# EMPIRICAL SEQUENTIAL TESTS

Lester S. Adelman, Martin Miller and Peter Nemenyi
State Univ. of N. Y., Downstate Medical College, Brooklyn

## Summary

Sequential likelihood ratio tests [1] require a knowledge of the power function of a test statistic. It often happens, especially in nonparametric procedures, that the power function of a statistic has not been tabulated.

A multi-stage procedure is proposed which can be used with any statistic whose Null-distribution is tabulated. The method is not optimal but offers appreciable average savings in sample size, at least in certain Normal situations.

In a modification, the tail probabilities $P_1$ at each step are converted to normal deviates and fed into a Wald sequential scheme.

## 1. General Description of the Proposed Method

Consider any statistic U designed to test a null hypothesis $H_o$ against an alternative $H_1$. Assume that U tends to be big when $H_1$ is true and that the full distribution of U under $H_o$ is known, so that any percentage point can be found.

The usual fixed sample size procedure is to calculate U from a sample of size n, reject $H_o$ at the α level if $U \geq U_\alpha$ and "accept $H_o$" or "reserve judgment" if $U < U_\alpha$. Here $U_\alpha$ is defined by $\Pr\{U \geq U_\alpha | H_o\} = \alpha$ if U is a continuous statistic (e.g. a sample mean) or $\leq \alpha$ if it is discrete.

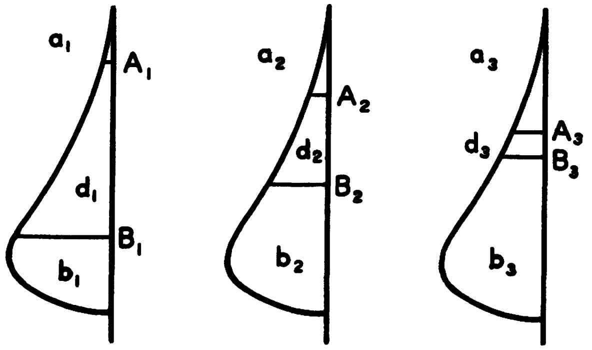
If $H_1$ is simple or is in the form of a prior distribution, and if the distribution of U under (each) $H_1$ is also known, then the power of the test, Pr $\{\text{rejecting } H_o | H_1\}$ can also be calculated.

The fixed sample size procedure can always be modified as follows to yield sequential or multi-stage tests:

Take a sample of size $n_1$ and calculate the value of the statistic, which will be called $U_1$. If $U_1 \geq A_1$, reject $H_o$; if $U_1 < B_1$, accept $H_o$; if $B_1 \leq U_1 < A_1$, take a small (independent) sample of size $n_2$ and proceed as before with subscripts 2 in place of 1. Continue in this way until the first time that $U_i \geq A_i$ or $< B_i$. At each step the probabilities $a_i = \Pr\{U_i \geq A_i | H_o\}$ and $b_i = \Pr\{U_i < B_i | H_o\}$ are known, $A_i$ and $B_i$ having been chosen in advance. The probability that sampling will be continued after a given step is $1-a_i-b_i$, which will be called $d_i$ for short.

The level of the test will be $\leq \alpha$ if the constants $n_i$ and $A_i \geq B_i$ are chosen in any manner such that

$$\alpha \geq a_1 + d_1 a_2 + d_1 d_2 a_3 + \dots \qquad (1).$$



The power is

$$1-\beta = a_1' + d_1' a_2' + d_1' d_2' a_3' + \dots \qquad (2).$$

The expected sample size is

$$\text{ASN} \doteq \begin{cases} n_1 + d_1 n_2 + d_1 d_2 n_3 + \dots & \text{under } H_o \\ n_1 + d_1' n_2 + d_1' d_2' n_3 + \dots & \text{under } H_1 \\ n_1 + d_1'' n_2 + d_1'' d_2'' n_3 + \dots & \text{under } H \end{cases}$$

where primed probabilities are calculated under the assumption that $H_1$ is true, and double prime refers to probabilities based on any other assumption H. The summations are carried to i = M, the point of truncation where $A_i$ is set = $B_i$ and sampling terminates automatically ($d_i = 0$).

β is known if the distribution of U under $H_1$ is known for fixed sample sizes; (the null distribution of U for fixed n is assumed to be known). The ASN can also be calculated under any assumption for which fixed sample size distributions of U are known.

It is possible to devise non-truncated strategies, for example by setting $a_i = a =$ constant, and $b_i = b =$ constant $(i = 1, 2, \ldots, \infty)$, $d = 1\text{-}a\text{-}b$, $\alpha = a/(1\text{-}d)$. This particular infinite strategy is poor, and others have not been explored. Perhaps the method proposed in this paper should therefore be called "multi-stage" rather than "sequential". In particular, we have given most attention to three-stage strategies, finding that these can do a good deal better than two-stage strategies but that we were unable to obtain substantial further gains by employing four stages with this method.

## 2. Comparison with Fixed Sample Size and S.L.R. Tests (General)

When $\alpha$, $\beta$ and the number of steps M have been specified, the problem arises how to choose the constants $n_i$, $a_i$ and $b_i$ $(i = 1, 2, \ldots, M)$ from all possible sets satisfying (1) and (2) in such a manner as to minimize the ASN under $H_o$ or $H_1$ (or some intermediate H). Alternatively it might be desired to specify $\alpha$, M (or $\sum_1^M n_i$) and the ASN under some H, and minimize $\beta$.

This kind of problem cannot be solved in general terms, as the solution depends on the sampling distributions of the statistic used. Even in the simpler special cases an optimum may be very difficult to find. But it can be shown that:

(a) the proposed M-step method can be designed so as to yield an ASN appreciably smaller than the fixed sample size necessary for the same level and power, at least in certain Normal situations, and

(b) no choice of constants can make the ASN as small as a sequential likelihood ratio test [1]. This is clear because the proposed method treats all outcomes in the zone of indifference $B_i \leq U_i < A_i$ alike while the Wald test takes full account of the actual value of $U_i$ within this interval. (The multi-stage tests frequently used in sampling inspection [2] are like Wald sequential tests in this respect).

Thus the proposed method is not optimal but can be recommended as a worthwhile improvement over fixed sample size methods when Wald-like sequential methods are not available, as for example in many nonparametric problems. The point of view is that if The Best is not available, more modest improvements should not be shunned.

The M-stage method is rather flexible. The sample sizes used in any one step, or from step to step, need not be equal. Also it does not matter if external conditions (e.g. variances) change from step to step; and if conditions call for it, different test statistics for the same hypothesis may be used at different stages so long as the decision to change is not made after inspection of the data with an eye on the outcome. And only the null distribution of test statistics needs to be known to make the procedure possible.

In making comparisons between the proposed multi-stage tests and either fixed sample size or s.l.r. strategies, it is desirable to compare their performance not only at two points specified by $H_o$ and a simple $H_1$ but over a whole range of parameter values, since various values may occur in practice. Thus we have set up comparisons in such a way as to juxtapose the ASN of a given multi-stage test at each parameter point with that fixed sample size that would yield the same power at this point; alternatively we compare the power of a given strategy with the power of a test using a fixed sample of size equal to the ASN calculated at that point. This type of comparison of two curves seems to be more realistic than the more usual comparison of a curve with the horizontal straight line based on a simple alternative.

In most problems, the values of $\beta$ and ASN are not readily available for s.l.r. strategies at parameter values other than $H_o$ and $H_1$. We therefore chose to make the comparison with s.l.r. tests in a simple binomial problem where we were able to get information on the s.l.r. test for at least one intermediate parameter value.

## 3. Example: Binomial p

Multi-stage - $H_o$: $p = .1$, $H_1$; $p > .1$.

$M = 3$, $n_1 = n_2 = n_3 = 48$.

$U_i$ = number of "successes" observed at i-th stage.

Step 1: Reject $H_o$ if $U_1 \geq 9$,
accept $H_o$ if $U_1 < 7$,
continue if $7 < U_1 \leq 9$.

Step 2: Reject $H_o$ if $U_2 \geq 8$,
accept $H_o$ if $U_2 < 7$,
continue if $U_2 = 7$.

Step 3: Reject $H_o$ if $U_3 \geq 8$,
accept $H_o$ if $U_3 < 8$.

In other words,

$$\begin{Bmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \end{Bmatrix} = \begin{Bmatrix} 9 & 8 & 8 \\ 7 & 7 & 8 \end{Bmatrix}.$$

From Bureau of Standards binomial tables,

$$\begin{Bmatrix} a_1 & a_2 & a_3 \\ 1\text{-}b_1 & 1\text{-}b_2 & 1\text{-}b_3 \end{Bmatrix} = \begin{Bmatrix} .046 & .102 & .102 \\ .200 & .200 & .102 \end{Bmatrix}.$$

d's by subtraction, (and $d_3 = 0$).

$\alpha = a_1 + d_1 a_2 + d_1 d_2 a_3 = .0575.$

At p = .2,

$$\left\{\begin{matrix} a_1' & a_2' & a_3' \\ b_1' & b_2' & b_3' \end{matrix}\right\} = \left\{\begin{matrix} .642 & .771 & .771 \\ .129 & .129 & .229 \end{matrix}\right\},$$

and $\beta = b_1'+d_1'b_2'+d_1'd_2'b_3' = .164$.

(Usually the probabilities $a_i$ and $b_i$ would be chosen first and cutoff points $A_i$ and $B_i$ derived; but with discrete statistics, integer A's and B's must be chosen, of course with the probabilities in mind, as these really constitute the strategy).

<u>Wald strategy</u>: $H_o$: p = .1, $H_1$: p = .2.

$\alpha = .0575$, $\beta = .164$;

i.e. the s.l.r. test is equivalent to our strategy at these points as far as level and power is concerned. For this particular s.l.r. test the value of β at p = .1462 can be found in Dixon and Massey [3]; it is .6049.

A rough comparison is carried out in the table below:

COMPARISON OF THREE-STEP, S.L.R., AND FIXED-N STRATEGIES IN A BINOMIAL EXAMPLE

OPERATING CHARACTERISTIC

| p | 3-step | s.l.r. |
|---|---|---|
| (0) | (1.0) | (1.0) |
| .1 | 0.9425 | 0.9425 |
| .12 | 0.3273 | |
| .1462 | | 0.6049 |
| .15 | 0.5582 | |
| .18 | 0.2280 | |
| .20 | 0.1640 | 0.1640 |
| .30 | 0.0041 | |
| 1.0 | 0 | 0 |

AVERAGE SAMPLE NUMBER

| p | 3-step | s.l.r. | equiv.<br>fixed-N |
|---|---|---|---|
| (0) | (48.0) | (14.7) | |
| .1 | 56.101 | 39.847 | 74.69 |
| .12 | 60.980 | | 70.70 |
| .1462 | | 56.902 | |
| .15 | 65.044 | | 73.90 |
| .18 | 63.231 | | 74.753 |
| .20 | 60.074 | 44.663 | 68.48 |
| .30 | 49.109 | | 37.77 |
| 1.0 | 48.0 | 3.861 | |

Equivalent fixed sample sizes were calculated from the normal-approximation formula based on values of β actually calculated for the three-stage test at the various values of p.

Rough graphical interpolation of power functions indicates that the OC curves are fairly close together at least for $0 < p \leq .3$. Hence a point-for-point comparison of expected sample sizes is meaningful. Expected sample sizes are compared in fig. 1. For values of p up to about .24, the ASN of the three-stage test is intermediate between Wald's ASN and the fixed sample sizes calculated to produce the same OC. For larger values of p our strategy becomes more expensive than a fixed sample, due to the requirement that at least the first 48 individuals be drawn.

## 4. Example for a Normal Mean, Variance Known

Comparison with fixed-sample-size strategies is easiest in the problem of testing hypotheses about the mean of a normal distribution with known variance $\sigma^2$. We have tried out a number of three-step strategies and a few two- and four-step strategies, seeking to find out empirically what the better strategies look like and how they behave. The table below shows calculations and results for the best three-stage strategy we found so far:

CALCULATION OF THE OC AND ASN

THREE-STEP TEST FOR A NORMAL MEAN

$H_o$: $\mu = \mu_o$. $H_1$: $\mu = \mu_o + \delta\,(\delta > 0)$. $\alpha = .05$

$a_1 = .035$ $1-b_1 = .177857$ $d_1 = .142857$

$a_2 = .07$ $1-b_2 = .141429$ $d_2 = .071429$

$a_3 = .49020$

| $\delta\frac{\sqrt{n}}{\sigma}$ | 0.1 | 0.5 | 2.0 | 4.0 |
|---|---|---|---|---|
| $A_1$ | 1.712 | 1.312 | -0.188 | -2.188 |
| $B_1$ | 0.824 | 0.424 | -1.076 | -3.076 |
| $A_2$ | 1.376 | 0.976 | -0.524 | -2.524 |
| $B_2$ | 0.974 | 0.574 | -0.926 | -2.926 |
| $A_3$ | -0.051 | -0.451 | -1.951 | -3.951 |
| | | | | |
| $a_1'$ | .0434496 | .0947618 | .5745971 | .9856646 |
| $1-b_1'$ | .2049726 | .3357847 | .8590355 | .9989509 |
| $d_1'$ | .1615230 | .2410229 | .2844864 | .0132863 |
| $a_2'$ | .0844134 | .1645351 | .6998585 | .9941981 |
| $1-b_2'$ | .1650311 | .2829862 | .8227741 | .9982830 |
| $d_2'$ | .0806177 | .1184511 | .1229156 | .0040849 |
| $a_3'$ | .5203371 | .6740045 | .9744709 | .9999611 |
| | | | | |
| $1-\beta$ | .0638599 | .1536610 | .8077315 | .9989281 |
| ASN | 1.1745 | 1.2696 | 1.3194 | 1.0133 |
| N* | 1.4811 | 1.5575 | 1.5808 | 1.3891 |
| % saved | 20.7 | 18.5 | 16.5 | 27.1 |

*N = fixed sample size required for $\alpha = .05$ and the same β as that of 3-step test for the same δ.

The following table of % saved by the three-step strategy was calculated in the same way

| $\delta\frac{\sqrt{n}}{\sigma}$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 1.0 |
|---|---|---|---|---|---|---|
| % saved | 20.7 | 19.9 | 20.2 | 19.4 | 18.5 | 14.1 |

| $\delta\frac{\sqrt{n}}{\sigma}$ | 1.5 | 2.0 | 2.5 | 3.0 | 4.0 |
|---|---|---|---|---|---|
| % saved | 13.0 | 16.5 | 22.1 | 26.1 | 27.1 |

The comparison of equivalent sample sizes is illustrated graphically in Figure 2.

The results obtained so far suggest the following conclusions:

(1) Three-step strategies do substantially better than two-step strategies of the sort discussed in this paper (which yield only diminutive savings and frequehtly losses). Soon after three steps a point of diminishing returns is reached: We have been unable to find a four-step strategy which does much better than a good three-step. (A computer will be used to investigate this further).

(2) The more efficient strategies are of the "converging" kind for which the upper and lower cutoff points move closer together at successive steps. (If the evidence in earlier steps has been suggestive enough to warrant continued sampling, it seems logical to reach a decision on the basis of weaker evidence subsequently than would have been required at the first step.) Thus even a strategy as extreme - (in the sense that the third step tests at the 72% level) - as the following, is quite good:

$$a_1 = .03 \quad 1-b_1 = .1966667 \quad d_1 = .1666667$$

$$a_2 = .06 \quad 1-b_2 = .1433334 \quad d_2 = .0833334$$

$$a_3 = .71994 = 1-b_3.$$

| $\delta\frac{\sqrt{n}}{\sigma}$ | 0.1 | 0.5 | 1.0 | 1.5 | 2.0 | 4.0 |
|---|---|---|---|---|---|---|
| % saved | 24.6 | 19.9 | 13.2 | 11.7 | 15.2 | 29.2 |

(3) The choice of $a_1$ is rather crucial in determining a good strategy, as $a_1$ is "undiluted" by factors $d_1$, $d_2$ in the formulae for $\alpha$ and $\beta$. Similarly $n_1$ must be quite important, being undiluted in the formula for the ASN. (However we have only tried out strategies with $n_1 = n_2 = \dots$.)

(4) As might be expected, well-chosen multi-stage strategies are most economical for parameter values very close to those specified by $H_o$ and values quite far from them, and not very helpful for moderate deviations. However, for values extremely far from $H_o$, any given multi-stage strategy becomes uneconomical because of the requirement that no less than $n_1$ readings be taken; (in practice this merely means that the $n_i$'s should be made small if extremely large deviations from $H_o$ are to be anticipated).

## 5. A Nonparametric Example

To illustrate the use of the multi-stage sampling plans with distribution-free statistics, we consider a rank test.

$H_o$: Two continuous populations are identical

$H_1$: Translation.

A sample of 5 is drawn from each population at each step. The statistic "U" is Wilcoxon's rank sum statistic, (not in the form of a Mann-Whitney U, although critical values were obtained by transformation from the Mann-Whitney table).

Step 1: Reject $H_o$, accept $H_o$ or continue according as $U_1 \geq 37$, $U_1 < 32$ or $32 \leq U_1 < 37$;

Step 2: Reject $H_o$, accept $H_o$ or continue according as $U_2 \geq 36$, $U_2 < 33$ or $33 \leq U_2 < 36$;

Step 3: Reject $H_o$ or accept $H_o$ according as $U_3 \geq 24$ or $U_3 < 24$.

From a table of the Wilcoxon distribution (or rather, by conversion from the Mann-Whitney table), we find that the a's are .028, .048 and .726, the (1-b)'s are .210, .155 and .726 and hence the d's are .182, .107 and 0. The idea was to approximate what seemed like a good three-stage strategy with $\alpha = .05$ as closely as possible with the discrete distribution at hand. From the a's and d's we calculate $\alpha = .05087$.

The power function for this rank test cannot be computed without a knowledge of the power function of Wilcoxon's two-sample statistic for fixed sample sizes 5. (The Dutch school worked it out for sizes up to about 3). Therefore we can only study the performance of the strategy - i.e. a strategy using the same $a_i$'s and $b_i$'s - in the parametric comparison of two means with (let us say) known equal variances. The table below shows some percentage savings in the parametric case, calculated exactly as in the example of Section 4:

| d/σ | % Saved by 3-Step |
|---|---|
| 0.06324 | 27.14 |
| 0.3162 | 20.77 |
| 0.6324 | 12.90 |
| 0.6956 | 11.82 |
| 0.7905 | 10.66 |
| 0.8854 | 10.17 |
| 0.9486 | 10.25 |
| 1.0118 | 10.61 |
| 1.1067 | 11.68 |
| 1.2016 | 13.26 |
| 1.2648 | 14.55 |
| 1.8972 | 27.74 |

## 6. A Modification and Refinement

A possible refinement would be to subdivide the zone of indifference at a given step i into subintervals. Sampling is continued whenever $U_i$ falls into any of the subintervals of the middle zone, but the boundaries at the next step are made to depend on the subinterval.

In practice such a subdivision at several successive steps would lead to hopelessly complicated calculations. But the problem can be simplified by going to the limit:

Let the probability of at least the observed $U_i$ at the i-th step given $H_o$ be $P_i$. (In effect we are doing the probability integral transformation). Let $Z_i = \Phi^{-1}(P_i)$ where $\Phi$ is the cumulative standard normal distribution function. Then $Z_i$ at each step is standard normal and is independent of the previous Z's (except that the existence of $Z_i$ (i.e. of an i-th sample) is conditional on earlier Z's being medium-sized). Stop sampling and reject $H_o$ at step M if $\sum_1^M Z_i/\sqrt{M}$ gets big enough. Stop sampling and accept $H_o$ if $\Sigma Z_i/\sqrt{M}$ gets small enough. For intermediate values of $\Sigma Z_i/\sqrt{M}$ continue sampling. For example, the boundaries ("big enough" and "small enough") could be set by an s.l.r. procedure.

Although the method has now been Waldianized in a sense, it cannot generally be made equivalent to Wald's test. This would mean using $n_i$'s = 1 (at least from a certain point on), which sometimes leads to absurdities. E.g. in sign tests, taking one pair at a time contributes a probability p = 1/2 or 1 at each step, thus $Z_i = 0$ (no contribution) or $-\infty$, so that $H_1$ is almost inevitably accepted even if $H_o$ is true. It is thus necessary that several readings be taken at each step, enough to yield something of a probability distribution each time.

. . . . . . . . .

## References

1. Wald, A., Sequential Analysis, Wiley, 1947.

2. Hastay, W.; Wallis, W. A.; and Eisenhardt, C. Selected Techniques of Statistical Analysis, McGraw-Hill, 1947.

3. Dixon, W. J., and Massey, F. J., Introduction to Statistical Analysis, McGraw-Hill, 2nd Ed., 1957. (See p. 305).

4. National Bureau of Standards, Tables of the Binomial Probability Distribution, U. S. Gov't Printing Office, 1949.

5. Davies, O. L., Design and Analysis of Industrial Experiments, Hafner, 1956.

6. Armitage, P., Sequential Medical Trials, Charles C. Thomas, Springfield, 1960.

7. Chilton, N. W.; Fertig, J. W.; and Kutcher, A. H., "Studies in the Design and Analysis of Dental Experiments, III Sequential Analysis, (Double Dichotomy)", J. Dental Research, 40, No. 2, 1961, 331-340.

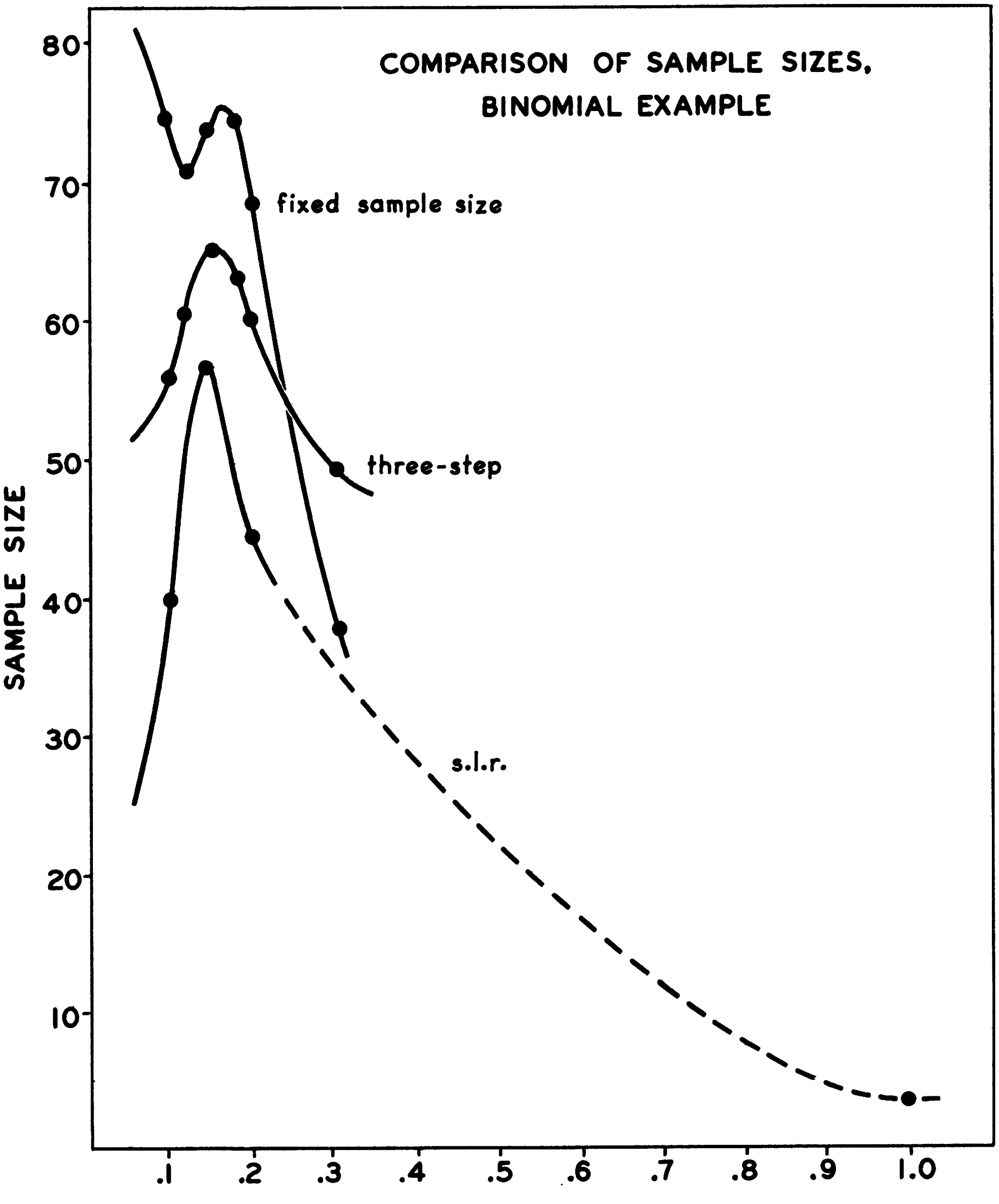

COMPARISON OF SAMPLE SIZES,
BINOMIAL EXAMPLE
SAMPLE SIZE
80
70
60
50
40
30
20
10
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
TRUE p
fixed sample size
three-step
s.l.r.

FIGURE 1.

## COMPARISON OF 3-STEP ASN AND EQUIVALENT FIXED SAMPLE SIZE IN TESTS FOR A NORMAL MEAN

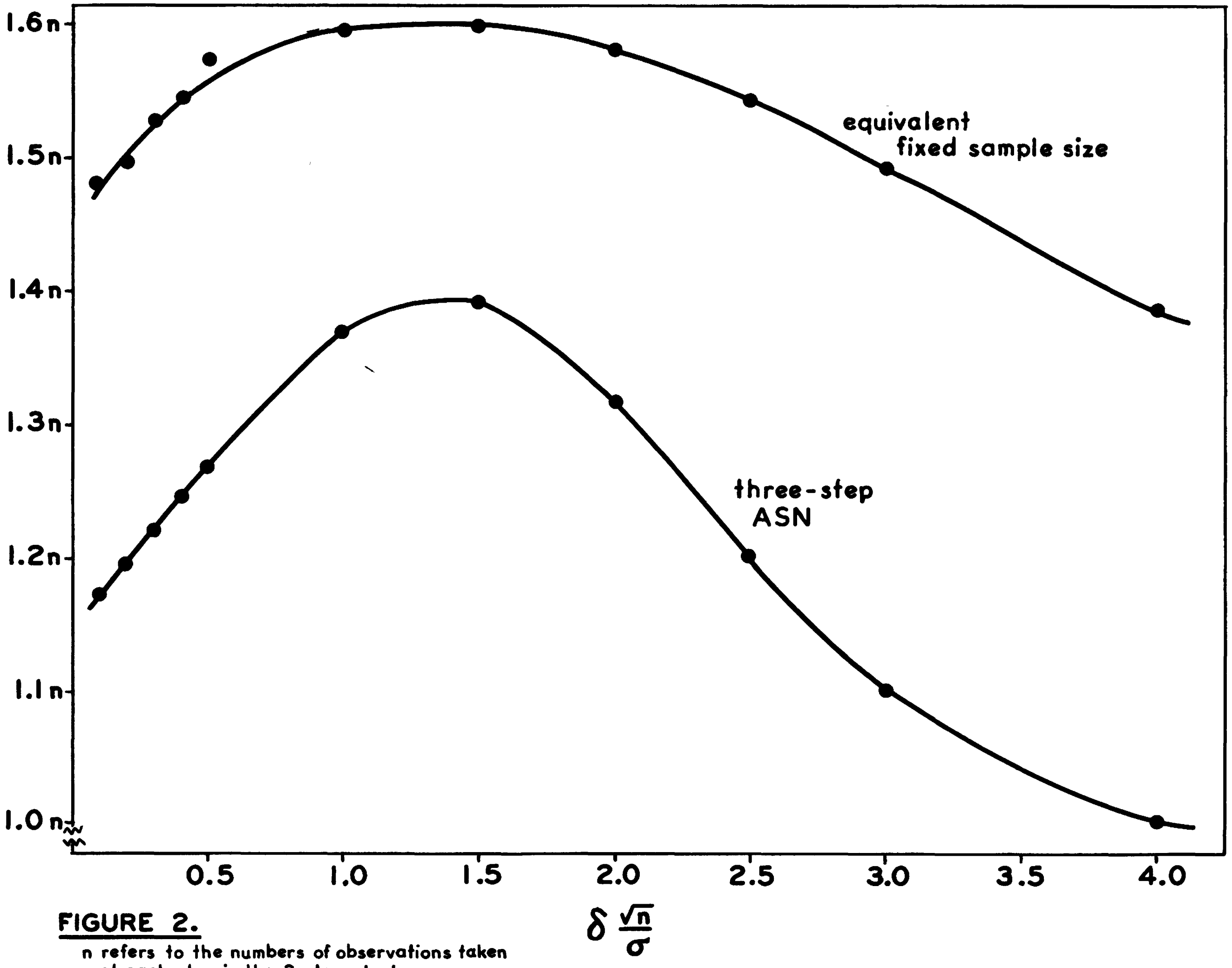


FIGURE 2.

n refers to the numbers of observations taken at each step in the 3-stage test.