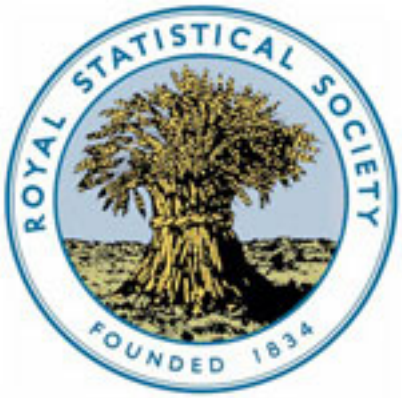

Repeated Significance Tests on Accumulating Data

By P. ARMITAGE, C. K. MCPHERSON and B. C. ROWE

Department of Medical Statistics and Epidemiology, London School of Hygiene and Tropical Medicine

SUMMARY

If significance tests at a fixed level are repeated at stages during the accumulation of data the probability of obtaining a significant result when the null hypothesis is true rises above the nominal significance level. Numerical results are presented for repeated tests on cumulative series of binomial, normal and exponential observations.

1. INTRODUCTION

THE general effect of performing repeated significance tests at different stages during the accumulation of a body of data is well known. If the null hypothesis is true and if each significance test is performed at the same nominal level, the probability that at some stage or another the test criterion is significant may be substantially greater than the nominal value. Feller (1940) discussed the possibility that some of the more significant results in card-guessing experiments in extra-sensory perception might be attributed to "optional stopping" at particularly favourable stages during an investigation. The law of the iterated logarithm shows that a test criterion which takes the form of a standardized cumulative sum of deviations from expectation divided by its standard error will, with probability one, eventually reach any pre-assigned value. Thus, in many common situations a result as highly significant as one desires can be obtained by sufficiently extensive sampling. Robbins (1952) and Anscombe (1954) provide further discussion of this point. The desire to control the error of the first kind, as well as the power of a test procedure, was of course one of the motivations of sequential analysis (Wald, 1947).

More recently the practical relevance of this phenomenon has been called into question. Anscombe (1954) had pointed out that inferences based on likelihoods or, through likelihoods, on posterior probabilities were unaffected by stopping rules. The contrast between this property and the extreme sensitivity of frequency-type inferences to the stopping rule explains why sequential analysis is a topic of such contention between adherents of different viewpoints (Birnbaum, 1964; Cornfield, 1966; Armitage, 1967). The exchanges of opinion on these matters have been remarkable for the lack of quantitative information about the optional stopping effect. It has not, for example, been possible to answer questions such as the following.

(a) What is the probability of obtaining a result "significant" at a certain nominal level, within the first 50 tests?

(b) Does the enhancement of the probability of obtaining a significant result reach a noticeably high level only after a very large number of tests?

(c) What is the effect of repeated tests when the null hypothesis is not true?

The purpose of the present paper is to repair some of these gaps in our knowledge without indulging in further discussion of inferential problems. We consider sequential observations of three different distributional forms: binomial, normal and exponential. In the binomial case exact results are obtained by direct calculation of

probabilities. In the normal case approximate results are obtained by two methods—numerical integration and simulation; and in the exponential case by numerical integration. The present results all concern repeated tests of a null hypothesis which is true; the emphasis is therefore on questions like (a) and (b) above. Later publications will report results for the non-null situation, as in question (c). We hope to use all these results to formulate proposals for sequential sampling plans which can be interpreted either from the frequency point of view, with specified probabilities of errors, or as repeated significance tests at a specified level, or perhaps as having a stopping rule defined in terms of likelihood or posterior probability (Armitage, 1967).

2. THE BINOMIAL CASE

Some numerical results published by Armitage (1960, Table 1.2) are incorrect. Corrected values were published by Armitage (1967) and are extended below.

TABLE 1

The probability of being absorbed at or before the nth observation in binomial sampling with repeated tests at a nominal two-sided significance level 2α. *The number of opportunities of hitting the boundary is shown in parentheses*

n \ 2α	0·01	0·02	0·03	0·04	0·05
10	0·00781 (1)	0·01563 (1)	0·02930 (2)	0·05469 (2)	0·05469 (2)
15	0·01538 (3)	0·02844 (3)	0·03955 (3)	0·08191 (4)	0·08191 (4)
20	0·01840 (4)	0·04033 (5)	0·05248 (5)	0·08970 (5)	0·10662 (6)
25	0·02266 (6)	0·04811 (7)	0·06538 (7)	0·10326 (7)	0·12140 (8)
30	0·02746 (8)	0·05586 (9)	0·07631 (9)	0·11488 (9)	0·13355 (10)
35	0·03087 (10)	0·06083 (11)	0·08527 (11)	0·12438 (11)	0·14328 (12)
40	0·03406 (12)	0·06531 (13)	0·09255 (13)	0·13731 (14)	0·15351 (14)
45	0·03687 (14)	0·07034 (15)	0·09848 (15)	0·14416 (16)	0·16399 (16)
50	0·03931 (16)	0·07432 (17)	0·10508 (17)	0·15125 (18)	0·17117 (18)
60	0·04319 (20)	0·08205 (21)	0·11616 (22)	0·16227 (22)	0·18583 (22)
70	0·04679 (24)	0·08757 (25)	0·12436 (26)	0·17289 (26)	0·19973 (27)
80	0·05090 (29)	0·09402 (30)	0·13083 (30)	0·18298 (31)	0·20889 (31)
90	0·05345 (33)	0·09900 (34)	0·13867 (35)	0·19062 (35)	0·21941 (35)
100	0·05586 (37)	0·10320 (38)	0·14436 (39)	0·19881 (40)	0·22731 (40)
120	0·06085 (46)	0·11125 (47)	0·15515 (48)	0·21280 (50)	0·24187 (49)
140	0·06462 (55)	0·11790 (56)	0·16388 (57)	0·22211 (58)	0·25503 (58)
150	0·06619 (59)	0·12133 (61)	0·16833 (62)	0·22670 (62)	0·26108 (63)

An experiment consists of a series of independent binomial trials. On the null hypothesis the probability of success in each trial is $\frac{1}{2}$. After each trial the experimenter tests the significance of S_n, the total number of successes, using the test which would have been appropriate if the number of trials had been fixed in advance. He stops as soon as he gets a significant result. In other words, he stops the first time the inequality

$$a_n < S_n < b_n$$

fails to be satisfied, where a_n and b_n are the lower and upper two-sided 2α points of the binomial distribution $\mathrm{Bi}(n, \frac{1}{2})$. That is, b_n is the lowest integer for which in unrestricted sampling

$$P(S_n \geqslant b_n) \leqslant \alpha,$$

and $a_n = n - b_n$. The value of n at which the experiment stops is denoted by m.

We required the probability distribution of m for various values of α. A general program for the evaluation of the probabilities for arbitrary $\{a_n, b_n\}$, and arbitrary probability of success was written by C. K. M., and a separate sub-routine used for the calculation of the particular $\{a_n, b_n\}$ needed here. The main program proceeded along familiar lines, the number of admissible paths for each set of values $\{S_n\}$ being determined recursively from those for the $\{S_{n-1}\}$.

Results for $2\alpha = 0{\cdot}01\,(0{\cdot}01)\,0{\cdot}05$ and various values of $n \leqslant 150$ are given in Table 1. Because of the discrete nature of the variable, one can stop at only a restricted number of values of m. As we shall see, this seems to be relevant in comparing the binomial results with those for normal sampling described below. The number of possible stopping points for each value of n is shown in brackets in Table 1.

3. The Normal Case

An experiment consists of a series of observations $X_1, X_2, \ldots, X_n$, on random variables which, on the null hypothesis, are independently and normally distributed with zero mean and unit variance. After each observation the experimenter uses the cumulative sum

$$S_n = \sum_{i=1}^{n} X_i \tag{1}$$

to test the null hypothesis. He stops sampling when, for the first time,

$$|S_n| \geqslant y_n, \tag{2}$$

where $y_n = k\sqrt{n}$ for some constant k. The value of n at which the experiment stops is denoted by m.

In this investigation k has been chosen to correspond to a non-sequential test for S_n at a selected significance level. For two-sided level 2α, with the usual notation for the normal integral,

$$\Phi(k) = 1 - \alpha.$$

3.1. *Solution by Quadrature*

The problem is to find the distribution of m. Denote by $f_n(s_n)$ the probability density function of S_n in the sequential procedure. Then

$$f_n(s_n) = \begin{cases} \displaystyle\int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\,\frac{1}{\sqrt{(2\pi)}}\exp\{-\tfrac{1}{2}(s_n-u)^2\}\,du, & -y_n \leqslant s_n \leqslant y_n \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The density function f_n can thus be defined recursively by (3), f_1 being the standardized normal density function. The probability of absorption at or before n observations (that is, the probability that $m \leqslant n$) is then

$$P_n = 1 - \int_{-y_n}^{y_n} f_n(u)\,du. \tag{4}$$

Alternative formulae are

$$P_n - P_{n-1} = 2\int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\,\{1-\Phi(y_n-u)\}\,du \tag{5}$$

and

$$P_n - P_{n-1} = 2\int_{y_n}^{\infty}\int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\frac{1}{\sqrt{(2\pi)}}\exp\{-\tfrac{1}{2}(v-u)^2\}\,dv\,du.$$

The basic method was to evaluate the right-hand side of (3) at points on a grid of width h, that is, for $s_n = h, 2h, \ldots, \lambda_n h$, where $\lambda_n h \leqslant y_n < (\lambda_n+1)h$, and also for $s_n = \frac{1}{2}(\lambda_n h + y_n)$ and $s_n = y_n$. This was done by application of a Newton–Cotes formula of the second order over the range

$$-\lambda_{n-1}h \leqslant u \leqslant \lambda_{n-1}h$$

and of Simpson's three-point formula over the range

$$\lambda_{n-1}h \leqslant |u| \leqslant y_{n-1}.$$

P_n was evaluated by (4), using the same method of quadrature.

Results for selected values of $n \leqslant 200$, for $2\alpha = 0{\cdot}10, 0{\cdot}05, 0{\cdot}02$ and $0{\cdot}01$, based on a grid width $h = 0{\cdot}1$, are shown in Table 2 in the columns headed Q. Calculations for $2\alpha = 0{\cdot}05$, $n \leqslant 100$ and $h = 0{\cdot}2$ and $0{\cdot}05$ gave values of P_n differing at the most by 1 unit in the fifth decimal place from those for $h = 0{\cdot}1$. Similarly, for $100 \leqslant n \leqslant 200$, changing h from $0{\cdot}1$ to $0{\cdot}2$ affects P_n by, at the most, 2 units in the fifth place. Results for P_n based on (4) and (5) with $h = 0{\cdot}2$ agree to within 1 unit in the fourth place; (5) is likely to be the less accurate formula. The results in Table 2 for $n \leqslant 200$ are therefore likely to be almost fully accurate to five places.

The lower part of Table 2 (for $n > 200$) shows some further results for which the calculations were made using $h = 0{\cdot}5$. These figures are less accurate than those for $n \leqslant 200$ but at $n = 200$ they differ by at the most three digits in the fourth decimal place. Using such a large value for h, although placing some doubt on the accuracy of the computation, enables large values of n to be considered without excessive use of computer time; the calculations for $n \leqslant 1000$ and for four boundaries used a total of 26 min. on an IBM 360/65.

An alternative approach is to work with the quantities

$$p_{K,n} = \int_{(K-1)h}^{Kh} f_n(u)\,du,$$

using the recurrence relation

$$p_{K,n} = \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\,[\Phi(Kh-u) - \Phi\{(K-1)h-u\}]\,du,$$

where $f_{n-1}(u)$ in the integrand is approximated by $p_{(u/h)+\frac{1}{2},n-1}/h$ and the integral is evaluated numerically using a standard method of approximation to the normal

integral. This method has no advantage over that described earlier, and calculations with varying h showed the first method to be both more accurate and less time-consuming.

TABLE 2

The probability of being absorbed at or before the nth observation in sampling from a normal distribution with known variance, with repeated tests at a nominal two-sided significance level 2α *(i.e. standardized normal deviate* k*)*†

	2α	0·10		0·05		0·02		0·01	
	k	1·645		1·960		2·326		2·576	
n		Q	S	Q	S	Q	S	Q	S
1		0·10000	0·0970	0·05000	0·0545	0·02000	0·0230	0·01000	0·0135
2		0·16015	0·1650	0·08312	0·0885	0·03453	0·0385	0·01766	0·0235
3		0·20207	0·1980	0·10726	0·1115	0·04561	0·0510	0·02366	0·0275
4		0·23399	0·2295	0·12617	0·1260	0·05454	0·0610	0·02858	0·0345
5		0·25963	0·2590	0·14169	0·1420	0·06201	0·0675	0·03274	0·0390
10		0·34169	0·3425	0·19336	0·1925	0·08776	0·0905	0·04738	0·0525
15		0·38973	0·3870	0·22509	0·2195	0·10419	0·1070	0·05692	0·0625
20		0·42319	0·4300	0·24791	0·2455	0·11628	0·1200	0·06403	0·0695
25		0·44861	0·4550	0·26567	0·2665	0·12586	0·1295	0·06971	0·0775
30		0·46896	0·4765	0·28016	0·2855	0·13379	0·1360	0·07444	0·0815
35		0·48584	0·4925	0·29238	0·3015	0·14054	0·1430	0·07850	0·0865
40		0·50020	0·5045	0·30293	0·3150	0·14643	0·1550	0·08205	0·0905
45		0·51266	0·5170	0·31220	0·3205	0·15165	0·1630	0·08520	0·0960
50		0·52364	0·5260	0·32045	0·3295	0·15633	0·1700	0·08805	0·1000
60		0·54223	0·5460	0·33464	0·3435	0·16446	0·1795	0·09300	0·1070
70		0·55754	0·5595	0·34653	0·3560	0·17113	0·1850	0·09722	0·1105
80		0·57051	0·5735	0·35674	0·3655	0·17733	0·1920	0·10090	0·1130
90		0·58170	0·5855	0·36568	0·3730	0·18260	0·1975	0·10416	0·1170
100		0·59152	0·5975	0·37362	0·3830	0·18732	0·2015	0·10708	0·1195
120		0·60659		0·38722		0·19549		0·11216	
140		0·62292		0·39857		0·20238		0·11647	
160		0·63315		0·40829		0·20834		0·12022	
180		0·64301		0·41677		0·21359		0·12353	
200		0·65165		0·42429		0·21828		0·12649	
250		0·670		0·440		0·228		0·133	
500		0·720		0·487		0·259		0·152	
750		0·746		0·513		0·276		0·164	
1,000		0·763		0·530		0·288		0·172	

† Columns headed Q calculated by quadrature, those headed S by simulation. Grid-width $h = 0{\cdot}1$ for $n \leqslant 200$, $h = 0{\cdot}5$ for $n > 200$.

3.2. *Simulation Results*

Before the results discussed in Section 3.1 were available an alternative approach based on simulation was attempted using programs written by B. C. R. Recent work

has made the generation of pseudo-random numbers on a digital computer a straightforward process (see MacLaren and Marsaglia, 1965, or Hull and Dobell, 1962). The following expression was found to give satisfactory results for the generation of a long sequence of rectangular deviates:

$$X_{i+1} = 5^5 X_i \quad (\text{modulo } 2^{26}).$$

These rectangular deviates were then used to generate normal deviates using an algorithm due to Schafer (1962).

Two main runs were carried out with these sequences. In each case a tabulation was done of the results from 1,000 experiments. Each experiment involved the generation of a random deviate with zero mean and unit variance which was added to a running sum. The value of this sum was then compared with sets of values that formed boundaries representing the nominal levels of significance. The test was made both for an upper crossing ($S_n \geqslant y_n$) and a lower crossing ($S_n \leqslant y_n$). One hundred deviates were generated for each experiment, thus providing results comparable with those obtained by quadrature for $n \leqslant 100$.

These two runs differed in the start number used for the series of rectangular deviates. Various statistics including the mean, variance and auto-correlation of lag one were calculated for the two sets of 100,000 normal deviates used in the two runs, and for the sets of rectangular deviates used in generating the normal deviates, and proved satisfactory.

The results from these two runs have been combined in Table 2. In addition, the results for the two sides of the test have been combined. These results are thus directly comparable with those obtained by quadrature. Virtually all the figures agree at the first significant digit and many agree to two or even three significant digits. The standard error of most of the entries is less than 0·01.

The simulation results showed no advantage over those obtained by quadrature in the amount of machine time required. The quadrature results summarized in Table 2, with output for all values of $n \leqslant 200$, required the following times on an IBM 360/65: 7 min. 40 s.; 9 min. 5 s.; 10 min. 20 s.; and 11 min. 30 s. The two runs of 1,000 simulation experiments with $n \leqslant 100$, summarized in Table 2, required about 14 min. each on Atlas. The simulation results could, however, be made more efficient by the use of a better system and by increasing the efficiency of sampling by the use of correlated variables.

4. THE EXPONENTIAL CASE

An experiment consists of a series of observations $X_1, X_2, \ldots, X_n$ on random variables which, on the null hypothesis, are independently distributed according to the exponential distribution $e^{-y}\,dy$, and after each observation the cumulative sum

$$S_n = \sum_{i=1}^{n} X_i$$

is used to test the null hypothesis. In general $2S_n$ follows the χ^2 distribution on $2n$ degrees of freedom and therefore the lower and upper $100\,\alpha$ per cent probability points for S_n are

$$\tfrac{1}{2}\chi^2_{2n,1-\alpha} \quad \text{and} \quad \tfrac{1}{2}\chi^2_{2n,\alpha},$$

where $\chi^2_{\nu,p}$ is the value of χ^2 on ν degrees of freedom for which there is probability p of a more extreme value. Thus the experiment continues so long as

$$Y_{1,n} < S_n < Y_{2,n},$$

where $Y_{1,n} = \frac{1}{2}\chi^2_{2n,1-\alpha}$ and $Y_{2,n} = \frac{1}{2}\chi^2_{2n,\alpha}$, for a fixed α. The value of n at which the experiment stops is denoted by m.

4.1. *Solution by Numerical Integration*

As the probability density function is greater than zero only on the positive real line it is not satisfactory to use numerical quadrature as in Section 3.1. At each point on a grid there is not necessarily an odd number of positive ordinates of the conditional probability density for the previous n, and quadrature formulae involving an even number of function values are much less accurate than those involving an odd number. Consequently the distribution of m was calculated using approximate integration of small grids of depth h.

Thus if

$$p_{k,n} = \int_{(k-1)h}^{kh} f_n(u)\,du,$$

where $f_n(S_n)$ is the probability density function of S_n, $p_{k,n}$ can be calculated using the recurrence relation

$$p_{k,n} = \int_{Y_{1,n-1}}^{kh} f_{n-1}(u)\{e^{u-(k-1)h} - e^{u-kh}\}\,du,$$

where $f_{n-1}(u)$ is evaluated at the mid-points of the grids using the approximation $f_{n-1}(u) \approx p_{(u/h)+\frac{1}{2},n-1}/h$ for integral values of $u/h+\frac{1}{2}$. If $kh > Y_{2,n-1}$ then the upper limit of integration is $Y_{2,n-1}$. Special allowance has to be made for the integrand near to the limits of integration, where incomplete grid widths are used.

Table 3 shows the results for four values of 2α and $n \leqslant 100$, calculated using $h = 0{\cdot}1$.

These results are very similar to those of the normal distribution and, in general, the first two significant digits are the same. One possible cause of the irregularity of the discrepancies between Tables 2 and 3 is the rounding errors in the boundary values of $\frac{1}{2}\chi^2$ which were copied from tables with three decimal places. (This explains the entry 0·051 instead of 0·050 for $n = 1$.) The cumulative effect of these errors on the distribution of m could be enough to affect even the second significant digit for large m. It is hoped in future work to use an algorithm for the value of χ^2 which will reduce the effect of these errors.

5. DISCUSSION

Tables 1, 2 and 3 show that the probabilities of exceeding critical levels in repeated significance tests on accumulating data can be substantially above the nominal significance levels even for only moderate amounts of data. It would not be surprising to find an investigator examining his data on, say, five occasions which divided the accumulating data into approximately equal parts. The probabilities given in Table 2 for $n = 5$ exceed the nominal levels by a multiple of between $2\frac{1}{2}$ and 4. The probability of a Type I error could be controlled at a specified level by paying attention only to results significant at a somewhat higher level. For a Type I error probability of 0·05, for instance, the individual test should be carried out at a level between 0·01 and 0·02.

Table 4 shows this value as 0·015. The entries in Table 4 were obtained by inverse interpolation from calculations at appropriately chosen values of k.

In sampling from $N(\mu, 1)$, an investigation with $n \leqslant 200$, in which significance tests are applied at $n = 50$, 100, 150 and 200, would be equivalent to one with $n \leqslant 4$ and

TABLE 3

The probability of being absorbed at or before the nth observation in sampling from an exponential distribution with repeated tests at a nominal two-sided level 2α

n	2α			
	0·10	0·05	0·02	0·01
1	0·100	0·051	0·020	0·010
2	0·161	0·084	0·034	0·018
3	0·203	0·108	0·045	0·024
4	0·236	0·128	0·054	0·029
5	0·261	0·143	0·062	0·033
10	0·343	0·195	0·088	0·048
15	0·390	0·227	0·104	0·057
20	0·423	0·249	0·116	0·065
25	0·449	0·267	0·126	0·070
30	0·469	0·281	0·134	0·075
35	0·486	0·293	0·139	0·079
40	0·500	0·304	0·147	0·083
45	0·512	0·313	0·152	0·086
50	0·523	0·321	0·157	0·089
60	0·542	0·335	0·165	0·094
70	0·557	0·347	0·172	0·098
80	0·570	0·358	0·178	0·102
90	0·581	0·366	0·183	0·105
100	0·591	0·374	0·188	0·108

TABLE 4

Cumulative sampling of normal observations with known variance. Values of k *and* 2α *giving a probability of absorption at or before the nth observation equal to* 0·05

n	k	2α
1	1·96	0·050
5	2·42	0·015
10	2·56	0·010
15	2·64	0·008
20	2·68	0·007
50	2·80	0·005
100	2·88	0·004
200	2·96	0·003

tests at $n = 1, 2, 3$ and 4. In some circumstances, even though an investigator would usually do only a small number of intermediate tests, he might be prepared to test at more frequent intervals if the most recent tests or an informal analysis of current data suggested that a significance criterion was likely to be reached. This would be very early equivalent to a procedure in which a test is carried out after each observation, and the entries in the lower parts of Tables 2 and 4 become relevant. To preserve an overall Type I error rate of, say, 0·05 the individual tests have to be well beyond the 0·01 level.

It is also interesting to note that even for large values of the nominal significance level the asymptotic cumulative probability of absorption (equal to unity) is approached only after a very large number of repeated investigations. For instance, experiments with repeated significance tests at the 5 per cent level will have a median number of tests before absorption of 613 and a high probability that they will continue beyond the 1,000th test when the null hypothesis is true. Alternatively, if an investigator stops at the 1,000th test whether or not a significant boundary is reached, the average number of tests in such an investigation under the null hypothesis will be 537.

The results shown in Tables 2 and 4 relate to tests on cumulative sums of observations from $N(\mu, 1)$ and are clearly appropriate for any known value of the variance of the normal distribution. The results for exponential sampling are very close to those for normal sampling. The analogous results for various non-normal distributions and for t-tests on normal observations with unknown variance would be of interest and we hope to study some of these situations in future work. If the results for binomial series in Table 1 are compared with those for normal observations in Table 2 the tabulated values for a given n are seen to be considerably lower in Table 1. This might be expected partly because the binomial tests are conservative, having size not greater than 2α at each value of n, and also because boundaries can be crossed at only a restricted number of values of n. The results in Table 1, in fact, correspond reasonably well to those in Table 2 if the number of crossing points shown in parentheses in Table 1 is regarded as equivalent to the n of Table 2. The following extracts from Tables 1 and 2 illustrate this point.

	2α		
	0·05	0·02	0·01
Normal ($n = 30$)	0·2802	0·1338	0·0744
Binomial ($n = 80$)	0·2089 (31)	0·0940 (30)	0·0509 (29)

The excess of the entries in the first line over those in the second is of much the same magnitude as the mean difference between nominal significance level and the size of a binomial test.

ACKNOWLEDGEMENT

We are grateful to Miss Judith Bailey for help in programming the simulation experiments and in checking the validity of the random numbers used. We thank the Director of the Institute of Computer Science, London, for the use of the Atlas computer and the Director of the Computer Centre, University College London, for the use of the IBM 360/65.

REFERENCES

ANSCOMBE, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, **10**, 89–100.

ARMITAGE, P. (1960). *Sequential Medical Trials.* Oxford: Blackwell.

—— (1967). Some developments in the theory and practice of sequential medical trials. In *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, **4**, 791–804.

BIRNBAUM, A. (1964). The anomalous concept of statistical evidence: axioms, interpretations, and elementary exposition. *Technical Report IMM*, Courant Institute of Mathematical Science, New York University.

CORNFIELD, J. (1966). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *J. Amer. Statist. Ass.*, **61**, 577–594.

FELLER, W. (1940). Statistical aspects of ESP. *J. Parapsychol.*, **4**, 271–298.

HULL, T. E. and DOBELL, A. R. (1962). Random number generators. *SIAM Rev.*, **4**, 229–254.

MACLAREN, M. D. and MARSAGLIA, G. (1965). Uniform random number generators. *J. Ass. Comp. Mach.* **12**, 83–89.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, **58**, 527–535.

SCHAFER, D. (1962). Norm Dev (Algorithm 121). *Commun. Ass. Comp. Mach.*, **5**, 482.

WALD, A. (1947). *Sequential Analysis.* New York: Wiley.