BAYESIAN SEQUENTIAL ANALYSIS IN PSYCHOLOGICAL RESEARCH

UWE MORTENSEN

University of Canterbury

Bayesian sequential analysis provides a means to decide between possible therapeutic treatments or, more generally, between competing hypotheses as economicaly as possibe. A method due to Anscombe is described and its advantages are discussed with respect to conventional procedures.

Very often the clinical psychologist has to decide which one of two or more available treatments is to be preferred, and the problem is to arive at such a decision with minimum costs and maximum safety. The costs are money, time and the moral costs if patients are to be treated with some inferior form of psychotherapy.

A frequently used and very straightforward way of testing which of (say) two treatments is the better is simply to form two groups of patients, have the members of group I treated with therapy A and the members of group II treated with therapy B. If the number of successes in each group happens to be different for the two groups, an appropriate statistical procedure is found to guide the decision on which of the therapies should be preferred. If one of them appears to be 'significantly' superior to the other, the problem has been settled. If the null hypothesis, that both forms of therapy are equally effective, cannot be rejected, the psychologist can proceed in either of two ways: If he doubts the result of his analysis or his experiment, he may set up another experiment, perhaps more refined, and possibly ends up with a significant result, or he can assume that both therapies are indeed equally effective, and that it is therefore a matter of taste which one is applied, given that the costs of applying them are equal.

It has long been felt that this approach is unsatisfactory, mainly because the number of patients treated inadequately is not minimised. The sample size, that is the number of patients in each experimental group, should have been chosen in advance with respect to the chosen Type I and Type II errors, and even if one starts to suspect that one treatment is superior to the other one while the experiment is still under way one has to carry it out completely so that a test of significance can be made. Furthermore, a particular therapy may indeed yield slightly better results than another one, but the costs of applying it may not make it worthwhile to use it. More recently, the whole idea of tests of significance has been severely criticised (Lindley, 1970), so that even the most practically-minded psychologist is forced to look for some more sophisticated tools when making decisions.

Statisticians have offered the method of sequential sampling as a solution to the problem. A pair of patients comparable with respect to their disorder is chosen as a trial and each of them is treated with one of the competing therapies. We may assume that the effect of the therapies can be scored numerically and that a higher score reflects a better effect. If one of the therapies has appeared to be superior to the other one, after a number of such trials, the experiment is stopped and a decision is made; otherwise the experiment is continued. In medical and psychological contexts the maximum number of trials, Nmax, is usually also fixed in advance. If Nmax trials have to be made and neither of the therapies then appears to be superior to the other, the null hypothesis (both therapies are equivalent with respect to their effect) is accepted.

The method of sequential testing as it is known today was developed by Wald and Barnard (see Wetherill, 1968) and has been adapted to clinical trials by Armitage (1960) within the framework of the Neyman-Pearson approach to decision making, where the sample size is determined according to the choice of the probability of a Type I error (i.e. the probability of rejecting H_o though it is correct), and the probability of Type II error (the probability of rejecting H_1 falsely). However, these procedures have also been criticised (for instance, see Anscombe, 1963, and Cornfield, 1966), since the decisions made with these procedures do not take into account the costs of accepting certain alternatives, and even worse, depend on the stopping rule, i.e. on the choice of the Type I and Type II errors and the resulting estimation of the sample size. Versions of sequential testing methods based on the Bayesian approach of decision making are offered and will be outlined later. Firstly, some of the relevant notions and concepts will be introduced.

The Sequential Probability Ratio Test (SPRT)

Let p be the cure rate of a therapy, and let T be the random variable representing the number of patients cured by it. Then T has a probability distribution f(T, p), which will be the binomial distribution if the parameter p is constant and the patients are cured independently.

In the following treatment the numbers in brackets refer to equations given in the appendix.

For brevity we shall write f(t,p) for f(T=t/p). Let $f(t/p_1)$ represent the probability that t patients have been cured under the condition that $p=p_1$, and let $f(t/p_o)$ be the same probability under the condition $p=p_o$. Wald (1947) considered the ratio (1) where n is the number of observations, i.e. the total number of pairs of patients examined so far. It is obvious that the hypothesis that $p=p_1$ becomes unlikely if 1_n becomes small, and likely with large 1_n. In particular, Wald's decision rule is as follows: Continue sampling as long as (2) holds, and as soon

as 1_n exceeds B, stop sampling and decide in favour of P_o. If 1_n exceeds A, stop sampling and decide in favour of p_1.

A and B are called 'rejection boundaries'; they are constants and can be chosen such that they correspond to the error probabilities of a (Type I) and b (Type II) such that the less efficient therapy is chosen only with probability a, or that the competing therapy is chosen falsely only with probability b. Wald has shown that approximately (3) holds.

For a more complete description of the SPRT see, for instance, Wetherill (1968, p. 14). But it should be noted here, as Wald has shown, that for the SPRT the expectation of the sample size takes a minimum as compared to other sequential methods.

A disadvantage of the SPRT is that it implies a great variance of the sample size, n; under certain conditions n can become very large, possibly infinite. This is a severe disadvantage in medical and psychological trials, where one wants to experiment with a possibly inferior therapy as little as posible. Armitage (1960) therefore provided 'restricted' procedures, which have to be considered shortly.

Let us assume that for a certain disorder two forms of therapy are available. We want to know whether they are equally effective or not, for example whether the distribution of differences between the treatments has a mean x of zero or not. This means that we have to carry out a 2-sided test of the two hypotheses H_o and H_1 (4), where the standard deviation s is used as a unit of measurement; m_1 is a critical value for the observed mean.

The larger the sample size, the larger the number of failures of a therapy will be if its cure rate is constant over time. We want to discard a therapy if the number of failures becomes larger than a critical number, but since the failure rate depends on the sample size, the critical number should also depend on the sample size. The sets of critical numbers for sample sizes $n=1, 2, 3, \ldots$ are called rejection boundaries, and Armitage suggested we should use (5), where A stands for 'upper boundary' and B for 'lower boundary', and the coefficients a and b are determined in a way outlined below.

Let d_i be the difference of the effect of the two therapies for the i-th pair of patients, and let y_n be the sum of the first n differences.

A decision for the hypothesis H_1 is made if y_n becomes larger than the corresponding $-z_n$. The hypothesis H_o is accepted, if the maximum sample size Nmax has been reached without y_n leaving the boundaries (see Figure 1.).

In order to proceed in this way, the coefficients c and d in (5) have to be determined. Armitage suggests we consider the likelihood ratio for $x = O$ to $x \neq O$. If $x \neq O$ we can always write $x = Cs$, i.e. express the non-zero value of x as a multiple or fraction C of the standard-deviation s. Armitage suggests that the likelihood ratio for $x \neq O$ to $x = O$ along the upper boundary A should be (6) which leads to (7)

for the coefficients in (5). In order to determine the maximum sample size Nmax one may assume that according to the central limit theorem the distribution of the observed differences is normal. Using the normal distribution one can compute the probabilities P(D/c,d,Nmax), where D is a value for a difference, and in particular one has for the difference D = C the equations (8) which can be used to compute an approximation for the value of Nmax (Wetherill, 1968, p. 77).

Armitage's method, though elegant, has been criticised by statisticians sharing the Bayesian viewpoint (Anscombe, 1963; Cornfield, 1966). The criticisms put forward have been indicated in the introduction of this paper, namely that Armitage's methods depend, like the SPRT, on the stopping rule, which depends on the error probabilities, which are taken unconditionally over the whole sample space, which again implies that decisions are based partially on what was not observed rather than on what was observed, which may lead to absurdities. Bayesian statisticians require that a decision should only be made dependent upon the likelihood of the observed data and that furthermore the costs of the decision should be taken into account. In the following, the Bayesian approach will be outlined.

Bayesian decision making

Here the decision for a hypothesis is based on the posterior probability of that hypothesis. According to Bayes' theorem the posterior probability is proportional to the product of the prior probability of the hypothesis and the likelihood of the data with respect to this hypothesis.

Furthermore, the losses, which are the costs of the decision, are taken into account. (There are, however, Bayesians who get along without doing this).

Let us assume that a certain therapy has a cure rate p. Let $W_0(p)$ be the costs of accepting the therapy and $W_1(p)$ the costs of rejecting it. Let $P_{ac}(p/S)$ be the probability of accepting the therapy under the sampling plan S, and $P_{rej}(p/S)$ be the probability of rejecting it. Of course, $P_{ac}(p/S) + P_{rej}(p/S) = 1$. Let E(n/p,S) be the expected sample size. The loss function is then defined as (9) where the costs W_0 and W_1 are usually considered to be proportional to the number of cases examined. It is clear that one wants to make a decision such that R(p/S) is minimised. Let Pr(p) be the prior distribution of p. The risk of a decision is then defined as (10) i.e. the risk is the expectation of the loss with respect to the prior distribution chosen. A sequential sampling plan is called a Bayes solution if it is the result of minimising the overall risk with respect to some prior distribution. Wetherill (1968) shows (p. 94) that under these conditions the terminal decision only depends on the posterior distribution of p. The question remains of how the prior distribution is chosen and how the loss function has to

be defined. In general, the definition of the prior distribution and of the loss-function will depend upon the particular situation. In the following, an example of a method of sequential testing in the framework of Bayesian decision theory will be given; the example is due to Anscombe (1963).

Anscombe assumes that there are only two treatments under investigation. The responses of patients to the treatments are assumed to be assessable in numerical terms. It is further assumed that the two treatments do not differ in side effects or in costs and that they have the same relative effectiveness for all patients. The observed variable is the difference of scores for a pair of patients, where one patient has been treated by T1 and the other by T2. The differences are assumed to be normally distributed with known variance, (in particular unit variance is assumed) but with unknown mean p. A high response score reflects a high treatment effect; therefore, T1 will be preferred to T2 if x the mean difference is positive, and T2 will be preferred to T1 if the mean difference is negative. Assume that a decision is reached after n observations, and let y be the sum of the n differences observed.

Let X be a random variable and let the notation N(X) indicate that X is normally distributed. In particular, set $X = y/n$. Let M be the 'true' mean of the differences; we can then assume that $X = y/n$ is normally distributed about M with standard deviation the inverse root of n. Then (11) is the normal density for the difference between the true observed means. This density can be taken as the likelihood-function for the true mean M since it allows us to compute the probability of the observed mean under the condition that M is the true mean. We now want to estimate M and therefore we have to find the posterior distribution for M. This requires an assumption about the prior distribution of M, and Anscombe suggests the uniform distribution, mainly to keep the computations simple; a different choice of the prior does not change the basic argument. (It should be noted here that classical statistics also contain the assumption of a uniform prior in an implicit and disguised form.) An alternative to the assumption of a uniform prior would be the normal distribution (Sasaki, 1969, p. 358). Under the assumption of a uniform prior the posterior distribution for M can be shown to be (12). If M is greater than O, T1 will be preferred to T2, otherwise T2 to T1. If the wrong decision was made the loss would depend upon the value of M. Anscombe defines the loss function as follows. If n pairs of patients have been investigated, n patients have been given the inferior treatment. The costs are then $n.g(|M|)$, where g is some function of $|M|$. In particular, $g(|M|) = |M|$ may be chosen. At the end of the experiment, after n pairs have been observed, the expected loss is then given by (13) where f(M) is the distribution of M.

This definition of the costs is not complete. Anscombe argues that the experiments will be published and therefore will have some influence, since patients will be given the therapy which appeared to be

superior in the experiment. If the decision is made after the experiment was wrong this will lead to an increase of the loss or the costs. If k patients will be treated in the future the loss is (14) where 'sgn y' stands for 'signum of y'; see (15), i.e. the loss resulting from M and y having opposite signs is considered. The total loss is found by adding (13) and (-14) and evaluating the expectations; the necessary computations are somewhat lengthy. The resulting expression for the total loss is then (16), where $F(-|y|/n)$ denotes the area under the normal curve up to $X=|y|/n$. This integral does not exist in closed form but its value can be found using numerical methods and it has been tabulated.

The problem now is to find a stopping rule that minimises R(n,y). If the value for k, the number of future patients treated with the preferred therapy is known, then the determination is always possible in principle; it may be relatively difficult, though. But k is usually not known, and Anscombe (1963) discusses the resulting intricacies. He offers approximate solutions for two cases, (i) where k is assumed to be given and (ii) where it is assumed that we are given, both n, the number of patients examined during the experiment, and k the number of future patients treated.

For a given sum y, R(n,y) is a function of n, the sample size. We optimise our decision by finding that value of n for which R(n,y) has a minimum. This value is found by differentiating R(n,y) with respect to n, setting the derivative equal to O and solving for n. It can be shown that R(n,y) has minima where (17) is satisfied, and the loss for a minimising value of n for a given sum y is given by (18).

If after n observations n and y satisfy (17), no more observations should be made, since this would lead to an increase of (18). The decision procedure can be summarised as follows:

1. Examine the n-th pair of patients, $n = 1, 2, 3, \ldots$ find the difference d_n of the treatment effects and add to y_{n-1}:
   $$y_n = y_{n-1} + d_n. \quad \text{For } n = 1, \text{ set } y_{n-1} = O.$$
2. Compute $n/(k + 2n)$ and $-y_n/n = y'_n$
3. Use y'_n to find the value of $y''_n = F(y')$, i.e. the value of the area under the normal curve up to y'_n.
4. Compare y''_n and $n/(k + 2n)$; if they are approximately equal stop sampling and decide for treatment E1 or T2 depending on the value of y being positive or negative. Otherwise go back to step 1.

Equation (17) defines the boundaries for the decision problem. Anscombe (1963) tabulates some values of the boundaries (17), for the case k fixed and for the case $k + 2n$ chosen fixed.

Anscombe's suggestions were given in some detail to give the flavour of the Bayesian approach. Novick and Grizzle (1965) suggested another sequential sampling plan, where categorised data are considered: a treatment is either effective or ineffective, which leads

to the binomial distribution of successes for a given sample size. Novick and Grizzle discuss the effect of choosing different prior distributions and suggest a graphical rather than a numerical decision procedure. However, details of their plan will not be given here and the interested researcher is referred to their original paper.

Summary

It appears to be of great importance to test competing forms of treatment as economically as possible, which means that the sample size has to be kept as small as possible under the restriction that the decision should be made with maximum security. Sequential sampling plans seem to offer a solution and have been adapted to medical and psychological trials by Armitage (1960). His decision procedures, however, are based on the Neyman-Pearson approach of decision making, which implies the disadvantage that the decision depends on the particular stopping rule chosen and might not be optimal with regard to costs. On the other hand, Bayesian decision making does not depend on the stopping rule and takes the costs into account. The Bayesian approach is outlined and its adaptation to sequential testing as suggested by Anscombe is described.

Uwe Mortensen is Postdoctoral Fellow, Department of Psychology, 1972-74.

REFERENCES

Anscombe, F. J. Sequential medical trials. *Journal of the American Statistical Association,* 1963, 58, 365-387.

Armitage, P. *Sequential medical trials.* Oxford: Blackwell, 1960.

Cornfield, J. A Bayesian Test of some classical hypotheses—with application to sequential medical trials. *Journal of the American Statistical Association,* 1965, 61, 577-594.

Lindley, D. V. Bayesian approach to statistics. In *Formulation and assessment of statistical models for experimental psychology.* Proceedings of the NUFFIC international Summer Session in Science, August, 1970.

Novick, M. R. and Grizzle, J. E. A Bayesian approach to the analysis of data from clinical trials. *Journal of the American Statistical Association,* 1965, 60, 81-96.

Sasaki, K. *Statistics for modern business decision making.* Belmont: Wadsworth, 1969.

Wetherill, G. B. *Sequential methods in statistics.* London: Methuen, 1968.

Wald, A. *Sequential analysis.* New York: John Wiley, 1947.

APPENDIX

Equations used in the text

(1) $1_n = f(t|p_1)/f(t|p_o)$

(2) $B < 1_n < A$

(3) $A = (1\text{-}b)/a$, $B = b/(1\text{-}a)$

(4) H_o: $x = O$, H_1: $|x/s| > m_1$

(5) A: $z_n = c + d_n$, B: $-z_n = -c - d_n$

(6) $p(\text{data}|H_1)/p(\text{data}|H_o) = (1\text{-}b)/a$

(7) $c = \log((1\text{-}b)/a)C$, $d = C.s/2$

(8) $P(C|c,d,N) = 1\text{-}b$, $P(O|c,d,N) = a$

(9) $R(p,s) = E(n|p,S) + P_{ac}(p|S)W_o(p) + P_{rej}(p|S)W_1(p)$

(10) $D_i = \int Pr(p)W_i(p)dp$

(11) $N((y - nM)/n)$

(12) $Pr(M|y) = n^{\frac{1}{2}}N(Mn^{\frac{1}{2}} - y/n^{\frac{1}{2}})dM$

(13) $nE(|M|) = n\int f(M)|M|dM$

(14) $W = kE(\max(O, -\text{sgn } y))$

(15) $\text{sgn } y = +1$, if $y > O$ -1, if $y < O$

(16) $R(n,y) = |y| + ((k+2)/n)(N(y/n^{\frac{1}{2}}) - (y/n^{\frac{1}{2}})F(-|y|/n))$

(17) $F(-|y|/n) = n/(k + 2n)$

(18) $R(n,y) = ((k+2n)/n)\Phi(y/n)$

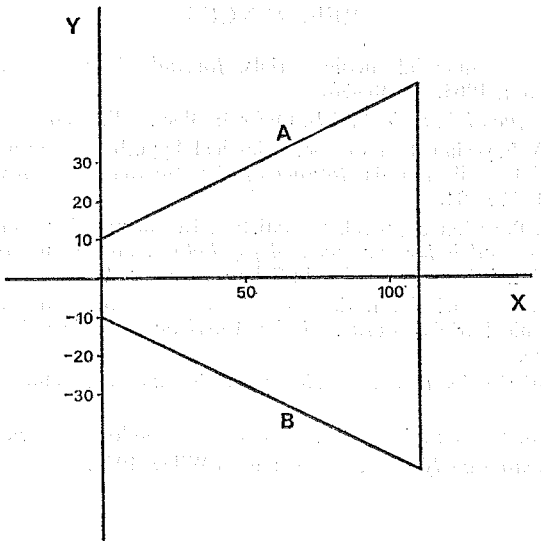

Figure 1. Some possible rejection boundaries according to Armitage. The ordinate is the cumulative sum of differences in response to the drug. The X-axis is the sample size as pairs of patients. A and B are boundary values as in formula (5).