 hold.

Then the posterior density of $x = f(x_1, x_2, \ldots, x_k)$ is asymptotically normal with mean value $f(r_1, r_2, \ldots, r_k)$ and variance s^2/n as given in (78).

This theorem with similar conditions was stated by v. Mises in 1935 without proof.[1] It is much simpler to prove than the analogous "direct" theorem (see Chapter XII, Section 5.4). A more general version where the original collective is not assumed arithmetical does not seem difficult to prove but has not yet been proved.

D. Rare Events (Section 7)

7. *Inference on the Probability of Rare Events*

7.1. *Problem and solution.* The passage to the limit $n \to \infty$ as carried out in the foregoing section loses its usefulness as an approximation of

[1] Deux nouveaux théorèmes de limite dans le calcul des probabilités." *Rev. Fac. Sci. Univ. Istanbul* 1 (1935), pp. 61–80.

$q_n(x)$ in such cases where, in spite of large n, the number n_1 of successes is so small that the value of r is close to zero. Let us derive, in analogy to the Poisson formula in the Bernoulli problem for rare events, a similar expression giving an approximation to the inference probability $q_n(x)$ for large n and moderate $n_1 = rn$.

We rewrite Eq. (8) using $u = nx$ as new variable

$$q_n(x) = \text{const.}\, p_0(x)(1-x)^{n-n_1}x^{n_1} = C_n p_0\left(\frac{u}{n}\right)\left(1-\frac{u}{n}\right)^n\left(1-\frac{u}{n}\right)^{-n_1}u^{n_1}, \tag{94}$$

where the factor n^{-n_1} has been absorbed in C_n. Again, the assumption is made that $p_0(x)$ is non-vanishing and continuous at the decisive point, which is now $x = 0$. Then, the first factor after the constant tends toward $p_0(0)$, the following toward e^{-u} as n increases toward infinity while $|u| < U_0$, where U_0 is an arbitrarily large but fixed number. The third expression consists of a finite number, n_1, of terms, each of which has the limit 1; therefore, it goes toward unity. We thus find

$$\lim_{n\to\infty}\frac{q_n(u/n)}{C_n p_0(u/n)} = u^{n_1}e^{-u}. \tag{95}$$

If we use this as an approximation for large n, where $p_0(0)$ is absorbed in C_n', we have

$$q_n(x) \sim C_n' u^{n_1}e^{-u} = C_n' e^{-nx}(nx)^{n_1}, \tag{95'}$$

the constant C_n' is computed from

$$1 = \int_0^1 q_n(x)\,dx = C_n'\int_0^1 e^{-nx}(nx)^{n_1}\,dx = C_n'\frac{1}{n}\int_0^n u^{n_1}e^{-u}\,du. \tag{96}$$

As $n \to \infty$, the upper limit of the last integral also becomes infinite. The integral then represents the factorial or the Γ function

$$\int_0^\infty e^{-u}u^{n_1}\,du = n_1! = \Gamma(n_1+1), \qquad n_1 \text{ being an integer,}$$

and we obtain with

$$C_n'\frac{n_1!}{n} = 1, \qquad C_n' = \frac{n}{n_1!} \tag{97}$$

the result

$$\lim_{n\to\infty}\frac{1}{n}q_n\left(\frac{u}{n}\right) = \frac{u^{n_1}}{n_1!}e^{-u} \tag{98}$$

or asymptotically,

$$q_n(x) \sim \frac{n}{n_1!}e^{-nx}(nx)^{n_1} = \frac{n}{n_1!}u^{n_1}e^{-u}. \tag{98'}$$

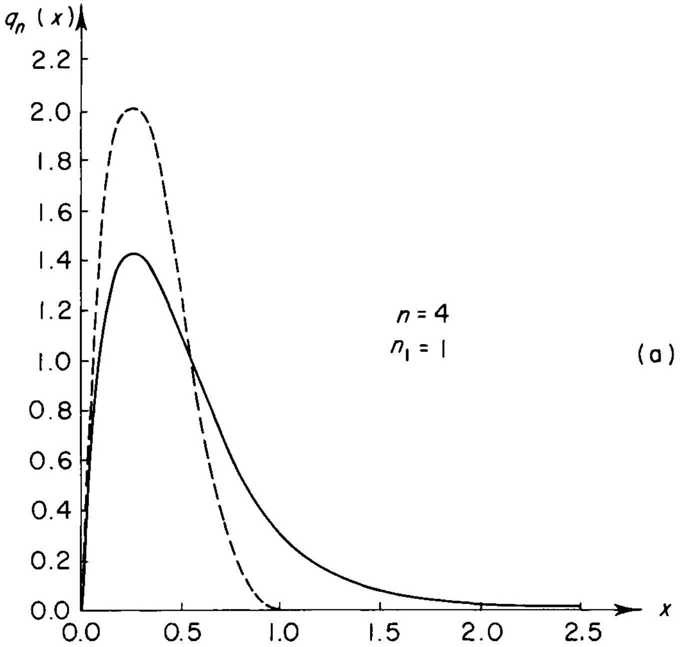
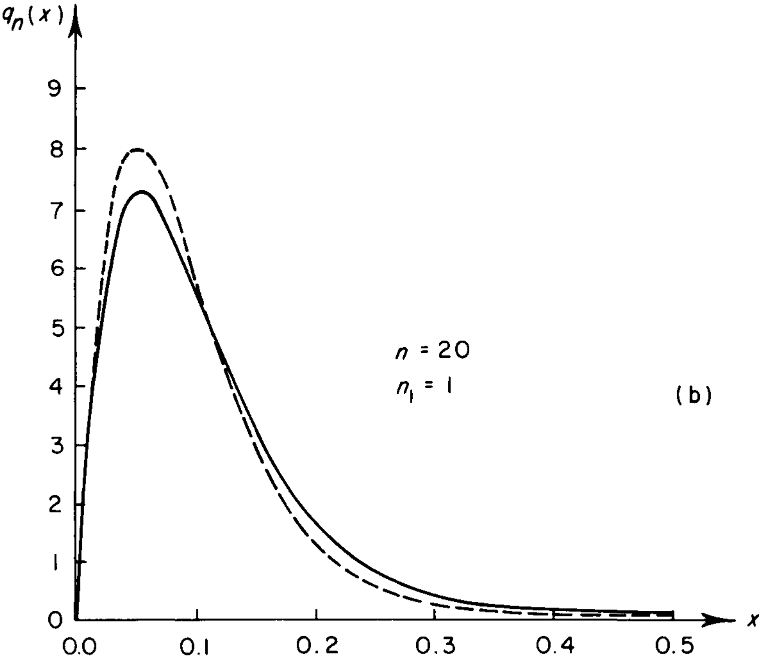
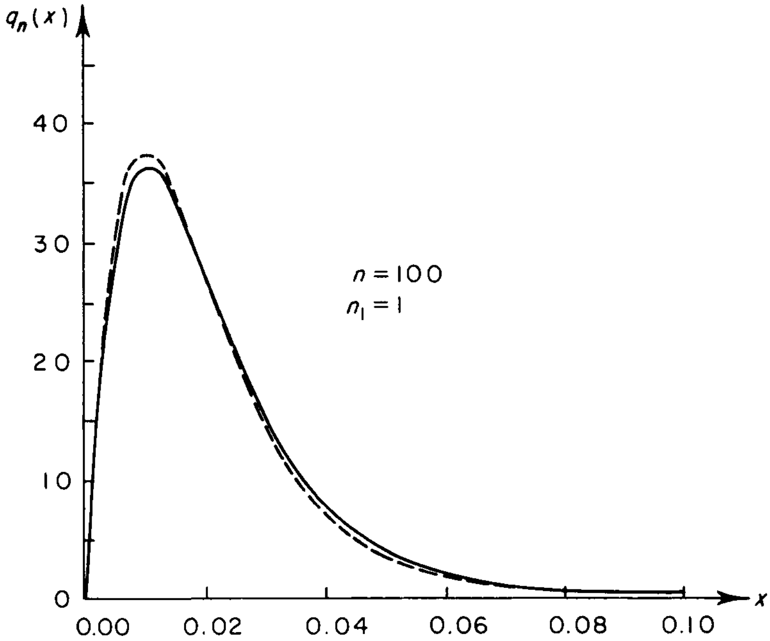


FIG. 21. Inference on rare events. Solid lines approximation; dashed lines exact curves.





(c)

As in preceding similar computations one has to show, if C_n' from (97) is substituted into (96), that $\lim_{n\to\infty} \int_0^1 q_n(x)\,dx = 1$. The integration of both sides of (98) from 0 to ∞ must be justified since we had assumed $|u| < U_0$. The simple additional considerations may be found in the paper quoted in footnote 2 on p. 366 or in v. Mises [21], p. 165.

In (98'), $q_n(x)$ is the probability density with respect to x. It is however, more convenient to refer the density to $u = nx$, to use a function $q_n'(u)$ defined by $q_n'(u)\,du = q_n(x)\,dx$, and to write (98') as

$$q_n'(u) \sim \frac{1}{n_1!} e^{-u} u^{n_1} . \qquad (98'')$$

As x is the probability of a single success, $nx = u$ is the expectation of the number of successes in n trials. Our result can be interpreted as *an inference from an observed number of rare events to the probability of the expected value of this number.*

We formulate: *If in a great number n of observations an event has appeared n_1 times where n_1 is a moderate number, then the probability* *density of the unknown probability x of the event is given by* (98′) *and that of the expectation u of the event is given by* (98″).

Figures 21a, b, c, show the correct curves computed by (94) with p_0 = constant as compared with the approximation (98′), for $n_1 = 1$ and $n = 4, 20, 100$, respectively. The last one shows very close agreement with the correct line.

7.2. *Discussion.* The distribution defined by the right-hand side of (98″) has the form of the so-called chi-square distribution. The density starts with zero at $u = 0$, reaches a maximum value at $u = n_1$, and tends asymptotically to zero as u increases indefinitely. The mean value is determined by

$$a = \int_0^\infty u q_n'(u)\, du = \frac{1}{n_1!}\int_0^\infty e^{-u} u^{n_1+1}\, du = \frac{(n_1+1)!}{n_1!} = n_1 + 1.\ ^1 \qquad (99)$$

In the same way, the variance is found as

$$s^2 = \int_0^\infty u^2 q_n'(u)\, dx - a^2 = \frac{(n_1+2)!}{n_1!} - (n_1+1)^2 = n_1 + 1. \qquad (100)$$

The corresponding cumulative distribution function, i.e., the probability that the expected value is not greater than u,

$$Q_n'(u) = Q_n(x) = \int_0^u q_n'(u)\, du = \int_0^x q_n(x)\, dx \qquad (101)$$

depends on the so-called incomplete gamma function. If we write

$$\int_0^\rho e^{-u} u^{\lambda-1}\, du = \Gamma(\lambda, \rho), \qquad (102)$$

we have

$$Q_n'(u) = Q_n(x) = \frac{1}{n_1!}\Gamma(n_1 + 1, u). \qquad (103)$$

Tables for $\Gamma(\lambda, \rho)$ are available (K. Pearson, *Tables of the Incomplete Γ-Function*, London, 1922). Moreover, approximations can easily be computed by Simpson's rule, etc.[2]

[1] Note that a does not coincide with the mode n_1.

[2] The results of this section were first given by H. GEIRINGER, "Rückschluss auf die Wahrscheinlichkeit seltener Ereignisse." *Z. Angew. Math. Mech.* 5 (1925), pp. 493–501.

Other aspects and generalizations of the problem of probability inference will be discussed in the second part of this book which is devoted to the theory of statistics, in particular, in Chapter X which is primarily based on Bayesian inference.

Problem 11. In Germany in the year 1911, 6 persons died at an age of over 102 years. What was the probability that the expected value of such cases fell in the interval 5 to 7 ? (Compute the densities for 5, 6, 7 and use Simpson's rule.)