PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

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_{vw} . 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 $E = f + g_w + h_{vw}$. 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_v , can be expressed as the sum of the true score, T_v , and its error of measurement, E_v [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'_v may enter as a multiplier in the relationship between obtained and true score, $X_v = E'_v T_v$, X_v may also be written $X_v = T_v + E_v$, where $E_v = (E'_v - 1)T_v$. 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 $X_v = T_v + E_v$ may be obtained as follows:

$$\text{Mean} = \lim_{N\to\infty} [\sum^{N} X_v/N] = \lim_{N\to\infty} [\sum^{N} (T_v + E_v)/N] = \bar{T} + \bar{E}.$$

$$\text{Variance} = \lim_{N\to\infty} [\sum^{N} x_v^2/N] = \lim_{N\to\infty} [\sum^{N} (t_v + e_v)^2/N]$$

$$= \sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e \, .$$

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

TABLE 1

Parameters of Universes of True, Error and Obtained Scores

Universe	Mean	Variance
True scores T_v	$\bar{T}$	σ_t^2
Error scores E_v	$\bar{E}$	σ_e^2
Obtained scores X_v	$\bar{T} + \bar{E}$	$\sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$

in which cases one or more of the parameters will be common to the universes of true and obtained scores.

When error is absent, the mean = $\bar{T}$ and variance = σ_t^2 (Case 1). When error is constant $\bar{E} = f > 0$, $\sigma_e^2 = 0$, $\rho_{te} = 0$; hence the mean of fallible scores = $\bar{T} + f$, and variance = σ_t^2 (Case 2). Where error is variable its distribution may be either random or nonrandom. (In either case, the variances of error about different true score values may be homogeneous or heterogeneous. Heterogeneity of variance permits nonzero correlation between true scores and the *variance* of errors about them, but, as in random sampling, this correlation is independent of ρ_{te}. Heterogeneity of error variance should, of course, be taken into account in any analysis of variance [2].) If errors occur at random about T_v, then $\bar{E} = 0$, $\sigma_e^2 > 0$, and $\rho_{te} = 0$; hence mean = $\bar{T}$, and variance = $\sigma_t^2 + \sigma_e^2$ (Case 3). If errors are randomly distributed about $T_v + f$, $\bar{E} = f + 0$, $\sigma_e^2 > 0$, $\rho_{te} = 0$; hence mean = $\bar{T} + f$, and variance = $\sigma_t^2 + \sigma_e^2$ (Case 4). Where errors are distributed randomly about $T_v + g_w$, then $\bar{E} = \bar{g} + 0$, $\sigma_e^2 > 0$, $\rho_{te} > 0$, and hence mean = $\bar{T} + \bar{g}$, and variance = $\sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$ (Case 5). With nonrandom distribution of errors, generally one would find $\bar{E} > 0$, $\sigma_e^2 > 0$, and $\rho_{te} > 0$. Whether errors are distributed about T_v, $T_v + f$, or $T_v + g_w$, mean = $\bar{T}$ + error, and variance = $\sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$. All cases of nonrandom distribution of error arè here referred to as Case 6.

The six cases are summarized in Table 2 to enable comparison of the parameters of fallible score universes with those of the true score universe. In no case are both parameters the same as those in Case 1; however, Case 2 has the same variance, and Case 3 the same mean. Cases 1 and 2 are unlikely to occur in practice. Most experiments aim to achieve the conditions of Case 3, but the intrusion of constant errors, scale biases, and other nonrandom errors makes Cases 4, 5, and 6 quite common. The following discussion will center on Cases 1 and 3, with incidental comment on the others.

TABLE 2

Parameters of Universes of True and Fallible Scores

Case	Mean	Variance
1	$\bar{T}$	σ_t^2
2	$\bar{T} + f$	σ_t^2
3	$\bar{T}$	$\sigma_t^2 + \sigma_e^2$
4	$\bar{T} + f$	$\sigma_t^2 + \sigma_e^2$
5	$\bar{T} + \bar{g}$	$\sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$
6	$\bar{T}$ + error	$\sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$

Comparison of the Design Models

With the universes of true and fallible scores defined, it becomes possible to compare the sensitivity of tests of significance applied in given cases. For comparative purposes the analysis of variance for Case 1 will be described. Then two analyses for Case 3 will be considered—the first reflecting common practice, the second involving random replication of measurement to increase reliability and hence sensitivity.

Notation and plan for Case 1

Consider the comparison of means of independent random samples of true scores obtained at different levels of a single-treatment classification. Let $i = 1, 2, \cdots, a$ represent any one of the treatment levels within the treatment classification A. Let $j = 1, 2, \cdots, b$ represent any one subject in a sample of subjects B. Then X_{ij} is the true score of the subject j in the treatment level or group i. As subjects are randomly sampled, j represents number only, not rank within a group. Let a dot in place of a subscript represent summation across the class indicated by the subscript replaced, e.g.,

$$\sum_{j=1}^{b} X_{ij} = X_{i.}\,, \qquad \sum_{i=1}^{a} \sum_{j=1}^{b} X_{ij} = X_{..}\,.$$

The sample values of X_{ij} and the sums are represented in Table 3.

TABLE 3

Plan of Obtained Scores of Subjects Within Random Samples Allocated to Independent Treatment Groups

A				B Subjects			
Treatments	1	2	.	j	.	b	Sum
1	X_{11}	X_{12}	.	X_{1j}	.	X_{1b}	$X_{1.}$
2	X_{21}	X_{22}	.	X_{2j}	.	X_{2b}	$X_{2.}$
.	.	.	.	.	.	.	.
i	X_{i1}	X_{i2}	.	X_{ij}	.	X_{ib}	$X_{i.}$
.	.	.	. .	.	.	.	.
a	X_{a1}	X_{a2}	.	X_{aj}	.	X_{ab}	$X_{a.}$
							$X_{..}$

Analysis of variance for Case 1

The total variance of the ab sample values of X_{ij} can be expressed in terms of two sources of variation: between treatments, A, and between subjects within treatment levels, B_A. A given deviation score may be written as

$$x_{ij} = (X_{ij} - \bar{X}_{..}) = (\bar{X}_{i.} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.}).$$

The total sum of squares is

$$\mathrm{SS}_T = \sum_{i=1}^{a} \sum_{j=1}^{b} (X_{ij} - \bar{X}_{..})^2 = b \sum_{i=1}^{a} (\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} (X_{ij} - \bar{X}_{i.})^2.$$

The degrees of freedom pertaining to these components are Total $= (ab - 1)$, $A = (a - 1)$, $B_A = a(b - 1)$. From the SS and df, the mean squares, S^2, may be obtained as unbiased estimates (on the null hypothesis) of a common population variance.

Expectation of mean squares for Case 1

To determine what is estimated by a given S^2, one takes the expectation according to the model involved. As Case 1 involves a universe of true scores, Model 1 can be written as

$$X_{ij} = A_i + B_{ij} \; .$$

A_i is the class of treatment parameters of which the sampled treatment means are estimators. The distribution of A_i will vary according as treatments are fixed constants or randomly sampled. For convenience the case of random A_i with variance σ_A^2 will be considered. B_{ij} is the class of true score deviations from A_i, which are normally distributed with zero mean and variance σ_t^2. To find the expected values of SS and then S^2, one substitutes model values in the analysis of sample variance and thereby determines the limiting value of a given component.

(*i*) *Expectation of* S_A^2

$$(\bar{X}_{i.} - \bar{X}_{..}) = (A_i - \bar{A}_.) + (\bar{B}_{i.} - \bar{B}_{..});$$

$$E\left\{b \sum_{i=1}^{a} (\bar{X}_{i.} - \bar{X}_{..})^2\right\} = E\left\{b \sum_{i=1}^{a} (A_i - \bar{A}_.)^2 + b \sum_{i=1}^{a} (\bar{B}_{i.} - \bar{B}_{..})^2\right\}$$

$$= b(a - 1)\sigma_A^2 + b(a - 1)\sigma_t^2/b.$$

Thus $S_A^2 = b \sum_{i=1}^{a} (\bar{X}_{i.} - \bar{X}_{..})^2/(a - 1) \rightarrow b\sigma_A^2 + \sigma_t^2$.

(*ii*) *Expectation of* S_{BA}^2

$$(X_{ij} - \bar{X}_{i.}) = (B_{ij} - \bar{B}_{i.}),$$

and

$$S_{BA}^2 = \sum^{a} \sum^{b} (X_{ij} - \bar{X}_{i.})^2/a(b - 1) \rightarrow \sigma_t^2 \, .$$

TABLE 4

Analysis of Variance for Model 1:
Single Treatment Classification Design with
b Randomly Sampled Subjects for Each of *a* Levels (True Scores)

Number	Source	Sum of Squares	df	S^2	Expectation of S^2
1	A	$b \sum^{a} (\bar{X}_{i.} - \bar{X}_{..})^2$	$(a-1)$	S_A^2	$b\sigma_A^2 + \sigma_t^2$
2	B within A	$\sum^{a} \sum^{b} (X_{ij} - \bar{X}_{i.})^2$	$a(b-1)$	$S_{B_A}^2$	σ_t^2
3	Total	$\sum^{a} \sum^{b} (X_{ij} - \bar{X}_{..})^2$	$(ab-1)$		

These outcomes for the analysis of variance are summarized in Table 4. On the null hypothesis $\sigma_A^2 = 0$. One rejects the null hypothesis if the ratio $F_1 = S_A^2/S_{B_A}^2$ with $\text{df}_1 = (a - 1)$ and $\text{df}_2 = a(b - 1)$ exceeds F_α, the tabled value for the chosen level of significance.

Case 3

It is common practice in psychological experimentation to use a design superficially similar to the one just described. That is, one has a series of random samples of subjects allocated to treatment levels and for each subject one has a single score. If, as is usually the case, the scores are fallible, then Model 1 is inapplicable and instead one must write the model to include error of measurement. Assuming that the scores have been drawn from a Case 3 universe, there will be two designs according as one has or has not random replication of measurement on a given subject. For common practice, which provides no measurement replication, Model 3*a* is

$$X_{ij} = A_i + B_{ij} + \Gamma_{ij} .$$

A_i and B_{ij} have been defined above; Γ_{ij} is the random error of measurement component, normally distributed with zero mean and variance σ_e^2. The summary of the analysis of variance for Model 3*a* is given in Table 5. For the test of significance, the null hypothesis is $\sigma_A^2 = 0$. One rejects the null hypothesis if the ratio $F_{3a} = S_A^2/S_{B_A}^2$ with $\text{df}_1 = (a - 1)$ and $\text{df}_2 = a(b - 1)$ exceeds the tabled value of F for the chosen level of significance..

One may note that the terms σ_A^2 and σ_t^2 are common to the expectations of S_A^2 for Models 1 and 3*a*. In addition, the df_1 and df_2 are the same for F_1 and F_{3a}. This enables comparison of the sensitivity of the two tests. The power of the F_1 test is Prob $\{F_1 > F_\alpha \sigma_t^2/(b\sigma_A^2 + \sigma_t^2)\}$; and the power of F_{3a} is Prob $\{F_{3a} > F_\alpha(\sigma_t^2 + \sigma_e^2)/(b\sigma_A^2 + \sigma_t^2 + \sigma_e^2)\}$. The smaller the value to the right of $>$, the greater the power of the test. As $\sigma_t^2/(b\sigma_A^2 + \sigma_t^2) < (\sigma_t^2 + \sigma_e^2)/$

TABLE 5

Analysis of Variance for Model 3a:
Single Treatment Classification Design with
b Randomly Sampled Subjects for Each of a Levels (Fallible Scores)

Number	Source	Sum of Squares	df	S^2	Expectation of S^2
1	A	$b\sum^{a}(\bar{X}_{i.} - \bar{X}_{..})^2$	$(a-1)$	S_A^2	$b\sigma_A^2 + \sigma_t^2 + \sigma_e^2$
2	B within A	$\sum^{a}\sum^{b}(X_{ij} - \bar{X}_{i.})^2$	$a(b-1)$	$S_{B_A}^2$	$\sigma_t^2 + \sigma_e^2$
3	Total	$\sum^{a}\sum^{b}(X_{ij} - \bar{X}_{..})^2$	$(ab-1)$		

$(b\sigma_A^2 + \sigma_t^2 + \sigma_e^2)$, the power of F_1 is greater than the power of F_{3a}. That is, analysis in accordance with Model 3*a* provides a less sensitive test of the hypothesis $\sigma_A^2 > 0$ than does Model 1; the loss of sensitivity is due to the intrusion of random error of measurement.

Model 3*a* allows for the acknowledgement of the presence of error variance, but there is no provision for its isolation. To achieve this, one has to add random replication of measurement for each subject. That is, instead of a single score for each subject one has a number of scores. This introduces a source of variation in addition to those already accounted for; accordingly the notation and plan presented above have to be expanded. Let $k = 1, 2, \cdots, c$ represent any one measure or score in a sample of scores C. Then X_{ijk} is the kth score of subject j at treatment level i. As measures on subjects are randomly sampled, k represents number only, not rank. Now Model 3*b* may be written as

$$X_{ijk} = A_i + B_{ij} + \Gamma_{ijk} .$$

A_i and B_{ij} have been defined above; and Γ_{ijk} is defined as was Γ_{ij} . That is, Model 3*a* is the special case of Model 3*b* in which $k = 1$. The summary of the analysis of variance for Model 3*b* is given in Table 6. This analysis provides two tests of significance.

For the first, the null hypothesis is $\sigma_t^2 = 0$. One rejects the null hypothesis if the ratio $F_{3b} = S_{B_A}^2/S_{C_B}^2$ with $\text{df}_1 = a(b - 1)$ and $\text{df}_2 = ab(c - 1)$ exceeds the tabled value of F for the chosen level of significance. If the null hypothesis is not rejected, the outcome is consistent with the homogeneity of experimental subjects, and in that sense one has zero reliability of measurement. If the null hypothesis is rejected, an estimate of the reliability of measurement may be obtained. With the Case 3 universe, the population value of the reliability coefficient [1] is $\rho_{xx} = \sigma_t^2/(\sigma_t^2 + \sigma_e^2)$, which may be estimated by

TABLE 6

Analysis of Variance for Model 3b:
Single Treatment Classification Design with
c Random Measures on each of
b Randomly Sampled Subjects for each of
a Levels (Fallible Scores)

Number	Source	Sum of Squares	df	S^2	Expectation of S^2
1	A	$bc \sum^{a} (\bar{X}_{i..} - \bar{X}_{...})^2$	$(a-1)$	S_A^2	$bc\sigma_A^2 + c\sigma_t^2 + \sigma_e^2$
2	B within A	$c \sum^{a} \sum^{b} (\bar{X}_{ij.} - \bar{X}_{i..})^2$	$a(b-1)$	$S_{B_A}^2$	$c\sigma_t^2 + \sigma_e^2$
3	C within B	$\sum^{a} \sum^{b} \sum^{c} (X_{ijk} - \bar{X}_{ij.})^2$	$ab(c-1)$	$S_{C_B}^2$	σ_e^2
4	Total	$\sum^{a} \sum^{b} \sum^{c} (X_{ijk} - \bar{X}_{...})^2$	$(abc-1)$		

$$r_{xx} = (S_{BA}^2 - S_{CB}^2)/[S_{BA}^2 - S_{CB}^2(1 - c)].$$

For the second, the null hypothesis is $\sigma_A^2 = 0$. One rejects the null hypothesis if the ratio $F'_{3b} = S_A^2/S_{BA}^2$ with $\mathrm{df}_1 = (a - 1)$ and $\mathrm{df}_2 = a(b - 1)$ exceeds the tabled value of F for the chosen level of significance.

Comparison of the power of the F'_{3b} test

$$\text{Prob } \{F'_{3b} > F_\alpha(c\sigma_t^2 + \sigma_e^2)/(bc\sigma_A^2 + c\sigma_t^2 + \sigma_e^2)\}$$

with the powers of F_1 and F_{3a} shows that as

$$\frac{\sigma_t^2}{b\sigma_A^2 + \sigma_t^2} < \frac{c\sigma_t^2 + \sigma_e^2}{bc\sigma_A^2 + c\sigma_t^2 + \sigma_e^2} < \frac{\sigma_t^2 + \sigma_e^2}{b\sigma_A^2 + \sigma_t^2 + \sigma_e^2}$$

then power $F_1 >$ power $F'_{3b} >$ power F_{3a}.

While analysis by the Model $3b$ allows for isolation of an estimate of σ_e^2 , it is important to note that one may *not* convert F'_{3b} to F_1 by subtracting $S_{CB}^2 \rightarrow \sigma_e^2$ from the numerator and denominator of F'_{3b} and making appropriate adjustments for the weights b and c. F is the ratio of two independent χ^2 variates—the independence is negated by such a procedure. The only way to achieve the standard of sensitivity of the F_1 test with the given number of subjects is to use error-free measurement. As this is an ideal towards which one can do no more than strive, one has to be satisfied with a less sensitive test. Of the two remaining experimental designs, assuming that one can achieve measurement replication, that which provides the $3b$ form of analysis is to be recommended for general practice. It yields estimates of measurement error variance and reliability, for the latter a test of significance, as well as providing a more sensitive test of treatment effects than the 3*a* design using the same number of subjects. These contentions apply with equal force to the design situations where the *t* test is ordinarily applied. Finally, while the argument has been in terms of the single treatment classification design, it may be generalized to multiple classification designs.

REFERENCES

[1] Alexander, H. W. The estimation of reliability when several trials are available. *Psychometrika,* 1947, 12, 79-99.

[2] Ehrenberg, A. S. C. The unbiased estimation of heterogeneous error variances. *Biometrika,* 1950, **37**, 347-357.

[3] Walker, Helen M. and Lev, J. *Statistical inference.* New York: Holt, 1953.

Manuscript received 1/14/57
Revised manuscript received 4/30/57