TECHNOMETRICS© VOL. 15, NO. 1 FEBRUARY, 1973

Errors of Measurement, Precision, Accuracy and the Statistical Comparison of Measuring Instruments

FRANK E. GRUBBS

U.S. Army Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland

A very important and yet widely misunderstood concept or problem in science and technology is that of precision and accuracy of measurement. It is therefore necessary to define the terms precision and accuracy (or imprecision and inaccuracy) clearly and analytically if possible. Also, we need to establish and develop appropriate statistical tests of significance for these measures, since generally a relatively small number of measurements will be made or taken in most investigations.

In this paper a discussion is given of some of the pertinent literature for estimating variances in errors of measurement, or the "imprecisions" of measurement, when two or three instruments are used to take the same observations on a series of items or characteristics. Also, present techniques for comparing the imprecision of measurement of one instrument with that of a second instrument through the use of statistical tests of significance are reviewed, as well as procedures for detecting the significance of the difference in biases or levels of measurement of two instruments. Finally, we indicate methods of extending present theory to the case of three measuring instruments, for which rather sensitive statistical test of significance are developed for dealing with the precision and accuracy problem.

An example for the three instrument case is given to illustrate the suggested methodology of analysis.

KEY WORDS

Statistical Comparison of Measuring Instruments
Errors of Measurement
Precision
Accuracy

I. INTRODUCTION

Each measurement made by an instrument or measuring device consists of the true, unknown level of the characteristic or item measured plus an error of measurement. In practice it is important to know whether or not the variance in errors of measurement of an instrument, or the imprecision of measurement, is suitably small as compared to the variance of the characteristic or product measured, or the product variability. Indeed, for efficiency of the measuring process, the variance in errors of measurement should be several or many times smaller than the variability of the characteristic measured or product variance. Methods of estimating true product variability and true variances in errors of measurement were studied by the author (1948). In this connection, it was shown that two or more measuring instruments were required to separate and hence estimate true variances in instrumental errors on one hand and the product variance on the other. When three or

Received July 1971; revised March 1972

more instruments are used for simultaneous or common readings on a series of items then estimates of precision of measurement and their variances can be obtained which are entirely free of product or item variability. Tests of statistical significance, however, were not easily developed then because of the dependence or high degree of correlation between the measurements of different instruments. Of course, the Pitman–Morgan (1939) test could have been used for the two-instrument case, although such a test seemed to be rather insensitive to detecting differences in variances of errors of measurements of the instruments, except for suitably large sample sizes, since masking due to product variability is present. The recent results of Maloney and Rastogi (1970) for the two-instrument case and the particular model of Hahn and Nelson (1970), therefore, provide some motivation for studying significance tests for the case of three different measuring devices.

By using two instruments to measure a series of items, and always taking a second reading with one of the instruments, Hahn and Nelson (1970) were able to develop tests of significance for comparing (1) variances in errors of measurement, and (2) levels of measurement for the two instruments by using an appropriate Student t-test. Such tests of significance, nevertheless, depend on the assumptions of independence of variates used, which is correct in the Hahn and Nelson two-instrument, three measurements model because the differences in duplicate readings of the second instrument, and the differences in readings of the first instrument and the average of the duplicate readings of the second instrument, are independent for normally distributed variates. As indicated below, the validity of statistical tests of significance for imprecision and level may easily be extended to three different measuring devices. Also, if from past data the relative sizes of variances in errors of measurement are known for two of the three instruments, then the precision of measurement of the third or "test" instrument may be judged in a significance test, although the computations are somewhat more involved. In case the variances of errors of measurement for two of the instruments are the same, a very desirable practical situation, then the precision of measurement of the third instrument may be tested statistically using the results of Hahn and Nelson (1970). Even though the variances in errors of measurement of the two "standard" instruments may not be equal, and their ratio not known precisely, the Pitman–Morgan (1939) test for correlated variances may still be used to advantage to pick up differences in imprecision of measurement between the third or test instrument and the two standards as indicated in the sequel.

For the two-instrument (one measurement by each) case, the significance tests on imprecisions developed by Maloney and Rastogi (1970) are appropriate.

The significance tests of Thompson (1963) and Jaech (1971) may also be used to advantage for the two instrument case.

As will be seen, we have some special interest in the three instrument case, and in particular the simultaneous use of two "standard" instruments to check a third or "test" instrument for precision and accuracy.

2. Preliminary Statement of Model for Studying Relative Precision of Measurement, Instrumental Bias and Accuracy

We consider three instruments which are used to make simultaneous or the same measurements on each of a series of n items. Their measurements are indicated by the quantities

$$r_i = \beta_1 + x_i + e_{i1} \qquad s_i = \beta_2 + x_i + e_{i2} \qquad t_i = \beta_3 + x_i + e_{i3} \tag{1}$$

where $i = 1, 2, \cdots, n$; β_1, β_2 and β_3 are respectively the (unknown) instrumental biases; e_{i1}, e_{i2}, and e_{i3} the random errors of measurement of the three instruments, and x_i the true, unknown value of the ith item or characteristic measured in a series of measurements on n items. For the significance tests indicated below, the e_{ij} are assumed to be normally and independently distributed of each other and independent of the x_i over the range of consideration.

We note that the expected variance in readings or measurements of the first instrument, for example, is seen to be

$$\sigma_r^2 = \sigma_x^2 + \sigma_{e1}^2$$

so that the variance in errors of measurement σ_{e1}^2—or the "imprecision" of measurement—should be suitably small as compared to the product variability or variance σ_x^2 for good precision of measurement. Moreover, for good accuracy, then it is easy to see that both the imprecision of measurement σ_{e1}^2 and the instrumental bias β_1 should be small or approach zero.

For the common or simultaneous measurements made by the three instruments on items $i = 1, 2, \cdots n$ the following three columns of differences:

$$\begin{aligned} v_{1i} &= r_i - s_i = \beta_1 - \beta_2 + e_{i1} - e_{i2} \\ w_{1i} &= s_i - t_i = \beta_2 - \beta_3 + e_{i2} - e_{i3} \\ z_{1i} &= t_i - r_i = \beta_3 - \beta_1 + e_{i3} - e_{i1} \end{aligned} \tag{2}$$

are very informative as they involve only differences in instrumental biases and errors of measurements, with the product level, x_i, eliminated. Clearly, these quantities give basic insight for studying precision and accuracy.

Instruments 1 and 2 may be considered preliminarily as "standards", "acceptable" instruments, or instruments of known precision, and our interest centers around whether the "test" instrument 3 may be precise enough, i.e. possesses sufficiently small variance in errors of measurement, σ_{e3}^2, and whether instrument 3 reads at the correct level or has some bias as compared to the two standards, or needs calibration. Methods are given by Grubbs (1948) for estimating the true product variance σ_x^2 and the true variances in errors of measurement for the three instruments, or the "imprecisions", σ_{e1}^2, σ_{e2}^2 and σ_{e3}^2. These variance estimates which unfortunately sometimes turn out to be negative, nevertheless, give an initial idea of the relative sizes of σ_{e1}^2, σ_{e2}^2 and σ_{e3}^2 preliminary to a test of significance, or for other comparative purposes.

Instruments 1 and 2 also may possibly as a result of previous tests be regarded as "reference" instruments, since some appropriate information may have been gathered on their precision of measurement characteristics and they might have been properly calibrated with respect to level. We would thus be particularly interested in comparing the imprecision of the third instrument, or σ_{e3}^2, with the average imprecision of the two standards, i.e. $(\sigma_{e1}^2 + \sigma_{e2}^2)/2$, and a statistical test of significance to do this would of course be of considerable utility. For the two standard instruments alone, we would be interested in whether σ_{e1} and σ_{e2} are equal or not. The estimated variances (or standard deviations) in errors of measurement, moreover, give us needed measures of imprecision, and hence relative precision of measurement of the instruments.

Accuracy of measurement depends on both the precision (or imprecision) of the instrument and also on its level of measurement, i.e. its bias, or just how well the instrument has been "calibrated." For the two "standard" instruments we are interested in whether $\beta_1 - \beta_2$ is zero, positive or negative (i.e. its size), and hence whether recalibration is called for. Of course, in order to detect the difference between β_1 and β_2 the variances in errors of measurement will have to be suitably small, or the sample size suitably large, so that a statistical study of accuracy depends on both the imprecision of measurement and the levels of measurement or the biases of the instruments. For the third or "test" instrument, our interest in accuracy centers around whether $\beta_3 - (\beta_1 + \beta_2)/2$ is zero, positive or negative (i.e. its size), and to make such a comparison the imprecisions of measurement σ_{e1}^2, σ_{e2}^2 and σ_{e3}^2 necessarily come into play. Appropriate significance tests for the required comparisons are developed below.

With the above preliminary discussion of the basic model, we may now illustrate with actual data the application of the theory which is developed in some detail in Sections IV and V below.

3. Example*

The problem considered here arose from a NATO study on velocity chronographs submitted for acceptance or standardization. In this connection, it was apparently desirable to use two reference or "standard" chronographs, which would seem to be better than a single reference instrument, to judge a third chronograph submitted for test. Perhaps it was considered that such a procedure would result in more confidence and provide some checks on the test results. Of course, the choice of the two standards for initial tests is somewhat arbitrary indeed, although pair-wise comparisons of the three instruments can be made using the procedure below by permuting the instruments (i.e. the r_i, s_i and t_i) as desired.

Three velocity-measuring chronographs, the "Fotobalk", the "Counter" and the "Terma" instruments, are used simultaneously to determine velocities of each of twelve successive rounds fired from a 155mm Gun. The velocities are recorded in meters per second (m/s), and the individual velocity measurements are given on Table I. Recorded also on Table I are the sample variances, sample covariances, sample correlation coefficients and other quantities which are useful in significance tests or are of some interest otherwise. We assume here that no past data are available on precision of measurement for the "standard" instruments, the Fotobalk and the Counter, and our purpose here is to check out the precision and accuracy of measurement for the Terma or "test" instrument. Equation (9) of the first Reference is used to estimate the standard deviations in errors of measurement for each of the three instruments, the computations being shown on Table II. It is seen that the estimated standard error of measurement for the Terma chronograph (.468 m/s from Table II) seems larger than that of the other two chronographs. We check this value later, however, after first checking out the two "standards", the Fotobalk and Counter, for relative precision and agreement in level, or for bias.

First, we find the sums, $y_i = r_i + s_i$, and differences $v_{1i} = r_i - s_i$ of the velocities for the Fotobalk and Counter instruments and compute $S^2(y) = 7.508$, $S^2(v_1) = .0590$, $S(yv_1) = .1748$, so that $r(yv_1) = .2626$, and from (28), Section V

$$t(n-2, \sigma_{e1} = \sigma_{e2}) = r(yv_1)\sqrt{n-2}/[1 - r^2(yv_1)]^{\frac{1}{2}} = .861$$

for Student's t to compare σ_{e2} and σ_{e1}, whereas $t_{.975}(10) = 2.228$ and $t_{.95}(10) = 1.812$. We therefore conclude that the Fotobalk and Counter chronographs have equal precision of measurement even though for 12 rounds $\hat{\sigma}_{e1} = .026$ m/s and $\hat{\sigma}_{e2} =$

* The author is indebted to Mr. James F. O'Bryon of the U. S. Army Ballistic Research Laboratories for the data of this example, and illuminating discussions which stimulated much of this study.

TABLE I

Simultaneous Velocities of the Fotobalk, Counter and Terma Chronographs on Each of Twelve Successive Rounds in Meters Per Second (m/s)

Round No.	Foto r	Counter s	Terma t	$(r+s)-1580=y$	$r-s=y_1$	$t-(r+s)/2=u_1$
20	793.8	794.6	793.2	8.4	-.8	-1.00
21	793.1	793.9	793.3	7.0	-.8	- .20
22	792.4	793.2	792.6	5.6	-.8	- .20
23	794.0	794.0	793.8	8.0	.0	- .20
24	791.4	792.2	791.6	3.6	-.8	- .20
25	792.4	793.1	791.6	5.5	-.7	-1.15
26	791.7	792.4	791.6	4.1	-.7	- .45
27	792.3	792.8	792.4	5.1	-.5	- .15
28	789.6	790.2	788.5	-0.2	-.6	-1.40
29	794.4	795.0	794.7	9.4	-.6	.00
30	790.9	791.6	791.3	2.5	-.7	+ .05
31	793.5	793.8	793.5	7.3	-.3	- .15

$S^2(y) = [n\Sigma y_i^2 - (\Sigma y_i)^2]/n(n-1) = [12(488.89) - (66.3)^2]/132 = 7.508$

$S^2(v_1)\ [12(5.09) - (-7.3)^2]/132 = .0590$

$S^2(u_1) = [12(4.6925) - (-5.05)^2]/132 = .2334$

$S(yv_1) = [n\Sigma y_i v_{1i} - (y_i)\ (\Sigma v_{1i})]/n(n-1) = [12(-38.41)-(66.3)(-7.3)]/132 = .1748$

$S(u_1v_1) = [12(3.325) - (5.05)(7.3)]/132 = .0230$

$r(yv_1) = (.1748)/\sqrt{(7.508)(.0590)} = .2626$

$r(u_1v_1) = (.0230)/\sqrt{(.2334)(.0590)} = .1959$

Mean $(r-s) = \bar{v}_1 = -.608$ m/s $\quad \bar{u}_1 = -.421$ m/s

.229 m/s from Table II. This computation involves the data from only the two "standard" instruments. Since, however, we have data from three measuring devices, an alternative and generally more sensitive test to compare σ_{e1} and σ_{e2} is found from formula (29) below and is

$$t(n-2, \sigma_{e1} = \sigma_{e2}) = \frac{[S^2(w_1)/S^2(z_1) - 1]\sqrt{n-2}}{[4(1 - r^2\{w_1z_1\})S^2(w_1)/S^2(z_1)]^{\frac{1}{2}}}$$

$$= \frac{[(.2711)/(.2252) - 1]\sqrt{10}}{[4(1 - \{.8847\}^2)(.2711)/(.2252)]^{\frac{1}{2}}} = .63$$

which, however, still does not reject the hypothesis that $\sigma_{e1} = \sigma_{e2}$. We conclude therefore on the basis of the twelve rounds that the Foto and Counter may have equal precision of measurement.

Next we check the agreement in levels of measurement for the Fotobalk and Counter. This is accomplished by computing Student's t from Formula (31) below, i.e.

$$t_0(n-1, \beta_1 = \beta_2) = \bar{v}_1\sqrt{n}/S(v_1) = -.608\sqrt{12}/(.2429) = -8.67$$

which for $n - 1 = 11$ d.f. is very highly significant. Thus, we would ordinarily look for the cause of this disagreement, run a retest of the two "standards" or calibrate them since the Fotobalk reads .61 m/s lower than the Counter. In this case, however, the sample variance of the differences in errors of measurement is very small, i.e. $S^2_{e1-e2} = .0590$ (or $S_{e1-e2} = .2429$) and our t-test is sensitive enough to easily pick up a difference of 0.61 m/s in velocity levels. Actually, however, without making corrections or calibrating, we may proceed to check out the Terma Chronograph, as its imprecision and level will be compared with the average values of the Fotobalk and Counter anyway, thereby not really making any difference.

To ascertain whether the variance in errors of measurement of the Terma chronograph is equal to that of the average of the Fotobalk and Counter instruments, we use formula (32) of Section V, i.e.

$$t_0[n - 2, \sigma^2_{e3} = (\sigma^2_{e1} + \sigma^2_{e2})/2] = \frac{[S^2(u_1)/S^2(v_1) - .75]\sqrt{n-2}}{[3(1 - r^2\{u_1 v_1\})S^2(u_1)/S^2(v_1)]^{\frac{1}{2}}}$$

$$= \frac{[(.2334)/(.0590) - 0.75]\sqrt{10}}{[3(1 - \{1959\}^2)(.2334)/(.0590)]^{\frac{1}{2}}} = 3.00$$

We therefore conclude that the Terma chronograph is not as precise as the ("average" of the) Fotobalk and Counter instruments, since $t_{.95}(10) = 1.812$.

We note that the standard deviation in errors of measurement for the Terma chronograph is estimated as .468 m/s and this instrument is measuring an estimated standard deviation in true velocity of 1.42 m/s, so that it may be sufficiently precise for the measurements taken here, nevertheless. Hence, we may want to check on level of velocity of the Terma chronograph by using formula (33). Here, we find

$$t_0[n - 1, \beta_3 = (\beta_1 + \beta_2)/2] = \bar{u}_1\sqrt{n}/S(u_1) = -.421\sqrt{12}/(.483) = -3.02$$

and since $-t_{.95}(11) = -1.796$ we conclude that the Terma chronograph reads low by .421 m/s as compared to the average of the Fotobalk and Counter. (Note that the bias of .61 m/s between the two "standards" is even a bit larger.)

The variance in errors of measurement of the Terma or third chronograph may be estimated from

$$\text{Est } \sigma^2_{e3} = S^2(u_1) - S^2(v_1)/4 = .2334 - .0590/4 = .2187 \text{ m/s}$$

which agrees, of course, with the value of .2186 computed by the equivalent formula on Table II.

The standard deviation of the mean velocities listed in the fifth column of Table II [i.e. from the model $x_i + (\beta_1 + \beta_2 + \beta_3)/3 + (e_{i1} + e_{i2} + e_{i3})/3$] is found to be 1.43 m/s as compared to the estimated true value of 1.42 m/s. In summary, therefore, we conclude that the variance in errors of each measuring instrument is appreciably smaller than the (population) variance of the velocities of the rounds. Nevertheless, some calibration of instruments may be desirable or even needed here.

We remark as an extension of the above example that the results in Section IV on Theory below may be generalized to provide a significant ranking of any number of measuring instruments with regard to both precision and accuracy, which would amount to a very important and highly practical accomplishment indeed. It is highly desirable in this connection that the significance tests developed should point out the particular instruments which are relatively imprecise or inaccurate as attempted herein.

TABLE II

Estimates of Precision of Measurement Based on Three Simultaneous Velocity Measurements of the Fotobalk, Counter, and Terma Chronographs

Round No.	Foto r	Counter s	Terma t	Mean Velocity m/s	r-s $=v_1$	s-t $=w_1$	t-r $=z_1$
20	793.8	794.6	793.2	793.87	-0.8	+1.4	-0.6
21	793.1	793.9	793.3	793.43	-0.8	+0.6	+0.2
22	792.4	793.2	792.6	792.73	-0.8	+0.6	+0.2
23	794.0	794.0	793.8	793.93	0.0	+0.2	-0.2
24	791.4	792.2	791.6	791.73	-0.8	+0.6	+0.2
25	792.4	793.1	791.6	792.37	-0.7	+1.5	-0.8
26	791.7	792.4	791.6	791.90	-0.7	+0.8	-0.1
27	792.3	792.8	792.4	792.50	-0.5	+0.4	+0.1
28	789.6	790.2	788.5	789.43	-0.6	+1.7	-1.1
29	794.4	795.0	794.7	794.70	-0.6	+0.3	+0.3
30	790.9	791.6	791.3	791.27	-0.7	+0.3	+0.4
31	793.5	793.8	793.5	793.60	-0.3	+0.3	0.0

$S^2(v_1) = S^2_{e1-e2} = .0590$ $\qquad$ $\beta_1 - \beta_2 + \bar{e}_1 - \bar{e}_2 = -.608$

$S^2(w_1) = S^2_{e2-e3} = .2711$ $\qquad$ $\beta_2 - \beta_3 + \bar{e}_2 - \bar{e}_3 = .725$

$S^2(z_1) = S^2_{e3-e1} = .2252$ $\qquad$ $\beta_3 - \beta_1 + \bar{e}_3 - \bar{e}_1 = +.117$

Using equations (9) and (26) of the first reference:

$\text{est}(\sigma^2_{e1}) = .5(.0590 + .2252 - .2711) = .0065$ or est $\sigma_{e1} = .0255$ m/s (Foto)

$\text{est}(\sigma^2_{e2}) = .5(.0590 - .2252 + .2711) = .0525$ or est $\sigma_{e2} = .229$ m/s (Counter)

$\text{est}(\sigma^2_{e3}) = .5(-.0590 + .2252 + .2711) = .2186$ or est $\sigma_{e3} = .468$ m/s (Terma)

$\text{est}(\sigma_x) = 1.42$ m/s = Estimated standard deviation of the true velocities of the rounds.

$r(w_1 z_1) = .8847$ = Correlation coefficient between w_1 and z_1

4. THEORY

In order to develop the theory applied to the above example, we refer to the model summarized in Section II, particularly Equations (1) and (2), and proceed as follows.

It is well known that if the covariances of normally distributed quantities are zero then such quantities are independently distributed. Hence, we see that the covariance of $(\beta_1 - \beta_2 + e_{i1} - e_{i2})$ and the quantity $1_1(\beta_2 - \beta_3 + e_{i2} - e_{i3}) + 1_2(\beta_3 - \beta_1 + e_{i3} - e_{i1})$, where 1_1 and 1_2 are constants, is given by

$$-1_1\sigma^2_{e2} - 1_2\sigma^2_{e1} \tag{3}$$

and this is zero if and only if

$$-1_1/1_2 = \sigma^2_{e1}/\sigma^2_{e2} \qquad 1_2\ ,\ \sigma^2_{e2} \neq 0 \tag{4}$$

Thus, if $\sigma^2_{e1} = \sigma^2_{e2}$, then we may use weights of $-\frac{1}{2}$ and $+\frac{1}{2}$ as follows for the second and third pairs of differences.

$$\begin{aligned} u_{1i} &= -.5(\beta_2 - \beta_3 + e_{i2} - e_{i3}) + .5(\beta_3 - \beta_1 + e_{i3} - e_{i1}) \\ &= -.5(s_i - t_i) + .5(t_i - r_i) = t_i - (r_i + s_i)/2 \\ &= (\beta_3 + e_{i3}) - (\beta_1 + \beta_2 + e_{i1} + e_{i2})/2 \end{aligned} \tag{5}$$

and this quantity is therefore, for normally distributed variates, independent of

$$v_{1i} = \beta_1 - \beta_2 + e_{i1} - e_{i2} \tag{6}$$

which, since $S^2(u_1)$ and $S^2(v_1)$ are then independent variances, establishes (i) the validity of an F-test for the ratio of these two variances which can be used to compare σ_{e3}^2 with $(\sigma_{e1}^2 + \sigma_{e2}^2)/2 = \sigma_{e1}^2 = \sigma_{e2}^2$ as seen from Formula (34) below, and (ii) the usual t-test for the comparative level of measurement of the third instrument, as found equivalently by Hahn and Nelson (1970).

On the other hand, even though σ_{e1} and σ_{e2} are not equal, then the u_{1i} and v_{1i} of (5) and (6) may still be used to advantage. In this case, the population covariance of (5) and (6) is clearly

$$\sigma(u_1 v_1) = (\sigma_{e2}^2 - \sigma_{e1}^2)/2 \tag{7}$$

or the true correlation coefficient is

$$\rho(u_1 v_1) = (\sigma_{e2}^2 - \sigma_{e1}^2)/[(\sigma_{e1}^2 + \sigma_{e2}^2)(4\sigma_{e3}^2 + \sigma_{e1}^2 + \sigma_{e2}^2)]^{\frac{1}{2}} \tag{8}$$

which in many practical instances may be somewhat near zero. In any event, we may compute the sample variances and the covariance

$$S^2(u_1) = \sum_{i=1}^{n} (u_{1i} - \bar{u}_1)^2/(n - 1) = [n \sum u_{1i}^2 - (\sum u_{1i})^2]/n(n - 1) \tag{9}$$

$$S^2(v_1) = \sum_{i=1}^{n} (v_{1i} - \bar{v}_1)^2/(n - 1) = [n \sum v_{1i}^2 - (\sum v_{1i})^2]/n(n - 1) \tag{10}$$

$$\begin{aligned} S(u_1 v_1) &= \sum_{i=1}^{n} (u_{1i} - \bar{u}_1)(v_{1i} - \bar{v}_1)/(n - 1) \\ &= [n \sum u_{1i} v_{1i} - (\sum u_{1i})(\sum v_{1i})]/n(n - 1) \end{aligned} \tag{11}$$

and use the Pitman–Morgan (1939) t-test procedure for testing the equality of correlated variances. That is, in effect we test the hypothesis that $\sigma^2(u_1) = \lambda\sigma^2(v_1)$, which are correlated (dependent), by using

$$t(n - 2) = \frac{[S^2(u_1)/S^2(v_1) - \lambda]\sqrt{n - 2}}{[4(1 - r^2)\lambda\{S^2(u_1)/S^2(v_1)\}]^{\frac{1}{2}}} \tag{12}$$

where $t(n - 2)$ is Student's t with $n - 2$ degrees of freedom, the observed or sample correlation coefficient r is given by

$$r = S(u_1 v_1)/[S(u_1)S(v_1)] \tag{13}$$

and the ratio of population variances of u_1 and v_1 is

$$\lambda = [\sigma_{e3}^2 + (\sigma_{e1}^2 + \sigma_{e2}^2)/4]/(\sigma_{e1}^2 + \sigma_{e2}^2) = \sigma^2(u_1)/\sigma^2(v_1) \tag{14}$$

Thus, to test the null-hypothesis that $\sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2$, we use the value $\lambda = \frac{3}{4}$ in formula (12), and refer the computed Student t-statistic $t = t[n - 2, \sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2]$, based on the observed data to a selected significance level of Student's t distribution. Formula (12) in this form is thus seen to give a very sensitive test for comparing imprecisions, especially as compared to the usual Pitman–Morgan

t-test for correlated variances (see the two-instrument case, formula 28 below.) We therefore have a rather direct comparison of σ_{e3}^2 and $(\sigma_{e1}^2 + \sigma_{e2}^2)/2$.

To proceed further, we note that if σ_{e1}^2 and σ_{e2}^2 are accurately known, perhaps from past experience, then it is easy to see that the purposely chosen quantity

$$\begin{aligned} u_{2i} &= [-\sigma_{e1}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)](\beta_2 - \beta_3 + e_{i2} - e_{i3}) \\ &\quad + [\sigma_{e2}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)](\beta_3 - \beta_1 + e_{i3} - e_{i1}) \\ &= (\beta_3 + e_{i3}) - [\sigma_{e2}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)](\beta_1 + e_{i1}) - [\sigma_{e1}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)](\beta_2 + e_{i2}) \end{aligned} \tag{15}$$

is also statistically independent of $(\beta_1 - \beta_2 + e_{i1} - e_{i2})$, so that a valid F test can therefore still be developed for the three different instruments here. In this case of known σ_{e1} and σ_{e2}, we define and compute the following:

$$\begin{aligned} u_{2i} &= [-\sigma_{e1}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)](\beta_2 - \beta_3 + e_{i2} - e_{i3}) \\ &\quad + [\sigma_{e2}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)](\beta_3 - \beta_1 + e_{i3} - e_{i1}) \end{aligned} \tag{16}$$

$$v_{2i} = (\beta_1 - \beta_2 + e_{i1} - e_{i2}) = v_{1i} \tag{17}$$

$$S^2(u_2) = \sum_{i=1}^{n} (u_{2i} - \bar{u}_2)^2/(n - 1) \tag{18}$$

$$S^2(v_2) = \sum_{i=1}^{n} (v_{2i} - \bar{v}_2)^2/(n - 1) \tag{19}$$

$$S(u_2v_2) = \sum_{i=1}^{n} (u_{2i} - \bar{u}_2)(v_{2i} - \bar{v}_2)/(n - 1) \tag{20}$$

The expected value of $S^2(u_2)$ is given by

$$\sigma^2(u_2) = \sigma_{e3}^2 + \sigma_{e1}^2\sigma_{e2}^2/(\sigma_{e1}^2 + \sigma_{e2}^2) \tag{21}$$

With the above formulations, it is clear that

$$(n - 1)S^2(u_2)/[\sigma_{e3}^2 + \sigma_{e1}^2\sigma_{e2}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)] = \chi^2(n - 1), \tag{22}$$

or Chi-square with $(n - 1)$ d.f.

$$(n - 1)S^2(v_2)/(\sigma_{e1}^2 + \sigma_{e2}^2) = \chi^2(n - 1), \tag{23}$$

or Chi-square with $(n - 1)$ d.f. and therefore that

$$\frac{S^2(u_2)\cdot[\sigma_{e1}^2 + \sigma_{e2}^2]}{S^2(v_2)\cdot[\sigma_{e3}^2 + \sigma_{e1}^2\sigma_{e2}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)]} = F(n - 1, n - 1) \tag{24}$$

which is distributed as Snedecor's F with $(n - 1)$ and $(n - 1)$ degrees of freedom.

Thus, to test the hypothesis $\sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2$, we compute

$$F(n - 1, n - 1) = S^2(u_2)/S^2(v_2)[\tfrac{1}{2} + \sigma_{e1}^2\sigma_{e2}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)^2] \tag{25}$$

and compare this observed value of F with a desired probability level or percentage point of the F-distribution.

When $\sigma_{e1} = \sigma_{e2}$, and putting $u_2 = u$, $v_2 = v$, then it may be noted that (25) reduces to special case

$$F(n - 1, n - 1) = 4S^2(u)/3S^2(v) \tag{26}$$

which is the known result of Hahn and Nelson (1970). Of course, it is a rather desirable condition for applications to have $\sigma_{e1} = \sigma_{e2}$, or use standard instruments with equal precision of measurement, for independence is assured and the computations are somewhat easier.

Finally, it is clear that there are no problems of dependence involved in using t-tests to judge differences in levels of measurement, for then sample means and sample variances are independently distributed for normal variates, and the ordinary Student "t" tests apply therefore.

5. Summary of Useful Formulas and Results

A. *Two Instrument Case*

For the two-instrument case, we record here for reference the results of Maloney and Rastogi (1970). Here, we take the readings from the two instruments, i.e.

$$r_i = \beta_1 + x_i + e_{i1} \qquad s_i = \beta_2 + x_i + e_{i2}$$

and form their sums and differences:

$$y_i = r_i + s_i = \beta_1 + \beta_2 + 2x_i + e_{i1} + e_{i2}$$

$$v_i = r_i - s_i = \beta_1 - \beta_2 + e_{i1} - e_{i2}$$

We then compute the sample correlation coefficient of y and v

$$\begin{aligned} r(yv) &= S(yv)/S(y)S(v) \\ &= [S^2(r) - S^2(s)]/[\{S^2(r) + S^2(s) + 2S(rs)\}\{S^2(r) + S^2(s) - 2S(rs)\}]^{\frac{1}{2}} \end{aligned} \tag{27}$$

Finally, we use the Pitman–Morgan (1939) procedure to compute the quantity

$$\begin{aligned} t(n - 2, \sigma_{e1} = \sigma_{e2}) &= r(yv)\sqrt{n-2}/[1 - r^2(yv)]^{\frac{1}{2}} \\ &= \frac{[S^2(r)/S^2(s) - 1]\sqrt{n-2}}{\sqrt{4(1 - r^2\{rs\})S^2(r)/S^2(s)}} \end{aligned} \tag{28}$$

which, with the assumption of normally distributed x_i, is distributed as Student's t with $n - 2$ degrees of freedom when the population correlation coefficient $\rho(yv) = 0$.

We note that the population correlation of y and v is

$$\rho(yv) = (\sigma_{e1}^2 - \sigma_{e2}^2)/[(4\sigma_x^2 + \sigma_{e1}^2 + \sigma_{e2}^2)(\sigma_{e1}^2 + \sigma_{e2}^2)]^{\frac{1}{2}}$$

which is zero if and only if $\sigma_{e1} = \sigma_{e2}$. Thus, the t-statistic of (28) is precisely a test of whether $\sigma_{e1} = \sigma_{e2}$ at the α probability level if the observed $t = t_0$ is such that

$$-t_{1-\alpha/2} \le t_0 \le t_{1-\alpha/2}$$

where $t_{1-\alpha/2}$ is the upper $\alpha/2$ probability level of t with $n - 2$ d.f.

We accept the hypothesis $\sigma_{e1} < \sigma_{e2}$ in a one-sided test at the α probability level if the observed

$$t_0 < -t_{1-\alpha}$$

or we accept $\sigma_{e1} > \sigma_{e2}$ at the α probability level if the observed $t_0 > t_{1-\alpha}$.

A more powerful or sensitive test for comparing the imprecisions of the two standards, i.e. whether $\sigma_{e1} = \sigma_{e2}$, may be obtained by using the data from all three instruments and applying the Pitman–Morgan type test to the correlated variates w_{1i} and z_{1i} of (2), so that the characteristic measured, x_i, is not involved. In fact, looking at (12) we now have

$$\lambda = \sigma^2(w_1)/\sigma^2(z_1) = (\sigma_{e2}^2 + \sigma_{e3}^2)/(\sigma_{e1}^2 + \sigma_{e3}^2)$$

so that a Pitman–Morgan test of whether $\sigma_{e1} = \sigma_{e2}$ is obtained by putting $\lambda = 1$ in the formula

$$t(n - 2, \sigma_{e1} = \sigma_{e2}) = \frac{[S^2(w_1)/S^2(z_1) - \lambda]\sqrt{n-2}}{[4\lambda(1 - r^2\{w_1z_1\})S^2(w_1)/S^2(z_1)]^{\frac{1}{2}}} \tag{29}$$

By looking at the development of the power of this test by Morgan (1939), and especially his formula (24), it can be shown through some algebraic manipulation that (29) is a more powerful test than (28) as would be expected.

Maloney and Rastogi (1970) develop an approximate test of whether $\sigma_{e1} = 0$, or whether $\sigma_{e2} = 0$. To test whether $\sigma_{e1} = 0$, they use

$$-2 \ln \lambda = -2 \ln [(S_r^2 S_s^2 - S_{rs}^2)/(S_r^2\{S_r^2 + S_s^2 - 2S_{rs}\})]^{n/2} \tag{30}$$

where S_r^2, S_s^2 are the ordinary sample variances and S_{rs} the sample covariance of r_i and s_i based on $n - 1$ d.f. The quantity $-2 \ln \lambda$ for sufficiently large n is distributed as Chi-square with 1 d.f. as shown by S. S. Wilks. We therefore reject the hypothesis of $\sigma_{e1} = 0$ if the quantity $-2 \ln \lambda > \chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}(1)$ is the upper α probability level of the Chi-square distribution with 1 d.f.

To test whether $\sigma_{e2} = 0$, we simply interchange the r_i and s_i.

A t-test for the hypothesis that $\beta_1 = \beta_2$, or that the levels of measurement for or the biases of instruments 1 and 2 agree, may be found by computing (from the v_i) the observed

$$t_0(n - 1, \beta_1 = \beta_2) = \bar{v}_1\sqrt{n}/S(v_1) \tag{31}$$

Accept $\beta_1 < \beta_2$, $\beta_1 = \beta_2$ or $\beta_1 > \beta_2$ (in individual significance tests) at the α probability level, depending respectively on whether $t_0 < t_{1-\alpha}$, $-t_{1-\alpha/2} \le t_0 \le t_{1-\alpha/2}$, or $t_0 > t_{1-\alpha}$.

B. *Three-Instrument Case*

First, and as a routine for three simultaneous or common measurements, we might well use formulas (28) or (29) and (31) to check the performance of instruments 1 and 2 before proceeding with tests on instrument 3. Indeed, it may be desirable to calibrate instrument 1 or 2, or both. Otherwise, we proceed to test instrument 3 as follows:

B1. Test Procedure When Sizes of σ_{e1} and σ_{e2} are Unknown.

In case it is not known whether $\sigma_{e1} = \sigma_{e2}$, and the relative sizes therefore of σ_{e1} and σ_{e2} are unknown, then use $t_i - (r_i + s_i)/2$ and $r_i - s_i$ to compute (9), (10), and (11). Then use the Pitman–Morgan (1939) test for correlated variances, Formula (12) above, based on the observed value of $t = t_0$ for $n - 2$ d.f. and $\lambda = \frac{3}{4}$, i.e., we test the hypothesis $\sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2$ by computing

$$t_0[n - 2, \sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2] = \frac{[S^2(u_1)/S^2(v_1) - 0.75]\sqrt{n-2}}{[3(1 - r^2)\{S^2(u_1)/S^2(v_1)\}]^{\frac{1}{2}}} \tag{32}$$

On the basis of this observed value, which is referred to a table of percentage points of Student's t, we decide in individual tests whether

$$\sigma_{e3}^2 < (\sigma_{e1}^2 + \sigma_{e2}^2)/2, \qquad \sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2, \quad \text{or} \quad \sigma_{e3}^2 > (\sigma_{e1}^2 + \sigma_{e2}^2)/2$$

at the α probability level, depending respectively on whether $t_0(n - 2) < t_{1-\alpha}$, $-t_{1-\alpha/2} < t_0(n - 2) \le t_{1-\alpha/2}$, or $t_0(n - 2) > t_{1-\alpha}$. We remark that (12) could be used to find confidence bounds on λ or even on the quantity $2\sigma_{e3}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)$, or its square root.

For judging the level of measurement of instrument 3 as compared to the average level of instruments 1 and 2 for this case, i.e. the size of β_3 versus $(\beta_1 + \beta_2)/2$, then compute the observed value of t for $n - 1$ d.f. given by

$$t_0[n - 1, \beta_3 = (\beta_1 + \beta_2)/2] = \bar{u}_1\sqrt{n}/S(u_1) \tag{33}$$

We conclude in individual tests at the α probability level that

$$\beta_3 < (\beta_1 + \beta_2)/2, \quad \beta_3 = (\beta_1 + \beta_2)/2, \quad \text{or} \quad \beta_3 > (\beta_1 + \beta_2)/2,$$

according respectively to whether $t_0(n - 1) < t_{1-\alpha}$, $-t_{1-\alpha/2} \le t_0(n - 1) \le t_{1-\alpha/2}$ or $t_0(n - 1) > t_{1-\alpha}$. We could then make corrections to or calibrate instrument 3 accordingly.

B2. Test Procedure When $\sigma_{e1} = \sigma_{e2}$

If it should happen $\sigma_{e1} = \sigma_{e2}$, then use the quantities (5) and (6), i.e. $t_i - (r_i + s_i)/2$ and $r_i - s_i$, to compute $S^2(u_1)$ and $S^2(v_1)$ from (9) and (10). Then compute

$$F_0(n - 1, n - 1, \sigma_{e3} = \sigma_{e1}) = 4S^2(u_1)/3S^2(v_1) \qquad \text{(Hahn and Nelson)} \tag{34}$$

which allows for possible differences in levels of measurement for instruments 1 and 2, and refer F_0 to the table of percentage points of F. In individual tests we then accept, remembering $(\sigma_{e1}^2 + \sigma_{e2}^2)/2 = \sigma_{e1}^2$, that

$$\sigma_{e3}^2 < (\sigma_{e1}^2 + \sigma_{e2}^2)/2, \qquad \sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2, \quad \text{or} \quad \sigma_{e3}^2 > (\sigma_{e1}^2 + \sigma_{e2}^2)/2$$

at the α level of probability depending respectively on whether we find $F_0 < 1/F_{1-\alpha}(n - 1, n - 1)$, $1/F_{1-\alpha/2}(n - 1, n - 1) \le F_0 \le F_{1-\alpha/2}(n - 1, n - 1)$, or $F_0 > F_{1-\alpha}(n - 1, n - 1)$.

For a t-test when $\sigma_{e1} = \sigma_{e2}$ concerning the level of measurement of instrument 3, or that is the relative sizes of β_3 and $(\beta_1 + \beta_2)/2$, then compute the observed t from

$$t_0[n - 1, \beta_3 = (\beta_1 + \beta_2)/2] = \bar{u}_1\sqrt{n}/S(u_1) \tag{35}$$

and in individual tests accept

$$\beta_3 < (\beta_1 + \beta_2)/2, \quad \beta_3 = (\beta_1 + \beta_2)/2 \quad \text{or} \quad \beta_3 > (\beta_1 + \beta_2)/2$$

at the α probability level, depending on whether $t_0 < t_{1-\alpha}$, $-t_{1-\alpha/2} \le t_0 \le t_{1-\alpha/2}$, or $t_0 > t_{1-\alpha}$. Thus, instrument 3, if not measuring at the proper level, could be calibrated if its imprecision of measurement, σ_{e3}, is suitably small.

We note that σ_{e3} may be estimated from

$$\hat{\sigma}_{e3}^2 = S^2(u_1) - S^2(v_1)/4$$

if this quantity is positive and otherwise $\hat{\sigma}_{e3}$ is taken as equal to zero.

Hahn and Nelson (1970) provide also one-sided lower and upper $1 - \alpha$ confidence bounds on

$$\sigma_{e3}^2/[(\sigma_{e1}^2 + \sigma_{e2}^2)/2] = \sigma_{e3}^2/\sigma_{e1}^2$$

which may be obtained respectively from

$$\frac{2S^2(u_1)}{F_{1-\alpha}(n - 1, n - 1)S^2(v_1)} - \frac{1}{2} \tag{36}$$

and

$$2F_{1-\alpha}(n - 1, n - 1)S^2(u_1)/S^2(v_1) - \tfrac{1}{2} \tag{37}$$

B3. Test Procedure When σ_{e1} and σ_{e2} are Known But Unequal

Finally, if $\sigma_{e1} \neq \sigma_{e2}$, and we know their relative sizes, $\sigma_{e1} = k\sigma_{e2}$ say, from experience, then use (16), (17), (18) and (19) to compute the observed value of F in (25), i.e.

$$F_0[n-1, n-1, \sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2] = S^2(u_2)/S^2(v_2)[\tfrac{1}{2} + \sigma_{e1}^2\sigma_{e2}^2/(\sigma_{e1}^2 + \sigma_{e2}^2)^2] \quad (38)$$

Refer this observed F_0 to the table of percentage points of Snedecor's F-distribution to decide in individual tests if

$$\sigma_{e3}^2 < (\sigma_{e1}^2 + \sigma_{e2}^2)/2, \qquad \sigma_{e3}^2 = (\sigma_{e1}^2 + \sigma_{e2}^2)/2 \quad \text{or} \quad \sigma_{e3}^2 > (\sigma_{e1}^2 + \sigma_{e2}^2)/2,$$

at the α probability level, depending respectively on whether

$F_0 < 1/F_{1-\alpha}(n-1, n-1)$, $1/F_{1-\alpha/2}(n-1, n-1) \leq F_0 \leq F_{1-\alpha/2}(n-1, n-1)$
or

$$F_0 > F_{1-\alpha}(n-1, n-1)$$

Concerning tests on the bias β_3 of instrument 3, we note that

$$E(\bar{u}_2) = \beta_3 - (k^2\beta_1 + \beta_2)/(1 + k^2)$$

since we assumed that $\sigma_{e1} = k\sigma_{e2}$. Thus, to find whether

$$\beta_3 <, =, \text{ or } > (k^2\beta_1 + \beta_2)/(1 + k^2)$$

in individual tests, we compute the observed t from

$$t_0[n-1, \beta_3 = (\beta_1 + \beta_2)/2] = \bar{u}_2\sqrt{n}/S(u_2)$$

and decide on the above order respectively according to whether in single tests

$$t_0 < -t_\alpha, \quad -t_{1-\alpha/2} \leq t_0 \leq t_{1-\alpha/2}, \quad \text{or} \quad t_0 > t_{1-\alpha}.$$

For the three-instrument case, it should be noted that significance tests depend on only the errors of measurement e_{ij} being normally distributed, whereas the x_i represent the common or same true values measured by each of the three instruments.

Since there is an interest in comparisons of specific measuring instrument precisions, a referee has suggested that an F-test on $\hat{\sigma}_{e1}^2/\hat{\sigma}_{e2}^2$, etc., for three or more instruments might be developed using Satterthwaite's approximation to distributions. Also, as another referee so aptly pointed out, it might be easy to generalize the results to any number of measuring instruments for the case where all instruments are on an equal footing, since simultaneous comparisons may be obtained from multivariate statistical theory. Indeed, the significant paper of Hanumara and Thompson (1968) will be very valuable for placing simultaneous confidence bounds on the product variability σ_x and the imprecisions of measurement, σ_{e1}, σ_{e2}, σ_{e3}, etc. For example, for the data of Table I, we find from the results of Hanumara and Thompson that 95% simultaneous confidence intervals for the product variability and the imprecisions of measurement are given by

$$.77 \leq \sigma_x \leq 3.57 \ m/s \qquad .00 \leq \sigma_{e1} \leq 1.22 \ m/s$$
$$.00 \leq \sigma_{e2} \leq .92 \qquad .00 \leq \sigma_{e3} \leq 1.98$$

Note how seemingly wide the confidence bounds on the imprecisions of measurement appear to be for only the 12 rounds. The work of Thompson (1963) for the two-instrument case to test hypotheses such as $\sigma_x/\sigma_{e1} = \Delta$, etc., may also have some practical applications.

Our purpose here has been to use an approach which will pinpoint the instruments which have poor precision of measurement and inaccuracy or bias, so that specific actions may be taken in the measuring process. Also, we desire to point out that for important measurements it may well be desirable to use more than a single reference instrument or standard, and that the analyses given herein when made as a routine give an account of the actual performance of instruments for the test under consideration. In particular, it might be wise to continually check on the performance of standard instruments, as they might perhaps become imprecise or require calibrations as we have seen in our example. We believe that these principles may be especially important in accurate determination of fundamental physical constants such as the velocity of light or radio wave propagation.

6. Acknowledgment

The author expresses his sincere appreciation to the Editor and referees for excellent suggestions on preparation of the paper.

References

[1] Grubbs, F. E. (1948). On Estimating Precision of Measuring Instruments and Product Variability, *J. Amer. Statist. Assoc. 43*, 243–264.
[2] Hahn, J. H. and Nelson, W. (1970). A Problem in the Statistical Comparison of Measuring Devices. *Technometrics, 12, 1*, 95–102.
[3] Hanumara, R. C. and Thompson, W. A., Jr. (1968). Percentage Points of the Extreme Roots of a Wishart Matrix, *Biometrika, 55*, 505–512.
[4] Jaech, J. L. (1971). Further Tests of Significance for Grubbs' Estimators. *Biometrics, 27*, 1097–1101.
[5] Maloney, C. J. and Rastogi, S. C. (1970). Significance Test for Grubbs' Estimators. *Biometrics, 26*, 671–676.
[6] Morgan, W. A. (1939). A Test for the Significance of the Difference Between Two Variances in a Sample from a Normal Bivariate Population, *Biometrika, 31*, 13–19.
[7] Pitman, E. J. G. (1939). A Note on Normal Correlation. *Biometrika, 31*, 9–12.
[8] Thompson, W. A. (1963). Precision of Simultaneous Measurement Procedures, *J. Amer. Statist. Assoc. 58*, 474–479.