ERROR OF MEASUREMENT AND THE SENSITIVITY OF A TEST OF SIGNIFICANCE

J. P. SUTCLIFFE*

UNIVERSITY OF SYDNEY

Implications of random error of measurement for the sensitivity of the F test of differences between means are elaborated. By considering the mathematical models appropriate to design situations involving true and fallible measures, it is shown how measurement error decreases the sensitivity of a test of significance. A method of reducing such loss of sensitivity is described and recommended for general practice.

In the statistical theory of sampling, explicit attention is given to sampling error, which refers to fluctuations in the composition of samples drawn at random from a defined universe. A second form of error, largely ignored in this context, is measurement error. This applies to the individual sampling units and is thus related to the definition of the universe rather than sampling outcomes. Applications of sampling theory have proceeded on the implicit assumption that the sampling units which make up the defined universe are error free, that (in psychometric terms) the universe consists of true scores. This assumption is not justified in practice, where measurement is seldom free from error. Parameters, such as the mean and the variance, of a universe of fallible scores will differ from those of a universe of true scores; tests of significance of a given effect will not necessarily be the same in the two cases. This paper elaborates the implications of measurement error for the simple case of the F test of difference between means. By setting up the mathematical models appropriate to the relevant design situations, it is shown how measurement error (relative to the parallel true score case) decreases the sensitivity of the test of significance. Sensitivity refers to the likelihood of detecting a nonzero population effect at a given level of significance. Through its inverse, proneness to Type II error, it is usually expressed quantitatively as power. A method of reducing such loss of sensitivity is described.

Definition of Universes of Scores

The scale or range of application of a measuring instrument comprises a number of units of measurement. Let \( w \) represent any one unit or subrange of the scale and \( v \) any one occasion of measurement. Errors of measurement

*I wish to express my thanks in acknowledgement that the present form of this paper has benefited from editorial comment, and from the advice of Dr. H. Mulhall of the Department of Mathematics, University of Sydney.
constant for all units of the scale on all occasions of testing will be designated \( f \); errors constant for all occasions of measurement with a particular unit, but variable from unit to unit will be designated \( g_w \); errors variable from occasion to occasion and from unit to unit will be designated \( h_{uw} \). For example, a carpenter's tape may be incorrectly calibrated uniformly over the whole scale; then unevenly stretched over the first few feet which are most commonly used; and finally subject to random error on any given application. For this case the total error of measurement \( E = f + g_w + h_{uw} \). Analogous errors of measurement occur with psychological tests [3], but these will not be discussed here; while knowledge of the source of error can facilitate its control, it is rather the mode of operation of error which is relevant to the statistical argument.

Most generally, an obtained fallible measure or score, \( X \), can be expressed as the sum of the true score, \( T \), and its error of measurement, \( E \), [3]. This holds whether measurement error is unitary, or complex in the sense illustrated above. The additive relationship also holds whatever other relationship may be shown to obtain between true score and error for a universe of obtained scores. For instance, while \( E' \) may enter as a multiplier in the relationship between obtained and true score, \( X_s = E'T_s \), \( X \), may also be written \( X_s = T_s + E_s \), where \( E_s = (E' - 1)T_s \). Other assumptions about the nature of error and its relationship to true score are tenable, but the additive assumption is adopted here because it simplifies the subsequent analysis.

The mean and variance of an infinite universe of fallible scores \( X_s = T_s + E_s \) may be obtained as follows:

\[
\text{Mean} = \lim_{N \to \infty} \left[ \frac{1}{N} \sum X_s / N \right] = \lim_{N \to \infty} \left[ \frac{1}{N} \sum (T_s + E_s) / N \right] = \bar{T} + \bar{E}.
\]

\[
\text{Variance} = \lim_{N \to \infty} \left[ \frac{1}{N} \sum X_s^2 / N \right] = \lim_{N \to \infty} \left[ \frac{1}{N} \sum (T_s + E_s)^2 / N \right] = \sigma_t^2 + \sigma_e^2 + 2 \rho_{te} \sigma_t \sigma_e.
\]

These outcomes are summarized in Table 1. Depending upon the mode of operation of error, cases may arise where any or all of \( \bar{E} \), \( \sigma_e^2 \), and \( \rho_{te} \) are zero,

<table>
<thead>
<tr>
<th>Universe</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>True scores ( T )</td>
<td>( \bar{T} )</td>
<td>( \sigma_t^2 )</td>
</tr>
<tr>
<td>Error scores ( E )</td>
<td>( \bar{E} )</td>
<td>( \sigma_e^2 )</td>
</tr>
<tr>
<td>Obtained scores ( X )</td>
<td>( \bar{T} + \bar{E} )</td>
<td>( \sigma_t^2 + \sigma_e^2 + 2 \rho_{te} \sigma_t \sigma_e )</td>
</tr>
</tbody>
</table>