ON THE CORRECT USE OF BAYES' FORMULA

By R. v. Mises
Harvard University

The problem that we try to solve by using Bayes' formula consists in making an inference from an observed statistical value upon the unknown value of a parameter, and in examining the chance of this inference being correct. One may call this the principle problem of practical statistics or the estimation problem, or, as the author put it in German (Rueckschluss-Wahrscheinlichkeit) problem of inference probability; at any rate we encounter this kind of problem in various forms in almost every branch of statistical investigation. It will be convenient to base the following discussion on a concrete question in quite specified form which will allow us to see clearer the points that are to be stressed in this paper.

1. The problem. In examining the quality of water supplies with respect to the number of bacterias of a certain kind they contain, a definite procedure is usually adopted. One takes $n = 5$ samples out of the water, each sample of exactly 10 ccm. Then by a certain biological test one finds out whether or not each sample contains at least one bacteria of the kind under consideration. The number x (zero to five) of positive tests is the observed value from which an inference is drawn upon the probability θ for a sample containing at least one bacteria. It is assumed that this θ is connected with the average number λ of bacterias per 10 ccm by

$$\theta = 1 - e^{-\lambda}; \qquad \theta = \theta_1 = 0.63 \qquad \text{for } \lambda = 1 \tag{1}$$

according to Poisson's law. A particular question which we want to answer is this: What is the chance of being right, if we conclude from the observed fact $x = 0$, (in other cases from $x = 1$) that θ lies between 0 and $\theta_1 = 0.63$ (or λ between 0 and 1)?

For a given θ the probability of getting x positive tests out of n tests is according to Bernoulli's formula

$$p(x \mid \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}. \tag{2}$$

The chance of having a θ-value between 0 and θ_1 when x positive tests are observed is according to Bayes' formula

$$P_x(\theta_1) = \frac{\int_0^{\theta_1} p(x \mid \theta)\, dP(\theta)}{\int_0^1 p(x \mid \theta)\, dP(\theta)} \tag{3}$$

where $P(\theta)$ is a distribution function, monotonically increasing from 0 to 1 and usually known as the *a priori probability.*

2. The apriori. The function $P(\theta)$ is generally considered as a troublemaker. As one uses to call P the a priori probability most people think that it has something to do with those absurd conceptions of non-empirical, a priori known probabilities that cannot be tested by any experiments etc. This cannot be strongly, enough refuted. In our particular case the meaning of $P(\theta)$ is the following. Each probability statement refers, as we know, to a certain infinite sequence of experiments or trials which form a kollektiv. If we ask for the chance $P_x(\theta_1)$ of having a θ-value between zero and θ_1 when a certain x has been observed, we have in mind a sequence of trials each consisting of two steps, first, picking out one particular water supply, and then testing the number x of samples that contain bacillas. Among the first N trials of this kind we shall have N_1 cases where the θ-value for the water supply picked out lies between 0 and θ_1, then we shall have N_x cases where the number of positive tests is x, and finally in a number N_{1x} of cases both conditions will be fulfilled. The chance $P_x(\theta_1)$ we ask for is then by definition

$$P_x(\theta_1) = \lim_{N\to\infty} \frac{N_{1x}}{N_x}, \tag{4}$$

while the so-called a priori probability is

$$P(\theta_1) = \lim_{N\to\infty} \frac{N_1}{N}. \tag{5}$$

Later on we shall also use the probability

$$Q_x = \lim_{N\to\infty} \frac{N_x}{N}. \tag{6}$$

All these magnitudes are to the same extent empirical or non-empirical. They are "empirical," since we get approximate values for them out of a long sequence of experiments, and they may be considered as something super-empirical since the concepts of an infinite sequence and of a limit are used in the definition—as each theory must involve a certain amount of "idealization."

In order to avoid the above mentioned equivocation the author had suggested a long time ago[1] to call the probabilities corresponding to $P(\theta)$ and $P_x(\theta)$ respectively the *initial* and the *final* probability. Another expression which could be used in connection with the distribution function $P(\theta)$ is *overall distribution,* since it means the distribution of θ-values within the total mass of samples, not regarding what the values of x are in each case.

3. No randomness required. Now, the first remark we have to make is the following: In the Bayes' formula (3) the existence of a function $P(\theta)$ is presup-

[1] Cf. reference [2], p. 152.

posed, i.e. we assume that in the sequence of successive trials the frequency of those cases in which θ falls into a certain region has a definite limit. But nothing is assumed about this limit being independent of a place selection. The sequence of trials must fulfill the first condition of a kollektiv, with respect to θ but not the second; in other words *the randomness in the succession of θ-values is not required.* Thus we may say that θ is not supposed to be a chance variable in the usual sense of this term. Sometimes people are shocked by the idea that in Bayes' theory the individual cases are supposed to be picked out at random, and it is often considered as a superiority of the method of confidence intervals that here such assumption is avoided.

It is true that in the latter method even the existence of the frequency limit is not required,[2] but this does not seem to make any essential difference. The fact is that, if we want to make an inference upon the value of θ i.e. an assertion about the chance of θ falling into a certain interval, we have to assume that in the long run different θ-values may occur with certain frequencies.

It may be useful to have different expressions for the two cases where a frequency limit is or is not supposed to be independent of an arbitrary place selection. As we use the word probability in the first case it seems suitable to apply the word *chance* in the second. Thus, if $P(\theta)$ is the initial or the over all chance of θ we would say that $P_x(\theta_1)$ is the final chance of θ being smaller than or equal to θ_1 for a certain observed x-value. When $P(\theta)$ is supposed to be a probability, i.e. to fulfill the condition of randomness, then $P_x(\theta_1)$ will have this property too and has to be called probability.

4. Inequalities for the final chance $P_x(\theta)$. A much better founded objection against the practical application of Bayes' formula consists in saying that in most cases we have no sufficient information about the function $P(\theta)$. This undeniable fact leads often to an incorrect simplification of the formula by replacing in it $dP(\theta)$ by $d\theta$ which means an a priori probability of constant density. It is obvious that this is no solution: if you do not know what $P(\theta)$ is, to assume it equal to θ. On the other hand, if we accept Bayes' formula as correct (and there is no reason for not doing so) we learn that the value $P_x(\theta)$ we ask for *depends essentially on* $P(\theta)$, and is undetermined as far as $P(\theta)$ is undetermined. The only consequence in this situation is, first to use all information we can get about $P(\theta)$, and then to make the answer as vague or undetermined as the incompleteness of this information requires.

One way to do this consists in setting up inequalities for $P_x(\theta)$ based on certain inequalities for $P(\theta)$. A formula which turns out to be useful, at least in a well-known asymptotic problem is the following:

Let us consider the general case where θ stands for several variable parameters, and let A be the set of all possible values of θ. We are interested in the final probability P_C of a subset C of A given by

---

[2] Cf. reference [4], p. 201.

$$P_C = \frac{\int_{(C)} p(x \mid \theta)\, dP(\theta)}{\int_{(A)} p(x \mid \theta)\, dP(\theta)}, \tag{6}$$

where x is supposed to be known.

Let P_c' be the value of P_c under the assumption of a constant initial density and denote by P_B, P_B' the analogous values for a subset B which includes C so as to have

$$C \prec B \prec A. \tag{7}$$

The quantities P_B' and P_C' depend only on the function $p(x \mid \theta)$ and the sets B and C while P_B and P_C change with $P(\theta)$.

If we assume that the initial density $p(\theta)$ has the limits

$$\begin{aligned} m &\leqq p(\theta) \leqq M \quad \text{within } B \\ m' &\leqq p(\theta) \leqq M' \quad \text{within } A - B, \ (A \text{ minus } B) \end{aligned} \tag{8}$$

it can easily be shown that

$$\frac{m}{M} P_B' + \frac{m'}{M}(1 - P_B') \leqq \frac{P_C'}{P_C} \leqq \frac{M}{m} P_B' + \frac{M'}{m}(1 - P_B'). \tag{9}$$

We may consider the following application of these inequalities.

If we are concerned with a case where a great number n of trials is involved, the function $p(x \mid \theta)$—which determines the P' values—shows an increasing concentration at a certain point of the set A. In other words, for large n we have a subset B more and more reducing to one single point for which P_B' is as near to 1 as we want. If we then assume that the density $p(\theta)$ is continuous and bounded, the difference between m and M tends to zero, and if m is supposed to have a positive lower bound, both the first and the last expression in (9) tend to unit or P_C approaches P_C'. This is a generalized form of the statement which the author proved for the first time in 1919,[3] that in the original Bayes' problem where we are concerned with n repetitive observations of an alternative, *the final probability becomes more and more independent of the initial probability* $P(\theta)$ as the number n of observations involved increases.

5. Using previous experience. The inequalities (9) may be of use in many cases. But to be sure, in general, they are not the basis upon which practical estimation judgments rest. Everybody acquainted with the conditions of testing water supplies takes it for granted that the outcome $x = 0$ (no positive test) supplies a sufficient reason for the statement $\theta \leqq \theta_1 = 0.63$ (less than one

[3] Cf. reference [1], p. 84.

bacteria per 10 cc). But, if nothing were known about the initial distribution $P(\theta)$, we could assume $P(\theta)$ in the form

$$P(\theta) = \theta^m, \qquad p(\theta) = m\theta^{m-1} \qquad \text{for } 0 \leqq \theta \leqq 1,$$

with a large value of m. With $n = 5, x = 0$ equations (2) and (3) give $P_0(\theta_1) = 0.50$ for $m = 10$, and $P_0(\theta_1) = 0.88$ if m is 5. These values are much too low to justify any recommendation of a water supply for which x was found to be zero. Thus we have to ask: What is the *real source of the confidence* we put in the inference from $x = 0$ upon $\theta \leqq \theta_1$?

There is no doubt, that this confidence is based on previous experience. We know that the water supplies subjected to the routine test in the past formed a class of rather clean than dirty water and we rely that a new sample will belong to the same class. The author was given the following information about the results under the jurisdiction of Massachusetts during the last decade. Out of a total of $N = 3420$ examinations there were found

3086 cases with $x = 0$ (no positive test)

279 cases with $x = 1$ (one positive test)

32 cases with $x = 2$

15 cases with $x = 3$

5 cases with $x = 4$

3 cases with $x = 5$

The overwhelming majority of cases with $x = 0$ is evident. The question is only how we can use these statistics of past experiments for obtaining a numerical inference upon the value of $P_x(\theta)$.

If the initial distribution $P(\theta)$ were known, we could find the probability Q_x of getting x positive tests out of n:

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

Using the numbers N_1, N_x, N_{1x} introduced in section 2 the probability $Q(x)$ is defined by equation (6).

If the number N of past examinations is considered as sufficiently large, we can take the ratios 3086/3420, 279/3420 etc. as approximate values for Q_0, Q_1 etc. Now, according to the well-known identities

$$\frac{1}{n} \sum_{x=0}^{n} x \binom{n}{x} \theta^x (1-\theta)^{n-x} = \theta. \tag{11}$$

$$\frac{1}{n(n-1)} \sum_{x=0}^{n} x(x-1) \binom{n}{x} \theta^x (1-\theta)^{n-x} = \theta^2, \tag{12}$$

and using (10) we can derive from the values $Q_0, Q_1, \cdots, Q_n$ *the first and second moments* of the distribution function $P(\theta)$:

$$M_1 = \int_0^1 \theta \, dP(\theta) = \frac{1}{n} \sum_{x=0}^{n} Q_x$$

(13)

$$M_2 = \int_0^1 \theta^2 \, dP(\theta) = \frac{1}{n(n-1)} \sum_{x=0}^{n} x(x-1) Q_x .$$

If we introduce here the above mentioned empirical ratios for Q_x we find the approximate values for the first and second moments of $P(\theta)$:

(13′) $$M_1 = 0.02474 \qquad M_2 = 0.00401.$$

6. Determination of a distribution function by its first moments. In an earlier paper the author showed [3] how the exact upper and lower bounds for a distribution function $P(\theta)$ can be found, if the expected values of two functions $f(\theta)$ and $g(\theta)$ are known. The only condition was that the curve represented in a Cartesian coordinate system by $x = f(\theta)$, $y = g(\theta)$ is convex. Let us take

$$f(\theta) = g(\theta) = 0 \qquad \text{for } \theta < 0$$

(14)

$$f(\theta) = \theta, \qquad g(\theta) = \theta^2 \qquad \text{for } 0 \leqq \theta \leqq 1$$

$$f(\theta) = g(\theta) = 1 \qquad \text{for } \theta > 1.$$

In this case the condition is fulfilled and the expected values of $f(\theta)$ and $g(\theta)$ are the moments M_1, M_2, respectively. The results obtained in the paper quoted above take the following form:

First, we have to derive from the given values M_1 and M_2 two points θ' and θ'' of the internal $0 \leqq \theta \leqq 1$

(15) $$\theta' = \frac{M_1 - M_2}{1 - M_1}, \qquad \theta'' = \frac{M_2}{M_1}.$$

Then the limits for $P(\theta)$ are:

$$0 \leqq P(\theta) \leqq \frac{M_2 - M_1^2}{M_2 - 2M_1\theta + \theta^2} \qquad \text{for } 0 \leqq \theta \leqq \theta'$$

(16)

$$1 - M_1 - \frac{M_1 - M_2}{\theta} \leqq P(\theta) \leqq 1 - \frac{M_1\theta - M_2}{\theta - 1} \qquad \text{for } \theta' \leqq \theta \leqq \theta''$$

$$\frac{(M_1 - \theta)^2}{M_2 - 2M_1\theta + \theta^2} \leqq P(\theta) \leqq 1 \qquad \text{for } \theta'' \leqq \theta \leqq 1.$$

In our case we find $\theta' = 0.0213$, $\theta'' = 0.1619$ and the point $\theta_1 = 0.6321$ falls into the third interval θ'', 1. The lines $O\ A\ B\ C$ and $O\ D\ E\ F\ G$ in Fig. 1 show (slightly distorted) the lower and upper bounds for $P(\theta)$.

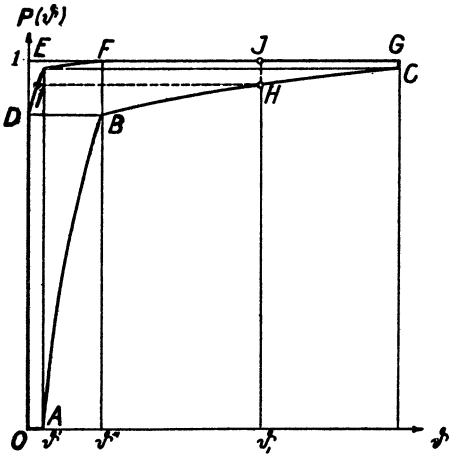

FIG. 1

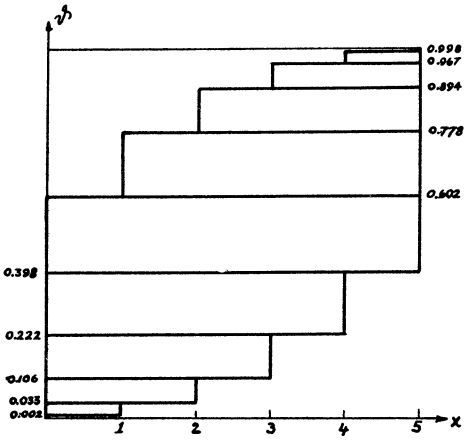

FIG. 2

FIG. 1. The limits of the overall distribution function
FIG. 2. The 99% region in the methods of confidence intervals

7. Application to Bayes' formula. The inequalities (16) enable us to find in a simple way a lower bound for the end probability $P_x(\theta_1)$ defined by (2) and (3) in the case $x = 0$. Let us denote by A the numerator in (3) and by B the supplementary integral

$$B = \int_{\theta_1}^{1} p(x \mid \theta)\, dP(\theta), \tag{17}$$

so as to have $A + B$ for the denominator in (3). If the subscripts min and max denote a lower and upper bound respectively we can write

$$P_x(\theta_1) = \frac{A}{A+B} \geqq \frac{A_{\min}}{A_{\min} + B_{\max}}. \tag{18}$$

Now, taking $x = 0$ we find by product integration

$$A = P(\theta_1)(1 - \theta_1)^n + n \int_0^{\theta_1} P(\theta)(1 - \theta)^{n-1}\, d\theta. \tag{19}$$

Therefore, $A_{\min}$ is found when we introduce in this expression the lower lim for $P(\theta)$ as given in (16). If we do this and use the values for M_1 and M_2 according to (13′), numerical computation leads to $A_{\min} = 0.712$.
In the same way we obtain B in the form

$$B = -P(\theta_1)(1 - \theta_1)^n + n \int_{\theta_1}^{1} P(\theta)(1 - \theta)^{n-1}\, d\theta. \tag{20}$$

The upper bound $B_{\max}$ is reached, if we introduce in the integral $P(\theta) = 1$ and in the first term the minimum value for $P(\theta_1)$ following from (16). The second

term becomes thus equal to $(1 - \theta_1)^n$ and the numerical result is $B_{\max} = 0.0000607$. Therefore the inequality (18) supplies

$$P_0(\theta_1) \geqq \frac{0.712}{0.71206} = 0.99915. \tag{18'}$$

The final outcome secured in this way can be formulated as follows: *If we assume that in continuing the experiments the distribution of test results will be about the same as it has been in the past* 3420 *cases, we have a chance of more than* 99.9% *of being right, when we state in each case of no positive test that the density of bacterias is less than* 1 *per* 10 *ccm.*

The high value of 99.9% for $P(\theta_1)$ is of course strictly bound to the assumption that the entire mass of water supplies to be tested is homogeneous and sufficiently characterized by the distribution of test results found in the past. If e.g. we had to assume that the six possible values for x (0 to 5) in the long run appear with equal frequencies so as to have $Q_0 = Q_1 = \cdots Q_5 = \frac{1}{6}$, the same method would give $M_1 = \frac{1}{2}$, $M_2 = \frac{1}{3}$, then $\theta' = \frac{1}{3}$, $\theta'' = \frac{2}{3}$, and the final result would be $P_0(\theta_1) \geqq 0.73$. The assumption of a constant initial density $P(\theta) = \theta$ would give $P_0(\theta_1) = P_0'(\theta_1) = 0.9975$, a little less than the value found above in (18').

8. The case $x = 1$. The results are less favorable in the case of one positive test, $x = 1$. Here we have

$$p(1 \mid \theta) = n\theta(1 - \theta)^{n-1} = 5\theta(1 - \theta)^4, \tag{21}$$

and the derivative of p is first positive, then negative. We can conclude from Fig. 1 that the minimum value for A and the maximum for B will be reached when the distribution function $P(\theta)$ is represented by the line $O\ D\ I\ H\ J\ G$ where $I\ H$ is horizontal and H the point on $B\ C$ with abscissa θ_1. The abscissa θ_0 of I is determined by the equation

$$\frac{M_2 - M_1^2}{M_2 - 2M_1\theta_0 + \theta_0^2} = \frac{(M_1 - \theta_1)^2}{M_2 - 2M_1\theta_1 + \theta_1^2}, \tag{22}$$

which supplies $\theta_0 = 0.0190$. We then have

$$A_{\min} = \int_0^{\theta_0} p(1 \mid \theta)\, dP(\theta), \tag{23}$$

with the value $p(1 \mid \theta)$ from (21) and with

$$P(\theta) = \frac{M_2 - M_1^2}{M_2 - 2M_1\theta + \theta^2}$$

according to (16). On the other hand $B_{\max}$ is found, as in the former case, to be

$$B_{\max} = p(1 \mid \theta_1)[1 - P(\theta_1)], \tag{24}$$

where we have to take for $P(\theta_1)$ its minimum value according to (16). The numerical computation yields $A_{\min} = 0.0062$ and $B_{\max} = 0.00052$ so as to give

$$P(\theta_1) \geqq \frac{62}{67.2} = 0.92.$$

The result is that—under the assumption above mentioned—*we have more than 92% chance of being right*, if we predict each time one out of five tests has been positive that the density of bacilli is less than 1 per 10 ccm.—The chance computed under the assumption of a uniform initial distribution $P(\theta) = \theta$ would be 0.97.

9. The method of confidence intervals. One may ask what kind of answer to our questions can be deduced from the principle of confidence intervals. This method has undeniably to its credit that no use is made here of the initial distribution $P(\theta)$ and that, therefore, all its statements are completely independent of what is assumed about $P(\theta)$.

In order to apply this method[4] we have to select for a given degree of confidence, say $\alpha = 0.99$, a region of acceptance, i.e. an area in the two dimensional x, θ plane limited by two lines $x_1(\theta)$ and $x_2(\theta)$ so as to have for each θ

$$\text{Prob}\,\{x_1(\theta) \leqq x \leqq x_2(\theta)\} = \alpha. \tag{25}$$

The region is, of course, not uniquely determined by (25). In our case, however, one will generally agree that the best way to determine the region consists in assuming for $x_1(\theta)$ and $x_2(\theta)$ two step lines with steps at the integer values $x = 0, 1, 2, \cdots$ as indicated in Fig. 2. Then the formula (2) for $p(x \mid \theta)$ combined with (25) supplies the abscissae of the steps, if x is given. If we transform the limits for θ into limits for λ using equation (1), the final outcome reads as follows:

Whatever the initial distribution $P(\theta)$ may be, we have a chance of 99% of being right, if we predict:

each time $x = 0$ is observed that λ lies between 0 and 0.92,
each time $x = 1$ is observed that λ lies between 0.002 and 1.51,
each time $x = 2$ is observed that λ lies between 0.036 and 2.24,
each time $x = 3$ is observed that λ lies between 0.112 and 3.41,
each time $x = 4$ is observed that λ lies between 0.25 and 8.48,
each time $x = 5$ is observed that λ lies between 0.51 and ∞.

It is true that in this way we obtain a result independent of any assumption on $P(\theta)$. But it is essential that the chance of $\alpha = 99\%$ holds only *for the six joint statements as a whole*. This means it may happen that for instance the first assertion (that λ is smaller than 0.92 in the case $x = 0$) is correct but very seldom or even never, while other assertions (e.g. those for $x = 4$ and 5) have

[4] Cf. reference [5] and reference [4], p. 203.

a much greater chance than 99% of being correct. Whether this happens or not depends on the initial distribution $P(\theta)$. As long as we know nothing about $P(\theta)$ we are not in the position to conclude, by using the method of confidence intervals, that the particular statement "$\lambda \leqq 0.92$ if $x = 0$" has a chance of 99% or even any chance at all of being correct. On the other hand, when $x = 0$ has been observed we are in no way interested in consequences that may be drawn in the case $x = 4$ or $x = 5$ or in a set of statements that includes the cases $x = 4$ and $x = 5$. The only practical question that is relevant to the purpose for which the tests are made is this: *What can we conclude from the fact that in a certain instance* $x = 0$ *has been observed* (or in another instance $x = 1$)? It seems that the method of confidence intervals, discarding any consideration of the initial distribution, can supply no contribution towards the answering *this particular question.*

REFERENCES

[1] R. v. MISES, "Fundamentalsaetze der Wahrscheinlichkeitsrechnung," *Math. Zeit.*, Vol. 4 (1919), pp. 1–97.

[2] R. v. MISES, *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und Theoretischen Physik*, Leipzig und Wien, 1931.

[3] R. v. MISES, "The limits of a distribution function if two expected values are given," *Annals of Math. Stat.*, Vol. 10 (1939), pp. 99–104.

[4] R. v. MISES, "On the foundation of probability and statistics," *Annals of Math. Stat.* Vol. 12 (1941), pp. 191–205.

[5] J. NEYMAN, *Roy. Stat. Soc. Jour.*, Vol. 97 (1934), pp. 590–592.