ON BAYES PROCEDURES FOR A PROBLEM WITH CHOICE OF OBSERVATIONS[1]

By T. W. Anderson

Columbia University

1. Introduction. In establishing statistical means to decide between two hypotheses, H_0 and H_1, an experimenter may have the choice of observing a variable X alone or of observing two variables X and Y. While observation of the two variables is more informative than observation of the one variable, it is also more expensive. The question is whether it is worthwhile for the experimenter to pay the greater cost necessary for the two variables. He must make the decision whether to observe Y before the observation on X is made.

As an example, in a medical study X may be a sufficient statistic for a sample of m individuals who have been treated with a drug and are undergoing observation. Because the treatment and observation take a long period of time, before he has completed study of the first m individuals treated the investigator may consider treating n additional individuals, Y being the sufficient statistic for the second sample. This example is covered by the general study if X and Y are the sums (or means) of the observations in the two samples, respectively, and the measurements are normally distributed with known and common variance.

The situation considered here is different from the usual two-sample or sequential situation in that the decision whether to observe Y is made independently of the observation X. As a matter of fact, this problem arose as a simple analogue of a problem of finding Bayes and admissible procedures for deciding between two hypotheses H_0 and H_1 when observations are taken sequentially and after the decision to stop observation has been taken, m more observations (corresponding to X) are obtained, as for instance in clinical trials [Anderson (1964)].

This study also applies to a problem of classification in multivariate statistical analysis. Suppose that an investigator wants to classify an individual as coming from one of two populations. The measurements he may make have joint normal distributions in the two populations; the populations are the same in variances and correlations, but differ in means. The investigator may be able to observe either the set of measurements $z_1, \cdots, z_m$ or the set of measurements $z_1, \cdots, z_{m+n}$; for example, the first m measurements may be made by one device and are required, and the last n measurements may be made by another device and are optional. Does it pay the investigator to observe the second set of measurements in addition to the first set? This problem will be treated explicitly at the end of Section 2 as a special case of the general problem.

We formulate our problem more precisely by assuming that there is a loss

Received 17 December 1963; revised 4 March 1964.

[1] Research sponsored by Contract AF 41(609)-1534 between the USAF School of Aerospace Medicine and Teachers College, Columbia University.

W_0 for rejecting H_0 when H_0 is true and a loss W_1 for rejecting H_1 when H_1 is true. Let C be the cost of observing Y (or the difference between the cost of observing X and Y and the cost of observing X). We shall assume that X and Y are independently distributed with densities (or probability functions) $f_k(x)$ and $g_k(y)$ under H_k, $k = 0, 1$, respectively. Then the best procedures (admissible or Bayes) are based on the likelihood ratios $f_1(x)/f_0(x)$ and $f_1(x)g_1(y)/[f_0(x)g_0(y)]$ in the two cases. If only X is observed, the decision will be to accept H_0 or H_1 according to whether $\log[f_1(x)/f_0(x)]$ is less than or greater than a number a. If both X and Y are observed, the decision will be to accept H_0 or H_1 according as $\log[f_1(x)/f_0(x)] + \log[g_1(y)/g_0(y)]$ is less than or greater than some number b. In each case there may be randomization if the logarithm of the likelihood ratio is equal to the constant.

If g_0 is an a priori probability of H_0 and g_1 $(= 1 - g_0)$ is the corresponding a priori probability of H_1, the expected loss in each case is minimized by taking the constant as $\log[W_0g_0/(W_1g_1)]$. The minimum expected loss when X is observed, say $\rho(g_0)$, is the weighted average of the two probabilities of errors in that case; the minimum expected loss when X and Y are observed, say $\sigma(g_0)$, is the weighted average of the losses in that case plus C. We ask for what values of g_0 $\rho(g_0)$ is less than, equal to, or greater than $\sigma(g_0)$, respectively. The specific problem considered here is whether we can characterize the Bayes solutions in a simple way.

If $g_0 = 0$, the Bayes procedure in either case is to accept H_1, and then $\rho(0) = 0$ and $\sigma(0) = C$. Similarly, $\rho(1) = 0$ and $\sigma(1) = C$. Thus for a small value of g_0 or a large value of g_0 the Bayes procedure is based on observing X alone. The question we raise is whether the values of g_0 for which the Bayes procedures are based on X and Y constitute an interval.

The answer to the question depends on the distribution of X and Y. In Section 2 we show that for normal distributions (with variances common to H_0 and H_1) the values of g_0 for which the Bayes procedures use X and Y is an interval. In Section 3 we give an example in which that set of g_0 consists of several intervals.

In most problems the set of Bayes procedures is also the set of admissible procedures. In the type of problem considered here, if the set of g_0 for which the Bayes procedures use X and Y is an interval, the admissible procedures can be classified into five cases. For small probabilities of Type II error and relatively large probabilities of Type I error admissible procedures for our problem are admissible procedures for the hypothesis-testing problem with X alone (corresponding to g_0 small); for intermediate probabilities of Type I and Type II errors, admissible procedures for our problem consist of admissible procedures for the hypothesis-testing problem with both X and Y observed (corresponding to intermediate values of g_0); for relatively large probabilities of Type II error and small probabilities of Type I error admissible procedures here are again admissible based on X alone (corresponding to g_0 large). Between each successive pair of the above classes is a class of admissible procedures, each of which is a randomization between an admissible procedure based on X alone and an ad-

missible procedure based on both X and Y (each such class corresponding to a single value of g_0).

We treat in this paper only simple hypotheses H_0 and H_1. Some problems involving composite hypotheses can be related to these; for instance, if the variance of a normal distribution is known, good procedures for deciding whether the mean is negative or positive may be approximated by good procedures for deciding whether the mean is a given negative number or a given positive number. However, the more common statistical problem of deciding about the mean when the variance is unknown is not easily approximated by a problem of simple hypotheses.

2. Normal distributions. Suppose that X and Y are independently normally distributed with means $-\gamma$ and $-\mu$ under H_0 and γ and μ under H_1 and variances τ^2 and σ^2, respectively. The logarithm of the likelihood ratio for X is $2\gamma x/\tau^2$ (Section 6.4 of [1], for example), and the logarithm of the likelihood ratio for X and Y is $2\gamma x/\tau^2 + 2\mu y/\sigma^2$. Then $2\gamma x/\tau^2 = u$, say, has the normal distribution $N(\pm\lambda, 2\lambda)$, where $\lambda = 2\gamma^2/\tau^2$, and $2\mu y/\sigma^2 = v$, say has the normal distribution $N(\pm\nu, 2\nu)$, where $\nu = 2\mu^2/\sigma^2$.

A procedure based on u alone will accept H_0 if $u < a$ and accept H_1 if $u > a$. Then the probability of rejecting H_0 when it is true is

$$\Pr\{u > a \mid H_0\} = \Phi\left(\frac{-a-\lambda}{(2\lambda)^{\frac{1}{2}}}\right), \tag{1}$$

where

$$\Phi(w) = \int_{-\infty}^{w} \phi(t)\,dt, \qquad \phi(t) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}t^2}. \tag{2}$$

Similarly, the probability of rejecting H_1 when it is true is

$$\Pr\{u < a \mid H_1\} = \Phi\left(\frac{a-\lambda}{(2\lambda)^{\frac{1}{2}}}\right). \tag{3}$$

Given g_0, the Bayes procedure is to accept H_0 if $u < \log [W_0 g_0/(W_1 g_1)]$ and accept H_1 if $u > \log [W_0 g_0/(W_1 g_1)]$. The expected loss (or risk) using u alone is

$$W_0 g_0 \Phi\left(\frac{-\log[W_0 g_0/(W_1 g_1)] - \lambda}{(2\lambda)^{\frac{1}{2}}}\right) + W_1 g_1 \Phi\left(\frac{\log[W_0 g_0/(W_1 g_1)] - \lambda}{(2\lambda)^{\frac{1}{2}}}\right). \tag{4}$$

Similarly, the expected loss (or risk) using u and v is

$$W_0 g_0 \Phi\left(\frac{-\log[W_0 g_0/(W_1 g_1)] - (\lambda+\nu)}{[2(\lambda+\nu)]^{\frac{1}{2}}}\right) + W_1 g_1 \Phi\left(\frac{\log[W_0 g_0/(W_1 g_1)] - (\lambda+\nu)}{[2(\lambda+\nu)]^{\frac{1}{2}}}\right) + C. \tag{5}$$

The difference between the two expected losses [(4)–(5)] as a function of g_0 is $-C$ at $g_0 = 0$ and $g_0 = 1$. We show that the set of g_0 for which the difference is

nonnegative is an interval by showing that the derivative of the difference between the two expected losses is positive in an interval $0 < g_0 < k$ for some k and negative in an interval $k < g_0 < 1$. Hence the difference is a function of g_0 that is monotonically increasing in the interval $0 < g_0 < k$, has a maximum at $g_0 = k$, and is monotonically decreasing in the interval $k < g_0 < 1$.

If the logarithm of the likelihood ratio Z in any Bayes problem has a density under each of the two hypotheses, the Bayes procedure is defined by a value a to minimize

$$W_0 g_0 \Pr\{Z > a \mid H_0\} + W_1 g_1 \Pr\{Z \leqq a \mid H_1\}; \tag{6}$$

this value of a makes the derivative of (6) 0. Insertion of this value as a function of g_0, say $a(g_0)$, makes (6) the Bayes risk. Then the derivative of the Bayes risk with respect to g_0 is

$$\begin{aligned} &W_0 \Pr\{Z > a(g_0) \mid H_0\} - W_1 \Pr\{Z \leqq a(g_0) \mid H_1\} \\ &+ \left[W_0 g_0 \frac{d}{da} \Pr\{Z > a \mid H_0\} \Big|_{a=a(g_0)} + W_1 g_1 \frac{d}{da} \Pr\{Z \leqq a \mid H_1\} \Big|_{a=a(g_0)} \right] \\ &\cdot \frac{da(g_0)}{dg_0} = W_0 \Pr\{Z < a(g_0) \mid H_0\} - W_1 \Pr\{Z \leqq a(g_0) \mid H_1\}. \end{aligned} \tag{7}$$

The derivative of the difference between (4) and (5) is

$$\begin{aligned} &W_0 \Phi\left(\frac{-\log[W_0 g_0/(W_1 g_1)] - \lambda}{(2\lambda)^{\frac{1}{2}}}\right) - W_1 \Phi\left(\frac{\log[W_0 g_0/(W_1 g_1)] - \lambda}{(2\lambda)^{\frac{1}{2}}}\right) \\ &- W_0 \Phi\left(\frac{-\log[W_0 g_0/(W_1 g_1)] - (\lambda + \nu)}{[2(\lambda + \nu)]^{\frac{1}{2}}}\right) \\ &+ W_1 \Phi\left(\frac{\log[W_0 g_0/(W_1 g_1)] - (\lambda + \nu)}{[2(\lambda + \nu)]^{\frac{1}{2}}}\right). \end{aligned} \tag{8}$$

The derivative of (8) with respect to $w = \log [W_0 g_0/(W_1 g_1)]$ is

$$\begin{aligned} &- \frac{W_0}{(2\lambda)^{\frac{1}{2}}} \phi\left(\frac{-w - \lambda}{(2\lambda)^{\frac{1}{2}}}\right) + \frac{W_0}{[2(\lambda + \nu)]^{\frac{1}{2}}} \phi\left(\frac{-w - (\lambda + \nu)}{[2(\lambda + \nu)]^{\frac{1}{2}}}\right) \\ &- \frac{W_1}{(2\lambda)^{\frac{1}{2}}} \phi\left(\frac{w - \lambda}{(2\lambda)^{\frac{1}{2}}}\right) + \frac{W_1}{[2(\lambda + \nu)]^{\frac{1}{2}}} \phi\left(\frac{w - (\lambda + \nu)}{[2(\lambda + \nu)]^{\frac{1}{2}}}\right) \\ &= (2\pi)^{-\frac{1}{2}}([2(\lambda + \nu)]^{-\frac{1}{2}} \exp\{-[w^2 + (\lambda + \nu)^2]/[4(\lambda + \nu)]\} \\ &- (2\lambda)^{-\frac{1}{2}} \exp\{-[w^2 + \lambda^2]/[4\lambda]\})(W_0 e^{-\frac{1}{2}w} + W_1 e^{\frac{1}{2}w}). \end{aligned} \tag{9}$$

Study of (9) shows that it is 0 for $w = -\infty$, is positive for w in an interval $-\infty < w < -h$, is negative in an interval $-h < w < h$, is positive in an interval $h < w < \infty$, and is 0 for $w = \infty$. Then we see that (8) is 0 for $g_0 = 0 (w = -\infty)$, is positive in an interval $0 < g_0 < k$ (corresponding to a value of w between $-h$ and h), is negative in an interval $k < g_0 < 1$, and is 0 at $1 (w = \infty)$.

Thus the difference (4)–(5) is $-C$ at $g_0 = 0$, increases to a maximum at $g_0 = k$, and decreases to $-C$ at $g_0 = 1$. If the difference is positive at $g_0 = k$ the set of g_0 for which it is positive is an interval; otherwise there is no g_0 at which it is positive.

For example, if $W_0 = W_1$, we can take $W_0 = W_1 = 1$. Then $k = \frac{1}{2}$, and the maximum difference [(4)–(5)] is

$$\Phi\{-[\tfrac{1}{2}\lambda]^{\frac{1}{2}}\} - \Phi\{-[\tfrac{1}{2}(\lambda+\nu)]^{\frac{1}{2}}\} - C. \tag{10}$$

Given values of λ and ν (that is, γ, τ^2, μ and σ^2) and C, (10) can be evaluated to determine whether it is positive. If $W_0 \neq W_1$, the value of g_0 making (8) equal to 0 could be approximated numerically by trial and error; then the difference (4)–(5) could be evaluated at this point.

In general, let the interval for which observation on both X and Y is preferred be (g_0^L, g_0^U). The endpoints can be determined numerically by trying out values in (4) and (5). Since the difference is increasing in the neighborhood of g_0^L and decreasing in the neighborhood of g_0^U and the evaluation of the derivative (8) involves the same function, numerical procedures are relatively easy to carry out.

An interesting case is where the investigator has the choice of observing m independent observations from a normal distribution, say $z_1, \cdots, z_m$, or $m+n$ observations, say $z_1, \cdots, z_{m+n}$, at an additional cost of $C = nc$. Then $X = \sum_1^m z_\alpha$, $Y = \sum_{m+1}^n z_\alpha$, $\lambda = 2m\theta^2$, $\nu = 2n\theta^2$, H_0 is the hypothesis that the mean of the distribution divided by the standard deviation is $-\theta$, and H_1 is the hypothesis that this ratio is θ. Then if the difference (4)–(5) is positive it is worthwhile to observe the additional n observations. The main result of this section shows that the set of g_0 for which the difference is positive is an interval. This is equivalent to the statement that if it is worthwhile to take the second set of observations for two values of g_0, it is also worthwhile to do so for any value of g_0 between those two values.

The main result derived here as well as the method depends on properties of the normal distribution, specifically (9). This will be made evident by the example in Section 3.

A more general problem can be reduced to the above. Suppose that the vector $z^{(1)}$ of m components has the distribution $N(-\mu^{(1)}, \Sigma_{11})$ under H_0 and $N(\mu^{(1)}, \Sigma_{11})$ under H_1 and that z has the distribution $N(-\mu, \Sigma)$ under H_0 and $N(\mu, \Sigma)$ under H_1, where

$$z = \begin{pmatrix} z^{(1)} \\ z^{(2)} \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{11}$$

Let C be the difference in cost between observing z and $z^{(1)}$. The statistic for testing H_0 against H_1 based on $z^{(1)}$ is the logarithm of the likelihood ratio, $u = 2\mu^{(1)\prime}\Sigma_{11}^{-1} z^{(1)}$, which has the distribution $N(\pm\lambda, 2\lambda)$, where $\lambda = 2\mu^{(1)\prime}\Sigma_{11}^{-1}\mu^{(1)}$. Observation of z is equivalent to observation of $z^{(1)}$ and $w = z^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}z^{(1)}$. The two vectors are independent and w has the distribution $N[\pm(\mu^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mu^{(1)}), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})]$. The statistic for testing H_0 against H_1 based on $z^{(1)}$ and w is

the logarithm of the likelihood ratio,

$$(12)\quad u + v = 2\mu^{(1)\prime}\Sigma_{11}^{-1}z^{(1)} + 2(\mu^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mu^{(1)})'(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21})^{-1}w.$$

The linear function of w, namely v, has the distribution $N(\pm\nu, 2\nu)$, where

$$(13)\quad \nu = 2(\mu^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mu^{(1)})'(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21})^{-1}(\mu^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mu^{(1)}).$$

Thus u and v here are equivalent to u and v at the beginning of the section. The statistic (12) is the classification statistic $2\mu'\Sigma^{-1}z$, and $2(\lambda + \nu) = 4\mu'\Sigma^{-1}\mu$ is the distance between the two distributions. (See Section 6.4 of Anderson (1958), for example.)

3. Some nonnormal distributions. We shall give an example of discrete distributions in which the set of g_0 for which observation on both X and Y is preferred is not an interval. The distributions are given in Table 1 below. The probabilities for the likelihood ratio for X alone are given in Table 1 and for X and Y in Table 2. The cumulative distributions of the likelihood ratios are given in Table 3.

Let the cumulative distribution of the likelihood ratio be $P_k(z)$ under H_k when X is observed and $Q_k(z)$ under H_k when X and Y are observed ($k = 0, 1$). A Bayes procedure based on a given likelihood ratio z when $W_0 = W_1 = 1$ is to accept H_0 if $z < g_0/g_1$, accept H_1 if $z > g_0/g_1$, and accept either hypothesis if $z = g_0/g_1$. To make the procedure unique we shall accept H_0 in the last case. Then the expected losses if X is observed and if X and Y are observed are,

TABLE 1

Distributions of two random variables

X				Y			
Value	Prob. under H_0	Prob. under H_1	Likelihood ratio	Value	Prob. under H_0	Prob. under H_1	Likelihood ratio
a	1/3	1/6	3/6	A	1/2	1/3	2/3
b	1/3	2/6	6/6	B	1/2	2/3	4/3
c	1/3	3/6	9/6				

TABLE 2

Probabilities of the likelihood ratio for two random variables

Value	X	Y	Probability under H_0	Probability under H_1
2/6	a	A	1/6	1/18
4/6	a	B	2/6	4/18
	b	A		
6/6	c	A	1/6	3/18
8/6	b	B	1/6	4/18
12/6	c	B	1/6	6/18

TABLE 3
Cumulative distributions of likelihood ratios

Likelihood ratio	X under H_0	X under H_1	X and Y under H_0	X and Y under H_1
0	0	0	0	0
2/6	0	0	1/6	1/18
3/6	1/3	1/6	1/6	1/18
4/6	1/3	1/6	3/6	5/18
6/6	2/3	3/6	4/6	8/18
8/6	2/3	3/6	5/6	12/18
9/6	1	1	5/6	12/18
12/6	1	1	1	1

TABLE 4
Expected loss functions

g_0	$g_0/(1-g_0)$	$\rho(g_0)$	$\sigma(g_0) - C$	$\rho(g_0) - \sigma(g_0) + C$
0	0	0	0	0
1/4	2/6		1/4	0
1/3	3/6	1/3		1/54
2/5	4/6		11/30	0
1/2	6/6	5/12	7/18	1/36
4/7	8/6		8/21	1/42
3/5	9/6	2/5		1/30
2/3	12/6		1/3	0
1	∞	0	0	0

respectively,

$$
\begin{aligned}
\rho(g_0) &= g_0\left[1 - P_0\left(\frac{g_0}{1-g_0}\right)\right] + (1-g_0)P_1\left(\frac{g_0}{1-g_0}\right), \\
\sigma(g_0) &= g_0\left[1 - Q_0\left(\frac{g_0}{1-g_0}\right)\right] + (1-g_0)Q_1\left(\frac{g_0}{1-g_0}\right) + C.
\end{aligned} \tag{14}
$$

Since the cumulative distribution functions are constant on intervals of the likelihood ratio, $\rho(g_0)$ and $\sigma(g_0)$ are linear on intervals of g_0. In Table 4 these functions are specified by giving their values at the end points of the intervals. The difference $\rho(g_0) - \sigma(g_0)$ is, therefore, linear on intervals, and it is also specified in Table 4.

As can be seen from Table 4, if C, the cost of observing Y, is 0, the use of X and Y together has a smaller expected loss than the use of X alone except on the interval $0 < g_0 \leqq 1/4$, at the point $g_0 = 2/5$, and on the interval $2/3 \leqq g_0 \leqq 1$. However, if C has a value satisfying $0 < C < 1/54$, the expected loss using X and Y together is less than the expected loss using X alone for g_0 in an interval

between 1/4 and 2/5 and in an interval between 2/5 and 2/3, and the expected loss using X and Y together is greater in an interval between 0 and 1/3, an interval between 1/3 and 1/2 and an interval between 3/5 and 1.

This example shows that the set of a priori probabilities g_0 for which it is worthwhile to pay for, and use, another observation is not necessarily an interval. Thus, the nature of the solution to the problem of choice of observations depends on the distributions.

This example could be stated in terms of densities by replacing the probability of a, b, or c, by a constant density of X over (0, 1), (1, 2) or (2, 3), respectively, and the probability of A or B by a constant density of Y over (0, 1) or (1, 2), respectively. Moreover, small modifications of the constant densities would yield likelihood ratios with densities. Thus the discreteness of the example is not essential.

REFERENCES

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

ANDERSON, T. W. (1964). Sequential analysis with lagged observations. To be published.