A SEQUENTIAL TEST OF WEAK INEQUALITIES: THE PREDICTION OF PRESIDENTS

R. A. Weitzman, Naval Postgraduate School

0. ABSTRACT

Important potential test situations exist in which hypotheses of equality or strong inequality are inadmissible. One example is the prediction of election results in a two-candidate contest. Both candidates cannot win. Using as an illustration the results of 39 Presidential elections with majorities ranging between .652 to just over .500, this article describes a sequential test of weak inequalities (H_1: $p < p_0$ vs. H_2: $p > p_0$) which has a specifiable total probability of error (α_T). In the illustration presented, the meaning of this probability is as follows: If the test with equal α_T had been used in all 39 elections, prediction would have been in error in approximately $39\alpha_T$ of them. Monte Carlo trials provided a vehicle to evaluate the test. On 1,000 of these trials with $\alpha_T = .100$, the proportion of erroneous predictions was .105, and the average sample number over all 39 elections was 2,188.

1. INTRODUCTION

Do differences observed in samples reflect true population differences? A statistical test, which in the Neyman-Pearson formulation [3] can answer this question either yes (with the risk of a Type I error) or no (with the risk of a Type II error), can also, in the Fisher [1, Chapter 2] formulation, fail to answer the question (an insignificant result) or, answering it, answer it only in the affirmative (a significant result). Neyman [2] reviews the controversy between these two opposing formulations. Though the tendency over the years has been increasingly to adopt the Neyman-Pearson formulation in both textbooks and research reports, the practice in specific subject-matter areas has not always been consistent. In psychology, for example, while textbooks typically present the Neyman-Pearson formulation, research reports continue to reflect the influence of Fisher in such statements as "The result is significant ($p < .05$)" or "The result is not significant ($p > .05$)," where p indicates the probability that the result (or a more extreme result) is simply due to sampling error. The purpose here, however, is not to evaluate either formulation, especially relative to the other, but rather to present a hybrid formulation applicable particularly when hypotheses of equality or strong inequality are inadmissible. Tests so formulated turn out to be adaptations of the sequential methods developed by Wald [4] to the remaining choice between complementary weak inequalities.

2. A HYBRID FISHER - NEYMAN-PEARSON FORMULATION

Inadmissibility of hypotheses of equality or strong inequality is rather common. Two-candidate elections constitute a familiar example. Ties are inadmissible: One candidate must win, one must lose. The null hypothesis (H_0) that the two candidates are equally popular must thus be false.

If this hypothesis is false, conversely, then one candidate must be more popular than the other. Deciding that one candidate is more popular than the other when the reverse is true is under these conditions an all-inclusive error having unconditional, or total, probability

$$\alpha_T = \alpha_1 P_1 + \alpha_2 P_2 , \tag{2.1}$$

where α_i $(i = 1, 2)$ is the conditional probability of incorrectly deciding that candidate i is more popular and P_i $(i = 1, 2)$ is the (prior) probability that candidate i is in fact less popular. Fairness to both candidates requires that $\alpha_1 = \alpha_2 = \alpha$ so that

$$\alpha_T = \alpha(P_1 + P_2) . \tag{2.2}$$

If equal popularity is inadmissible, $P_1 + P_2 = 1$; therefore, $\alpha_T = \alpha$: The total probability of error is equal to either one of the two (equal) conditional probabilities of error.

This formulation thus resembles Fisher's in its exclusion of the acceptability of H_0 and Neyman-Pearson's in its inclusion of the probability of error.

3. SEQUENTIAL TESTING

Application in the form of a statistical test requires sequential data collection until the test statistic reaches a value that favors one hypothesis (Candidate 1 is more popular) or the other (Candidate 2 is more popular).

As developed by Wald [4], sequential tests typically apply to hypotheses of strict equality: $H_1 : \theta = \theta_1$ and $H_2 : \theta = \theta_2$. The basic test statistic is the likelihood ratio

$$L_n = \frac{f_2(\underline{x}_n)}{f_1(\underline{x}_n)} , \tag{3.1}$$

where $\underline{x}_n = \{x_1, x_2, \ldots, x_n\}$ is the n-valued observation vector and f_i $(i = 1, 2)$ is the probability (density) of this vector if hypothesis i is true. (In this formulation, the scalar x's are n successive observations on a single variable.) If α_1 is the maximal probability of error in accepting hypothesis 1 and α_2 is the maximal probability of error in accepting hypothesis 2, then sampling continues (n increases) until L_n is smaller than $\alpha_2/(1-\alpha_1)$ or larger than $(1-\alpha_2)/\alpha_1$. Acceptance of hypothesis 1 occurs in the first case, of hypothesis 2 in the second.

Sampling is typically independent so that

$$f_1(\underline{x}_n) = \prod_{j=1}^{n} f_1(x_j) \tag{3.2}$$

and

$$f_2(\underline{x}_n) = \prod_{j=1}^{n} f_2(x_j) . \tag{3.3}$$

If the population is dichotomous with $H_1 : p = p_1$ and $H_2 : p = p_2$, for example, then

$$f_i(\underline{x}_n) = \prod_{j=1}^{n} p_i^{x_j}(1-p_i)^{1-x_j} \quad (i=1,2) \tag{3.4}$$

where X is a 0-1 binary random variable and $p = E(X)$. In adapting sequential testing to the choice between weak inequalities, we shall confine ourselves to this case of independent sampling from a dichotomous population.

The adaptation requires setting $\alpha_2 = \alpha_1 = \alpha$ and determining $f_1(\underline{x}_n)$ for $p < p_o$ and $f_2(\underline{x}_n)$ for $p > p_o$. If $p_{11}, p_{12}, \ldots, p_{1N_1}$ constitute all the observed or known values of $p < p_o$ and $p_{21}, p_{22}, \ldots, p_{2N_2}$ constitute all the observed or known values of $p > p_o$, then

$$f_1(\underline{x}_n) = N_1^{-1} \sum_{i=1}^{N_1} \prod_{j=1}^{n} p_{1i}^{x_j}(1-p_{1i})^{1-x_j} \tag{3.5}$$

and

$$f_2(\underline{x}_n) = N_2^{-1} \sum_{i=1}^{N_2} \prod_{j=1}^{n} p_{2i}^{x_j}(1-p_{2i})^{1-x_j} . \tag{3.6}$$

These equations are, in fact, raw-data forms of general distributional equations presented by Wald [4] for sequential tests of composite hypotheses. The case in which the p's constitute all the known (as opposed to theoretical) values is essentially an empirical Bayes case. This is the case that we shall consider in our illustration.

Computational note: When the true value of p is near .5, some of the products in L_n may become so small as to cause a computer underflow. Multiplication of both the numerator and the denominator by a number greater than one will correct this problem without changing the value of L_n. Continued multiplication may be necessary if this problem recurs, and, when this is the case, another problem may occur: Some of the products may become so large as to cause a computer overflow. Long before this problem occurs, however, the smallest products may be set equal to zero without noticeably changing the value of L_n.

4. THE PREDICTION OF ELECTION RESULTS

Voters in two-candidate elections constitute a dichotomous population. The history of N two-candidate elections for a particular office, like the United States' Presidency, provides a record of complementary values of p_{1i} and p_{2i} such that $p_{2i} = 1 - p_{1i}$ $(i = 1, 2, \ldots, N)$. Table 1 (Tables follow the References) shows values of p_{2i} based on the top two candidates $(p_{2i} > p_{1i})$ in 39 $(N = 39)$ successive Presidential elections beginning in 1824 (Jackson vs. Adams), the first Presidential election for which there was a popular vote, and ending in 1976 (Carter vs. Ford), the most recent Presidential election. Monte Carlo analysis using these data illustrates how the test just developed works.

In this analysis, with $\alpha_T = .10$, the 39 elections had an equal probability (1/39) of selection on each of 1,000 trials. Each trial ended in the choice of one candidate or the other depending on the value of L_n at the conclusion of the test on that trial:

$$L_n = \frac{\sum_{i=1}^{N} \prod_{j=1}^{n} p_i^{x_j}(1 - p_i)^{1 - x_j}}{\sum_{i=1}^{N} \prod_{j=1}^{n} (1 - p_i)^{x_j} p_i^{1 - x_j}} , \tag{4.1}$$

where, with $p_i = p_{2i}$, $x_j = 1$ if the j-sampled voter favored Candidate 2 and $x_j = 0$ if the j-sampled voter favored Candidate 1. Of the 1,000 choices, 105 were in error, which is close to the nominal error rate of 100/1,000 $(\alpha_T = .10)$.

The average sample number for these 1,000 trials was 2,188. Table 2 shows the average sample number, together with the number of trials, for each of the 39 elections. Since the distribution is highly skewed, the median average sample number, 239, would seem to be more representative than the over-all average sample number. On approximately half of all the elections, the test required sampling no more than 239 voters.

In the case of several elections, however, sampling many more than this number of voters tended to be necessary. In the election of 1880 (Garfield vs. Hancock), for example, the average sample number was 18,730. This is, comparatively, a large number, but the number required for a corresponding 90% confidence interval that excludes .500 is even larger: 4,337,189. This number, indeed, is only slightly smaller than half the total 1880 electorate (8,891,083)!

Table 2 also presents the observed error rate for each election. Different from classical tests or sequential tests of point hypotheses, this rate varies systematically around the nominal error rate (.10). The correlation between majority and observed error rate is, in fact, -.71. The error rate has a rather pronounced tendency to be greater for majorities close to

.500 than for majorities far from .500. Over all elections, however, the error rate tends (as noted earlier) to approximate its nominal value.

5. DISCUSSION

The preceding example well illustrates the results obtainable from a sequential weak-inequalities test (SWIT). If a SWIT were applied with $\alpha_T = .10$ to all 39 Presidential elections for which there was a popular vote, then the results would tend to be in error on no more than 10% of these elections. (If the test were applied to these elections with $\alpha_T = .025$, the results would tend to be in error on less than a single election.) A SWIT, like Baysian analysis, systematically takes past experience into account. Confidence-interval estimation, by contrast, refers to a hypothetical future. If the sampling procedure were to be repeated innumerable times to construct a 90% confidence interval, for example, the intervals constructed would contain the population value on approximately 90% of the repetitions. Every time one of these intervals contained .5, no decision would be possible. A SWIT always results in a decision. Being sequential, a SWIT shares advantages of other sequential tests, particularly regarding sample size. The average sample number of a sequential test is, as Wald [4] has shown, uniformly and often substantially smaller than the sample size required by a corresponding classical procedure. Perhaps the most important advantage of a SWIT has to do with the probability of error. In a classical test, not only does this probability generally have a different value for each of the two possible decisions, but also the value for only one of these decisions is known. The probability of error in a SWIT, which is in fact the total error probability, has the same, known value for each of the two possible decisions.

The usefulness of a SWIT for the prediction of election results depends, of course, on the resolution of practical sampling problems. Useful application may require more information, particularly about average sample numbers, than provided by the illustration presented here. The Monte Carlo analysis with $\alpha_T = .10$ required over 295 minutes of computer time. The time required for extending this analysis to smaller values of α_T would be prohibitive. This time depends not only on the value of α_T but also on the distribution of p values. Rather than the entire observed distribution, a pollster may wish to direct his inference to only a subset of the p values--for example, the subset corresponding to elections in which the current President is seeking a second term. (Occurring in the election of 1888, the lowest p value for this subset is .504.) The time required for a 1,000-trial Monte Carlo analysis may in this case be no greater than 300 minutes even for values of α_T smaller than .10. If the times for analysis are about the same, then the average sample numbers also ought to be about the same. In applications of particular interest, therefore, average sample numbers for SWITs in which $\alpha_T = .05$ or $\alpha_T = .01$ may not differ substantially from the average sample number obtained here for $\alpha_T = .10$.

The intention of the illustration presented was not to provide practical information, however, but to facilitate the description of a SWIT and to indicate at least one area of potential applicability. The requirements of a SWIT in this area are, taken together, somewhat unique: Independent sampling from a dichotomous population with empirically known prior probabilities. SWITs applied to other areas will generally have to meet different sets of requirements.

REFERENCES

[1] Fisher, R. A., The Design of Experiments, 7th ed., New York: Hafner, 1960.

[2] Neyman, J., "Silver Jubilee of my Dispute with Fisher," Journal of the Operations Research Society of Japan, 3, No. 3 (1961), 145-54.

[3] Neyman, J., and Pearson, E. S., "On the use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," Biometrika, 20A (July 1928), 175-240 (Part I) and 263-94 (Part II).

[4] Wald, A. "Sequential Tests of Statistical Hypotheses," Annals of Mathematical Statistics, 16, No. 2 (1945), 117-86.

1. Popular Vote for President

Date	Winner	Vote	Loser	Vote	Majority
1824	Jackson	155,872	Adams	105,321	.597
1828	Jackson	647,231	Adams	509,097	.560
1832	Jackson	687,502	Clay	530,189	.565
1836	Van Buren	762,678	Harrison	548,007	.582
1840	Harrison	1,275,017	Van Buren	1,128,702	.530
1844	Polk	1,337,243	Clay	1,299,068	.507
1848	Taylor	1,360,101	Cass	1,220,544	.527
1852	Pierce	1,601,474	Scott	1,386,578	.536
1856	Buchanan	1,927,995	Fremont	1,391,555	.581
1860	Lincoln	1,866,352	Douglas	1,375,157	.576
1864	Lincoln	2,216,067	McClellan	1,808,725	.551
1868	Grant	3,015,071	Seymour	2,709,615	.527
1872	Grant	3,597,070	Greeley	2,834,079	.559
1876	Hayes	4,284,757	Tilden	4,033,950	.515
1880	Garfield	4,449,053	Hancock	4,442,030	.500
1884	Cleveland	4,911,017	Blaine	4,848,334	.503
1888	Harrison	5,540,050	Cleveland	5,444,337	.504
1892	Cleveland	5,554,414	Harrison	5,109,802	.517
1896	McKinley	7,035,638	Bryan	6,467,946	.521
1900	McKinley	7,219,530	Bryan	6,358,071	.532
1904	Roosevelt	7,628,834	Parker	5,084,491	.600
1908	Taft	7,679,006	Bryan	6,409,106	.545
1912	Wilson	6,286,214	Roosevelt	4,216,020	.599
1916	Wilson	9,129,606	Hughes	8,538,221	.517
1920	Harding	16,152,200	Cox	9,147,353	.638
1924	Coolidge	15,725,016	Davis	8,385,586	.652
1928	Hoover	21,392,190	Smith	15,016,443	.588
1932	Roosevelt	22,821,857	Hoover	15,761,841	.591
1936	Roosevelt	27,751,597	Landon	16,679,583	.625
1940	Roosevelt	27,243,466	Wilkie	22,304,755	.550
1944	Roosevelt	25,602,505	Dewey	22,006,278	.538
1948	Truman	24,105,812	Dewey	21,970,065	.523
1952	Eisenhower	33,936,252	Stevenson	27,314,992	.554
1956	Eisenhower	35,585,316	Stevenson	26,031,322	.578
1960	Kennedy	34,227,096	Nixon	34,108,546	.501
1964	Johnson	43,126,506	Goldwater	27,176,789	.613
1968	Nixon	31,785,480	Humphrey	31,275,166	.504
1972	Nixon	47,165,234	McGovern	28,168,110	.626
1976	Carter	40,825,839	Ford	39,147,770	.510

Source: These data come from The World Almanac and Book of Facts 1978 (published in 1977 by Newspaper Enterprise Association, New York), p. 286.

2. Average Sample Number (ASN) and Error Rate in Monte Carlo Analysis

Majority[a]	ASN	Frequency	Error
.500	18,730	23	.522
.501	37,562	28	.393
.503	1,884	18	.556
.504	5,759	26	.385
.504	8,565	24	.292
.507	1,070	20	.550
.510	2,353	18	.333
.515	1,457	24	.040
.517	751	30	.100
.517	565	24	.167
.521	797	34	.088
.523	791	37	.162
.527	464	20	.100
.527	386	26	.154
.530	447	23	.087
.532	317	25	.080
.536	483	30	.100
.538	184	30	.067
.545	281	38	.026
.550	258	22	.000
.551	179	18	.111
.554	193	26	.038
.559	239	25	.000
.560	142	31	.000
.565	134	21	.048
.576	106	26	.000
.578	115	20	.000
.581	120	28	.000
.582	111	26	.038
.588	70	20	.000
.591	100	23	.000
.597	81	26	.000
.599	82	27	.000
.600	75	33	.000
.613	60	24	.000
.625	54	20	.000
.626	47	33	.000
.638	41	33	.000
.652	47	19	.000

[a]As in Table 1, the majorities indicated here are only 3-place approximations; for example, .500 is an approximation of the actual majority, 4,449,053/(4,449,053 + 4,442,030).