ON BIAS IN ESTIMATION DUE TO THE USE OF PRELIMINARY TESTS OF SIGNIFICANCE

By T. A. Bancroft

Iowa State College

I. INTRODUCTION

In problems of statistical estimation we often express the joint frequency distribution of the sample observations \( x_1, x_2, \cdots x_n \) in the form

\[
(1) \quad f(x_1, \cdots, x_n ; \alpha, \beta, \gamma, \cdots) \Pi dx_i, \quad (i = 1, \cdots, n)
\]

where the functional form, \( f \), is assumed known, and \( \alpha, \beta, \gamma, \cdots \) are certain population parameters whose values may or may not be known. Given this specification, statistical theory provides routine mathematical processes for obtaining estimates of the parameters \( \alpha, \beta, \gamma, \cdots \) from the observations \( x_1, x_2, \cdots, x_n \).

In performing tests of significance we often assume that the data follow some distribution

\[
(2) \quad f_1(x_1, \cdots, x_n ; \alpha, \beta, \gamma, \cdots) \Pi dx_i, \quad (i = 1, \cdots, n)
\]

where \( f_1 \) is a known function or family of functions. We may wish to test the hypothesis that the data follow the more specialized distribution

\[
(3) \quad f_2(x_1, \cdots, x_n ; \alpha', \beta', \gamma', \cdots) \Pi dx_i, \quad (i = 1, \cdots, n)
\]

where \( f_2 \) is some member or sub-group of the family \( f_1 \). Given this specification, statistical theory provides routine mathematical processes for testing such hypotheses.

In the application of statistical theory to specific data, there is often some uncertainty about the appropriate specifications in equations (1), (2) and (3). In such cases preliminary tests of significance have been used, in practice, as an aid in choosing a specification. We shall give several examples from the literature of statistical methodology.

(1) In an analysis of variance, in order to obtain a best estimate of variance, we may be uncertain as to whether two mean squares in the lines of the analysis may be assumed homogeneous, [1]. Suppose that it is desired to estimate the variance \( \sigma_1^2 \), of which an unbiased estimate \( s_1^2 \) is available. In addition, there is an unbiased estimate \( s_2^2 \) of \( \sigma_2^2 \), where from the nature of the data it is known that either \( \sigma_2^2 = \sigma_1^2 \) or \( \sigma_2^2 < \sigma_1^2 \). As a criterion in making a decision the following rule of procedure is used frequently: test \( s_1^2/s_2^2 \) by the \( F \)-test, where \( s_1^2 \) and \( s_2^2 \) are the two mean squares. If \( F \) is not significant at some assigned significance level use \( (n_1s_1^2 + n_2s_2^2)/(n_1 + n_2) \) as the estimate of \( \sigma_1^2 \). If \( F \) is significant at the assigned significance level, use \( s_2^2 \) as the estimate of \( \sigma_1^2 \).

(2) After working out the regression of \( y \) on a number of independent variates we may be uncertain as to the appropriateness of the retention of some one of