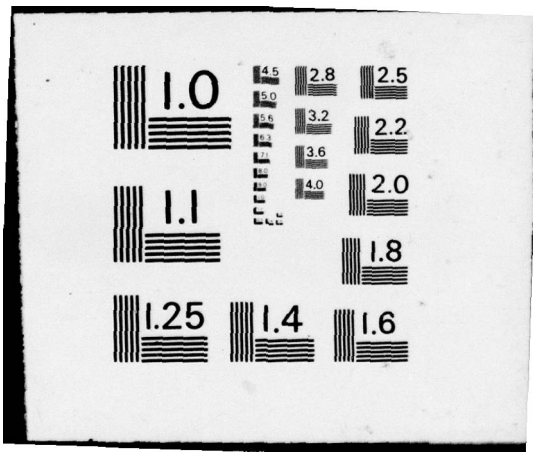
AD-A031 337 NAVAL POSTGRADUATE SCHOOL MONTEREY CALIF F/G 12/2
SUPERIORITY OF FIT.(U)
SEP 76 R A WEITZMAN
UNCLASSIFIED NPS-54WZ76091 NL

1 OF 1
ADA031337

END

DATE
FILMED
12 - 76


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6

12

NPS54Wz76091

NAVAL POSTGRADUATE SCHOOL

Monterey, California

SUPERIORITY OF FIT

by

R. A. Weitzman

September 1976

NAVAL POSTGRADUATE SCHOOL
Monterey, CA. 93940

Rear Admiral Isham Linder
Superintendent

Jack R. Borsting
Provost

The work reported herein was supported by the Naval Postgraduate School.

This report was prepared by:

R. A. WEITZMAN
Associate Professor

Reviewed by:

C. R. JONES, Chairman
Department of
Administrative Sciences

Released by:

ROBERT FOSSUM
Dean of Research

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPS54Wz76091
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): Superiority of Fit
5. TYPE OF REPORT & PERIOD COVERED: Technical Report (Final)
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): R. A. Weitzman
8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Dept. of Administrative Sciences Code 54
Naval Postgraduate School
Monterey, CA 93940
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Naval Postgraduate School
Monterey, CA 93940
12. REPORT DATE: Sept 1976
13. NUMBER OF PAGES: 12p.

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): NPS-54Wz76091
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Statistical test
Goodness of Fit
Superiority of Fit
Hypothesis testing

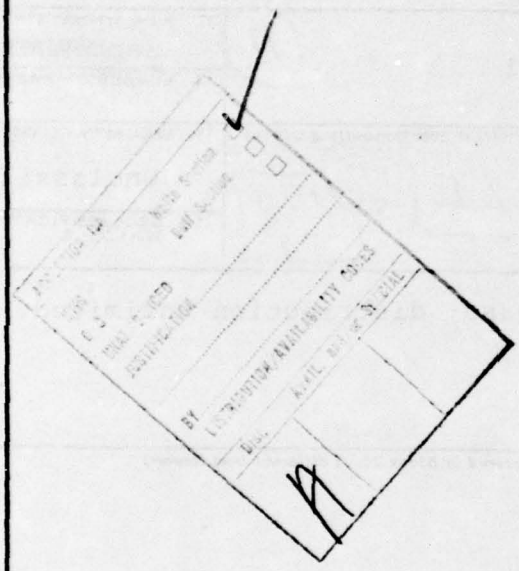
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
In the context of both the Fisher and the Neyman-Pearson formulations, this article develops a method of testing superiority of fit to determine which of two different numerical predictions is more nearly accurate. Based on the assumption that the two predictions are not equally nearly accurate, the method is sequential, continuing until a decision is reached in favor of one prediction or the other. The error ~~of~~ probability of this
OVER

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

251 450

decision is specifiable in advance of the test.



SUPERIORITY OF FIT

1. INTRODUCTION

Do discrepancies between observations and predictions indicate true population differences? A statistical test, which in the Neyman-Pearson formulation [4] 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 formulation [1, Chapter 2], fail to answer the question (an insignificant result) or, answering it, answer it only in the affirmative (a significant result). Neyman [3] 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 to the evaluation of numerical predictions.

2. A HYBRID FISHER - NEYMAN-PEARSON FORMULATION

This formulation rests on a widespread belief among philosophers of science [e.g., 2, Chapter 4, particularly p. 78] that no numerical prediction is precisely accurate. Testing the

accuracy of a single numerical prediction thus makes no sense: Use of a large enough sample will always lead to the rejection of the prediction as inaccurate. To rule out tests of goodness of fit, however, is not to rule out tests of superiority of fit. Testing the relative accuracy of two different numerical predictions on the same set of observations does make sense. The null hypothesis (H_0) of such a test (the hypothesis to be nullified, in Fisher's terminology) is simply that the two predictions are equally inaccurate. Equal inaccuracy implies that the population value is midway between the two predictions so that their mean is itself a precisely accurate prediction. The null hypothesis of equal inaccuracy must thus be false.

If this hypothesis is false, however, then one of the two predictions must be more accurate that the other. Deciding that one prediction is more accurate than the other when the reverse is true is thus 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 prediction i is less accurate and P_i (i=1,2) is the (prior) probability that prediction i is in fact more accurate. Fairness to both predictions requires that $\alpha_1 = \alpha_2 = \alpha$ so that

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

If equal inaccuracy is impossible, $P_1 + P_2 = 1$, and thus $\alpha_T = \alpha$: 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. TESTING SUPERIORITY OF FIT

Application in the form of a statistical test requires specification of the null hypothesis and sequential data collection until the rejection of this hypothesis occurs.

Since the null value (θ) is midway between the two predicted values $(\theta_1$ and $\theta_2)$, the null hypothesis is $H_0: \theta = (\theta_1 + \theta_2)/2$. The equality sign in this hypothesis shows that the test is two-tailed. Rejection of H_0 occurs when the test statistic falls in either tail of the sampling distribution that the test statistic would have if H_0 were true. The decision that follows, that one or the other prediction is more accurate, depends on which tail this is. Either decision has a probability of error equal to α, the area under each tail, which is also the total probability of error.

Sequential data collection is necessary to avoid the acceptance of H_0, which, according to the belief that no prediction is precisely accurate, is impossible. Sampling thus proceeds one sampling unit at a time. Computation of the test statistic (or an equivalent value) T follows the sampling of each sampling unit along with the determination (if necessary) of appropriate critical values, t_1

and t_2 ($t_1 < t_2$). The decision depends on the relationship between T and t_1 and t_2: If $T < t_1$, the decision is that prediction 1 is more accurate than prediction 2; if $T > t_2$, the decision is that prediction 2 is more accurate than prediction 1. If $t_1 < T < t_2$, however, the decision is to continue sampling.

4. AN ILLUSTRATION OF THE METHOD

On the first day of school, an instructor of a large class administers a preliminary examination consisting of many items, each scorable as correct or incorrect. Automatic scoring immediately following the examination shows that the mean proportion of items answered correctly is π. Knowing π, the instructor then asks a student selected randomly from the class a question selected randomly from the examination in a process that continues, if necessary, for a parallel succession of students and questions until the answer received is correct. Recording the number (X) of answers received prior to the correct answer, the instructor repeats the process for trial after trial in order to determine whether a geometric distribution with parameter π or a Poisson distribution with parameter $(1 - \pi)/\pi$ more accurately predicts the variance of X. Each of these distributions predicts the same mean: $\mu = (1 - \pi)/\pi$.

If the variance of X is σ_1^2 for the Poisson and σ_2^2 for the geometric distribution, the null hypothesis is $H_0: \sigma^2 = (\sigma_1^2 + \sigma_2^2)/2$, where

$$\sigma_1^2 = (1 - \pi)/\pi \qquad (4.1)$$

and

$$\sigma_2^2 = (1 - \pi)/\pi^2. \qquad (4.2)$$

The test statistic is chi square divided by its degress of freedom:

$$T = \sum^{N} (X - \mu)^2/N\sigma^2, \qquad (4.3)$$

which (μ being known) has N degrees of freedom after the sampling of N sampling units. Using μ and σ^2 as constants, an electronic calculator determines T for each pair of X and N values, and the instructor plots these values on a graph on which two lines join the critical values t_1 and t_2 for successive values of N. Figure A shows the results for $\pi = .2$ and $\alpha_T = .05$. After five ~~questions~~ trials, yielding the succession of X values 2, 5, 4, 6, and 4, the instructor rejects H_0, deciding with a probability of error equal to .05 that the Poisson distribution predicts the variance more accurately than the geometric distribution.

5. DISCUSSION OF THE ILLUSTRATION

Since the succession of incorrect and correct answers constitutes a Bernoulli process, the distribution of X ought to be geometric, not Poisson. The result, however, is not one of the five errors that can occur in every one hundred repetitions of the test. The illustration is fictitious. The product of simulation, the five observations tend in fact to follow a Poisson distribution with parameter equal to 4. The histogram in Figure B describes this distribution. As a general rule, for samples as small as five, the probability of error approximates its nominal value to the extent that the observations follow a normal distribution. Since the histogram in Figure B tends to be unimodal and symmetrical like a normal

FIGURE A. TEST STATISTIC (DOTS) AND CRITICAL VALUES (t_1 AND t_2) AS A FUNCTION OF TRIAL NUMBER (n)

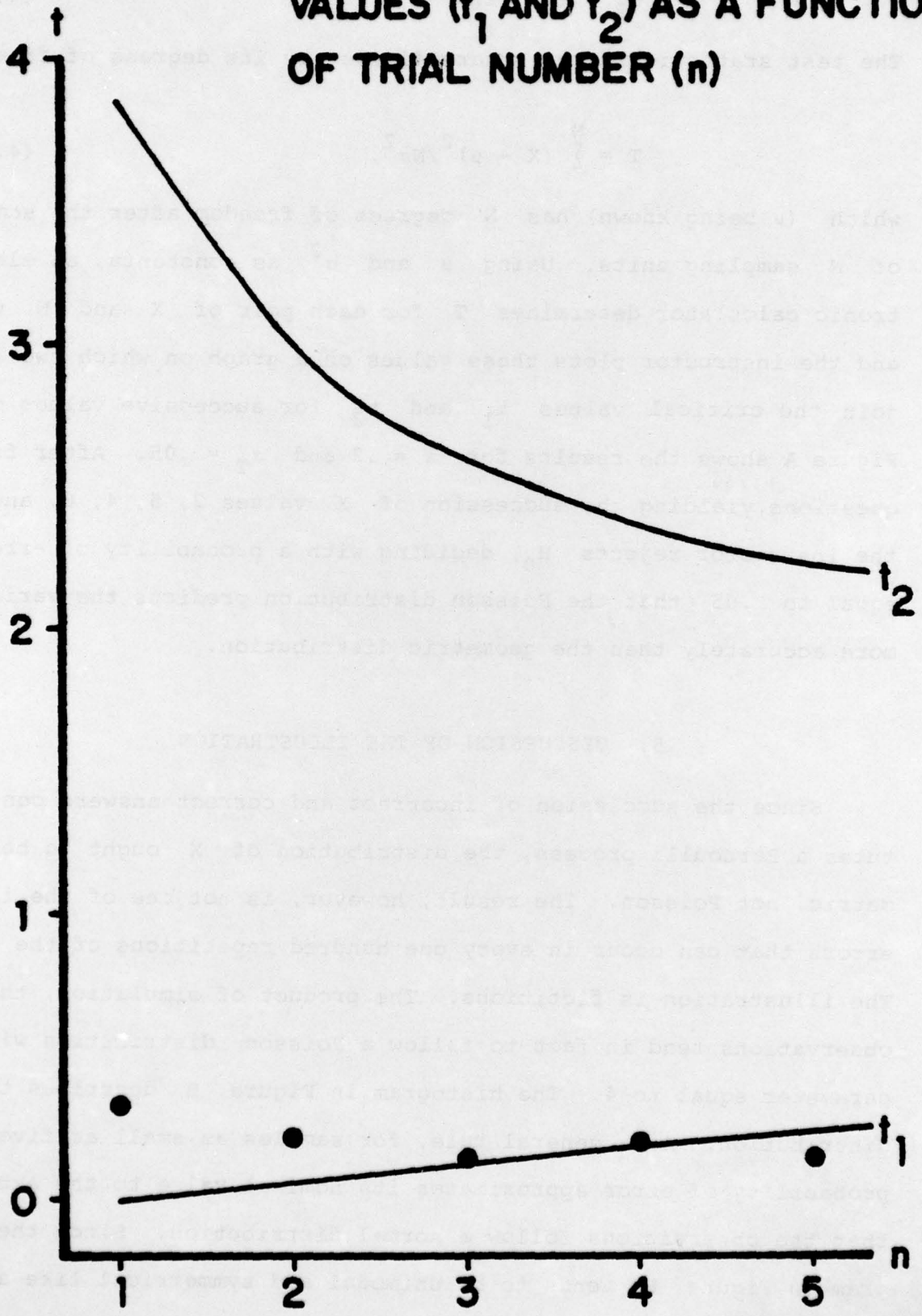
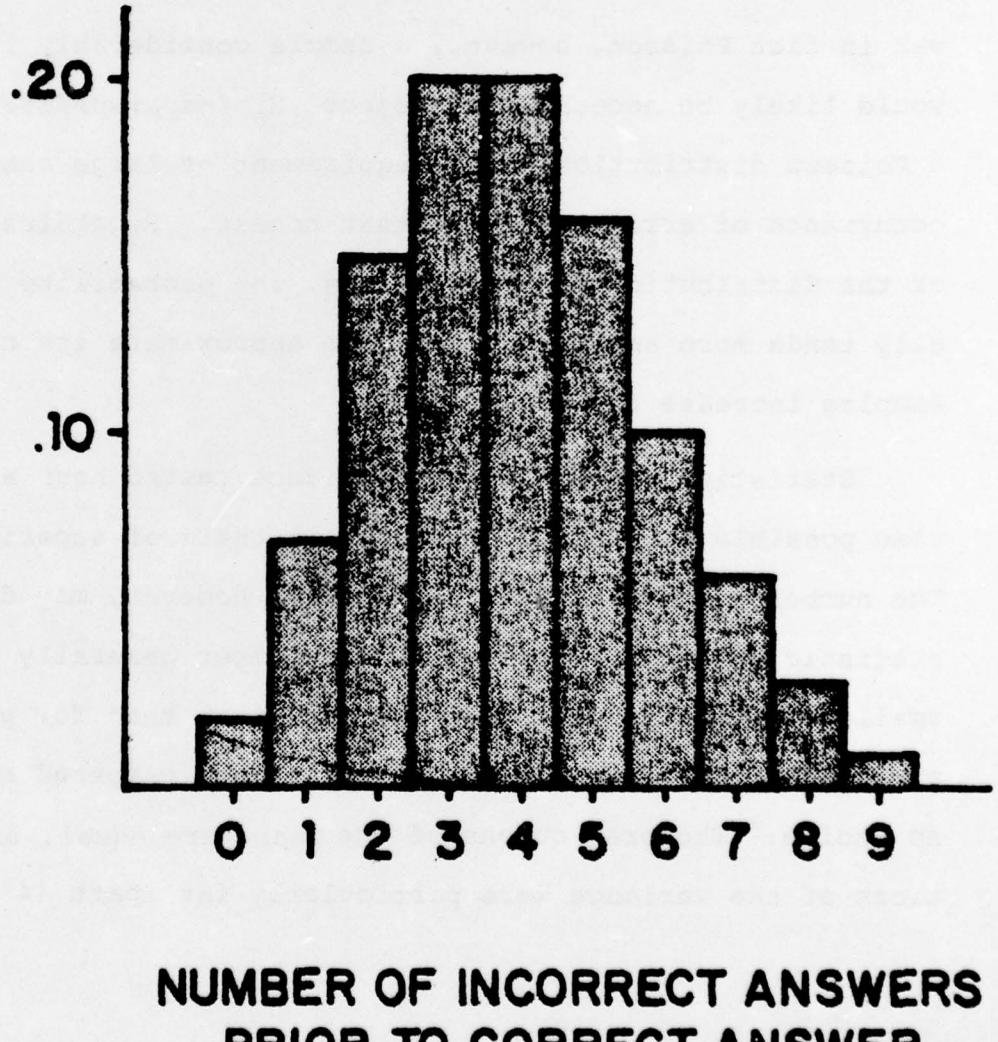


FIGURE B. HISTOGRAM OF POISSON DISTRIBUTION WITH PARAMETER 4



curve, therefore, the total probability of error ought to be close to its nominal value of .05 despite the small sample size.

Such early rejection of H_0 in a sequential test is ordinarily not so defensible. If the sampling distribution of X were closer to the geometric than the Poisson, for example, the histogram in Figure B might be skewed sufficiently to have a substantial effect on the total probability of error, particularly for low N. Since five observations were necessary to reject H_0 in favor of a Poisson distribution even when the distribution of the observations was in fact Poisson, however, a sample considerably larger than five would likely be necessary to reject H_0 *inappropriately* in favor of a Poisson distribution. The requirement of large samples for the occurrence of error keeps the test honest. Regardless of the form of the distribution of observations, the probability of error generally tends more and more closely to approximate its nominal value as samples increase in size.

Statistics other than the variance tested here are, of course, also possible targets of inference in tests of superiority of fit. The number of sampling units required, however, may depend on the statistic chosen for testing. This number generally ought to be smaller for predictions that are far apart than for predictions that are close together. The two distributions compared here thus allowed no choice: The predictions of the mean were equal, and the predictions of the variance were particularly far apart (4 versus 20).

REFERENCES

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

[2] Kemeny, J. G., A Philosopher Looks at Science, Princeton, N.J.: Van Nostrand, 1959.

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

[4] 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).

INITIAL DISTRIBUTION LIST

	No. Copies
Defense Documentation Center Cameron Station, Building 5 5010 Duke Street Alexandria, VA 22314	12
Dean of Research Code 023 Naval Postgraduate School Monterey, CA 93940	1
Library (Code 0212) Naval Postgraduate School Monterey, CA 93940	2
Library (Code 54) Naval Postgraduate School Monterey, CA 93940	2
Professor D. R. Barr (Code 55) Professor D. P. Gaver, Jr. (Code 55) Professor C. R. Jones (Code 54) Professor R. A. Weitzman (Code 54) Naval Postgraduate School Monterey, CA 93940	1 1 1 20
Dr. Gordon M. Becker Department of Psychology University of Nebraska Omaha, Nebraska 68132	1
Dr. Harold Gulliksen Department of Psychology Princeton University Princeton, NJ 08540	1